Statistical Models to Analyse a Baseline Survey on Rural KwaZulu-Natal Adults' HIV Prevalence and Associated Risk Factors

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and not necessarily to be attributed to the NRF.

Kameshan Moodley

December, 2020

Statistical Models to Analyse a Baseline Survey on Rural KwaZulu-Natal Adults' HIV Prevalence and Associated Risk Factors

by

Kameshan Moodley

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

Thesis Supervisor: Professor T.T. Zewotir

Thesis Co-supervisor: Ms. D.J. Roberts

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

Declaration - Plagiarism

I, Kameshan Moodley, declare that

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

Kameshan Moodley (Student)	Date
Professor T.T. Zewotir (Supervisor)	Date
Ms. D.J. Roberts (Co-supervisor)	Date

Disclaimer

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

Abstract

South Africa is at the global epicentre of the HIV-AIDS pandemic. Though there has been an increase in prevention and control measures that has led to a significant reduction in HIV-AIDS mortality rates globally, South Africa has experienced a high share of the HIV burden. HIV-AIDS imposes a substantial economic burden on both individuals and governments. It has had a considerable effect on poverty by affecting potentially economically active citizens who would otherwise have entered the workforce and contributed to the local and national economy. This has hindered economic growth and development in South Africa. The 2016 UNAIDS Gap Report estimates that in 2015 there were seven million people living with HIV in South Africa and that this resulted in 180,000 AIDS related deaths in the same year. The same year saw an unprecedented 380,000 new reported infections. The prevalence of HIV-AIDS in South Africa remains high at 19.2% among the general population. This study was an investigation into the determinants of HIV in adults in the age group 15-49 years. The study used the HIV Incidence Provincial Surveillance System (HIPSS) to collect data between June 2014 and June 2015. The final data set comprised 9,804 observations and consisted of explanatory variables pertaining to individuals' socio-economic, socio-demographic and behavioural circumstances. The response variable was binary indicating whether a participant tested positive or negative for HIV. Incorporating survey weights into the data owing to the complex sample design, necessitated the use multilevel regression procedures. To this end, survey logistic regression and the generalised linear mixed models were employed. The results emanating from these models revealed that factors encompassing socio-

economic, demographic and selected behavioural characteristics were significantly associated with HIV prevalence in the study location. In some instances, it is possible that households in close proximity exhibit some similarities with the inevitable result of spatial autocorrelation requiring the use of geographically weighted regression techniques able to account for spatial autocorrelation. The application of a spatial multilevel model showed that the influence between households in close proximity is greater than between those further away, a phenomenon that would be ignored in conventional multilevel models.

Acknowledgements

1. Firstly, I wish to thank the Lord Almighty for granting me the courage, strength, fortitude and willpower to complete this research.

2. My paternal grandmother, Mrs. D. Gengan whose keen interest in my studies and constant encouragement fueled its completion.

3. My parents Mr. M. Moodley and Mrs. S. Moodley and my brother Luvashan for their unwavering support and encouragement.

4. My supervisor, Professor T.T. Zewotir and co-supervisor Mrs. D.J. Roberts, for their professional advice and guidance when sought.

5. The Health Economics and HIV and AIDS Research Division (HEARD), of the University of KwaZulu-Natal (UKZN) for the provision of the HIV Incidence Provincial Surveillance System (HIPSS) data set.

6. The National Research Foundation (NRF) for their generous financial assistance towards the completion of this research.

Contents

List of Figures

List of Tables

Abbreviations

AGQ	Adaptive Gaussian Quadrature
AIC	Akaike Information Criterion
AIDS	Acquired Immune Deficiency Syndrome
BIC	Bayes Information Criterion
CA	Correspondence Analysis
CoV	Coronavirus
COVID-19	Coronavirus Disease - 2019
CI	Confidence Interval
CPI	Consumer Price Index
EA	Enumeration Area
GHQ	Gaussian Hermite Quadrature
GIS	Geographic Information System
GLMM	Generalized Linear Mixed Model
HIV	Human Immunodeficiency Virus
HIPSS	HIV Incidence Provincial Surveillance System
IoMD	Indices of Multiple Deprivation
MCA	Multiple Correspondence Analysis
NHI	National Health Insurance
OECD	Organisation for Economic Cooperation and Development
PHF	Primary Healthcare Facility
PLWHIV	People Living with HIV
PSU	Primary Sampling Unit
SARS	Severe Acute Respiratory Syndrome
SGLMM	Spatial Generalized Linear Mixed Model
SLR	Survey Logistic Regression
SSU	Secondary Sampling Unit
STI	Sexually Transmitted Infection
TSU	Tertiary Sampling Unit

Chapter 1

Introduction

1.1 Background

Commencing in the early 1980s the world has been gripped by the unfolding AIDS pandemic of which South Africa has borne the heaviest share. This makes South Africa an epicentre; an all-important site to examine HIV over the individual's lifespan (Houle et al., 2018). According to Statistics South Africa (2020), approximately 13%, nearly eight million South Africans, were living with HIV-AIDS in 2019. Of particular concern are individuals aged 15-49 years. Statistics South Africa (2020) estimates that these individuals constitute 18.7% of the total population of people living with HIV (PLWHIV) in South Africa. However, while they constitute the majority of HIV positive individuals, Rosenberg et al. (2017) found that advancing age increases the risk of HIV infection. A further area of noticeable HIV infection and incidence is among females in South Africa and sub-Saharan Africa at large; the statistics are usually observed to be gendered as noted by Gregson & Garnett (2000), Glynn et al. (2001) and MacPhail et al. (2002).

A prominent feature of early research into HIV-AIDS was that most research focused on understanding the clinical aspects surrounding HIV-AIDS infection. While society has benefited from the cornucopia of research produced in this regard, there

has been a growing need to understand HIV-AIDS infection rates among populations from a socio-economic, socio-demographic and psycho-social perspective. This is supported by evidence from a study conducted by Probst et al. (2016) which shows that the HIV mortality rate among people living in low socio-economic conditions exceeded the rate of people in higher socio-economic conditions by more than half. This can in part be attributed to the high economic and financial burdens placed on households by medical expenditures required for an HIV-AIDS-symptomatic patient, which Steinberg et al. (2002) found accounts for more than a third of monthly income. This gives impetus to the assertion by (Fenton, 2004, p. 2) that *"reducing poverty will be at the core of a long term sustainable solution to HIV-AIDS."*

An intrinsic driver of socio-economic inequality is unequal access to education primarily affecting those of lower socio-economic status. It thus stands to reason that advancing academically is vital in maintaining a low risk of HIV infection which, Bärnighausen et al. (2007) describes as a protective effect. This is evidenced by a 7% reduction in HIV incidence attributed to an additional year of formal education in KwaZulu-Natal. This was a departure from the findings of Hargreaves & Glynn (2002) in which the link between advanced academic attainment and HIV prevention was acknowledged, although the study stopped short of agreeing that this was a universal occurrence. The study appeared to agree with, though with qualification, the findings of Pettifor et al. (2005) which, without deference to academic qualifications concluded that this was true among young adults. Furthermore, while Bärnighausen et al. (2007) acknowledges the role of educational attainment in reducing the risk of HIV infection, there is little support for the sentiments of Fenton (2004) regarding poverty reduction and its link with HIV incidence reduction.

Socio-economic status, while still a main driver of HIV infection, does not contribute independently to the risk of HIV infection. Studies sometimes theorise that behavioural and cognitive factors are deterministic in HIV infection. These factors

extend, but are not limited to perception of HIV infection, the use of contraceptions to prevent HIV infection, knowledge of HIV transmission and prevention, and espousal of HIV stigma. Simbayi et al. (2019) found no gender disparity in knowledge about HIV transmission but recognise that younger individuals are more cognisant of HIV transmission than their older counterparts. In addition, among individuals surveyed there was almost unanimous rejection of HIV stigma harboured against HIV positive persons. Furthermore, although there is evidence of an upscale in knowledge acquisition among high-risk groups, Simbayi et al. (2019) note that this was not commensurate with the level of accurate knowledge that people should possess.

What is striking though unsurprising, is that like socio-economic status, behavioural and cognitive factors are seemingly not seen to exist independently. Booysen (2004) concludes that risky sexual behaviour is gendered but with a dual caveat. The study observed that females in the higher socio-economic echelons usually reject the use of condoms while women who in lower socio-economic conditions also reject the use of condoms but because of a lack of knowledge. This shows the clear link in the extremities between socio-economic status, lack of education, risky sexual behaviour, and their significance in predicting HIV infection. Sexually transmitted infections can in some ways be a consequence of risky sexual behaviour. Sexually transmitted infections and HIV incidence share a complex and bi-directional relationship according to Kharsany et al. (2020) in which the high association between an STI diagnosis and HIV prevalence is delineated. The intense burden that this clinical diagnosis places on a public healthcare system leads the authors to advocate the stated goals of the United Nations Programme on HIV/AIDS (2017) that HIV prevention programms incorporate early diagnosis and treatment of STIs in an effort to subvert the spread of HIV. These sentiments were also mooted by (Galvani et al., 2018).

Perceived susceptibility to HIV infection attempts to gauge respondents on how much at risk of infection they believe themselves to be. The perception of HIV risk

often determines an individual's conduct in relation to HIV prevention. Manjengwa et al. (2019) found that, *inter alia*, a low perceived susceptibility leads to eschewing contraception; a finding that is in accord with the results of Muchiri et al. (2017). In addition, Manjengwa et al. (2019) found that eschewing contraception as a result of a low perception to the risk of HIV infection could largely be attributed to individuals who seek to avoid the stigma attached to HIV infection.

1.2 HIV Incidence Provincial Surveillance System (HIPSS)

The results presented herein arise from an analysis of the data from the HIV Incidence Provincial Surveillance System (HIPSS) investigation. An HIPSS study was conducted in two sub-districts of Vulindlela and Greater Edendale in the uMgungundlovu municipality of the KwaZulu-Natal Province in South Africa.

The stated purpose of the HIPSS study was to initiate population-level HIV incidence in cohorts in these two sub-districts to examine HIV incidence in conjunction with the upscaling of prevention efforts implemented in a "real-world", non-trial setting. Two sequential cross-sectional surveys, conducted one year apart and comprising approximately 10,000 participants, consisting of both male and female participants, were selected at random in the age group 15-49 years.

This study focuses on the baseline data of the HIPSS study in which participants were surveyed between June 11, 2014 and June 22, 2015. Measurements that inform the baseline survey were collected by means of structured questionnaires and biological specimens. In addition 6,400 HIV uninfected participants in the 15-35 year age bracket were selected from a representative sample of households for a longitudinal follow-up study conducted twelve months after the first assessment for HIV infection. This was done to measure HIV incidence at the population level. The HIPSS study made further provision for the evaluation of laboratory tests of recent infections (TRIs) using a recent infection testing algorithm (RITA).

1.3 The Study Setting

The uMgungundlovu District Municipality is a conurbation located in the central region of the KwaZulu-Natal (KZN) Province of the Republic of South Africa (Figure 1.1). The Municipality comprises a mixture of urban and rural areas, from small towns such as Mooi River, Richmond and Impendle to much larger urban metropolises such as Howick and Pietermaritzburg. The latter serves as the provincial capital. The uMgungundlovu District Municipality encompasses a diverse range of human settlements ranging from traditional homesteads and farmlands to informal settlements coupled with a mix of urban and rural dwellings (Kharsany et al., 2015).

The uMgungundlovu Health District was selected as pilot district for the National Health Insurance (NHI) and consists of forty-six fixed clinics, seventeen mobile clinics, and one state-aided clinic. These healthcare facilities cater to 1,052,730 residents which is approximately 10% of the population of KwaZulu-Natal. A strengthened healthcare system is considered vital in uMgungundlovu where in 2016, a fifth of the population were HIV positive (uMgungundlovu District Municipality, 2020). Along with the eThekwini Metropolitan Municipality where the HIV prevalence was 16.8%, these two districts were among those with the highest prevalence in South Africa (George et al., 2020).

The uMgungundlovu District Municipality endures extreme climatic conditions where flooding, storms, droughts, heatwaves and land fires are common occurrences. These climatic extremities are attributed to the impact of climate change in the area and disproportionately affects the destitute and vulnerable population. In recent years these conditions have shown no sign of subsiding as the Municipality, which lies inland in KZN, has seen a rise in temperatures which has delayed summer rains thus resulting in flash flooding. Dry weather conditions also have an adverse effect on local agriculture thus impacting on local farmers (uMgungundlovu District Municipality, 2020).

An HIV Incidence Provincial Surveillance System (HIPSS), as already mentioned, was established in two sub-districts of the uMgungundlovu District Municipality, namely the Vulindlela and Greater Edendale sub-districts. Vulindlela is a chiefly rural area situated within the uMsunduzi and uMgeni municipal boundaries. The majority of the land is tribal, held in trust by traditional authorities, and ruled by local chieftains while the remaining land is governed by directly elected local councils. The population of Vulindlela exceeds 150,000 residents who communicate primarily in isiZulu.

The second sub-district under study is the Greater Edendale area which is the centre of economic activity in the uMgungundlovu municipality and is situated south-west of the uMsunduzi city centre. A dual carriage-way, which links the uMsunduzi municipality with Greater Edendale and other areas, facilitates the movement of goods and services - providing a channel for investment and growth in the area. Similarly, the Greater Edendale location consists of two constituent areas which are in part traditionally-owned land and partly administered by the KZN provincial government and the South African government. The Greater Edendale area has a population of 210,000 which is roughly 36% of the city's population residing in densely developed areas which comprise both formal and informal housing.

In the Vulindlela and Greater Edendale sub-districts there are seven and nine primary healthcare facilities (PHF) respectively. In these PHFs trained healthcare professionals provide a wide array of primary health care services such as family planning, voluntary HIV counselling and testing, treatment of sexually transmitted diseases, antenatal care, treatment of opportunistic infections, and other minor ailments. First responders are employed to link these PHFs with regional referral hospitals which provide tertiary care to patients. Grey's and Edendale hospitals are the primary referral hospitals for these PHFs as they are able to provide more comprehensive care.

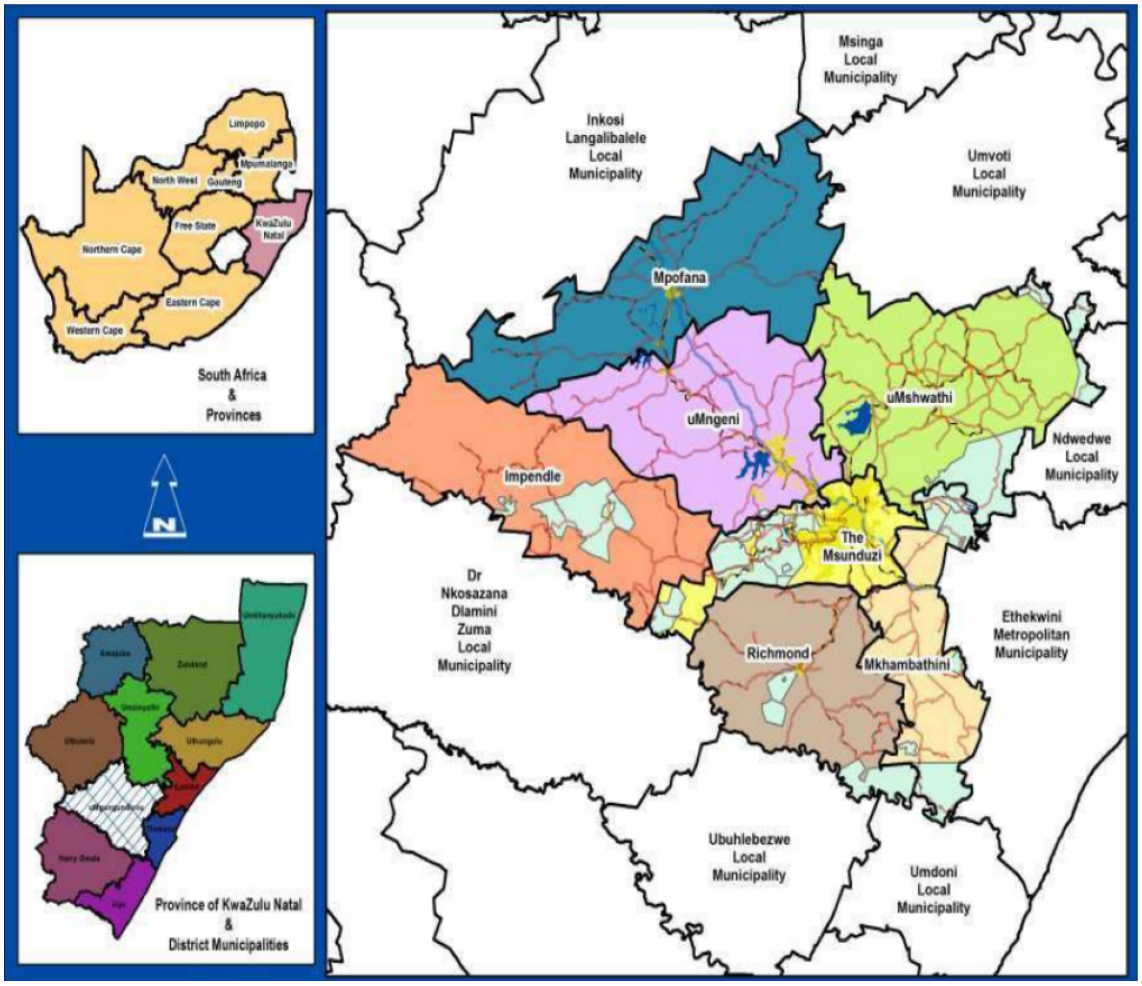

Figure 1.1: Map of the uMgungundlovu District in the KwaZulu-Natal Province of The Republic of South Africa (uMgungundlovu District Municipality, 2020)

1.4 Sampling Procedure and Data Collection Methods

The two sub-districts, Vulindlela and Greater Edendale, served as the strata while the enumeration area (EA) was the primary sampling unit. The secondary and tertiary sampling units were the household and the eligible individual respectively. The Vulindlela and Greater Edendale areas comprise a total of 591 enumeration areas of which 221 were selected. Using systematic sampling and following a serpentine pattern, 50 households were selected in each enumeration area.

Once a household was selected, the study staff, using a global positioning system (GPS) receiver, recorded the geographical coordinates of the household. This pro-

cedure continued until the stopping criteria of the requisite number of households were selected such that approximately 10,000 individuals were selected. For a household that could not be selected owing to residents being absent for a protracted time, or refusal to participate or the household was abandoned, the household to the right of the main entrance was selected. In each selected household, one individual was selected provided they fulfilled the age eligibility criterion.

1.4.1 Data Collection

In a selected household, the household head or a resident designated as such, was identified and presented with relevant information pertaining to the study after which verbal consent was sought. On completion of all the household procedures and enumeration of household members, a personal digital assistant (PDA) randomly selected a household resident in the 15-49 year age group irrespective of gender. Informed consent was sought from the participants who were selected and who were 18 years and older while participants aged 15 - <18 were required to provide assent plus parental consent. In the absence of a parent, consent was sought from an *in loco parentis* as per the Children's Amendment Act of South Africa (2007). Within the household, two questionnaires were administered; a household questionnaire and an individual questionnaire. The information collected in each questionnaire is detailed below.

1.4.2 Household Questionnaire

Field workers administered a structured questionnaire to the head or designated head of the household. These questionnaires collected information pertaining to the household socio-economic circumstance, access to reliable water supply, electricity and sanitation facilities. In addition, the household head was questioned about the household income, household food security and residential access to health service. In addition, socio-demographic information such as age, gender, highest level of education, employment status and access to social support grants was obtained from each household member.

1.4.3 Individual Questionnaire

For information pertaining to the individual, confidentiality was strictly maintained and no personal information was made known. Each participant was assigned a unique number which linked their questionnaires and biological samples. The questionnaire administered obtained a range of socio-demographic and clinical information such as age, sex, marital status, employment and educational status; psychosocial information including knowledge of HIV transmission and infection and motivational issues, social norms related to sexual risk behaviours; behavioural information including number of sex partners, condom use, knowledge of HIV status of own and sex partner(s), questions about HIV testing history including date of last HIV test, HIV results, current HIV treatment and medical male circumcision (MMC) status.

1.5 Ethical Considerations

The study protocol, informed consent and data collection forms were reviewed and approved by the University of KwaZulu-Natal (UKZN) Biomedical Research Ethics Committee (BF269/13), the Associate Director of Science of the Center for Global Health (CGH) at the United States Centers for Disease Control and Prevention (CDC) in Atlanta, and the Department of Health in the Province of KwaZulu-Natal (HRKM 08/14).

1.6 Thesis Objectives

This study aimed at producing a concise set of statistical models to effect understanding of HIV prevalence, particularly in a high risk area of KwaZulu-Natal (KZN). The specific objectives of this dissertation were to:

- investigate the prevalence of HIV in adults in a high risk area of the KwaZulu-Natal Province, and

- investigate the associated risk factors of HIV in adults in a high risk area of the KwaZulu-Natal Province.

1.7 Thesis Structure

Chapter 1 provides an introduction to the thesis and describes the preliminary considerations and ethical procedures governing the collection of data. Chapter 2 details the variables of interest and conducts data explorations to infer the intricate relationships of and between the variables under study.

This is followed by a process of statistical analysis in Chapter 3 to model the participants' HIV status. The method employed is a survey logistic regression model, which accommodates a binary response variable in the presence of sampling weights implemented as a consequence of multistage sampling. Chapter 4 offers a generalised linear mixed model to explore the addition of a random component based on a partition of the data set according to clusters, arguing that households within the same cluster exhibit similarities.

In Chapter 5 a spatially weighted generalised linear mixed model is applied to account for spatial autocorrelation. An *a priori* consideration is to detect the presence of spatial autocorrelation before accounting for the covariance structure in the generalised linear mixed model. Chapter 6, concludes the thesis by reflecting on the results obtained and makes recommendations for future studies.

Chapter 2

Weighted Exploratory Data Analysis

The importance of exploratory data analysis cannot be underestimated in the realm of statistical analysis. Exploratory data analysis is a well-established statistical tradition that provides valuable insight using computational tools allowing researchers to discover patterns which aid hypothesis development and refinement (Behrens, 1997). The development of efficient, widely available and user-friendly statistical software has increased the visibility of exploratory data analysis thus complementing confirmatory data analysis.

Jean-Paul Benzécri, a French statistician and linguist whose work in correspondence analysis is discussed in this chapter, was a noted scholar in one of two schools that Husson et al. (2016) call the French school of thought, the other being the Dutch school. Benzécri, according to (Husson et al., 2016, p. 1) advocated *"letting the data speak for itself"* and promulgated the idea of exploratory data analysis:

> *"The model must follow the data, not the other way around . . . What we need is a rigorous method which extracts structures from the data."*

A popular way of collecting data in the field of public healthcare is via public health

surveillance systems. The Centers for Disease Control & Prevention (2014) define a public health surveillance system as an ongoing and systematic collection, analysis and interpretation of health related data essential to planning, implementation and evaluation of public health practice. A prominent application of surveillance systems is in the field of HIV research. Buthelezi et al. (2016) state that HIV surveillance systems is a public health initiative that allows for the understanding of transmission patterns, gives proper direction of financial resources to vulnerable geospatial locations, and can predict and identify future infection rates.

Thus, one cannot detract from the important role that accurate and reliable information disseminated from survey data can play in the implementation of targeted intervention programmes. Exploratory data analysis assists in inferring the intricate relationships that exist in a large population from those observed in a small sample; a useful precursor to statistical modeling.

This chapter presents an overview of the variables of interest as well as how some of the variables were created using the available data. In addition, the results of the exploratory data analysis and correspondence analysis are presented. The variables used in the exploratory data analysis was weighted to account for the sampling design in which a multistage sampling technique was used to obtain the data, resulting in an unequal probability of selection.

2.1 Variables of Interest

The response variable is binary indicating whether or not a participant aged between 15-49 years tested positive or negative for HIV. The explanatory variables comprise a range of socio-economic, socio-demographic, behavioural and cognitive variables, as shown below:

1. Household deprivation

2. Gender
3. Highest level of education
4. Marital status
5. Knowledge of prevention
6. Perceived risk of HIV
7. Engaged in sexual intercourse
8. HIV stigma
9. Used contraception
10. HIV Information acquisition
11. Diagnosed with STI

2.2 Baseline Household Deprivation Formation

The household index of multiple deprivation (IoMD) is a measure of relative deprivation that encompasses a range of an individual's living standards. Poverty and deprivation are sometimes conflated; however, there is a distinct difference between these two concepts. People are considered to be living in poverty if they lack the financial resources to meet their needs whereas deprivation is considered the lack of any resource, not only income. Indices of multiple deprivation can be used to gauge the distribution of deprivation within the study location thereby providing a relative measure of the socio-economic standing of the study participants.

A household's basic requirements will grow with increasing household composition. However, owing to economies of scale, such growth is not always proportional (OECD, 1998). In this respect, the use of *equivalence scales* have become commonplace in assessing a household's needs alongside changing household size. Whilst

a wide variety of equivalence scales have been in use, EUROSTAT, the statistics authority of the European Union (EU) has adopted the Organisation for Economic Co-operation and Development's (OECD) modified equivalence scale first proposed by Hagenaars et al. (1994).

In conjunction with the OECD modified equivalence scale, the indices of multiple deprivation as discussed in Noble et al. (2006) is adapted in this study to inform the baseline household exploratory data analysis. These indices, which form part of larger domains, are considered to be important in identifying areas of deprivation in a location specific manner. In this respect, the Guttmann scale, a method of cumulative scoring proposed by Guttman (1944), is employed in which responses are coded as one (”1”) for an affirmative response and zero (”0”) for a non-affirmative response. The formation of these indices are detailed below, and Table 2.1 shows the method of scoring used to assess the scale of deprivation.

2.2.1 Formation of the Domain and Indices

Income and Material Deprivation

The OECD modified equivalence scale is determined by the household member's age and overall household composition. The OECD equivalence scale assigns a value of one to the head of the household and a value of 0.5 to each adult member. A value of 0.3 is assigned to each member younger than fourteen years of age in the household. Once each member is designated their appropriate value, the sum of the scores is multiplied against the total household income adjusted for inflation to obtain the OECD modified level of income which reflects what the total household income ought to be in relation to household size.

Adjusting household income for inflation accounts for the prevailing economic trends since the engagement of the respondent. The process of adjusting for inflation involved multiplying household income by an inflation factor which was the ratio of

the average rural consumer price index (CPI) for the period June 2014 to July 2015 to the average CPI for January 2020 to June 2020. The equation to adjust for inflation, as adapted from the United States Census Bureau (2020), can be presented as follows:

$$\text{Adjusted Income} = \text{Income}_{2014/2015} \times \frac{\overline{\text{CPI}}_{2019/2020}}{\overline{\text{CPI}}_{2014/2015}} \tag{2.1}$$

A household was deemed to be income deprived if the OECD modified income level was less than the average household income of all households in the study location. Furthermore, in accordance with the income and material deprivation domain, household ownership of a radio, television, fridge or freezer, and a functional motor vehicle for private use was considered.

Living Standards Deprivation

The focal point in the formation of the living standards domain was the standard of household infrastructure. Individuals residing in an environment lacking access to a reliable water supply or sanitation facility, or being reliant on a primitive energy source, were considered to reside in households deprived of adequate infrastructures for human habitation. Overcrowded households also contribute to the gradual decline of overall household infrastructure. In this respect, households were classified as overcrowded if the household size exceeded the national average household size of 3.3 members. Furthermore, cellular communication is deemed to contribute to an enhanced standard of living, thus household access to a mobile cellular device was also investigated.

Nutritional Deprivation

The Food and Agriculture Organization (1996) states that food security exists when all people, at all times, have physical and economic access to sufficient, safe and nutritious food that meets their dietary needs and preferences for an active life. Food security rests on four pillars, namely food availability, food access, food use, and food stability. When one of these pillars are rendered unstable or non-existent, people may live in a state of food insecurity.

A nutritional deprivation domain was constructed to investigate the level of food security in each household. It was investigated whether the household ever cuts the size of meals owing to food shortage, if the household skips meals owing to food shortage in the past year, and/or if the household eats a smaller variety of food because there is not an adequate food supply. In addition, it was investigated whether the household engaged in subsistence farming and if the household was considered to be below the food poverty line by comparing the bench mark household per capita income against the food poverty line of ZAR 547.00 for the year June 2014 to June 2015 as determined by (Statistics South Africa, 2018a).

Household Financial Security

Access to financial services are vital in attaining economic sustenance. Klasen (2000) argues that poor access to financial services limit a household's ability to sustain varying income streams resulting in financial risk and uncertainty. In addition, a household will often turn to financial institutions to seek assistance in times of financial strain and often default on the repayments. This study investigates whether the household is indebted to a financial institution by ascertaining if the amount owed to financial institutions exceeds the total household income. The study also seeks to investigate whether the household had been bankrupt during the preceding twelve months, if the household had savings in a financial institution, and if any household member attended any courses on financial education.

Constructing the Deciles of Multiple Deprivation

Following the formation of the indices of multiple deprivation, a scoring method was again applied, according to the postulates of Guttman (1944). This involved assigning a value of either one or nil depending on the nature of the responses from the head of the household. Once a score was allocated to each index per domain, the scores were summed and ranked according to their deciles which in turn was used to determine the overall scale of household deprivation. The method used to allocate a score to the indices of the relevant household are detailed in Table 2.1.

2.3 Baseline Household Deprivation Analysis

The HIPSS baseline study was conducted in 224 enumeration areas in the study location enrolling 11,289 households from which one eligible resident per household was enrolled for the study. Owing to participants refusing to participate, statistical sampling errors, and laboratory error, this number was revised down to an enrolled 9,812 households. As a consequence of missing data, eight observations were deleted and analysis was conducted on 9,804 households.

Figure 2.1 shows the distribution of households within each decile of deprivation. The majority of households in the study location are observed as either minorly or significantly deprived; these categories comprised more than 18% of the dwellings. There were no households observed to be moderately or majorly deprived. There were approximately 9% of households in the study area that were observed to be within the 10% of extremely deprived households.

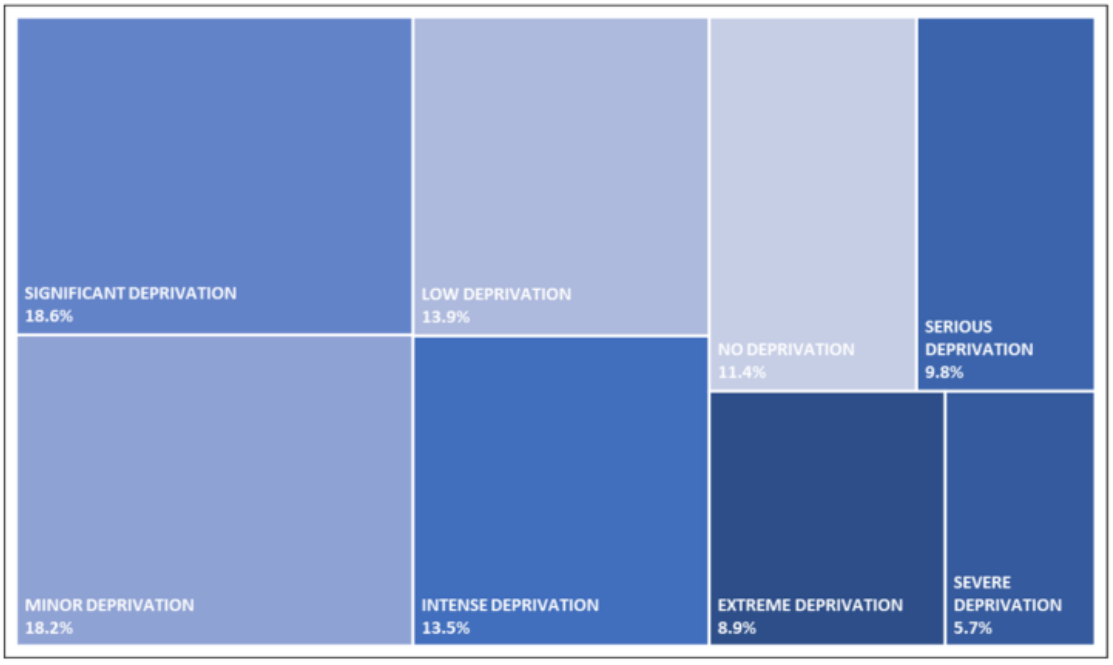

Figure 2.1: Distribution of households according to their intensity of deprivation

Table 2.1: Formation of the Household Indices of Multiple Deprivation (IoMD)

Domain	Indices of Deprivation	Method of Scoring
Income and Material Deprivation	The household is OECD income deprived The Household does not own a television The household does not own a private motor vehicle The household does not own a radio The household does not own a fridge/freezer	Yes=1, No=0
Living Standards Deprivation	The main source of water in the household The type of toilet facility in household The type of electricity source Does the household own a cellphone The household is considered overcrowded	Primitive Infrastructure = 1 Habitable Infrastructure = 0 Yes = 1, No = 0
Nutritional Deprivation	Household cut meal sizes Household skipped meals Household ate smaller food varieties Household does not engage in subsistence farming Household is below the food poverty line	Yes=1, No=0
Financial Security Deprivation	Does the household owe money to a financial institution? Does the household have money saved in a financial institution? Did the household run out of money in the last 12 months? Did any household member attend a financial education class? Is the household in financial debt?	Yes = 1, No = 0

Focusing on each domain individually, it can be observed from Figure 2.2 that as per the indices of multiple deprivation, the overall socio-economic conditions in the study area may not be characterised as dire since most households across each domain experience low levels of household deprivation.

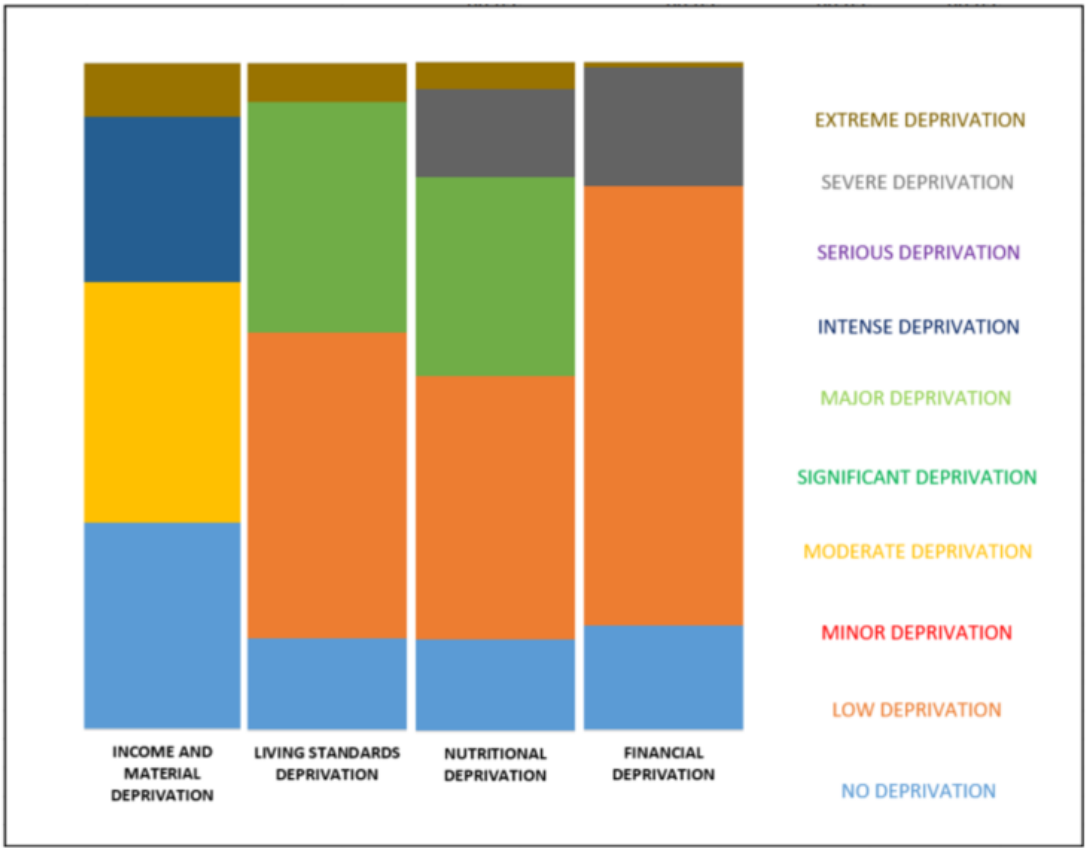

Figure 2.2: The proportional intensity of household deprivation within each domain

However, one cannot ignore the fact that approximately 10% of households were reported to experience extreme levels of income and material deprivation and that one-fifth of households within this domain were categorised as intensely deprived. An investigation into the standard of household infrastructure and access to reliable mobile telephone services revealed that 59.6% of households had habitable infrastructures capable of human habitation together with relative ease of access to reliable cellular communications. However, scarce cases of extreme deprivation were also noted in this domain.

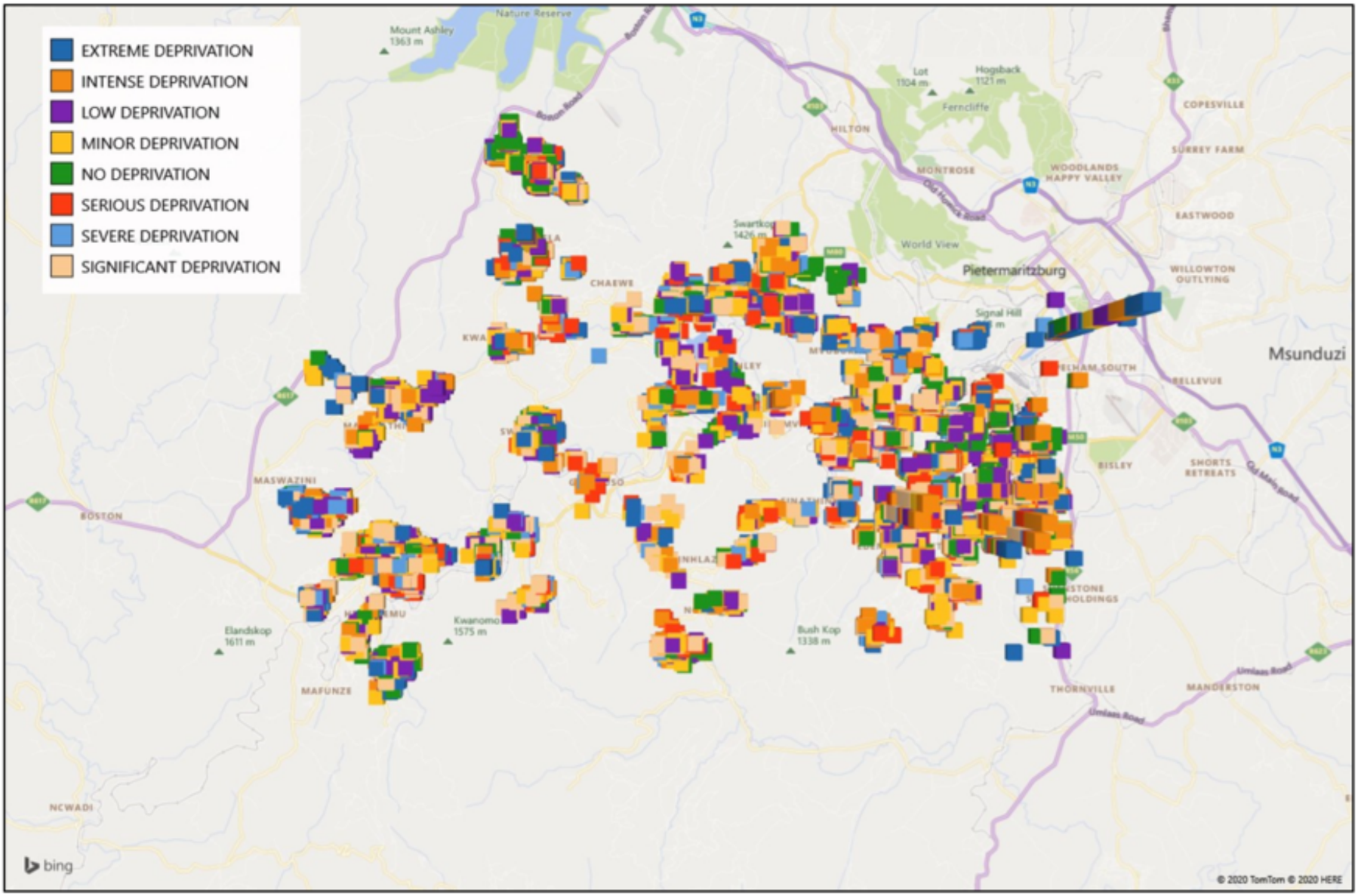

Figure 2.3: Geographical distribution of the intensity of deprivation in the study area.

Assessing household food security and household financial position in the nutritional and financial deprivation domains, a majority of households, 53.2% and 81.4% respectively, were observed to be within the 20% of least deprived households. Furthermore, it was noticed that slightly more than 17% of households within these domains were considered to be within the 20% of most deprived households. As previously stated, the relative deprivation of a household is not solely determined by the breadwinner's earning capacity. If one is to examine the per capita income of households prior to and post adjustment for inflation, the widening wealth gap in the study setting becomes evident. Between July 2014 and June 2015, the average of the poverty line threshold was ZAR 547.00 and between January 2019 and June 2020, the average poverty line threshold was ZAR 810.00 (Statistics South Africa, 2018b).

Adjusting the per capita income for inflation as per Equation 2.1 to obtain present day currency values and bench marking these against the poverty threshold, shows increasing income inequality. Figure 2.4 illustrates the projected widening of the wealth gap of households in the study location.

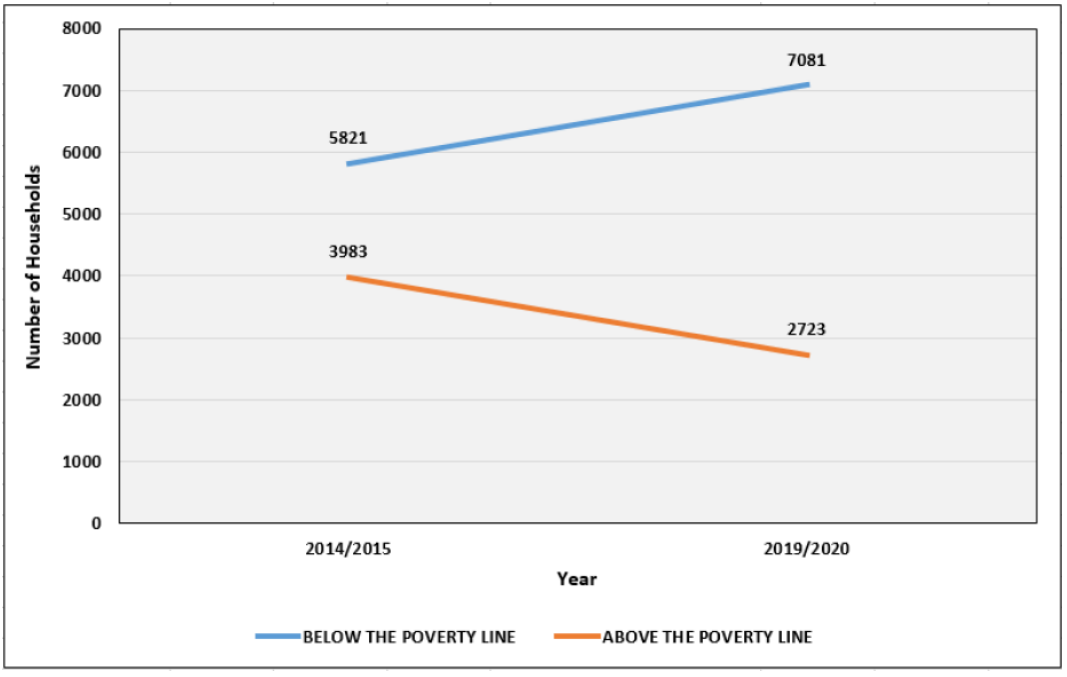

Figure 2.4: Illustration the projected widening of the household wealth gap.

In the year spanning June 2014 to June 2015, 5,821 or 59% of households were below the poverty line. This is projected to increase to 7,081 which will account for 72% of households in the year June 2019 to June 2020. Furthermore, 3,983 or 41% of households were above the poverty line in the year June 2014 to June 2015 which is projected to decrease in the June 2019 to June 2020 financial year to 2,723 which constitutes 28% of households in the study location. It can thus be expected that over a protracted period of time, the poverty level in the study area will increase. The inevitable result of this situation is that households will become financially destitute, and basic household commodities will be rendered unaffordable.

2.4 Baseline Individual Exploratory Data Analysis

The HIPSS baseline study comprised a total of 11,289 households subsequently enrolling 9,812 respondents between the ages of fifteen and forty nine years. Owing to missing data and other statistical anomalies, eight observations were discarded from the data. The data set used in the analysis thus comprised a total of 9,804 observations.

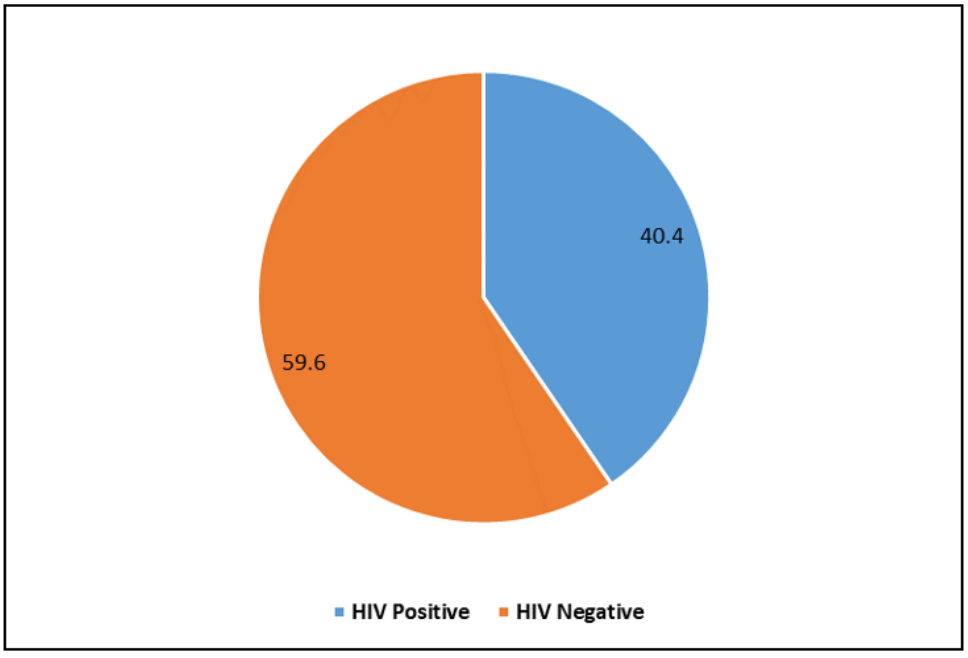

Figure 2.5: Observed prevalence of HIV among the study participants

Figure 2.5 above represents the observed prevalence of HIV among the study participants. A little more than 40% of the participants were HIV positive while approximately 60% of participants were HIV negative. Table 2.3 displays the weighted distribution of the sample variables at individual level. In respect of age demographics, more than one-fifth of the respondents were in the 20-24 year age group while participants at or approaching midlife (40-44 and 45-49 years) accounted for a tenth of all the respondents.

The final gender composition of the survey was 3,544 males and 6,260 females, representing 36.1% and 63.9% of the data set respectively. Examining the highest level of academic qualification in individuals, it was found that 43.1% of the participants had advanced beyond primary level education but had not completed secondary school compared with approximately 41% of the participants who graduated from secondary school. Tertiary graduates accounted for 5.6% of all the participants.

A further consideration in this study centered around examining the participants' knowledge of HIV prevention and acquisition of HIV information. In this respect the participants were gauged on nine relatively well-known, self-implementing measures to contain the spread of HIV and thereby reduce incidences of HIV infection. The participants' prowess at obtaining clinical and preventative HIV information from a variety of sources was also examined. We hypothesise that the more variation there is in the source of information, the more adequately informed the participants are of vital information pertaining to HIV.

Emphasis was also placed on investigating the prevailing social attitudes in the study area. To this end the level of HIV stigmatisation was measured by asking respondents about their perceptions of individuals living with HIV. They were, *inter alia*, asked if they would maintain a friendship with an HIV positive individual, whether HIV positive individuals ought to be ashamed of themselves, and whether they were deserving of their predicament.

Table 2.2: Score classification of behavioural characteristics

Domain	Score Interval	Classification
Knowledge of HIV Prevention	0 - 3	Lacking Knowledge
	4 - 6	Moderately Knowledgeable
	7 - 9	Highly Knowledgeable
HIV Information Acquisition	0 - 4	Lacking Information
	5 - 9	Moderately Well Informed
	10 - 16	Well Informed
HIV Stigmatisation	0	No Stigma
	1	Moderate Stigma
	2	Mild Stigma
	3	Severe Stigma

For the purpose of this study, the Guttmann scale, as detailed in Section 2.2 was employed to inform our findings. A response in the affirmative was recorded as one, a non-affirmative response was recorded as zero, and the sum of the scores were calculated. The scores were then classified as detailed in Table 2.2. On scoring the qualitative data detailed in Table 2.2, it was observed that the sum of preventative knowledge among participants was a cause for concern. It was found that there was a large degree of dis-proportionality in this respect as a combined percentage of 50.2% of participants were deemed moderately knowledgeable and highly knowledgeable of HIV preventative measures.

This closely matched the 49.9% of participants who lacked sufficient preventative knowledge, and further emphasised the need for knowledge and preventative information to form an integral component of targeted intervention measures. The disproportionality observed in the knowledge domain is commensurate with that of the participants' prowess in obtaining HIV preventative information in that a little more than 68% of the participants lacked adequate information pertaining to HIV prevention whilst a negligible proportion of participants were characterised as well informed. An overwhelming majority of participants, 81%, reported that they do not stigmatise an HIV positive person.

The vast majority of the participants were single, having never married and never cohabited. The HIV negative participants were observed to be largely risk averse when questioned about their perceived risks of HIV infection. While a considerable percentage of respondents, 84.6%, attested to having engaged in sexual intercourse prior to participating in the study, a clear majority, though less stark, attested to using contraceptions. In addition, 94.4% claimed not to have been diagnosed with a sexually transmitted infection.

Table 2.3: Distribution of the sample variables at individual level

Variable	Distribution (%)
Gender	
Male	36.1
Female	63.9
Age Group	
15-19	16.5
20-24	21.2
25-29	17.2
30-34	13.2
35-39	11.9
40-44	10.0
45-49	10.1
Highest Level of Education	
No Schooling	4.3
Primary	6.2
Incomplete Secondary	43.1
Completed Secondary	40.9
Tertiary	5.6
Knowledge of Prevention	
Lacking Knowledge	49.9
Moderately Knowledgeable	30.9
Highly Knowledgeable	19.2
HIV Information Acquisition	
Lacking Adequate Information	68.2
Moderately Well Informed	31.6
Well Informed	0.2
Marital status	
Legally Married	8.8
Separated - Legally Married	0.2
Cohabiting	2.4
Single - Never Married or Cohabited	83.8
Divorced	0.2
Single - Live in Partner	3.8
Widowed	0.8

Continued on next page

Continued from previous page

Variables	Distribution (%)
Perceived Risk of HIV	
Assured Infection	4.2
Probable Infection	21.0
Probable Non-Infection	37.0
Assured Non-Infection	15.3
Already HIV Positive	22.5
HIV Stigma	
No Stigma	81.0
Mild Stigma	15.4
Moderate Stigma	3.0
Severe Stigma	0.6
Engaged in Sexual Intercourse	
No	15.4
Yes	84.6
Diagnosed with an STI	
No	94.4
Yes	5.6
Used Contraception	
No	59.3
Yes	40.7

2.5 Exploring the Observed Prevalence of HIV

On investigating the observed prevalence of HIV across the socio-economic, socio-demographic and behavioural factors, an accord was observed between the results of the exploratory analyses and that of the regression analyses. Figure 2.6 depicts the scale of HIV prevalence in participants residing in households across varying intensities of household deprivation. One may discern from Figure 2.6 that the scale of deprivation is not a proximate cause of HIV incidence. The reason for this obser-

vation is that the prevalence of HIV among residents in households of deepening deprivation is largely the same. While households that experience no form of deprivation have relatively fewer residents who are HIV positive, the prevalence of HIV remains largely the same as the scale of deprivation becomes more intense. Additionally, no participants resided in households that were moderately or majorly deprived as no households were observed to fall within these deciles. Households characterised as extremely deprived housed the most HIV positive participants.

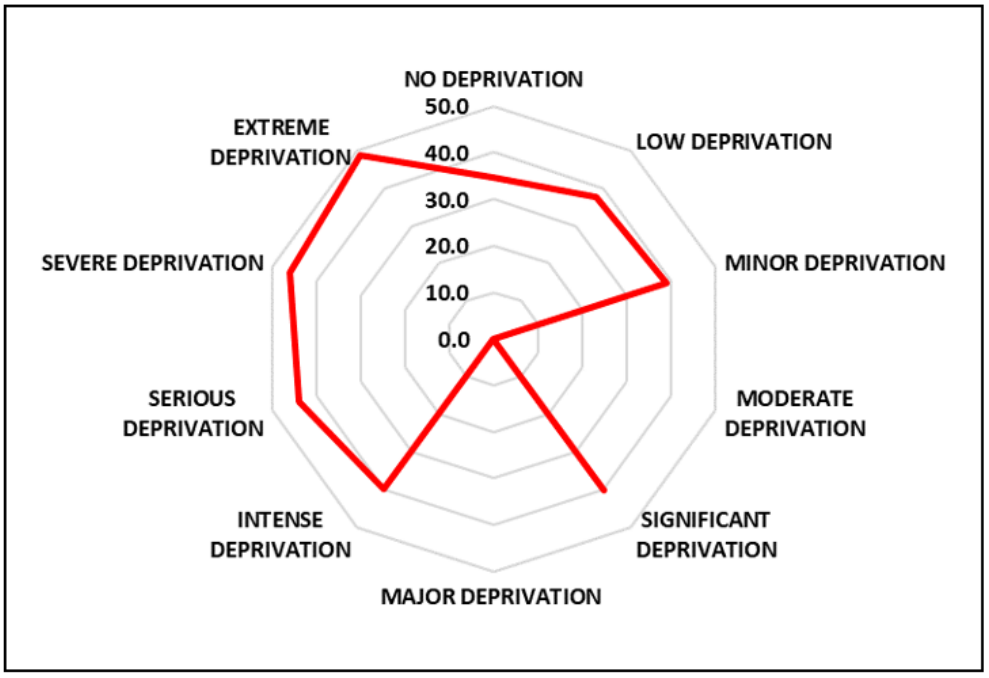

Figure 2.6: Observed prevalence of HIV of participants resident in household of varying deprivation levels

An infection rate of 47.2% among female participants was observed to be approximately twice that of male participants of whom 28.6% were HIV positive. As per Figure 2.8, participants considered to be in early to middle adulthood (25-44 years old) displayed the highest prevalence of HIV infection in their respective age groups accounting for more than 60% in some instances.

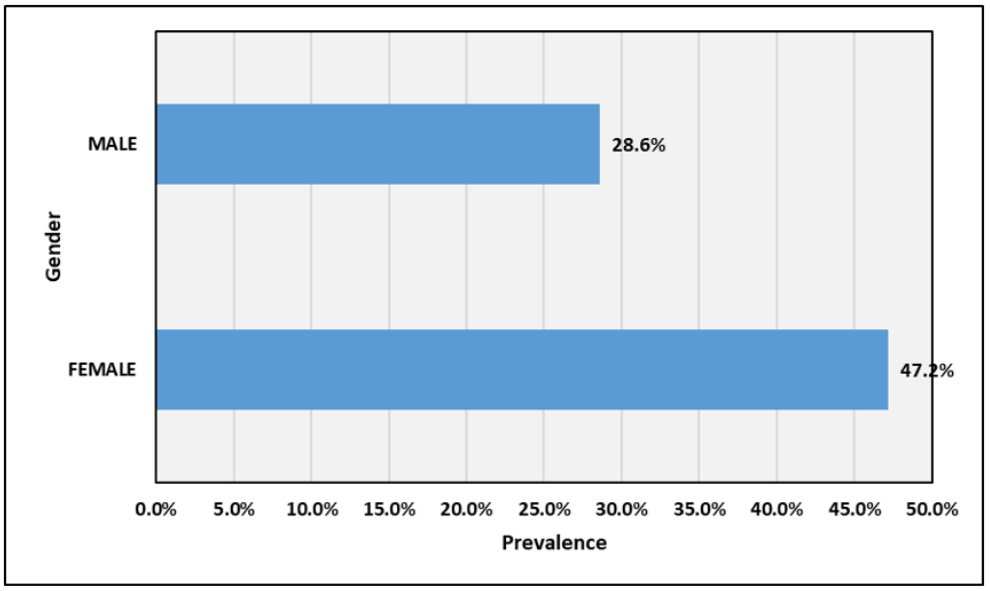

Figure 2.7: Observed prevalence of HIV with respect to gender.

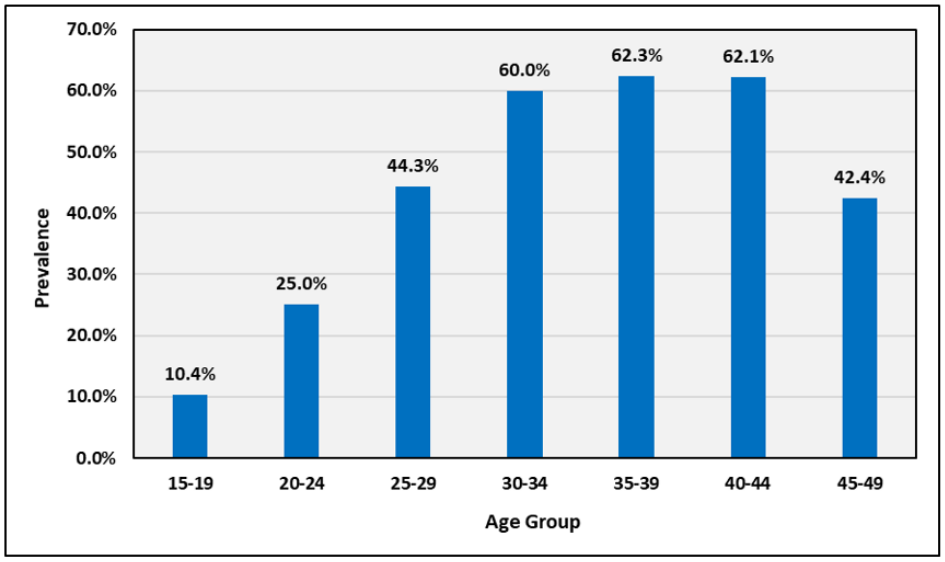

Figure 2.8: Distribution of HIV prevalence with respect to age category.

Participants who advanced beyond primary levels of education but did not complete secondary levels of education were observed to have the highest prevalence of HIV at 42.6%, which reduced to 38% for participants who completed secondary school.

Tertiary graduates were the least infected group with an observed HIV prevalence of 26.4% as per Figure 2.9. There appeared to be no noticeable outliers in terms of HIV prevalence when measured against a participant's marital status as can be seen in Figure 2.10.

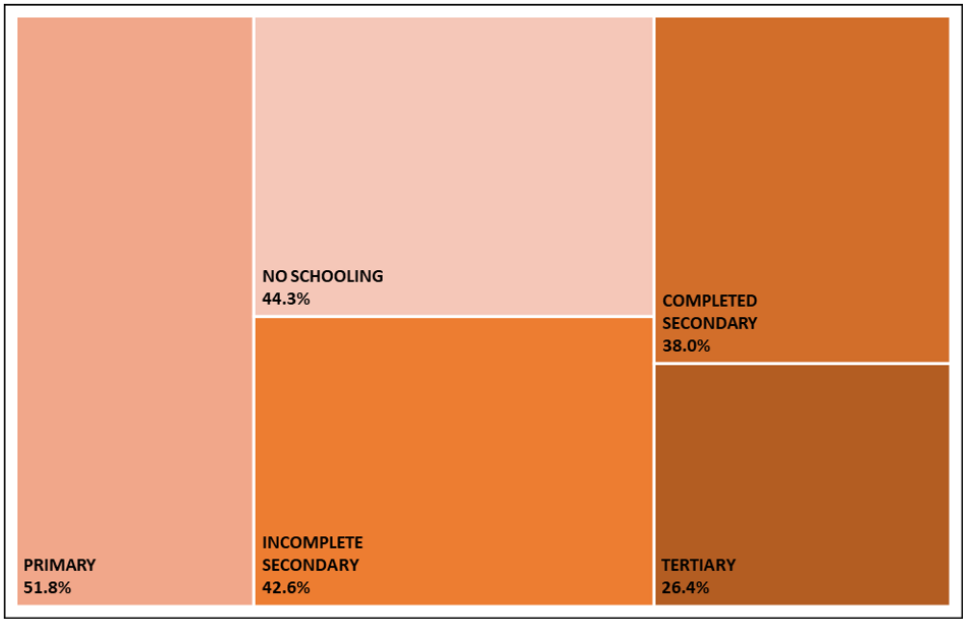

Figure 2.9: Distribution of HIV prevalence with respect to highest level of education.

A majority of participants who were single (with a live-in partner), widowed, separated but remained married, and participants who were cohabiting, were HIV positive. The observed prevalence within these respective groups exceeded 50% while approximately 40% of participants who were HIV positive were legally married or were single, having never married or cohabited.

Figure 2.11 depicts the observed prevalence of HIV with respect to selected behavioral characteristics. The observed prevalence of HIV according to an individual's knowledge of HIV prevention appeared equal irrespective of the category they fall into as per Table 2.2.

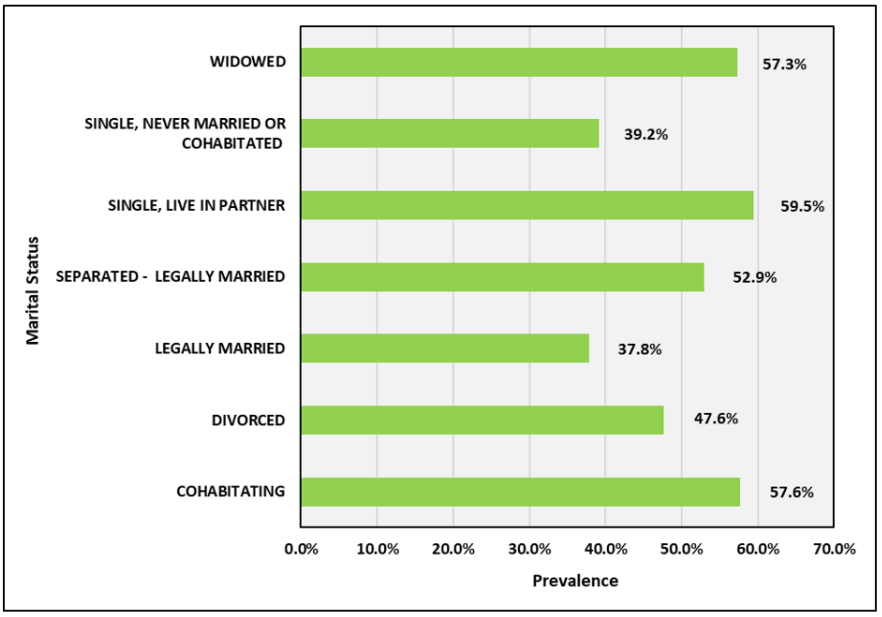

Figure 2.10: Distribution of HIV prevalence with respect to marital status

There appeared to be no discernible decrease, only a negligible increase, in the observed prevalence for individuals who were higher in terms of their knowledge rank. In this respect, the observed prevalence across the knowledge scale stood at approximately 40%.

In addition, the prevalence of HIV according to the participants' espousal of HIV stigmastisation prevalent in the study area also appeared to show no noticeable fluctuations in respect of HIV prevalence according to Figure 2.12. Among participants who did not exhibit any form of stigmatisation, the HIV prevalence was approximately 42% while the prevalence among those who had displayed severe forms of HIV stigmatisation was 38.3%.

Likewise, when measuring the scale of information possessed by respondents, it was noticed that the observed prevalence among those who were lacking adequate information or were in possession of moderate levels of information was inordinately high. This is in contrast to those who were well informed having been in

possession of vast amounts of information supplementing their knowledge on HIV transmission and infection.

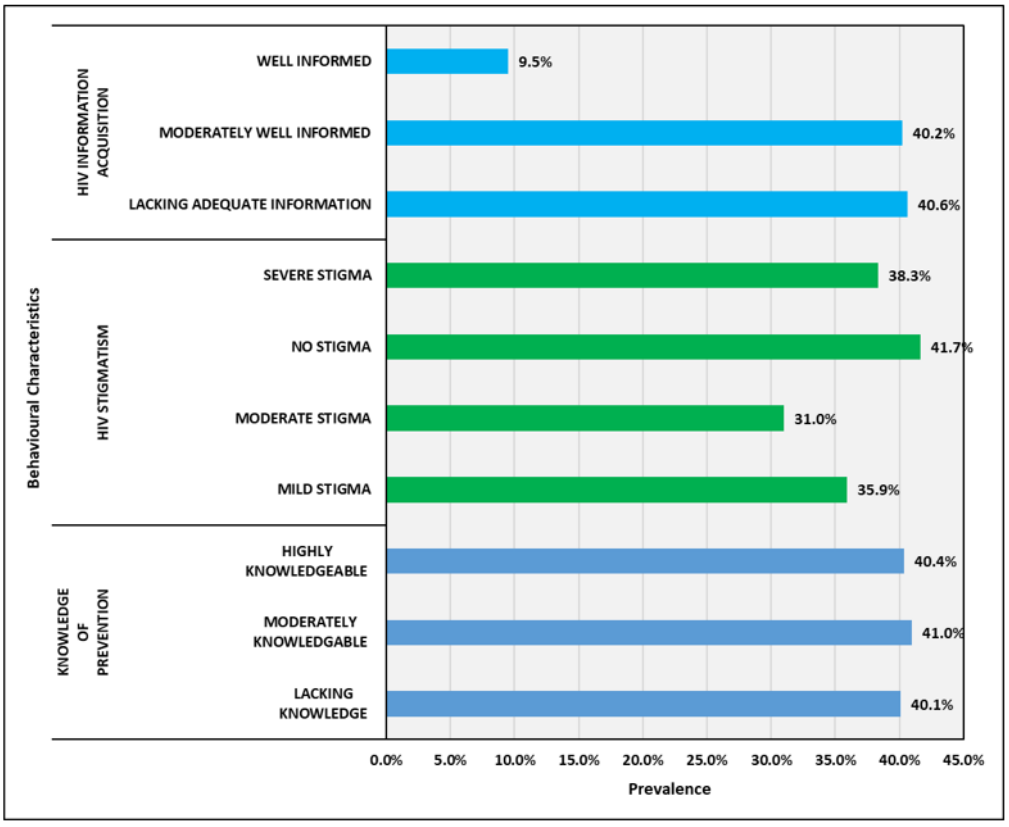

Figure 2.11: Distribution of HIV prevalence with respect to selected behavioural characteristics.

Assessing participant sexual behavioural characteristics, more than 45% of HIV positive respondents affirmed that they had engaged in sexual intercourse prior to participating in the study while more than half of the individuals diagnosed with a sexually transmitted infection, were HIV positive. Furthermore, a little over 47% of individuals who attested to using a condom at their sexual debut were observed to be HIV positive.

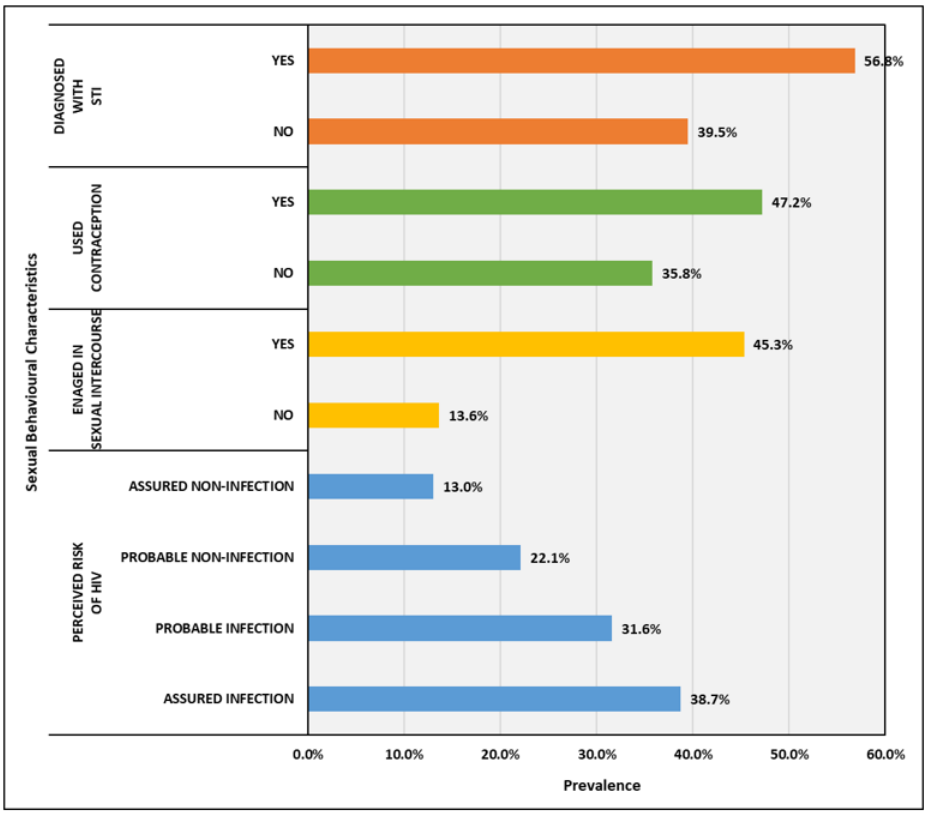

Figure 2.12: Distribution of HIV prevalence with respect to selected sexual behavioural characteristics.

2.6 Exploring Gender Based Characteristics

To determine if there was any gender disparity with respect to HIV prevalence, an investigation was conducted on selected socio-demographic and behavioural characteristics. The selected characteristics were participants' age, highest level of education, sexual behaviour and clinical characteristics. An almost symmetrical pattern of HIV prevalence was noted; female participants displayed a higher prevalence of HIV in young adulthood and males displayed a higher prevalence from middle adulthood onward. Furthermore, HIV was not seemingly prevalent in adolescence with both male and female participants displaying an observed prevalence not exceeding 5% whereas participants in the 30-34 year age band accounted for most cases of HIV at baseline. In this age group male and female participants had an observed prevalence of 21.2% and 19% respectively.

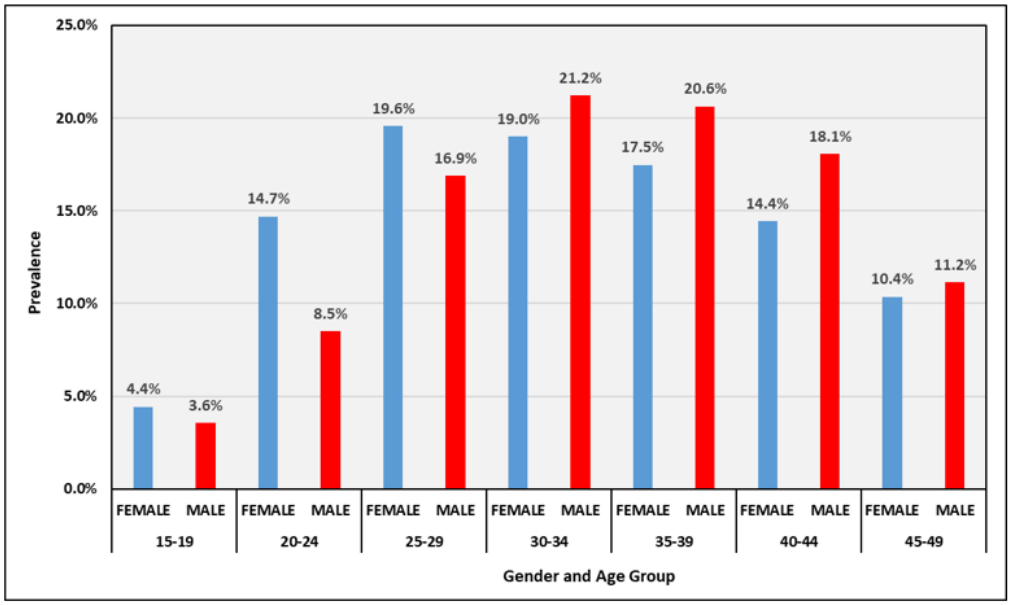

Figure 2.13: Distribution of HIV prevalence by gender and age category.

An inordinately high prevalence of HIV was noted in both male and female individuals who had advanced beyond primary school. However, for both male and female participants the prevalence was higher among those who had not completed secondary school compared with participants who had not. On inspection, one would note the level of extremity in terms of academic qualification. This is in contrast with participants who had no formal schooling or who had not sought any further education beyond primary school.

Furthermore, one may discern from Figure 2.14 that male and female participants who had no formal schooling, together with tertiary graduates are among the least infected group with respect to academic attainment with an observed prevalence under 5%.

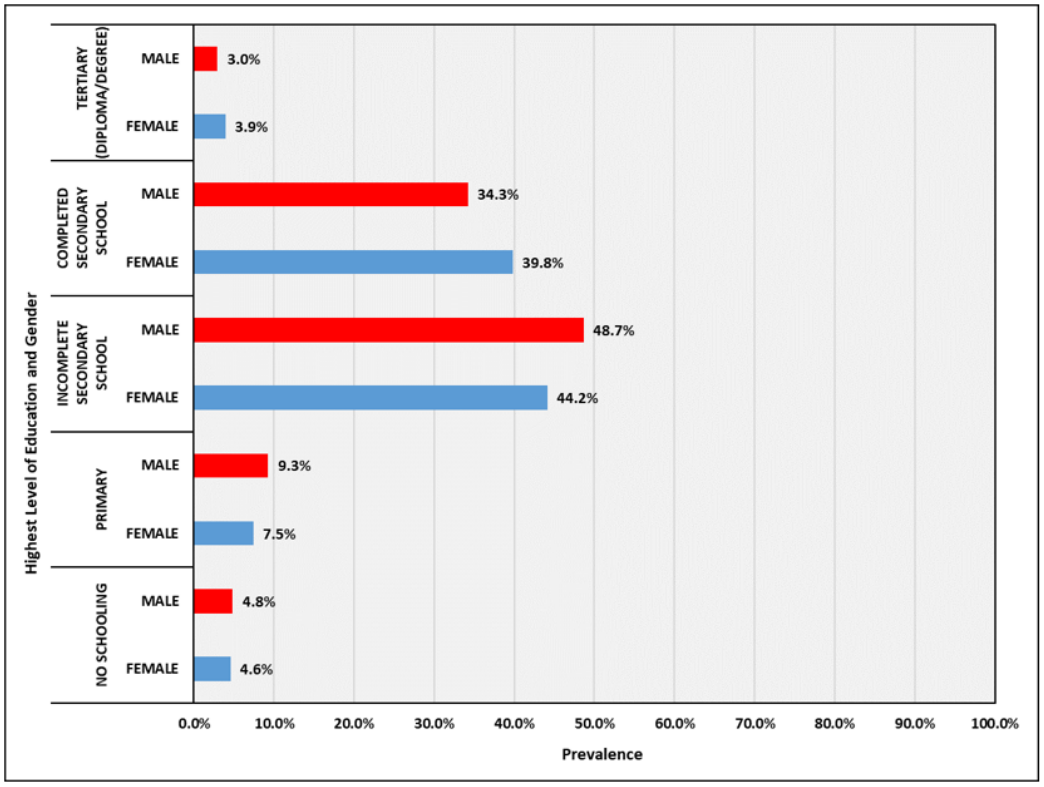

Figure 2.14: Distribution of HIV prevalence by gender and highest level of education.

Males who made use of contraception were observed to have a lower prevalence of HIV than those who did not. Male participants recorded a prevalence of more than 70% for those who eschewed the use of contraception during their first sexual encounter, and the prevalence decreased by more than half to 29.8% for those who did not eschew the use of a condom. This is in contrast to female participants among whom the prevalence increased from 46.3% for those who did not use contraception to more than 50% for those who used contraception.

Moreover, the clinical diagnosis of sexually transmitted infection does not appear to be linked with HIV positive diagnosis as can be seen in Figure 2.15, showing an inordinately high prevalence among male and female participants who were not diagnosed with HIV. Conversely, engaging in sexual intercourse appeared to be associated with HIV infection owing to the disproportionately high prevalence of HIV in both male and female participants.

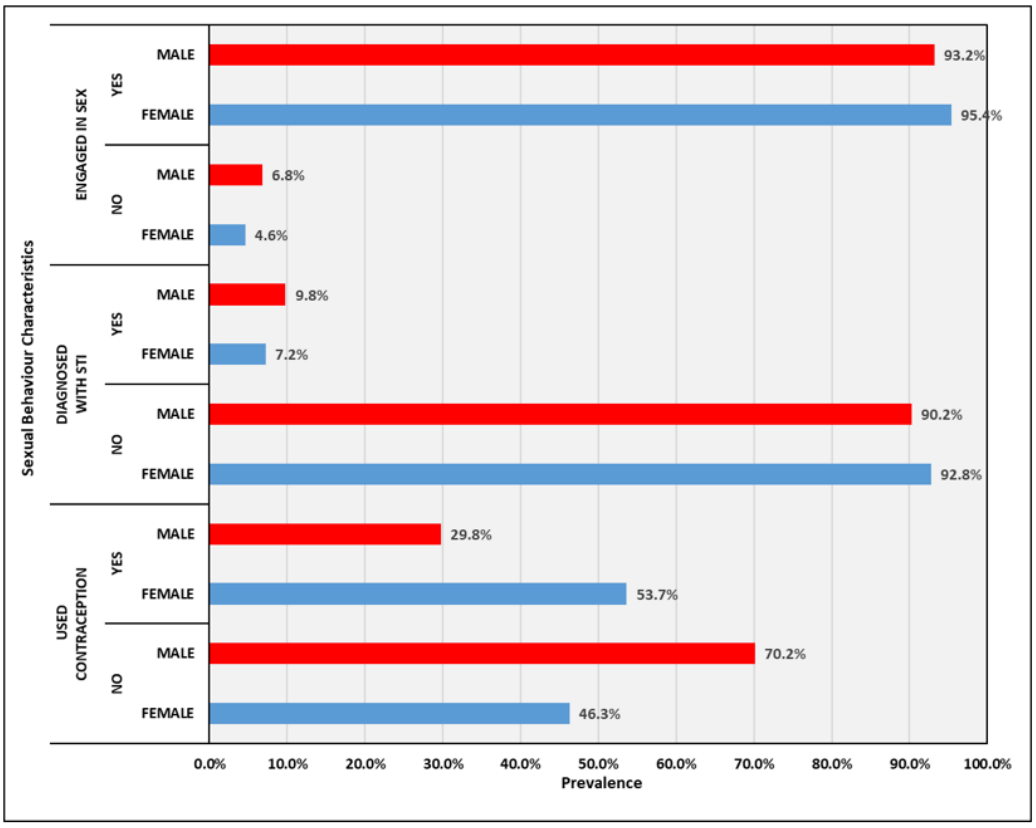

Figure 2.15: Distribution of the prevalence of HIV by gender and sexual behaviour.

Another common method employed in the area of exploratory data analysis is that of correspondence analysis, particularly multiple correspondence analysis. This technique allows one to explore the intricate relationships that exist between different categorical variables under consideration. A vital factor that motivates the use of multiple correspondence analysis is that it is able to simplify complex data into a user- friendly contingency table while still retaining valuable information within the data set. The following section details the theoretical considerations of correspondence analysis and thereafter multiple correspondence analysis since the former is vital in facilitating our understanding of the latter.

2.7 Introduction to Correspondence Analysis

The foundations of correspondence analysis can be traced as far back as the early twentieth century. De Leeuw (1983) partially credits Karl Pearson, in Pearson (1906), for laying the foundation of correspondence analysis. De Leeuw (1983) qualifies this by noting that Pearson (1906) did not make the link between presenting data in a contingency table and *singular value decomposition*, despite Beltrami (1873), Sylvester (1889) and Jordan (1874) studying singular value decomposition earlier than Pearson. Instead Pearson developed a correlation coefficient for a two-way contingency table by employing linear regression. The origins of correspondence analysis are thus rooted in techniques that were algebraic rather than geometric; the approach which was subsequently taken.

Research into correspondence analysis took a quantum leap forward from the 1960s when the geometric approach was given meaning by Jean-Paul Benzécri, a French statistician leading a team of mathematical statisticians at the Collége de France, who coined the French term for correspondence analysis as *l'analyse des correspondence* in the spring of 1963. Later, Benzécri would go on to conceptualise the technique of correspondence analysis further by developing aids to the interpretation and the deployment of correspondence analysis in software programmes in his laboratory at the Université Pierre-et-Marie-Curie in Paris during the last three decades of the twentieth century. This culminated in the publication of Benzécri's journal *Les Cahiers de l'Analyse des Données* (Journal of Data Analysis) the central focus of which was the repositioning of statistical thought in the light of computer based statistics becoming increasingly dominant - perhaps marking the beginning of supervised machine learning in mainstream statistical analysis.

Further development of correspondence analysis was carried out by a student of Jean-Paul Benzécri, Michael Greenacre, a South African statistician who gave further impetus to correspondence analysis resulting in its rapid expansion and use. In his

work, Greenacre (1984) sought to transcribe the French text of Benzécri's work making it more accessible to English speakers hence the enduring influence of Benzécri's style and methodology in correspondence analysis to the present day.

The development and conceptualization of correspondence analysis has since, over many years, contributed in leaps and bounds to societal understanding of a range of fields encompassing medicine, economics, psychometric analysis, linguistics and biometry to name a few; testifying to the versatility of correspondence analysis. This is encapsulated by (Kendall, 1972, p. 194) who states:

> *"It is hard to think of any subject which has not made some kind of contribution to statistical theory - agriculture, astronomy, biology and chemistry and so on through the alphabet. The remarkable thing, perhaps, is that these lines of development remained relatively independent for so long and only in the present century have been seen to have a common conceptual content."*

2.8 Classical Correspondence Analysis

In this section we develop, in detail, the theoretical considerations of *classical* correspondence analysis proposed chiefly by Greenacre (1984). Classical correspondence analysis is sometimes termed *simple* correspondence analysis. However, this must not be misconstrued as being easy to understand and implement. It *simply* implies that the theory developed for it is centered around the *simplest* form of representing a data set; a *two-way contingency table*.

The theoretical considerations of classical correspondence analysis presented in this section precedes that of multiple correspondence analysis. Multiple correspondence analysis extends the application of simple correspondence analysis from two-way contingency tables to multi-way contingency tables.

2.8.1 The Data Structure

Suppose an $I \times J$ two-way contingency table, *N*, wherein the $(i,j)^{th}$ element is denoted by n_{ij} for $i \in (1,2,....,I)$ and $j \in (1,2,....,J)$. Define the *grand total* of *N* as *n* and ***P*** as the probability or *correspondence* matrix wherein the $(i,j)^{th}$ element is given by $p_{ij} = \frac{n_{ij}}{n}$ where, as expected, the following result holds true

$$\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1 \tag{2.2}$$

Define the i^{th} row and j^{th} column marginal or conditional probabilities respectively as

$$p_{i\cdot} = \sum_{j=1}^{J} p_{ij} \quad \text{and} \quad p_{\cdot j} = \sum_{i=1}^{I} p_{ij} \tag{2.3}$$

It thus follows that

$$\sum_{i=1}^{J} p_{i\cdot} = \sum_{j=1}^{I} p.j = 1 \tag{2.4}$$

These marginal values are collectively called *masses* where the row marginal probabilities are referred to as *row masses* and the column marginal probabilities are referred to as *column weights*. Furthermore, D_i and D_j are defined as diagonal matrices in which the elements are the row and column masses respectively.

2.8.2 Profiles

Suppose one wishes to measure the degree of association between two *row* categories, *i* and *i'*. In the event that a particular cell value has a large number of observations, then it would have a large cell probability. As such, within the realm of correspondence analysis, one does not effect cell probabilities to measure the degree of comparison. Instead, one divides each row element by its respective row marginal value. This produces the *row profile* of the contingency table. Likewise, the

construction of the *column profile* follows analogously. Thus, the row profile, $r_{i\cdot}$, and the column profile, $r_{\cdot j}$, of the contingency table is given respectively as

$$r_{i\cdot} = \frac{p_{ij}}{p_{i\cdot}} \quad \text{for} \quad j \in (1, 2, ..., J) \tag{2.5}$$

and

$$r_{\cdot j} = \frac{p_{ij}}{p_{\cdot j}} \quad \text{for} \quad i \in (1, 2, ..., I) \tag{2.6}$$

2.8.3 Total Inertia

According to De Leeuw & Mair (2009) the concept of inertia and the implementation thereof in correspondence analysis is rooted in its conceptualisation in the field of mechanics. The term *moment of inertia*, from which the term inertia is derived, is rooted in Newtonian laws of motion where an object with a mass *m* and distance *d* has a *center of gravity* relative to a certain position in space.

Despite the widespread use of the χ^2 test as a measure variation from complete independence of the contingency table, it suffers a drawback in the case of proportionality. If the grand total, *n*, is doubled (or artificially inflated), there is a commensurate change to the χ^2 test. However, there is no change in variation between the rows and columns. As such, correspondence analysis corrects for this anomaly by using $\frac{\chi^2}{n}$ as a measure of variation rather than χ^2.

This ratio represents the *total inertia* whose constituent components comprise the contribution of each of the axes referred to as the *principle inertia*. Thus, correspondence analysis is deemed to produce an analysis concerned primarily with the correspondence matrix $\boldsymbol{P}$ than that of the primitive matrix N. It can be shown, via an amalgamation of the row profile and column profile, that the *total inertia* of the contingency table is given by

$$\text{Total Inertia} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(r_{ij} - r_j c_j)^2}{r_j c_j} \tag{2.7}$$

where c_j and r_j are the column and row masses respectively and are defined as

$$r_j = \frac{n_{i\cdot}}{n} \qquad \text{and} \qquad c_j = \frac{n_{\cdot j}}{n} \tag{2.8}$$

2.8.4 Singular Value Decomposition

The aim of correspondence analysis is to measure the strength of association between two categories. As with many multivariate techniques, a score serves as an indicator of this measure of association or disassociation between two categories in the rows or columns of the contingency table. In this endeavour, one may also wish to measure the association between rows and columns.

Correspondence analysis can be applied in such instances to a contingency table by using the property of complete independence between the rows and columns

$$p_{ij} = p_{i\cdot}p_{\cdot j} \quad \text{for} \quad i \in (1, 2, ..., I) \quad \text{and} \quad j \in (1, 2, ..., J) \tag{2.9}$$

Now, the condition of strict independence is not always satisfied and in order to account for this departure from the expected norm, a generic multiplicative measure is required in Equation 2.9, and denoted by

$$p_{ij} = \alpha_{ij}p_{i\cdot}p_{\cdot j} \tag{2.10}$$

where in the case of complete independence, α_{ij} = 1. Furthermore, as it is already acknowledged that complete independence between the rows and columns cannot always exist, one may determine instances where it does not exist by ascertaining when $\alpha_{ij} \neq 1$. This can be done by rearranging p_{ij}, which is called the *Pearson ratios*, as such

$$\alpha_{ij} = \frac{p_{ij}}{p_{i\cdot}p_{\cdot j}} \tag{2.11}$$

The process now proceeds to determining the scores of the rows and columns that will ultimately act as an indicator of the strength of association between the rows and columns. This is achieved by partitioning the Pearson ratios by a method of *singular value decomposition* such that

$$\alpha_{ij} = \sum_{m=0}^{M^*} a_{im}\lambda_m b_{jm} \tag{2.12}$$

where

$$\sum_{i=1}^{I} p_{i\cdot} a_{im} a_{im'} = \begin{cases} 1 & m = m', \\ 0 & m \neq m'. \end{cases} \tag{2.13}$$

and

$$\sum_{j=1}^{J} p_{\cdot j} b_{jm} b_{jm'} = \begin{cases} 1 & m = m', \\ 0 & m \neq m'. \end{cases} \tag{2.14}$$

and M^* = min(I, J) $-$ 1

Consider the RHS of Equation 2.12, if $\{a_{iu}, i = 1, 2, ..., I\}$ is defined as the u^{th} *left generalised vector* associated with the row categories and similarly, $\{b_{jv}, j = 1, 2, ..., J\}$ as the v^{th} *right generalised vector* associated with the column categories then the generalised basic vectors may be referred to as *singular vectors*.

The elements of λ_{m} are real and positive and are the first M^* *singular values* arranged in descending order as

$$1 = \lambda_0 \geq \lambda_1 \geq ... \geq \lambda_m^* \geq 0 \tag{2.15}$$

and which may be calculated as

$$\lambda_m = \sum_{i=1}^{I}\sum_{j=1}^{J} a_{im} b_{jm} p_{ij} \tag{2.16}$$

A consequence of the ordinal nature of the singular values is that the first value is *trivial* and they have a minimum value of zero. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are defined as singular vectors with trivial solutions for the set of values $\{a_{i0}\}$ and $\{b_{j0}\}$ equal to one, then in matrix notation, the correspondence analysis is defined as

$$\boldsymbol{D}_I^{-1}\boldsymbol{P}\boldsymbol{D}_J^{-1} = \boldsymbol{A}\boldsymbol{D}_\lambda\boldsymbol{B}' \tag{2.17}$$

where

$$\boldsymbol{A}'\boldsymbol{D}_I\boldsymbol{A} = \boldsymbol{I} \tag{2.18}$$

$$\boldsymbol{B}'\boldsymbol{D}_J\boldsymbol{B} = \boldsymbol{I} \tag{2.19}$$

where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{D}_I$ and $\boldsymbol{D}_J$ are the diagonal matrices whose elements comprise the row and column masses respectively. The matrix $\boldsymbol{A}$, the left generalised basic vector, has dimension $I \times M^*$ and contains the first M^* set of row scores while $\boldsymbol{B}$, the right generalised basic vector, with dimension $J \times M^*$ contains the first set M^* column scores. The matrix of singular values is the diagonal matrix, D_λ.

An alternative decomposition method is to use orthogonal polynomials instead of singular vectors of $\{a_{im}\}$ and $\{b_{jm}\}$. To extract the triviality referred to earlier, Equation 2.12 is rewritten as

$$\alpha_{ij} = 1 + \sum_{m=1}^{M^*} a_{im}\lambda_m b_{jm} \tag{2.20}$$

Now, using $\alpha_{ij} - 1$, the *Pearson contingencies*, under the usual limits for i and j, and applying a simple mathematical manipulation, Equation 2.11 becomes,

$$\frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{p_{i\cdot}p_{\cdot j}} = 1 + \sum_{m=1}^{M^*} a_{im}\lambda_m b_{jm} \tag{2.21}$$

2.9 The Correspondence Plot

As previously stipulated, correspondence analysis is a graphical statistical procedure that is applied to contingency tables. It is thus able to visualise the association between the rows and columns which respectively represent a particular attribute or characteristic. There are two methods of determining the coordinates that will be plotted to form the correspondence plot. Greenacre (1984) calls the first system the *standard coordinate*, which the singular vector system as detailed in Subsection 2.8.4, and the second the *classical coordinate* system. The difference between these two methods lie in their characterisation of the contribution of the principal inertia to the total inertia. These are explained in detail below.

2.9.1 Standard Profile Coordinates

To visualise the associations between the row categories or column categories, the singular vectors, $\{a_{im}\}$ and $\{b_{jm}\}$ are projected onto a *correspondence plot*. For M^* number of dimensions, the correspondence plot is called an *optimal plot* where the total inertia may be expressed in terms of the singular vector as

$$\frac{\chi^2}{n} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(p_{ij} - p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}}$$

Squaring the binomial and extracting $p_{i\cdot}p_{\cdot j}$ as a common factor, the following expression is derived

$$= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(p_{i\cdot}p_{\cdot j})^2 \left(\frac{p_{ij}}{p_{i\cdot}p_{\cdot j}} - 1\right)^2}{p_{1\cdot}p_{\cdot j}}$$

Now, from Equation 2.21

$$= \sum_{i=1}^{I}\sum_{j=1}^{J} p_{i\cdot}p_{\cdot j} \left(\sum_{m=1}^{M} a_{im}\lambda_m b_{jm}\right)^2$$
$$= \sum_{m=1}^{M} \lambda_m^2 \left(\sum_{i=1}^{I} p_{i\cdot}a_{im}^2\right)\left(\sum_{j=1}^{J} p_{\cdot j}b_{jm}^2\right)$$

which simplifies, as per Equation 2.13 and Equation 2.14, to

$$\frac{\chi^2}{n} = \sum_{m=1}^{M^*} \lambda_m^2 \tag{2.22}$$

Hence, it follows that the total variation in the contingency table may be partitioned into M^* components; the principle inertia values. Furthermore, these principle inertia values may be decomposed further to illustrate how a particular row or column contributes to the principle axis.

An M-dimensional correspondence plot comprises M principle axes for $M < M^*$. The x-axis is called the *first principle axis* while the y-axis is called the *second principle axis*. These components come together to produce the correspondence plot where the row profile $\{a_{im}\}$ is the coordinate of the I^{th} row category and $\{b_{jm}\}$ is the coordinate of the J^{th} column category, both of which are plotted on the m^{th} principle axis. This, however, does not take into consideration the strength of the contribution of the rows and the columns as these respective axes have a unit inertia associated with them.

2.9.2 Classic Profile Coordinates

Eschewing the use of singular vectors, define the row and column coordinates as:

$$f_{im} = a_{im}\lambda_m \tag{2.23a}$$

$$g_{jm} = b_{jm}\lambda_m \tag{2.23b}$$

Thus, 2.13 becomes

$$\sum_{i=1}^{I} p_{i\cdot} \left(\frac{f_{im}}{\lambda_m}\right)\left(\frac{f_{im'}}{\lambda_{m'}}\right) = \begin{cases} 1 & m = m', \\ 0 & m \neq m'. \end{cases} \tag{2.24}$$

which by a simple algebraic manipulation may be expressed as

$$\begin{cases} \sum_{i=1}^{I} p_{i\cdot} f_{im}^2 = \lambda_m^2 & \text{if} \quad m = m' \\ \\ \sum_{i=1}^{I} p_{i\cdot} f_{im} f'_{im} = 0 & \text{if} \quad m \neq m' \end{cases} \tag{2.25}$$

Similarly, 2.14 becomes,

$$\sum_{i=1}^{I} p_{\cdot j} \left(\frac{g_{jm}}{\lambda_m} \right) \left(\frac{f_{jm'}}{\lambda_{m'}} \right) = \begin{cases} 1 & m = m', \\ \\ 0 & m \neq m'. \end{cases} \tag{2.26}$$

which, again, by a rudimentary mathematical manipulation becomes

$$\begin{cases} \sum_{j=1}^{J} p_{\cdot j} g_{jm}^2 = \lambda_m^2 & \text{if} \quad m = m' \\ \\ \sum_{i=1}^{J} p_{i\cdot} g_{jm} g'_{im} = 0 & \text{if} \quad m \neq m' \end{cases} \tag{2.27}$$

Now, the system of coordinates in Equation 2.23 makes use of the singular value method detailed in Section 2.8.4. As per Equation 2.25 and Equation 2.27 the rows and columns have an associated unit inertia, λ_m^2, associated with them. Thus, the first principle axis with inertia of λ_1^2 is considered the most important axis as λ_1 is the largest singular value.

Hence, in classical correspondence analysis, a correspondence plot containing more than two dimensions will see the first two principle axis being more descriptive than any other axes included. Consider the row profile and column profile coordinates given by $\{f_{im}\}$ and $\{g_{jm}\}$ respectively. From their respective relationships to the principle inertia, as outlined in Equation 2.25 and Equation 2.26, their contributions to the m^{th} principle inertia is, respectively

$$\lambda^2_{m(i)} = p_{i\cdot} f^2_{im} \tag{2.28}$$

so that

$$\lambda^2_m = \sum_{i=1}^{I} \lambda^2_{m(i)} \tag{2.29}$$

And, similarly for the column profile

$$\lambda^2_{m(j)} = p_{\cdot j} g^2_{jm} \tag{2.30}$$

so that

$$\lambda^2_m = \sum_{j=1}^{J} \lambda^2_{m(j)} \tag{2.31}$$

Now, from Equation 2.22, Equation 2.29 and Equation 2.31, it follows for the row profile that

$$\frac{\chi^2}{n} = \sum_{i=1}^{I} \sum_{m=1}^{M} p_{i\cdot} f^2_{im} \tag{2.32}$$

and for the column profiles

$$\frac{\chi^2}{n} = \sum_{j=1}^{J} \sum_{m=1}^{M^*} p_{\cdot j} g^2_{jm} \tag{2.33}$$

From Equation 2.32 and Equation 2.33 it can be seen that the points (or profile coordinates) close to the origin are not contributory to the variation of the data. Profile coordinates further from the origin, however, are contributory to any variation in the data.

Alternatively, using the property of orthogonality and multiplying Equation 2.21 by $p_{\cdot j} b_{jm}$, the row profile and column coordinates may be respectively expressed as

$$f_{im} = \sum_{j=1}^{J} \frac{p_{ij}}{p_{i\cdot}} b_{jm} \tag{2.34}$$

and

$$g_{jm} = \sum_{i=1}^{I} \frac{p_{ij}}{p_{\cdot j}} a_{im} \tag{2.35}$$

where Equation 2.34 and Equation 2.35 are the weighted sum of the i^{th} row profile and j^{th} column profile respectively.

2.10 Distance

2.10.1 Centering of Profile Coordinates

It can be shown that the row and column profile coordinates, centered about the *centroid* (origin) of the correspondence plot, is where the expected values $\{p_{i\cdot}p_{\cdot j}\}$ lie. In this, we show that

$$\sum_{i=1}^{I} p_{i\cdot} f_{im} = 0 \qquad \forall m \in (1, 2, ..., M^*) \tag{2.36}$$

In order to prove this, consider the expression for the row profile coordinate as per 2.23a

$$\begin{aligned} \sum_{i=1}^{I} p_{i\cdot} f_{im} &= \sum_{i=1}^{I} p_{i\cdot} a_{im} \lambda_m \\ &= \lambda_m \sum_{i=1}^{I} p_{i\cdot} a_{im} \qquad (2.37) \\ &= 0 \qquad (2.38) \end{aligned}$$

as per Equation 2.13

Using the same deductive reasoning, it can be shown that the column profile is centered about the origin, where $\forall m \in (1, 2, ..., M^*)$ the following holds true;

$$\sum_{j=1}^{J} p_{\cdot j} g_{jm} = 0 \tag{2.39}$$

2.10.2 Distance from the Origin

The distance of the i^{th} row profile from the origin in an M^* dimensional correspondence plot is commensurate with the variation between the profile of the i^{th} row and the average column profile. This is illustrated below using the row profile.

The same conclusion can be drawn using the squared distance of column profile coordinates from the origin. From the origin, the squared Euclidean distance of the i^{th} row profile is

$$d_I(i,0)^2 = \sum_{j=1}^{J} \frac{1}{p_{\cdot j}} \left(\frac{p_{ij}}{p_i} \cdot - p_{\cdot j} \right)^2$$

By the Pearson ratio in Equation 2.20

$$\begin{aligned} &= \sum_{j=1}^{J} p_{\cdot j} (\alpha_{ij} - 1)^2 \\ &= \sum_{j=1}^{J} p_{\cdot j} \left(\sum_{m=1}^{M} a_{im} \lambda_{im} b_{im} \right)^2 \\ &= \sum_{m=1}^{M^*} \left(\sum_{j=1}^{J} p_{\cdot j} b_{jm}^2 \right) a_{im}^2 \lambda_m^2 \end{aligned}$$

This then simplifies to

$$d_I^2(I,0) = \sum_{m=1}^{M^*} f_{im} \tag{2.40}$$

Thus, Equation 2.32 may be expressed as

$$\frac{\chi^2}{n} = \sum_{i=1}^{M^*} p_{i\cdot} d_I^2(I,0) \tag{2.41}$$

Hence, points further from the origin are indicative of increased deviation from the expectation under complete independence. Points centered about the origin are not indicative of a deviation from the hypothesis of complete independence.

2.10.3 Within Variable Distances

In addition to computing the distance from the origin, a correspondence plot is able to graphically represent the association between two profiles of the same variables. Using the row profile argument, it is illustrated below that when two categorical profiles (row or column profiles) are in close proximity in a correspondence plot, they are deemed to be similar to those positioned at a distance. This is illustrated below using the row profile argument.

The squared distance (Euclidean distance) between two row profiles, i and i' in a correspondence plot is given by

$$d_I^2(i,i') = \sum_{j=1}^{J} \frac{1}{p_{\cdot j}} \left(\frac{p_{ij}}{p_{i\cdot}} - \frac{p_{i'j}}{p_{i'\cdot}} \right)^2 \tag{2.42}$$

In order to show that row categories at a distance are dissimilar, we aim to show that the distance between two row profiles can be expressed in terms of the row profile coordinates, f_{im} and $f_{i'm'}$ along the m^{th} principle axis.

$$\begin{aligned} d_I^2(i,i') &= \sum_{j=1}^{J} \frac{1}{p_{\cdot j}} \left(\frac{p_{ij}}{p_{i\cdot}} - \frac{p_{i'j}}{p_{i'\cdot}} \right)^2 \\ &= \sum_{j=1}^{J} p_{\cdot j} \left(\frac{\mathrm{p}_{ij}}{\mathrm{p}_{i\cdot}} - \frac{p_{i'j}}{p_{i'\cdot} p_{\cdot j}} \right)^2 \end{aligned}$$

Now, from Equation 2.20 and Equation 2.21

$$\begin{aligned} &= \sum_{j=1}^{J} p_{\cdot j} \left[\sum_{m=0}^{M^*} a_{im}\lambda_m b_{jm} - a_{i'm}\lambda_m b_{jm} \right]^2 \\ &= \sum_{j=1}^{J} p_{\cdot j} \left[\sum_{m=0}^{M^*} \lambda_m b_{jm}(a_{im} - a_{i'm}) \right]^2 \\ &= \sum_{j=1}^{J} \sum_{m=0}^{M^*} p_{\cdot j}\lambda_m^2 b_{jm}^2 \Big(a_{im} - a_{i'm} \Big)^2 \\ &= \sum_{m=0}^{M^*} \left(\sum_{j=1}^{J} p_{\cdot j} b_{jm}^2 \right) \Big(\lambda_m a_{im} - \lambda_m a_{i'm} \Big)^2 \end{aligned} \tag{2.43}$$

For m=0, the distance between two row profiles is given by

$$d_I^2(i,i') = \sum_{m=1}^{M^*} \Big(f_{im} - f_{i'm} \Big)^2 \tag{2.44}$$

Thus, the proximity between two row profiles along the m^{th} principal axis is the Euclidean distance given by

$$d_{I(m)}^2(i,i') = \sum_{m=1}^{M^*} \Big(f_{im} - f_{i'm} \Big)^2 \tag{2.45}$$

Using this logic, the proximity between two column profiles as a measure of their association, is given by

$$\begin{aligned} d_J^2 &= \sum_{i=1}^{I} \frac{1}{p_{\cdot}} \left(\frac{p_{ij}}{p_{\cdot j}} - \frac{p'_{ij}}{p_{\cdot j'}} \right)^2 \\ &= \sum_{m=1}^{M^*} \Big(g_{jm} - g_{j'm} \Big)^2 \end{aligned} \tag{2.46}$$

Hence, it follows that researchers may use correspondence analysis to discern how profiles within a variable relate, or more aptly, correspond to one another.

2.10.4 Interpretation of the Correspondence Plot

Inferring a correspondence plot by gauging the distance between the row and column points is considered controversial according to Roberts Jr (2000). However, it is a generally accepted principle that a measure of similarity between categorical profiles is indicated by a chi-squared distance between these two points. Hence, if two points in a correspondence plot are in close proximity they are considered to have similar profiles, and points that are at precisely the same location have the same profile. This is called the principle of *distributional equivalence*. Thus, if two rows have similar profiles, their distribution is similar across columns.

Within the scope of correspondence analysis, the rows and columns play a symmetric role as there is usually equal variation among their factors. Consequently, this allows for the row factors to be derived from the column factors and *vice-versa*. However, the proximity between the row points and column points, as a measure of their association, is more complex but owing to the *barycentric principle*, interpretation can be effected. The barycentric principle is encapsulated by (Husson et al., 2016, p. 4), as

> *"A column category point is, apart from scaling factors, the centroid of observations belonging to that category and a row point is also, apart from scaling factors, at the barycenter of the categories it belongs to."*

A barycenter refers to the average profile, and points in close proximity to the barycenter are similar to their average profile, while points located away from the barycenter are distinctly different from their average profile. Other terms used to describe a barycenter are *centre of gravity*, *centre of mass*, *mean vector*, *centroid* or *origin*. Nishisato (1980) argues that the barycentric principle is another way of introducing multiple correspondence analysis known as *dual scaling*. The method of multiple correspondence analysis is further discussed in Section 2.11.

It is therefore fairly evident to the researcher that simultaneous representation of the row and column profile in correspondence analysis is crucial. This is because it allows for effective comparison between the categories of qualitative variables represented in the rows and columns. In this respect, direct comparison between the row and column profile is subject to the conclusion drawn from the *transition formulae*. The transition formulae, according to Hill (1973), are equations that allow for the computation of profile coordinates where one set of coordinates may be calculated from the coordinates of the remaining variable.

2.10.5 Between Variable Distance

Consider the correspondence analysis problem in Equation 2.17 with the usual constraints as per Equation 2.18 on the singular vectors. As a way of linking the correspondence plot with the Pearson chi-squared statistics, the scores are re-scaled as per Equation 2.23. Using Equation 2.34 and Equation 2.23b:

$$f_{im}\lambda_m = \sum_{j=1}^{J} \frac{p_{ij}}{p_i.} b_{jm}\lambda_m \tag{2.47}$$

$$= \sum_{j=1}^{J} \frac{p_{ij}}{p_i.} g_{jm} \tag{2.48}$$

Thus, when the column profile coordinates are known, the row profile may be obtained as:

$$f_{im} = \frac{1}{\lambda_m} \sum_{j=1}^{J} \frac{p_{ij}}{p_i.} g_{jm} \tag{2.49}$$

and, *vice-versa*, the column profile coordinates on knowing the row profile

$$g_{jm} = \frac{1}{\lambda_m} \sum_{i=1}^{I} \frac{p_{ij}}{p.j} f_{im} \tag{2.50}$$

Thus, Equation 2.49 and Equation 2.50 are transition formulae and demonstrate how one profile is obtained from the other. The transition formulae may be expressed in vector form as

$$\boldsymbol{F}\boldsymbol{D}_\lambda = \boldsymbol{D}_I\boldsymbol{P}\boldsymbol{G} \tag{2.51}$$

$$\boldsymbol{G}\boldsymbol{D}_{\boldsymbol{\lambda}} = \boldsymbol{D}_J\boldsymbol{P}^T\boldsymbol{F} \tag{2.52}$$

A consequence of a particular categorical profile being a scaled function of the other, however, is that for a relatively large p_{ij} the column profile, g_{im} is heavily weighted,

directly influencing the row profile f_{im}. For a direct comparison between each categorical profile, consider the result of Goodman (1986) in which it was proved that instead of using the row and column profile as previously defined, it is more advantageous to use

$$\tilde{f}_{im} = \lambda_m^{\gamma} a_{im} = \frac{f_{im}}{\lambda_m^{\delta}} \qquad \tilde{g}_{jm} = \lambda_m^{\delta} b_{jm} = \frac{g_{jm}}{\lambda_m^{\gamma}} \tag{2.53}$$

for

$$\gamma + \delta = 1 \tag{2.54}$$

should one wish to effect a comparison between both the row and column profiles projected onto the same correspondence plot.

There are two scenarios in which Equation 2.54 is satisfied, namely when $\gamma = 1$ and $\delta = 0$, in which case the row profile coordinates are as Equation 2.23a, and the column profile are projected onto the correspondence plot using their standard profiles. In the same manner, when $\gamma = 0$ and $\delta = 1$, the column profile is as per Equation 2.23b where the rows are plotted using their standard profiles.

Now, for a direct comparison between the i^{th} row profile and the j^{th} column profile: the result of Goodman (1986) expressed in Equation 2.53 shows that from Equation 2.34 and Equation 2.35 a re-parameterised version of the row and column profiles in Equation 2.49 and Equation 2.50 become

$$\tilde{f}_{im} = \frac{1}{\lambda_{m^{2\delta}}} \sum_{j=1}^{J} \frac{p_{ij}}{p_{i\cdot}} \tilde{g}_{jm} \tag{2.55a}$$

$$\tilde{g}_{jm} = \frac{1}{\lambda_{m^{2\gamma}}} \sum_{i=1}^{I} \frac{p_{ij}}{p_{\cdot j}} \tilde{f}_{im} \tag{2.55b}$$

For the special case when $\gamma = \delta = \frac{1}{2}$, then Equation 2.55a is Equation 2.34 and

Equation 2.55b is Equation 2.35. The transition formulae in Equation 2.55 shows the relationship between the row and column, due to their shared entry within a contingency table, and subsequently projected onto a correspondence plot. Thus, one infers the comparison between the row and column profile via the transition formulae.

A comparison between the row and column profile for a large cell entry shared by a column and row implies that these categorical variables will be closer to one another, while the converse is true for a relatively small cell entry. In Section 2.11 we examine a multiple correspondence analysis which is able to accommodate several categorical variables in a higher order contingency table; in essence, multiple correspondence analysis is an extension of correspondence analysis.

2.11 Multiple Correspondence Analysis

The majority of early research conducted in the area of correspondence analysis focused on classical correspondence analysis applied to two-way contingency tables without much consideration given to higher order contingency tables. However, beginning in the early 1940s Louis Guttmann, a mathematician and sociologist, initiated discussion in this respect by exploring the field of *dual scaling*, more prominently referred to as *optimal scaling* postulating his findings in Guttman (1941). This would later become known as *multiple correspondence analysis*. Henceforth, the application of multiple correspondence analysis is conducted by transforming a contingency table into an *indicator matrix* or a *Burt matrix*. The latter is applied in Section 2.14.

This section commences by detailing the initial considerations for a three-way contingency table. Multiple correspondence analysis may be applied to any *multi-way* contingency table. However, to avoid any unwarranted complexities, the three-way

contingency table is considered. There are brief deliberations on the application of generalised singular value decomposition to two-way contingency tables as another method of conducting multiple correspondence analysis. We also briefly discuss, without mathematical rigour, the Tucker3 model by Tucker (1966), the PARAFAC model by Harshman (1970) and the CANDECOMP model by Caroll & Chang (1970).

2.11.1 The Data Structure

For a three-way contingency table, N, which is comprised of I rows, J columns and K tubes as per Kroonenberg (1989) or layers according to Kendall & Stuart (1979), the total number of observations with respect to these three variables is n. Within N, the $(i,j,k)^{th}$ entry is n_{ijk} $\forall\, i \in (1,2,...,I), j \in (1,2,....,J)$ and $k \in (1,2,...,K)$.

Let the probability associated with the $(i,j,k)^{th}$ cell be defined as

$$p_{ijk} = \frac{p_{ijk}}{n} \tag{2.56}$$

such that $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{ijk} = 1$

Furthermore, let $p_{i..}$ be the marginal probability for the i^{th} row with $\sum_{i=1}^{I} p_{i..} = 1$. In a similar manner, let $p_{.j.}$ be the marginal probability of the j^{th} column with $\sum_{j=1}^{J} p_{.j.} = 1$ and $p_{..k}$ be the marginal probability of the k^{th} layer where $\sum_{k=1}^{K} p_{..k} = 1$.

2.12 Multiple Correspondence Analysis - The Burt Matrix

A contingency table may be represented in the form of an *indicator matrix* in which the elements, n_{ijk}, are either 1 or 0. In these types of matrices, the rows represent individuals who have been categorised into the primitive contingency table while the columns are the categorical responses into which the individual was classified. To illustrate this point, consider the arbitrary 2×3 contingency table below which rep-

Table 2.4: Arbitrary two-way contingency table

Individual	Column 1	Column 2	Column 3	Total
Row 1	2	2	3	7
Row 2	1	1	2	4
Total	3	3	5	11

resent the frequency of individuals classified according to arbitrary attributes represented by the rows and columns.

The corresponding indicator matrix for Table 2.4, $\boldsymbol{Z}$, is presented in Table 2.5 and consists of eleven rows, as there are eleven individuals and five columns; one for each row and column.

Table 2.5: Indicator matrix of Table 2.4

Individual	Row 1	Row 2	Column 1	Column 2	Column 3
1	1	0	1	0	0
2	1	0	0	0	1
3	1	0	0	1	0
4	1	0	1	0	0
5	1	0	0	0	1
6	1	0	0	1	0
7	1	0	1	0	0
8	0	1	0	0	1
9	0	1	0	0	1
10	0	1	0	0	1
11	0	1	0	1	0

Examining Table 2.5, each row in the indicator matrix has two ones and three zeros as an indicator matrix of a multi-way (m-way) contingency table which consists of m $1^{'}s$ and the remaining entries are $0^{'}s$. Once the indicator matrix is constructed,

correspondence analysis is carried out on the two-way contingency table, N, using the indicator matrix $\boldsymbol{Z}$.

However, Burt (1950) states that a correspondence plot can be produced by directly analysing the actual results obtained, by way of a contingency table *in lieu* of the identity matrix. In this method of correspondence analysis, the *Burt matrix*, which is a product-sum matrix, B_t is analysed. The Burt-matrix has the following form.

$$B_t = \boldsymbol{Z}'\boldsymbol{Z} \tag{2.57}$$

The product-sum matrix, according to Burt (1950) is similar to the matrix of covariances between the categorical profiles, bar the fact that cell entries in $\boldsymbol{Z}$ are not standardised. For the analysis of the two-way contingency table, Equation 2.57 may be represented by

$$\boldsymbol{B}_t = \boldsymbol{Z}'\boldsymbol{Z} \begin{bmatrix} Z_1'Z_1 & Z_1'Z_2 \\ Z_2'Z_1 & Z_2'Z_2 \end{bmatrix} \tag{2.58}$$

An alternative form of the Burt matrix, for a two-way contingency table is

$$\boldsymbol{B}_t = \boldsymbol{Z}'\boldsymbol{Z} \begin{bmatrix} nD_I & N \\ N' & nD_J \end{bmatrix} \tag{2.59}$$

The elements of the off-diagonal in Equation 2.59 are sub-matrices, Z_q^T and Z_q, where $q \neq q'$. The product of the off-diagonal, $Z_q^T Z_q$, is a two-way contingency table which summarises the association attributes q and q' for the n individuals.

The corresponding Burt matrix for the arbitrary data presented in Table 2.4 is, as per Equation 2.59, given in Table 2.6. On closer examination of the Burt matrix above, and in accordance with Greenacre (2007), it is observed that the sub-tables have the same row margins in each set of horizontal tables and the same column margins in each set of vertical tables.

Table 2.6: Burt matrix for the arbitrary contingency Table in 2.4

Individual	Row 1	Row 2	Column 1	Column 2	Column 3
Row 1	7	0	2	2	3
Row 2	0	4	1	1	2
Column 1	2	1	3	0	0
Column 2	2	1	0	3	0
Column 3	3	2	0	0	5

2.12.1 Total Inertia

Multiple correspondence analysis using an indicator matrix creates several binary columns for each variable by partitioning the indicator matrix with the provision that only one column contains the 1 values. This method of coding leads to *artificial inflation* of the dimensions as one categorical variable is coded with several columns, the direct result of which is inflation of the inertia. This in turn, *deflates* the principle inertia of the first dimension, thus *underestimating* it.

To mitigate this two corrections are used, the first proposed by Benzécri (1979) and the second by Greenacre (1993). The corrections proposed account for the eigenvalues (squared singular) that are less than $\frac{1}{m}$, which is a consequence of the coding that led to the aforementioned additional dimensions. The first correction is known as Benzécri's correction and is detailed below. If λ_l^I is the eigenvalue of the indicator matrix, then the corrected eigenvalues, ${}_c\lambda_l^I$ are obtained as

$$\lambda_l^I = \begin{cases} \left[\left(\frac{m}{m-1}\right)\left(\lambda_l^I - \frac{1}{m}\right)\right]^2 & \text{if} \quad \lambda_l^I > \frac{1}{m} \\ 0 & \text{if} \quad \lambda_l^I > \frac{1}{m} \end{cases} \tag{2.60}$$

Using this formula provides a better estimate for the inertia extracted by each eigenvalue. A further correction, called Greenacre's correction, proposed by Greenacre (1993), involves evaluating the percentage of inertia with respect to the average inertia of the off-diagonal of the Burt matrix B_t.

Now, Greenacre (1984) showed that there is a link between the m^{th} singular values of the Burt matrix and the indicator matrix, λ_m^B and λ_m^Z respectively. Greenacre (1994) notes that the Benzécri's correction is quite liberal when adjusting for the quality of fit and proposes the following relationship

$$\lambda_m^B = \left(\lambda_m^Z\right)^2 \tag{2.61}$$

There are I responses for the row categories and J responses for the column categories. If Q is defined as the total number of categories, then $Q = I + J$. Hence, for Q categories, the total inertia for the diagonal sub-matrices in Table 2.6 as defined by Greenacre (1988) is

$$\mathcal{I}_{B_t} = \frac{Q-m}{m^2} \tag{2.62}$$

Using Greenacre's correction and denoting the average inertia by $\bar{\mathcal{I}}$, Greenacre's correction is

$$\bar{\mathcal{I}}_{B_t} = \frac{m}{m-1} \times \left(\sum_l \lambda_l^2 - \frac{Q-m}{m^2} \right) \tag{2.63}$$

The contribution to the percentage of inertia is obtained by the following ratio

$$\tau_c = \frac{{}_c\lambda}{\bar{\mathcal{I}}} \tag{2.64}$$

where ${}_c\lambda$ are the corrected eigenvalues.

2.12.2 Singular Value Decomposition

Previously, we discussed correspondence analysis by means of a two-way contingency table and applying singular value decomposition. There are many approaches to conducting correspondence analysis of multiple categorical data, and in this section a generalised form of the singular value decomposition in presented.

Consider the correspondence matrix from a three-way contingency table devoid of independence:

$$p_{ijk} = p_{i\cdot\cdot}p_{\cdot\mathrm{j}\cdot}p_{\cdot\cdot\mathrm{k}} \tag{2.65}$$

As with correspondence analysis, complete independence is not always guaranteed. Thus, as per Equation 2.11, a degree of deviation from complete independence, called *Pearson's three-way ratio*, is introduced in Equation 2.65 as follows

$$p_{ijk} = \alpha_{ijk}p_{i\cdot\cdot}p_{\cdot\mathrm{j}\cdot}p_{\cdot\cdot\mathrm{k}} \tag{2.66}$$

Insofar as the restrictions on the rows, columns and layers as per Section 2.11 is upheld, complete independence is achieved when $\alpha_{ijk} = 1$.

Furthermore, this degree of deviation is quantified by

$$\alpha_{ijk} = \frac{p_{ijk}}{p_{i\cdot\cdot}p_{\cdot j\cdot}p_{\cdot\cdot k}} \tag{2.67}$$

It should be stated that singular value decomposition as established under two-way contingency tables is not possible under three-way or multi-way contingency tables. However, strides have been made in the area of generalised singular value decomposition (GSVD) which seeks to extend two-way singular value decomposition to accommodate three-way and multi-way contingency tables.

To this end, Darroch (1974) proposes an *additive*, three-way interaction model

$$\frac{p_{ijk}}{p_{i\cdot\cdot}p_{\cdot j\cdot}p_{\cdot\cdot \mathrm{k}}} = \alpha_{ij} + \beta_{ik} + \gamma_{jk} \tag{2.68}$$

for some $\{\alpha_{ij}\}$, $\{\beta_{ik}\}$, $\{\gamma_{jk}\}$ which measure the departure from complete independence from a two variable independence for three variables.

There have been many attempts at conducting multiple correspondence analysis through GSVD of a multi-way contingency table that may be employed. The first such attempt was by Ledyard R.Tucker (1966) a psychometrician, which with orthogonal factors, is a three-way principal component analysis that allows for the extraction of a different number of factors in each mode.

Other strides into the field of GSVD saw the development of the **PARAFAC** (**PARA**llel **FAC**tor analysis) model proposed by Harshman (1970) and Harshman & Lundy (1984). This model is a generalisation of principle component analysis to higher order arrays originating in psychometric analysis. The **CANDECOMP** (**CAN**onical **DECOMP**osition) model was another such attempt at GSVD of multi-way contingency tables; it is similar to the **PARAFAC** model and was proposed independently by Caroll & Chang (1970), and is applied in the area of multidimensional scaling.

2.13 Interpretation of Multiple Correspondence Analysis

Interpretation of a multiple correspondence analysis plot, as with that of correspondence analysis, is contentiously based on the proximity between two points on the correspondence plot. Further similarity is noted in the interpretation of correspondence analysis and multiple correspondence analysis in that interpretation only becomes meaningful for points of the same set (comparison between rows with rows and columns with columns). The complexity lies in the comparison between variables wherein two scenarios need to be taken into consideration:

1. The proximity between the levels of different nominal variables implies that such levels appear closer together in observations.

2. Levels of the same nominal variable do not usually occur together, hence the proximity between levels indicate that groups of observations that share these characteristics are similar.

As such, interpretation of the multiple correspondence analysis plot is based on points found in the same direction from the centroid and in the same location of the Euclidean space. However, as Greenacre (1988) and Greenacre & Hastie (1987) note, distance as a measure of association is not a universally accepted principle in multiple correspondence analysis. The geometry of multiple correspondence analysis is therefore not a generalisation of the geometric principles governing correspondence analysis.

2.14 Application of Multiple Correspondence Analysis

The procedure **PROC CORRESP** with the **MCA** option invokes the multiple correspondence analysis procedure in **SAS** Version 9.4. Listing the variables using the **TABLES** statement without delimiting each variable with a comma produces a symmetric Burt table which is displayed in the output using the **OBSERVED** option. The **OUTC** option creates an output coordinate data set. The inertiae are adjusted according to Greenacre (1988) using the **GREENACRE** option resulting in a more realistic percentage of the inertia explained along each axis.

The inertia of component describes the amount of variation explained by that component. The inertia of a column describes how much the values for a specific category differs from the expected value on the assumption that there is no multicollinearity. The Greenacre adjusted inertia decomposed into fourteen components is presented in Table 2.7. The total inertia explained by the fourteen components is 73.59%. An interesting point to note is that the total inertia of the fourteen compo-

nents is not 1 or 100%. This is ascribed to the adjustment for inertia using Greenacre's correction detailed in Section 2.12.1.

Multiple correspondence analysis using the Burt matrix with an adjustment for inertia is able to explain at least 62.91% of the total inertia for the fourteen components. This may not be sufficient thus necessitating the addition of two more components; the third and fourth components increasing the cumulative proportion of inertia to 70.70%. The interpretations effected herein are based on points that are found in the same direction from the barycenter and in the same region of the Euclidean space.

A multiple correspondence plot projects all categories onto a Euclidean space with the first two dimensions plotted to assess the association among categories. The first dimension accounts for 7.668% of the variation in the data while the second dimension accounts for 5.596% of the variation. These inertiae appear low and indicate possible instability in the individual axes. From Figure 2.16 we observe that an HIV negative status in the lower hemisphere, is mostly associated with younger adult participants enrolled in the study. These individuals are also observed to be less optimistic regarding their perceived risk of infection as most of these individuals do not foresee themselves at an imminent risk of infection. These individuals have, in most circumstances, attained a higher education qualification from a tertiary institution placing them in a relatively unique position to be well informed and more knowledgeable about HIV preventative measures than most of their study counterparts. Furthermore, they are not observed as being involved in marital relations of any kind and do not advocate any HIV stigma that may be perpetuated by others within the study location. The converse is true to an extent for HIV positive individuals who are observed as being associated with older participants who are involved in marital relations, whether separated or legally married. In this respect, these individuals have a lower level of education and are associated with lacking adequate preventative HIV information.

Furthermore, participants who are deemed to lack knowledge of HIV prevention are associated with male adolescents who have not completed a secondary education. These participants are associated with moderate levels of stigmatisation. Highly knowledgeable individuals are observed to be associated with adults in their mid- to late- twenties, who appear pessimistic about their risk of infection, and possess moderate levels of HIV preventative information.

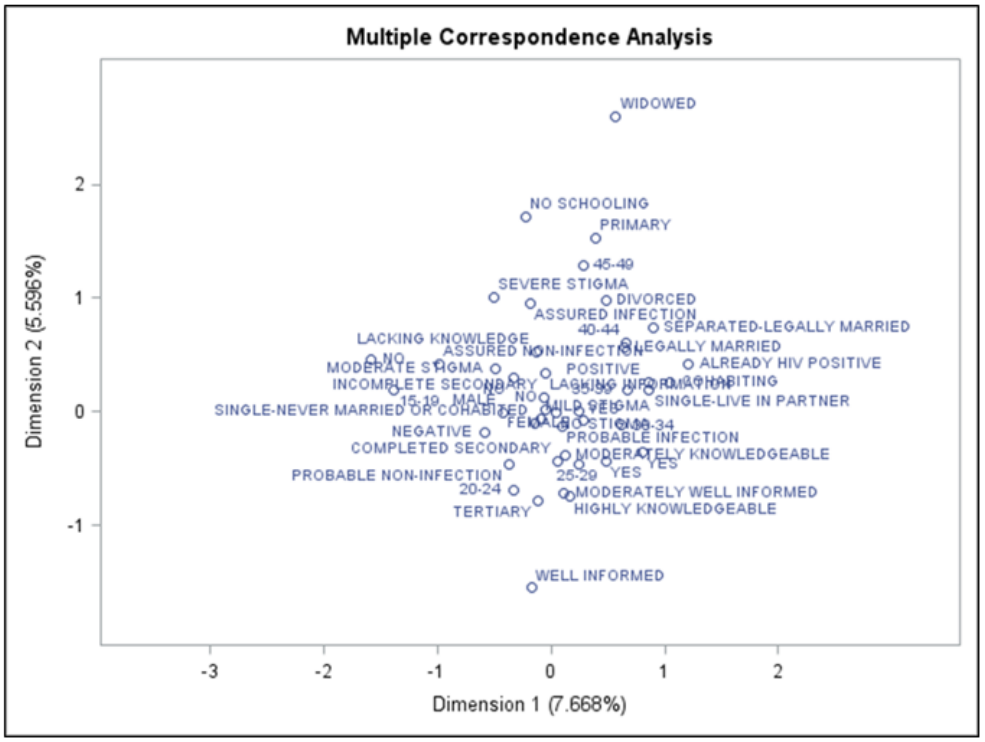

Figure 2.16: Multiple correspondence analysis for dimensions one and two

Table 2.7: Adjusted inertiae adjusted by Greenacre's correction.

Greenacre adjusted inertia decomposition					
Axis	Principal Inertia	Adjusted Inertia	Percent	Cumulative Percent	----+10----+20----+30----+40----+50
1	0.20449	0.01747	48.55	48.55	**********************************
2	0.14923	0.00517	14.36	62.91	*******
3	0.11997	0.00160	4.44	67.35	***
4	0.11517	0.00121	3.35	70.70	**
5	0.10418	0.00052	1.44	72.14	*
6	0.09709	0.00023	0.63	72.76	*
7	0.09566	0.00018	0.50	73.26	*
8	0.09023	0.00006	0.16	73.42	*
9	0.08824	0.00003	0.08	73.50	*
10	0.08700	0.00002	0.04	73.55	*
11	0.08655	0.00001	0.03	73.58	*
12	0.08487	0.00000	0.01	73.59	*
13	0.08425	0.00000	0.00	73.59	
14	0.08387	0.00000	0.00	73.59	

2.15 Summary and Discussion

This chapter investigated the patterns inherent in the data. Through exploratory data analysis, we attempted to provide an insightful view into the data and to give meaning to the sentiment expressed by (Tukey, 1977, p. vi):

> *"The greatest value of a picture is when it forces us to notice what we never expected to see."*

Applying the Guttmann scale which uses quantitative scaling, a score was constructed for each household to measure its socio-economic status. The rationale for applying the Guttmann scale was that there were a large number of variables pertaining to the socio-economic status of the household. Applying the Guttmann scale using Noble et al. (2006) provided a quantitative approach method to qualitative data, and the PIMD provided a method that makes use of the most salient household variables. In this endeavour, we were able to obtain a clear depiction of the socio-economic conditions in the study area. While the overall level of deprivation in the study area was not observed to be dire, approximately 9% of households were found to be within the 10% of severely deprived households. There were however, pockets of extreme deprivation noted in specific domains that constituted an index of multiple deprivation.

At the individual level there was serious disparity in terms of HIV prevalence noted with respect to gender, though not with respect to age. The overwhelming majority of participants attained only a basic education and a higher education was not sought by many study participants. In this regard, a further investigation into whether this is linked with certain socio-economic factors could be conducted.

On an inquiry into the behavioural and cognitive characteristics inherent in the population, it can be concluded that the importance of being well informed about HIV prevention measures cannot be understated. This is ascribed to the skewed preva-

lence of HIV between well informed individuals compared to those lacking adequate information or those moderately well informed. The prevalence of HIV among individuals who failed to practice safe sexual intercourse was higher than those who did.

With a keen focus on the individual level, further investigation was conducted to ascertain where the disparity in gender with respect to HIV prevalence lay. The socio-demographic characteristics related to gender revealed older males displaying higher prevalence with the same holding true for younger females. There was a negligible difference in HIV prevalence with respect to education which saw male tertiary graduates reporting the lowest prevalence while males who did not complete high school reported the highest HIV prevalence.

Multiple correspondence analysis proved to be a versatile, exploratory statistical technique for visualising contingency tables. From the application of multiple correspondence analysis, we inferred that HIV infection was closely associated with older participants who were not as highly educated as HIV negative individuals. These individuals were mostly involved in marital relations but lacked HIV preventative knowledge and adequate HIV preventative information. Individuals who were HIV negative were younger, better educated with exceptional knowledge and information gathering skills on HIV prevention.

The following three chapters give an overview of the statistical models that were applied to the HIPSS baseline data as well as a presentation of the results of each model.

Chapter 3

The Survey Logistic Regression Model

3.1 Introduction

Multistage sampling is a technique often employed in survey designs for reasons largely attributed to administrative purposes and cost effectiveness. Multistage sampling is a tiered sampling technique that involves, first, selecting an initial sample, called a primary sampling unit (PSU). Suppose a population is aggregated, then the PSU involves selecting such aggregates with a view to selecting individual units within the PSU. Once the individual elements are sampled from within these aggregations, it results in a secondary stage sampling unit (SSU). Repetitive sampling in this manner is termed multistage sampling.

An unavoidable and inevitable consequence of multistage sampling is the unequal probability of selection of sampling units at all or at some stage in the sampling process resulting in sampling bias (Pffeferman et al., 1998). As such, *sampling weights*, which may be thought of as a number of observations represented by a unit in a population, are introduced in order to compensate for the bias associated with unequal probability selection (Wang, 2013).

The role of sampling weights is a topic of contention among researchers since there are numerous complexities associated with their introduction into a regression model. Estimating what may be considered rudimentary concepts in the absence of sampling weights, such as sample mean and standard error may prove complex when sampling weights are involved. Furthermore, there is much deliberation among researchers as to what actually constitutes a sampling weight. It is a generally accepted principle that a sample weight is the reciprocal of the probability of selection. This, however, is not a view shared by everybody. Establishing sample weights is subject to adjusting for non-responses, post-stratification or other ancillary adjustments by perusing supplemental data information.

Under linear regression and binary logistic regression, maximum likelihood estimation is usually applied to estimate the regression coefficients. This method, however, is well suited to predictors that follow the assumption of normality which is not always the case. Furthermore, according to Skinner et al. (1989), a sampling design that involves cluster sampling may induce correlation among observations, and ignoring these may result in inconsistent statistical tests and consequently, superfluous results.

To develop a survey logistic regression model, a weighted maximum likelihood function is employed to provide estimates for the regression coefficients. An *a priori* consideration however is that sampling weights have to be adjusted to compensate for non-responses and other pre and post data collection discrepancies with respect to the population. The process does not terminate at the construction of a predictive model; the ultimate aim of any modelling process. Assessing the model's validity by testing the model's goodness-of-fit allows one to gauge the predictive accuracy of the model.

3.2 Probability Sampling Weights

Sampling weights or survey weights are positive values that are linked to individual observational units within a sample. As previously stated, the justification of a sampling weight is rooted in the unequal chance of selection of a sampling unit. When constructing an individual weight, it must be borne in mind that a sample weight ought to represent the frequency which that particular sampling unit represents in the population from which it is drawn. It thus stands to reason that the sum of the sample weights should provide an estimate of the population size N.

Sampling weights are usually considered to be the reciprocal of the probability of selection of a particular observation. If the i^{th} unit in a population has a probability p_i of selection into a sample, then the weight, w_i, associated with the i^{th} unit is, according to Kish (1965), given by:

$$w_i = \frac{1}{p_i} \tag{3.1}$$

However, drawing a sample using simple random sampling ensures that there is an equal likelihood of selection for each sampling unit and Equation 3.1 holds true. It is thus less complex to estimate population means, population proportions, and population totals. However, other sampling techniques could be used resulting in an arbitrary probability of selection for each sampling unit. As such, Horvitz & Thompson (1952), propose an unbiased estimator of the population total and by extension, the population mean detailed below. This is referred to as the Horvitz-Thompson estimator.

Let T represent the population total and $\hat{T}$, the associated unbiased estimator of T. Furthermore, let $(x_1, x_2, ..., x_n)$ be a sample drawn from a finite population of size N. The Horvitz-Thompson estimator for the population total is thus given by

$$\hat{T} = \sum_{i=1}^{n} w_i x_i \tag{3.2}$$

Hence, it follows from Equation 3.2 that an estimator for the population mean, $\bar{x}$, will be given by

$$\frac{\hat{T}}{N} = \frac{\sum_{i=1}^{n} w_i x_i}{N} = \bar{x} \tag{3.3}$$

However, the sum of the individual weights is the population size N. Hence, the Horvitz-Thompson estimator for the population mean, $\bar{x}$ will be given by

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{3.4}$$

Whereas no formal procedural protocol exists for computing the sample weights, the statistical justification is quite evident. However, this process can be quantified in three stages and are listed below as per the Global Adult Tobacco Survey Collaborative Group (2010).

- The first stage is to obtain the design weights or sampling weights such that the sum of these weights correspond to the population size.
- Thereafter, as non-responses are inevitable in survey data, *adjusting* these weights is fundamental to compensate for the loss in data due to non-responses and other both pre- and post- data collection inconsistencies.
- The final stage is referred to as *calibration* wherein design weights (and those adjusted for non-responses) are adjusted to match the population totals from which the sample is drawn. Thus, distinct homogeneous groups within the sample may match their respective population totals. This is also referred to as *post-stratification*, a special case of *calibration*.

3.3 Adjusted Weights

Evident from the weighting process listed above, adjusting sample weights to account for non-responses and other sampling inconsistencies, is imperative in the weighting process. These adjusted weights will be denoted by w_i^*. However, prior to detailing the estimation of population parameters under adjusted weights, the rationale of a *super-population* is first introduced. Cassel et al. (1977) argue that inferential statistics based on a super-population is vital to effect understanding of the process under investigation.

The concept of a super-population was first proposed by Deming & Stephan (1941), describing it as an infinite population from which the finite population is drawn and is in itself a sample (Graubard & Korn, 2002). It is usually the case that the target of inferential statistics is the super-population based on results emanating from the finite population. Hence, weighting in the context of survey sampling is crucial if inferences are to be projected to a population accurately. (Pfeffermann et al., 1998, p. 1087) regard the super-population as an intrinsic factor in the collection of survey data stating:

> *"Survey data may be viewed as the outcome of two random processes: The process of generating the values in the finite population, often referred to as the super-population model, and the process of selecting the sample data from the finite population values, known as the sample selection mechanism."*

The process surrounding the generation of population data (the super-population model) and the process employed to select the sample from the population are thus co-requisites for any realistic statistical model arising from sample survey data.

Having defined the rationale behind the super-population, we derive the estimators of the mean and variance of the super-population as delineated by Potthoff et al. (1992).

Let y_i denote the response of the i^{th} sampling unit of the super-population. For a super-population model, the responses per sampling unit are assumed to be independent random variables. The following expressions are pre-defined for the mean, m and variance, v, for the i^{th} sampling unit

$$m_i = E(y_i) \qquad v_i = var(y_i) \quad \text{for i} \in [1, n] \tag{3.5}$$

where n is the sample size.

Within a super-population, there are two sources of stochasticity. These arise from the randomisation of survey processes and the parameters which change over time, hence m_i and v_i are interpreted instantaneously. Within the finite population model, sampling all the units in the population would yield a variability of zero thus violating the stochastic assumption that $v_i > 0$.

Define the mean, m, and variance, v respectively for the super-population as follows

$$m = \frac{\sum_{i=1}^{n} w_i m_i}{\sum_{i=1}^{n} w_i}, \qquad v = \frac{\sum_{i=1}^{n} w_i^2 v_i}{\sum_{i=1}^{n} w_i^2} \tag{3.6}$$

If y_i is used as an unbiased estimator for m_i then an unbiased estimator for m, denoted by $\hat{m}$ as a function of y_i is given by

$$\hat{m} = \left(\frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \right) \tag{3.7}$$

where y_i is the realized value and is the normalization factor such that $\hat{m}$ is an unbiased estimator of m.

As such, from Equation 3.7 it is deduced that

$$var(\hat{m}) = var\left(\frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}\right) = \left(\frac{\sum_{i=1}^{n} w_i^2}{\left(\sum_{i=1}^{n} w_i\right)^2}\right) v_i \tag{3.8}$$

Now, consider a new set of weights defined by

$$w_i^* = \left(\frac{\hat{n} w_i}{\sum_{i=1}^{n} w_i}\right) \tag{3.9}$$

where $\hat{n}$ is the effective sample size and is defined as the estimated sample size to effect the same precision if simple random sampling were to be used. The effective sample size, $\hat{n}$, as proposed by Kish (1965) is

$$\hat{n} = \left(\frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} w_i^2}\right) \tag{3.10}$$

Through an elementary algebraic manipulation applied simultaneously on Equation 3.9 and Equation 3.10 it can be shown that $\hat{n}$ has the following property

$$\hat{n} = \sum_{i=1}^{n} w_i = \sum_{i=1}^{n} w_i^2 \tag{3.11}$$

Using the relationship in Equation 3.11, Equation 3.6 together with Equation 3.7 and Equation 3.8 may be expressed respectively as

$$m = \frac{\sum_{i=1}^{n} w_i^* m_i}{\hat{n}}, \quad \hat{m} = \frac{\sum_{i=1}^{n} w_i^* y_i}{\hat{n}}, \quad v = \frac{\sum_{i=1}^{n} (w_i^*)^2 v_i}{\hat{n}}, \quad var(\hat{m}) = \frac{v}{\hat{n}} \tag{3.12}$$

3.4 The Survey Logistic Regression Model

Categorical outcomes, commonplace in surveys, are modelled using binary logistic regression. This provides a probability that relates to the likelihood of an event subject to a set of covariates. However, not all surveys select sampling units using simple random sampling. Multistage sampling, often employed in a hierarchical manner, introduces varying sampling techniques at each sampling stage.

The possibility of correlation among observations is often highly likely owing to the peculiarities of such varying sampling techniques. In this instance, analysis using the ordinary logistic regression becomes redundant; suitably adjusting the ordinary logistic model for the cluster effect is thus necessary (Wilson & Lorenz, 2015) and (Perera et al., 2014).

Furthermore, Rao & Scott (1984) state that employing complex survey designs with post-stratification, adjustments for non-responses, clustering and/or unequal weighting, produces inconsistent estimates unless specialised techniques are used. These techniques are examined below in the context of the survey logistic regression model for complex survey designs as documented by Hosmer et al. (2013) and Roberts et al. (1987).

Suppose that a population is divided into $k = 1, 2, ..., K$ strata within which there are $j = 1, 2, ..., M_k$ primary sampling units (PSU). If we assume that the observed data comprises n_{kj} elements from m_k PSUs drawn from the k^{th} stratum, then the total sample size is given by $n = \sum_{k=1}^{K} \sum_{j=1}^{m_k} n_{kj}$. The sample weight for the kji^{th} observation will be denoted by w_{kji}. Let π_{kji} be the probability of selection of the i^{th} sampling unit from the j^{th} PSU located in the k^{th} strata. Thus, the dichotomous outcome, y_{kji} is related to the vector of covariates, $\boldsymbol{x}'_{kji}$ by the survey logistic regression model stated below.

$$\ln\left(\frac{\pi_{kji}}{1-\pi_{kji}}\right) = x'_{kji}\boldsymbol{\beta}, \quad y \in [0,1] \tag{3.13}$$

where

$$\pi_{kji} = \frac{\exp(x'_{kji}\boldsymbol{\beta})}{1 + \exp(x'_{kji}\boldsymbol{\beta})} \tag{3.14}$$

and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_{t+1})'_{(t+1)\times 1}$ is a column vector of regression coefficients.

3.5 Weighted Maximum Likelihood Estimation

Complex sample design often involves multistage cluster sampling which results in differential sampling weights. Adjustments for non-responses and post-stratification are among the reasons for differential sampling weights. As such, the method of maximum likelihood estimation will not, in general, suffice and the method of pseudo-maximum likelihood estimation is employed (Graubard et al., 1997) and (Skinner et al., 1989).

The pseudo-likelihood function, $\mathcal{L}_p$, is then constructed as per the product of the individual contributions to the likelihood function. Thus, the pseudo-likelihood function, as adapted from Archer et al. (2007) and Zhang et al. (2018), will be given by

$$\mathcal{L}_p(\boldsymbol{\beta}) = \prod_{k=1}^{K}\prod_{j=1}^{m_k}\prod_{i=1}^{n_{kj}} \pi_{kji}^{w_{kji}\times y_{kji}}(1-\pi_{kji})^{w_{kji}\times(1-y_{kji})} \tag{3.15}$$

As with the method of maximum likelihood estimation, the partial derivative of the pseudo log-likelihood function with respect to $\boldsymbol{\beta}$ is obtained to obtain $\widehat{\boldsymbol{\beta}}$, an unbiased estimator for $\boldsymbol{\beta}$. The process in which the regression coefficients are obtained is detailed below. The method is initiated by obtaining the pseudo log-likelihood function which maximises Equation 3.15 to obtain the best linear unbiased estimate of $\boldsymbol{\beta}$

$$\ln \mathcal{L}_p(\boldsymbol{\beta}) = \sum_{k=1}^{K}\sum_{j=1}^{m_k}\sum_{i=1}^{n_{kj}} [w_{kji} \times y_{kji}] \times \ln(\pi_{kji}) + [w_{kji} \times y_{kji}] \times \ln(1-\pi_{kji}) \tag{3.16}$$

Obtaining the partial derivative with respect to $\boldsymbol{\beta}$, and the unknown regression coefficients in Equation 3.16, the following system of $t+1$ score equations are obtained.

$$\frac{\partial \ln \mathcal{L}_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \left\{ \begin{array}{l} f_{\text{new}}(\beta_0) = \sum_{k=1}^{K} \sum_{j=1}^{m_k} \sum_{i=1}^{n_{kj}} X_{kj1} w_{kj1} (y_{kj1} - \pi_{kj1}) = 0 \\ f_{\text{new}}(\beta_1) = \sum_{k=1}^{K} \sum_{j=1}^{m_k} \sum_{i=1}^{n_{kj}} X_{kj2} w_{kj2} (y_{kj2} - \pi_{kj2}) = 0 \\ \quad . \\ \quad . \\ \quad . \\ f_{\text{new}}(\beta_{t+1}) = \sum_{k=1}^{K} \sum_{j=1}^{m_k} \sum_{i=1}^{n_{kj}} X_{kji} w_{kji} (y_{kji} - \pi_{kji}) = 0 \end{array} \right\} = \boldsymbol{X'W(Y-\pi)} = 0 \tag{3.17}$$

Employing the Newton-Raphson iterative method we obtain an estimate for $\widehat{\boldsymbol{\beta}}$.

$$\widehat{\boldsymbol{\beta}}_{(t+1)} = \widehat{\boldsymbol{\beta}}_{(t)} - \boldsymbol{H}^{-1}\mathbf{q} \tag{3.18}$$

where $\boldsymbol{H} = \boldsymbol{X'DX}$ and $\boldsymbol{q} = \boldsymbol{X'D(Y-\pi(t))}$.

The resulting estimate for $\boldsymbol{\beta}$ is

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{(t+1)} &= \widehat{\boldsymbol{\beta}}_{(t)} + \boldsymbol{(X'DX)^{-1}X'D(Y-\pi)} \\ &= \boldsymbol{(X'DX)^{-1}X'D(X\beta + D^{-1}(Y-\pi))} \end{aligned} \tag{3.19}$$

Hence, by solving Equation 3.19, the following solution is deduced for $\widehat{\boldsymbol{\beta}}$

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{(X'DX)^{-1}X'DU} \tag{3.20}$$

where $\boldsymbol{U} = \boldsymbol{X\beta + D^{-1}(Y-\pi)}$, $\boldsymbol{D} = \boldsymbol{WV(t)}$ and;

$$\boldsymbol{X} = \begin{pmatrix} X_{110} & X_{111} & \cdots & X_{11i} \\ X_{210} & X_{211} & \cdots & X_{22i} \\ \vdots & \vdots & \ddots & \vdots \\ X_{kj0} & X_{kj1} & \cdots & X_{kji} \end{pmatrix}_{n\times(t+1)} , \boldsymbol{Y} = \begin{pmatrix} Y_{111} \\ Y_{112} \\ \vdots \\ Y_{kji} \end{pmatrix}_{n\times 1}$$

Furthermore,

$$\boldsymbol{\pi(t)} = \begin{pmatrix} \pi_{111} \\ \pi_{112} \\ \vdots \\ \pi_{kji} \end{pmatrix}_{n\times 1} , \widehat{\boldsymbol{\beta}}\boldsymbol{(t)} = \begin{pmatrix} \widehat{\beta}_0(t) \\ \widehat{\beta}_1(t) \\ \widehat{\beta}_2(t) \\ \vdots \\ \widehat{\beta}_{t+1}(t) \end{pmatrix}_{(t+1)\times 1}$$

and,

$$\boldsymbol{V(t)} = \begin{pmatrix} \pi_{kj1}(1-\pi_{kj1}) & & & \\ & \pi_{kj2}(1-\pi_{kj2}) & & \\ & & \ddots & \\ & & & \pi_{kji}(1-\pi_{kji}) \end{pmatrix}_{n\times n}$$

The sample weight, w_{kji}, is the weight associated with the kji^{th} sample unit and is expressed as the following diagonal array.

$$\boldsymbol{W} = \begin{pmatrix} W_{kj1} & & & \\ & W_{kj2} & & \\ & & \ddots & \\ & & & W_{kji} \end{pmatrix}_{n\times n}$$

Consequently, an appropriate estimate for the $\mathrm{Var}(\boldsymbol{\beta})$ is:

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X'DX})^{-1}\boldsymbol{S}(\boldsymbol{X'DX})^{-1} \tag{3.21}$$

3.6 Test for Model Goodness-of-Fit

The likelihood function, expressed in Equation 3.16, represents an approximation of the true likelihood function. It therefore stands to reason that inferences about regression parameters should be based on the univariable or multivariable Wald test statistic (Hosmer et al., 2013). Conventional methods such as the Wald test used in ordinary logistic regression tend to conflate the survey weights with actual observations and therefore do not correctly assess the model fit.

The use of an adjusted Wald test accounting for complex survey sample design by Thomas & Rao (1987) and Korn & Graubard (1990) resulted in a test with improved adherence to α level of significance compared to the standard Wald test. Hence, complex sample surveys introduce a new dynamic in the execution of the Wald test, necessitating the use of a modified Wald test.

Furthermore, under ordinary logistic regression, Hosmer & Lemeshow (1980) proposed grouping cases in ascending order as per their predicted probability. Thereafter, the observed and expected frequencies are computed and the Pearson chi square test is applied. However, the pitfall of this test lies in the grouping as there exist no theoretical guide in this respect. This was not apparent until conventional statistical software allowed users to specify the number of groups rather than classifying them into deciles by default (Allison, 2013).

Assessing the overall model fit is an important stage in statistical modelling and assists in deciding whether a model is correctly specified. There are multiple tests

in this respect. Conducting these tests is relatively simple in the context of ordinary sampling procedures. Under a multistage sample design, certain modifications are incorporated into tests for goodness-of-fit that improve overall suitability. Two such modifications to existing goodness-of-fit tests are detailed below.

3.6.1 The Modified Wald Test

Archer et al. (2007) proposed an extension to the decile of risk test which incorporates a modified form of the Wald test to assess the model fit. This process is described below as detailed by Hosmer et al. (2013).

1. Let $s = \sum_{k=1}^{K} m_k - K$ represent the total number of primary sampling units per strata less the total number of strata. To conduct the modified Wald test, construct at most g, for $g = 1, 2, ..., 10$, $(s+2)$ groups such that the total sample weight per group is approximately 10% of the total sample weight.

2. Calculate the weighted mean of the model's residuals using the sample weights in each of the groups formed in step one above.

$$\widehat{M}_k = \frac{\sum_{k=1}^{K}\sum_{j=1}^{m_k}\sum_{i=1}^{n_{kj}} w_{kji} r_{kji}}{\sum_{k=1}^{K}\sum_{j=1}^{m_k}\sum_{i=1}^{n_{kj}} w_{kji}} \quad \text{for k} \in [1, s+2) \tag{3.22}$$

   The model's residuals are estimated from $\hat{r}_{kji}$ and is calculated as follows

$$\widehat{r}_{kji} = (y - \widehat{\pi}_{kji}) \tag{3.23}$$

   This is considered a crucial step in assessing the model fit; if the weighted means of the residuals differ significantly from 0, the model is not a good fit of the responses.

3. From Equation 3.22 a linearised estimator, as detailed by Archer (2001), of the covariance matrix $\widehat{\boldsymbol{V}}(\widehat{\boldsymbol{M}})$ is constructed from $\widehat{M}_k = (M_1, M_2, ..., M_k)'$.

From this the modified Wald test statistic is formulated and given by

$$\widehat{W}_{\hat{M}} = \widehat{\boldsymbol{M}}'[\widehat{\boldsymbol{V}}(\widehat{\boldsymbol{M}})]^{-1}\widehat{\boldsymbol{M}} \tag{3.24}$$

4. Using Equation 3.24, the hypothesis

$$H_0 : M_1 = M_2 = ... = M_k = 0$$
$$H_1 : \text{At least one } M_k \neq 0 \quad \text{for k} \in [1, s+2)$$

is tested as per the following test statistic

$$F_{\hat{M}} = \frac{s-g+2}{s \times g} \sim F_{(g-1),(f-g+2)} \tag{3.25}$$

5. The rejection region, RR, given by RR $= \{F_{(g-1),(f-g+2)} > F_{\hat{M}}\}$, is used to test the hypothesis and draw it to an appropriate conclusion.

3.6.2 The Rao-Scott Correction to the Likelihood Ratio Test

It is inevitable that samples within the sample PSU, drawn under complex survey designs, will result in highly correlated observations. Adjustments to compensate for non-responses and other sample anomalies can result in differing sample weights, violating the assumption of independence and identical distribution among the responses (Scott, 2007). Selecting a simple representation of the data and assessing the viability of the model thus proves difficult for samples drawn under complex sampling schemes.

This section examines the Rao-Scott first corrections to the chi-squared tests for contingency tables wherein cell proportions are obtained by complex sampling procedures adapted from the SAS Institute Inc. (2016).

Testing the Global Null Hypothesis

Suppose we have the cumulative model whose parameters are expressed as $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$ where $\boldsymbol{\alpha}$ represent the model parameters and $\boldsymbol{\beta}$, the regression coefficients. Let r denote the number of restrictions imposed on $\boldsymbol{\theta}$, then $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_k)'$ where $r = k$.

The *global null hypothesis* refers to the null hypothesis that is investigated to determine the significance of all the explanatory variables. Testing this null hypothesis determines if all explanatory variables may be excluded from the model and thus contain only the intercept terms. The global null hypothesis tested will be given by

$$\mathrm{H}_0 : \boldsymbol{\beta} = 0 \tag{3.26}$$

$$\mathrm{H}_\mathrm{a} : \boldsymbol{\beta} \neq 0 \tag{3.27}$$

Denote $\hat{V}_{rr}(\hat{\boldsymbol{\theta}})$ to represent the estimated covariance matrix of $\hat{\boldsymbol{\theta}}$ of the sample design and $\hat{V}_{rr}^{srs}(\hat{\boldsymbol{\theta}})$ to represent the estimated covariance matrix of $\hat{\boldsymbol{\theta}}$ under simple random sampling. As such the *design effect matrix*, E, is defined as

$$E = \hat{V}_{rr}(\hat{\boldsymbol{\theta}}) \left(\hat{V}_{rr}^{srs}(\hat{\boldsymbol{\theta}}) \right)^{-1} \tag{3.28}$$

where

$$\hat{V}_{rr}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} var(\hat{\theta}_1, \hat{\theta}_1) & cov(\hat{\theta}_1, \hat{\theta}_2) & \cdots & cov(\hat{\theta}_1, \hat{\theta}_k) \\ cov(\hat{\theta}_2, \hat{\theta}_1) & var(\hat{\theta}_2, \hat{\theta}_2) & \cdots & cov(\hat{\theta}_2, \hat{\theta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\hat{\theta}_k, \hat{\theta}_1) & cov(\hat{\theta}_k, \hat{\theta}_2) & \cdots & var(\hat{\theta}_k, \hat{\theta}_k) \end{pmatrix}$$

$$\hat{V}_{rr}^{srs}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} var(\hat{\theta}_1, \hat{\theta}_1) & cov(\hat{\theta}_1, \hat{\theta}_2) & \cdots & cov(\hat{\theta}_1, \hat{\theta}_k) \\ cov(\hat{\theta}_2, \hat{\theta}_1) & var(\hat{\theta}_2, \hat{\theta}_2) & \cdots & cov(\hat{\theta}_2, \hat{\theta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\hat{\theta}_k, \hat{\theta}_1) & cov(\hat{\theta}_k, \hat{\theta}_2) & \cdots & var(\hat{\theta}_k, \hat{\theta}_k) \end{pmatrix}_{SRS}$$

Furthermore, the estimated covariance matrices partitioned by the r slope parameters are given below. Let r^* be the rank of E wherein the positive eigenvalues of E are given by $\delta_k \ni \delta_1 \geq \delta_2 \geq ... \geq \delta_r^* > 0$. Now, the likelihood ratio test under the ordinary logistic regression procedure is

$$Q_{\chi^2} = -2 \log \left[\frac{L(\widehat{\boldsymbol{\theta}})}{L(\widehat{\boldsymbol{\theta}}_{H_0})} \right] = -2[\log L(\widehat{\boldsymbol{\theta}}) - \log L(\widehat{\boldsymbol{\theta}}_{H_0})] \sim F_{(r,\infty)} \tag{3.29}$$

However, to account for the influence of clustering or stratification, Rao & Scott (1984) and Rao & Scott (1987) propose adjustments to the likelihood ratio test. These are discussed below.

The Rao-Scott First Order Design Correction

To address the influence of complex survey design on the likelihood ratio test, Rao & Scott (1984) suggest the first-order correction to the chi-square statistic as

$$Q_{RS_1} = \frac{Q_{\chi^2}}{\bar{\delta}} \sim \chi^2_{r^*} \tag{3.30}$$

where $\bar{\delta}$ is the first-order design correction and is the average of the positive eigenvalues of E, expressed as follows

$$\bar{\delta} = \sum_{k=1}^{r^*} \frac{\delta_k}{r^*} \tag{3.31}$$

The corresponding F test statistic is subsequently given by

$$F_{RS_1} = \frac{Q_{RS_1}}{r^*} \sim F_{(r^*, df \times r^*)} \tag{3.32}$$

The Rao-Scott Second Order Design Correction

Rao & Scott (1987) further suggest the second order Rao-Scott correction to the chi-square statistic as follows

$$Q_{RS_2} = \frac{Q_{RS_1}}{(1+\hat{a}^2)} \tag{3.33}$$

where Q_{RS_1} is the first-order Rao-Scott chi-square statistic and the *second order design correction* , $\hat{a}^2$, is given below, and is determined from the coefficient of variation of the eigenvalues of the design matrix E as

$$\hat{a}^2 = \frac{1}{r^* - 1} \sum_{k=1}^{r^*} \frac{(\delta_k - \bar{\delta})}{\bar{\delta^2}} \tag{3.34}$$

The corresponding F statistic is

$$F_{RS_2} = \frac{Q_{RS_2(1+\hat{a}^2)}}{r^*} \sim F\left(\frac{r^*}{1+\hat{a}^2}, \frac{df \times r^*}{1+\hat{a}^2}\right) \tag{3.35}$$

in which *df* is the design degrees of freedom whose computation is guided as follows

$$df = \begin{cases} \tilde{n} - K & \text{if the design contains clusters,} \\ n - K & \text{if the design does not contain clusters.} \end{cases}$$

where $\tilde{n}$ is the total number of clusters, if clustering is present in the survey design. Furthermore, n is the total sample size and K is the number of strata if stratification is present, else $K = 1$.

3.7 Survey Logistic Regression Applied to the HIPSS Baseline Data

The analyses presented in this section were conducted using **SAS** Version 9.4. The procedure **PROC SURVEYLOGISTIC** is employed in **SAS** to account for a multistage sampling design. The **PROC SURVEYLOGISTIC** procedure makes provision for in-

clusion of clusters strata and sampling weights to be specified using the **CLUS** – **TER**, **STRATA** and **WEIGHT** commands respectively. The socio-economic and socio-demographic variables explored in Chapter 2 formed the explanatory variables of the model that was used to model an individual's HIV status at baseline.

In order to construct the model, all fixed effect predictor variables were established and included in the model together with selected two-way interactions, and consequently selected higher order interactions were explored and included in the model. Table 3.1 summarises the final survey logistic regression model.

In constructing a regression model, assessing the overall goodness-of-fit and the predictive accuracy of the model is a crucial stage in the modelling process. Akaike's Information Criterion (AIC) and Bayes Information Criterion (BIC) are two approaches used in assessing the model's relative goodness-of-fit (or lack thereof). Within the context of weighted logistic regression, the Archer-Lemeshow test as detailed in Section 3.6 is used to assess the model's goodness-of-fit. This, as of late, has not been incorporated into the **PROC SURVEYLOGISTIC** procedure in **SAS** Version 9.4.

Assessing the goodness-of-fit under a complex survey design thus is a slight deviation from conventional assessments of the model fit in which the likelihood ratio test is employed. To account for the influence of clustering and stratification, the Rao-Scott second order design correction is applied. The *corrected* likelihood ratio test is then assessed as to the goodness-of-fit. In this study a Rao-Scott second order design correction, which is produced by **SAS** Version 9.4 in the **PROC SURVEYLOGISTIC** procedure, of 0.0002 was applied to the likelihood ratio test. Consequently, at a 5% level of significance, the global null hypothesis in Equation 3.26 is rejected, indicating that the model was an appropriate fit to the data.

The predictive power of a survey logistic regression model is an additional consideration that one has to take into account. The predictive accuracy of the model is

indicative of the probability that the outcome is correctly predicted subject to the independent regressors. In this instance, the concordance index, *c*, which is equal to the area under the receiver operating curve (ROC), is used to gauge the predictive power under survey logistic regression. The concordance index, *c*, is produced in the **PROC SURVEYLOGISTIC** procedure and is given by

$$c = [n_c + 0.5(t - n_c - n_d)t^{-1}] \tag{3.36}$$

Within the context of ordinary logistic regression, the area under ROC is used to assess the predictive ability of the model. Thus, in the case of a binary response variable, the concordance index provides an estimate of the area under the ROC curve (Hanley & McNeil, 1982) and (Agresti, 2007). With respect to Equation 3.36, n is the total number of observations in the data set and t refers to the number of pairs in the data given as follows

$$t = \frac{n(n-1)}{2} \tag{3.37}$$

And n_c refers to the number of concordant pairs. A concordant pair of observations may be described as a pair of observations wherein a lower order response has a higher predicted mean than that of those responses which are higher ordered. Conversely, n_d is the number of discordant pairs and $t - n_c - n_d$ is the number of tied pairs. These are pairs of observations with different responses which cannot be characterised as concordant or discordant. N is the sum of the observed frequencies.

Mathematically this may be expressed as follows. Consider a pair of observations (r_1, c_1) and (r_2, c_2). According to Rudolfer (2002) and Agresti (1990) (r_1, c_1) and (r_2, c_2) are

$$\begin{aligned} \textbf{concordant} \quad & \textbf{if} \quad (r_1, c_1)(r_2, c_2) > 0 \\ \textbf{discordant} \quad & \textbf{if} \quad (r_1, c_1)(r_2, c_2) < 0 \\ \textbf{tied} \quad & \textbf{if} \quad (r_1, c_1)(r_2, c_2) = 0 \end{aligned}$$

A concordant pair is thus one in which a subject ranking higher on *r* also ranks *higher* on *c*, and discordant if the subject ranking *higher* on *r* ranks *lower* on *c* and tied otherwise. According to Agresti (2007) and Uno et al. (2019) the concordance index must be within the range of 0.5 to 1, with the former an indication of no correlation and the latter, complete correlation.

In addition to the concordance index, three other methods can be used to gauge the model's predictive accuracy; Somer's D (SD), Goodman-Kruskal Gamma (GKG), and Kendall's Tau-a (KT) - all of which are produced in the **PROC SURVEYLOGI − STIC** procedure.

$$SD = (n_c - n_d)t^{-1}$$

$$GKG = (n_c - n_d)(n_c + n_d)^{-1}$$

$$KT = (n_c - n_d)[0.5N(N-1)]^{-1}$$

To gauge the predictive accuracy of a model using the aforementioned indicators, Lee et al. (2019) argue that the greater these indicators, the better the forecasting accuracy which can be regarded as a measure of correlation intensity. A value for the Goodman-Kruskal Gamma must lie between -1 and 1, and a value close to or at one, indicates that all pairs are concordant and thus there is complete correlation between both the forecasted values and the actual values. Somer's D and Kendall's Tau-a are also measures of correlation intensity ranging from 0 to 1 with a value close to one indicative of complete correlation and concordance among the observed pairs.

A survey logistic regression model of the fixed effects was then constructed and all possible two way interactions were explored. Thereafter, selected two-way interaction, and consequently, selected higher order interactions which pertained to the domain and were found to result in a significant decrease in the deviance, were investigated, all of which were found to be significant under survey logistic regression. Table 3.1 summarises the fixed effects and all two-way and higher order interactions that inform the model that was fitted.

A Taylor series approximation was used for the variance estimation of the survey logistic regression model. The Taylor series approximation is the default under the **PROC SURVEYLOGISTIC** procedure. The concordance index (c) of the final model was 0.872 which indicates that 87.2% of positive HIV cases were correctly predicted, and consequently the predictive accuracy of the model is found to be within an acceptable range.

After fitting the survey logistic regression model, it was found that HIV stigma and the use of contraception were insignificant in predicting a participant's HIV status while variables of a socio-economic, socio-demographic and behavioral nature were found to be significant. In terms of the two-way interaction effects, the association between a participant's highest educational qualification attained and their knowledge of HIV prevention measures together with a participant's knowledge of HIV prevention measures and their acquisition of information and knowledge of HIV prevention, were found to be significant.

Furthermore, a higher order three-way interaction between participants highest educational attainment, their knowledge of HIV prevention and their acquisition of HIV information, was found to be significant. All variables, either significant or insignificant, were deemed as such at a 5% level of significance. The parameter estimates together with the adjusted odds ratio (aOR) and their respective 95% confidence intervals and p-values are given in Table 3.2.

From a socio-economic perspective, there were increased odds of HIV among participants residing in households wherein increased levels of household destitution were observed. This could arguably be ascribed to participants, resident within these households, being unable to access basic provisions such as primary healthcare that would enable them to mitigate the spread of HIV. Participants residing in households that were classified as extremely deprived (aOR=1.381, 95% CI:1.007;

Table 3.1: Type III analysis of the fixed effects for the SLR Model

Effect	F-Value	P-Value
Household Deprivation	2.25	0.0273
Gender	60.97	<.0001
Age Group	37.56	<.0001
Highest Level of Education	22.96	<.0001
Marital Status	8.06	<.0001
Knowledge of Prevention	1.70	<.0001
Perceived Risk of HIV	144.24	<.0001
Engaged in Sexual Intercourse	12.83	0.0003
HIV Stigma	0.45	0.7179
Used Contraception	1.83	0.1763
HIV Information Acquisition	215.73	<.0001
Diagnosed with STI	5.82	0.0159
Highest Level of Education*Knowledge of Prevention	10.14	<.0001
Knowledge of Prevention*HIV Information Acquisition	5.83	0.0001
Highest Level of Education*Knowledge of Prevention*HIV Information Acquisition	7.29	<.0001

1.894) (10% of most deprived households) were increasingly likely to be at risk of HIV than participants residing in households experiencing significant deprivation. The odds of HIV infection were almost twice as high for females (aOR=1.964, 95% CI:1.658; 2.326) than for males.

Advancement in age was observed to be indicative of an increased odds of HIV infection. Participants in young adulthood (30-34 years) were found to be more than twice as likely to be HIV positive than participants aged 45-49 years. A similar observation was made for participants in the 40-44 year age group. There was no significant difference in the odds of HIV infection were observed for 25-29 year old participants when compared to 45-49 year old participants.

With respect to the participants' marital status, the odds of HIV was observed to be low among those who were legally married (aOR=0.251, 95% CI:0.106; 0.595) compared with those who were widowed. Legally married participants and participants who were separated but still legally married were found to be significantly different from widowers as evidenced by their confidence intervals ranging from 0.106 to 0.595 and from 0.037 to 0.764 respectively. There also appeared to be an increasing likelihood of HIV infection among divorcees (aOR=0.824, 95% CI: 0.209 ; 3.239) when compared to widowers.

Some of the possible contributory factors to HIV that were investigated in addition to the socio-economic and socio-demographic determinants were factors encompassing intellectual and social dynamics. Among these were participants' perceived risk of HIV and the level of HIV stigmatisation prevalent among participants. In this respect, individuals who were questioned about their perceived risk of HIV were observed to be risk averse compared to individuals who were already HIV positive irrespective of their perceptions surrounding their chances of HIV infection.

Table 3.2: Odds ratio estimates (OR) and corresponding 95% confidence intervals (CI) for the variables not included in interactions for the SLR Model

Parameter	Odds Ratio (95% CI)
Household Deprivation Level (ref = Significant Deprivation)	
No Deprivation	0.757 (0.557 ; 1.030)
Low Deprivation	0.793 (0.602 ; 1.043)
Minor Deprivation	0.815 (0.636 ; 1.046)
Intense Deprivation	0.876 (0.669 ; 1.147)
Serious Deprivation	0.935 (0.694 ; 1.259)
Severe Deprivation	1.046 (0.721 ; 1.516)
Extreme Deprivation	1.381 (1.007 ; 1.894)*
Gender (ref = Male)	
Female	1.964 (1.658 ; 2.326)*
Age Group (ref = 45-49)	
15-19	0.298 (0.198 ; 0.451)*
20-24	0.643 (0.455 ; 0.911)*
25-29	1.157 (0.818 ; 1.635)
30-34	2.029 (1.432 ; 2.875)*
35-39	1.976 (1.390 ; 2.810)*
40-44	2.213 (1.545 ; 3.170)*
Marital status (ref = windowed)	
Legally Married	0.251 (0.106 ; 0.595)*
Separated - Legally Married	0.168 (0.037 ; 0.764)*
Cohabiting	0.556 (0.216 ; 1.430)
Single - Never Married or Cohabited	0.626 (0.270 ; 1.451)
Divorced	0.824 (0.209 ; 3.239)
Single Live-in Partner	0.674 (0.275 ; 1.649)

Continued on next page

Continued from previous page

Variables	Odds Ratio (95% CI)
Perceived Risk of HIV (ref = Already HIV Positive)	
Assured Infection	0.018 (0.011 ; 0.029)*
Probable Infection	0.014 (0.009 ; 0.020)*
Probable Non-Infection	0.011 (0.007 ; 0.016)*
Assured Non-Infection	0.008 (0.005 ; 0.012)*
HIV Stigma (ref = Severe Stigma)	
No Stigma	0.775 (0.362 ; 1.662)
Mild Stigma	0.753 (0.342 ; 1.656)
Moderate Stigma	0.620 (0.256 ; 1.505)
Engaged in Sexual Intercourse (ref = Yes)	
No	0.553 (0.400 ; 0.765)*
Diagnosed with an STI (ref = Yes)	
No	0.667 (0.481 ; 0.927)*
Used Contraception (ref = Yes)	
No	1.130 (0.947 ; 1.348)

* *Significant at a 5% level of significance*

Studies have shown that HIV stigma and discrimination is contributory to HIV vulnerability. Individuals facing increased odds of HIV often encounter such stigmatisation and discrimination based on their actual or perceived health status, socio-economic standing and other socio-demographic characteristics. Such individuals are sometimes shunned by family members, peers and the wider community. The odds of HIV infection by respondents who espoused no stigmatisation (aOR=0.775, 95% CI:0.362; 1.662) and mild stigmatisation (aOR=0.753, 95% CI:0.342; 1.656) was relatively low when compared to those who espoused severe levels of HIV stigmatisation. Overall, it was observed that varying levels of HIV stigmatisation prevalent in the study area were not largely contributory to HIV incidence and that no significant difference in the odds of HIV infection were observed with respect to different levels of HIV stigma.

Furthermore, selected behavioral and clinical determinants were observed to be associated with vulnerability to HIV infection. Participants who did not engage in sexual intercourse (aOR=0.553, 95% CI:0.400; 0.765) had lower odds of HIV infection than those who engaged in sexual intercourse. In addition, participants who affirmed that they did not make use of contraceptive methods at sexual debut (aOR=1.130, 95% CI:0.947; 1.348) were observed as having higher odds of HIV infection than those who used contraceptive methods at sexual debut. A non-diagnosis of a sexually transmitted infection (STI) by a certified healthcare worker in individuals appeared non-indicative of increased odds of HIV infection.

Figure 3.1 depicts the higher order interaction between a participant's knowledge of HIV prevention, highest level of education and acquisition of HIV information. The odds of HIV infection appeared relatively high for participants lacking adequate HIV information irrespective of their level of academic qualification and knowledge of HIV prevention. There were large fluctuations in the odds of HIV infection noted among participants who were moderately well informed and who either had no formal schooling, up to a primary level, or did not complete secondary school.

There were, once more, large fluctuations in odds of HIV infection noted among individuals who were well informed with information pertaining to HIV infection. Individuals who were highly knowledgeable about HIV prevention and did not complete a secondary education were observed to have the lowest odds of HIV infection. Participants who were moderately knowledgeable of HIV prevention and completed a secondary education were at a higher odds of HIV infection; slightly higher than participants with the same qualification but highly knowledgeable about HIV prevention measures.

The odds of HIV infection remained relatively low and stable among participants who were in possession of a tertiary qualification and who were well informed irrespective of their level of knowledge in relation to HIV preventative measures.

A similar observation was noted in individuals who did not possess any formal schooling and were well informed but displayed a moderate to high degree of knowledge about HIV prevention.

The interaction effect decisively proves that knowledge of HIV prevention measures is vital to ensure that the odds of HIV among individuals is kept to a minimum. However, a participant's level of academic qualification is not indicative of reduced odds of infection. Furthermore, it is evident that obtaining HIV information plus a participant's knowledge of HIV information, is commensurate with a reduction in the odds of HIV infection. These factors, examined as part of higher order interactions, also show that being well informed about HIV prevention measures is a driver of reduced odds of HIV infection among participants. This is observed irrespective of a persons knowledge of HIV prevention and of their academic and/or scholarly achievements.

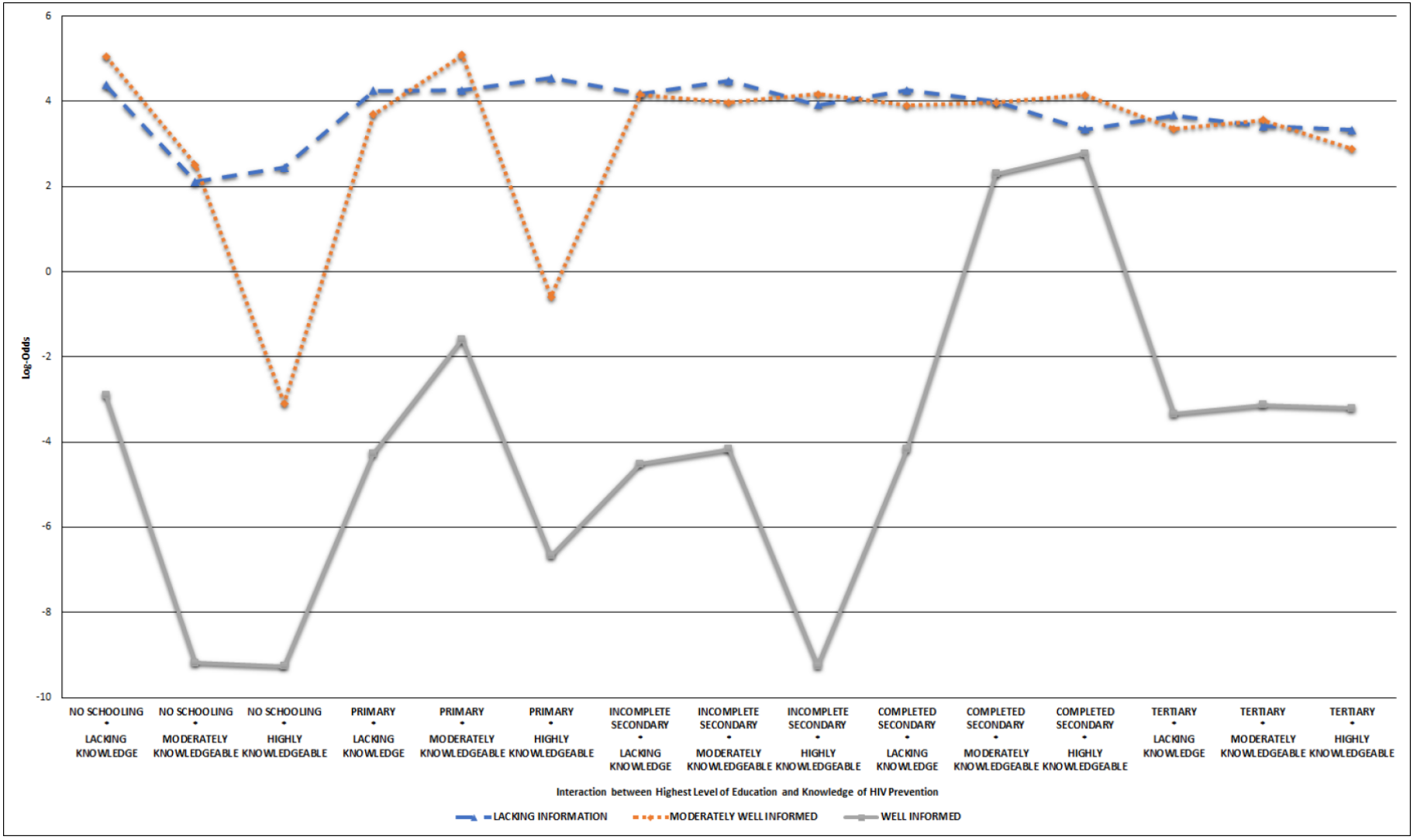

Figure 3.1: The estimated log-odds associated with the interaction of the highest level of education, knowledge of HIV prevention and HIV information acquisition

3.8 Summary and Discussion

The presence of sample weights as a consequence of complex survey design is a contentious issue among researchers. Incorporating survey weights into likelihood based models does not always involve taking the reciprocal of the probability of inclusion to represent the weight. The logistics extend far beyond this and not without its complexities as (Gelman, 2007, p. 153) states unequivocally:

> *"Survey weighting is a mess. It is not always clear how to use weights in estimating anything more complicated than a simple mean or ratio, and standard errors are tricky even with simple weighted means."*

Adjustments for certain characteristics and factors ought to be accounted for by using an amalgamation of probability calculations. Consideration of the sampling weights is paramount if each individual does not have an equal chance of selection into the sample, failing which, the statistical inferences drawn will be redundant and inconclusive. Survey logistic regression, applied in this instance, concurrently accounts for both survey weights and a binary response variable.

The results presented in this chapter, accounting for the survey design, show conclusively that socio-economic circumstances together with socio-demographic and selected behavioural characteristics were associated with HIV infection. Furthermore, the two-way and three-way joint effects were observed to be significant in the model.

If the sampling design is ignored and simple random sampling is assumed to be the sampling design and the standard logistic regression model is applied, then, both the point estimates and their standard errors will be erroneously calculated. While the survey logistic regression model does incorporates the survey design, it does not compensate for the effect of clustering. The effect of clustering refers to the possibility that participants within the same cluster exhibit similar traits and characteristics to participants from different clusters. This renders the potential for correlation in

the outcome among participants a very real possibility resulting in a violation of the independence assumption, of which the inevitable result is statistical bias.

The presence of statistical bias inevitably results in inconsistent estimates and hence the results will be considered superfluous. Oltean & Gagnier (2015) state that if clustering is predetermined to be a realistic consideration, it should be accounted for in the analysis. Accordingly, in the following chapter, the generalised linear mixed model (GLMM) is employed to account for the effect of clustering.

Chapter 4

The Generalised Linear Mixed Model

4.1 Introduction

Partitioning a data set into smaller enumerated groups, wherein constituents of the groupings exhibit a particularly common but unmeasured characteristic, increases the number of fixed effects parameters as the sample size increases for fixed effects. Such models are frequently applied in an array of disciplines such as medicine, public health, ecology and evolutionary biology where responses are sometimes clustered and non-normal.

The generalised linear mixed model (GLMM), first conceptualised by Breslow & Clayton (1993), is used to analyse responses that are correlated as a result of clustered observations. This is achieved by amalgamating the theoretical considerations of the generalised linear model (GLM) and the linear mixed model (LMM) (Broström & Holmberg, 2011; Rich, 2018). What differentiates the GLMM from the GLM is the structure of the GLMM in which provision is made for a *random effect* in the linear predictor. The random effect is able to accommodate correlation in non-normally distributed data making the application of GLMMs indispensable in many practical fields.

The objective of this study was to investigate the determinants of HIV wherein the response variable is binary, that is, whether a participant is HIV positive or negative. For a model containing only fixed effects, a generalised linear model will suffice. However, with randomisation involved in the selection of the primary sampling unit (PSU), these are included in the model as random effects. Furthermore, as there is only one random factor under consideration, the most simplest form of the GLMM, the random intercept model, is employed to assess the effect of clustering in modelling HIV status.

4.2 The Generalised Linear Mixed Model

Suppose a sample n is drawn from n independent clusters denoted by $\mathbf{y}$ such that $\mathbf{y} = (\mathbf{y}_1^T, \mathbf{y}_2^T, ..., \mathbf{y}_n^T)^T$ where $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{it_i})$ for $i \in [1, n]$. Now, within clusters, it is often the case that responses are correlated. Define $\mathbf{b}_i$ as the cluster random effect, whose inclusion in the model may be ascribed to shared characteristics among subjects; it is the condition on which within cluster responses, $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{it_i})$, are conditioned on such that

$$y_{i1}, y_{i2}, ..., y_{it_i}|\mathbf{b}_i \quad \underset{\sim}{\text{ind}} \quad f(y_{ij}|\mathbf{b}_i) \tag{4.1}$$

where $\mathbf{b}_i = (\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_n)'$ and $E(y_{ij}|\mathbf{b}_i) = \mu_{ij}^c$ is the mean of the conditional density function $f(y_{ij}|\mathbf{b}_i)$ which follows a generalised linear model (GLM) that incorporates a random factor. This is referred to as the generalised linear mixed model (GLMM) and is expressed mathematically as

$$g(\mu_{ij}^c) = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{b}_i \tag{4.2}$$

And more compactly in vector form as

$$g(\boldsymbol{\mu}_i^c) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i \tag{4.3}$$

where $\boldsymbol{\mu}_i^c = E(Y_i|\mathbf{b}_i)$ and $\mathbf{X}_i$ is a $t_i \times p$ design matrix of the fixed effects constructed from $\mathbf{x}'_{ij}$ and $\mathbf{Z}_i$ is $t_i \times q$ design matrix of the random effects constructed in a similar manner to the computation of $\mathbf{X}_i$. These respective matrices are detailed below.

$$\mathbf{X}_i = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1t_i} \\ 1 & X_{21} & X_{22} & \cdots & X_{2t_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{i1} & X_{i2} & \cdots & X_{it_i} \end{pmatrix}_{n\times(t_i+1)} , \mathbf{Z}_i = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1t_i} \\ z_{21} & z_{22} & \cdots & z_{2t_2} \\ \vdots & \vdots & \ddots & \vdots \\ z_{i1} & z_{i2} & \cdots & z_{it_i} \end{pmatrix}_{t_i\times q}$$

Furthermore, we define $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{t_i+1})'_{(t_i+1)\times 1}$ as the column vector of regression coefficients.

The $q \times 1$ vector of random effects $\mathbf{b}_i$, over n clusters is assumed to have a mean of $\mathbf{0}$ and a variance-covariance matrix $\mathbf{G}(\boldsymbol{\Psi})$ such that

$$\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_n \underset{\sim}{\text{iid}} f(\mathbf{b}_i, \boldsymbol{\Psi}) \tag{4.4}$$

The common choices of the distribution of the density function $f(\mathbf{b}_i, \boldsymbol{\Psi})$ are:

- The normal distribution
- A conjugate distribution for the conditional distribution of y_{ij}, or
- f may be left unspecified or several non-parametric approaches are possible.

There are two approaches in computing the regression coefficients of a GLMM: a Bayesian approach and a maximum likelihood based approach. The general preference for the maximum likelihood approach over the Bayesian approach is the relative ease with which maximum likelihood produces unbiased estimates (Browne & Draper, 2006). Furthermore, owing to the optimal properties of the maximum likelihood-based approach, it is usually the preferred method of fitting a multilevel model (Searle et al., 2006). Thus, the maximum likelihood-based approach will be examined.

4.3 Maximum Likelihood Estimation

The i^{th} cluster contributes to the likelihood function through the marginal density of $\mathbf{y}_i$ which is given by $f(\mathbf{y}_i; \boldsymbol{\beta}, \boldsymbol{\Psi})$. This marginal density is however, not specified directly in the model and must by computed from the conditional density of $\mathbf{y}_i$ given $\mathbf{b}_i$ and the marginal density of $\mathbf{b}_i$. This is done by first computing the joint density of $\mathbf{y}_i$ and $\mathbf{b}_i$ and consequently deriving the marginal density of $\mathbf{y}_i$.

The joint density function of $\mathbf{y}_i$ and $\mathbf{b}_i$ is given by:

$$f(\mathbf{y}_i, \mathbf{b}_i, \boldsymbol{\beta}, \boldsymbol{\Psi}) = f(\mathbf{b}_i; \boldsymbol{\Psi}) f(\mathbf{y}_i | \mathbf{b}_i; \boldsymbol{\beta}) \tag{4.5}$$

$$= f(\mathbf{b}_i; \boldsymbol{\Psi}) \prod_{j=1}^{t_i} f(y_{ij} | \mathbf{b}_i; \boldsymbol{\beta}) \tag{4.6}$$

Thus, the marginal density of $\mathbf{y}_i$ is given by

$$f(\mathbf{y}_i, \boldsymbol{\beta}, \boldsymbol{\Psi}) = \int f(\mathbf{y}_i, \mathbf{b}_i; \boldsymbol{\beta}, \boldsymbol{\Psi}) \tag{4.7}$$

$$= \int f(\mathbf{b}_i; \boldsymbol{\Psi}) \prod_{j=1}^{t_i} f(y_{ij} | \mathbf{b}_i; \boldsymbol{\beta}) d\mathbf{b}_i \tag{4.8}$$

As a result of the independence of the clusters, the likelihood function is thus the product of the joint and marginal density functions and is expressed as

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\Psi}, \mathbf{y}) = \prod_{i=1}^{n} \int f(\mathbf{b}_i; \boldsymbol{\Psi}) \prod_{j=1}^{t_i} f(y_{ij} | \mathbf{b}_i; \boldsymbol{\beta}) d\mathbf{b}_i \tag{4.9}$$

The likelihood function of the linear mixed model is considered simplistic as a solution exists in closed form and can be evaluated using numerical integration. This in large part, may be ascribed to the linearity of the model. However, Jiang (2007) contends that the evaluation of the likelihood function is computationally more demanding in the GLMM as the integral in $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\psi})$ has to be numerically evaluated but no solution exists in closed form. Thus, the application of iterative techniques is employed to produce an approximated solution to the likelihood function.

4.3.1 Evaluation of the Log-Likelihood Function of the GLMM

Evaluating the log-likelihood function may be quantified in three approaches:

- Numerical integration methods wherein the integral is approximated using quadrature techniques.
- Analytic approximation of the integrand using Laplace approximation.
- Monte Carlo integration which uses simulation-based techniques in the evaluation of the log-likelihood function.

Numerical integration techniques are focused on quadrature methods that account for sample weights and are usually an *a priori* consideration in survey methodology. Other common methods employed are the marginal quasi-likelihood (MQL), penalised quasi-likelihood (PQL) functions, and Laplacian approximations. That these methods are not investigated as part of this study is in large part owed to the findings of Rodríguez & Goldman (1995), Pinheiro & Bates (1995) and Tierney & Kadane (1986) who found significant bias in the estimation of the model parameters employing these respective methods, particularly the MQL and PQL which produced these estimates in the presence of large variance components.

4.3.2 Numerical Integration Techniques

The Gauss-Hermite Quadrature (GHQ)

Quadrature techniques approximate an integral using a finite weighted sum of the integrand evaluated at a set of values of the variable to be integrated out. There are two common types of quadrature approaches, namely the Gauss-Hermite quadrature and the adaptive Gaussian quadrature. The Gauss-Hermite quadrature evaluates integrals of the form

$$\int_{-\infty}^{+\infty} \exp^{-x^2} f(x)dx \tag{4.10}$$

where $f(\cdot)$ is a smooth function which is approximated by a polynomial. The Gauss-Hermite quadrature uses the principle that an integral can be thought of as an infinite weighted sum of the function being integrated (the integrand). Hence, Equation 4.10 can be approximated as

$$\int_{-\infty}^{+\infty} \exp^{-x^2} f(x)dx \approxeq \sum_{j=1}^{R} w_j f(x_j) \tag{4.11}$$

where x_j is the quadrature points (called: *abscissas*), and the solution of the L^{th} order to the Hermitian polynomial $f(\cdot)$ and w_j is the quadrature weights. The number of quadrature points, R, is selected such that the accuracy of the approximation is linked with an increasing R. Furthermore, if $f(\cdot)$ is a $2(R-1)$ degree polynomial, then the R point GHQ is an exact approximation. The application of the Gaussian-Hermite quadrature is practical for a model with random effects as a result of the weight function, $\exp(x^2)$ being proportional to the normal density.

As stated in Section 4.1 there is only a single random effect in the model under consideration. Hence, for illustrative purposes, the simplistic GLMM with only the random intercept will be considered to explicate the process of numerical integration and approximation.

The rudimentary form of the GLMM with a single random effect is

$$g(\mu_{ij}) = \mathbf{x}'_{ij} + b_i, \qquad \{b_i\} \underset{\sim}{iid} N(0, \Psi) \tag{4.12}$$

Hence, the likelihood contribution from the i^{th} cluster will be given by

$$\int_{-\infty}^{+\infty} f(b_i|\Psi) \prod_{j=1}^{n_i} f(y_{ij}|b_i)db_i \tag{4.13}$$

In order to evaluate Equation 4.13, the change of variable technique to a standard normal distribution is conducted. The substitution $u_i = \dfrac{b_i}{\sqrt{\Psi}}$ is made such that

Equation 4.13 becomes

$$\int_{-\infty}^{+\infty} \phi(u_i) \prod_{j=1}^{n_i} f(y_{ij}|\sqrt{\Psi}u_i)du_i \tag{4.14}$$

where

$$\phi(u_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u_i^2}{2}\right) \tag{4.15}$$

Now, applying the GHQ as per Equation 4.11, yields the following relationship

$$\int_{-\infty}^{+\infty} \phi(u_i) \prod_{j=1}^{n_i} f(y_{ij}|\sqrt{\Psi}u_i)du_i \approxeq \sum_{j=1}^{R} w_j \prod_{j=1}^{n_i} f(y_{ij}|\sqrt{\Psi}a_j) \tag{4.16}$$

where

$$w_j \equiv \frac{w_j^*}{\sqrt{\pi}}, \qquad a_j \equiv \sqrt{2}a_j^* \tag{4.17}$$

where the quadrature points x_j are the roots of the L^{th} order Hermitian polynomial with sampling weights w_j.

The GHQ may be efficient for a small R, that is, for an integrand well approximated by a polynomial. However, it is often the case that for GLMM, $R < 10$ can be inaccurate and an $R > 20$ is necessary (McCullogh & Searle, 2008) . Notwithstanding these recommendations for R, the pitfalls of the GHQ are compounded when clusters are too large and there is increased variability in the random effects. These problems can be mitigated by employing the adaptive Gaussian quadrature (AGQ) which tailors (rescales) the quadrature points and weights to the function integrated (Hartzel et al., 2001). Furthermore, the AGQ can be applied with arbitrary degrees of accuracy that can lead to nearly unbiased estimates. However, the increased level of complexity in its computation is sometimes the reason for its avoidance (Pinheiro & Bates, 1995) and (Pinheiro & Chao, 2006).

The Adaptive Gaussian Quadrature (AGQ)

The following relationship holds true for the function in Equation 4.13

$$\phi(u_i)\prod_{j=1}^{n_i} f(y_{ij}|\sqrt{\Psi}u_i) \quad \propto \quad f(b_i|y_{ij}) \tag{4.18}$$

This density can be approximated by a normal density $\phi(u_i;\mu_i,\tau_i^2) \sim N(\mu_i,\tau_i^2)$. The prior density $\phi(u_i)$, instead of being viewed as the weight function, may be rewritten as

$$\int_{-\infty}^{+\infty} \phi(u_i;\mu_i,\tau_i^2) \left\{ \frac{\phi(u_i)\prod_{j=1}^{n_i} f(y_{ij}|\sqrt{\Psi}u_i)}{\phi(u_i;\mu_i,\tau_i^2)} \right\} du_i \tag{4.19}$$

which approximates the posterior density as the weight function for the quadrature.

Changing the variable of integration from u_i to $z_i = \dfrac{(u_i-\mu_i)}{\tau_i}$ yields the following result

$$f(\mathbf{y}_i) = \int_{-\infty}^{+\infty} \frac{\phi(z_i)}{\tau_i} \left\{ \frac{\phi(\tau_i z_i+\mu_i)\prod_{j=1}^{n_i} f(y_{ij}|\sqrt{\Psi}(\tau_i z_i+\mu_i))}{\dfrac{\exp\left(\dfrac{-z_i^2}{2}\right)}{\sqrt{2\pi\tau_i^2}}} \right\} \tag{4.20}$$

Applying the GHQ as stipulated in Equation 4.11

$$f(\mathbf{y}_i) = \int_{-\infty}^{+\infty} \frac{\phi(z_i)}{\tau_i} \left\{ \frac{\phi(\tau_i z_i+\mu_i)\prod_{j=1}^{n_i} f(y_{ij}|\sqrt{\Psi}(\tau_i z_i+\mu_i))}{\dfrac{\exp\left(\dfrac{-z_i^2}{2}\right)}{\sqrt{2\pi\tau_i^2}}} \right\} \tag{4.21}$$

$$\approx \sum_{j=1}^{R} w_j \left\{ \frac{\phi(\tau_i a_j + \mu_i) \prod_{j=1}^{n_i} f(y_{ij}|\sqrt{\Psi}(\tau_i a_j + \mu_i))}{\frac{\exp(-a_j^2/2)}{\sqrt{2\pi\tau_i^2}}} \right\} = \sum_{j=1}^{R} \pi_{ij} \prod_{j=1}^{n_i} f(y_{ij}|\sqrt{\Psi}\alpha_{ij}) \tag{4.22}$$

where

$$\alpha_{ij} \equiv \tau_i a_j + \mu_i \tag{4.23}$$

are the shifted and re-scaled quadrature points with weights, and

$$\pi_{ij} \equiv \sqrt{2\pi}\tau_i \exp\left(\frac{w_j}{2}\right) \phi(\tau_i a_j + \mu_i) w_j \tag{4.24}$$

In the GHQ, the quadrature points x_j and the weights w_j are fixed and independent of $\phi(u_i).f(u_i)$, the function describing the contribution to the likelihood of each observation, whereas in the AGQ the quadrature points and weights are *adapted* to support $\phi(u_i).f(u_i)$. Thus, the term *adaptive* refers to the scaling of the function being integrated using the Hessian function at optimal points as is similarly done in Laplacian approximation. A typical drawback of the AGQ is that the application is time consuming and computationally complex since calculating the quadrature points and weights are dependent on the fixed effects and variance components. In addition, the parameters need to be updated at each iteration and the computation is further complicated when an increasing number of random effects are included in the model (Handayani et al., 2017).

4.4 Model Selection Criteria

An important feature of model evaluation is selecting an appropriate model to accurately capture the relationship between the outcome and the explanatory effects. This process is referred to as model selection and makes use of the concept of an

information criterion. This is a technique designed to minimise the amount of information required to express data as a concise regression model which is an accurate representation of the data (Wasserman, 2000) and (Acquah, 2010). In practice, two penalised criteria are usually employed in selecting an appropriate model.

The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) have been used extensively in this endeavour. The two methods differ in that the AIC is concerned with finding the most simplistic model while the BIC is designed to identify the true model. Bozdogan (1987) and Acquah (2010) note that as the AIC does not depend directly on sample size, the property of asymptotic consistency is absent in AIC but present in BIC, resulting in the BIC sometimes being favoured to produce more simple models.

Extending on the basic principles of the AIC and BIC has been the focal point of research over many years resulting numerous revisions to existing methods and the development of new information criterion for model selection. However, the simplistic approach and application of the AIC and BIC methods has seen widespread use in applied statistical investigation. The formulation of the AIC and BIC is detailed below as delineated by Hilbe (2009).

4.4.1 Akaike Information Criterion - AIC

The AIC is one of the most prominent information criterion used in model selection first proposed by Hirotugu Akaike (1974). The AIC technique selects a model by minimising the negative log-likelihood function conditioned on the number of parameters in the regression model. The original formulation of the AIC is given as follows

$$\text{AIC} = -\frac{2}{n}\left(\ln \mathcal{L}(\boldsymbol{\beta}) - k\right) \tag{4.25}$$

where $\ln \mathcal{L}(\boldsymbol{\beta})$ represents the log-likelihood function and k is the number of predictor

variables included in the model including the model intercept term. Doubling the number of predictor variables, that is, the $2k$ term is referred to as the penalty term that adjusts for the size and complexity of the model. Thus, if more parameters are introduced into the model, any bias will be mitigated. The consideration of the sample size, n allows for a per observation contribution. Thus, larger samples of n produces a smaller AIC which is favoured as an indication of a better model fit.

4.4.2 Bayes Information Criterion - BIC

The Bayes information criterion (BIC) proposed by Raftery (1995) is another method of model selection that features prominently alongside the AIC. Similar to the AIC, models that minimize the BIC are considered favourable. Unlike the AIC however, the BIC is conceptualised within a Bayesian framework and hence is designed to find the most probable model.

The Bayes information criterion was originally defined by Raftery (1995) as

$$\text{BIC} = D - df \ln(n) \tag{4.26}$$

where D is the model deviance statistic, n represents the number of model observations and df, the model degrees of freedom. Whilst various simulation studies have been conducted to examine the superiority of the BIC or the AIC, these are considered superfluous as the theoretical aspects between both methods differ to a degree. As such, for purposes of application, it is accepted that the stated aim of the AIC and BIC is to aid in the selection of a concise and accurate model (Acquah, 2010).

4.5 Generalized Linear Mixed Models Applied to the HIPSS Baseline Data

The analysis presented herein was conducted using **SAS** Version 9.4. The procedure, **PROC GLIMMIX** allows a generalised linear mixed model (GLMM) to be fitted to

the HIPSS data. Under conventional circumstances in which the **PROC GLIMMIX** statement is employed, the **METHOD** statement, which specifies the method of approximation of the likelihood function, would be the Laplace method of approximation. However, the inclusion of survey weights in the data necessitates the Gauss-Hermite quadrature (GHQ) method be used in the approximation in order to produce a weighted maximum likelihood function to determine the regression coefficients. The study setting consisted of 221 enumeration areas from which fifty households were randomly selected from each PSU. Further to this, a logit link function was used in conjunction with a binary distribution. Established methods of model selection such as the Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC) was used as the GHQ is likelihood based. The **RANDOM** statement specifies the random effect (PSU) that is to be included in the model. In order to account for heterogeneity between clusters, the inclusion of a cluster varying intercept term in the model resulted in a random intercept model.

Additionally, within the **RANDOM** and **MODEL** statements, consideration is given to the survey weights, particularly individual weights. Within the realm of GLMM, the primary sampling unit (PSU) corresponds to the **SUBJECT** statement while the **WEIGHT** statement refers to the survey weight variable.

Furthermore, the need for a random intercept was evaluated using the **COVTEST** procedure which produces likelihood ratio tests for the covariance parameters. Table 4.1 below summarises the results of this test when a weighted GLMM is fitted to the HIPSS data. At a 5% level of significance the null hypothesis of the covariance parameters equaling zero was rejected indicating that the inclusion of the random cluster effect was highly significant in the model.

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

Label	DF	-2Log Likelihood	χ^2	P-Value
No G - side effects	2	186,956	275.30	<0.0001

An *a priori* consideration to model selection of the fixed effects is fitting a covariance structure for **G**. Four such covariance structures were fitted; the initial one being the Variance Components (VC) covariance structure which is the default in **SAS** Version 9.4. The other covariance structures fitted were the Unstructured (UN), Compound Symmetry (CS) and the AR(1) covariance structures. Table 4.2 displays the different covariance structures fitted plus their corresponding AIC values. The Variance Components (VC) and the Unstructured (UN) covariance structures produced the lowest AIC value. Owing to the less complex nature of fitting the structure, the Variance Components covariance structure was selected.

Table 4.2: AIC Goodness of Fit for the GLMM

Covariance Structure	AIC
Variance Components (VC)	**186,822.7***
Compound Symmetry (CS)	186,825.3
Autoregressive (AR)	186,824.7
Unstructured (UN)	**186,822.7**

* *Selected covariance structure*

The Pearson Chi-Square statistic over its degrees of freedom was 0.67 which is relatively close to one. This is an indicator that the variability in the data was properly modelled thus mitigating the effects of any residual over-dispersion. Furthermore, the estimate for the variance component for the cluster effect was 0.8687 with a standard error of 0.4502. This estimate exceeds zero and further justifies the inclusion of a random effect in the model.

A GLMM of the fixed effects was produced and all two-way interactions were explored. Thereafter, selected two way and consequently, selected higher order three-way interactions were explored and included in the model producing the final GLMM which is summarised in Table 4.3. The denominator degrees of freedom was calcu-

lated to be 9,754. The results emanating as a consequence of the inclusion of a random effect concurred with the results produced by the SLR model. Consistent with the SLR model, all socio-economic and socio-demographic variables were significant while variables such as HIV stigmatisation and the use of contraception, variables that may be described as behavioural predictors, were not significant in predicting HIV infection.

The two-way interaction between highest level of education and knowledge about HIV prevention and the highest level of education and the participants' acquisition of HIV clinical and preventative information, were significant. Furthermore, the higher order three-way interaction between participants' highest level of education, knowledge about HIV prevention and their acquisition of HIV information, was significant. All variables deemed to be either significant or not significant were done at a 5% level of significance. Table 4.4 summarises the adjusted odds ratio and their corresponding 95% confidence interval.

Examining the socio-economic dynamics of the study setting under the GLMM, household deprivation appeared consequential of an increased odds of HIV infection among those resident in such houses. The result largely concurred with the SLR model wherein it was observed that participants residing in households with extreme levels of socio-economic deprivation displayed a higher odds of HIV infection. Under the GLMM as in the SLR model, the odds of HIV infection among residents of the bottom 10% of extremely deprived households (aOR=1.399, 95% CI:0.9796; 2.006) was twice that of residents in significantly deprived households. Participants residing in seriously deprived households (aOR=1.078, 95%CI:0.703; 1.652) were also observed to be vulnerable to HIV infection when compared to households that are considered significantly deprived. At an individual level, and in line with the SLR model, female participants (aOR=2.216, 95% CI:1.786; 2.750) were observed to be at a disproportionately higher likelihood of HIV infection than male participants.

Table 4.3: Type III analysis of the fixed effects of the GLMM

Effect	F-Value	P-Value
Household Deprivation	2.35	0.0211
Gender	52.18	<.0001
Age Group	23.92	<.0001
Highest Level of Education	10.60	<.0001
Marital Status	7.72	<.0001
Knowledge of Prevention	15.72	<.0001
Perceived Risk of HIV	70.05	<.0001
Engaged in Sexual Intercourse	11.73	0.0006
HIV Stigma	0.56	0.6385
Used Contraception	1.90	0.1682
HIV Information Acquisition	58.28	<.0001
Diagnosed with STI	6.20	0.0159
Highest Level of Education*Knowledge of Prevention	7.67	<.0001
Knowledge of Prevention*HIV Information Acquisition	12.49	0.0001
Highest Level of Education*Knowledge of Prevention*HIV Information Acquisition	6.85	<.0001

Table 4.4: Odds ratio estimates (OR) and corresponding 95% confidence intervals (CI) for the variables not included in interactions for the GLMM

Parameter	Odds Ratio (95% CI)
Household Deprivation Level (ref = Significant Deprivation)	
No Deprivation	0.729 (0.515 ; 1.032)
Low Deprivation	0.720 (0.529 ; 0.980)*
Minor Deprivation	0.793 (0.597 ; 1.053)
Intense Deprivation	0.826 (0.609 ; 1.121)
Serious Deprivation	1.078 (0.703 ; 1.652)
Severe Deprivation	0.899 (0.641 ; 1.263)
Extreme Deprivation	1.399 (0.976 ; 2.006)
Gender (ref = Male)	
Female	2.216 (1.786 ; 2.750)*
Age Group (ref = 45-49)	
15-19	0.268 (0.167 ; 0.427)*
20-24	0.609 (0.409 ; 0.906)*
25-29	1.161 (0.786 ; 1.715)
30-34	2.271 (1.522 ; 3.388)*
35-39	2.169 (1.453 ; 3.237)*
40-44	2.651 (1.740 ; 4.040)*
Marital status (ref = Widowed)	
Legally Married	0.191 (0.071 ; 0.571)*
Separated - Legally Married	0.143 (0.023 ; 0.896)*
Cohabiting	0.508 (0.176 ; 1.470)
Single - Never Married or Cohabited	0.559 (0.216 ; 1.446)
Divorced	0.634 (0.131 ; 3.059)
Single - Live in Partner	0.621 (0.226 ; 1.705)

Continued on next page

Continued from previous page

Variables	Odds Ratio (95% CI)
Perceived Risk of HIV (ref = Already HIV Positive)	
Assured Infection	0.010 (0.005 ; 0.020)*
Probable Infection	0.007 (0.004 ; 0.014)*
Probable Non-Infection	0.006 (0.003 ; 0.010)*
Assured Non-Infection	0.004 (0.002 ; 0.007)*
HIV Stigma (ref = Severe Stigma)	
No Stigma	0.764 (0.314 ; 1.859)
Mild Stigma	0.738 (0.295 ; 1.847)
Moderate Stigma	0.570 (0.207 ; 1.569)
Engaged in Sexual Intercourse (ref = Yes)	
No	0.534 (0.373 ; 0.765)*
Diagnosed with an STI (ref = Yes)	
No	0.625 (0.432 ; 0.905)*
Used Contraception (ref = Yes)	
No	1.152 (0.942 ; 1.408)

**Significant at a 5% level of significance*

Advanced age also appeared to be indicative of an increased odds of HIV infection which became increasingly apparent for individuals approaching their midlife, agreeing with the results of the SLR model. Adolescent participants reported the lowest odds of HIV infection (aOR=0.268, 95% CI:0.167; 0.427) in comparison with their study counterparts aged between 45-49. This, while participants in their early to mid forties were increasingly likely to be HIV positive (aOR=2.651, 95% CI:1.740; 4.040) when compared to those who were aged between 45 and 49 years. The likelihood of HIV infection in participants within the 30-34 years (aOR=2.271, 95% CI:1.522; 3.388) and 35-39 years (aOR=2.169, 95% CI:1.453; 3.237) age bracket were found to be markedly higher compared to participants aged between 45-49 years old. In respect of marital status, no significant difference in the odds of HIV infection were

observed among participants who were cohabiting, single (either living with a partner or have never having married or cohabited) or divorcees. This was in accordance with the findings under the SLR model.

On consideration of the behavioural and clinical factors, participants who did not make use of contraception at their sexual debut were 1.152 (95% CI: 0.942; 1.408) times more likely to be HIV positive than those who did. Furthermore, significant differences in the odds of HIV infection was observed among participants who practiced abstinence and those who did not. Additionally, individuals who were not diagnosed with sexually transmitted infections were viewed to be less susceptible to HIV infection than those who attested to a prior STI diagnosis.

Challenging societal discrimination that is sometimes inherent where traditionalist attitudes prevail cannot be understated. In these societies HIV related stigma can be found with negative psychological impacts on the people it is directed at. In measuring the level of HIV stigmatisation prevalent in the study setting, the results revealed that HIV stigmatisation was not associated with rising odds of HIV infection as there was no significant differences noted across the increasingly intensifying levels of stigmatisation espoused by the participants.

Figure 4.1 depicts the higher order three-way interaction between a participant's highest level of education, knowledge about HIV prevention and acquisition of HIV information. In Figure 4.1 it can be observed that the participants who appear to be lacking in adequate clinical and preventative information, irrespective of their knowledge of HIV prevention and level of educational attainment, were vulnerable to HIV infection.

Furthermore, there appeared to be mutually high odds of HIV infection between these participants and moderately well informed participants with academic qual-

ifications ranging from incomplete secondary school to having completed a secondary education. In addition, these participants possessed moderate to high levels of knowledge about HIV prevention. Respective participants who were in receipt of moderate to advanced levels of HIV preventative information and having graduated from a tertiary institution, were observed as having lower odds of HIV infection. This observation was noted among those who were moderately informed and well informed about HIV infection having sought information from a variety of sources.

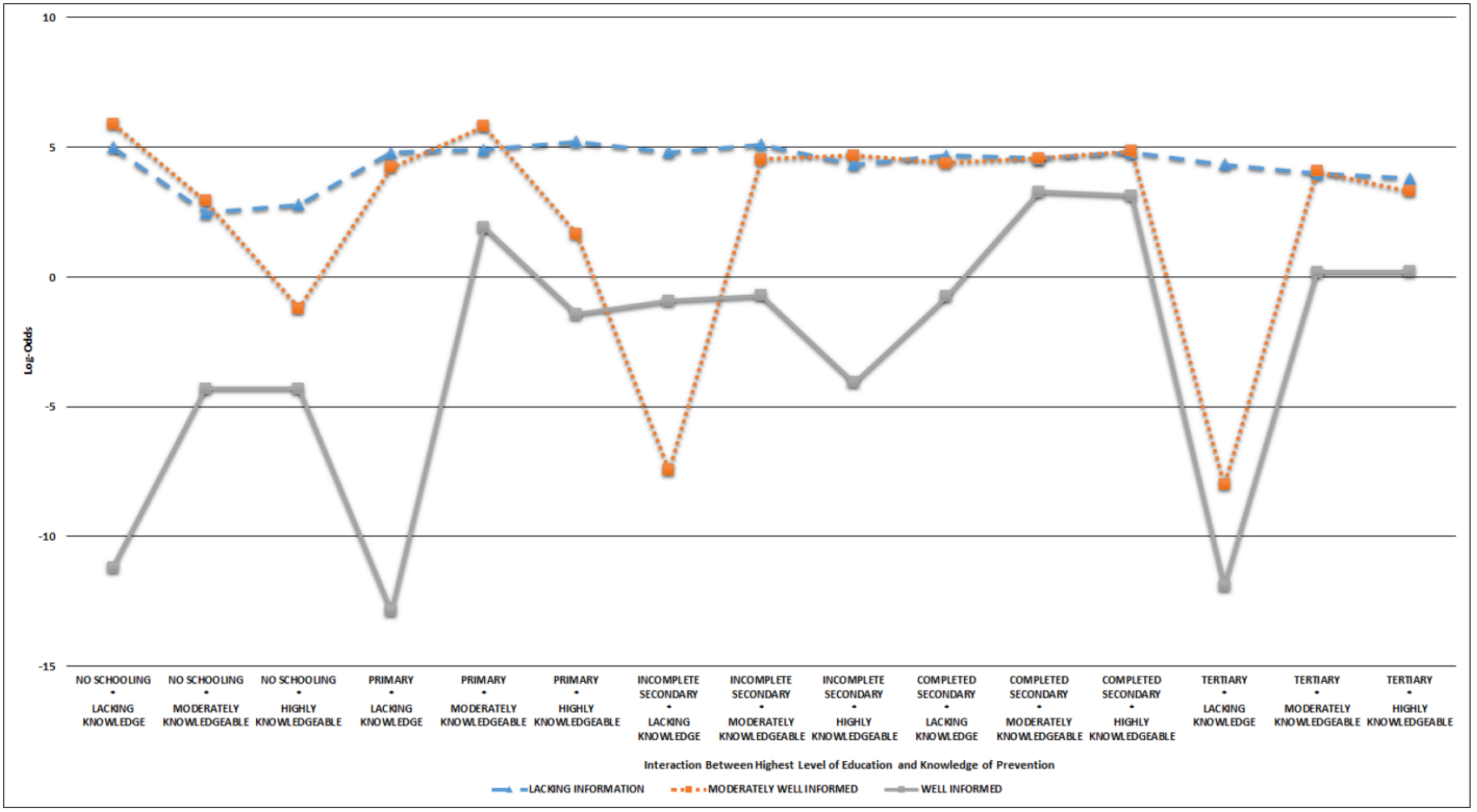

Figure 4.1: The estimated log-odds associated with the interaction of the highest level of education, knowledge of HIV prevention and HIV information acquisition

4.6 Summary and Discussion

A data set partitioned into smaller subgroups due to a common unmeasured characteristic is commonplace in many of research fields. These subgroups are referred to as clusters and to account for within cluster correlation, a generalised linear mixed model is employed in which the clusters serve as random component. The use of the Gaussian quadrature, a likelihood estimation technique well suited for weighted observations, is employed for the purpose of parametric estimation.

The results obtained as a consequence of the inclusion of a random effect largely concur with the results produced by the survey logistic regression model. Moreover, the socio-economic background of participants in tandem with certain socio-demographic and behavioural determinants of HIV infection were observed to be significant. These determinants encompassed the scale of household deprivation, gender, age group, highest academic attainment, marital status, knowledge about HIV prevention, perceived risk of HIV, whether or not the participant engaged in sexual intercourse and their acquisition of information pertaining to HIV.

Furthermore, building on the fixed effects, relevant and significant two-way, and subsequently, three-way interactions were identified and investigated. The two-way interactions were associations between highest level of education and knowledge of prevention together with participants' knowledge about HIV prevention and their acquisition of HIV information. The three-way interaction investigated the association between the highest level of education, knowledge about HIV prevention, and acquisition of HIV information. These joint effects were significant under the GLMM as was the case under the SLR model.

It is usually the case that spatial autocorrelation is present between selected enumeration areas and, subsequently, selected households. This necessitates the use of geographically weighted regression techniques that are able to compensate for spa-

tial variability. Failure to account for the correlation structure renders the estimates arising from the analysis inconsistent due to underestimated standard errors. The next chapter investigates the effect of spatial variability between primary sampling units (PSUs).

Chapter 5

The Spatial Generalised Linear Mixed Model

5.1 Introduction

When data can be partitioned into smaller subsets of a population, the constituent subjects will exhibit similar characteristics and other traits. Data with accompanying geospatial information will show the results of spatial dependencies which intensify with observations in close proximity. Access to spatial information of data is therefore considered vital to allow one to take into consideration both spatial interactions and spatial externalities in the analysis. Analysing the spatial structure allows one to address, if necessary, any violation of hypotheses and to confirm the assumptions of spatial independence.

Autocorrelation refers to the measure of *correlation* (*relatedness*) of a variable with itself when observations are considered in terms of a time lag or in space, a *spatio-temporal shift*. *Spatial autocorrelation* is the correlation (whether positive or negative) of a variable with itself as a consequence of the spatial location of the variable. Spatial autocorrelation is often the result of undetectable and sometimes complex processes that are unable to be quantified, thus giving rise to spatial structuring.

From a statistical viewpoint, analyses are conducted on the hypotheses of independence among the variables. However, a variable that is spatially autocorrelated violates this hypothesis and challenges the validity of the results. It should be noted that spatial autocorrelation and spatial structure do not exist independently of one another and that analysis of spatial autocorrelation enables analysis of spatial structures.

To codify the process of accounting for spatial variability, one first detects for spatial autocorrelation among the regression residuals. The rationale behind investigating residual autocorrelation allows the researcher to mitigate any structural defects by accounting for autocorrelation so that the resultant model is free of same and thus will not violate the assumption for normality. Thereafter, once spatial autocorrelation is established within the residuals, the presence of autocorrelation can be accounted for in the GLMM (Chen, 2016). The resulting model, once survey weights are accounted for, is termed a spatially weighted generalised linear mixed model. The results emanating from this are compared and contrasted with the SLR model and the GLMM.

5.1.1 The Weight Matrix

Prior to assessing spatial autocorrelation, one first needs to formally define what constitutes two observations being in close proximity. In this instance, one must conceptualise a measure of *contiguity* or *spatial adjacency* between two observations presented in the form of a $n \times n$ weight matrix, w_{ij}. Each w_{ij} reflects the *spatial influence* of observation *i* on *j*. The weight matrix is based on the centroid distance d_{ij} between the pairs of spatial units *i* and *j*.

The weight matrix, w_{ij}, may be mathematically described as per the following forms according to (Leung et al., 2000).

K-Nearest Neighbour Weight Matrix

Let each centroid distance, d_{ij}, from spatial observation *i* to spatial observation *j*, $\forall$ $j \neq i$, be ranked as $d_{ij(1)} \leq d_{ij(2)} \leq .... \leq d_{ij(n-1)}$. Then for each $k = 1, ..., (n-1)$, the set $N_k(i) = \{j(1), j(2), ..., j(k)\}$ contains the *k* closest units to *i*. For a given *k*, the *k-nearest neighbour* weight matrix, w_{ij}, expressed mathematically, has the form

$$w_{ij} = \begin{cases} 1 & \text{if} \quad j \in N_k(i), \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

The Gaussian Weight Function

The *Gaussian* function is one of the commonly used weighting functions. The principle behind the Gaussian weighting function is that observations close to the centre are more important and points further away from the centre are considered insignificant. If *i* is a spatial location, an observation, *j*, close to *i* will exert more influence on *i* than any observation further away from *i* than *j*. Therefore, any analysis should place more emphasis on *j* owing to its proximity to *i* than any other observation within the geospatial location.

The Gaussian weight function is given below

$$w_j(i) = \exp(-\theta d_{ij}^2) \quad j = 1, 2, ..., n \tag{5.2}$$

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

The Distance Weight Function

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

$$w_i(j) = \begin{cases} \left(1 - \frac{d_{ij}^2}{d^2}\right)^2 & \text{if} \quad d_{ij} \leq d, \\ 0 & \text{if} \quad d_{ij} > d \end{cases} \tag{5.3}$$

5.2 Measures of Spatial Autocorrelation Among Residuals

As stated in the introduction, the focus will be on testing for spatial autocorrelation among the residuals, $\mathbf{r} = (r_1, r_2, ..., r_n)'$. The null and alternative hypothesis for testing for spatial autocorrelation among the residuals is, as per (Leung et al., 2000):

H_0 : There exists no presence spatial autocorrelation among the residuals

H_1 : There exists the presence of spatial autocorrelation among the residuals

Alternatively, the null hypothesis, H_0 may be expressed as

$$\mathrm{H}_0 : \mathrm{Var}(r) = E(rr') = \sigma^2 \boldsymbol{I} \tag{5.4}$$

The alternative hypothesis is that there is spatial autocorrelation among the residuals (either positive or negative) and among the residuals with respect to a weight matrix, $\boldsymbol{W}$, which is governed by the underlying spatial structure which in itself is determined by the degree of spatial contiguity or spatial adjacency (Leung et al., 2000). The appropriate rejection region for the hypotheses test for Moran's I and Geary's C are addressed in Subsection 5.2.1 and Subsection 5.2.2 respectively.

In seeking to test for the presence of spatial autocorrelation, two methods are frequently employed. These are Moran's I and Geary's C proposed by Moran (1950) and Geary (1954) respectively. While Moran's I is a more global measurement and though demonstrably proved by Cliff & Ord (1975) and Cliff & Ord (1981) in simulation studies to be more consistent and powerful than Geary's C, the latter is employed in this study owing to its sensitivity to differences in smaller neighbourhoods. Largely, however, one should expect the same conclusion irrespective of the methodology used. Moran's I and Geary's C are presented below as individual methods employed to detect the presence of spatial autocorrelation among the residuals as delineated by Leung et al. (2000).

5.2.1 Moran's Index (I)

For a vector of residuals r, estimated by $\hat{r}$, where $\hat{r} = (\hat{r}_1, \hat{r}_2, ..., \hat{r}_n)'$ and $\boldsymbol{W}$, a prespecified spatial weight matrix, $\boldsymbol{W} = (w_{ij})$, Moran's I is expressed mathematically as

$$I = \frac{n}{s}\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\hat{r}_i\hat{r}_j}{\sum_{i=1}^{n}\hat{r}^2}\right) = \frac{n}{s}\left(\frac{\hat{\boldsymbol{r}}'\boldsymbol{W}\hat{\boldsymbol{r}}}{\widehat{\boldsymbol{r}'\boldsymbol{r}}}\right) \tag{5.5}$$

where

$$s = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} \tag{5.6}$$

Now, $\boldsymbol{W}$ is frequently used in a row standardised form in which the rows of $\boldsymbol{W}$ sum to one, thereby rendering $\boldsymbol{W}$ as asymmetric. In this instance, $\boldsymbol{W}^*$ is defined as a new symmetric spatially weighted matrix, where

$$\boldsymbol{W}^* = (w_{ij}^*) = \frac{1}{2}(\boldsymbol{W} + \boldsymbol{W}') \tag{5.7}$$

As such it can be assumed that without the loss of generality, $\boldsymbol{W}$ is also symmetric and that

$$\hat{\boldsymbol{r}}'\boldsymbol{W}'\hat{\boldsymbol{r}} = \hat{\boldsymbol{r}}'\boldsymbol{W}\hat{\boldsymbol{r}} \tag{5.8}$$

The term $\frac{n}{s}$ is a scaling factor and thus may be excluded without any ensuing repercussion on the *p*-value for $\boldsymbol{W}$, the symmetric spatial weight matrix of the n^{th} order. Hence, from Equation 5.8, the following holds true

$$\frac{\hat{\boldsymbol{r}}\boldsymbol{W}'\hat{\boldsymbol{r}}}{\hat{\boldsymbol{r}}'\hat{\boldsymbol{r}}} = \frac{\hat{\boldsymbol{r}}\boldsymbol{W}\hat{\boldsymbol{r}}}{\hat{\boldsymbol{r}}'\hat{\boldsymbol{r}}} \tag{5.9}$$

Thus, Moran's I, as per Equation 5.5, may be expressed in terms of the residuals as

$$I = \frac{\hat{\boldsymbol{r}}'\boldsymbol{W}\hat{\boldsymbol{r}}}{\hat{\boldsymbol{r}}'\hat{\boldsymbol{r}}} \tag{5.10}$$

For large values of I, it follows that there is evidence of positive autocorrelation among the residuals with the converse being true. If i is defined as the observed value for I, then, with respect to the *p*-values of I, the following holds true for the null and alternative hypotheses respectively, where if;

$$p = \begin{cases} \mathrm{P}(I \geq i) & \mathrm{H}_0 \text{ is not rejected,} \\ \mathrm{P}(I \leq i) & \mathrm{H}_0 \text{ is rejected} \end{cases} \tag{5.11}$$

Hence, following the existing framework as per Section 5.2 it can be concluded that at an α level of significance and with the presence of spatial autocorrelation (H_0 not rejected) the following assumption is justified

$$r = (r_1, r_2, ..., r_n)' \quad \sim \quad \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \tag{5.12}$$

5.2.2 Geary's Coefficient (C)

For the residuals, $\hat{r} = (\hat{r}_1, \hat{r}_2, ..., \hat{r}_n)'$ as described in Subsection 5.2.1 and a pre-specified spatial weight matrix $\boldsymbol{W} = (w_{ij})$, Geary's C is expressed mathematically as

$$C = \frac{n-1}{2s}\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(\hat{r}_i - \hat{r}_j)^2}{\sum_{i=1}^{n}\hat{r}_i^2}\right) \tag{5.13}$$

Through a rudimentary mathematical manipulation, the numerator in Equation 5.13 may be expressed in vector notation as

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(\hat{r}_i - \hat{r}_j)^2 = \hat{\boldsymbol{r}}'(\boldsymbol{D} - 2\boldsymbol{W})\hat{\boldsymbol{r}} \tag{5.14}$$

where the sum of the i^{th} row and i^{th} column are respectively given by

$$w_{i.} = \sum_{j=1}^{n} w_{ij} \quad \text{and} \quad w_{.j} = \sum_{i=1}^{n} w_{ji} \tag{5.15}$$

$\boldsymbol{D}$ is the diagonal of the weight matrix expressed as

$$\boldsymbol{D} = \text{diag}(w_{1.} + w_{.1}w_{2.} + w_{.2}w_{3.} + ... + w_{n.} + w_{.n}) \tag{5.16}$$

Thus, substituting Equation 5.14 into Equation 5.13 and excluding the scale factor $\frac{n-1}{2s}$, Geary's C may be expressed as

$$C = \frac{\hat{\boldsymbol{r}}'(\boldsymbol{D} - 2\boldsymbol{W})\hat{\boldsymbol{r}}}{\hat{\boldsymbol{r}}'\hat{\boldsymbol{r}}} = \frac{\hat{\boldsymbol{r}}'\boldsymbol{A}\hat{\boldsymbol{r}}}{\hat{\boldsymbol{r}}'\hat{\boldsymbol{r}}} \tag{5.17}$$

where

$$\boldsymbol{A} = (\boldsymbol{D} - 2\boldsymbol{W}) \tag{5.18}$$

and if $\boldsymbol{W}$ is symmetric, then

$$\boldsymbol{D} = 2\ \text{diag}(w_{1.}, w_{2.}, ..., w_{n.}) = 2\boldsymbol{D}^* \tag{5.19}$$

Thus, if $\boldsymbol{D}^*$ is the diagonal matrix wherein each element along the main diagonal corresponds to the row sum of $\boldsymbol{W}.$, then $\boldsymbol{A} = 2(\boldsymbol{D} - \boldsymbol{W})$ is symmetric. For a $\boldsymbol{W}$, where $\boldsymbol{W}$ is row standardised, then $\boldsymbol{A}^*$ is defined as

$$\boldsymbol{A}^* = \frac{1}{2}(\boldsymbol{A} + \boldsymbol{A}^T) \tag{5.20}$$

$$= \boldsymbol{D} - (\boldsymbol{W} + \boldsymbol{W}^T) \tag{5.21}$$

Then, if $\boldsymbol{A}^*$ is symmetric, it can be assumed, without loss of generality, that $\boldsymbol{A}$ is also symmetric and that

$$\hat{\boldsymbol{r}}'\mathbf{A}^*\hat{\boldsymbol{r}} = \hat{\boldsymbol{r}}'\boldsymbol{A}\hat{\boldsymbol{r}} \tag{5.22}$$

Now, for a pre-specified weight matrix, $\boldsymbol{W}$, a relatively low value for C, implies that the alternative hypothesis holds true and that there is a presence of spatial autocorrelation among the residuals, $\hat{r}$, while the converse is true. Suppose c is used to denote the observed value for C and P, the p-value of C. Thus, the test for the null hypothesis, H_0, can be encapsulated as

$$p = \begin{cases} \mathrm{P}(C \geq c) & \mathrm{H}_0 \text{ is not rejected,} \\ \mathrm{P}(C \leq c) & \mathrm{H}_0 \text{ is rejected} \end{cases} \tag{5.23}$$

Hence, following the existing framework as per Section 5.2 it can be concluded that at an α level of significance and with the presence of spatial autocorrelation (H_0 not rejected) the following assumption is justified

$$r = (r_1, r_2, ..., r_n) \quad \sim \quad \mathrm{N}(\mathbf{0}, \sigma^2\boldsymbol{I}) \tag{5.24}$$

5.3 Point Referenced Modelling

5.3.1 Preliminary Considerations

Suppose there is a spatial process, $\boldsymbol{S}$, with the outcome $Y(\boldsymbol{S})$, a mean $\mu(s)$=$E(Y(s))$ and associated variance σ_s^2 valid $\forall\ \boldsymbol{S} \in D$, the spatial domain. The outcome is said to be *Gaussian* if, for $n \geq 1$, and for spatial locations denoted by $\boldsymbol{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_n\}$, $\boldsymbol{Y} = (Y(\boldsymbol{s}_1), Y(\boldsymbol{s}_2), ..., Y(\boldsymbol{s}_n))$ follows a multivariate normal distribution.

As with time series models, spatial models also exhibit *stationarity* where changes over time and space (*spatio-temporal shifts*) do not result in a change of distribution. Stationarity can be quantified into three unique categories listed below. Consider n spatial locations $\forall\ n \geq 1\ \boldsymbol{h} \in \Re^D$. Then, a spatial process is deemed:

- ***Strictly Stationary***: if the distribution of $(Y(\boldsymbol{s}_1), Y(\boldsymbol{s}_2), ..., Y(\boldsymbol{s}_n))$ and $(Y(\boldsymbol{s}_1 + \boldsymbol{h}), Y(\boldsymbol{s}_2 + \boldsymbol{h}), ..., Y(\boldsymbol{s}_n + \boldsymbol{h}))$ does not change despite a *spatio-temporal* shift.
- ***Weakly Stationary***: if $\mu(\boldsymbol{S}) \equiv \mu$. This implies that the process has a constant mean. Furthermore, a process is weakly stationary if $Cov(Y(\boldsymbol{S}), Y(\boldsymbol{S} + \boldsymbol{h}))$=C($\boldsymbol{h}$) (Cressie, 1993).
- ***Instrinsically Stationary***: if a process is considered intrinsically stationary then $Var(Z(\boldsymbol{S}+\boldsymbol{h}) - Z(\boldsymbol{S}))$ is not dependent on the position of $\boldsymbol{S}$ but rather on the distance between them. In addition, Matheron (1965) expresses the mean of an intrinsically stationary process as $E\left[(r(\boldsymbol{S} + \boldsymbol{h}) - r(\boldsymbol{S}))^2\right]$. In other words, the variance is only determined by the lag distance $\boldsymbol{h}$ and there is constant expectation. Hence, the concept of a *variogram* arises allowing the following relationship to be stated

$$Var(Z(\boldsymbol{S} + \boldsymbol{h}) - Z(\boldsymbol{S})) = 2\gamma(h) \tag{5.25}$$

The premise behind stationarity is that if we regard the outcome from a spatial location as a *regionalised phenomenon*, it is obvious that for each spatial location there

is only one realisation (outcome). Matheron (1965) thus contends that for inferential geostatistics to be plausible, there should be additional assumptions made about the random function governing $Y(\boldsymbol{S})$. As a result, this has led to stationarity assumptions which are *a priori* considerations and summarised as per (Matheron et al., 2019, p. 46) [1]

> *"A stationary random function is, in a way, repeating itself in space, and this repetition gives a new opportunity for statistical inference from a single realisation"*

5.3.2 The Variogram Procedure

Consider an intrinsically stationary spatial process, $\boldsymbol{S}$, in which the spatial locations, $\boldsymbol{S} = (s_1, s_2, ..., s_n)'$, are governed by the restriction $[Z(\boldsymbol{S}), \boldsymbol{S} \in D, D \subset \Re^d]$. The process of fitting a variogram model commences by first fitting an *empirical semi-variogram* (Matheron, 1963). The empirical semi-variogram is a rudimentary, nonparametric estimate of the semi-variogram which is then compared against a series of theoretical semi-variogram models which are further discussed in Section 5.3.4. The empirical semi-variogram is

$$\widehat{\gamma}(\boldsymbol{s}_i - \boldsymbol{s}_j) = \frac{1}{2N(h)} \sum_{(\boldsymbol{s_i}, \boldsymbol{s_j}) \in N(h)} [Y(\boldsymbol{s}_i) - Y(\boldsymbol{s}_j)]^2 = \widehat{\gamma}(h) \tag{5.26}$$

The constituent components of an intrinsically stationary process is referred to as *isotropic* if $\gamma(\boldsymbol{s}_i - \boldsymbol{s}_j)$ is a function of the Euclidean distance, $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ between two observed locations allowing the simpler notation, $\gamma(h)$ to be used to specify the semi-variogram ($2\gamma(h)$ would be the variogram). A valid semi-variogram must be conditionally negative and must satisfy the following condition

$$\sum_{i=1}^{n} \sum_{i=1}^{n} w_{ij} \gamma(\boldsymbol{s}_i - \boldsymbol{s_j}) \leq 0 \tag{5.27}$$

[1] This book was a posthumous publication of the lectures and seminal works of Georges Matheron and published 19 years after Matheron's demise in August 2000.

for each pair of locations $s_1, s_2, ..., s_n$ and all weights $w_1.w_2., ..., w_n. \ni \sum_{i=1}^{n} w_i = 0.$

Now, for a sample of given realisations from $Y(\cdot)$, the empirical variogram is the unbiased estimator of the isotropic variogram and is expressed as

$$2\widehat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{(\boldsymbol{s_i}, \boldsymbol{s_j}) \in N(h)} [Y(\boldsymbol{s}_i) - Y(\boldsymbol{s}_j)]^2 \tag{5.28}$$

where

$$N(h) = Card\{(\boldsymbol{s_i} - \boldsymbol{s_j}) : \|\boldsymbol{s_i} - \boldsymbol{s_j}\| = h : \forall i, j \in [1, 2, ..., n]\} \tag{5.29}$$

$\|N(h)\|$ is the number of distinct pairs in *N(h)*. Furthermore, the semi-variogram may be estimated in different directions in a particular spatial location thus highlighting its *aniostrophic* property.

Geo-statistical research suggests the use of several theoretical models that are equipped to fit the data to a sample variogram. For the purpose of this study, we consider six theoretical models which are further examined in Section 5.3.4 *in concert* with how these models link with the parameters of the theoretical variogram.

5.3.3 Construction of the Theoretical Variogram Model

As stated earlier, an important consideration in variograms is that of *isotropy*. Banerjee et al. (2004) state that if the semi-variogram, $\gamma(\boldsymbol{h})$, depends only on the separation vector via its length $\|\boldsymbol{h}\|$ then this process is called *isotropic*, and if not, it s referred to as an *aniostrophic* process. Hence, an isotropic process, $[\gamma(\boldsymbol{h}) \in \Re^D]$ which is also univariate, can be expressed as $\gamma(\|\boldsymbol{h}\|)$.

A sample variogram involves using the input distance h between two points to produce a variogram estimate $\gamma(h)$ which explains the variation in $\boldsymbol{Y}$ over these two points. The construction of the sample variogram can be quantified into the four steps detailed below (West, 2001):

1. The range of distances between two sampled points is divided into a set of intervals wherein each lag or interval h is sufficiently large and contains enough point pairs for estimation in all intervals. While a benchmark of at least thirty point pairs are deemed sufficient, it is advantageous to have as many pairs as possible for plotting $\gamma(h)$ against h.

2. Compute the distance between every pair of sample points together with the squared difference in $\boldsymbol{Y}$ values.

3. Assign each pair of sampled points to an interval; this results in accumulated variance in each level.

4. Thereafter, once each pair of points has been assigned an interval, compute the average variation per interval. This value will be $\gamma(h)$ and is then plotted against the midpoint distance of each interval h.

The resulting plot consists of as many points as there are intervals with an estimate for each distance. Computing a variogram to model all possible distances rather than relying solely on a plot of the midpoints of the intervals is a crucial step. In this endeavour, theoretical variogram models are employed to fit the variogram.

5.3.4 Calculation of the Theoretical Variogram Model

A significant advantage of isotropic processes is their simplicity of interpretation and that numerous parametric forms of isotropic processes are available for implementation in geostatistical models (Banerjee et al., 2004). Discussed below are the different parametric forms of isotropic processes. Whilst a variety of these parametric forms are available, a select few, namely those that are used in the analysis and will eventually inform the covariance structure are addressed below as per (Banerjee et al., 2004).

If τ^2 is defined to be the *nugget* which is often viewed as a *non-spatial effect variance*, and σ^2 as the *sill* or *spatial effect variance* then along with ϕ, as the range parameter

(sometimes called the *decay* parameter), the following variogram forms are considered.

Linear: The *linear* isotropic process is a straight line semi-variogram. As $\|\boldsymbol{h}\| \to 0$, $\gamma(h) \to \infty$. This linear semi-variogram is not weakly stationary despite being intrinsically stationary.

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2\|\boldsymbol{h}\|, & \|\boldsymbol{h}\| > 0, \tau^2 > 0, \sigma^2 > 0 \\ 0, & \text{otherwise} \end{cases} \tag{5.30}$$

Spherical: The validity of the *spherical* semi-variogram does not extend beyond the third dimension; as for fourth and beyond, the covariance matrix will not be positive definite. An advantageous feature to fitting a spherical variogram is the clearly defined nugget, sill, and range; the three components that constitute the semi-variogram.

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2, & \|\boldsymbol{h}\| \geq \frac{1}{\phi} \\ \tau^2 + \sigma^2 \left\{ \dfrac{3\phi\|\boldsymbol{h}\|}{2} - \frac{1}{2}(\phi\|\boldsymbol{h}\|)^3 \right\}, & 0 < \|\boldsymbol{h}\| \geq \frac{1}{\phi} \\ 0, & \text{otherwise} \end{cases} \tag{5.31}$$

Exponential: The *exponential* model is considered to be more advantageous than the spherical variogram in that it is of a simpler functional form in all dimensions. The exponential variogram is one that approaches the sill gradually but never converges with the sill.

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi\|\boldsymbol{h}\|), & \|\boldsymbol{h}\| > 0 \\ 0, & \text{otherwise} \end{cases} \tag{5.32}$$

Gaussian: The *Gaussian* variogram is analytic and produces smooth realisations of the spatial process. It is similar to the exponential function and is observed to ascend gradually from the nugget prior to which it rapidly advances toward the sill.

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi^2\|\boldsymbol{h}\|^2)), & \|\boldsymbol{h}\| > 0 \\ 0, & \text{otherwise} \end{cases} \tag{5.33}$$

Powered Exponential: The *powered exponential* variogram produces a family of valid variograms. On inspection, it can be deduced that for p =1,2, the powered exponential variogram is the exponential and Gaussian variograms respectively.

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-|\phi\|\boldsymbol{h}\||^p)), & \|\boldsymbol{h}\| > 0 \\ 0, & \text{otherwise} \end{cases} \tag{5.34}$$

Matérn: The *Matérn* variogram stems from Matérn (1960). The parameter v is called the smoothness parameter, $\Gamma(.)$ is the usual gamma function, and K_v is the modified Bessel function of order v.

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2 \left[1 - \frac{(2\sqrt{v}\phi)^v}{2^{v-1}\Gamma(v)} K_v(2\sqrt{v}\|\boldsymbol{h}\|\phi)\right] & \|\boldsymbol{h}\| > 0 \\ 0, & \text{otherwise} \end{cases} \tag{5.35}$$

5.3.5 Components of the Semi-Variogram

The significance of the variogram in spatial statistics cannot be understated as it plays a significant role in facilitating one's understanding of the observations and underlying spatial autocorrelation structure. It can therefore be unequivocally stated that the variogram is a focal point of geostatistical methods.

However, the empirical variogram and accompanying properties, as detailed in Section 5.3, cannot be directly applied as some of these properties are not satisfied. In order to allow for its implementation in geostatistical models, various adjustments are required using theoretical models that are well equipped for geostatistical analysis. Prior to detailing these adjustments, we proceed by presenting the classical form of the variogram and detailing the constituent components that constitute the variogram. Thereafter, we address the functions that satisfy the aforementioned properties by detailing how these are configured to effect meaning to the concepts detailed below.

The structure of the variogram is such that it initially increases up to a certain level forming a plateau. At this level the value of h corresponding to the plateau is called the *range*. The range is understood to be the relationship between the covariance function and the semi-variogram and represents the point at which total variability is reached.

Furthermore, to understand why the semi-variogram is increasing at this point consider the covariance function of two observations h distance apart. The covariance relates to the semi-variogram as per the following function

$$2\gamma(\boldsymbol{h}) = Var[Y(\boldsymbol{S}+\boldsymbol{h}) - Y(\boldsymbol{S})] \tag{5.36}$$

which, via an elementary mathematical manipulation is

$$2\gamma(\boldsymbol{h}) = 2[C(0) - C(h)] \tag{5.37}$$

thus yielding the following result

$$\gamma(h) = C(\mathbf{0}) - C(\boldsymbol{h}) \tag{5.38}$$

As the covariance is a decreasing function for increasing distance, it can be deduced

from Equation 5.36 that the variogram is an increasing function for a decreasing covariance function.

At a certain distance after the peak of the range, the covariance is cancelled out beyond which there is no possible relationship between the observed values. With respect to the semi-variogram, the values beyond the range are constant implying that C=$C(0)$=σ^2. This point is referred to as the *sill* and is the point at which, for large values of h, corresponds to the variance of the observation(s). Now, for $h = 0$, the value of the variogram is zero, however, for values close to zero, the variogram takes on values larger than zero resulting in a level of discontinuity at or near the origin. This phenomenon, which can be regarded as one of the limits of the variogram, is called the *nugget*. The nugget can be regarded as the intercept of the variogram or a representation of measurement error at separations smaller than the sample distance. Drawing on the explanation of the semi-variogram by Matheron (1965), the nugget effect can be encapsulated as the variation between two measurements made at infinitely close locations from which arises two effects:

- Variability of the measuring instrument: the nugget is partly a measurement of statistical errors that are spatially dependent.
- The real nugget effect: a sudden change in the measured parameter possibly ascribed to sampling error and short-scale variability resulting in unusual variation between two observations.

These components come together to form the variogram which is shown in Figure 5.1.

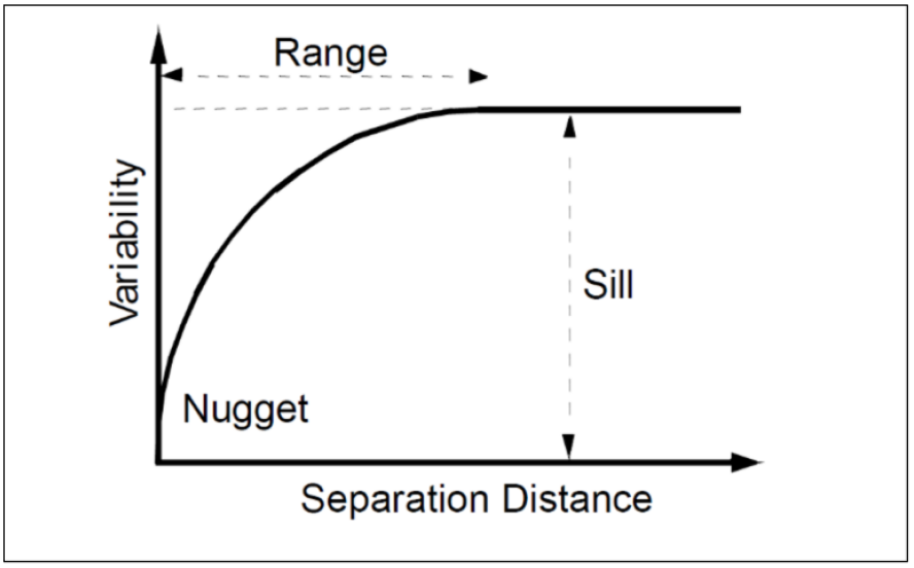

Figure 5.1: Components of the semi-variogram (Arnold, 2013)

5.4 Multilevel Spatial Models

Spatial modelling has become indispensable in contributing to societal understanding of numerous facets of everyday life. Advances in multiple spheres of scientific, societal and economic research can be attributed to the conceptualisation of spatial regression. Spatial data is considered a realisation from a random field, and the focus is largely based on predicting a random variable at a new location or region (Shojaei et al., 2018). In this respect, *Kriging*, a geostatistical technique developed in the late 1950s by Matheron (1963) building on the work of Danie Krige, a South African mining engineer who proposed innovative methods for mining estimation, is a frequent topic of discussion and application (West, 2001). As the primary focus of this section is modelling spatial variation, *Kriging* is not discussed any further.

Whereas random effects have the benefit of accounting for spatial correlations, these are usually unobserved and cannot be predicted without the benefit of prior information (Fortin, 2013) and (Skrondal & Rabe-Hesketh, 2009). In this respect, a generalised linear mixed model with spatially correlated random effects is employed in the analysis of discrete or continuous data. The random effects are usually a zero mean Gaussian random field (GRF) (Shojaei et al., 2018).

The spatial generalised linear mixed model (SGLMM) is presented as per Bonat & Ribeiro (2016) and Fortin (2013), and together with the maximum likelihood estimation technique, the Gauss-Hermite quadrature (GHQ) is adapted to include the spatial element for observations located within a spatial domain. The GHQ is employed because it is equipped for the approximation of regression coefficients in the presence of spatial weights.

5.4.1 The Spatial Generalized Linear Mixed Model

Suppose that $\boldsymbol{y} = (y_1, y_2, ..., y_n)'$ are recorded observations from spatial locations $\boldsymbol{S} = (s_1, s_2, ..., s_n)'$ and that $\boldsymbol{y}$ is the realisation of $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)'$. Then the spatial generalised linear mixed model (SGLMM) is given by

$$\begin{aligned} \boldsymbol{Y}(\mathbf{S})|\boldsymbol{S}(s) &\sim f(\cdot, \mu(\boldsymbol{S}), \psi) \\ f(\mu(\boldsymbol{S})^c) &= x_{ij}^T\boldsymbol{\beta} + \sigma z_{ij}(\boldsymbol{S}, \gamma) + \epsilon R \\ &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i(\boldsymbol{S}) + \epsilon R \\ \boldsymbol{S}(s) &\sim N(\mathbf{0}, \Psi) \end{aligned} \tag{5.39}$$

The SGLMM is formulated on the assumption that $Y_1, Y_2, ..., Y_n$ are conditionally independent for a Gaussian spatial process $\boldsymbol{S}(s)$ which is distributed as $f(\cdot, \mu(\boldsymbol{S}), \psi)$. The SGLMM comprises two sets of parameters, $\mu(\boldsymbol{S})^c$ and ψ; the conditional mean and the dispersion or precision parameter respectively. The conditional mean is related to the linear predictor through the link function $g(\cdot)$. The dispersion parameter is included in the probability density function but treated independently as an additional parameter when evaluating the likelihood function.

Furthermore, the spatial process comprises a spatially dependent process $\boldsymbol{Z}(\boldsymbol{S})$ and spatially independent process (R), each with unit variance and scaled parameter σ and ϵ respectively. The linear predictor consists of the fixed effects, $\boldsymbol{X}_i\boldsymbol{\beta}$, the spatially

correlated random effects $Z(\boldsymbol{S})$ and the spatially uncorrelated random component $\epsilon R = \delta \sim N(0, \epsilon I)$. The design matrix $\boldsymbol{X}$ of the potential covariates, and $\boldsymbol{\beta}$, the column vector of regression parameters, are given by

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_{12} & \cdots & X_{1j} \\ 1 & X_{22} & \cdots & X_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{i2} & \cdots & X_{ij} \end{pmatrix}_{n\times(t+1)} , \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{t+1} \end{pmatrix}_{(t+1)\times 1}$$

In the realm of geostatistics, $\boldsymbol{Z}(\boldsymbol{S})$, is a unit variance Gaussian random field (GRF) with correlation function $\rho(h, \gamma)$, for $\rho \in D$. The correlation function which is composed of h, is the distance between two spatial locations given by the Euclidean distance $\|s_i - s_j\|$ and γ, the parameter measuring spatial correlation.

5.4.2 Maximum Likelihood Estimation

The i^{th} cluster contributes to the likelihood function through the marginal density of $\mathbf{y}$ which is given by $f(y_i, \boldsymbol{\beta}, \boldsymbol{\Psi})$. As stated in Section 4.3 and reiterated in this section, the marginal density cannot be specified from the model itself, but, must be computed from the conditional density of $\boldsymbol{y}_i$ given the spatial domain $\boldsymbol{S}$ and the marginal density of $\boldsymbol{S}$. This process commences by computing the marginal density of $\boldsymbol{y}_i$ and $\boldsymbol{S}$ and thereafter deriving the marginal density of $\boldsymbol{y}_i$.

The joint density function of $\boldsymbol{y}_i$ and $\boldsymbol{S}$ is

$$f(\boldsymbol{y}_i, \boldsymbol{S}, \boldsymbol{\beta}, \boldsymbol{\Psi}) = f(\boldsymbol{S}; \boldsymbol{\Psi}) f(\boldsymbol{y}_i | \boldsymbol{S}; \boldsymbol{\beta}) \tag{5.40}$$

$$= f(\mathbf{S}; \boldsymbol{\Psi}) \prod_{j=1}^{t_i} f(y_{ij}, |\mathbf{S}; \boldsymbol{\beta}) \tag{5.41}$$

Thus, the marginal density of $\mathbf{y}_i$ is

$$f(\mathbf{y}_i, \boldsymbol{\beta}, \boldsymbol{\Psi}) = \int f(\boldsymbol{y}_i, \boldsymbol{S}, \boldsymbol{\beta}, \boldsymbol{\Psi}) \tag{5.42}$$

$$= \int f(\boldsymbol{S}; \boldsymbol{\Psi}) \prod_{j=1}^{t_i} f(y_{ij}|\boldsymbol{S}; \boldsymbol{\beta}) d\boldsymbol{S} \tag{5.43}$$

Now, owing to the independence of the clusters, the likelihood function is the product of the joint and marginal density functions and can be expressed as

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{y}) = \prod_{i=1}^{n} \int f(\boldsymbol{S}; \boldsymbol{\Psi}) \prod_{j=1}^{t_i} f(y_{ij}|\boldsymbol{S}; \boldsymbol{\beta}) d\boldsymbol{S} \tag{5.44}$$

Despite the mathematical rigour and computationally demanding nature of the Gauss-Hermite quadrature, it is presented as a means of evaluating the log-likelihood function as per Fortin (2013).

Evaluating the Log Likelihood Function

In Equation 5.39, let $(\boldsymbol{S}, \gamma) = u_i$, then Equation 5.39 can be rewritten as

$$E(\mathrm{Y}(\boldsymbol{S})|x_{ij}) = \int f(x_{ij}\boldsymbol{\beta} + z_{ij}u_i)\mathrm{PDF}(u_i, \boldsymbol{\Psi}_u) du_i \tag{5.45}$$

where $\mathrm{PDF}(u_i, \boldsymbol{\Psi}_u^2)$ is the probability density function of the normal distribution with a mean of zero and a variance $\boldsymbol{\Psi}$. The Gaussian quadrature method assumes the function to be integrated over the variable v_i may be expressed as

$$f(v_i) = w(v_i)h(v_i) \tag{5.46}$$

where $w(v_i)$ is the weighting function and $h(v_i)$ is the polynomial to be approximated. In this respect, $f(v_i)$ can be approximated as

$$\int f(v_i) dv_i \approx \sum_{k=1}^{r} w_j' h(v_{i,j}) \tag{5.47}$$

where r is the number of *quadrature points* and w'_j and $h(v_{i,j})$ represent the weight function and associated value of the weight function associated with the j^{th} observation respectively. As the number of quadrature points, p, is increased, the more precise the approximation is in Equation 5.47. However, increasing the number of quadrature points exponentially is computationally more demanding. The weights, w'_j, are dependent on the quadrature points as well as the weighting function $w(v_i)$.

The Gauss-Hermite quadrature is a special case of the Gaussian quadrature and the weight function $w(v_i) = e^{-v_i^2}$ which is directly applied to Equation 5.45 for a Gaussian PDF. Applying the change of variable technique, we let $u_i = \sqrt{2\boldsymbol{\Psi}_u^2}v_i$, hence $\frac{du_i}{dv_i} = \sqrt{2\boldsymbol{\Psi}_u^2}$ and consequently, $du_i = \sqrt{2\boldsymbol{\Psi}_u^2}dv_{i.}$. The Gauss-Hermite quadrature for the spatial generalised linear mixed model is derived thus

$$
\begin{aligned}
f(Y(\boldsymbol{S}|x_{ij})) &= \int f\left(x_{ij}\boldsymbol{\beta} + z_{ij}u_i\right) \text{PDF}(u_i, \boldsymbol{\Psi}_u)du_i \\
&= \int f\left(x_{ij}\boldsymbol{\beta} + z_{ij}u_i\right) \frac{\exp\left(\dfrac{-u_i^2}{2\boldsymbol{\Psi}_u}\right)}{\sqrt{2\pi\boldsymbol{\Psi}_u}}du_i \\
&= \int f\left(x_{ij}\boldsymbol{\beta} + z_{ij}\sqrt{2\boldsymbol{\Psi}_u}v_i\right) \frac{\exp(-v_i)^2}{\sqrt{\pi}}dv_i \\
&= \int \exp(-v_i^2)\frac{g\left(x_{ij}\boldsymbol{\beta} + z_{ij}\sqrt{2\boldsymbol{\Psi}_u}v_i\right)}{\sqrt{\pi}}dv_i
\end{aligned}
$$

Applying the Gauss-Hermite approximation as per Equation 5.47

$$
\approx \frac{1}{\sqrt{\pi}}\sum_{j=1}^{r} w'_j f(x_{ij}\boldsymbol{\beta} + z_{ij})\sqrt{2\boldsymbol{\Psi}_u}v_{i,j} \tag{5.48}
$$

At point j, the weight, w'_j, is calculated as

$$
w'_j = \frac{2^{r-1}r!\sqrt{\pi}}{r^2[H_{r-1}(v_{i,j})]^2} \tag{5.49}
$$

where H_{r-1} is the Hermitian polynomial of order $r-1$.

5.5 Accounting for Spatial Variability

5.5.1 Examining Residual Autocorrelation

The Variogram Procedure

Prior to constructing a model that accounts for spatial autocorrelation, it must first be established whether there is autocorrelation present in the data by detecting the presence of spatial autocorrelation in the residuals. This process is conducted by constructing and examining the structure of the empirical semi-variogram which measures the distance between dwelling units by using GIS coordinates.

A prerequisite requirement to constructing an empirical semi-variogram is to group the spatial locations into intervals in accordance with the common distance between them. In **SAS** Version 9.4, the procedure **PROC VARIOGRAM**, which produces an empirical semi-variogram, determines the distance between each spatial location using the uniqueness of the GIS coordinates for a particular location. The procedure also requires that the user specifies both the size of the lag class and maximum number of lags using the **LAGDISTANCE** and **MAXLAG** commands respectively.

Table 5.1: Pairwise information for 50 classes

Number of Lags	51
Lag Distance	0.0071
Maximum Data Distance in Latitude	0.18
Maximum Data Distance in Longitude	0.31
Maximum Data Distance	0.36

Specifying the appropriate number of intervals usually happens at the discretion of the researcher and there is no interval that is specific to any given data set. The arbitrary nature of the number of interval classes has thus been subject to examination by Banerjee et al. (2004) and Journel & Huijbregts (1978), recommending that inter-

vals be wide enough to capture at least thirty pairwise observations per interval. For the purpose of the construction of the variogram it was decided that spatial locations would be grouped across fifty classes. Based on this, the following pairwise information, presented in Table 5.1 was collated. With the information presented in Table 5.1, a common lag distance of 0.0071 was specified accompanied by a maximum number of thirty-eight intervals. The figures presented were obtained as follows:

$$\begin{aligned}\text{Common Lag-Distance} &= \text{[Upper Bound - Lag (n+1)]} - \text{[Upper Bound - Lag (n)]} \\ &= 0.01068 - 0.00356 \\ &= 0.0071\end{aligned}$$

There was a sufficient number of pairwise observations per class up to and including the 38^{th} lag, which had a class interval of (0.26697 ; 0.27409). Hence, the following calculation justifies the maximum number of lags:

$$\text{Maximum Number of Classes} = \frac{0.27409}{0.0071} = 38$$

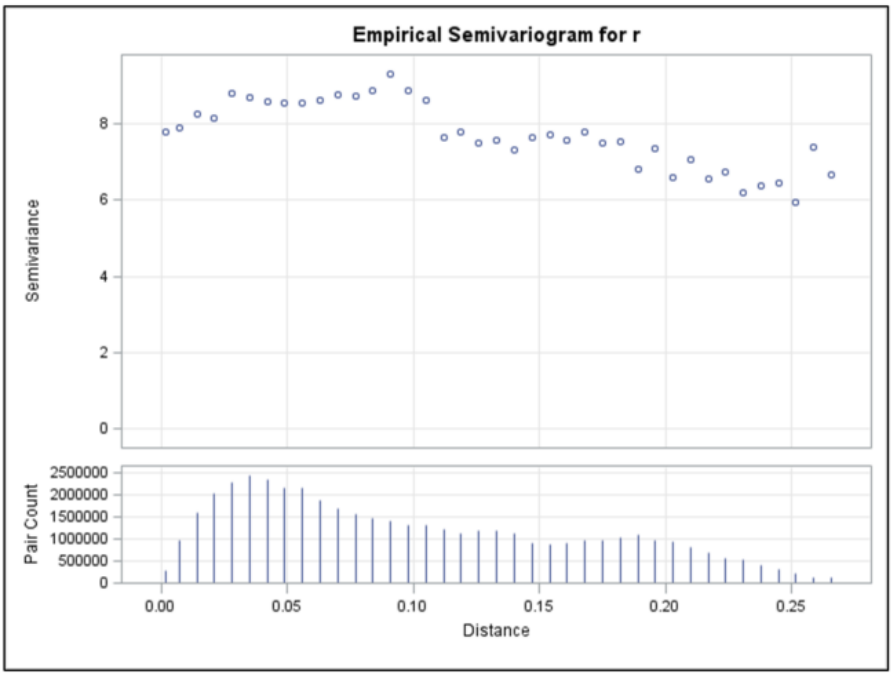

Figure 5.2: Empirical semi-variogram of the HIPSS data

Using the information contained in Table 5.1 together with the calculations above, the empirical semi-variogram is constructed and presented in Figure 5.2. On inspection of the semi-variogram, one is able to discern that with increasing distance between households in the study location, the semi-variogram decreases. This suggests that there is progressively less variation among spatial units (households) the further apart properties become. The implication herein is that spatial units in close geographic proximity exhibit spatial autocorrelation.

Detecting Spatial Autocorrelation Using Geary's C

As discussed in Section 5.2, Geary's coefficient, often referred to as Geary's *C*, is employed in the detection of spatial autocorrelation, though less frequently than Moran's *I* owing to the latter's affinity to detect spatial autocorrelation on a global scale. Notwithstanding this, an advantage of Geary's *C* is the localised nature of the methodology and its affinity for detecting differences in smaller spatial locations.

Table 5.2: Geary's *C* for the presence of spatial autocorrelation

Assumption	Coefficient	Observed	Std. Dev.	Z	Pr.$\|Z\|$
Normality	Geary's *C*	1.0044101	0.000384	11.49	$<$.0001

As the spatial location of the study forms part of a larger conurbation, Geary *C* is well suited in its ability to detect spatial autocorrelation and is thus employed for this purpose. As is evident in Table 5.2, using Geary's *C*, the null hypothesis as expressed earlier in Section 5.2 is refuted at a 5% level of significance. It may thus be concluded that there is evidence of spatial autocorrelation in the data concurring with the findings of the semi-variogram.

At this juncture, once the presence of residual autocorrelation has been established, there follows the modelling procedure that will now account for the presence of spatial correlation. The process now proceeds by fitting an appropriate spatial covariance structure to the GLMM. In order to determine the most appropriate spatial covariance structure, one must specify a range of variogram models in the **`VARIOGRAM`**

procedure and proceed to examine their AICs according to the *smaller-is-better* criterion. The results of this process are presented in Table 5.3 along with their corresponding AIC. As evidenced in Table 5.3, the *spherical* variogram is the most appropriate structure and will be the selected and specified under the **`TYPE`** command in the **`PROC GLIMMIX`** procedure.

Table 5.3: Fit statistics of the spatial covariance structures for the semi-variogram

Spatial Model	**Spherical**	Exponential	Power	Matérn	Gaussian
AIC	**338.92417**	338.92417	338.92418	340.92417	338.92417

Figure 5.3 depicts a comparison of the spherical model with Gaussian and exponential models. As one is able to discern from the figure, the spherical model is linear at the origin and the range parameter is exactly the correlation length. This is one of the salient features of the spherical variogram model; the ease of interpretation associated with it. In the following section the spherical variogram is employed in the **`PROC GLIMMIX`** procedure to account for the spatial variability that was statistically shown, via the residuals, to be inherent in the data.

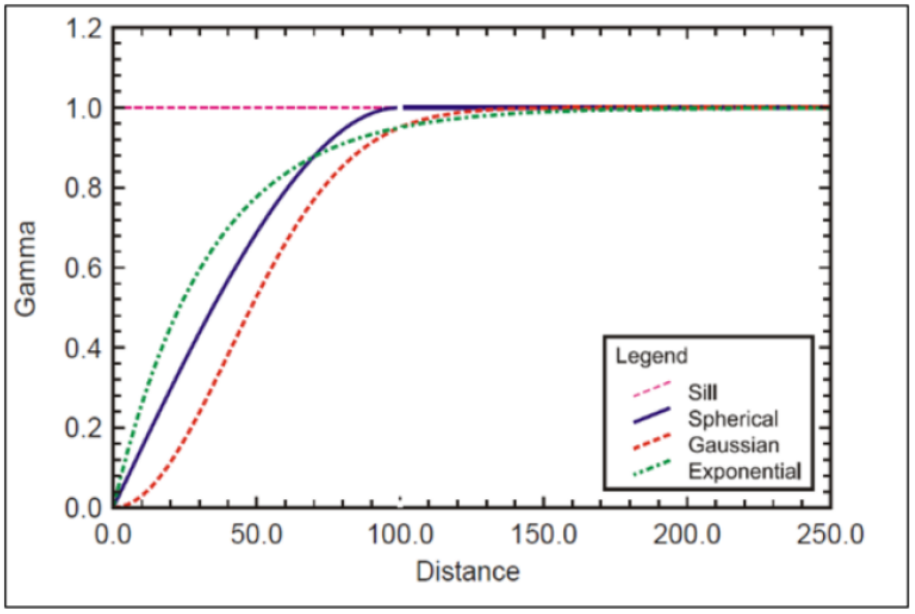

Figure 5.3: Comparison of the spherical variogram model to the Gaussian and exponential model (Arnold, 2013)

5.6 Spatial Generalised Linear Mixed Models Applied to the HIPSS Baseline Data

The analyses presented herein were conducted using **SAS** Version 9.4. The procedure **PROC GLIMMIX** was employed to fit a generalised linear mixed model to the HIPSS data. Accounting for the inclusion of survey weights, the Gaussian Hermite quadrature method was specified in the **METHOD** command *in lieu* of the Laplacian approximation which is conventionally used in the absence of survey weights to iteratively determine the regression coefficients. Furthermore, the logit link function was employed together with a binary distribution. Methods of model selection such as the Akaike Information Criterion (AIC) and Bayes Information Criterion informed the selection of the model owing to the likelihood-based nature of the Gauss-Hermite quadrature.

The **RANDOM** statement specifies a random effect, which is the primary sampling unit (PSU) that will be included in the model. Additionally, to account for spatial heterogeneity, a cluster varying intercept is added thus producing a random intercept model. The **RANDOM** statement accommodates the inclusion of subject specific weights for multilevel models, particularly in the case of our data, the individual weights.

As previously detailed in Section 4.5, the **COVTEST** procedure evaluates the inclusion of a random intercept by producing likelihood ratio tests for the covariance parameters. As per Table 5.4 the null hypothesis is refuted that the covariance parameter equates to zero at a 5% level of significance. This indicates that accounting for clustering is significant in the model.

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

Label	DF	-2Log Likelihood	χ^2	P-Value
No G - side effects	2	180,218	232.00	<0.0001

The ratio of the Pearson Chi-Square statistic to its degrees of freedom was 0.68; an indicator that the variability in the data was properly modelled and that there were no consequences of residual over-dispersion. A precursor to model selection of the fixed effects is fitting a covariance structure for **G**. This is done in the form of fitting an appropriate variogram using the *smaller-is-better* criterion of the AIC method of model selection. As per Table 5.3, the *spherical* variogram model is selected. The estimate for the variance component of the cluster effect was 0.8170 which represents the partial *sill*. The estimated *range*, which appears as **`SP(SPH)`**, is **`14.0000`**. This implies that observations more than fourteen units apart are not spatially correlated.

It should be noted that the *sphercial* covariance structure precisely defines the *range*, but that some structures as detailed in 5.3.4 do not. In such instances, spatial researchers have to rely on a predetermined *effective range*. The *effective range* is defined as the distance at which the semi-variogram attains 95% of the *sill*, or alternatively, the distance at which spatial autocorrelation declines to, or below 0.05 which is considered negligible autocorrelation. However, the effective range is considered a point of contention among spatial experts. (Stroup, 2012, p. 445) disclaims:

> *"Our use of 0.05 as the effective range is for the purpose of discussion. In practice, effective range is a matter of judgment and, even then, spatial experts may differ."*

The final SGLMM which accounted for spatial variation with both the fixed effects and interaction effects is summarised in Table 5.5 while Table 5.6 summarises the adjusted odds ratio (aOR) and their corresponding 95% confidence interval (CI).

An SGLMM of the fixed effects including the two-way and higher order three-way interactions explored as per the SLR model and GLMM, were examined. The denominator degrees of freedom was calculated to be 9,490. The results emanating from the inclusion of a random effect and accounting for spatial variability, largely

concurred with the results of the SLR model and the GLMM at a 5% level of significance. Variables that one would consider socio-economic and socio-demographic were found to be significant in the SGLMM. Predictors that are behavioural in nature, such as the use of contraception, prevalence of HIV stigmatisation and participants knowledge about HIV prevention measures, were not contributory in modelling HIV prevalence. The latter, while not independently significant, was significant when jointly modelled with the participants' highest levels of education and in a three-way interaction between itself, a participant's highest level of education, and their ability to acquire HIV information.

In a marked departure from the SLR model and the GLMM, the joint effect between knowledge about HIV prevention and the participants' acquisition of HIV information, was not deemed significant in modelling HIV status. This despite one of the constituent components, HIV information acquisition, being significant in the model.

On inspection, there is no noticeable shift in the risk of HIV infection in participants that can be directly attributed to their socio-economic circumstances. In agreement with the SLR and GLMM, participants living is households characterised as extremely deprived (aOR=1.393, 95% CI:0.968; 2.004), that is the 10% of most deprived households, were observed to be at an inordinately higher risk of HIV infection than participants residing in households characterised as significantly deprived.

Furthermore, and in line with the results of the SLR and GLMM, female participants (aOR=2.146, 95% CI:1.731; 2.660) were observed to be more than twice as likely to be HIV positive than their male counterparts. While aging was observed to be associated with increased likelihood of HIV infection, the odds of infection among the different age categories remained largely the same across the SLR, GLMM and SGLMM.

Table 5.5: Type III analysis of the fixed effects for the SGLMM

Effect	F-Value	P-Value
Household Deprivation	2.46	0.0159
Gender	48.49	<.0001
Age Group	22.27	<.0001
Highest Level of Education	10.64	<.0001
Marital Status	7.79	<.0001
Knowledge of Prevention	1.16	0.3145
Perceived Risk of HIV	67.12	<.0001
Engaged in Sexual Intercourse	11.79	0.0006
HIV Stigma	0.59	0.6194
Used Contraception	1.65	0.1987
HIV Information Acquisition	68.62	<.0001
Diagnosed with STI	6.68	0.0097
Highest Level of Education*Knowledge of Prevention	5.44	<.0001
Knowledge of Prevention*HIV Information Acquisition	2.15	0.0724
Highest Level of Education*Knowledge of Prevention*HIV Information Acquisition	2.82	<.0001

Table 5.6: Odds ratio estimates (OR) and corresponding 95% confidence intervals (CI) for the variables not included in interactions for the SGLMM

Parameter	Odds Ratio (95% CI)
Household Deprivation Level (ref = Significant Deprivation)	
No Deprivation	0.723 (0.510 ; 1.025)
Low Deprivation	0.695 (0.510 ; 0.964)*
Minor Deprivation	0.804 (0.604 ; 1.071)
Intense Deprivation	0.839 (0.615 ; 1.144)
Serious Deprivation	0.898 (0.638 ; 1.262)
Severe Deprivation	1.108 (0.723 ; 1.698)
Extreme Deprivation	1.393 (0.968 ; 2.004)
Gender (ref = Male)	
Female	2.146 (1.731 ; 2.660)*
Age Group (ref = 45-49)	
15-19	0.259 (0.161 ; 0.416)*
20-24	0.573 (0.383 ; 0.858)*
25-29	1.111 (0.749 ; 1.649)
30-34	2.158 (1.438 ; 3.238)*
35-39	2.107 (1.403 ; 3.162)*
40-44	2.433 (1.591 ; 3.720)*
Marital status (ref = Widowed)	
Legally Married	0.156 (0.059 ; 0.413)*
Separated - Legally Married	0.139 (0.019 ; 0.993)*
Cohabiting	0.402 (0.140 ; 1.155)
Single - Never Married or Cohabited	0.457 (0.179 ; 1.164)
Divorced	0.513 (0.108 ; 2.442)
Single - Live in Partner	0.511 (0.188 ; 1.386)

Continued on next page

Continued from previous page

Variables	Odds Ratio (95% CI)
Perceived Risk of HIV (ref = Already HIV Positive)	
Assured Infection	0.010 (0.005 ; 0.010)*
Probable Infection	0.008 (0.004 ; 0.014)*
Probable Non-Infection	0.006 (0.003 ; 0.011)*
Assured Non-Infection	0.004 (0.002 ; 0.008)*
HIV Stigma (ref = Severe Stigma)	
No Stigma	0.770 (0.318 ; 1.865)
Mild Stigma	0.570 (0.208 ; 1.562)
Moderate Stigma	0.736 (0.295 ; 1.834)
Engaged in Sexual Intercourse (ref = Yes)	
No	0.526 (0.365 ; 0.759)*
Diagnosed with an STI (ref = Yes)	
No	0.616 (0.426 ; 0.819)*
Used Contraception (ref = Yes)	
No	1.142 (0.933 ; 1.398)

* *Significant at a 5% level of significance*

As a participant approached midlife, there was a slight though noticeable decrease in the odds of HIV infection upon comparison between the random effects model which accounted for spatial variability and the model that did not. It was observed that individuals in the 35-39 year and 40-44 year age groups were more than twice as likely to be HIV positive compared to 45-49 year old participants. However, this was not an occurrence peculiar to the model accounting for spatial variability, as it was also the case in both the SLR model and the GLMM. With respect to marital status, no significant difference was observed among participants who were cohabiting, single (either with a live-in partner, or who have never married or cohabited) or divorcees as observed under the GLMM and SLR model.

The participants who were questioned about their perceptions of their risk of HIV infection appeared to be risk averse with no observable association between a person's perceived risk of infection and those who claimed that they were HIV positive. A major area of concern is the prevalence of HIV stigma which is sometimes observed in certain sectors of society. Research has shown that HIV stigmatisation can have adverse consequences on HIV positive individuals who may become withdrawn from society. However, the results show that HIV stigma is not statistically significant and that however severe the levels of HIV stigma espoused by participants, it was not associated with increased (or decreased) odds of HIV infection.

The participants who did not use contraception (aOR=1.142, 95% CI:0.933; 1.398) at their sexual debut were 1.142 times more likely to be HIV positive compared to those who used contraception. In this instance, however, no significant difference was observed. Furthermore, participants who had not engaged in sexual intercourse (aOR=0.526, 95% CI:0.365; 0.759) were less susceptible to HIV infection compared to those who had engaged in sexual intercourse. Additionally, participants who had not been diagnosed with a sexually transmitted infection (aOR=0.616, 95% CI:0.426; 0.819) had significantly lower odds of HIV infection compared to those who had been diagnosed with a sexually transmitted infection.

Figure 5.4 depicts the higher order three-way interaction between a participant's highest educational qualification, their knowledge about HIV prevention and their acquisition of HIV clinical and preventative information. An observation of generally decreased odds of HIV infection was observed among participants who had acquired a vast amount of information on HIV prevention compared to individuals lacking information and those moderately informed. This was noted irrespective of their levels of education.

Participants who had no formal schooling and who were moderately knowledgeable to highly knowledgeable had a low susceptibility to HIV infection. Similarly, well

informed individuals who were highly knowledgeable and did not complete secondary school had the lowest odds of HIV infection. Tertiary graduates, irrespective of the tier of knowledge they were categorised into displayed a fairly stable odds of HIV infection but far lower than high school graduates who were at the higher end of the knowledge scale. A participant who either did not attend school or who only attained up to a primary level and lacked information or who was moderately well informed, displayed fluctuating odds of HIV infection. This observation was made across varying levels of HIV preventative knowledge. A participant who attested to attaining education beyond primary level, displayed a stable and almost equal but high odds of HIV infection. The trend observed in the aforementioned interactions, when accounting for spatial variability, was in most respects analogous to the SLR model and the GLMM.

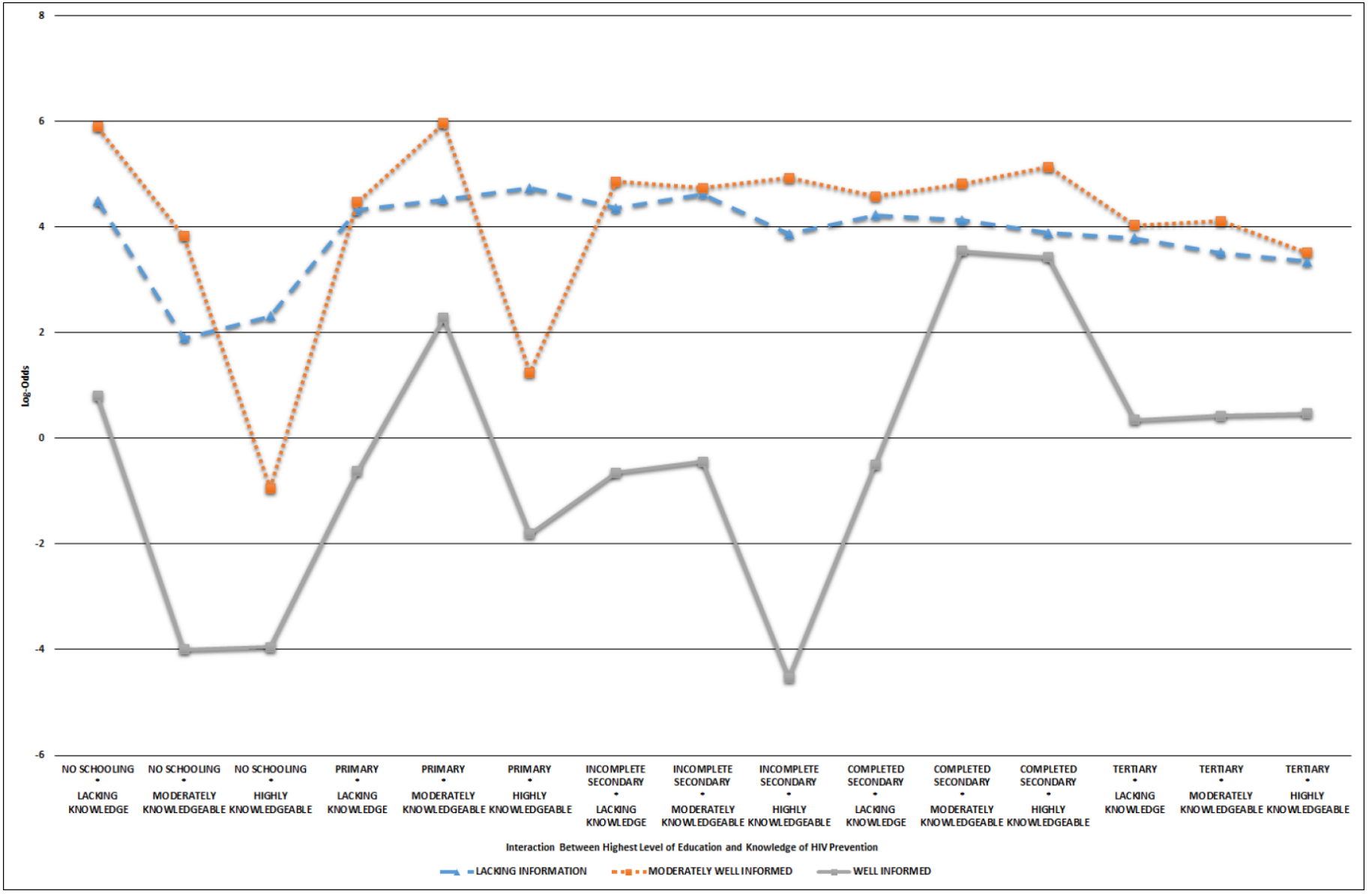

Figure 5.4: The estimated log-odds associated with the interaction of the highest level of education, knowledge of HIV prevention and HIV information acquisition

5.7 Summary and Discussion

The use of geographically weighted regression procedures to explore spatial non-stationarity has been well developed over time. An important assumption in this respect is that of disturbance terms (residuals) being uncorrelated and characterised by a common variance. The presence of spatial autocorrelation and failure to compensate for it, however, challenges the validity of the assumption of homoscedasticity in the residuals with the inevitable result of superfluous inferences. To this end Geary's C, considered well suited to detect spatial autocorrelation at a localised level, is employed to detect for the presence of spatial autocorrelation following which a well suited spatial covariance structure is employed in spatially weighted generalised linear mixed models to account for the presence of spatial autocorrelation. The presence of spatial autocorrelation was compensated for in the analysis.

The results are centered around applying spatial generalised multilevel models, as detailed above in assessing the socio-economic, socio-demographic and behavioural determinants of HIV in adults in a rural setting in the Western region of KwaZulu-Natal, a province of the Republic of South Africa. Research has shown that failure to account for spatial variability could severely skew results and produce biased estimators with inappropriate decisions and conclusions an inevitable consequence (Jossart et al., 2020). From the process outlined in this chapter, the process of accounting for spatial variability can be informally quantified into three stages

$$\text{Suspect} \quad \rightarrow \quad \text{Detect} \quad \rightarrow \quad \text{Represent}$$

In the *suspect* stage, the researcher *suspects* that there is a degree of autocorrelation within the data. At this point the researcher will detect the presence of spatial autocorrelation using either methods outlined in 5.2 and applied in 5.5.1 or one of the many other existing measures for this purpose such as the variogram outlined in 5.3.2 and demonstrated in 5.5.1. Thereafter, they would represent this degree of au-

tocorrelation within their modelling by accounting for the spatial structure of the data under study.

For this study, it was demonstrably proved using Geary's *C* and an empirical semivariogram procedure that there was inherent spatial autocorrelation within the regression residuals. This is because the residuals are able capture vital information that pertains to the structure of the data which allows one to verify whether the normality assumption had been violated or not.

On modelling the data and compensating for the spatial variability to produce an SGLMM, it was observed that the results were in many respects comparable to that produced having accounted for spatial weights and the inclusion of the random factor. Selected predictors characterised as behavioural predictors such as prevailing levels of HIV stigmatisation, and the use of contraception were not significant under the SLR model, GLMM and SGLMM. Predictors such as participant knowledge about HIV prevention were significant under the SLR model and GLMM but not under SGLMM, though in the latter, it was considered contributory to HIV infection when jointly interacting with highest level of education and in a three-way interaction with highest level of education and the acquisition of HIV information by participants.

Increasing one's perception of HIV risk is vital in effecting understanding and acceptance of HIV infection. The results in this chapter revealed that whatever a participant's perceived vulnerability to HIV infection was, it was not observed to be associated with an increased likelihood of HIV infection. Furthermore, the eradication HIV stigma is one of the stated goals of the United Nations and its associated member states as it is seen as a barrier to universal access to treatment and support. The result of the SGLMM revealed that espousal of HIV stigma is not indicative of increased susceptibility to HIV infection. On the subject of sexual behaviour, though

engaging in sexual intercourse and being clinically diagnosed with a sexually transmitted infection is not linked to the likelihood of HIV infection, eschewing contraception is likely to result in HIV infection.

The joint effect investigated the interaction between participants' levels of HIV information acquired, highest level of education and their levels of HIV preventative knowledge. It was unsurprising that the level of knowledge in well informed participants was central to their risk of HIV infection and that level of education was inconsequential in this regard.

Chapter 6

Concluding Remarks

The focus of this study was to construct statistical models in an effort to analyse the prevalence and risk factors of HIV among adults in the age group of 15-49 years in a rural setting in the KwaZulu-Natal Province of South Africa. To achieve this the SLR, GLMM and SGLMM were employed. The results that arose out of the analysis of the baseline study employing these methods showed that socio-economic circumstance, selected socio-demographic, behavioural and cognitive characteristics were significantly associated with HIV infection in the study location. Furthermore, the results emanating from the analysis largely concurred across each method employed.

At the time of writing this dissertation, the world is witnessing the unfolding of another global pandemic, the SARS-CoV-2 (COVID-19) pandemic. Like the HIV-AIDS pandemic, much research is now being coordinated to understand the etiology of the novel coronavirus with the stated aim of producing a vaccine of maximum efficacy thus providing maximum immunity to the population. As the HIV pandemic spreads globally, the research that was initially conducted centered on investigating the etiology of the virus by examining the clinical factors associated with HIV incidence. However, much of the research conducted from the second decade of the pandemic to date has been largely dual focused, encompassing both clinical and socio-economic determinants of HIV. A significant amount of research undertaken in

this regard has shown the inextricable link between HIV incidence and individuals living in destitute economic and environmental conditions. Pellowski et al. (2013) perhaps encapsulate this by unequivocally stating that HIV is a disease rooted in socio-economic inequality.

Using the provincial index of multiple deprivation as a guide in variable selection and the Guttmann scale as a guide for scoring, we constructed household indices of multiple deprivation. These indices assessed household income and material deprivation, infrastructure and living standards, food security and financial security. Employing the household indices of multiple deprivation and profiling households according to their decile of deprivation was not wholly effective in definitively identifying the disparity between the poor and wealthy with respect to their risk of HIV infection.

Deeply ingrained poverty prevalent in rural societies is often generational and is, *inter alia*, observed to be a barrier to accessing amenities and services that are widely available to wealthier populous in urban areas. An inevitable consequence of this is that the rate of HIV infection in South Africa, particularly the province of KwaZulu-Natal, has not been subdued in this, the fourth decade of the HIV pandemic. Wabiri & Taffa (2013) explain that an enduring criticism of an index constructed using household assets is that the line between what constitutes a poor household and what constitutes a poorer household are blurred. This can potentially be rectified by the suggestion of Rutstein (2008) who advocates a dual approach of using variables that are suited to measuring economic status in urban and rural locations. These variables should be adequate in describing a socio-economic situation by differentiating between these two economic groups thereby allowing a composite index to be calculated.

In addition to the plenitude of research conducted on the link between socio-economic

status and HIV incidence, society now has the benefit of more than three decades of behavioural research and its link to HIV infection. The Joint United Nations Programme on HIV/AIDS (2014) has identified HIV stigma as a barrier to universal access to HIV preventative treatment, care and support. It has advocated intervention programmes to curb the scourge of HIV stigma to give credence to the words of the former the South African President Nelson Mandela (2000) who remarked

> *"We need to break the silence, banish stigma and discrimination, and ensure total inclusiveness within the struggle against AIDS."*

South Africa has had a long history of HIV denial which has often been reflected in the once high levels of societal stigmatisation attached to HIV positive individuals. While the majority of participants in this study espoused no stigmatisation toward HIV positive individuals one cannot ignore the fact that high levels of HIV stigmatisation are espoused by those who were themselves HIV positive. This, at a time in which South Africa has achieved near universal coverage of anti-retroviral treatment (ART), is an area of concern especially since HIV preventative information is widespread in the public domain. An analysis into the extent to which HIV stigma contributes toward HIV infection however, found that it was largely inconsequential and that varying levels of HIV stigma was not contributory to an increased likelihood of HIV infection across the differing methodologies applied.

Survey data is not without intricate patterns hidden in individual responses. More than inferring inherent data patterns in isolation, what ought to be studied are the associations between two or more variables contributing to HIV prevalence and the associated risk factors. Multiple correspondence analysis is always a foremost method implemented to understand these associations. Employing multiple correspondence analysis showed the stark association between HIV infection and a diverse range of variables across the social spectrum. These relationships might not have been inferred by simply examining our predictors independently.

On the analytical front, three statistical models were employed to model the HIV status of participants. To account for the presence of survey weights and the binary nature of the response variable, a survey logistic regression model was used. The inclusion of a random factor and in order to compensate for spatial variability, the generalized linear mixed model and a spatial generalized linear mixed model were respectively employed. The results emanating from these models seemingly concurred with each other, where a combination of socio-economic, socio-demographic, behavioural and cognitive factors were found to be significantly associated with HIV infection.

Within the realm of the three models applied, the risk of HIV infection remained largely the same across the varying levels associated with each effect. This was observed in the results as household socio-economic circumstance was indicative of an increased likelihood of HIV infection. There was an increased likelihood of HIV infection among participants resident in households that were classified as increasingly deprived. With the link between socio-economic status and HIV infection being well established, targeted intervention programmes aimed at poverty alleviation are now more than ever required; Baker (2019) shows inequality becoming deeply entrenched in post-apartheid South Africa. Additionally, the risk of HIV infection is observed to be gendered as female participants were at a disproportionately higher risk of HIV infection than males. The risk of HIV infection also appeared to be higher in elder participants than younger participants.

Acquiring preventative knowledge about and adopting the right attitude to HIV infection is vital in stemming the tide of HIV infection. In the age of the internet superhighway, social media is firmly at the centre of knowledge sharing and acquisition. This has contributed to, *inter alia*, societal understanding (or misunderstanding) of the scourge of HIV incidence. Shamu et al. (2020) conducted a study that gauged the level of influence of social media in shaping individuals' cognitive and behavioural

decisions and their relation to HIV risk. There appeared to be a significant relationship between social media use and HIV knowledge, non-condom use and HIV knowledge, and high-risk sexual behaviours and less HIV knowledge.

Our study is not without limitations, the first being the self-reported nature of the data in which the responses were not subject to a reliability test such as Crohnbach's alpha. Additionally, not including race as a factor which may be a contributory factor to HIV infection, may have been a significant drawback. Whilst there is an appreciation of the significance of race as a vital factor in measuring socio-economic status, particularly in an epidemiological setting, we believed that most of the variables we included in the measures of household deprivation were able to capture the relationship between HIV prevalence and socio-economic status. A study of the racial make-up of the study setting found that a significant number of participants were black African. Thus, given the racial demographics of the study region, we did not believe that race could explain HIV prevalence not explicable by the socio-economic indicators of HIV.

The very nature of the cross-sectional data set was a limitation in itself. Cross-sectional studies are a preferred method of data collection in a plethora of studies across the social scientific, epidemiological and public health spectrum. This is true as cross-sectional studies are a direct measure of exposure and outcome. However, the nature of cross-sectional data is such that there is no provision for inferring temporal relationship unless there exists a strong plausibility for one of the directions of associations (Kestenbaum, 2019).

Furthermore, a serious limitation centered around joint effects. While subject to the intuition of the researcher, there was not as many joint effects included in the analysis as there should have been. The studies, as outlined in Section 1.1 are observed to link socio-economic status with behavioral and cognitive factors. In this regard a

separate study should be undertaken in which selected though relevant joint effects are investigated regarding their contribution to predicting HIV infection.

This study did not consider a variety of clinical factors as contributory to HIV infection. Comorbidities associated with HIV infection are defined by diseases diagnosed outside of an Acquired Immune Deficiency Syndrome - associated illness. Lorenc et al. (2014) cite diabetes mellitus, cardiovascular disease (CVD e.g. hypertension), respiratory illnesses (e.g. chronic obstructive pulmonary disease, pneumonia and tuberculosis) and hepatic diseases (e.g. hepatitis B and hepatitis C). Additionally, psychiatric disorders such as depression, anxiety, schizophrenia and cognitive impairment are diagnosed in HIV positive individuals. These conditions tend to worsen with worsening HIV severity. Despite the wide variety of research conducted into HIV infection and associated comorbidities, the need for more research in this area is encouraged, as for example conducted by Nlooto (2017). This study found that coinfections and comorbidities are prevalent in HIV positive individuals.

In acknowledging these limitations, we also acknowledge that our study does not terminate here. As stated, society has had the benefit of research into HIV infection not only from a clinical perspective, but a societal one as well. These contributions to our understanding of HIV infection are by no means exhaustive, and to gain more knowledge and insight from many perspectives is always encouraged, particularly in a fast moving and ever-changing world. To chart a future direction for this study, we recommend that the limitations addressed be incorporated in a longitudinal study. A longitudinal study which incorporates the spatial effects of this study will be advantageous as it will address developmental changes such as a *spatio-temporal shift* over a sustained time period rather than at a single time point. This could contribute to the sum of societal knowledge of HIV infection in perpetuity.

References

Acquah, H. D. G. (2010). Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship. *Journal of Development and Agricultural Economics*, *2*(1), 001–006.

Agresti, A. (1990). *Categorical Data Analysis*. John Wiley and Sons Inc. New Jersey.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis (2nd ed.)*. John Wiley and Sons Inc. New Jersey.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723.

Allison, P. D. (2013). Why I dont trust the Hosmer-Lemeshow test for logistic regression. https://statisticalhorizons.com/hosmer-lemeshow. Accessed: 2020-06-19.

Archer, K. J. (2001). Goodness-of-fit tests for logistic regression models developed using data collected from a complex sampling design. PhD Thesis, Ohio State University.

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

Arnold, B. W. (2013). Introduction to Geostatistics for Site Characterization and Safety Assessment. Tech. rep., Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).

Baker, A. (2019). What South Africa can teach us as worldwide inequality grows. *Time Magazine*.

Banerjee, S., Gelfand, A. E., & Carlin, B. P. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton: Florida.

Bärnighausen, T., Hosegood, V., Timaeus, I. M., & Newell, M.-L. (2007). The socio-economic determinants of HIV incidence: evidence from a longitudinal, population-based study in rural South Africa. *AIDS*, *21*(Suppl 7), S29–S38.

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods*, *2*(2), 131–160.

Beltrami, E. (1873). Sulle funzioni bilineari. *Giornale di Matematiche ad Uso degli Studenti Delle Universita*, *11*(2), 98–106.

Benzécri, J. (1979). Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire, addendum et erratum à [bin. mult.]. *Cahiers de l'Analyse des Données*, *4*(3), 377–378.

Bonat, W. H., & Ribeiro, P. J. (2016). Practical likelihood analysis for spatial generalized linear mixed models. *Environmetrics*, *27*(2), 83–89.

Booysen, F. l. R. (2004). HIV/AIDS, poverty and risky sexual behaviour in South Africa. *African Journal of AIDS Research*, *3*(1), 57–67.

Bozdogan, H. (1987). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370.

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

Broström, G., & Holmberg, H. (2011). Generalized linear models with clustered data: Fixed and random effects models. *Computational Statistics & Data Analysis*, *55*(12), 3123–3134.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*(3), 473–514.

Burt, C. (1950). The factorial analysis of qualitative data. *British Journal of Statistical Psychology*, *3*(3), 166–185.

Buthelezi, U. E., Davidson, C. L., & Kharsany, A. B. M. (2016). Strengthening HIV surveillance: Measurements to track the epidemic in real time. *African Journal of AIDS Research*, *15*(2), 89–98.

Caroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of" eckart-young" decomposition. *Psychometrika*, *35*, 238–319.

Cassel, C., Särndal, C. E., & Wretman, J. H. (1977). *Foundations of Inference in Survey Sampling*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. John Wiley and Sons Inc.

Centers for Disease Control, & Prevention (2014). Introduction to public health. In: Public Health 101 Series, Atlanta, GA: US department of Health and Human Services, CDC. https://www.cdc.gov/publichealth101/surveillance.html. Accessed: 2020-07-13.

Chen, Y. (2016). Spatial autocorrelation approaches to testing residuals from least squares regression. *PloS one*, *11*(1), e0146865.

Children's Amendment Act of South Africa (2007). Children's Amendment Act 41 of 2007.

Cliff, A. D., & Ord, J. K. (1975). The choice of a test for spatial autocorrelation. *Display and analysis of spatial data*, *54*, 77.

Cliff, A. D., & Ord, J. K. (1981). Spatial Processes: Models and Applications, Pion Limited.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Blackwell, New York.

Darroch, J. N. (1974). Multiplicative and additive interaction in contingency tables. *Biometrika*, *61*(1), 207–214.

De Leeuw, J. (1983). On the prehistory of correspondence analysis. *Statistica Neerlandica*, *37*(4), 161–164.

De Leeuw, J., & Mair, P. (2009). Simple and canonical correspondence analysis using the r package anacor. *Journal of Statistical Software*, *31*(5), 1–18.

Deming, W. E., & Stephan, F. F. (1941). One the Interpretation of Cencus as Sample. *Journal of the American Statistical Association*, *36*(213), 45–49.

Fenton, L. (2004). Preventing HIV/AIDS through poverty reduction: the only sustainable solution? *Lancet (London, England)*, *364*(9440), 1186–1187.

Food and Agriculture Organization (1996). *Rome Declaration on World Food Security and World Food Summit Plan of Action: World Food Summit 13-17 November 1996, Rome, Italy*. FAO.

Fortin, M. (2013). Population-averaged predictions with generalized linear mixed-effects models in forestry: an estimator based on gauss- hermite quadrature. *Canadian Journal of Forest Research*, *43*(2), 129–138.

Galvani, A. P., Pandey, A., Fitzpatrick, M. C., Medlock, J., & Gray, G. E. (2018). Defining control of HIV epidemics. *The Lancet HIV*, *5*(11), 667–670.

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

Gelman, A. (2007). Struggles with Survey Weighting and Logistic Regression. *Statistical Science*, *22*(2), 153–164.

George, G., Cawood, C., Puren, A., Khanyile, D., Gerritsen, A., Govender, K., Beckett, S., Glenshaw, M., Diallo, K., Ayalew, K., Reddy, T., Maurai, L., Kufa-Chakezha, T., & Kharsany, A. o. (2020). Evaluating DREAMS HIV prevention interventions

targeting adolescent girls and young women in high HIV prevalence districts in South Africa: protocol for a cross-sectional study. *BMC Women's Health*, *20*(1), 1–11.

Global Adult Tobacco Survey Collaborative Group (2010). Global Adult Tobacco Survey (GATS): Sample Weights Manual, Version 2.0.

Glynn, J. R., Caraël, M., Auvert, B., Kahindo, M., Chege, J., Musonda, R., Kaona, F., Buvé, A., & The Study Group on the Heterogeneity of HIV Epidemics in African Cities (2001). Why do young women have a much higher prevalence of HIV than young men? a study in kisumu, kenya and ndola, zambia. *AIDS*, *15*, 51–60.

Goodman, L. A. (1986). Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *International Statistical Review/Revue Internationale de Statistique*, (pp. 243–309).

Graubard, B. I., & Korn, E. L. (2002). Inference for Super-Population Parameters Using Sample Means. *Statistical Science*, *17*(1), 73–96.

Graubard, B. I., Korn, E. L., & Midthune, D. (1997). Testing goodness-of-fit for logistic regression with survey data. In *Proceedings of the American Statistical Association*.

Greenacre, M. (1993). *Correspondence Analysis in Practice*. London Academic Press.

Greenacre, M. (1994). Multiple and joint correspondence analysis. *Correspondence analysis in the social sciences*, (pp. 141–161).

Greenacre, M. (2007). *Correspondence Analysis in Practice*. Taylor and Francis Group, LLC.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. London (UK) Academic Press.

Greenacre, M. J. (1988). Correspondence analysis of multivariate categorical data by weighted least-squares. *Biometrika*, *75*(3), 457–467.

Greenacre, M. J., & Hastie, T. (1987). The geometric interpretation of correspondence analysis. *Journal of the American Statistical Association*, *82*(398), 437–447.

Gregson, S., & Garnett, G. P. (2000). Contrasting gender differentials in HIV-1 prevalence and associated mortality increase in eastern and southern Africa: artefact of data or natural course of epidemics? *AIDS*, *14*(3).

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. *The Prediciton of Personal Adjustment*.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*(2), 139–150.

Hagenaars, A. J. M., De Vos, K., & Zaidi, M. A. (1994). Poverty statistics in the late 1980s: Research based on micro-data.

Handayani, D., Notodiputro, K. A., Sadik, K., & Kurnia, A. (2017). A comparative study of approximation methods for maximum likelihood estimation in generalized linear mixed models (glmm). In *AIP Conference Proceedings*, vol. 1827, (p. 020033). AIP Publishing LLC.

Hanley, J. A., & McNeil, B. J. (1982). The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, *143*, 29–36.

Hargreaves, J. R., & Glynn, J. R. (2002). Educational attainment and HIV-1 infection in developing countries: a systematic review. *Tropical Medicine & International Health*, *7*(6), 489–498.

Harshman, R. A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phoenetics*, *16*, 1–84.

Harshman, R. A., & Lundy, M. E. (1984). The PARAFAC model for three-way factor analysis and multidimensional scaling. *Research methods for multimode data analysis*, *46*, 122–215.

Hartzel, J., Agresti, A., & Caffo, B. (2001). Multinomial logit random effects models. *Statistical Modelling*, *1*(2), 81–102.

Hilbe, J. M. (2009). *Logistic Regression Models*. CRC press.

Hill, M. O. (1973). Reciprocal averaging: an eigenvector method of ordination. *The Journal of Ecology*, (pp. 237–249).

Horvitz, D. G., & Thompson, D. J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, *47*(260), 663–685.

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

Hosmer, D. W. J., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*, vol. 3. John Wiley and Sons Inc. Hoboken, New Jersey.

Houle, B., Mojola, S. A., Angotti, N., Schatz, E., Gómez-Olivé, F. X., Clark, S. J., Williams, J. R., Kabudula, C., Tollman, S., & Menken, J. (2018). Sexual behavior and HIV risk across the life course in rural South Africa: trends and comparisons. *AIDS care*, *30*(11), 1435–1443.

Husson, F., Josse, J., & Saporta, G. (2016). Jan de Leeuw and the French school of data analysis. *Journal of Statistical Software*.

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

Joint United Nations Programme on HIV/AIDS (2014). Reduction of HIV-related stigma and discrimination. *UNAIDS Information Production Unit*, (pp. 1–18).

Jordan, C. (1874). Mémoire sur les formes bilinéaires. *Journal de Mathématiques Pures et Appliquées*, *19*, 35–54.

Jossart, J., Theuerkauf, S. J., Wickliffe, L. C., & Morris Jr, J. A. (2020). Applications of spatial autocorrelation analyses for marine aquaculture siting. *Frontiers in Marine Science*, *6*, 806.

Journel, A. G., & Huijbregts, C. J. (1978). *Mining Geostatistics*. Academic Press.

Kendall, M., & Stuart, A. (1979). *The Advanced Theory of Statistics*. Charles Griffin, London.

Kendall, M. G. (1972). The history and future of statistics. *Statistical Papers in Honor of George W. Snedecor*, (pp. 193–210).

Kestenbaum, B. (2019). *Cross-Sectional Studies in Epidemiology and Biostatistics*. Springer.

Kharsany, A. B. M., Cawood, C., Khanyile, D., Grobler, A., Mckinnon, L. R., Samsunder, N., Frohlich, J. A., Karim, Q. A., Puren, A., Welte, A., George, G., Govender, K., Toledo, C., Chipeta, Z., Zembe, L., Glenshaw, M. T., Madurai, L., Deyde, V. M., & Bere, A. (2015). Strengthening HIV surveillance in the antiretroviral therapy era: Rationale and design of a longitudinal study to monitor HIV prevalence and incidence in the umgungundlovu District, Kwazulu-Natal, South africa. *BMC Public Health*, *15*(1), 1149.

Kharsany, A. B. M., McKinnon, L. R., Lewis, L., Cawood, C., Khanyile, D., Maseko, D. V., Goodman, T. C., Beckett, S., Govender, K., George, G., Kassahun, A. A., & Toledo, C. (2020). Population prevalence of sexually transmitted infections in a high HIV burden district in Kwazulu-Natal, South Africa: Implications for HIV epidemic control. *International Journal of Infectious Diseases*, *98*, 130–137.

Kish, L. (1965). *Survey Sampling*. Wiley Series of Survey Methodology. John Wiley and Sons Inc.

Klasen, S. (2000). Measuring poverty and deprivation in South Africa. *Review of income and wealth*, *46*(1), 33–58.

Korn, E. L., & Graubard, B. I. (1990). Simultaneous testing of regression coefficients with complex survey data: Use of bonferroni t statistics. *The American Statistician*, *44*(4), 270–276.

Kroonenberg, P. M. (1989). Singular value decompositions of interactions in three-way contingency tables. In *Multiway Data Analysis*, (pp. 169–184). North-Holland Publishing Co., Amsterdam.

Lee, C. C., Liang, C. M., & Liu, Y. T. (2019). A comaprison of the predictive powers of tenure choices between property ownership and renting. *International Journal of Strategic Property Management*, *23*(2), 130–141.

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

Lorenc, A., Ananthavarathan, P., Lorigan, J., Banarsee, R., Jowata, M., & Brook, G. (2014). The prevalence of comorbidities among people living with HIV in Brent: A diverse London Borough. *London Journal of Primary Care*, *6*(4), 84–90.

MacPhail, C., Williams, B. G., & Campbell, C. (2002). Relative risk of HIV infection among young men and women in a South African township. *International Journal of STD & AIDS*, *13*(5), 331–342.

Mandela, N. R. (2000). Closing address by Nelson Mandela at 13th International AIDS conference, Durban, 14 July 2000. http://www.mandela.gov.za/mandela_speeches/2000/000714_aidsconf.htm.

Manjengwa, P. G., Mangold, K., Musekiwa, A., & Kuonza, L. R. (2019). Cognitive and behavioural determinants of multiple sexual partnerships and condom use in South Africa: Results of a national survey. *Southern African Journal of HIV Medicine*, *20*(1), 1–9.

Matérn, B. (1960). Spatial variation-stochastic models and their application to some

problems in forest surveys and other sampling investigations. Meddelanden fran statens skogsforskningsintitut, Almaenna foerlaget, Stockholm. (1986), 49 (5).

Matheron, G. (1963). Principles of geostatistics. *Economic geology*, *58*(8), 1246–1266.

Matheron, G. (1965). *Les variables régionalisées et leur estimation: une application de la théorie des fonctions aléatoires aux sciences de la nature*. Masson et CIE.

Matheron, G., Pawlowsky-Glahn, V., & Serra, J. (2019). *Matheron's Theory of Regionalized Variables*. International Association for Mathematical Geology Studies i. Oxford University Press.

McCullogh, C. E., & Searle, S. R. (2008). *Generalized Linear and Mixed Models 2nd Edition*. John Wiley and Sons.

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

Muchiri, E., Odimegwu, C., & De Wet, N. (2017). HIV risk perception and consistency in condom use among adolescents and young adults in urban Cape Town, South Africa: a cumulative risk analysis. *Southern African Journal of Infectious Diseases*, *32*(3), 105–110.

Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. University of Toronto Press, Toronto.

Nlooto, M. (2017). Comorbidities of HIV infection and health care seeking behavior among HIV infected patients attending public sector healthcare facilities in Kwazulu-Natal: A cross sectional study. *PLoS One*, *12*(2), e0170983.

Noble, M., Babita, M., Barnes, H., Dibben, C., Magasela, W., Noble, S., Ntshongwana, P., Phillips, H., Rama, S., Roberts, B., Wright, G., & Zungu, S. (2006). The provincial indices of multiple deprivation for South Africa 2001.

OECD (1998). OECD Project on income distribution and poverty. http://www.oecd.org/social/inequality.htm.

Oltean, H., & Gagnier, J. J. (2015). Use of clustering analysis in randomized controlled trials in orthopaedic surgery. *BMC Medical Research Methodology*, *15*(1), 17.

Pearson, K. (1906). On certain points connected with scale order in the case of the correlation of two characters which for some arrangement give a linear regression line. *Biometrika*, *5*(2), 176–178.

Pellowski, J. A., Kalichman, S. C., Matthews, K. A., & Adler, N. (2013). A pandemic of the poor: Social disadvantage and the us hiv epidemic. *American Psychologist*, *68*(4), 197–209.

Perera, A. A. P. N. M., Sooriyarachchi, M. R., & Wickramasuriya, S. L. (2014). A goodness of fit test for the multilevel logistic model. *Communications in Statistics-Simulation and Computation*, *45*(2), 643–659.

Pettifor, A. E., Kleinschmidt, I., Levin, J., Rees, H. V., MacPhail, C., Madikizela-Hlongwa, L., Vermaak, K., Napier, G., Stevens, W., & Padian, N. S. (2005). A community-based study to examine the effect of a youth HIV prevention intervention on young people aged 15–24 in South Africa: results of the baseline survey. *Tropical Medicine & International Health*, *10*(10), 971–980.

Pfeffermann, D., Krieger, A. M., & Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, *8*(4), 1087–1114.

Pffeferman, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for Unequal Selection in Multilevel Models. *J.R Statist Soc*, *60*(1), 23–40.

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of computational and Graphical Statistics*, *4*(1), 12–35.

Pinheiro, J. C., & Chao, E. C. (2006). Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics*, *15*(1), 58–81.

Potthoff, R. F., Woodbury, M. A., & Manton, K. G. (1992). "Equivalent Sample Size" and "Equivalent Degrees of Freedom" Refinements for Inference Using Survey Weights under Superpopulation Models. *Journal of the American Statistcal Association*, *87*(418), 383–396.

Probst, C., Parry, C. D. H., & Rehm, J. (2016). Socio-economic differences in HIV / AIDS mortality in South Africa. *Tropical Medicine & International Health*, *21*(7), 846–855.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, (pp. 111–163).

Rao, J. N., & Scott, A. J. (1984). On Chi Square Tests for Multiway Contigency Tables with Cell Proportions Estimated from Survey Data. *The Annals of Statistics*, *12*, 46–60.

Rao, J. N. K., & Scott, A. J. (1987). On simple adjustments to chi-square tests with sample survey data. *The Annals of Statistics*, (pp. 385–397).

Rich, J. L. (2018). Comparison of generalized linear mixed model estimation methods.

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

Roberts Jr, J. M. (2000). Correspondence analysis of two-mode network data. *Social Networks*, *22*(1), 65–72.

Rodríguez, G., & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *158*(1), 73–89.

Rosenberg, M. S., Gómez-Olivé, F. X., Rohr, J. K., Houle, B. C., Kabudula, C. W., Wagner, R. G., Salomon, J. A. S., Kahn, K., Berkman, L. F., Tollman, S. M., & Bärnighausen, T. (2017). Sexual behaviors and HIV status: a population-based

study among older adults in rural South Africa. *J Acquir Immune Defic Syndr.*, *74*(1), e9–e17.

Rudolfer, S. M. (2002). Diagnosis of Carpel Tunnel Syndrome using Logistic Regression.

Rutstein, S. O. (2008). The DHS wealth index: approaches for rural and urban areas. 2008, Washington, DC: Macro International Inc.

SAS Institute Inc. (2016). *SAS/STAT 14.2 User's Guide Cary, NC: SAS Institute Inc*. SAS Institute Inc.

Scott, A. (2007). Rao-Scott corrections and their impact. In *Proceedings of the 2007 joint statistical meetings, Salt Lake City, Utah*.

Searle, S. R., Casella, G., & McCulloch, C. E. (2006). *Variance components*, vol. 391. John Wiley & Sons.

Shamu, S., Khupakonke, S., Farirai, T., Slabbert, J., Chidarikire, T., Guloba, G., & Nkhwashu, N. (2020). Knowledge, attitudes and practices of young adults towards HIV prevention: An analysis of baseline data from a community-based HIV prevention intervention study in two high HIV burden districts, South Africa. *BMC Public Health*, *20*(1), 1–10.

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

Simbayi, L., Zuma, K., Zungu, N., Moyo, S., Marinda, E., Jooste, S., Mabaso, M., Ramlagan, S., North, A., Van Zyl, J., Mohlabane, N., Dietrich, C., Naidoo, I., & the SABSSM V Team (2019) (2019). South African National HIV Prevalence, Incidence, Behaviour and Communication Survey, 2017: Towards achieving the UNAIDS 90-90-90 targets.

Skinner, S. J., Holt, D., & Smith, T. M. F. (1989). *Analysis of Complex Surveys*. John Wiley and Sons Inc.

Skrondal, A., & Rabe-Hesketh, S. (2009). Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(3), 659–687.

Statistics South Africa (2018a). National poverty lines. *Statistical Release P03101*.

Statistics South Africa (2018b). National poverty lines. http://www.statssa.gov.za/publications/P03101/P031012018.pdf. Accessed: 2020-07-05.

Statistics South Africa (2020). Mid-year population estimates: 2019. *Statistics South Africa*.

Steinberg, M., Johnson, S., Schierhout, G., Ndegwa, D., Hall, K., Russell, B., & Morgan, J. (2002). Hitting home: How households cope with the impact of the HIV/AIDS epidemic. *A survey of household affected by HIV/AIDS in South Africa. ABT Associates*.

Stroup, W. W. (2012). *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. CRC Press.

Sylvester, J. J. (1889). On the reduction of a bilinear quantic of the nth order to the form of a sum of n products by a double orthogonal substitution. *Messenger of Mathematics*, *19*(6), 42–46.

Thomas, D. R., & Rao, J. N. K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, *82*(398), 630–636.

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*(393), 82–86.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, *31*(3), 279–311.

Tukey, J. W. (1977). *Exploratory Data Analysis*, vol. 2. Reading, MA.

uMgungundlovu District Municipality (2020). Integrated Development Plan (IDP) - 2020/2021 Review.

United Nations Programme on HIV/AIDS, J. (2017). Making the end of AIDS real: Consensus building around what we mean by "epidemic control"—a meeting convened by the UNAIDS Science Panel—Glion, Switzerland, 4-6 October 2017. Geneva: Joint United Nations Programme on HIV. *AIDS*.

United States Census Bureau (2020). Current versus constant (or real) dollars. https://www.census.gov/topics/income-poverty/income/guidance/current-vs-constant-dollars.html.

Uno, H., Cai, T., Pencina, M. J., Agostino, R. B., & Wei, L. J. (2019). On the C-statistics for evaluating overall accuracy of risk prediction procedures with censored survival data. *Statistics in Medicine*, *30*(10), 1105–1117.

Wabiri, N., & Taffa, N. (2013). Socio-economic inequality and HIV in South Africa. *BMC Public Health*, *13*(1), 1037.

Wang, J. (2013). *Incorporating Survey Weights into Logistic Regression Models*. Masters Thesis, Worcester Polytechnic Institute.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*(1), 92–107.

West, B. T. (2001). Spatial analysis of a small area problem. *Master's Thesis: University of Michigan*.

Wilson, J. R., & Lorenz, K. A. (2015). *Modeling Binary Correlated Responses using SAS, SPSS and R*, vol. 9. Springer.

Zhang, D., Ren, N., & Hou, X. (2018). An Improved Logistic Regression Model based on a Spatially Weighted Technique (ILRBSWT v1. 0) and its Application to Mineral Prospectivity Mapping. *Geoscientific Model Development*, *11*(6), 2525–2539.