



**A Frequentist and a Bayesian Approaches to
Estimating HIV Prevalence Accounting for
Non-response using Population-based Survey Data**

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

Doctor of Philosophy

in the

School of Mathematics, Statistics and Computer Science

of the

University of KwaZulu-Natal

by

Amos Chinomona

February 2016

Published Articles from the Research

1. Chinomona, A. and Mwambi, H. (2015). Multiple Imputations for Non-response when Estimating HIV Prevalence. BMC Public Health, 15:1059 DOI 10.1186/s12889-1-1.
2. Chinomona, A. and Mwambi, H. (2015). Estimating HIV Prevalence in Zimbabwe using Population-based Survey Data. PlosOne, DOI: 10.1371/journal.pone.0140896.

Declaration

I declare that the research work described in this thesis is my original work carried out in the School of Statistics, Mathematics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg under the supervision of Professor H. Mwambi. The work has not been submitted in any form for any qualification to any University and where use has been made of other authors' work, it has duly been acknowledged.

Signature:

Date:

Amos Chinomona

Signature:

Date:

Professor H. Mwambi

Dedication

This research is dedicated to my wife Perpetua and son Kutenda.

Acknowledgments

I would like to express my sincere appreciation to my supervisor, Professor H. Mwambi for your unwavering support and encouragement throughout the journey that culminated into this thesis. It would not have been possible to come this far. A special thanks goes to the Staff and postgraduate students in the School of Mathematics, Statistics and Computer Science at the Pietermaritzburg campus for the words of encouragement from the informal chats that we had. Additionally, many thanks to Christel and Bev for your assistance with all the administrative issues.

Many thanks go to my colleagues in the Department of Statistics at Rhodes University for all the encouragement. I would also like to acknowledge the financial support I got from the Rhodes University Research Council grant that enabled me to travel to the University of KwaZulu Natal to conduct my research. Special thanks goes to the Demographic and Health Surveys (DHS) for providing me with the data that I used in this research.

To my friends Albert, Wilbert and Tinashe, your support and words of encouragement did not go unnoticed, thank you. To my guys in Pietermaritzburg, Tariro, Einstein, Tafadzwa, Terence and Prudence, and my sister Joanah and your family, I sincerely appreciated your support. To my lovely wife Perpetua and son Kutenda, sincere appreciation for your unwavering support and understanding especially during those lengthy periods of my absence from home. To all those whom I could not mention by name, your contributions at various levels are greatly appreciated. Lastly but not least, I would like to give glory and honour to God Almighty for His grace and unconditional love that He showed me during the tough times I went through working on this research.

Abstract

Enhanced and novel frequentist and Bayesian approaches to estimating disease measures such as HIV prevalence utilizing the recent advances in statistical computing software are explored and applied making use of population-based complex survey data. In particular design-consistent estimates and logistic regression models for HIV prevalence are respectively computed and fitted using each of the approaches.

Practical survey data are rarely obtained using simple random sampling schemes, instead complex sampling designs, that are designed to reflect complex underlying population structures, are employed. These designs usually involve stratification, multistage sampling and unequal selection probability of sampling units giving rise to data that are hierarchical (multilevel), clustered, and hence correlated. This is particularly true for large-scale population-based surveys. Consequently this often gives rise to units that are correlated within clusters as well as multiple sources of variability rendering standard statistical methods based on the assumption of independence of units inappropriate. Survey logistic regression models built from a generalized linear modelling framework were used to explain the variation in HIV prevalence accounting for the non-independence of the units. In addition, a hierarchical logistic regression model built from a generalized linear mixed modelling framework was used to capture the variability and correlation of the units within clusters and further determine how different layers interact and impact on a response variable. In particular, the logistic regression models for HIV prevalence on demographic, behavioural and socio-economic variables were developed from a frequentist and a Bayesian perspective.

Statistical methods that incorporate prior known information about unknown parameters are vital in most scientific and biological research especially in studies where replicative experimental investigations are not possible. The Bayesian statistical paradigm offers a framework upon which a prior distribution of a parameter can be combined with the likelihood of the observed data to obtain a posterior distribution for explaining the

stochastic variation in a response variable. Computer-intensive simulation-based algorithms such as the Markov chain Monte Carlo (MCMC) methods were used to draw samples from the posterior distribution for inference purposes. A Bayesian logistic regression model for HIV prevalence on demographic and socio-economic variables was fitted from a generalized linear modelling framework using the MCMC algorithms.

Furthermore, practical complex survey data are often characterized by missing observations due to non-response, a phenomenon that is true to the data used for the current research. Often, the analyses of such data take a complete case approach, that is taking a list-wise deletion of all cases with missing observations, assuming that missing values are missing completely at random (MCAR). In the current research, we systematically simulate or generate multiple values for the missing observations under a multiple imputation method accounting for the structure of the data. A rectangular complete data set is produced and the variability or uncertainty induced by the very process of imputing the values for the missing observations is accounted for.

The study utilizes complex (multi-layered and clustered data with missing values) survey data obtained from the 2010-11 Zimbabwe Demographic and Health Surveys (2010-11ZDHS). The results show that HIV prevalence varies considerably across subgroups of the population. All the analyses are done using **R** statistical software packages.

Contents

1	Introduction	1
1.1	Background	1
1.2	Sampling under complex surveys	5
1.2.1	Sampling weights	8
1.2.2	The design effect (deff)	9
1.2.3	Variance estimation under complex sampling	10
1.3	Approaches to statistical analysis of survey data	11
1.3.1	Design-based approach	12
1.3.2	Model-based and model assisted approaches	13
1.3.3	Bayesian approach	14
1.4	Missing data in surveys	16
1.5	Statistical computing	17
1.6	Objectives of the study	18
1.7	Significance of the study	19
1.8	Thesis layout	21
2	Exploratory data analysis	23
2.1	Introduction	23
2.2	The data	23
2.3	The 2010-11ZDHS data	26

2.4	The key variables	26
2.5	Basic relationships between the explanatory variables	31
3	Estimating HIV prevalence in Zimbabwe using population-based survey data	34
3.1	Introduction	35
3.2	Methods	38
3.2.1	The data	38
3.2.2	Statistical computing	39
3.2.3	Statistical methods	39
3.3	Results	49
3.3.1	Descriptive analysis	50
3.3.2	Logistic regression analysis	55
3.4	Discussion	62
3.5	Potential strengths and limitations of study	63
3.6	Conclusion	64
4	Multiple imputation for non-response when estimating HIV prevalence using population-based survey data	66
4.1	Introduction	68
4.2	Methods	72
4.2.1	Types of missingness	72
4.2.2	Multiple imputations	73
4.2.3	The analysis model	76
4.2.4	The data	77
4.2.5	Statistical computations	78
4.3	Results	80
4.3.1	Prevalence estimation results	80

4.3.2	Logistic regression results	86
4.4	Discussion	89
4.5	Potential strength and limitations of the study	92
4.6	Conclusion	93
5	Hierarchical logistic regression for estimating risk of HIV using population-based survey data	94
5.1	Introduction	95
5.2	Methods	99
5.2.1	Hierarchical modelling	99
5.2.1.1	Hierarchical modelling for a binomial response variable	102
5.2.1.2	Generalized linear mixed effects model	103
5.2.2	Handling missing data via multiple imputation	106
5.2.2.1	Types of missingness	106
5.2.2.2	The multiple imputation procedure	107
5.2.3	The Data	109
5.2.4	Statistical Software	111
5.3	Results and discussion	112
5.4	Conclusion	118
6	A Bayesian logistic regression for estimating risk of HIV using population-based survey data	120
6.1	Introduction	121
6.2	Methods	124
6.2.1	An overview of the Bayesian methods	124
6.2.1.1	The prior distributions	124
6.2.1.2	The likelihood	125
6.2.1.3	The posterior distributions	126

6.2.1.4	Determining the posterior distribution	127
6.2.1.5	Convergence diagnostics for a Markov chain	130
6.2.2	Bayesian logistic regression	132
6.2.3	Statistical computing	135
6.3	Results and discussion	136
6.4	Conclusion	144
7	Bayesian hierarchical logistic regression for estimating risk of HIV using population-based survey data	145
7.1	Introduction	147
7.2	Methods	149
7.2.1	Hierarchical Bayesian modelling	149
7.2.2	Statistical computing	151
7.3	Results and discussion	152
7.4	Conclusions	155
8	A predictive model for HIV using a semi-parametric spline approach	157
8.1	Introduction	158
8.2	Methods	160
8.2.1	Generalized additive models	160
8.2.2	Logistic generalized additive regression	163
8.2.3	Statistical computing	164
8.2.4	The data	164
8.3	Results and discussion	165
8.4	Conclusion	170
9	Conclusions and future research	171
9.1	Introduction	171
9.2	Summary of conclusions	171

9.3	Strengths and limitations of the study and future research	173
9.3.1	Strengths	173
9.3.2	Limitations	174
9.3.3	Direction for future research	175
A	R code for Chapter 3	196
B	R code for Chapter 4	198
C	R code for Chapter 5	199
D	R code for Chapter 6	200
E	R code for Chapter 7	202
F	R code for Chapter 8	203

List of Figures

3.1	HIV prevalence across the different categories of marital status	53
3.2	HIV prevalence across the different categories of five-year age-groups . .	54
3.3	HIV prevalence in the different administrative provinces of the country	54
3.4	The average HIV prevalence per a given age versus age	56
3.5	HIV prevalence estimates and their respective 95% confidence intervals by age group for males and females separately	58
4.1	Estimates of HIV prevalence by marital status obtained using complete case analysis and multiple imputations	84
4.2	Estimates of HIV prevalence per age group obtained using complete case analysis and multiple imputations	85
4.3	Estimates of HIV prevalence by province using complete case analysis and multiple imputations	86
5.1	Three level hierarchical data structure	96
6.1	Trace plots and density plots for the first six coefficients from the poste- rior distribution	138
6.2	Geweke plots for the first six coefficients from the posterior distribution	139
8.1	Plot showing the relationship between HIV prevalence and age	159

8.2	Plot of the final logistic GAM, a semi-parametric model of HIV with a smooth function, $\hat{f}(\text{age})$, of age together with the factors of gender, marital status, place of residence and literacy levels and their respective 95% confidence intervals.	168
-----	--	-----

List of Tables

2.1	Crude summary of the HIV test results for the respondents	27
2.2	Distribution of the respondents by education level	27
2.3	Summary of the literacy levels of the respondents	28
2.4	Summary of the five year age groups for the respondents	28
2.5	Summary of the marital status for the respondents	29
2.6	Summary of access to information for the respondents	29
2.7	Summary of the employment status of the respondents	30
2.8	Summary of the wealth indices of the respondents	30
2.9	Cross tabulation of gender by age group	32
2.10	Contingency table for marital status by age group	33
3.1	Crude design-based subgroup estimates of HIV prevalence along with their respective (crude) ORs and 95% confidence intervals	51
3.2	Rao-Scott (F – based) test statistics and p-values, for association of individual predictor variables and HIV status	57
3.3	Estimated adjusted ORs and crude ORs together with their respective 95% confidence intervals for the parameter estimates for the logistic regression model	60
4.1	Variables and percentages of missing data	78
4.2	Crude subgroup estimates and their standard errors of HIV prevalence for (a) complete case analysis and (b) multiple imputation.	82

4.3	Parameter estimates and standard errors of logistic regression models under (a) complete case analysis and (b) multiple imputation analysis	87
4.4	ORs for the estimates of the survey logistic regression models under (a) complete case analysis and (b) multiple imputation analysis and their respective 95% confidence intervals	89
5.1	Hierarchical structure of the 2010-11 ZDHS data	110
5.2	Percentages of missing data per variable	111
5.3	Parameter estimates, standard errors and p-values for the generalized linear mixed models under (a) multiple imputation analysis (b) complete case analysis	114
5.4	Odds ratios and their corresponding 95% confidence intervals for the best models under multiple imputations and complete case analysis	116
6.1	Parameter estimates, standard errors and p-values for the posterior distribution for the observed data using a non-informative prior distribution	137
6.2	Summary measures (mean, median, standard deviation and credibility intervals) for the coefficients obtained from the posterior distribution .	140
6.3	Odds ratios and credibility intervals for parameter estimates obtained from simulating from the posterior distribution	142
7.1	The marginal posterior distribution for the fixed effects	153
7.2	Variance components for the random effects	154
8.1	Results of a generalized additive model (GAM) with (a) the parametric terms and (b) a smooth term of age	166
8.2	Parameter estimates, standard errors and p-values for an ordinary logistic GLM	167
8.3	Analysis of the AICs for the GAM and the GLM	167
8.4	Analysis of the deviances for the GAM and the GLM	168

8.5 Odds ratios and their corresponding 95% confidence intervals for the parametric effects	170
--	-----

Chapter 1

Introduction

The focus of this study is on sound methods for estimating disease prevalence as a measure of disease burden. In particular we consider the estimation and modelling of HIV prevalence in Zimbabwe using novel frequentist and Bayesian methods.

1.1 Background

Zimbabwe ranks among the countries that have been worst affected by the HIV and AIDS epidemic in sub-Saharan Africa. Although studies, see for example Humphrey et al. (2010) have shown that HIV prevalence has been falling since the late 90s, a recent estimate from the National HIV estimates of 2010, of approximately 15% among the country's sexually active population (15 years and above) is still considered to be too high. The factors that have contributed to the observed decline in HIV prevalence since the late 90s include robust prevention programs, significant behaviour change that has reduced new infections and the successful implementation of the Prevention of Mother to Child Transmission (PMTCT) program, see for example Humphrey et al. (2010), Gonese et al. (2010) and Gregson et al. (2006). Recent studies have also revealed that the relatively high literacy level among the country's population has allowed rapid and effective dissemination of information on HIV and AIDS awareness and appropriate

preventive measures. These have in turn resulted in increased and consistent use of condoms, reduction in casual sex, reduction in extramarital partners and in commercial sex.

According to the United States' Center for Disease Control (CDC), HIV prevalence is defined as the percentage of the population of people with HIV infection who are alive at a given point in time regardless of whether they have or they have not progressed to AIDS. The center also defines the HIV prevalence rate as the number of people living with the infection at a given time per a hundred thousand population. Although it is argued that prevalence does not indicate how long a person has had the virus, it can be used to estimate the probability that a person selected at random from a population has the virus. In the studies of HIV/AIDS, a good understanding of HIV prevalence is essential for monitoring the epidemic and for assessing and evaluating the effectiveness of prevention programs. In addition, HIV prevalence is used as a measure of disease burden and for monitoring global mortality, see Mathers & Loncar (2006).

HIV prevalence has had a fairly significant amount of attention in previous studies in understanding the HIV epidemic. In a Zimbabwean context, Mahomva et al. (2006) studied HIV prevalence among pregnant women attending Antenatal Clinic in Zimbabwe and reported that the HIV prevalence rate declined substantially from 32.1% in 2000 to 23.9% in 2004. The 2009 National Survey of HIV and Syphilis Prevalence Among Women Attending Antenatal Clinic reported a decline in HIV prevalence from 25.7% in 2002 to 16.1% in 2009. At a regional level, Freeman & Glynn (2004) researched on how other sexually transmitted infections affect HIV concordance and/or discordance in married couples in four African countries. Other notable studies of HIV at the sub-Saharan African context include for instance Hallett et al. (2006), Szwarcwald et al. (2008), Ramjee & Eleanor (2002), Bwayo et al. (1991) and Welz et al. (2007). Susser et al. (1993) studied HIV prevalence in Psychiatric Patients in a New York City Men's shelter and reported on the importance of identifying and responding

to the spread of the epidemic in this population.

Estimation of HIV prevalence has been based mainly on data obtained from studying a specific subset of the population. Specifically the estimates have been derived mainly from sentinel surveillance systems that monitored HIV rates among pregnant women and high-risk populations such as drug users, truck drivers, men who have sex with other men and commercial sex workers. Mahomva et al. (2006), Hallett et al. (2006) and Szwarcwald et al. (2008) estimated HIV prevalence using data obtained from surveys of pregnant women attending antenatal clinics. Scott & Holmberg (1996) used data from injection drug users and men who have sex with other men, Ramjee & Eleanor (2002) used data obtained from a survey of truck drivers visiting sex workers in KwaZulu-Natal, Bwayo et al. (1991) also used data on knowledge and attitudes pertaining to sexually transmitted diseases which cause body wasting and death in Mombasa. Lyerla et al. (2006) provided estimates of HIV prevalence by studying populations which are most exposed to HIV in countries with low and concentrated epidemics, Welz et al. (2007) estimated HIV prevalence through a population based longitudinal study in rural KwaZulu-Natal and Fabiani et al. (2003) obtained HIV prevalence estimates by combining sero-survey data and hospital discharge records. Buseh et al. (2002) studied the influence of knowledge of and perceived seriousness of HIV/AIDS, and cultural and gender norms on the spread of the virus in Swaziland using focus groups. These studies were mainly based on data obtained from a subset of the population making inference about the entire population less accurate.

As at July 2011, Zimbabwe's population was estimated at 12 084 304 with approximately 41.9% being below the age of 15 years, 54.3% being in the age group 15 - 64 years and only 3.8% were 65 years and above. The urban dwellers constitute about 38% of the population. Approximately 98% of the population are people of an African origin (of whom 82% are Shona, 14% are Ndebele and 2% are other minority groups), 1% are of Asian and mixed origins and whites of European origin constitute less than

1%. The Zimbabwean population is relatively well educated with a literacy rate (those 15 years and above who can read and write English) of 90%.

Zimbabwe has a wealth of literature on HIV/AIDS surveillance based on a host of a variety of data sources ranging from Antenatal Clinic (ANC) surveys ran since 1990, behavioural data that are available from the 2001/2002 Young Adult Survey (YAS), the 2001 and 2003 Population Services International (PSI) Youth Survey, the Zimbabwe Demographic Health Surveys (1988, 1994, 1999 and 2005/6 with the latest one being 2010-2011), the World Health Organization (WHO) and the UNAIDS.

Zimbabwe's HIV epidemic, similar to other sub-Saharan African countries, is mainly driven by heterosexual transmission. This has mainly been linked to networks of multiple sexual relations, including concurrent relations in which the virus is passed on rapidly. Studies have shown that sexual relations outside marriage have been on the rise for most men in urban areas of Zimbabwe, Gregson et al. (2002). The socio-economic and socio-political situations in the country brought about by the recent hyper inflationary environment have also exacerbated the spread of the virus through forced family separation, commercial sexual exploitation and trafficking of women. High unemployment rates have witnessed an increase in the number of commercial sex workers in recent years, a pervasive problem that is argued to originate from poverty, Mbirimtengerenji (2007). It is argued that poverty has been a key driver of the spread of HIV in many sub-Saharan African countries. In particular, it (poverty) has been linked to deprivation, constrained choices, unfulfilled capacity and interrelated features of well-being that impact upon the standard of living and the quality of life, thereby forcing people to indulge in risky behaviours such as commercial sex to bring basic survival resources Mbirimtengerenji (2007). Despite aggressive awareness programs, studies still show that commercial sex has contributed significantly to new cases and the spread of the virus, see Makondo & Makondo (2014).

Studies have also established that a leading cause of HIV infections in new-born

babies is the MTCT, Mahomva et al. (2006), Coutsooudis et al. (1999), Wiegert et al. (2014) and Bertozzi et al. (2006). The 2001 National HIV Sentinel Surveillance Survey estimated an antenatal nationwide HIV sero-prevalence of 29.5% acquire HIV from their mothers annually in Zimbabwe (Perez et al. (2004)). Maheswaran & Bland (2009) reported that MTCT before, during and after delivery may result in the acquisition of HIV for 30-35% of infants of HIV-infected mothers. During breastfeeding, HIV transmission is likely associated with an elevated viral load in the breast milk, see Bertozzi et al. (2006) and Humphrey et al. (2010). Most HIV interventions target transmissions outside long-term partnerships, however studies point to rapid transmission of the virus in married couples, see for instance Freeman & Glynn (2004) and Mastro & de Vincenzi (2006). In Freeman & Glynn (2004), the study established that most cohabiting couples share related HIV strains and that transmission probability is enhanced by lack of such factors as male circumcision, lack of condom use, vaginal intercourse during menstruation and presence of other sexually transmitted infections.

The current study aims to develop techniques that can be used to provide national HIV prevalence estimates and estimates in subgroups (categorized geographically, socio-economically and demographically). Essentially the techniques account for the hierarchical data structure inherent in populations, for the complex sampling design and for the variability due to missing data. The proceeding sections of the current chapter outlines the fundamental statistical methods relevant for analyzing survey data and specific objectives of the study.

1.2 Sampling under complex surveys

Sampling is a statistical technique that involves random selection of a subset of a population to be used in analysis as a representative of the population. To ensure randomness in the selection, researchers usually employ the concepts of probability resulting

in probability samples or random samples. According to Chambers & Skinner (2003), Heeringa et al. (2010) and Lee & Forthofer (2006), selection of samples in practical surveys rarely involve simple random sampling (SRS), instead complex sampling procedures that include stratification and multistage selection of elements with stochastic assumptions involved in the formulation of the sampling schemes are usually utilized. The use of complex sampling schemes is aimed at improving the representativeness of the sample and capture the prominent features of the underlying population, see for example Heeringa et al. (2010), Fuller (1975), Chambers & Skinner (2003) and Lehtonen & Pahkinen (2004), to optimize the variance or cost ratio of the final design or to meet precision targets for subgroups of the target population. Some populations under study may reflect complex underlying structures with observations from different individuals dependent on each other and observations within clusters being correlated. In addition, the sampling designs are characterized by unequal selection probabilities of units, double sampling, and multiple frames, and estimation features such as imputations, adjustments and compensation for non-response and under-coverage. Survey data arising from complex sampling schemes or reflecting associated underlying complex population structures are referred to as complex survey data.

Although not primarily concerned with the scientific, medical or socio-economic interpretation of the facts, (despite being required to supply material adequate for such interpretation), sample surveys are concerned with the accurate ascertainment of the individual facts and observations from elements recorded and also with their compilation and summarization. The presence of covariates or auxiliary variables also has a bearing in the way the survey data are analyzed. It is common to meet with survey data which are disproportionate to the target population and this often necessitates some form of weighting to be employed. In this regard sampling weights are often used to reflect the unequal sample inclusion probabilities and compensate for differential non-response and frame under-coverage as explained by Pfeffermann (1996).

It is argued that the purpose of sampling theory is to make sampling more efficient. According to Cochran (1977), Hansen et al. (1953), Kish (1965) and Kish & Frankel (1974) sampling theory attempts to develop methods of sample selection and of estimation that provide, at the lowest possible cost, estimates that are ‘precise’.

Conventional analysis of survey data often ignore the complex sampling schemes with the assumption that all sample observations were independently selected with equal probabilities. However as by Lumley (2010), Skinner et al. (1989), Heeringa et al. (2010) and Lee & Forthofer (2006) the assumption rarely holds in practice, often leading to unbiased estimates and inaccurate conclusions.

In a typical complex sampling design, clustering, defined as the natural grouping of population elements that are relatively homogeneous, is mainly significant for reducing survey costs or for simplifying the logistics of the actual survey data collection, Lehtonen & Pahkinen (2004) and Heeringa et al. (2010). However it is worth noting that sampling schemes that incorporate clustering often result in standard errors, for survey estimates, that are generally greater than those from an SRS of equal size. Furthermore, special approaches to variance estimation are required whenever clustering is involved as sample units from the same cluster generally tend to be correlated see Binder (1983), Rust (1985) and Rust & Rao (1996). A common statistical measure of the homogeneity of observations within sample clusters is the intra-class correlation (ICC) denoted ρ , see Kish (1965), Heeringa et al. (2010) and Lehtonen & Pahkinen (2004).

The contribution of stratification, defined as the division of population elements into mutually exclusive and exhaustive non-overlapping subgroups, in a complex design is towards improving on statistical efficiency in estimation and inference. Stratification often gives smaller standard errors for sample estimates relative to SRS. Specifically as noted by Heeringa et al. (2010), since stratified sampling selects independent random samples from each of the $h = 1, \dots, H$ strata of relative size $W_h = N_h/N$, any variance attributed to differences among strata is eliminated from the sampling variance of the

estimate. Thus the ideal for any stratification designed to increase sample precision is to form strata that are as “homogeneous within” as possible and as “heterogeneous between” as possible Heeringa et al. (2010).

In complex surveys that involve both stratification and clustering, the effect of the design falls between the high value for the clustered only design and the low value for the scenario where only stratification effects are considered. Thus complex surveys result in what is termed a “tug of war” between the variance inflation due to clustering and variance reduction due to stratification.

1.2.1 Sampling weights

Complex sampling schemes that involve varying sample inclusion probabilities for individual observations often employ sampling weights to “map” the sample back to an unbiased representation of the target population. Complex survey data are usually characterized by disproportionate representation of the observations in the population brought about to reflect the prominent underlying population structure in the sample. As a result sampling weights are essential in survey data analysis for adjusting for the differential representation of sample observations. Thus essentially the key purpose of sampling weights is to make the distribution of the variables in the sample data approximate the distribution of the variables in the target population, see for example Pfeffermann (1993), Heeringa et al. (2010) and Winship & Radbill (1994).

Formally, let π_i , for $i = 1, \dots, N$ denote the probability that unit i in the population is included in the sample, the sampling weight for any sampling design is defined as the reciprocal of π_i given by

$$w_i = \frac{1}{\pi_i}$$

In particular, for unit-specific weights in survey data analysis, weights reflect the number of population elements that is represented by the respective sample observation.

Following Lee & Forthofer (2006), two types of sampling weights are often encountered in practical complex surveys. These are the expansion weight, that is the reciprocal of the selection probability, and the relative weight, that is obtained by scaling down the expansion weight to reflect the sample size. The theory behind these two types of sample weights is quite varied in relation to different sampling designs, see for example Pfeffermann (1996), Pfeffermann (1993), Winship & Radbill (1994) and Lepkowski et al. (2006).

Ignoring sampling weights in the analysis of complex survey data often leads to biased estimates and model mis-specification (Pfeffermann (1993), Pfeffermann (1996), Lee & Forthofer (2006) and Rust (1985)). In addition, it underestimates the variance thereby resulting in incorrect standard errors of estimates and Type I errors, Winship & Radbill (1994).

1.2.2 The design effect (deff)

Most practical sampling designs, for instance those used in national demographic and health surveys, are rarely SRS; rather they involve stratification, clustering and disproportionate sampling of population elements. The design effect for a particular complex sampling design as defined by Gabler et al. (2006), Heeringa et al. (2010) and Kish (1965) is the net effect of of the combined influences of stratification, clustering and weighting relative to a SRS design. Often, the design effect is expressed as a ratio of the standard error (or variance) of an estimate obtained using a complex scheme compared to one obtained using a SRS scheme. For instance, for a population quantity θ , say, estimated by $\hat{\theta}$ obtained from a particular design, relative to the SRS of equal size, the net effect of a complex design on the standard error of $\hat{\theta}$ is given by

$$D^2(\hat{\theta}) = \frac{\left[\text{SE}(\hat{\theta})_{\text{complex}} \right]^2}{\left[\text{SE}(\hat{\theta})_{\text{srs}} \right]^2} = \frac{\text{Var}(\hat{\theta})_{\text{complex}}}{\text{Var}(\hat{\theta})_{\text{srs}}}$$

where

$D^2(\hat{\theta})$ = the design effect for $\hat{\theta}$;

$SE(\hat{\theta})_{complex}$ = the complex sample design standard error of $\hat{\theta}$;

$SE(\hat{\theta})_{srs}$ = the simple random sample standard error of $\hat{\theta}$;

$Var(\hat{\theta})_{complex}$ = the complex sample design variance of $\hat{\theta}$;

$Var(\hat{\theta})_{srs}$ = the simple random sample variance of $\hat{\theta}$.

The design effects are used mainly in computing confidence intervals, desired sample sizes and test statistics that incorporate the estimates of standard errors corrected for the complex sample design, Heeringa et al. (2010). Furthermore, the design effect is used to measure the relative loss (or gain) in precision achieved by using a given complex sampling design compared to an SRS, see Wolter (1985) and Heeringa et al. (2010). Analytic statistics such as the Rao-Scott Pearson χ^2 and the likelihood ratio χ^2 tests, as described by Rao & Scott (1981), Holt et al. (1980) and Bedrick (1983) also rely on the design effect to reflect the effect of the complex sampling design. The current research utilizes the design effect in all instances where the above-mentioned statistics and/or tests are computed.

1.2.3 Variance estimation under complex sampling

A basic requirement when analyzing complex survey data and a good survey practice is that a measure of precision be provided for each estimate derived from the survey data, see Binder (1983) and Wolter (1985). It is important to note that the variance of an estimate is usually unknown, thus it (the variance) is also estimated from the survey data. Hence the estimated variance is a function of both the form of the parameter estimate and the nature of the sampling design, see Berger & Skinner (2004), Rust (1985), Rust & Rao (1996), Lee & Forthofer (2006) and Wolter (1985). Parameter estimates obtained from complex survey data are often nonlinear and complex, and their complexity is induced by the sampling design used to obtain the survey data, Rust

& Rao (1996), Heeringa et al. (2010) and Wolter (1985). There are several methods for estimating the variance and selecting the most appropriate is often based on accuracy of the variance estimator, time constraints, cost, simplicity and other administrative considerations.

Often with descriptive analysis of survey data, variance estimates are used for constructing confidence intervals for the population parameters, thus a second criterion that considers the best confidence interval is normally employed. Some analyses of survey data may specify particular statistical methods for the data analyses, by which preference is given to the variance estimator that has the best statistical properties for the proposed analysis. Availability of software packages that are capable of computing the appropriate variance estimates is also of importance. In practical surveys that are multipurpose, such as national household surveys, there may be many variables and statistics of interest each requiring an estimate of its own respective variance. Therefore it may be necessary to use one, or at least a few variance estimating methods where a compromise must be made to arrive at a variance estimator that might not be optimal for any single statistic, but, as Wolter (1985) suggests, one that involves a tolerable loss of accuracy for all, or at least the most important statistic. Common approaches to variance estimation for sample estimates obtained from complex survey data are based on the Taylor series linearization and the re-sampling procedures such as the Jackknife repeated replication (JRR) and the Balanced repeated replication (BRR) methods, see Wolter (1985).

1.3 Approaches to statistical analysis of survey data

The current study considers two approaches to analyzing survey data; the frequentist (or the classical) and the Bayesian. The frequentist approach, based on the fundamental ideas envisaged by Neyman (1934), is further subdivided into design-based and the

model-based. The difference between the design-based and the model-based approaches lie in the sources of randomness that is responsible for giving the stochastic structure in the data as explained by Gregoire (1998) and Sarndal et al. (1978). The Bayesian paradigm has its basis in incorporating prior information about the model parameters in the analysis using the Bayes theorem relying on the subjective definition of probability.

1.3.1 Design-based approach

The design-based approach, as given by Chambers & Skinner (2003) and Heeringa et al. (2010), that was formalized by Neyman (1934), is a statistical framework in which the only source of random variation is that induced by the sampling mechanism, that is, the complex sampling design. Sarndal et al. (1978) described the source of randomness in design-based approach as the probability ascribed by the sampling design to the various subsets of the finite population. Under a design-based approach, the population whose data values are unknown but are regarded as fixed is specified and the observed sample is random due to the random selection, Lumley (2004). Hence stochasticity is introduced at the sampling stage.

Statistical inference under the design-based approaches rely on the sampling distribution of repeated samples generated by the sampling design. The analysis of the data incorporates the design features as well as the sample weights that are designed to reflect the design features, unequal probabilities of selection, non-response and post-stratification, (Shao & Chen (1998), Lehtonen & Pahkinen (2004) and Lee & Forthofer (2006)). Statistical inference is often then based on the normal approximation justified by large-sample arguments. Under complex sampling, classical statistical methods that assume independence of units become inappropriate, necessitating methods of analysis that take account of the sampling scheme as explained in Rao & Scott (1981), Rao & Thomas (2003), Lohr (2010) and Lee & Forthofer (2006).

1.3.2 Model-based and model assisted approaches

The model-based approach makes use of a probability distribution for the random variables of interest as described by Chambers & Skinner (2003) and Heeringa et al. (2010). Under this approach, the only source of variation is that induced by the model that is presumed to have generated the population values. Inference under the model-based approach considers the values of the population as a realized outcome of a random variable obtained from a super-population model, Sarndal et al. (1992) and Gregoire (1998). The sample is held fixed, even if it is generated by a probability sampling design. This approach ignores the probability distribution induced by the sampling mechanism. Thus the model generating the values is taken to be a mathematical abstraction used to describe reality, Brus & de Gruijter (1997). Tests of hypotheses and interval estimation are carried out via the maximum likelihood. Consequently the probability distribution of all possible realizations of the outcome is the basic tool for inference.

Unlike under the design-based approach where models are used to describe the population characteristics, under the model-based approach, models are used to describe the data generating process. Discussions around the fundamental differences between the design-based and the model-based approaches are well documented, see for example Sarndal et al. (1978), Hansen et al. (1983), Smith (1976) and Gregoire (1998).

In an analytic approach, there are many models that are used to analyze complex survey data. These include general regression models and generalized linear models (GLM) which are suitable for both normal and non-normal as well as counts data. For a binary response in particular, logistic regression models are often utilized, see Agresti (2007), Hosmer & Lemeshow (2000), Chambers & Skinner (2003) and Heeringa et al. (2010). On the other hand, the analysis may need to take account of different sources of variability often encountered in multi-layered clustered data. Models that are designed, often using a GLM framework, for this are generally called hierarchical or multilevel models as described by Goldstein (1991), Goldstein (2011), Goldstein & McDonald

(1988) and Snijders & Bosker (1999). Various extensions of GLMs are available that include random effects, giving rise to generalized linear mixed effects models (GLMM) as explained by McCulloch & Searle (2001).

However frequentist modelling approaches have a potential limitation in that the uncertainty in predicting parameters is not reflected in prediction inferences. In addition, evaluating high-dimensional numerical integrals that are commonly encountered under the frequentist approach is often not possible. Bayes models that propagate uncertainty about parameters are preferable especially in small samples.

1.3.3 Bayesian approach

Most practical scientific investigations, such as laboratory tests, are carried out as controlled learning process whereby various aspects of a research problem are iteratively illuminated as the study progresses. The process may involve tentative conjecture suggesting an experiment with appropriate analysis of the data leading to a modified conjecture which in turn leads to a new experiment, and so on. Quick and unambiguous convergence of the process is indicative of an efficient investigation. A Bayesian approach allows combining prior knowledge (that come as the conjecture or prior belief about model parameters) with the likelihood of the observed data in a scientific investigation to obtain a posterior distribution.

We consider a Bayesian approach to finite population inference as described by Little (2003), where the population values are assigned a prior distribution and then the posterior distribution is computed using the Bayes' theorem. Good practical illustrations where a Bayesian approach is used are given by Sedransk (2008). Application of the Bayesian analysis is often encountered in studies involving small area estimation, see for example Rao (2003) and Jiang & Lahiri (2006). Vail et al. (2001) applied the Bayesian paradigm approach in randomized clinical trials setting.

Unlike the frequentist approach, in Bayesian statistical inference, the uncertainty

about the true value of the unknown parameter is reflected by specifying a probability (the prior) distribution for the parameter. Of importance to note is that, under the Bayesian approach the observed data are regarded as fixed and the parameters are assumed random. The overarching theory underlying the formulation of the Bayesian analysis and inference has received considerable attention, see for example Press (1989), Bolstad (2007), Bernardo & Smith (1994), Dempster (1968) and Raiffa & Schlaifer (1961).

Key to proper formulation of a Bayesian analysis is the ability to quantify prior knowledge (or ignorance) in a statistical specification. Gelman et al. (2008) gave a detailed discussion on the formulation of prior distributions under Bayesian regression models. The underlying fundamental reasoning behind specifying prior distributions stems from the view that science generally comes about by learning and incorporating findings and information from previous research to inform the current studies. Often, use is made of non-informative priors in many practical applications in which the observed data are allowed to dominate the analysis and “speak for themselves”, see for example Box & Tiao (1973) and Sweeting (1981). Bedrick (1983) outlined formulation of informative prior distributions for GLMs through the use of conditional means priors.

There has been a renaissance in the development and application of Bayesian statistical methods, owing mostly to developments of powerful statistical software tools that render the specification and estimation of complex models feasible from a Bayesian perspective. In many practical applications, especially where high-dimensional parameters are involved, iterative simulation values are drawn from the posterior distribution and then inferences are based on these draws. Notable examples include Doucet et al. (2000), Gamerman (1997) and Lesaffre & Lawson (2012). In many practical research studies involving surveys, non-informative priors that reflect absence of strong prior information are often preferred. Consequently, the Bayesian approach often yields similar estimates and estimates of precision to the frequentist approach however the confidence

intervals have a different meaning.

1.4 Missing data in surveys

Most practical surveys are characterized by non-response resulting in missing data, for instance the national Demographic and Health Surveys. There are a variety of reasons why data are missing in survey research, see for example Baraldi & Enders (2010), Little & Rubin (1987a) and Schafer & Olsen (1998). Missing data in surveys may be due to what is termed unit non-response. This occurs when no survey data are collected for an individual selected for the sample. In addition, missing data may be a result of refusal to participate in a survey or non-contact or language barrier. Missing data may also occur when a sampled individual participates in the survey but fails to provide acceptable responses to one or more of the survey items. Ignoring missing data in statistical analyses results in biased estimates and incorrect statistical inference. Missing data are argued to be ubiquitous and pervasive in scientific and social research see Schafer (1997), Schafer & Olsen (1998), Schafer (1999), Rubin (1976) and Rubin (1987). A substantial effort has been directed towards finding methods of handling missing in surveys, for instance Pigott (2001), Schafer & Olsen (1998), Schafer (1999), Kalton & Brick (1996), Baraldi & Enders (2010) and Rubin (1987). Available methods range from ad hoc and traditional deletion and single imputation methods (such as mean and regression imputations) to more advanced and “state of the art” methods such as the likelihood and the multiple imputation (MI). Most of the advances in these methods have been enhanced by the advent of fast, powerful and sophisticated statistical packages.

In the analysis of data with missing observations, Rubin (1976) and his colleagues Little & Rubin (1987b) classified missing data according to what they termed “missing data mechanisms”. The classification is based on the relationship between the mea-

sured variables and the probability of missing data. Data are classified as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Data are MCAR if the probability of missing data on a given variable is unrelated to other measured variables and to the values of the variable itself. That is, missingness is completely unsystematic and the observed data can be regarded as a random sub-sample of the hypothetically complete data, Baraldi & Enders (2010). Under the MAR mechanism, missingness is related to other measured variables of the incomplete variable. On the other hand missing data that are systematically related to the hypothetical values that are missing are classified under the MNAR mechanism. Appropriate methods for handling missing data are often related to the respective missing data mechanism. Detailed discussion of the missing data and methods for handling missing data are presented in Chapter 4.

1.5 Statistical computing

All the analyses for the current research were done using packages available in **R** version **3.1.3**, by **R** Team (2013). In particular, the package **survey** by Lumley (2010) was used to compute the design-consistent crude estimates and the survey logistic regression models for HIV prevalence. The package allows specification of the survey design, the sampling weights and the appropriate variance estimation method. In addition the package supports the design-induced distortion of the asymptotic distribution of the Pearson and the likelihood ratio statistic that comes with the Rao-Scott test by Rao & Scott (1981). The Hosmer-Lemeshow test by Hosmer & Lemeshow (1980) that accounts for survey design and for the model goodness of fit for the logistic regression model was performed using the package **ResourceSelection** by Lele et al. (2015). The automated model building procedures for the survey logistic regression were done using the packages **glmulti** by Calcagno & de Mazancourt (2010) and **stepAIC** by

Park & Hastie (2010). The packages generate all possible models under constraints set by the analyst with specified response and explanatory variables and finding the best model according to some criteria such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Multiple imputations were done using the package **mi** by Gelman et al. (2015). The package uses chained equations approach to imputation and allows specification of conditional distribution of each variable with missing values conditioned on other variables in the data in a Bayesian framework. The procedure is an iterative algorithm that sequentially iterates through the variables to impute the missing values using the specified models. The analysis model, obtained by pooling estimates from the m ‘complete’ data sets, was computed using the **survey** package.

The hierarchical logistic regression models were computed using the package **lme4** by Bates et al. (2014). The package is designed to build models in a GLMM framework in which fixed and random effects can be specified explicitly. The Bayesian logistic regression was computed using the package **arm** by Gelman et al. (2010). Simulation draws from the posterior distributions were carried out using the **MCMCpack** by Martin et al. (2013). Assessment of the convergence of the Markov chain Monte Carlo (MCMC) was done via checking the convergence diagnostics using the package **mcmcplots** by Curtis et al. (2015). The generalized additive models (GAMs) were computed using the **mgcv** package by Wood (2006). Essentially the package uses a local scoring algorithm to iteratively fit the models in GAM framework via the back-fitting procedure. The specific functions that were used in the various approaches were given in the respective chapters.

1.6 Objectives of the study

The objectives of the research are to:

- develop an insight into the dynamics of HIV infection and prevalence in different categories of the population of Zimbabwe;
- utilize the theory of the analysis of survey data to estimate HIV prevalence at both the national and subgroup level of the population of Zimbabwe;
- identify risk factors associated with the HIV prevalence in the respective subgroups;
- explain the variation in HIV (using a modelling approach) using demographic and socio-economic factors of the population both at the individual and household level taking the clustering and hierarchical structure of the data via both a frequentist and a Bayesian approach;
- account for missing data often encountered in practical survey data due to non-response using sound techniques that account for the uncertainty due to the missing data themselves, the method of handling the missing data as well as the structure of the underlying population.

1.7 Significance of the study

Statistical methods have been widely used to assess the trends, patterns and the dynamics of HIV/AIDS epidemic. The focus of these methods range, depending on the objectives embedded in the research, from estimating prevalence, understanding the pathogenesis of HIV infection, assessing the potency of antiviral therapies, evaluation of treatment efficacy of viral dynamic models and assessing disease burden. Statistical models that estimate the magnitude and trajectory of the HIV/AIDS epidemic have also been constructed. These models have been used as tools to extract and convey as much information as possible from available data and provide accurate representation of both the knowledge (understanding origin and progression) and uncertainty about

the epidemic. The success of the statistical methods rely on the quality of the data available. In addition, the quality and reliability of the estimates calculated from the available data depend on the validity of the statistical analysis performed. Sampling and survey techniques have been used as bases for accurate data collection and provide sufficient theory behind consistent analysis of phenomena such as HIV epidemic. A key responsibility of the discipline of statistics is to provide science with theory and methods for objective evaluation of data-based evidence and measuring the strength of that evidence, Royall (2003).

It is under this background that a study such as the current one is used to inform researchers and policy makers regarding understanding, the progression of HIV, assess the intervention programs and develop epidemiological projections. In particular, a good understanding of prevalence of HIV can allow proper implementation of prevention and intervention programs such as implementation of antiretroviral treatment. It is argued that a good understanding of HIV infection and prevalence is critical, not only to develop appropriate surveillance system instruments, but also to understand the epidemic and implement appropriate intervention programmes, Bertozzi et al. (2006). Intervention programs can be designed to target the mode of transmission. The use of population-based survey data also allows researchers to focus attention to subgroups or domains within the population that are most at risk or exposed to HIV as concentrated epidemics. In addition, population-based estimates can also be combined with ANC estimates (as proxies) to give more accurate estimates of HIV prevalence. Cross-sectional studies of HIV prevalence (such as the current one) can complement mathematical models that track transmission dynamics in generating HIV prevalence projections. Studies have shown that in one way or the other every individual living in this day is either infected or affected by the HIV/AIDS pandemic. Thus a good understanding of HIV/AIDS can enhance provision of support structure for both the infected and the affected. To achieve the above desired objectives, novel methods of

estimating, enhancing and explaining variation in HIV prevalence are explored.

Most survey data, see for example Brick & Kalton (1996), Pigott (2001) and Raghunathan (2010) are characterized by missing data, especially in research involving HIV which is still regarded a sensitive issue in most of sub-Saharan African countries. Respondents are not at liberty to disclose their HIV status nor to consent for a test especially if they suspect or they know that they are positive. Hence the use of proper techniques of handling missing data that are supported by sound statistical theory can enhance accurate estimation of HIV prevalence.

1.8 Thesis layout

The thesis is organized into chapters in which each chapter is stand-alone motivated by the respective methods or approaches to estimating HIV. In particular, each approach attempts to explore and capture ways in which the variation in HIV prevalence can be explained, taking the design features, nesting or clustering data structures and non-response into account. Chapter 2 gives a detailed explanation of the data; the source, the variables and the sampling design. In addition, some basic exploratory analyses to explore elementary relationships between the variables are presented. Chapter 3 presents the details of how a design-consistent survey logistic regression model was computed to explain variation of HIV prevalence from a GLM perspective. In addition, design-consistent estimates of HIV prevalence are given at the national and domain (subgroup) level. Practical surveys are often characterized by non-response that results in missing data. Chapter 4 outlines the details of the nature of missing data in surveys and presents a multiple imputation technique used to ‘fill in’ missing data and at the same time account for the variability induced by the missing data themselves, the imputation process as well as the underlying structure of the population. Discussion of results obtained from a complete case analysis and a multiple imputations-based

analysis is presented.

Chapter 5 focuses on multilevel or hierarchical models built from a GLMM perspective. The multiple sources of variability resulting from the multi-layering of the data, and the clustering nature of the data brought about by the prominent features of the underlying population are captured. In addition, measures of within and between-grouping variability are provided. Chapter 6. provides details of a logistic regression model for HIV prevalence fitted from a Bayesian analysis framework. A posterior distribution as an empirical distribution is computed, and for inference purposes sampling from the posterior is done using the MCMC technique. Convergence diagnostics for the MCMCs were presented in the form of plots and summary statistics. A Bayesian hierarchical logistic regression model for estimating HIV prevalence using the GLMM framework is given in Chapter 7. A general additive model for HIV is provided using semi-parametric splines approach in Chapter 8. General concluding remarks and direction for further research are given in Chapter 9. In addition, the strengths and potential limitations of the study are also given in Chapter 9.

Chapter 2

Exploratory data analysis

2.1 Introduction

Most studies involving estimation of HIV prevalence have been based mainly on data obtained from studying a specific subset of the population. Specifically the estimates have been derived mainly from sentinel surveillance systems that monitor HIV rates (often) among pregnant women attending ANCs and high-risk populations such as drug users, truck drivers, men who have sex with other men and commercial sex workers. We present the data used for the analysis for the current research. In particular, the variables and their respective significance to HIV research are explicitly detailed.

2.2 The data

The data used for the study were obtained from the 2010-11 Zimbabwe Demographic and Health Surveys (2010-11ZDHS). The 2010-11ZDHS is one of a series of interview-based household surveys undertaken by the Zimbabwe National Statistics Agency (ZIMSTAT) under the auspices of the Zimbabwe National Household Survey Capability Programme (ZNHSCP) and the worldwide DHS programme, see Mutasa (2012). Similar surveys were also carried out previously in the years 1988, 1994, 1999, and in 2005-06.

Unlike the previous surveys, the 2010-11ZDHS includes a section on HIV/AIDS. In addition, the interviews (for the 2010-11ZDHS) were carried out using electronic personal digital assistants (PDAs) rather than the paper questionnaires for recording the responses used in previous surveys. The PDA data collection system that was developed by the DHS project is equipped with blue-tooth technology to enable remote electronic transfer of files. The DHSs in general are country-level population-based household surveys. The data obtained from these surveys are mainly aimed at providing current information for monitoring and impact evaluation of key indicators pertaining to population (socio-economic and demographic), health and nutrition.

Specifically, for the 2010-11ZDHS females aged 15 to 49 and males aged 15 to 54 were eligible for interview and collection of blood samples or specimens, using dried blood spot (DBS), for laboratory testing (which includes HIV testing). The collection of data for the 2010-11ZDHS's main interview section was done with the use of three questionnaires; the household, the women's and the men's questionnaires. The questionnaires were adopted from the DHS project and were aimed at reflecting country-specific population and health issues. The household questionnaire provides general household information and was also used to identify household members who were eligible for interview and for collection of blood samples. The women's questionnaire was used to collect information pertaining to women (fertility, marriage and family planning) whereas the men's questionnaire was designed to collect information regarding the men.

For HIV testing, five blood samples were collected on a special filter paper card using capillary blood from a finger prick. An "anonymized" antibody testing process was conducted at the National Microbiology Reference Laboratory (NMRL) in Harare. Bar coded labels were used to identify the DBS samples to ensure the anonymity and these were used to track the outcome of the testing procedure and the results. Laboratory testing of the blood specimens followed a standard laboratory algorithm designed to

maximize the sensitivity and specificity of the test results. In particular, the algorithm uses two different HIV antibody enzyme-linked immunosorbent assays (ELISAs) that are based on antigens. Discordant samples that were positive in the first test were retested using both ELISAs and discordant samples from the second round of testing were regarded as “indeterminate”. The “indeterminate” were then subjected to a western blot confirmatory test, in which the results are considered final. Written consent was sought from the respondents before the collection of the blood samples, and for the 15–17 year old (still minors) respondents further consent was also sought from their parents or responsible adult. Furthermore, consent was sought to store blood samples for future research. All participants were given information brochures pertaining to HIV/AIDS and giving details regarding the nearest facility providing voluntary counseling and testing (VCT). All HIV testing procedures were reviewed and approved by the ethical review boards of ORC Macro, a US-based company that provides technical assistance to DHS worldwide, the Centers for Disease Control (CDC) and the Medical Research Council of Zimbabwe (MRCZ).

For the current research the response variable is HIV status, which is a binary variable since an individual can either be HIV positive or negative. The socio-economic, demographic and behavioural factors (that were used as the predictors) were selected as those factors thought to influence HIV infection as informed by the proximate determinants conceptual framework by Boerma et al. (2003). The 2010 -11ZDHS sample was designed to yield a representative sample for the country as a whole, for urban and rural areas and for each of the ten provinces (Manicaland, Mashonaland East, Mashonaland West, Mashonaland Central, Midlands, Harare, Masvingo, Matabeleland South, Matabeleland North and Bulawayo). The same data are used in subsequent chapters and to avoid repetition of the data description reference is made of the current section whenever the data are described.

2.3 The 2010-11ZDHS data

Under the 2010-11ZDHS, a stratified two-stage cluster sampling design was used to collect the data using the 2002 population census figures as the sampling frame. Individuals were clustered within households which in turn were clustered within enumeration areas (EAs) and the country's ten administrative provinces were regarded as the strata. For stratification, provinces were split into rural or urban and used as stratifying variables, whereas for the multistage clustering, the primary sampling units (PSUs) were the EAs and the secondary sampling units (SSUs) were the households. The sampling design has 18 strata, 406 PSUs (169 in urban and 237 in rural areas) and SSUs of unequal size and of different number per PSU. The PSUs per stratum and SSUs within each PSU were selected using simple random sampling. All individuals aged 15 – 49 for females and 15 – 54 for males who were permanent residents of the selected households or who stayed in the household the night before the survey were eligible for interview and voluntary HIV testing.

The overall average household size was 4.1 people. Urban households were slightly smaller averaging 3.8 people as compared to 4.3 people in the rural areas. The sample consists of 17 434 respondents of whom 9 591 are female and 7 843 are males. The response rates were 93% and 86% for females and males respectively.

2.4 The key variables

The 2010-11 ZDHS sample was selected with the use of unequal probability sampling in order to ensure an adequate number of respondents that can allow for analysis in key domains of the population. We present the responses to the measured variables together with their respective frequency and percentage of missing data. Table 2.1 gives the crude HIV test results for the 17 434 individuals that make up the sample. 12 274 (70.40%) of the respondents tested negative, 2 388 (13.70%) of the respondents

tested positive and 2 772 (15.90%) had missing values for at least one of the measured variables. The current study focuses on the investigating the effects of demographic, socio-economic and behavioural factors on HIV. We explore each of the factors that are thought to be related to HIV. The information obtained is useful for understanding how these factors determine attitudes towards general health services and behaviours that may influence HIV infection.

Table 2.1: Crude summary of the HIV test results for the respondents

Test result	Number	Percentage (%)
Positive	2 388	13.70
Negative	12 274	70.40
Missing	2 772	15.90

Majority of the Zimbabwean population stay in the rural areas hence 11 144 (63.92%) of the respondents were drawn from the rural areas whereas 6 290 (36.08%) were drawn from the urban areas. The level of education attained by the respondents is shown in Table 2.2. Education level attained is regarded as an important characteristic of the respondents as it is associated with many factors that have a significant impact on health seeking behaviour and use of HIV preventive services. There were 597 non-respondents for this variable representing approximately 3.42% of the sampled individuals.

Table 2.2: Distribution of the respondents by education level

Level	Number	Percentage (%)
No education	329	1.89
Primary	4 634	26.58
Secondary	10 972	62.93
Higher	902	5.17
Missing	597	3.42

The results in the table show that the majority (approximately 94.96%) of the respondents have received some form of formal education. The population of Zimbabwe is characterized by a considerably high literacy level¹ as evidenced by the data in Table

¹Literacy was measured in terms of ability to read and write. Non-literate: those who cannot read

2.3.

Table 2.3: Summary of the literacy levels of the respondents

Attribute	Number	Percentage (%)
Literate	15 977	91.64
Partially literate	1 263	7.24
Non literate	22	0.13
Missing	172	1.00

Table 2.4 gives the number of respondents in each of the five-year age groups. The results in the table indicate that the proportion of respondents in each age group decreases with increasing age. Approximately 22.38% of the respondents were in the 15 – 19 years age group and approximately 18.71% were from the 20 – 24 years age group. Relatively smaller proportions of respondents were drawn from the older age groups in order to mirror the age structure of the Zimbabwean population (a broad base and a narrow top similar to most developing countries). Non-respondents constituted approximately 1.06% of the sampled individuals with regards to the age variable.

Table 2.4: Summary of the five year age groups for the respondents

Age group	Number	Percentages (%)
15 - 19	3 901	22.38
20 - 24	3 262	18.71
25 - 29	2 988	17.14
30 - 34	2 351	13.49
35 - 39	1 919	11.01
40 - 44	1 366	7.84
45 - 49	1 061	6.09
50 - 54	401	2.30
Missing	185	1.06

Table 2.5 gives the results of the marital statuses of the respondents. The data show that just over half (53.62%) of the respondents were married and those in categories single/never married, divorced and widowed constituted 32.77%, 5.49% and 3.75% of the

nor write; Partially: those that can read or write part of a sentence; Literate: those that can read or write full sentence

respondents. Those with missing data constituted approximately 4.37%. The marital status of an individual is significant in HIV research as it reflects the level of how sexually active the individual is. This is particularly an important factor as sexual contact is argued to be the key driver of HIV infection especially in sub-Saharan Africa.

Table 2.5: Summary of the marital status for the respondents

Marital Status	Number	Percentage (%)
Single	5 713	32.77
Married	9 348	53.62
Divorced	957	5.49
Widowed	654	3.75
Missing	762	4.37

The data show that a large proportion of the population has access to information through reading magazines, listening to the radio and watching television as shown in Table 2.6. The respondents' access to information measured in terms of how often one reads magazines, listen to the radio or watch television (TV) was linked to access to HIV prevention and education material. Of those sampled, approximately 21.51%, 38.3% and 37.22% read a magazine, listen to the radio and watch TV respectively, at least once a week. The percentages for those who read magazines, listen to the radio or watch TV less than once a week are respectively 27.88%, 23.4 and 18.26%. Those who do not have access to information are 27.49% for reading magazine, 29.93% for radio listening and 36.35% for TV watching. The percentages of missing values are 23.12% for magazine reading, 8.37% for listening to the radio and 8.17% for watching TV. Low levels of access to information could mainly be because the majority (63.92%) of the population reside in the rural areas.

Table 2.6: Summary of access to information for the respondents

Attribute	Reading magazine (%)	Listening to radio (%)	Watching TV (%)
≥ 1 a week	3 750 (21.51)	6 677 (38.30)	6 489 (37.22)
< 1 a week	4 861 (27.88)	4 080 (23.40)	3 183 (18.26)
Not at all	4 793 (27.49)	5 218 (29.93)	6 337 (36.35)
Missing	4 030 (23.12)	1 459 (8.37)	1 425 (8.17)

Literacy level (as shown in Table 2.3) and access to information combined were used to ascertain the level of access to mass media health messages. Approximately 45.52% of the respondents were employed whereas approximately 53.22% were unemployed with about 1.26% of the respondents having missing values as shown in Table 2.7.

Table 2.7: Summary of the employment status of the respondents

Employed	Number	Percentage (%)
No	9 279	53.22
Yes	7 936	45.52
Missing	219	1.26

Research has established that socio-economic status of a population is associated with its health status, Mbirimtengerenji (2007), Rutstein & Kiersten (2004) and Meer et al. (2003). For the 2010-11 ZDHS, the wealth index² was used as a measure of the socio-economic status of the population at household level expressed in the five quintiles shown in Table 2.8. Information on household assets was used to create an index regarding household wealth. Specifically the assets used include ownership of consumer goods such as television and vehicles as well as dwelling characteristics such as source of drinking water and sanitation facilities. It is also evident from the data in the table that the respondents are fairly evenly distributed across the categories of the wealth index as shown in Table 2.8.

Table 2.8: Summary of the wealth indices of the respondents

Level	Number	Percentage (%)
Poorest	3 062	17.56
Poorer	2 936	16.84
Middle	3 016	17.30
Richer	3 735	21.42
Richest	3 949	22.65
Missing	737	4.23

²Wealth Index is a composite measure of the household's cumulative living standard based on a household's ownership of selected assets such as televisions, vehicles as well as water access and sanitation facilities. It is generated using the principal component analysis and places households into five wealth quintiles

The percentages range from 16.84% for the poorer to 22.65% for the richest, and approximately 4.23% had missing values. In the DHS, the wealth index is a composite measure of a household's cumulative living standard generated using principal component analysis as described by Rutstein & Kiersten (2004). The wealth index is used to identify problems associated with access to health care services, and health inequalities and increased risk of infection with HIV, as described in Feinstein (1993).

2.5 Basic relationships between the explanatory variables

The basic relationships between the selected risk factors that are regarded as the explanatory variables are considered. In particular, cross tabulations for the complete cases of the variables by gender are presented in Table 2.9. For both males and females the proportions of respondents in each age group declines with increasing age. Over half of the respondents are married for both males (51.28%) and females (61.01%). A considerably greater proportion of males (43.78%) are single as compared to their female (24.6%) counterparts whereas a greater proportion of females (6.6%) are widowed than males (1.14%). There is a higher unemployment rate among the female respondents (64.97%) than among the male respondents (41.02%). A slightly lower proportion of females (65.79%) reside in urban areas as compared to males (70.63%). There are no marked differences in literacy and education levels between the females and the males.

Table 2.9: Cross tabulation of gender by age group

Age group	Female		Male	
	Number	Percentage (%)	Number	Percentage (%)
<i>Age group</i>				
15-19	1 739	21.29	1 556	24.61
20-24	1 617	19.79	1 132	17.91
25-29	1 535	18.79	987	15.61
30-34	1 139	13.94	822	13.00
35-39	918	11.24	683	10.80
40-44	657	8.04	481	7.61
45-49	564	6.90	328	5.19
50-54	-	-	333	5.27
<i>Marital status</i>				
Single	2 009	24.60	2 768	43.78
Married	4 984	61.01	3 242	51.28
Divorced	637	7.80	240	3.80
Widowed	539	6.60	72	1.14
<i>Employment status</i>				
Unemployed	5 307	64.97	2 593	41.02
Employed	2 862	35.03	3 729	58.98
<i>Place of residence</i>				
Rural	5 374	65.79	4 465	70.63
Urban	2 795	44.21	1 857	29.37
<i>Literacy Level</i>				
Non literate	537	6.57	322	5.09
Partially	574	7.03	529	8.37
Literate	7 058	86.40	5 471	86.54
<i>Education Level</i>				
No education	192	2.35	79	1.25
Primary	2 434	29.80	1 681	26.59
Secondary	5 214	63.83	4 177	66.07
Tertiary	329	4.03	385	6.09

Table 2.10 displays a marital status by age group contingency table. It is evident that the 15 – 19 years age group who are single/never married constitute almost 20% of the respondents whereas a substantial proportion of the married and the divorced respondents are in the middle age groups, that is 20 – 39. Generally the widowed are from the older age groups, 40 – 50.

Table 2.10: Contingency table for marital status by age group

	Marital status							
	Single		Married		Divorced		Widowed	
Age group	Number	%	Number	%	Number	%	Number	%
15-19	2 857	19.72	390	2.69	47	0.32	1	0.01
20-24	1 242	8.57	1 320	9.12	169	1.17	18	0.12
25-29	434	2.99	1 843	12.72	212	1.46	33	0.23
30-34	127	0.80	1 551	10.70	186	1.28	97	0.67
35-39	52	0.36	1 312	9.05	116	0.80	121	0.84
40-44	41	0.28	860	5.93	82	0.57	155	1.07
45-49	17	0.12	655	4.52	50	0.35	170	1.17
50-54	7	0.05	295	2.04	15	0.10	16	0.11

Chapter 3

Estimating HIV prevalence in Zimbabwe using population-based survey data

Abstract

Estimates of HIV prevalence computed using data obtained from sampling a subgroup of the national population may lack the representativeness of all the relevant domains of the population. These estimates are often computed on the assumption that HIV prevalence is uniform across all domains. Use of appropriate statistical methods together with population based survey data can enhance better estimation of national and subgroup level HIV prevalence and can provide improved explanations of the variation in HIV prevalence across the different domains. In the current study we computed design-consistent estimates of HIV prevalence, and their respective 95% confidence intervals at both the national and subgroup levels. In addition, we provided a multivariable survey logistic regression (which takes account of the complex sampling design) model from a generalized linear modelling perspective for explaining the variation in HIV prevalence

using demographic, socio-economic, socio-cultural and behavioural factors. Essentially, this study borrows from the proximate determinants conceptual framework which provides guiding principles upon which socio-economic and socio-cultural variables affect HIV prevalence through biological behavioural factors. We utilize the 2010-11 Zimbabwe Demographic and Health Survey (2010-11 ZDHS) data (which are population based) to estimate HIV prevalence in different categories of the population and for constructing the logistic regression model. It was established that HIV prevalence varies greatly with age, gender, marital status, place of residence, literacy level, belief on whether condom use can reduce the risk of contracting HIV and level of recent sexual activity whereas there was no marked variation in HIV prevalence with social status (measured using a wealth index), method of contraceptive and an individuals level of education.

3.1 Introduction

Zimbabwe ranks among the countries that have been worst affected by the HIV and AIDS epidemic in sub-Saharan Africa. Although studies have shown that HIV prevalence has been falling since the late 90s, an estimate from the National HIV estimates of 2010, of approximately 15% among the country's sexually active population (15 years and above) is still considered to be too high. The factors that have contributed to the observed decline in HIV prevalence since the late 90s include robust prevention programs, significant behaviour change that has reduced new infections and the successful implementation of the Prevention of Mother-to-Child Transmission (PMTCT) program. Recent studies have also revealed that the relatively high literacy level among the country's population has allowed rapid and effective dissemination of information on HIV and AIDS awareness and appropriate preventive measures, Gregson et al. (2006) and Hallett et al. (2006). These have in turn resulted in increased and consistent use

of condoms, reduction in casual sex, reduction in extramarital relationships and in commercial sex activities.

Zimbabwe has a wealth of literature on HIV/AIDS surveillance based on a variety of data sources. These range from Antenatal Clinic (ANC) surveys ran since 1990, Behavioural data obtained from the 2001/2002 Young Adult Survey (YAS), the 2001 and 2003 Population Services International (PSI) Youth Surveys, the Zimbabwe Demographic Health Surveys (1988, 1994, 1999 and 2005/6 with the latest one being 2010-2011), the World Health Organization (WHO) and the UNAIDS. Zimbabwe's HIV epidemic, just like in other sub-Saharan African countries, is mainly driven by heterosexual transmission, Champredon et al. (2013). This has mainly been linked to networks of multiple sexual relations, including concurrent relations in which the virus is passed on rapidly. The socio-economic and socio-political situations in the country have also exacerbated the spread of the virus. Despite aggressive awareness programs, studies have shown that increased commercial sex activities resulting from high unemployment rates have contributed significantly to new cases and the spread of the virus. Studies have also established that a leading cause of HIV infections in new-born babies is the MTCT, Wiegert et al. (2014). The 2001 National HIV Sentinel Surveillance Survey estimated an antenatal nationwide HIV sero-prevalence of 29.5% acquire HIV from their mothers annually in Zimbabwe, Perez et al. (2004). Maheswaran & Bland (2009) reported that MTCT before, during and after delivery may result in the acquisition of HIV for 30 – 35% of infants of HIV-infected mothers.

Statistical methods have been widely used to assess the trends, patterns and the dynamics of HIV/AIDS epidemic. The focus of these methods range, depending on the objectives embedded in the research, from estimating prevalence, understanding the pathogenesis of HIV infection, assessing the potency of antiviral therapies and evaluation of treatment efficacy of viral dynamic models, Anderson et al. (1999). Statistical models that estimate the magnitude and trajectory of the HIV/AIDS epidemic have

also been constructed. These models have been used as tools to extract and convey as much information as possible from available data and provide accurate representation of both the knowledge (understanding origin and progression) and uncertainty about the epidemic. The success of the statistical methods rely on the quality of the data available. Furthermore, the quality and reliability of the estimates calculated from the available data depends on the validity of the statistical analysis performed. Sampling and survey (including complex surveys) techniques have been used as bases for accurate data collection and provide sufficient theory behind consistent analysis of phenomena such as HIV epidemic. A key role of the discipline of statistics is to provide science with theory and methods for objective evaluation of data-based evidence and measuring the strength of that evidence, Royall (2003).

Previous studies involving national HIV prevalence estimation have mainly been based on data obtained from subgroups of the population. For instance, data obtained from pregnant women attending ANCs, from blood donors, from truck drivers, from commercial sex workers and from drug users, see Gregson et al. (2006), Anderson et al. (1999) and Ramjee & Eleanor (2002). The representativeness of these data to the target population has been argued to be inadequate, see Pettifor et al. (2005). The current study investigates how and describes the way in which HIV prevalence varies with demographic and socio-economic risk factors using population-based DHS data. We consider the extent to which the association between HIV status is affected by a person's age, gender and marital status, education, place of residence, wealth status, religion and behaviour towards HIV. In an attempt to enhance the quality of the estimates the study exploits the strength of statistical methods to provide estimates of HIV prevalence at the national and domain (subgroup) levels using nationally representative sample survey data. In addition, from a statistical modelling approach, a multivariable survey logistic regression model was computed. Essentially, the survey logistic regression is an extension of the ordinary logistic regression by accounting for the complex sampling

design.

3.2 Methods

3.2.1 The data

In addition to the data description given in section 2.2, for administration purposes, Zimbabwe is divided into ten provinces. During the 2002 population census (which was used as the sampling frame in the 2010-11 ZDHS) each province was subdivided into districts and each district is made up of wards and the wards consist of a number of enumeration areas (EAs). For the current research the response variable is HIV status, a binary variable indicating whether a respondent is HIV positive or negative. The study investigates the relationship between HIV and socio-economic, socio-cultural, demographic and behavioural factors (risk factors) of the population using a multivariable survey logistic regression model. In determining the risk factors, the study borrows from the proximate-determinants conceptual framework as explained in Boerma et al. (2003). Essentially, the underlying socio-economic, socio-cultural and environmental determinants operate through the proximate-determinants in order to affect an outcome such as HIV status. These factors include age, gender, marital status, education level, literacy level, economic status (wealth index), religion, province, method of contraceptive used, belief whether condom use works to reduce risk of HIV and recent sexual activities (measured in how sexually active a respondent has been in the previous four weeks) and place of residence (whether rural or urban). The sample consists of 17 434 respondents, 14 491 with non-missing values and an additional 2 943 with missing values for at least one measured variable. The current chapter assumes a complete case analysis where a list-wise deletion of cases with missing values is taken as explained in Little & Rubin (1987a). However the assumption of missing values being missing completely at random (MCAR) may be too restrictive. Future analyses may therefore

need methods that correct for the impact of the missing values and this is considered in Chapter 4 and Chapter 5.

3.2.2 Statistical computing

All the analyses were done using **survey** package by Lumley (2010) in **R** Team (2013). In particular, all the design features such as stratification, clustering and weighting were accounted for explicitly using the **svydesign** function. The function **svyglm** was used to describe the model by specifying the predictors and their functional form together with the link function. For automated model selection and multi-model inference, we utilized the **glmulti** package by Calcagno & de Mazancourt (2010), and **stepAIC** package by Park & Hastie (2010) respectively. The packages function by considering all possible explanatory variables and build unique models for the main effects and (optionally) the pairwise interactions. A stepwise, forward selection backward elimination procedure was used to select best predictor variables. This was done by utilizing the fundamental statistical modelling framework explained in Subsection 3.2.3 below. The model goodness of fit was done using **ResourceSelection** package by Lele et al. (2015). The function **hoslem.test** was used to perform the Hosmer-Lemeshaw (H-L) test for goodness of fit as explained by Hosmer & Lemeshow (2000).

3.2.3 Statistical methods

Novel design-consistent national and domain-level estimates of HIV prevalence and their respective measures of variability were computed using sound statistical techniques for analyzing complex survey data. In particular, point estimates in the form of proportions, their standard errors and $100(1 - \alpha)\%$ confidence intervals were computed. The Taylor series linearization variance estimation method as explained by Wolter (1985), that takes the design features into account was used.

For illustration of the underlying theory, we considered a complex sampling design

in which we have a finite population that can be broken into $k = 1, \dots, K$ strata, $j = 1, \dots, M_k$ primary sampling units (PSUs) in each stratum and $i = 1, \dots, N_{kj}$ elements in the $(k, j)^{\text{th}}$ PSU. Suppose that the observed data consist of n_{kj} elements from m_k PSU of stratum k , hence the total number of observations is $n = \sum_{k=1}^K \sum_{j=1}^{m_k} n_{kj}$. Suppose also that y_{kji} is a binary indicator for an attribute, for example 1 = HIV positive and 0 = HIV negative and let $\widehat{M} = \sum_{k=1}^K \sum_{j=1}^{m_k} \sum_{i=1}^{n_{kj}} w_{kji} y_{kji}$ and $\widehat{N} = \sum_{k=1}^K \sum_{j=1}^{m_k} \sum_{i=1}^{n_{kj}} w_{kji}$, where w_{kji} are the design-consistent weights. Here \widehat{M} is the design-consistent estimator of the total of all elements with the attributes of interest and \widehat{N} is the estimate of the population size. A design ratio mean estimator of the population proportion p , denoted \widehat{p} , of all the elements with the attribute of interest is given by

$$\widehat{p} = \frac{\sum_{k=1}^K \sum_{j=1}^{m_k} \sum_{i=1}^{n_{kj}} w_{kji} y_{kji}}{\sum_{k=1}^K \sum_{j=1}^{m_k} \sum_{i=1}^{n_{kj}} w_{kji}} = \frac{\widehat{M}}{\widehat{N}}. \quad (3.1)$$

The variance of the estimate is obtained by applying the Taylor series linearization as discussed by Wolter (1985) and Woodruff (1971) to the ratio estimator given in Equation 3.1 as

$$\text{var}(\widehat{p}) = \frac{\text{var}(\widehat{M}) + \widehat{p}^2 \text{var}(\widehat{N}) - 2 \times \widehat{p} \times \text{cov}(\widehat{M}, \widehat{N})}{\widehat{N}^2}, \quad (3.2)$$

and the standard error is given by $\text{se}(\widehat{p}) = \sqrt{\text{var}(\widehat{p})}$. Thus a 100(1 - α)% confidence interval for p is given by

$$CI(\widehat{p}) = \widehat{p} \pm t_{1-\alpha/2, df} \text{se}(\widehat{p}). \quad (3.3)$$

where α is the level of significance and t is the critical value of the Student's - t distribution. For the domain level, the estimator of the population proportion is given

as

$$\widehat{p}_k = \frac{\sum_{j=1}^{m_k} \sum_{i=1}^{n_{kj}} w_{ji} y_{ji}}{\sum_{j=1}^{m_k} \sum_{i=1}^{n_{kj}} w_{ji}}. \quad (3.4)$$

The standard error and the $100(1 - \alpha)\%$ confidence interval for the estimate given in Equation 3.4 are obtained in the same way as given in Equations 3.2 and 3.3.

The analysis of categorical variables (common with survey data), that is, the test for goodness of fit and of independence or of association, under complex design requires correction of the Pearson χ^2 test because the Pearson χ^2 test was developed under the assumption of multinomial or product-multinomial sampling. The modification needed involves adjusting the Pearson χ^2 statistic by a weighted sum of the cell and marginal design effects (**deff**), see for example Rao & Scott (1981), Holt et al. (1980) and Bedrick (1983). In particular, Holt et al. (1980) and Rao & Scott (1981) demonstrated, using simulations that the X^2 statistics for testing a null hypothesis of independence in contingency tables with complex survey data are asymptotically distributed as weighted sums of independent χ_1^2 variables.

Following Holt et al. (1980), for a test of goodness of fit, suppose the finite population can be split into k categories with population proportions p_1, \dots, p_k such that $\sum_{j=1}^k p_j = 1$. If we define $\mathbf{p} = (p_1, \dots, p_{k-1})'$, then the null hypothesis to be tested is

$$H_0 : \mathbf{p} = \mathbf{p}_0 = (p_1, \dots, p_{k-1})'.$$

The observed data produce unbiased estimates $\widehat{p}_1, \dots, \widehat{p}_k$ of the population proportions, thus, under a true H_0 , the ordinary χ^2 statistic is given as

$$\overline{X}^2 = n \sum_{j=1}^k \frac{(\widehat{p}_j - p_{0j})^2}{p_{0j}}.$$

If $\mathbf{P}_0 = \text{diag}(\mathbf{p}_0) - \mathbf{p}_0 \mathbf{p}_0'$ denotes the covariance matrix of $\widehat{\mathbf{p}}$ under simple random

sampling and assuming H_0 true, then an alternative form of \bar{X}^2 is given by

$$\bar{X}^2 = n (\hat{\mathbf{p}} - \mathbf{p}_0)' \mathbf{P}_0^{-1} (\hat{\mathbf{p}} - \mathbf{p}_0).$$

However, under a complex sampling design, such as the one used to obtain the data for the current research, that typically involves stratification and multistage sampling, the independence of observations assumption becomes inappropriate. Therefore, it is assumed that

$$\sqrt{n} (\hat{\mathbf{p}} - \mathbf{p}) \xrightarrow{L} N(\mathbf{0}, \mathbf{V}), \quad (3.5)$$

where \mathbf{V} is some positive-definite covariance matrix as $n \rightarrow \infty$, and \xrightarrow{L} denotes convergence to a multivariate normal distribution. Under the assumption in 3.5,

$$\bar{X}^2 \sim \sum_{i=1}^{k-1} d_i Z_i^2,$$

where Z_1, \dots, Z_{k-1} are asymptotically independent standard normal random variables and d_1, \dots, d_{k-1} are the eigenvalues of $\mathbf{D} = \mathbf{P}_0^{-1} \mathbf{V}$. Thus by Holt et al. (1980), the asymptotic distribution of \bar{X}^2 is a linear combination of χ_1^2 random variables and is exactly χ_{k-1}^2 under a multinomial case when all the d_i 's are equal to one.

For test for independence in contingency tables with complex survey data, suppose that the finite population can be cross-tabulated into r rows and c columns and let $\mathbf{p} = (p_{11}, \dots, p_{rc})'$ denotes a vector of cell probabilities where $\sum_{i=1}^r \sum_{j=1}^c p_{ij} = 1$. Suppose also that $\hat{\mathbf{p}} = (\hat{p}_{11}, \dots, \hat{p}_{rc})$ and that

$$\sqrt{n} (\hat{\mathbf{p}} - \mathbf{p}) \xrightarrow{L} N(\mathbf{0}, \mathbf{V}).$$

If we define the vectors of the marginal probabilities as $\mathbf{p}_r = (p_{1+}, \dots, p_{(r-1)+})'$, where $p_{i+} = \sum_{j=1}^c p_{ij}$ and $\mathbf{p}_c = (p_{+1}, \dots, p_{+(c-1)})'$, where $p_{+j} = \sum_{i=1}^r p_{ij}$, then the null hypothesis

of independence of rows and columns is given by

$$H_0 : p_{ij} = p_{i+}p_{+j} \quad (i = 1, \dots, r; j = 1, \dots, c).$$

According to Holt et al. (1980), the typical tests considered are of the form

$$h_{ij}(\hat{\mathbf{p}}) = \hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j} \quad (i = 1, \dots, r-1; j = 1, \dots, c-1),$$

Suppose that $\mathbf{h}(\mathbf{p}) = (h_{11}(\mathbf{p}), \dots, h_{(r-1)(c-1)}(\mathbf{p}))'$ and let $H(\mathbf{p})$ denotes the $(r-1)(c-1)$ matrix of partial derivatives $H(\mathbf{p}) = \partial\mathbf{h}(\mathbf{p})/\partial\mathbf{p}$, then under assumption 3.5

$$\sqrt{n}(h(\hat{\mathbf{p}}) - (\mathbf{p}-\mathbf{p}_0)) \xrightarrow{L} (\mathbf{0}, \mathbf{H}\mathbf{V}\mathbf{H}).$$

Under a complex sampling scheme the asymptotic chi-squared statistic is given by

$$\bar{X}_I^2 = \sum_{i=1}^{(r-1)(c-1)} \delta_i Z_i^2,$$

where the Z_i 's are the asymptotically independent standard normal random variables under H_0 and δ_i 's are the eigenvalues of

$$\mathbf{D}_I = (\mathbf{H}\mathbf{P}_0\mathbf{H}')^{-1} (\mathbf{H}\mathbf{V}\mathbf{H}').$$

A measure of the relative precision lost or gained by the use of specific complex design to an SRS is provided by the design effect denoted *deff*. The *deff*s are used mainly in computing confidence intervals, desired sample sizes and test statistics that incorporate the estimates of standard errors corrected for the complex sampling design, Heeringa et al. (2010). As mentioned in Chapter 1 above, analytic statistics such as the Rao-Scott Pearson χ^2 and the likelihood ratio χ^2 also rely on the design effect to reflect the effect of the complex sampling design compared to the SRS, see Rao &

Scott (1981) and Holt et al. (1980). A *deff* value greater than 1 indicates a gain in precision whereas a value less than 1 is indicative of loss in precision. Stratification tends to increase precision and clustering tends to decrease it, hence the overall *deff* depends on whether more precision is lost by clustering than gained by stratification, Lohr (2010), Bedrick (1983), Rao & Scott (1981) and Heeringa et al. (2010).

A survey logistic regression model for explaining the variation in HIV accounting for the underlying population structure, via a complex sampling design, from a GLM perspective was computed. A logistic regression model differs from an ordinary logistic regression in that it (the survey logistic) takes account of the complex sampling design. Essentially GLMs, as first introduced by Nelder & Wedderburn (1972) and further modified by McCullagh & Nelder (1989), are a flexible and unified class of models that are applicable to diverse types of response variables found in both normal and non-normal data including binary data. We discuss briefly the underlying theory behind logistic regression modelling from a GLM perspective.

As given in McCullagh & Nelder (1989) and Dobson & Barnett (2008), a GLM consists of three components; the random component, the systematic component and the link function. Essentially, the random component is the response variable and its probability distribution, that has to be a member of the exponential family of distributions. The systematic component represents the predictors whereas the link function links the random and the systematic components. In particular, a distribution for a response y belongs to the exponential family if

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

where θ is called the natural parameter, ϕ is called the dispersion parameter and $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions.

Specifically, let \mathbf{y} , a vector of observations on the response variable having n elements, be a realization of a random vector variable \mathbf{Y} whose elements are independent

and normally distributed with means $\boldsymbol{\mu}$. Under the general linear modelling framework,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \quad (3.6)$$

where:

\mathbf{X} is an $n \times p$ design matrix;

$\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters to be estimated from the data and;

the e_i 's are iid random variables with $e_i \sim N(0, \sigma^2)$ for $i = 1, \dots, n$.

The systematic part of the model gives a specification for the vector $\boldsymbol{\mu}$ in terms of a set of unknown parameters given in matrix notation as

$$\mathbf{E}(\mathbf{Y}) = \boldsymbol{\mu} = \mathbf{X}^T \boldsymbol{\beta}. \quad (3.7)$$

Under a generalized linear modelling approach the normality assumption is relaxed in order to include all models, such the binomial, Poisson and Gamma that belong to the exponential family. Instead of modelling $\boldsymbol{\mu} = \mathbf{E}(\mathbf{Y})$ directly as in Equation 3.7, some function $g(\boldsymbol{\mu})$ is modelled as

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}^T \boldsymbol{\beta}. \quad (3.8)$$

The function $g(\cdot)$ is called the link function, and this can be any monotonic differentiable function as described in Dobson & Barnett (2008) and McCullagh & Nelder (1989). The maximum likelihood estimates of the parameters $\boldsymbol{\beta}$ in the linear predictor $\boldsymbol{\eta}$ are obtained using the iterative least squares method utilizing numerical procedures such as the Newton-Raphson and the Fisher's scoring method. The maximum likelihood estimator, $\hat{\boldsymbol{\beta}}$, of the parameter vector $\boldsymbol{\beta}$ is asymptotically multivariate normal with mean $\boldsymbol{\beta}$ and covariance matrix \mathbf{I}_β^{-1} , that is, $\hat{\boldsymbol{\beta}} \sim \mathbf{N}(\boldsymbol{\beta}, \mathbf{I}_\beta^{-1})$, where \mathbf{I}_β is the Fisher information matrix, see McCullagh & Nelder (1989).

The logistic regression is used when the response variable follows a binomial distri-

bution. Let Y_i be a binary response variable assuming values 0 and 1 satisfying the binomial conditions, that is $Y_i \sim \text{Bin}(n_i, \pi_i)$, such that, (for instance)

$$y_i = \begin{cases} 1 & \text{if subject has attribute of interest, e.g the } i^{\text{th}} \text{ individual is HIV positive} \\ 0 & \text{if subject has no attribute of interest, e.g the } i^{\text{th}} \text{ individual is HIV negative} \end{cases}$$

It can be shown that the binomial distribution belongs to the exponential family of distributions, see McCullagh & Nelder (1989). Suppose that the probabilities π_i are dependent on a vector of observed covariates \mathbf{x}_i that are related to Y_i and can provide additional information for predicting Y_i . From a modelling point of view, the fundamental theory behind logistic regression seeks to construct a formal model thought to describe the variation in the probabilities π_i as a linear function of the covariates. That is

$$\pi(\mathbf{x}_i) = g^{-1}(\boldsymbol{\eta}) = g^{-1}(\mathbf{x}_i' \boldsymbol{\beta}), \quad (3.9)$$

However, there are potential problems with modelling probabilities using Equation 3.9, see Hosmer & Lemeshow (2000) and McCullagh & Nelder (1989): (a) linear models are unbounded and the right-hand side of Equation 3.9 can take the probabilities outside the unit interval; (b) in many practical applications, diminishing returns are observed, that is, changing π_i by the same amount requires a bigger change in x_i when π_i is already large (or small) than when π_i is close to 1/2 and linear models cannot accommodate that and (c) the assumptions of mean zero and constant variance of the errors that are made under linear models of the form of Equation 3.9 are not appropriate when the response variable is binary. As a remedy, a logit transformation can be performed.

Then the logistic regression can be expressed in terms of $\pi(\mathbf{x}_i)$ as

$$\text{logit}(\pi(\mathbf{x}_i)) = \log\left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)}\right) = \mathbf{x}_i' \beta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \quad (3.10)$$

The logit is nonlinear in $\pi(\mathbf{x}_i)$ but presumed to be linear in the parameters, which may be continuous and may range from $-\infty$ to $+\infty$, thus removing the floor limits, Hosmer & Lemeshow (2000). A logistic regression model is a GLM in that, the random component Y_i has a binomial distribution, having predictors \mathbf{x}_i and a logit link.

Equivalently, Equation 3.10 can be expressed as the odds of a positive response as

$$\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} = \exp\{\mathbf{x}_i' \beta\},$$

or as the probability of a positive response as

$$\pi(\mathbf{x}_i) = g^{-1}(g(\pi(\mathbf{x}_i))) = \frac{\exp\{\mathbf{x}_i' \beta\}}{1 + \exp\{\mathbf{x}_i' \beta\}}. \quad (3.11)$$

Under a complex sampling design however, a pseudo-maximum likelihood method for parameter estimation that take the structure of the underlying population, via the complex sampling design, into account is used. For a binary response variable satisfying the binomial conditions, the pseudo-maximum likelihood, as described by Breslow & Holubkov (1997), is given by

$$\prod_{k=1}^K \prod_{j=1}^{m_k} \prod_{i=1}^{n_{ki}} \pi(x_{kji})^{w_{kji} y_{kji}} [1 - \pi(x_{kji})]^{w_{kji} (1 - y_{kji})}.$$

Then the approximate log-likelihood function is given by

$$\sum_{k=1}^K \sum_{j=1}^{m_k} \sum_{i=1}^{n_{kj}} [w_{kji} \times y_{kji}] \times \ln[\pi(x_{kji})] + [w_{kji} \times (1 - y_{kji})] \times \ln[1 - \pi(x_{kji})].$$

Wald tests were employed to test the null hypothesis that a single coefficient is equal to zero, that is $H_0 : \beta_j = 0$ and confidence intervals are further used to provide

information on the potential magnitude and uncertainty associated with the estimated effects of individual predictor variables, Heeringa et al. (2010). For interval estimation, estimated design-based confidence intervals for the logistic parameter can be given as

$$CI_{1-\alpha}(\beta_j) = \hat{\beta}_j \pm t_{df, 1-\alpha/2} \cdot se(\hat{\beta}_j) \quad (3.12)$$

Alternatively, the significance of predictors can be carried out directly for the $\hat{\beta}'_j$ s on the log-odds scale. That is, in a logistic regression model with a single predictor, x_1 , an estimate of the odds ratio (OR) corresponding to a unit increase in the value of x_1 can be obtained by ‘exponentiating’ the estimated logistic regression coefficient giving $\hat{\psi} = \exp(\hat{\beta}_1)$. In a multivariable logistic regression model, $\hat{\psi}_j | \hat{\beta}_{k \neq j} = \exp(\hat{\beta}_j)$. This is an adjusted OR representing the multiplicative impact of a one-unit increase in the predictor variable x_j on the odds of the outcome variable being equal to one, controlling for the effects of the other variables. Confidence intervals can also be obtained for the adjusted ORs as

$$CI(\psi_j) = \exp\left(\hat{\beta}_j \pm t_{df, 1-\alpha/2} \cdot se(\hat{\beta}_j)\right)$$

Goodness of fit test were performed based on the Hosmer-Lemeshow (H-L) test proposed by Hosmer & Lemeshow (2000). Under the H-L test suppose that $J = n$, where n corresponds to the n values of the estimated probabilities arranged in ascending order, the H-L test is based on grouping the units into $k = 1, \dots, 10$ deciles called ‘deciles of risk’ using the estimated probabilities π_k . A H-L statistic is then given as

$$\hat{C} = \sum_{k=1}^g \frac{(o_k - n_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)}, \quad (3.13)$$

where n'_k is the total number of units in the k^{th} group, c_k is the number of covariate patterns in the k^{th} decile and $o_k = \sum_{j=1}^{c_k} y_j$ is the number of responses among the c_k covariate pattern and

$$\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \bar{\pi}_j}{n'_k}$$

is the average estimated probability. With the use of simulations, Hosmer & Lemeshow (1980) demonstrated that \hat{C} is well approximately $\chi^2_{(g-2)}$. For goodness-of-fit under complex sampling, the data are divided into weighted deciles of risk which have a weighted one-tenth of the n observations in the data-set in each group. That is n_1 observations with the smallest predicted probabilities are in the first group where n_1 is chosen so that $\sum_{i=1}^{n_1} w_{1i} / \sum_{i=1}^n w_i \simeq 0.1$; n_2 observations with the next smallest predicted probabilities are in the second group where n_2 is chosen so that $\sum_{i=1}^{n_2} w_{2i} / \sum_{i=1}^n w_i = 0.1$; until the tenth group is formed with n_{10} observations with the largest probabilities where n_{10} is chosen so that $\sum_{i=1}^{n_{10}} w_{10i} / \sum_{i=1}^n w_i = 0.1$. The w_{ki} is the sample weight for the i^{th} observation in the k^{th} (weighted) decile of risk. For the k^{th} decile, the weighted number of observations is $o_k = \sum_{i=1}^{n_k} w_{ki} y_{ki}$ and the weighted number of expected outcomes is $e_k = \sum_{i=1}^{n_k} w_{ki} \hat{\pi}_{ki}$. The H-L test tests H_0 : logistic regression model is adequate for the data versus H_1 : the logistic regression model is not adequate. A Wald test statistic for complex samples is based on $\hat{C}_d = (\mathbf{o} - \mathbf{e})' \mathbf{S}_d^{-1} (\mathbf{o} - \mathbf{e})$ where \mathbf{o} is the vector of weighted number of outcomes, \mathbf{e} is the vector of the weighted expected outcomes and \mathbf{S}_d is a design consistent estimator for the covariance matrix of $\mathbf{o} - \mathbf{e}$. The Wald test rejects the fit of the model when $\hat{C}_d > \chi^2_{g-2}$.

3.3 Results

This section presents the design-consistent descriptive estimates of HIV prevalence and their respective 95% confidence intervals at both national and domain levels. In addition, for analytic purposes a multivariable survey logistic regression model for HIV on the demographic, socio-economic, socio-cultural and behavioural factors was computed. Parameter estimates were expressed on the odds ratio (OR) scale in order to facilitate

interpretation of the logistic regression model. In particular, both the adjusted (for the effects of the other covariates in the model) and crude ORs are displayed together with their respective 95% confidence intervals. It is worthy pointing out that the crude descriptive estimates presented here differ slightly from those reported in the 2010-11 ZDHS report due to a number of factors. These factors may include the way in which the complex sampling design features are accounted for, the method used to handle missing data and the statistical software used.

3.3.1 Descriptive analysis

The estimated overall design-consistent HIV prevalence in the entire population was found to be $\hat{p} = 15.7\%$, 95% CI = 14.7 – 16.0%. The prevalence estimate is close to the $\hat{p} = 15.2$, 95% CI = 14.3 – 16.1% reported in the 2010-11 ZDHS report. The 95% CIs for the two estimates overlap showing that the difference is not statistically significant. HIV prevalence is known to vary considerably across population subgroups, hence in order to enhance the estimation and bring out the variation, we computed domain level estimates of HIV prevalence. The domains considered were based on the risk factors, as informed by the proximate determinant conceptual framework as explained in Boerma et al. (2003), such as gender, marital status and age that form the natural subgroups of the population. The crude design-consistent estimates for the different prominent subgroups of the population are given in Table 3.1. Specifically, for gender the results show that the females have a higher HIV prevalence rate ($\hat{p} = 17.7\%$, 95% CI = 16.6 – 18.7%) than the males ($\hat{p} = 12.8\%$, 95% CI = 11.8 – 13.7%).

Table 3.1: Crude design-based subgroup estimates of HIV prevalence along with their respective (crude) ORs and 95% confidence intervals

Risk Factor	Level	<i>n</i>	%	Est ³ (%)	95% CI	OR	95% CI
Gender:	Female	8 169	56.4	17.7	(16.6, 18.7)	Ref	
	Male	6 322	43.6	12.8	(11.8, 13.7)	0.690	(0.624, 0.763)
Marital status:	Single	4 777	33.0	5.6	(5.2, 6.6)	Ref	
	Married	8 226	56.8	16.7	(15.8, 17.7)	3.224	(2.796, 3.718)
	Divorced	877	6.1	28.8	(25.8, 32.3)	6.511	(5.320, 7.967)
	Widowed	611	4.2	54.4	(51.3, 60.0)	19.940	(16.038, 24.791)
Age group:	15 – 19	3 295	22.7	4.0	(3.2, 4.7)	Ref	
	20 – 24	2 749	19.0	7.9	(6.8, 8.9)	2.109	(1.664, 2.673)
	25 – 29	2 522	17.4	15.8	(14.2, 17.3)	4.360	(3.495, 5.439)
	30 – 34	1 961	13.5	23.2	(21.3, 25.2)	7.159	(5.756, 8.902)
	35 – 39	1 601	11.0	26.9	(24.5, 29.2)	8.556	(6.848, 10.690)
	40 – 44	1 138	7.9	25.5	(22.8, 28.3)	7.865	(6.210, 9.959)
	45 – 49	892	6.2	25.8	(22.7, 28.8)	8.427	(6.579, 10.793)
	50 – 54	333	2.3	18.7	(14.3, 23.1)	5.572	(3.919, 7.921)
Place of residence:	Rural	9 839	67.9	14.7	(13.9, 15.4)	Ref	
	Urban	4 652	32.1	16.8	(15.7, 18.0)	1.184	(1.069, 1.312)
Employment status:	Unemployed	7 900	54.5	13.5	(12.8, 14.3)	Ref	
	Employed	6 591	45.5	17.3	(16.3, 18.3)	1.350	(1.225, 1.487)
Literacy ¹	Non-literate	859	5.9	13.9	(11.5, 16.2)	Ref	
	Partially	1 103	7.6	19.8	(17.2, 22.3)	1.489	(1.153, 1.924)
	Literate	12 529	86.5	15.1	(14.4, 15.7)	1.083	(0.884, 1.326)
W/index ²	Poorest	2 811	19.4	15.8	(14.3, 17.2)	Ref	
	Poorer	2 652	18.3	14.6	(13.2, 16.1)	0.930	(0.795, 1.086)
	Middle	2 742	18.9	16.3	(14.9, 17.9)	1.075	(0.922, 1.254)
	Richer	3 134	21.6	16.0	(14.6, 17.4)	1.055	(0.911, 1.223)
	Richest	3 152	21.8	13.9	(12.6, 15.2)	0.859	(0.738, 1.001)
Education level	No education	271	1.9	17.0	(12.3, 21.7)	Ref	
	Primary	4 115	28.4	17.9	(16.4, 19.0)	1.053	(0.745, 1.488)
	Secondary	9 391	64.8	15.4	(14.6, 16.3)	0.891	(0.634, 1.252)
	Higher	714	4.9	12.6	(9.9, 15.3)	0.707	(0.468, 1.069)
Contraceptive	No method	8 076	56.1	14.4	(13.6, 15.3)	Ref	
	Traditional	107	0.7	21.0	(11.9, 30.0)	1.571	(0.909, 2.714)
	Modern	6 308	43.2	16.7	(16.6, 18.7)	1.269	(1.146, 1.406)
Religion	Apostolic	4 732	32.7	15.7	(14.4, 16.9)	Ref	
	Muslim	70	0.5	24.4	(14.5, 34.4)	1.748	(0.997, 3.046)
	None	2 035	14.0	17.2	(15.4, 19.1)	1.158	(0.993, 1.351)
	Other Christians	7 333	50.6	15.6	(14.7, 16.5)	1.005	(0.900, 1.121)
	Traditional	321	2.2	16.6	(12.3, 20.9)	1.073	(0.781, 1.476)
Believe condom works	Yes	12 190	84.1	16.8	(16.8, 17.6)	Ref	
	No	1 877	13.0	12.3	(10.7, 13.9)	0.696	(0.597, 0.813)
Recent sex activities	Don't know	424	2.9	8.2	3.5, 11.3	0.442	(0.294, 0.666)
	Never had sex	2 872	19.8	4.3	(3.5, 5.1)	Ref	
	Not currently active	3 680	25.4	22.2	(20.7, 23.7)	6.397	(5.211, 7.851)
	Postpartum	519	3.6	19.1	(15.5, 22.7)	5.268	(3.901, 7.112)
	Currently active	7 420	51.2	17.0	(16.0, 18.0)	4.584	(3.753, 5.598)

¹ Literacy was measured in terms of ability to read and write. Non-literate: those who cannot read nor write; Partially: those who can read or write part of a sentence; Literate: those who can read or write full sentence

² W/index is a composite measure of the household's cumulative living standard based on a household's ownership of selected assets such as televisions, vehicles as well as water access and sanitation facilities. It is generated using the principal component analysis and places households into five wealth quintiles

³ Est is the HIV prevalence as a percentage

In order to complement the descriptive statistics in Table 3.1, the HIV prevalence for the different domains in selected factors were presented graphically. Figures 3.1, 3.2 and 3.3 show the variation of HIV prevalence across different categories of marital status, five-year age-groups and across the administrative provinces of the country respectively. The results show that there are substantial differences in the HIV prevalence rates between the singles/never married, the married, the divorced and the widowed. The highest HIV prevalence (by marital status) was among the widowed, ($\hat{p} = 54.4\%$, 95% CI = 51.3 – 60.0%) and the lowest was among the single/never married individuals ($\hat{p} = 5.6\%$, 95% CI = 5.2 – 6.6%). The results also show that, for five-year age-group variable the highest prevalence is among the 35 – 39 years age-group ($\hat{p} = 26.9\%$, 95% CI = 24.5 – 29.2%) and the lowest prevalence is among the 15 – 19 years age-group ($\hat{p} = 4.0\%$, 95% CI = 3.2 – 4.7%).

Other notable variations can be observed for the place of residence variable where HIV prevalence is significantly higher among the urban dwellers ($\hat{p} = 16.8\%$, 95% CI = 15.7 – 18.0%) than among the rural residents ($\hat{p} = 14.7\%$, 95% CI = 13.9 – 15.4%). Regarding the recent sexual activity variable, those that never had sex have a significantly lower HIV prevalence ($\hat{p} = 4.3\%$, 95% CI = 3.5 – 5.1%) than the other categories, and those that were not sexually active in the previous month have the highest prevalence ($\hat{p} = 22.2\%$, 95% CI = 20.7 – 23.7%).

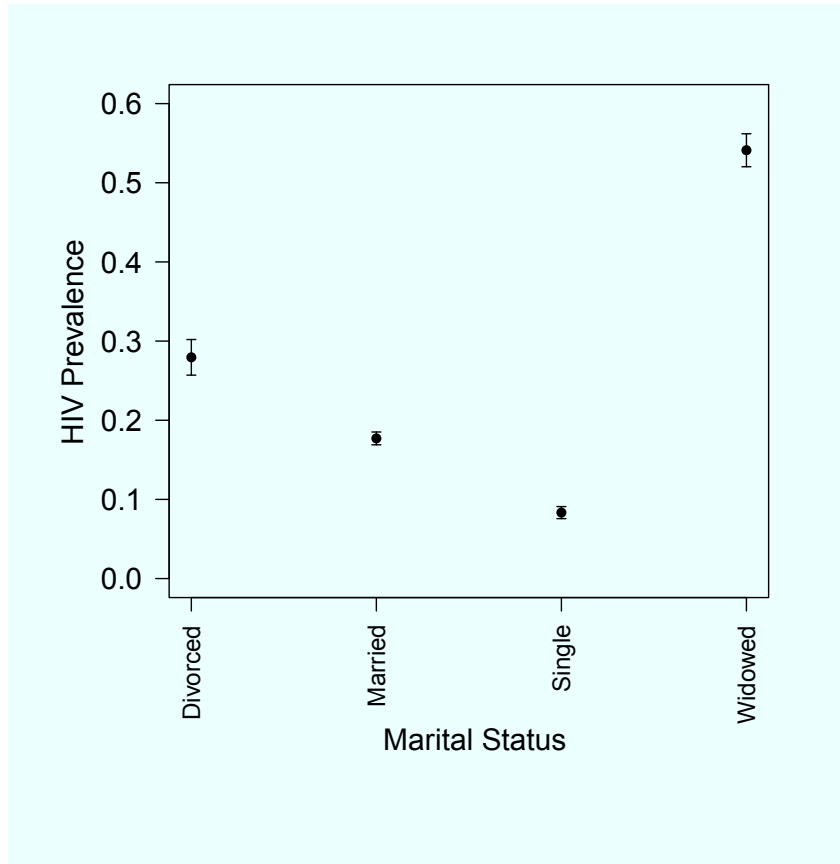


Figure 3.1: HIV prevalence across the different categories of marital status

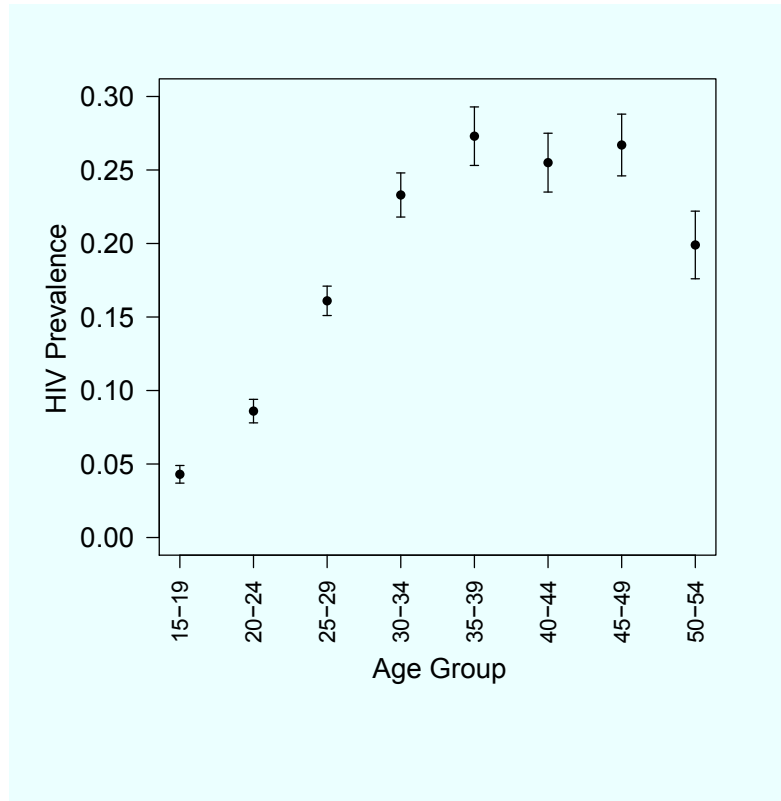


Figure 3.2: HIV prevalence across the different categories of five-year age-groups

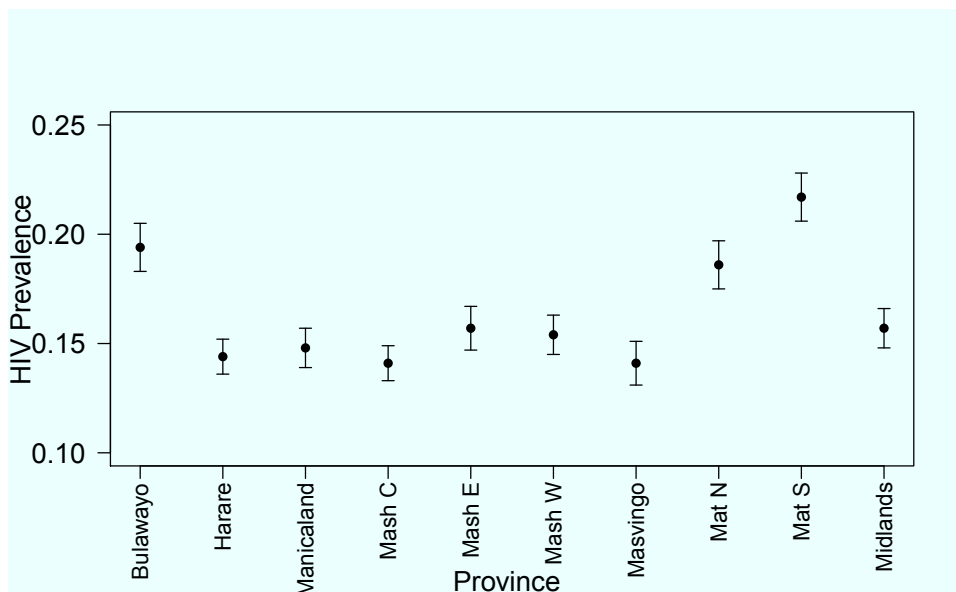


Figure 3.3: HIV prevalence in the different administrative provinces of the country

For the variation across provinces as shown in Figure 3.3, Matabeleland South

province has the highest ($\hat{p} = 21.1\%$, 95% CI = 19.5 – 24.0%) whereas Mashonaland Central province has the lowest ($\hat{p} = 13.8\%$, 95% CI = 12.5 – 15.8%). Matabeleland North province ($\hat{p} = 18.5\%$, 95% CI = 14.8 – 22.2%) and Bulawayo ($\hat{p} = 18.8\%$, 95% CI = 15.9 – 21.8%) also have relatively higher HIV prevalence than the rest of the provinces whose HIV prevalence range between $\hat{p} = 13.6\%$ to $\hat{p} = 15.6\%$. Table 3.1 also displays crude (unadjusted) ORs for each risk factor to quantify the odds of HIV for each category relative to their respective reference levels. For instance, for the factor gender, relative to the females, the results show that the males have higher odds of HIV (OR = 0.69, 95% CI = 0.624 – 0.763). For the marital status factor, with reference to those who are single/never married, the results show that the odds of HIV are higher (OR = 3.224, 95% CI = 2.796 – 3.718) for the married individuals, over six times higher for the divorced individuals (OR = 6.511, 95% CI = 3.495 – 5.439) and almost twenty times higher (OR = 19.94, 95% CI = 16.038 – 24.791) for widowed individuals. All the other odds ratios in Table 3.1 can be interpreted in a similar way.

3.3.2 Logistic regression analysis

This section presents details of how the logistic regression analysis was used to construct a model for HIV prevalence. For the preliminaries we considered the relationship between HIV prevalence and age as a continuous variable. Figure 3.4 gives a plot of the average HIV prevalence for a given age against the respective age. It is evident, from the plot that there is generally a positive linear relationship with some curvature among the old respondents.

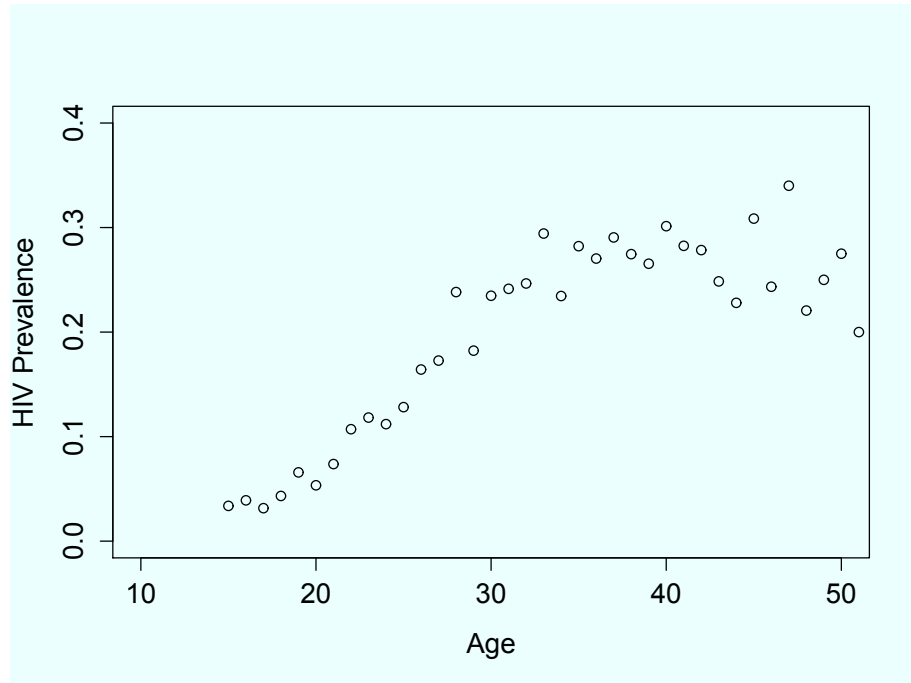


Figure 3.4: The average HIV prevalence per a given age versus age

Bivariate design-based tests for association between each of the categorical predictor variables and the response variable were also performed. In particular the Rao-Scott test by Rao & Scott (1981) that takes account of the design-induced distortion of the asymptotic distribution of the Pearson χ^2 test as explained in Subsection 3.2.3 above and the likelihood ratio statistics was used. Table 3.2 displays the test results. The results indicate that the factors religion and wealth index are not significantly associated with HIV whereas the rest of the factors are significantly associated based on the p-values. The non-significant variables were not dropped completely for the computations of the multivariable survey logistic regression model, however even after being included they were found to also contribute insignificantly in explaining the variation in HIV.

Table 3.2: Rao-Scott (F – based) test statistics and p-values, for association of individual predictor variables and HIV status

Variables	Levels	n	%	HIV pos.	F value	p-value
Gender	Female	8 169	56.4	1 521	82.92	< 0.001
	Male	6 322	43.6	867		
Marital status	Single	4 777	33.0	308	326.17	< 0.001
	Married	8 226	56.8	1 465		
	Divorced	877	6.1	277		
	Widowed	611	4.2	338		
Age Group	15 – 19	3 295	22.7	140	109.06	< 0.001
	20 – 24	2 749	19.0	252		
	25 – 29	2 522	17.4	443		
	30 – 34	1 961	13.5	490		
	35 – 39	1 601	11.0	444		
	40 – 44	1 138	7.9	310		
	45 – 49	892	6.2	244		
	50 – 54	333	2.3	65		
Employment	Yes	6 591	45.5	1 216	30.27	< 0.001
	No	7 900	54.5	1 172		
Place of Residence	Rural	9 839	67.9	1 551	5.29	0.0220
	Urban	4 652	32.1	837		
Education	No Education	271	1.9	48	5.80	0.001
	Primary	4 115	28.4	774		
	Secondary	9 391	64.8	1 475		
	Higher	714	4.9	91		
Wealth Index	Poorest	2 811	19.4	483	1.41	0.230
	Poorer	2 652	18.3	430		
	Middle	2 742	18.9	475		
	Richer	3 134	21.6	549		
	Richest	3 152	21.8	451		
Literacy	Non-literate	859	5.9	133	7.29	0.0008
	Partially	1 103	7.6	229		
	Literate	12 529	86.5	2 026		
Religion	Apostolic	4 732	32.7	754	1.955	0.099
	Muslim	70	0.5	20		
	Non	2 035	14.0	356		
	Other Christians	7 333	50.6	1 197		
	Traditional	321	2.2	61		
Contraceptive	Non	8 076	56.1	1 246	10.818	< 0.001
	Traditional	107	0.7	19		
	Modern	6 308	43.2	1 155		
Believe condom works	Yes	12 190	84.1	2 146	17.551	< 0.001
	No	1 877	13.0	228		
	Don't know	424	2.9	37		
Recent Sex Activities	Never had sex	2 872	19.8	128	121.445	< 0.001
	Not active last month	3 680	25.4	846		
	Postpartum	519	3.6	106		
	Currently active	7 420	51.2	1 329		

It was established that HIV prevalence varies considerably with gender (that is, a gender effect) for each age group as displayed in Figure 3.5. The plot shows that although HIV prevalence generally increases with age for both males and females, it rises faster in females than in males among the lower age groups, however the prevalence becomes higher among males than among females from the 40 – 44 year age group and older. This implies that the risk of HIV varies by gender across different age groups necessitating the inclusion of a gender by age group interaction effect.

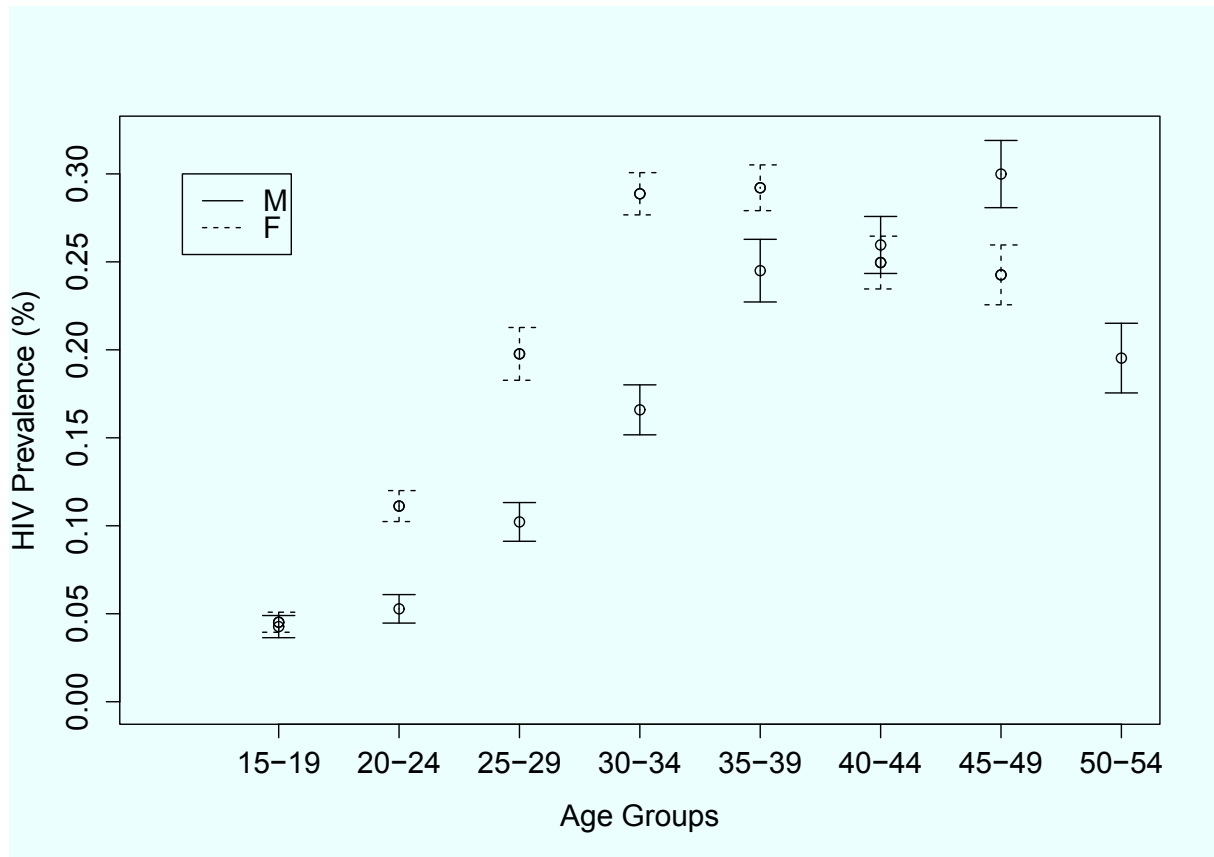


Figure 3.5: HIV prevalence estimates and their respective 95% confidence intervals by age group for males and females separately

From a logistic regression modelling perspective, the results show that HIV prevalence is dependent (conditionally) on gender, age, marital status, literacy level, place of residence, recent sexual activities, one’s opinion on whether condom use helps reduce risk of contracting HIV and a gender by age interaction effect (estimates of the model

not shown here). The H-L goodness of fit statistic $\hat{C} = 15.157$ on $g - 2 = 8$ d.f giving a p-value of 0.141 showing no evidence of lack of fit.

In order to facilitate the interpretation of the results of the computed model, we expressed the parameter estimates as adjusted (to control for the effect of other factors in the model) ORs and assess the effects of confounding. It is important to note that in the presence of confounding among the variables, the adjusted ORs take precedence. We presented the unadjusted ORs for the factors included in the final model. The results are displayed in Table 3.3. For the interpretation of the ORs we assume the reference cell method as explained by Breslow & Holubkov (1997), Rao & Scott (1981) and Agresti (2007).

Table 3.3: Estimated adjusted ORs and crude ORs together with their respective 95% confidence intervals for the parameter estimates for the logistic regression model

Parameter	Adjusted OR	95% CI	Crude OR	95% CI
Intercept	0.024	(0.016, 0.036)		
Gender				
Male	0.992	(0.665, 1.503)	0.696	(0.624, 0.763)
Age Group				
20 – 24	1.944	(1.386, 2.726)	2.109	(1.664, 2.673)
25 – 29	3.454	(2.487, 4.797)	4.360	(3.495, 5.439)
30 – 34	5.211	(3.739, 7.262)	7.159	(5.756, 8.902)
35 – 39	5.098	(3.626, 7.166)	8.556	(6.848, 10.690)
40 – 44	3.041	(2.082, 4.440)	7.865	(6.210, 9.959)
45 – 49	2.914	(1.939, 4.379)	8.427	(6.579, 10.793)
50 – 54	3.349	(2.050, 5.470)	5.572	(3.919, 7.921)
Marital Status				
Married	0.922	(0.727, 1.169)	3.224	(2.796, 3.718)
Divorced	1.801	(1.374, 2.359)	6.511	(5.320, 7.967)
Widowed	4.484	(3.315, 6.067)	19.940	(16.038, 24.791)
Literacy				
Partially	1.536	(1.145, 2.061)	1.489	(1.153, 1.924)
Literate	1.240	(0.982, 1.567)	1.083	(0.884, 1.326)
Place of Residence				
Urban	1.225	(1.087, 1.381)	1.184	(1.069, 1.312)
Recent Sex activities				
Not active last month	2.112	(1.651, 2.701)	6.397	(5.211, 7.851)
Postpartum	1.707	(1.200, 2.428)	5.268	(3.901, 7.112)
Currently active	1.826	(1.405, 2.373)	4.584	(3.753, 5.598)
Believe condom works				
No	0.881	(0.747, 1.038)	0.696	(0.597, 0.813)
Don't know	0.506	(0.350, 0.732)	0.442	(0.294, 0.666)
Age group*Gender				
20 – 24 : Male	0.462	(0.263, 0.797)		
25 – 29 : Male	0.484	(0.295, 0.792)		
30 – 34 : Male	0.550	(0.339, 0.894)		
35 – 39 : Male	0.919	(0.565, 1.496)		
40 – 44 : Male	1.726	(1.019, 2.922)		
45 – 49 : Male	2.289	(1.305, 4.015)		

The results in Table 3.3 show that the odds of HIV are almost equal (OR = 0.992,

95% CI = 0.665 – 1.503) for males and females, holding the effects of the other variables in the model constant. With reference to the 15 – 19 year old respondents, the results show that the odds of HIV increase with age, peaking among the 30 – 34 year old respondents (OR = 5.211, 95% CI = 3.739 – 7.262) before falling among the ‘older’ respondents, controlling for the other variables in the model. Relative to the single/never married, the results show that the odds of HIV are slightly lower (OR = 0.922, 95% CI = 0.727 – 1.169) for the married individuals, almost twice (OR = 1.801, 95% CI = 1.374 – 2.359) for the divorced and over four times (OR = 4.484, 95% CI = 3.315 – 6.067) for the widowed controlling for the effect of the other variables in the model.

With reference to the non-literate, the odds for HIV for the partially literate are approximately one and a half (OR = 1.536, 95% CI = 1.145 – 2.061) higher and slightly higher for the literate (OR = 1.240, 95% CI = 0.982 – 1.567) controlling for the effect of the other variables in the model. In relation to those respondents who have never had sex, those who were not sexually active (for reasons other than postpartum) in the previous four weeks have odds over twice higher (OR = 2.112, 95% CI = 1.651 – 2.701), whereas those who were not sexually active (for postpartum reasons) in the previous four weeks have higher odds of HIV (OR = 1.707, 95% CI = 1.200 – 2.428) and those who were currently sexually active have odds of HIV almost twice higher (OR = 1.826, 95% CI = 1.405 – 2.373). Regarding individuals’ opinion on whether condom use reduces the risk of contracting HIV, the odds of being HIV are less (OR = 0.575, 95% CI = 0.386 – 0.855) among those who say they do not know and the odds are slightly higher (OR = 1.136, 95% CI = 0.964 – 1.338) among those who say no than those who answered yes, holding other variables in the model constant.

The five-year age-group by gender interaction (effect modification) shows the additional effect of age on HIV prevalence for males in relation to females. The results show that among the lower age-groups (20 – 24, 25 – 29 and 30 – 34) age has a increasing

effect on the odds of HIV for females as compared to the males whereas among the older age-groups, age has an increasing effect on the odds of HIV for males as compared to the females. This implies that it is more likely for a young female person to be HIV positive as compared to a young male person.

3.4 Discussion

There are a variety of possible reasons for the observed variations in HIV prevalence across various categories of the risk factors of the respondents. These are mainly driven by biological, socio-economic and socio-cultural factors. It is important to note that sexual contact remains the key driver of HIV transmission among the sexually active population in sub-Saharan Africa. In particular the higher risk of HIV among the females as compared to the males could be explained by a number of factors that make females more vulnerable. These include different rates of sexual interaction and relationships between males and females brought about by socio-cultural practices such as polygamy that give rise to differing susceptible rates to HIV. In addition, socio-economic issues such as economic dependence of women on men, that is still highly prevalent in most sub-Saharan African countries, leaves women with limited negotiating power with regards to sexual matters putting them at more risk. It is worthy noting that during penile-vaginal intercourse, a woman's body is more susceptible to HIV infection than a man's. There is increased surface area of the body parts of a woman where HIV transmission can occur than on a man.

The relatively low prevalence among the single/never married individuals possibly points to the fact that HIV transmission is mainly driven by sexual contact, and as such most of those who are single/never married are most likely not yet sexually active. On the other hand the relatively high prevalence among the widowed may be a indication that the partner lost died due to AIDS-related causes. Similar to the variation across

marital status classes, the relatively low prevalence among the 15 – 19 is perhaps due to the fact that these are mainly young, possibly of school going age and relatively less sexually active. On the contrary the middle ages, 25–40 year old are often characterized by high sexual activities hence a possible explanation for the relatively high prevalence.

The observed urban-rural variation can be attributed to the fact that majority of urban residents are middle aged and are synonymous with relatively high sexual activities. In addition, commercial sex activities are also prominent in urban areas than in rural areas, argued as one of the key contributors of new HIV infections. The lack of sexual activity reported among those with high prevalence is possibly linked to their HIV status. Studies have shown that although reaction to the discovery of one's seropositivity vary, there is a general decline in sexual activity. The significantly high HIV prevalence in Matabeleland South province is possibly due to the border towns in the province, such as Beitbridge and Plumtree that are synonymous with increased commercial sex activities.

The additional effect of age on the gender effect on HIV prevalence that sees the young females being at higher risk of HIV than their young male counterparts agrees well with a general belief that young females engage in sexual activities with older males. From an epidemiological perspective, the variable gender is considered an effect modifier on age as a risk factor in explaining the variation in HIV. There is generally a similar trend in the variation in HIV prevalence across different categories of the risk factors between the adjusted and the unadjusted odds ratios except in the magnitude of the risk across the categories.

3.5 Potential strengths and limitations of study

The approach of the current chapter draws its main strength from the appropriate application of sound statistical methods coupled with the utilization of advanced statistical

computing software in estimating HIV prevalence. However, a potential drawback of the current study comes from the use of secondary data which often leaves the data analyst with limited control over the data collection process although this is not to downplay DHS data that are carefully collected by a team of highly trained statisticians with excellent expertise in survey methodology. Furthermore, the complete case analysis approach that we made regarding missing data often results in loss of statistical information especially if the distribution of observed data is different from that of the missing data. Thus the current chapter can be regarded as a base up on which future research involving properly handling missing data on estimation of HIV prevalence using population-based data can be built. This is the main idea of Chapter 4.

3.6 Conclusion

Estimating national HIV prevalence using data obtained from sampling a subgroup of the entire population is argued to have shortcomings in that the estimates might be biased especially if the subgroup is not representative of the target population. In addition the estimation does not display the variation across different domains of the population. This study has utilized population-based survey data supported by the use of sound statistical methodology for analyzing complex survey data to enhance better estimation of both national and domain level HIV prevalence. Furthermore, explaining variation in HIV using risk factors was aided by the use of population-based survey data that allow linking HIV status to demographic, socio-economic, socio-cultural and behavioural factors, and as well making use of the proximate determinants conceptual framework. The current chapter provided crude design-based estimates of HIV prevalence, at the national level as well as domain estimates based on the prominent risk factors in the population. From a modelling stand-point, a survey logistic regression model was fitted to provide a way of explaining the variation in HIV prevalence using

socio-economic, socio-cultural, behavioural and demographic variables.

The results from the study showed that HIV prevalence is dependent on one's age, gender, marital status, literacy level, level of recent sexual activities, belief about whether condom use reduces risk of HIV and place of residence. The results also showed that one's religion, education level and wealth status do not play a significant role in determining one's HIV status. It is further revealed that HIV prevalence is higher among females than in males (both crude and adjusted). Generally HIV prevalence increases with age. This shows that, for randomly selected 'older' person, the probability of obtaining an HIV positive individual is higher than for a 'younger' person. The results also showed that, as compared to the single or never married people, the married, divorced and widowed are more likely to be HIV positive. People in the urban areas have higher HIV prevalence as compared to their rural counterparts. This study found that there is an age by gender interaction effect on HIV prevalence, namely that the HIV prevalence increases with age at a faster rate in males than in females.

Chapter 4

Multiple imputation for non-response when estimating HIV prevalence using population-based survey data

Abstract

Missing data are a common feature in many areas of research especially those involving survey data in biological, health and social sciences research. Most of the analyses of the survey data are done taking a complete-case approach, that is taking a list-wise deletion of all cases with missing values assuming that missing values are missing completely at random (MCAR). Methods that are based on substituting the missing values with single values such as the last value carried forward, the mean and regression predictions (single imputations) are also used. However these methods often result in potential bias in estimates, in loss of statistical information and of distributional relationships between variables. In addition, the strong MCAR assumption is not always tenable in most practical instances.

Since missing data are a major problem in HIV research, the current research seeks

to enhance HIV prevalence estimation via the implementation of the multiple imputation procedure, as a method of handling missing data. This is particularly possible since the multiple imputation method is designed to draw multiple values for the missing observations from plausible predictive distributions for them and correctly account for the uncertainty brought about by the very process of imputing the missing data themselves. The proper handling of missing data is important especially in HIV research in sub-Saharan Africa where accurate collection of (complete) data is still a challenge. Specifically, national and subgroup estimates of HIV prevalence in Zimbabwe were computed using several imputed data sets for missing data in the 2010-11 Zimbabwe Demographic and Health Surveys (2010-11 ZDHS) data. A survey logistic regression model for HIV prevalence on demographic and socio-economic variables was used as the substantive analysis model. The results for both the complete-case analysis and the multiple imputation analysis are presented and discussed.

Across different subgroups of the population, the crude estimates of HIV prevalence are generally different but their variations are consistent between the two approaches (complete-case analysis and multiple imputation analysis). The respective estimates of the standard errors, and hence the lengths of the confidence intervals, vary considerably between the two approaches (multiple imputation and complete case). Similarly, the logistic regression adjusted odds ratios also exhibit great variations between the two approaches. The model based confidence intervals for the adjusted odds ratios are predominantly wider under the multiple imputation which is indicative of the inclusion of a combined measure of the within and between imputation variability.

The use of multiple imputations allows the uncertainty brought about by the imputation process to be measured. This consequently yields more reliable estimates of the parameters of interest and reduce the chances of declaring significant effects unnecessarily (type I error). In addition, the utilization of the powerful and flexible statistical computing packages in **R** enhances the computations.

4.1 Introduction

Most practical survey data, especially those obtained for scientific and social investigations, are often characterized by missing data as a result of non-response, see for example Brick & Kalton (1996), Raghunathan (2010), Raghunathan (2004) and Pigott (2001). In particular non-response is regarded as a pervasive and persistent problem in most social research studies. In practice, see for example Wang et al. (1992), Little (1988) and Kalton & Kasprzyk (1986) analyses of incomplete data, especially in longitudinal studies, often take a complete-case analysis approach despite the fact that current statistical software resources have capabilities to mitigate the problems. That is, a list-wise deletion approach in which cases with missing values are omitted from the analysis is often adopted. This is mainly based on the assumption that missing data are missing completely at random (MCAR) as described by Rubin (1987). However this assumption is generally difficult to justify in practice neither it is easily tenable. Furthermore, ad hoc methods that substitute the missing values with single values such as the last value carried forward, the mean and regression predictions (single imputation) are also often used. However these methods also have considerable drawbacks especially if the percentage of missing data is high, as explained by Rubin (1987) and Sterne et al. (2009). Biased results can be obtained if the complete data are not representative of the entire sample (MCAR assumption is violated) and possibly the target population, and also relationships among variables are lost. In addition, single imputation may yield unduly small standard errors since the uncertainty about the imputed values is not accounted for Sterne et al. (2009).

There are several reasons why data are missing in surveys, see for example Rubin (1987), Sterne et al. (2009), Little & Rubin (1987a), Kalton & Brick (1996) and Baraldi & Enders (2010). Essentially, missing data may be a result of an element in the target population not being included on the survey's sampling frame, resulting in what is called non-coverage. These elements have zero probability of being selected into the sample.

If a sampled element does not participate in the survey, this results in total/unit non-response. Total non-response may occur because of a participant's refusal to take part in the survey or due to language barrier or non-availability on the day of interview. The success of data collection in surveys, particularly in household surveys relies on the availability of participants on the day of interview. However participants are often unavailable giving rise to non-response and consequently missing data. Furthermore, a responding sampled element can fail to provide acceptable responses to one or more of the survey items resulting in what is called item non-response. The reasons for item non-response range from a respondent refusing to answer a question because it is too sensitive or does not know the answer or gives an answer that is inconsistent with answers to other questions Rubin (1987), Lohr (2010) and Schafer (1997). A non-response that falls between unit and item non-response is called partial non-response. Partial non-response occurs when a substantial number of item non-response occurs. This can occur, for instance, when a respondent cuts off the phone call in the middle of an interview or when a respondent in a multi-phase survey provides data for some but not all phases of data collection Rubin (1987), Kalton & Brick (1996) and Lohr (2010).

Missing data are classified according to the relationship between measured variables and the probability of missing data in what Rubin (1987), Little & Rubin (1987a) and Baraldi & Enders (2010) referred to as "missing data mechanisms". The missing data mechanism defines the distribution of missing data given the underlying data. The missing data can fall into one of three missing data mechanisms namely missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

Various methods have been developed in an attempt to compensate for non-response in survey data. The form of compensation depends on the source of the missing data. As described by Rubin (1987), Kalton & Brick (1996) and Little & Rubin (1987a) deletion, weighting adjustments and imputation methods are the most common ways used for

handling and/or compensating for non-response. In particular, compensation for total non-response and non-coverage is made by weighting adjustments. The respondents are assigned greater weight in the analysis so as to account for the shortfall resulting from the non-respondents. In the case of non-coverage, since the sample provides no information about the missing elements, weighting adjustments are based on external data sources. For the case of item non-response, compensation is done via imputation, see Rubin (1987). Imputation (which is the subject of the current chapter) involves systematically filling the missing value with new assigned values. Partial non-response can be compensated by both weighting adjustments and imputation.

Most statistical methods for data analysis assume a rectangular matrix with rows representing units and the columns representing variables measured (completely) for each unit. However this is often not the case in most practical scientific and social research including HIV studies due to missing data.

Originally suggested by Rubin (1987), multiple imputation is a Monte Carlo (or simulated based) technique that replaces each missing value with two or more plausible values. Essentially each missing value is imputed $m (\geq 2)$ different times using the same imputation method creating m data sets with no missing values. Each completed data set is analyzed using standard complete-data procedures as if the imputed data were real data obtained from the non-respondents. The results are later combined to produce estimates and confidence intervals that incorporate missing-data uncertainty. The estimates obtained are called multiple imputation estimates, Rubin (1987), Schafer (1997), Heeringa et al. (2010), Schafer & Olsen (1998) and Schafer (1999). The observed values are used to provide indirect evidence about the likely values of the unobserved ones averaging over the distribution of the missing data given the observed data Sterne et al. (2009). Thus for this reason multiple imputation falls under the MAR missingness mechanism as opposed to the MCAR. Key to this lies in correctly specifying the imputation model. In addition, the multiple imputation procedure is a computational

intensive analytic approach that accounts for the variability due to the missing values.

A Bayesian inference paradigm is utilized in the simulative multiple imputation procedure. Independent drawing of the parameters and the missing values from a posterior predictive distribution are carried out in a Bayesian framework, Rubin (1987) and Spratt et al. (2010). The multiple imputation procedure is carried out in three steps: 1) imputing data under an appropriate model to obtain m ‘complete’ data sets; 2) analyze each data set separately to obtain desired parameter estimates and standard errors; and 3) combining the results of the analyses from the m data sets by finding the mean of the m parameter estimates and a variance estimate that accounts for both the within-imputation and across-imputation variability using formulae given by Rubin (1987).

Since multiple imputation relies on a Bayesian paradigm, a prior distribution for the parameters is required. We used a non-informative prior distribution which is a default prior in most software packages. that correspond to a state of prior ignorance about model parameters, Lesaffre & Lawson (2012) and Press (1989). To simulate the draws from the posterior distribution for the missing values given the observed data, MCMC procedures were used as explained in Rubin (1987), Lesaffre & Lawson (2012) and Press (1989). The application of the multiple imputation comes with potential problems that are worthy noting as pointed out by Sterne et al. (2009). These include, challenges pertaining to ways for handling non-normally distributed variables, plausibility of the MAR assumption and how to handle data that are MNAR. For the current research, these are adequately accounted for in the statistical package **mi**, as explained in Section 4.2.5 below, that we used for the multiple imputation computations. The approach in this current chapter followed the guidelines outlined in strengthening the reporting of observational studies in epidemiology (STROBE) as outlined in von Elm et al. (2007). The MNAR approaches which rely on sensitivity analysis are not the focus of the current application.

The chapter is organized in the following format. Section 4.2 gives an overview of the data used for analysis, the underlying concepts of the multiple imputation procedure, a brief description of the missing data and the statistical computing package used for the analysis. Section 4.3 presents the results of the analyses in the form of descriptive and logistic regression analyses from both a complete case analysis and a multiple imputation analysis. Section 4.4 gives a detailed discussion of the findings. The potential strengths and limitations of the approach taken in this current chapter are presented in Section 4.5 and Section 4.6 gives the concluding remarks.

4.2 Methods

4.2.1 Types of missingness

Missing data are classified according to the relationship between measured variables and the probability of missing data in what Rubin (1987), Little & Rubin (1987a) and Little & Rubin (1987b) termed “missing data mechanisms”. Missing data can fall into one of three missing data mechanisms namely missing completely at random (MCAR), missing at random (MAR) and not missing at random (MNAR).

In order to illustrate the three missing data mechanisms, following Rubin (1987), suppose that $\mathbf{Y} = \{\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}\}$, where \mathbf{Y}_{obs} are the observed values and \mathbf{Y}_{mis} are the unobserved values and let \mathbf{M} be a missing data indicator matrix of the same dimension as \mathbf{Y} where the value in row i and column j is equal to 1 if the value in \mathbf{Y} is missing and 0 if the value is observed. Data are MCAR if $P(\mathbf{M}|\mathbf{Y}) = P(\mathbf{M})$ for all \mathbf{Y} , that is, the fact that the data are missing is not dependent on any values or potential values for any of the variables. That is the probability that a respondent does not report an item value for instance is completely independent of the true underlying values of all the observed and unobserved variables, Heeringa et al. (2010). Missingness is completely unsystematic and the observed data can be regarded as a random sub-sample of the

hypothetically complete data. Thus inference can be carried out with the observed data since they are representative of the complete sample as well as the target population.

Missing data are MAR if missingness is related to other measured or observed variables in the analysis, but not to the underlying values of the incomplete variable, that is the hypothetical values that would have resulted had the data been complete, Baraldi & Enders (2010). Thus MAR implies that $P(\mathbf{M}|\mathbf{Y}) = P(\mathbf{M}|\mathbf{Y}_{\text{obs}})$ for all \mathbf{Y} . The response mechanism responsible for MCAR and MAR is termed ignorable, Rubin (1987) and Pigott (2001).

Missing data are MNAR if they are neither MCAR nor MAR, that is if the missing data are not at least MAR. Specifically, missing data are MNAR if missingness depends on both the observed and unobserved values of \mathbf{Y} , that is $P(\mathbf{M}|\mathbf{Y}) = P(\mathbf{M}|\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$. The MNAR mechanism is also called non-ignorable missing data mechanism.

4.2.2 Multiple imputations

Multiple imputation was originally proposed by Rubin (1987), and is a Monte Carlo (or simulated based) technique that is used to ‘fill in’ each missing value with two or more plausible values. Essentially each missing value is imputed m (≥ 2) different times using the same imputation method creating a vector of m ‘complete’ data sets, that is with no missing values. Each of the data set is then analyzed using standard complete-data procedures as if the imputed data were real data obtained from the non-respondents. The results are later combined or pooled to produce estimates and confidence intervals that incorporate missing-data uncertainty. The overarching idea is to use the observed values to provide indirect evidence about the likely values of the unobserved ones.

There are several advantages of the multiple imputation procedure, see for example Rubin (1987), Schafer & Olsen (1998) and Schafer (1999). Key among them being that inferences obtained from MI are generally valid because they account for the uncertainty due to missing data, Schafer (1997). In particular, the multiple imputation procedure is

carried out in a repeated random draws fashion under a model for non-response. Thus (valid) inference that reflect the additional variability due to missing values under that model are obtained by combining complete-data inferences. Hence key to this lies in correctly specifying the imputation model.

There are assumptions, similar to any statistical method, on which the multiple imputation procedure is based. These assumptions pertain to (a) the data model (b) the prior distribution of parameters and (c) the mechanism of the non-response. Common probability models for the data (both observed and missing) range from multivariate normal, log-linear and general location models depending on software packages applied, see Schafer (1997). Since multiple imputation relies on a Bayesian paradigm, a prior distribution for the parameters is required. By default, most software packages utilize the non-informative prior distribution that correspond to a state of prior ignorance about model parameters, Schafer (1997). The Bayesian approach employs the Markov chain Monte Carlo (MCMC) procedure to simulate draws from the posterior distribution of the missing data given the observed data, see Rubin (1987) and Schafer (1999). The multiple imputation procedure assumes that missing data are MAR.

Formally, for the pooling of the estimates from the m multiply data sets and following Rubin (1987), we let θ be the population quantity to be estimated, $\hat{\theta} = \hat{\theta}(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$ denote the statistic that would be used to estimate θ if complete data were available and $U = U(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$ be its variance. In the presence of \mathbf{Y}_{mis} , suppose that we have $m \geq 2$ independent imputations, $\mathbf{Y}_{\text{mis}}^{(1)}, \dots, \mathbf{Y}_{\text{mis}}^{(m)}$, the imputed data estimates $\hat{\theta}^{(l)} = \hat{\theta}(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}^{(l)})$ along with their estimated variances $U^{(l)} = U(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}^{(l)})$, $l = 1, \dots, m$ were calculated. The overall estimate of θ was then given by the average

$$\bar{\theta} = \frac{1}{m} \sum_{l=1}^m \hat{\theta}^{(l)}. \quad (4.1)$$

The standard error of $\bar{\theta}$ was obtained from the estimated total variance given by

$$T = (1 + m^{-1}) B + \bar{U}, \quad (4.2)$$

where B is the between-imputation variance given by

$$B = \frac{\sum_{l=1}^m (\hat{\theta}^{(l)} - \bar{\theta})^2}{m - 1}$$

and \bar{U} is the within-imputation variance given by

$$\frac{\sum_{l=1}^m U^{(l)}}{m}.$$

Tests and confidence intervals were based on a Student's t – approximation

$$\frac{(\bar{\theta} - \theta)}{\sqrt{T}} \sim t_v$$

with degrees of freedom

$$v = (m - 1) \left[1 + \frac{\bar{U}}{(1 + m^{-1}) B} \right]^2$$

The validity of the multiple imputation methods, to give reasonable predictions of the missing data, is dependent on how well the m imputations were generated. Rubin (1987) suggested that the imputations be generated following a Bayesian approach. That is, specify a parametric model for the complete data, apply a prior distribution to the unknown model parameters and simulate m independent draws from the conditional distribution of \mathbf{Y}_{mis} given \mathbf{Y}_{obs} by the Bayes' theorem. Confidence intervals for descriptive population parameters are constructed from the multiple estimates, their standard errors and critical value from the Student's t – distribution as

$$CI(\theta) = \bar{\theta} \pm t_{\tilde{v}_{mi}, 1-\alpha/2} \times SE(\bar{\theta})$$

Here the degrees of \tilde{v}_{mi} are given by

$$\tilde{v}_{mi} = \left(\frac{1}{v_{mi}} + \frac{1}{\hat{v}_{obs}} \right)^{-1}$$

where

v_{mi} = the large sample MI degrees of freedom;

$\hat{v}_{obs} = \left(\frac{v_{com}+1}{v_{com}+3} \right) \times v_{com} \times (1 - \hat{\gamma}_{mi})$ = the degrees of freedom for the complete data;

v_{com} = the degrees of freedom for the complete case analysis;

$\hat{\gamma}_{mi}$ = the estimated fraction of missing information as defined by Rubin (1987).

4.2.3 The analysis model

For the analysis model for the complete case and the m multiple imputation data sets, we considered a multivariable survey logistic regression model from a generalized linear modeling (GLM), by McCullagh & Nelder (1989) framework. Specifically, suppose that Y_i is a binary response variable satisfying the binomial conditions, that is $Y_i \sim Bin(n_i, \pi_i)$ and we let \mathbf{x}_i be a vector of predictor variables related to Y_i and can provide additional information for predicting Y_i . From a GLM perspective, the logistic regression analysis seeks to come up with a model that explains the variation in the probabilities π_i , using the set of predictors in the form

$$\pi(\mathbf{x}_i) = g^{-1}(\mathbf{x}_i' \boldsymbol{\beta}) \quad (4.3)$$

where $\boldsymbol{\beta}$ is a set of parameters to be estimated from the data. Thus by a logit transformation

$$\log(\pi(\mathbf{x}_i)) = \log\left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)}\right) = \mathbf{x}_i' \boldsymbol{\beta} \quad (4.4)$$

Alternatively, Equation 4.4, which is a GLM with a logit link, can be expressed as odds of a positive response as

$$\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} = \exp\{\mathbf{x}'_i\boldsymbol{\beta}\}$$

or as the probability of a positive response as

$$\pi(\mathbf{x}_i) = \frac{\exp\{\mathbf{x}'_i\boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}'_i\boldsymbol{\beta}\}}$$

Under a complex sampling design, the parameters are estimated via a pseudo-likelihood estimation method. Design-based Wald test statistics are used to test the null hypothesis that $\beta_j = 0$ and design-based confidence intervals are provided in the form

$$CI_{1-\alpha}(\beta_j) = \hat{\beta}_j \pm t_{df, 1-\alpha/2} \cdot se(\hat{\beta}_j)$$

Alternatively, the individual predictors can be presented directly on the log-odds scale as $\hat{\psi} = \exp(\hat{\beta}_1)$. In a multivariable logistic regression model adjusted odds ratios of the form $\hat{\psi}_j | \hat{\beta}_{k \neq j} = \exp(\hat{\beta}_j)$ can be given and their respective confidence intervals are given as

$$CI(\psi_j) = \exp\left(\hat{\beta}_j \pm t_{df, 1-\alpha/2} \cdot se(\hat{\beta}_j)\right)$$

4.2.4 The data

With reference to the data description given in section 2.2, the sample consists of 17 434 respondents, 14 491 with non-missing values and an additional 2 943 with missing values in at least one of the measured variables. Table 4.1 gives the variables and their respective percentages of missing values.

Table 4.1: Variables and percentages of missing data

Variable	% Missing Values
HIV Status	15.90
Gender	0.00
Employment Status	1.26
Marital Status	4.37
Contraception	6.21
Wealth Index	4.23
Literacy Level	1.00
Religion	4.08
Educational Level	3.42
Place of Residence	0.99
Province	0.00
Age Group	1.06
Age	1.06

In the ZDHS data, missing data was mainly due to partial or incomplete reporting of information and inconsistent responses to different questions in the survey, see Mutasa (2012). All variables with missing values, as displayed in Table 4.1, were subjected to imputation taking their form (as explained in section 4.2.5 below) into account. The rural areas provided higher coverage of (83%) than urban areas (63%). Females had higher coverage (80%) than their male counterparts (69%) whereas coverage rates varied among provinces from a low of 51% in Bulawayo to 83% in Mashonaland Central. There was no marked coverage variations with regards to variables such as age, literacy levels and wealth status.

4.2.5 Statistical computations

We used the multiple imputation technique described in Section 4.2.2 above to obtain ‘complete’ data for for each of the variables and account for the variability about the missing data. We used the package `mi` in **R** by Gelman et al. (2015) and Su et al. (2011). The package uses a chained equation approach to the imputation, see van Buuren & Groothuis-Oudshoorn (2011). The approach allows specification of the con-

ditional distribution of each variable with missing values conditioned on other variables in the data, and the imputation algorithm sequentially iterates through the variables to impute the missing values using the specified models. This is the so called the fully conditional modelling approach van Buuren & Groothuis-Oudshoorn (2011). Depending on the variable type with missing values, Su et al. (2011) gave examples of conditional distributions. The multiple imputation procedure was performed using Markov chain Monte Carlo (MCMC) methods making use of an iterative data augmentation technique as explained by Schafer & Olsen (1998). In particular, as described by Su et al. (2011), the basic setup of the multiple imputation procedure in **mi** involves three steps; setup, imputation and analysis. The setup step involves a graphical display of missing data patterns, identifying structural problems in the data and pre-processing as well as specifying conditional models.

In the imputation step, the iterative imputation process was carried out based on the conditional models. The **mi** package handles ‘special’ types of variables with missing values as given by Su et al. (2011). With reference to the variables in Table 4.1 above which were used in the imputation model, the package can handle binary variables such as HIV status, place of residence, employment status; ordered categorical variables such as wealth index, literacy level, education level and age group; unordered categorical such as marital status, contraception and religion; and positive continuous such as age. In addition to the main effects we also considered potential interactions that are clinically reasonable and assessed their statistical significance as presented in Hosmer & Lemeshow (2000). Hence we established that there exists an age group by gender interaction effect and it was included in the conditional models. The **mi** package chooses the conditional models automatically according to the variable types identified. In particular, as given in Su et al. (2011) for binary, continuous and ordered categorical, **mi** fits the Bayesian versions of the GLMs (`bayesglm`). These models are slightly different from the classical GLMs in that they add a Student – t distribution on the

regression coefficients. In the current study we used the default Cauchy distribution as recommended by Gelman et al. (2008) as given in Su et al. (2011). Case sampling weights that account for the clustered sample design were included in the conditional models as predictors. Five complete data sets, as suggested in Schafer (1999) were obtained and analyzed separately using design consistent survey logistic regression models as the analysis models with details as given in Section 4.2.3 utilizing the package **survey** by Lumley (2010) in **R**. In addition, the **survey** package allows appropriate parameter estimates and their variance estimates, that account for the complex design, to be computed. We combined or pooled the results together using the formulae provided by Rubin (1987) as explained in Section 4.2.3 above.

4.3 Results

4.3.1 Prevalence estimation results

In this section we present crude design-consistent estimates for HIV prevalence obtained from both a complete case analysis and from the multiple imputed data sets. In the complete case analysis we considered a list-wise deletion of cases with missing values. In the multiple imputation case, the analyses are aimed at accounting for the variability brought about by both the complex sampling design and the imputation process. In particular, the variance estimates were computed in a way that allows reflecting the variability introduced by the imputation process and the variability required to account for the complex sampling design. Five imputed data sets were obtained.

Both approaches gave an overall HIV prevalence of approximately 15.7%. However the complete case analysis gave a standard error of the estimate of HIV prevalence of 0.32% as compared to 0.39% for the imputed case. The larger standard error for the multiple imputation approach correctly incorporate both the between and within imputation variances.

Results of the crude sub-group estimates of HIV prevalence are given in Table 4.2. The results displayed in the table show that the estimates obtained from both the complete case and the imputation are not overall identical. This is because of the additional 2943 cases that the multiple imputations have allowed to enter the analysis. However the differences are not significant as their respective 95% confidence intervals overlap.

Table 4.2: Crude subgroup estimates and their standard errors of HIV prevalence for (a) complete case analysis and (b) multiple imputation.

Variable	(a) Complete case analysis			(b) Multiple imputation		
	Estimate	S. E.	95% CI	Estimate	S. E.	95% CI
Overall	0.154	0.003	(0.147, 0.160)	0.157	0.004	(0.150, 0.164)
Gender						
Male	0.128	0.005	(0.119, 0.137)	0.131	0.004	(0.123, 0.139)
Female	0.177	0.005	(0.166, 0.185)	0.178	0.005	(0.169, 0.188)
Age Group						
15 – 19	0.040	0.004	(0.032, 0.047)	0.041	0.003	(0.035, 0.048)
20 – 24	0.079	0.005	(0.068, 0.089)	0.085	0.005	(0.076, 0.095)
25 – 29	0.158	0.008	(0.142, 0.173)	0.160	0.007	(0.146, 0.174)
30 – 34	0.232	0.010	(0.213, 0.252)	0.233	0.010	(0.214, 0.252)
35 – 39	0.269	0.012	(0.245, 0.292)	0.272	0.013	(0.251, 0.294)
40 – 44	0.255	0.014	(0.228, 0.283)	0.249	0.013	(0.227, 0.272)
45 – 49	0.258	0.015	(0.227, 0.288)	0.265	0.015	(0.236, 0.294)
50 – 54	0.187	0.022	(0.143, 0.231)	0.191	0.020	(0.154, 0.229)
Marital Status						
Single	0.056	0.003	(0.049, 0.063)	0.083	0.003	(0.076, 0.091)
Married	0.167	0.004	(0.159, 0.177)	0.169	0.004	(0.159, 0.179)
Divorced	0.288	0.016	(0.256, 0.319)	0.276	0.012	(0.259, 0.323)
Widowed	0.544	0.022	(0.500, 0.587)	0.551	0.020	(0.510, 0.587)
Wealth Index						
Poorest	0.151	0.007	(0.143, 0.172)	0.159	0.006	(0.142, 0.176)
Poorer	0.158	0.007	(0.132, 0.161)	0.148	0.005	(0.134, 0.162)
Middle	0.146	0.008	(0.149, 0.179)	0.138	0.007	(0.150, 0.187)
Richer	0.163	0.007	(0.146, 0.174)	0.170	0.006	(0.155, 0.184)
Richest	0.142	0.007	(0.126, 0.152)	0.142	0.007	(0.129, 0.154)
Literacy						
Non-lit.	0.139	0.012	(0.115, 0.162)	0.149	0.016	(0.122, 0.176)
Partially	0.198	0.014	(0.172, 0.223)	0.194	0.012	(0.171, 0.217)
Literate	0.151	0.003	(0.144, 0.157)	0.155	0.004	(0.147, 0.162)
Employment						
No	0.135	0.004	(0.128, 0.143)	0.139	0.004	(0.130, 0.147)
Yes	0.173	0.005	(0.163, 0.183)	0.177	0.005	(0.166, 0.187)
Place of Res						
Rural	0.147	0.004	(0.139, 0.154)	0.148	0.004	(0.138, 0.157)
Urban	0.168	0.006	(0.157, 0.180)	0.172	0.005	(0.160, 0.184)

We further present the design-consistent sub-group estimates of HIV prevalence obtained from the complete case analysis and the multiple imputation analysis graphically. Figure 4.1 gives the results for HIV prevalence estimates along with their respective 95%

confidence intervals by marital status of the respondents for the two analyses (complete case and multiple imputation). The results in the graph show that the estimates are significantly different among the single respondents as the confidence bands do not overlap, whereas the rest are not significantly different. Figure 4.2 gives a plot of the estimates of HIV prevalence by five-year age-groups from both the complete case analysis and the multiple imputation analysis. The results show that there are no significant differences in the location of HIV prevalence and width of the 95% confidence intervals across the age groups for the two analyses. Figure 4.3 presents the estimates of HIV prevalence across the country's administrative provinces together with their respective 95% confidence intervals for both analyses. Also all the estimates are not different as the confidence intervals overlap. Notable differences in the width of the confidence intervals can be seen for Bulawayo, Harare, Masvingo, Mashonaland East and Matabeleland North provinces. In particular the confidence intervals for multiple imputation analysis are narrower for Bulawayo, Harare, Mashonaland Central and Masvingo than for complete case analysis.

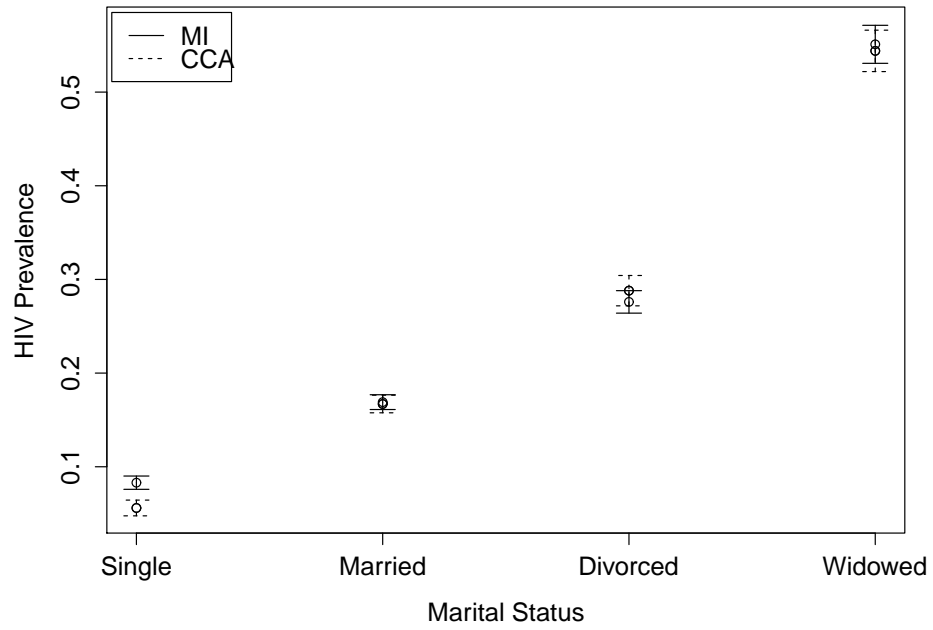


Figure 4.1: Estimates of HIV prevalence by marital status obtained using complete case analysis and multiple imputations

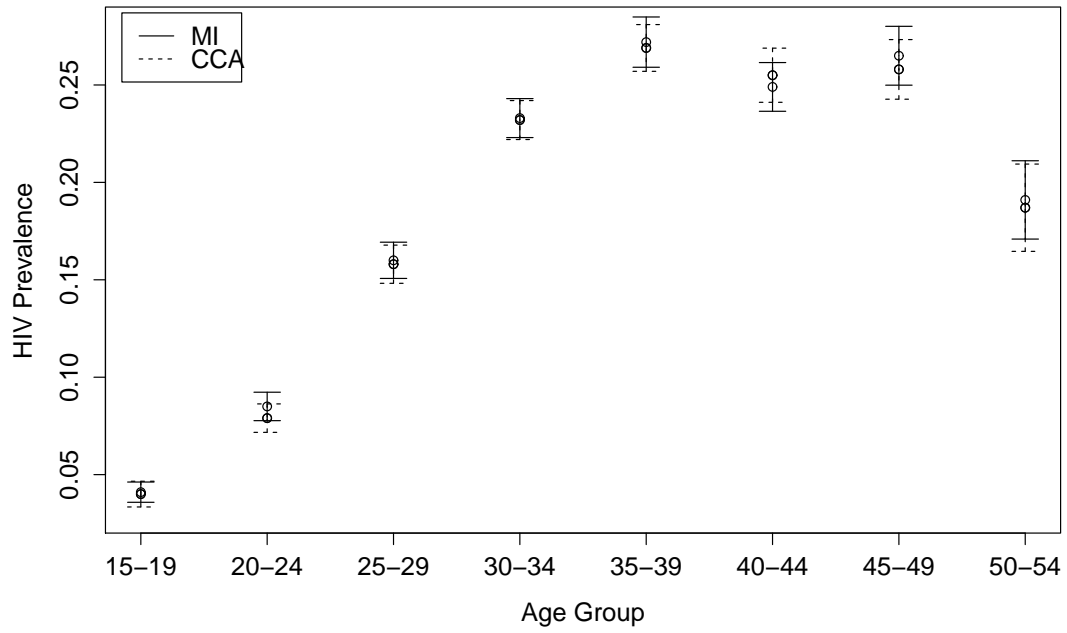


Figure 4.2: Estimates of HIV prevalence per age group obtained using complete case analysis and multiple imputations

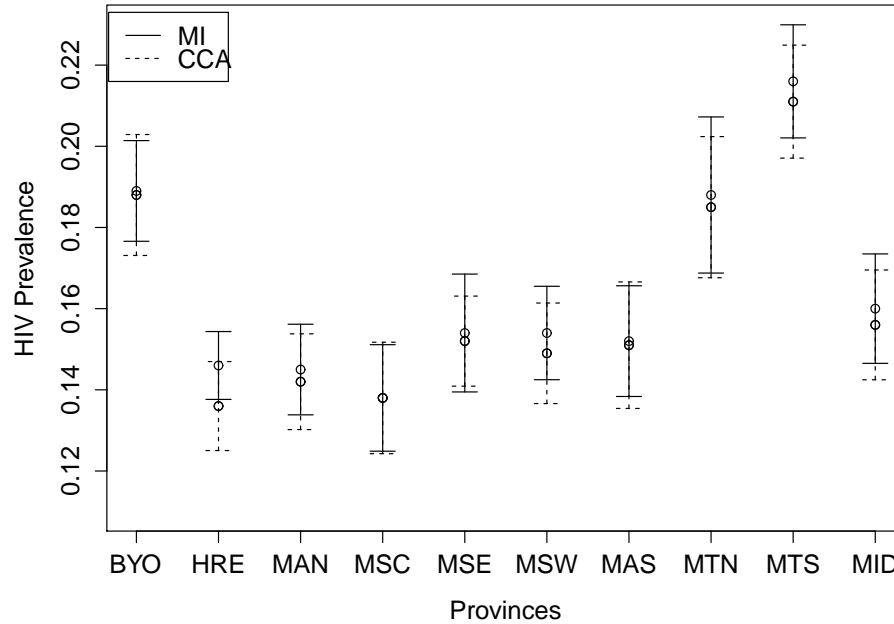


Figure 4.3: Estimates of HIV prevalence by province using complete case analysis and multiple imputations

BYO=Bulawayo, HRE=Harare, MAN=Manicaland, MSC=Mashonaland Central, MSE=Mashonaland East, MSW=Mashonaland West, MAS=Masvingo, MTN=Matabeleland North, MTS=Matabeleland South MID=Midlands

4.3.2 Logistic regression results

We also present the results of a logistic regression model with estimates and their standard errors pooled from the five imputed data sets as well as results from the complete case analysis. Specifically, multivariable survey logistic regression models for explaining the variation in HIV prevalence as a function of demographic and socio-economic variables were fitted. Table 4.3 gives the parameter estimates for the survey logistic regression models obtained from both approaches. The results show considerable variations in the parameter estimates between the two approaches. In particular, the married level of the marital status, the 45 – 49 :M level of the age group by gender interaction effect and the widow:M level of the marital status by gender interaction terms ceased to be statistically non-significant under the complete case analysis to being statistically

significant under multiple imputation. Also the divorced level of the marital status and the literate level of the literacy variable ceased to be significant under the complete case analysis to be non-significant under the multiple imputation. Larger standard errors under the multiple imputation analysis are due to the fact that they represent both between and within imputation variability.

Table 4.3: Parameter estimates and standard errors of logistic regression models under (a) complete case analysis and (b) multiple imputation analysis

Coefficient	(a) Complete case analysis			(b) Multiple imputation		
	Estimate	S. E.	p-value	Estimate	S. E.	p-value
Intercept	-3.450	0.166	< 0.001	-3.251	0.203	< 0.001
Age Group						
20 – 24	0.947	0.166	< 0.001	1.050	0.145	< 0.001
25 – 29	1.636	0.168	< 0.001	1.792	0.153	< 0.001
30 – 34	2.013	0.172	< 0.001	2.192	0.155	< 0.001
35 – 39	1.988	0.178	< 0.001	2.199	0.165	< 0.001
40 – 44	1.533	0.194	< 0.001	1.743	0.185	< 0.001
45 – 49	1.309	0.200	< 0.001	1.631	0.188	< 0.001
50 – 54	1.147	0.265	< 0.001	1.850	0.174	0.041
Gender						
Male	-0.150	0.194	0.439	-0.248	0.190	0.181
Marital Status						
Married	-0.046	0.124	0.708	-0.339	0.097	< 0.001
Divorced	0.676	0.147	< 0.001	0.238	0.181	0.168
Widowed	1.656	0.161	< 0.001	1.306	0.130	< 0.001
Literacy						
Partially	0.511	0.142	< 0.001	0.405	0.162	0.008
Literate	0.244	0.115	0.034	0.171	0.141	0.203
Place of Residence						
Urban	0.222	0.058	< 0.001	0.201	0.062	0.001
Age Group*Gender						
20 – 24 : Male	-1.033	0.275	< 0.001	-0.730	0.250	0.003
25 – 29 : Male	-0.919	0.273	0.001	-0.759	0.233	0.001
30 – 34 : Male	-0.870	0.288	0.003	-0.683	0.273	0.010
35 – 39 : Male	-0.415	0.295	0.160	-0.276	0.244	0.250
40 – 44 : Male	0.138	0.310	0.657	0.188	0.284	0.491
45 – 49 : Male	0.539	0.320	0.092	0.502	0.250	0.045
Marital status*Gender						
Married:M	0.525	0.203	0.010	0.444	0.186	0.003
Divorced:M	0.676	0.263	0.010	0.846	0.290	0.002
Widowed:M	0.654	0.360	0.069	0.638	0.279	0.033

Table 4.4 presents odds ratios for the estimates of the logistic regression model given in Table 4.3 above. Similar to the results in Table 4.3 above, it is also evident that the odds ratios for the complete case analysis and those obtained from the multiple imputation analysis are not identical but are consistent with regards to the significance of the parameters. As expected the results show a similar trend for the ORs of the parameter estimates, with the levels of the variables that ceased to be non-significant under complete case analysis to being significant under the multiple imputation analysis and those that ceased to be significant under complete case analysis to non-significant under the multiple imputation analysis as with the parameter estimates themselves.

Table 4.4: ORs for the estimates of the survey logistic regression models under (a) complete case analysis and (b) multiple imputation analysis and their respective 95% confidence intervals

Parameter	(a) Complete case analysis		(b) Multiple imputation	
	OR	95% CI	OR	95% CI
Intercept	0.032	(0.023, 0.044)	0.038	(0.026, 0.058)
Gender				
Male	0.860	(0.588, 1.259)	0.781	(0.538, 1.133)
Age Group				
20 – 24	2.577	(1.860, 3.571)	2.880	(2.166, 3.830)
25 – 29	5.137	(3.697, 7.139)	6.001	(4.443, 8.107)
30 – 34	7.486	(5.339, 10.497)	8.953	(6.607, 12.133)
35 – 39	7.298	(5.153, 10.336)	9.018	(6.526, 12.461)
40 – 44	4.633	(3.160, 6.792)	5.717	(3.980, 8.212)
45 – 49	3.703	(2.502, 5.481)	5.109	(3.536, 7.382)
50 – 54	3.150	(1.875, 5.292)	6.361	(4.523, 8.946)
Marital Status				
Married	0.955	(0.749, 1.216)	0.713	(0.590, 0.861)
Divorced	1.965	(1.474, 2.621)	1.269	(0.891, 1.808)
Widowed	5.236	(3.821, 7.175)	3.692	(2.862, 4.763)
Literacy				
Partially	1.666	(1.261, 2.201)	1.500	(1.091, 2.062)
Literate	1.276	(1.019, 1.597)	1.187	(0.900, 1.566)
Place of Residence				
	1.249	(1.114, 1.399)	1.223	(1.083, 1.381)
Age Group*Gender				
20 – 24:Male	0.356	(0.208, 0.610)	0.482	(0.295, 0.786)
25 – 29:Male	0.399	(0.234, 0.681)	0.468	(0.297, 0.740)
30 – 34:Male	0.419	(0.238, 0.737)	0.505	(0.296, 0.862)
35 – 39:Male	0.661	(0.371, 1.177)	0.759	(0.470, 1.226)
40 – 44:Male	1.148	(0.625, 2.109)	1.207	(0.692, 2.104)
45 – 49:Male	1.714	(0.915, 3.210)	1.652	(1.013, 2.694)
Marital status*Gender				
Married:M	1.691	(1.137, 2.516)	1.560	(1.083, 2.245)
Divorced:M	1.965	(1.174, 3.291)	2.330	(1.319, 4.116)
Widowed:M	1.923	(0.951, 3.891)	1.593	(1.095, 3.272)

4.4 Discussion

The results for the two approaches presented in Tables 4.2 through to 4.4 are not identical although they are generally consistent pertaining to the statistical interpretation of

the estimates. In particular, the crude estimates of HIV prevalence presented in Table 4.2 show no statistical significant differences between the two approaches. This is particularly so because the respective 95% confidence intervals for the estimates overlap. The results consistently show that the odds of HIV are lower among males ($\hat{p} = 12.8\%$, 95% CI = 11.8–13.7% for the complete case analysis and $\hat{p} = 13.1\%$, 95% CI = 12.3–13.9% for the multiple imputation) than among females ($\hat{p} = 17.7\%$, 95% CI = 16.6–18.7% for the complete case analysis and $\hat{p} = 17.8\%$, 95% CI = 16.9–18.8%). The differences are possibly due to the disparities in susceptibility to HIV between females and males especially in light of HIV infection through unprotected heterosexual intercourse. It has been reported that the risk of transmitting HIV from men to women is much higher than from women to men because women are exposed to considerable amounts of seminal fluids during vaginal sexual intercourse see Myer et al. (2003) and Coombs et al. (2003). Both approaches show a general increase in HIV prevalence with age peaking at the same age group 35–39. HIV prevalence is lowest among the single or never married for both approaches although the difference in the prevalence between the two is statistically significant as the 95% confidence intervals do not overlap. In particular, the prevalence is significantly lower ($\hat{p} = 5.6\%$, 95% CI = 4.9–6.3%) under the complete case analysis than under the multiple imputation ($\hat{p} = 8.3\%$, 95% CI = 7.6–9.1%). The widowed have the highest HIV prevalence for both approaches and there is no statistical significant difference in the prevalence between the two approaches as the 95% confidence intervals overlap. The interpretation of the results is the same for the other risk factors indicated in Table 4.2.

With reference to Table 4.4 both approaches show that the odds of HIV are less among the males (OR = 0.924, 95% CI = 0.631–1.354 under the complete case analysis and OR = 0.812, 95% CI = 0.516–1.175 under the multiple imputation) compared to the females controlling for the other covariates in the model. However both approaches show that the difference in the odds of HIV among males and females

are not significantly different, i.e. H_0 : odds HIV(males) = odds HIV(females). In other words H_0 : OR = 1. The results show that the odds of HIV increase with age for both approaches, however the multiple imputation results show higher odds of HIV at every age group. Relative to the single/never married, the married have slightly higher odds of HIV (OR = 1.182, 95% CI = 0.973–1.437) under the complete case analysis, whereas the married have slightly lower odds of HIV (OR = 0.842, 95% CI = 0.726–0.976) under the multiple imputation controlling for the other covariates in the model. The divorced have odds of HIV over twice higher (OR = 2.575, 95% CI = 1.990 – 3.230) under the complete case, whereas they have odds of HIV less than twice higher (OR = 1.658, 95% CI = 1.238 – 2.220) relative to the single/never married controlling for the other covariates in the model. The interpretations are the same for literacy levels and the place of residence.

The married level of marital status variable ceased to be non-significant under complete case analysis to being significant under multiple imputation whereas the literate level of the literacy variable ceased to be significant under the complete case analysis to being non significant under the multiple imputation analysis. The age by gender interaction effect shows that the risk of HIV is significantly higher, as evidenced by 95% confidence intervals that are not overlapping, in females than in males among the young age groups. However the risk is higher among males in age group 40 – 44 year olds and significantly higher among the 45 – 59 year olds in males than in females. These findings agree with a general perception in most sub-Saharan African countries that younger women engage in sexual activities with older men, a key driver of HIV infection in sub-Saharan Africa.

4.5 Potential strength and limitations of the study

The research draws its strength from the use of the multiple imputation technique to impute missing data in HIV research utilizing the powerful and advanced computational tools that are now available in statistical software such as **R**. Also noting that missing data are inevitable, pervasive and have severe consequences if not properly handled, use of sound statistical methods and computing resources to estimate disease measures of interest and appropriate measures of variability (that account for both the sampling mechanism and the imputation process) can enhance the validity of the statistical interpretations and inferences.

However a potential drawback of the current research comes from the use of secondary data which often leaves the data analyst with limited control over the data collection process. In addition, and particularly for the current research, a major drawback of using secondary is the limited knowledge about the reasons for the missing values. However this is not to downplay the importance of DHSs which are carefully designed, by a team of highly trained statisticians with excellent expertise in survey methodology, to collect population level information which is very important for public health policies. The package **mi**, although very powerful and flexible comes with its own limitations that it cannot allow users to alter the prior distributions for the conditional imputation models used under the Bayesian paradigm. Therefore further methodological and software developments research is necessary in order to make the approach even more flexible. Further work on the problem as a future extension is possible with inclusion of methods that allow for MNAR assumption by means of a sensitivity analysis.

4.6 Conclusion

Analysis of survey data that are characterized by missing data often take a complete case analysis approach where cases with missing values are excluded in the analysis. This often introduces bias in the estimates because of potential loss of information that occurs with the deletion of the cases with missing values. Alternatively, ad hoc approaches based on substituting the missing values with plausible ones such as the last value carried forward, the mean and the regression predictions (single imputations) can be used. However, these approaches may result in potential loss of the distributional relationships among variables and it is not possible to provide measures of uncertainty introduced by the imputation process. Hence we utilized the multiple imputation procedure to ‘fill in’ missing values in estimating HIV prevalence in Zimbabwe using the 2010-11 DHS data while at the same time accounting for the uncertainty about the missing data themselves. Crude design-consistent national and subgroup estimates of HIV prevalence were estimated under both the complete case analysis and the multiple imputation analysis. Survey logistic regression models were also fitted and the results showed considerable variation in the estimates obtained under the two approaches. The results of both the crude estimates and the survey logistic regression model show substantial differences in the estimates and the widths of the confidence intervals between the two approaches.

Chapter 5

Hierarchical logistic regression for estimating risk of HIV using population-based survey data

Abstract

Most practical complex survey data exhibit some multilevel or hierarchical structural form brought about by the prominent features of the sampling design and the underlying target population. These data are often obtained using stratified multistage clustered sampling designs and exhibit a ‘clustered’ or ‘nested’ structure that usually induce intra-class correlations of units within clusters. Appropriate statistical inference and correct conclusions based on such data require methods of analysis that take account of the hierarchical nature of the data. A hierarchical logistic regression model for HIV as a function of demographic, socio-economic and behavioural variables is built from a generalized linear modelling framework. The hierarchical models are capable of capturing the multiple sources of variability brought about by the layered structure of the data and determine how different layers interact and impact a response variable.

The research used data from the 2010-11 Zimbabwe Demographic and Health Surveys (2010-11ZDHS). The data are clustered (by household and enumeration area) and exhibit marked multi-layering, clustering and are characterized by missing observations. The results obtained from a rectangular data set with imputed values are presented together with those from a complete-case approach. The multiple imputation procedure accounted for the structure of the data by incorporating the sampling design features in the conditional models. It was established that there is a considerable household to household and cluster to cluster (enumeration area) heterogeneity of HIV prevalence. Notable differences in estimates of HIV prevalence were also observed between the multiple imputations and the complete-case approaches.

5.1 Introduction

Most practical complex survey data, especially those obtained for scientific and social investigations, often exhibit some multilevel structural form brought about by the prominent features of the underlying target population, see Khan & Shaw (2011). The data are usually obtained by stratified multistage clustered sampling designs that involve application of unequal selection probabilities to the sampling units. These designs often result in data that show a multilevel or hierarchical and nested or clustered structure and are often dependent. The clustered or nested nature of the data induces intra-class correlations among units in the same cluster rendering standard single level statistical methods, that are based on the assumption of independence inappropriate. The multi-layering also results in data with multiple sources of variation. Thus appropriate methods that account for the intra-class correlations are required. We compute hierarchical logistic regression models built from a GLMM framework to simultaneously capture the multiple sources of variability embedded in the multi-layered and clustered data structure. In addition, hierarchical modelling allows assessment of both between

group and within group variability.

A typical hierarchical data structure with three levels is given in Figure 5.1. The units at level one are nested within level two units which are in turn nested within level three units. Units in the same cluster tend to be more homogeneous than from different clusters resulting in cluster induced correlations (intra-class correlation). In addition, the presence of clusters at different levels of the hierarchy also introduces multiple sources of variability. Hence statistical methods for analyzing such clustered and hierarchical data need to take account of the different sources of variability.

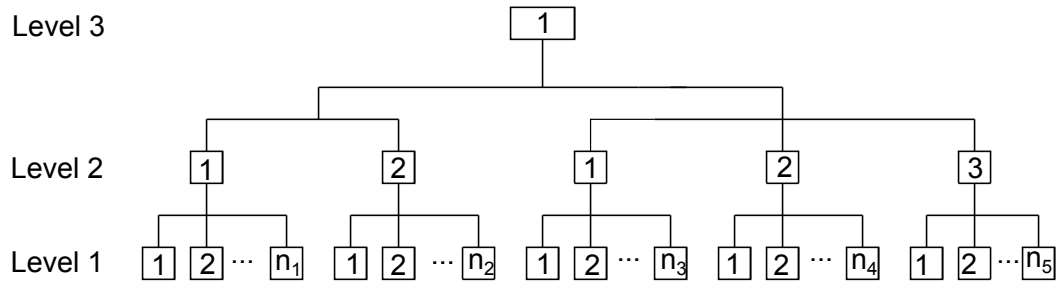


Figure 5.1: Three level hierarchical data structure

Practical instances where hierarchical data are encountered include where students are nested within classes that are nested within schools; individuals can be nested within households that are nested within districts; and patients can be nested within wards, that are nested within hospitals. It is argued that occurrence of data hierarchies in practical instances is neither accidental nor ignorable, Goldstein (2011). Ignoring hierarchical groupings, even though some hierarchies are random, risks overlooking the importance of the group effects.

We develop a hierarchical model for explaining the variation in HIV prevalence, while accounting for the individual level, household level and cluster (enumeration area) variability, in Zimbabwe using data obtained from the 2010-11 Demographic and Health Surveys (2010-11ZDHS). Epidemiological studies have shown that the spread,

and hence the prevalence of HIV has a ‘nesting’ or ‘clustering’ nature. The authors in Ugen et al. (1992) used an example on vertical transmission of HIV, which is essentially MTCT of HIV, in an attempt to bring out the clustering nature of HIV prevalence. Since HIV is spread through among other means, sexual contact, if say a husband contracts the virus, and if no intervention is put in place, automatically the wife is infected too giving data that are correlated. The spread of HIV in a community can be facilitated by social contacts and mixing patterns such as commercial sex activities as well as drug/needle sharing networks, Mossong et al. (2008). Social sexual mixing at the community level such as in mining communities, at growth point, along national roads and in border towns enhances ‘hot spots’ for cases of HIV resulting in clustered effect of HIV, Tanser et al. (2009). This produces patterns of HIV prevalence that are clustered and correlated at different levels giving rise to multilevel or hierarchical data structure, with individuals nested within families and families nested within communities.

Socio-economic differential factors linked to HIV also enhance a hierarchical data structure. HIV is argued to be embedded in social and economic inequality as pointed out by Perry (1998), in that HIV affects those of lower socio-economic status at a disproportionately high rate as compared to those of higher socio-economic status. In Adler (2006) it is pointed out that a lack of socio-economic resources is linked to the practice of riskier health behaviours such as early initiation of sexual activity and less frequent use of condoms. In general, health disparities imply differential disease severity. Hence the existence of a link between socio-economic status and HIV that results in a ‘clustered’ effect of HIV prevalence. Thus poorer communities tend to have higher rates of HIV prevalence than richer well to do communities. HIV/AIDS is also argued to be intertwined in a cause and effect fashion. According to Kedir (2001), HIV/AIDS is a disease of poverty in that poverty pushes men into single-sex migration, women into prostitution and children into under-nutrition.

Research has indicated stark geographical variation in HIV prevalence. The spatial

variation highlights a localized clustering in HIV transmission within micro-epidemics of varying scales and intensity. For example Tanser et al. (2009) and Cuadros et al. (2013) identified spatial clusters with high and low numbers of HIV in sub-Saharan African countries and measure the strength of clustering using a Kulldorff spatial scan statistic analysis under randomness.

Multilevel or hierarchical or random coefficient models are models that are designed to take account of and allow investigations of different sources of variation within and between clusters or within nested settings, and correctly estimate standard errors leading to more accurate inferential decisions, Khan & Shaw (2011). Furthermore, as given by Steenbergen & Jones (2002), hierarchical modelling provides a framework for building of models that capture the layered structure of multilevel data, and determine how different layers interact and impact a response variable of interest. In instances where discrete and/or non-normally distributed response variables are encountered, hierarchical generalized linear models, that make use of a linking function to linear predictor covariates are often used, Goldstein (2011) and Degenholtz & Bhatnagar (2009).

The hierarchical modelling approach encompasses random slopes models that allow the effect of a variable measured at a lower level to vary across different higher level units. They also include fixed effects models with dummy variables included for each higher level unit to capture any possible systematic variation. The hierarchical models also accommodate and allow modelling of differences in how predictors in a regression model influence an outcome of interest across clusters. In addition, hierarchical models provide a convenient framework for the systematic analysis of how covariates and how the interactions among these covariates measured at different levels affect the outcome variable, Guo & Zhao (2000). Estimation of variances and covariances of random effects at various levels, which are embedded well in multilevel models, enable decomposition of the total variance in the outcome variable into portions associated with each level. This in-turn allows measures of the strength of within cluster and between cluster

correlations to be computed.

Missing data that are brought about by non-response are a common occurrence in most practical survey data. There are various reasons why data are missing, see for example Rubin (1987) and Schafer (1997). Proper handling of missing data is key to valid statistical inference and conclusions that are based on applied data analysis. Most analyses of data that have missing observations take a complete-case analysis approach that involve a list-wise deletion of all cases with missing values on the assumption that missing values are missing completely at random (MCAR) as described by Rubin (1987). However these methods have potential drawbacks especially if the MCAR assumption does not hold. Hence we impose the multiple imputation as a method of handling missing data on the modelling approach by systematically substituting each of the missing value with $m \geq 2$ plausible values drawn from distribution of such values using a Bayesian paradigm. The strength of the multiple imputation procedure is in its ability to account for variability introduced by the very process of selecting the values for the missing data point.

The chapter is organized in the following format. Section 5.2 gives the underlying concepts of hierarchical models, generalized linear mixed effects models and the multiple imputations. In addition a description of the data and the statistical computing resources used are also given in this section. Section 5.3 presents the results and a detailed discussion of the findings. Then Section 5.4 gives the concluding remarks.

5.2 Methods

5.2.1 Hierarchical modelling

We consider a data set with three levels: cluster, household and respondent levels. This implies that the measurements on the response variable can be expressed as Y_{kji} , indexed as the i th respondent in the j th household within the k th cluster. Following

the underlying concepts of multilevel modelling given by Goldstein (1986), a full three level hierarchical or multilevel model can then be expressed in a generalized linear modelling (GLM) framework as

$$g [E (Y_{kji})] = \alpha_{kji}^* + \beta_{kj}^* + \gamma_k^*. \quad (5.1)$$

where $g(\cdot)$ is an appropriate link function, α_{kji}^* , β_{kj}^* and γ_k^* are the generic or implicit respondent, household and cluster level terms respectively. We consider the hierarchical GLMs (HGLMs), as defined by Lee & Nelder (1996), for the response conditional on the random effects at each level of the hierarchy. Under the general hierarchical linear modelling approach, a linear model is set up at each level of the hierarchy relating the terms in Equation 5.1 to functions of the predictor variables. In particular, at the cluster level, we have

$$\gamma_k^* = \gamma_0 + \gamma_1 w_{1,k} + \dots + \gamma_q w_{q,k} + v_k = \sum_{l=0}^q \gamma_l w_{l,k} + v_k, \quad (5.2)$$

where v_k is a random variable with $E(v_k) = 0$, $\text{var}(v_k) = \sigma_v^2$, and γ_l is the cluster level coefficient for the l th predictor variable $w_{l,k}$ for cluster k . At the household level we have

$$\beta_{kj}^* = \beta_0 + \beta_{1,k} z_{1,kj} + \dots + \beta_{p,k} z_{p,kj} + u_{kj} = \sum_{l=0}^p \beta_{l,k} z_{l,kj} + u_{kj}, \quad (5.3)$$

where u_{kj} is a random variable with $E(u_{kj}) = 0$, $\text{var}(u_{kj}) = \sigma_u^2$, and $\beta_{l,k}$ is the household level coefficient for the l th explanatory variable $z_{l,kj}$ for the household kj . Then at the individual level, we have

$$\alpha_{kji}^* = \alpha_0 + \alpha_{1,kj} x_{1,kji} + \dots + \alpha_{r,kj} x_{r,kji} + e_{kji} = \sum_{l=0}^r \alpha_{l,kj} x_{l,kji} + e_{kji}, \quad (5.4)$$

where e_{kji} is a random variable with $E(e_{kji}) = 0$, $\text{var}(e_{kji}) = \sigma^2$ and $\alpha_{l,kj}$ is the respondent or individual level coefficient of the l th predictor variable $x_{l,kji}$ for the

respondent kji . Alternatively, Equation 5.1 can be expressed in a more compact form by combining Equations 5.2, 5.3 and 5.4 to give

$$Y_{kji} = \alpha_0 + \gamma_0 + \beta_0 + \sum_{l=0}^r \alpha_{l,kj} x_{l,kji} + \sum_{l=0}^p \beta_{l,k} z_{l,kj} + \sum_{l=0}^q \gamma_l w_{l,k} + (v_k + u_{kj} + e_{kji}). \quad (5.5)$$

Each of the equations at each level of the hierarchy (Equations 5.2 to 5.4) is a mixed effects model hence Equation 5.5 is also a mixed effects model. In particular Equation 5.5 is the explicit hierarchical model with terms explicitly stated. If we assume that all the covariances in Equation 5.5 are zeros, then

$$\text{var}(Y_{kji}) = \sigma_v^2 + \sigma_u^2 + \sigma^2$$

implying that the overall variance of the response can be partitioned into components for the cluster, household and the respondent. Furthermore, it can be shown, see Goldstein (1986), that

$$\text{cov}(Y_{kji}, Y_{kj'i'}) = \sigma_v^2 + \sigma_u^2, \quad (i \neq i')$$

because both respondent i and i' share the same household and cluster. Note that $\text{cov}(Y_{kji}, Y_{kj'i'}) = \sigma_v^2$. Thus the household level intra-class correlation coefficient (ICC), as described by Goldstein (2011) is given by

$$\rho_{kj} = \frac{\sigma_v^2 + \sigma_u^2}{\sigma_v^2 + \sigma_u^2 + \sigma^2}$$

and the cluster level ICC is given by

$$\rho_k = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_u^2 + \sigma^2}$$

The ICC provides a measure of the variability ascribed to a particular level of the hierarchy. A non-zero ICC implies that the observations that share a given hierarchy are not independent. In particular, the ICC is the correlation between two randomly selected units in one randomly selected group or is the fraction of the total variability that is due to group level, Snijders & Bosker (1999), Goldstein & McDonald (1988), Goldstein (1986), Goldstein (2011) and Guo & Zhao (2000). A special function of the ICC under multilevel modelling is for justifying considering an hierarchical model instead of an ordinary linear or logistic regression model.

5.2.1.1 Hierarchical modelling for a binomial response variable

For a binary response variable, Y_{kji} that takes on values 0 or 1, the general modelling approach involves explaining the variation in π_{kji} , the probability of a positive response, $P(Y_{kji} = 1)$, using the explanatory variables. The natural choice of a model for such a binary response variable is a logistic regression model. That is, by a logit transformation, as given by Goldstein (2011), McCullagh & Nelder (1989) and Wong & Mason (1985) a basic logistic regression model expresses the logit as a linear function of the predictors as

$$\text{logit}(\pi_{kji}) = \log\left(\frac{\pi_{kji}}{1 - \pi_{kji}}\right) = \alpha_{kji}^* + \beta_{kj}^* + \gamma_k^* \quad (5.6)$$

Thus the success probabilities can then be expressed as

$$\pi_{kji} = \frac{\exp\{\alpha_{kji}^* + \beta_{kj}^* + \gamma_k^*\}}{1 + \exp\{\alpha_{kji}^* + \beta_{kj}^* + \gamma_k^*\}} \quad (5.7)$$

This formulation demonstrates the connection between hierarchical and logistic regression modelling and that will be the basis on which to enhance building of a hierarchical logistic regression model.

5.2.1.2 Generalized linear mixed effects model

The current research considers a multilevel logistic regression modelling formulated from a generalized linear mixed modelling (GLMM) framework. Essentially, GLMMs extend generalized linear models (GLMs) by the inclusion of random effects in the predictor and more importantly with the specific assumption that the random effects come from a Gaussian distribution. From a GLM perspective, following McCulloch & Searle (2001), Breslow & Clayton (1993) and Doran et al. (2007), we consider a statistical model in which the linear predictor for the response given as, $\eta_i = \mathbf{x}_i\boldsymbol{\beta}$ where \mathbf{x}_i is the i th row of the $n \times p$ model matrix \mathbf{X} , is related to the expected value of the response, μ_i , through a link function, g . That is

$$\eta_i = g(\mu_i) = \mathbf{x}_i\boldsymbol{\beta} \quad i = 1, \dots, n \quad (5.8)$$

and

$$\mu_i = g^{-1}(\mathbf{x}_i\boldsymbol{\beta}) \quad i = 1, \dots, n \quad (5.9)$$

Under a GLMM, the n -dimensional vector of linear predictors, $\boldsymbol{\eta}$, incorporates both fixed and random effects $\boldsymbol{\beta}$ and \mathbf{b} respectively to give

$$\boldsymbol{\eta} = \mathbf{g}(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} \quad (5.10)$$

where \mathbf{Z} is an $n \times q$ model matrix of covariates associated with the random effects. Each component of the random effects vector \mathbf{b} is associated with a level of a grouping or clustering factor. The distribution of the random effects is modelled as a multivariate normal distribution with mean $\mathbf{0}$ and $q \times q$ variance-covariance matrix Σ , that is

$$\mathbf{b} \sim N(\mathbf{0}, \Sigma(\boldsymbol{\theta}))$$

The parameter estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ are those that maximize the marginal likelihood of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ given the observed data. This likelihood is equivalent to the marginal density of \mathbf{y} given $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ given by

$$f(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\theta}) = \int_{\mathbf{b}} p(\mathbf{y}|\boldsymbol{\beta},\mathbf{b}) f(\mathbf{b}|\boldsymbol{\Sigma}(\boldsymbol{\theta})) d\mathbf{b} \quad (5.11)$$

where $p(\mathbf{y}|\boldsymbol{\beta},\mathbf{b})$ is the probability density function of \mathbf{y} , given $\boldsymbol{\beta}$ and \mathbf{b} , and $f(\mathbf{b}|\boldsymbol{\Sigma})$ is the probability density of \mathbf{b} . For parameter estimation for the fixed effects, the likelihood is given by

$$L = \int f(\mathbf{y}|\mathbf{b}) f(\mathbf{b}) d\mathbf{b} \quad (5.12)$$

where the integration is over the q -dimensional distribution of \mathbf{b} . The log-likelihood for the fixed effects is given by

$$l = \log \int f(\mathbf{y}|\mathbf{b}) f(\mathbf{b}) d\mathbf{b}. \quad (5.13)$$

It can be shown, see McCulloch & Searle (2001) that the derivative of the log-likelihood with respect to the fixed effects is given by

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \int \frac{\partial \log f(\mathbf{y}|\mathbf{b})}{\partial \boldsymbol{\beta}} f(\mathbf{b}|\mathbf{y}) d\mathbf{b} \quad (5.14)$$

Using the notation for an exponential family of distributions given by McCullagh & Nelder (1989), Equation 5.14 can be written as

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \mathbf{X}'\mathbf{E}[\mathbf{W}^*|\mathbf{y}] - \mathbf{X}'\mathbf{E}[\mathbf{W}^*\boldsymbol{\mu}|\mathbf{y}] \quad (5.15)$$

where $\mathbf{W}^* = \{[a(\phi)v(\boldsymbol{\mu})g(\boldsymbol{\mu})^{-1}]\}$. Here $a(\cdot)$ is a known function of a dispersion parameter ϕ , $v(\boldsymbol{\mu})$ is the variance function and $g(\boldsymbol{\mu})$ is a known link function. Hence

the likelihood equation for $\boldsymbol{\beta}$ is given by

$$\mathbf{X}'\mathbf{E}[\mathbf{W}^*|\mathbf{y}] = \mathbf{X}'\mathbf{E}[\mathbf{W}^*\boldsymbol{\mu}|\mathbf{y}] \quad (5.16)$$

Solving 5.16 involves iterative algorithms such as the expectation-maximization and the Laplace's approximation, see McCulloch & Searle (2001).

The parameters for the random effects, $\boldsymbol{\theta}$, are derived from $f(\mathbf{b})$ as

$$\frac{\partial l}{\partial \boldsymbol{\theta}} = \int \frac{\partial \log f(\mathbf{b})}{\partial \boldsymbol{\theta}} f(\mathbf{b}|\mathbf{y}) d\mathbf{b} \quad (5.17)$$

Further simplification require the form of the distribution of the random effects.

A drawback for the likelihood estimation method is that a closed form solution is not available, say when $p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{b})$ Equation 5.11 is binomial. Therefore a Laplace's approximation is used to evaluate high-dimensional integrals in the likelihood, see Doran et al. (2007). Essentially the Laplace's approximation is based on a second-order Taylor series expansion as given by McCulloch & Searle (2001). The basic idea is that, given values of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, the conditional modes of the random effects are determined by the penalized iteratively re-weighted least squares algorithm. Under this algorithm, the contribution of the parameters, $\boldsymbol{\beta}$, is incorporated as an offset, $\mathbf{X}\boldsymbol{\beta}$, and the contribution of the variance component $\boldsymbol{\theta}$ is incorporated as a penalty term in the weighted least squares fit.

An alternative to the Laplace's approximation is the expectation-maximization (EM) algorithm, which is an iterative method for finding maximum likelihood estimates when there are missing or unobserved data. The algorithm works by declaring \mathbf{b} to be missing data so that the 'complete data' are $\mathbf{M}' = (\mathbf{y}', \mathbf{b}')$. Then the EM proceeds by forming the log-likelihood of the complete data, calculating its expectation with respect to the conditional distribution of \mathbf{b} given \mathbf{y} and maximizing with respect to the parameters of interest. Since the algorithm is iterative, it is also called the Monte Carlo

expectation-maximization (MCEM).

Under the GLMM, test of hypotheses is based on the usual large-sample arguments. The likelihood ratio test for nested models is based on comparing $-2\log\Lambda$ to the chi-square distribution and Akaike information criterion (AIC), with a goal of selecting the most parsimonious model. Testing the significance of individual parameters is based on Wald test utilizing sample normality of the estimators, McCulloch & Searle (2001).

5.2.2 Handling missing data via multiple imputation

5.2.2.1 Types of missingness

Missing data are classified according to the relationship between measured variables and the probability of missing data in what Rubin (1987) and Little & Rubin (1987a) termed “missing data mechanisms”. Missing data can fall into one of three missing data mechanisms namely missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

Suppose $\mathbf{Y} = \{\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}\}$, where \mathbf{Y}_{obs} are the observed values and \mathbf{Y}_{mis} are the unobserved values and let \mathbf{M} be a missing data indicator matrix of the same dimension as \mathbf{Y} where the value in row i and column j is equal to 1 if the value in \mathbf{Y} is missing and 0 if the value is observed. Data are MCAR if $P(\mathbf{M}|\mathbf{Y}) = P(\mathbf{M})$ for all \mathbf{Y} , that is, the fact that the data are missing is not dependent on any observed values or potential unobserved values for any of the variables in \mathbf{Y} . This means that the probability that a respondent does not report an item value is completely independent of the true underlying values of all the observed and unobserved variables, Heeringa et al. (2010). Missingness is completely unsystematic and the observed data can be regarded as a random sub-sample of the hypothetically complete data. Thus inference can be carried out with the observed data since they are representative of the complete sample as well as the target population.

Missing data are MAR if missingness is related to other measured or observed vari-

ables in the analysis, but not to the underlying values of the incomplete variable, that is the hypothetical values that would have resulted had the data been complete, Baraldi & Enders (2010). Thus MAR implies that $P(\mathbf{M}|\mathbf{Y}) = P(\mathbf{M}|\mathbf{Y}_{\text{obs}})$ for all \mathbf{Y} . The response mechanism responsible for MCAR and MAR is termed ignorable, Rubin (1987) and Pigott (2001).

Missing data are MNAR if they are neither MCAR nor MAR, that is if the missing data are not at least MAR. Missing data are MNAR if missingness depends on both the observed and unobserved values of \mathbf{Y} , that is $P(\mathbf{M}|\mathbf{Y}) = P(\mathbf{M}|\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$. The MNAR mechanism is also called a non-ignorable missing data mechanism.

5.2.2.2 The multiple imputation procedure

Multiple imputation was originally proposed by Rubin (1987), and is a Monte Carlo (or simulated based) technique formulated to replace each missing value with two or more plausible values. Essentially each missing value is imputed m (≥ 2) different times using the same imputation method creating m complete data sets free of missing values. Each completed data set is analyzed using standard complete-data procedures as if the imputed data were observed data obtained from the non-respondents. The overarching idea is to use the observed values to provide indirect evidence about the likely values of the unobserved ones. Thus MI is based on the MAR ignorable assumption.

Essentially, the multiple imputation procedure is carried out in three steps: 1) imputing data under an appropriate model to obtain m ‘filled in’ data sets; 2) analyze each data set separately to obtain desired parameter estimates and standard errors. The estimates obtained are called multiply-imputed estimates, Rubin (1987), Pigott (2001) and Schafer & Olsen (1998). 3) combining the results of the analyses from the m data sets by finding the mean of the m parameter estimates and a variance estimate that accounts for the within-imputation and across-imputation variability using formulae given by Rubin (1987). At the pooling stage, the m analyses are combined to produce

unified estimates and confidence intervals that incorporate missing-data uncertainty.

There are several advantages of MI methods, see Rubin (1987), Schafer & Olsen (1998) and Schafer (1999). Key among them being that inferences obtained from MI are generally valid because they allow accounting for the uncertainty due to missing data, Schafer (1997). In particular, the MI procedure is carried out in a repeated random draws fashion under a model for the non-response. Thus (valid) inference that reflect the additional variability due to missing values under that model are obtained by combining complete-data inferences. Key to this lies in correctly specifying the imputation model.

Formally, following Rubin (1987), let θ be a population quantity to be estimated, $\hat{\theta} = \hat{\theta}(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$ denote the statistic that would be used to estimate θ if complete data were available and $U = U(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$ be its variance. In the presence of \mathbf{Y}_{mis} , suppose that we have $m \geq 2$ independent imputations, $\mathbf{Y}_{\text{mis}}^{(1)}, \dots, \mathbf{Y}_{\text{mis}}^{(m)}$, one can calculate the imputed data estimates $\hat{\theta}^{(l)} = \hat{\theta}(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}^{(l)})$ along with their estimated variances $U^{(l)} = U(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}^{(l)})$, $l = 1, \dots, m$. The overall estimate of θ is given by the average

$$\bar{\theta} = \frac{1}{m} \sum_{l=1}^m \hat{\theta}^{(l)}. \quad (5.18)$$

The standard error of $\bar{\theta}$ is obtained from the estimated total variance given by

$$T = (1 + m^{-1}) B + \bar{U}, \quad (5.19)$$

where B is the between-imputation variance given by

$$B = \frac{\sum_{l=1}^m (\hat{\theta}^{(l)} - \bar{\theta})^2}{m - 1}$$

and \bar{U} is the within-imputation variance given by

$$\bar{U} = \frac{\sum_{l=1}^m U^{(l)}}{m}.$$

Tests and confidence intervals are based on a Student's t – approximation

$$\frac{(\bar{\theta} - \theta)}{\sqrt{T}} \sim t_v$$

with degrees of freedom

$$v = (m - 1) \left[1 + \frac{\bar{U}}{(1 + m^{-1})B} \right]^2$$

The validity of the MI, to give reasonable predictions of the missing data, is dependent on how well the m imputations were generated. Rubin (1987) suggested that the imputations be generated following a Bayesian approach. That is, specify a parametric model for the complete data, apply a prior distribution to the unknown model parameters and simulate m independent draws from the conditional distribution of \mathbf{Y}_{mis} given \mathbf{Y}_{obs} . The current research regards the hierarchical logistic regression model as the substantive analysis model.

5.2.3 The Data

Further to the data description given in section 2.2 the full sample consists of 17 434 respondents, 14 491 with non-missing values and 2 943 with missing values for at least one measured variable. Table 5.1 gives the multilevel structure of the data for the cases with non-missing observations and Table 5.2 gives the variables and their respective percentages of missing values. The data depict an hierarchical structure in that individuals are nested within households and households are nested within EAs (clusters). For the current research, the response variable is HIV status, a binary variable taking on either positive or negative. The explanatory (covariates) variables at individual

level (level 1) are gender, age, education level, literacy level, employment status, place of residence and marital status. At the household level (level 2) the explanatory variables considered are wealth index and religion. The explanatory variables considered are those factors that are known to explain the variation in an individual's HIV status as informed by the proximate-determinants conceptual framework as explained by Boerma et al. (2003).

Table 5.1: Hierarchical structure of the 2010-11 ZDHS data

	Provinces Strata		Level 3 EAs (PSUs)		Level 2 HHs (SSUs)		Level 1 Individuals
2010-11 ZDHS Data	Bulawayo	—	43	—	558	—	942
	Harare	—	56	—	956	—	1801
	Manicaland	—	46	—	915	—	1650
	Mash Central	—	34	—	782	—	1550
	Mash East	—	38	—	834	—	1457
	Mash West	—	40	—	832	—	1630
	Masvingo	—	38	—	775	—	1249
	Mat North	—	36	—	699	—	1207
	Mat South	—	35	—	764	—	1326
	Midlands	—	40	—	829	—	1679
		Total 18 Units		406		7944	

Table 5.2: Percentages of missing data per variable

Variable	% Missing Values
HIV Status	15.90
Gender	0.00
Employment Status	1.26
Marital Status	4.37
Contraception	6.21
Wealth Index	4.23
Literacy Level	1.00
Religion	4.08
Educational Level	3.42
Place of Residence	0.99
Province	0.00
Age Group	1.06
Age	1.06

5.2.4 Statistical Software

We construct three level hierarchical logistic regression models using the **lme4** package by Bates et al. (2014) in **R**. In particular we utilized the function **glmer** to compute a hierarchical logistic regression model, from a GLMM perspective, for HIV on the demographic and socio-economic variables. The function allows specification of the link function such as the logit. Forward selection backward elimination model building strategy was employed to select predictor variables. Testing for the significant difference in nested models was done via assessing the difference in the likelihoods (using the deviances) with the G^2 test based on comparing $-2\log\Lambda$ to the chi-square distribution. The package accommodates multiple nested levels of variability by incorporating random effects at each level of the hierarchy.

In addition, we used the package **mi** in **R** by Gelman et al. (2015) to carry out the multiple imputation procedures. The package uses a chained equation approach to execute the imputations, see van Buuren & Groothuis-Oudshoorn (2011), using Markov chain Monte Carlo (MCMC) with application of an iterative data augmentation technique as explained by Schafer & Olsen (1998). The approach allows specification of

the conditional distribution of each variable with missing values conditioned on other variables in the data, and the imputation algorithm sequentially iterates through the variables to impute the missing values using the specified models. This is the so called fully conditional specification (FCS) modelling approach van Buuren & Groothuis-Oudshoorn (2011). Depending on the variable type with missing values, Su et al. (2011) gave examples of conditional distributions. For instance with variables that are continuous, binary and counts a Bayesian version of the GLMs with a student- t prior distribution are fitted. For the current research the prior distribution we used focuses on the Cauchy distribution with center 0, degree of freedom 1 and scale 2.5.

In particular, as described by Su et al. (2011), the basic structure of the multiple imputation procedure in **mi** package involves three steps after setup. These are the imputation, analysis and combining or pooling steps. The setup step involves a graphical display of missing data patterns, identifying structural problems in the data and pre-processing as well as specifying conditional models. In the imputation step, the iterative imputation process is carried out based on the conditional models. After obtaining the m complete data sets from the imputation step each of them is analyzed using the analysis model (the hierarchical logistic regression model in the current chapter) to answer the scientific question of interest. Finally, the m analyses are pooled together using the formulas provided by Rubin (1987) in the combining step to yield estimates appropriate for desired inference and measures of within and between imputation variability is provided.

5.3 Results and discussion

Separate hierarchical logistic regression models were fitted using both a complete case analysis and a multiple imputation analysis. The fixed effects considered are the individual level demographic and socio-economic variables presented in Table 5.2 above.

Since our response variable, HIV status is binary, a generalized linear mixed model with a binomial logit link was used. Household and cluster level random effect terms were included to account for variability at these higher levels of the hierarchy. The ‘best’ model obtained using the forward selection backward elimination method constructed for the complete case analysis is given. For this current chapter, under the multiple imputation analysis five data sets were imputed using the **mi** package in **R**. Table 5.3 gives the parameter estimates, standard errors and the p-values for the generalized linear mixed effects models obtained using each of the two approaches.

Table 5.3: Parameter estimates, standard errors and p-values for the generalized linear mixed models under (a) multiple imputation analysis (b) complete case analysis

	(a) Multiple imputation analysis			(b) Complete case analysis		
	Estimate	S.E.	p-value	Estimate	S. E.	p-value
<i>Fixed Effects</i>						
Intercept	-3.556	0.178	<0.001	-4.119	0.247	<0.001
Gender						
Male	-0.288	0.220	0.162	-0.348	0.194	0.073
Age group						
20 – 24	1.070	0.155	<0.001	0.747	0.170	<0.001
25 – 29	1.899	0.159	<0.001	1.543	0.179	<0.001
30 – 34	2.365	0.175	<0.001	2.090	0.193	<0.001
35 – 39	2.348	0.182	<0.001	1.878	0.195	<0.001
40 – 44	1.826	0.183	<0.001	1.479	0.213	<0.001
45 – 49	1.692	0.190	<0.001	1.204	0.222	<0.001
50 – 54	1.118	0.197	0.043	1.260	0.303	<0.001
Marital Status						
Married	-0.325	0.129	0.004	0.051	0.143	0.721
Divorced	0.434	0.316	0.149	1.076	0.187	<0.001
Widowed	1.708	0.184	<0.001	2.633	0.254	<0.001
Place of Residence						
Urban	0.190	0.084	0.012	0.203	0.084	0.015
Age*Gender						
20 – 24:Male	-0.740	0.255	0.003	-0.765	0.285	0.007
25 – 29:Male	-0.687	0.262	0.008	-0.810	0.297	0.006
30 – 34:Male	-0.666	0.327	0.024	-0.830	0.315	0.008
35 – 39:Male	-0.221	0.283	0.414	-0.055	0.324	0.864
40 – 44:Male	0.487	0.329	0.112	0.682	0.344	0.048
45 – 49:Male	0.637	0.291	0.026	0.924	0.366	0.012
Age*Mar. Sta.						
Married:Male	0.420	0.186	0.016	0.538	0.231	0.020
Divorced:Male	0.751	0.377	0.045	0.646	0.325	0.047
Widowed:Male	0.526	0.371	0.169	0.800	0.487	0.100
<i>Random Effects</i>						
Residual	0.461	0.281		0.875	0.247	
Household	1.375	1.086		3.805	1.174	
Cluster	0.206	0.329		0.034	0.351	

The results, displayed in Table 5.3, show that the estimates obtained from the two approaches are not necessarily identical but are in most of the cases consistent. Notable differences are observed in the marital status variable where the married level effect is

significant under multiple imputation but non-significant under complete case analysis, the divorced level effect is non-significant under the multiple imputation but significant under the complete case analysis. The 40 – 44 : Male level effect of the age by gender interaction effect is non-significant under the multiple imputation but significant under the complete case analysis.

Noting that the model presented in Table 5.3 is for a binomial response with a logit link, the coefficients generate log-odds for a positive response. Hence interpretation of the coefficients take account that slopes or differences in factor levels are with respect to the logit or log-odds function, that is, the marginal log-odds. To facilitate the interpretation of the coefficients, we present the odds ratios of a positive response, that is of being HIV positive, in relation to the reference level for each coefficient. The results are displayed in Table 5.4.

Table 5.4: Odds ratios and their corresponding 95% confidence intervals for the best models under multiple imputations and complete case analysis

Parameter	(a) Multiple Imputation		(b) Complete Case analysis	
	OR	95% CI	OR	95% CI
Intercept	0.029	(0.020, 0.040)	0.016	(0.010, 0.026)
Gender				
Male	0.750	(0.487, 1.155)	0.706	(0.483, 1.033)
Age group				
20 – 24	2.917	(2.154, 3.950)	2.111	(1.513, 2.945)
25 – 29	6.682	(4.898, 9.118)	4.679	(3.294, 6.645)
30 – 34	10.649	(7.559, 15.002)	8.085	(5.538, 11.802)
35 – 39	10.461	(7.322, 14.948)	6.540	(4.463, 9.585)
40 – 44	6.209	(4.338, 8.888)	4.389	(2.891, 6.662)
45 – 49	5.429	(3.740, 7.881)	3.333	(2.157, 5.151)
50 – 54	3.060	(2.078, 4.502)	3.525	(1.947, 6.385)
Marital Status				
Married	0.723	(0.562, 0.930)	1.052	(0.792, 1.398)
Divorced	1.543	(0.830, 2.869)	2.933	(2.033, 4.231)
Widowed	5.520	(3.850, 7.913)	13.915	(8.458, 22.893)
Place of Res.				
Urban	1.209	(1.026, 1.425)	1.225	(1.039, 1.444)
Age*Gender				
20 – 24:Male	0.477	(0.289, 0.786)	0.465	(0.266, 0.814)
25 – 29:Male	0.503	(0.301, 0.840)	0.445	(0.249, 0.796)
30 – 34:Male	0.514	(0.271, 0.974)	0.436	(0.235, 0.808)
35 – 39:Male	0.801	(0.460, 1.396)	0.946	(0.502, 1.786)
40 – 44:Male	1.627	(0.854, 3.100)	1.978	(1.008, 3.882)
45 – 49:Male	1.891	(1.069, 3.346)	2.519	(1.230, 5.162)
Age*Mar. Sta.				
Married:Male	1.522	(1.057, 2.191)	1.713	(1.089, 2.693)
Divorced:Male	2.120	(1.012, 4.437)	1.908	(1.009, 3.607)
Widowed:Male	1.692	(0.818, 3.501)	2.226	(0.857, 5.781)

The results show that males have lower odds of HIV than females, for instance for the multiple imputation (OR = 0.750, 95% CI = 0.487–1.155) and for the complete case analysis, (OR = 0.706, 95% CI = 0.483 – 1.033). However, the difference in the odds of HIV between males and females is not statistically significant as the confidence intervals contain a 1. Relative to the 15–19 age group, the odds of HIV increase with age, peaking at the 30 – 34 for both approaches; OR = 10.649, 95% CI = 7.559 – 15.002, for the

multiple imputation and $OR = 8.085$, $95\% CI = 5.538 - 11.802$ for the complete case analysis. Under the multiple imputation, with reference to the single/never married the odds of HIV are lower among the married ($OR = 0.723$, $95\% CI = 0.562 - 0.930$), and is higher among the divorced ($OR = 1.543$, $95\% CI = 0.830 - 2.869$) and the widowed ($OR = 5.520$, $95\% CI = 3.850 - 7.913$). However the odds of HIV for the married ($OR = 1.052$, $95\% CI = 0.792 - 1.398$), the divorced ($OR = 2.933$, $95\% CI = 2.033 - 4.231$) and the widowed ($OR = 13.915$, $95\% CI = 8.458 - 22.893$) are higher than the single/never married under the complete case analysis. The interpretation is similar for the place of residence variables. The age group by gender interaction indicates the additional effect of gender on age group in influencing the odds of HIV for a given level relative to the 15 – 19 year old females. In particular for the multiple imputation case, the odds of HIV are lower for the 20 – 24 year old males ($OR = 0.477$, $95\% CI = 0.289 - 0.786$), for the 25 – 29 year old males, ($OR = 0.503$, $95\% CI = 0.301 - 0.840$) and for the 30 – 34 year old males ($OR = 0.514$, $95\% CI = 0.271 - 0.974$) relative to the 15 – 19 year old females. However the odds of HIV are higher for the 40 – 44 year old males ($OR = 1.627$, $95\% CI = 0.854 - 3.100$) and for the 45 – 49 year old males ($OR = 1.891$, $95\% CI = 1.069 - 3.346$) as compared to the 15 – 19 year old females. Overall for the gender by age group interaction, there is an increasing effect of gender on age group on the odds of HIV among the young females as compared to the young males whereas there is an increasing effect of gender on age group on the chances of being HIV positive among the older males as compared to older females. Essentially, this implies that young females are at a higher risk of being HIV positive than young males whereas older males are at a relatively higher risk of being HIV positive compared to older females. This is in agreement with a general belief in the most sub-Saharan African countries that young females engage in sexual relationships with older men. There is a slight increasing effect of gender on marital status on the odds of HIV, for instance under the multiple imputation case for the married ($OR = 1.522$, 95%

CI = 1.057 – 2.191), divorced (OR = 2.120, 95% CI = 1.012 – 4.437) and widowed (OR = 1.692, 95% CI = 0.818 – 3.501) as compared to the single females. The results for the complete case analysis can be interpreted in a similar way.

For the random effects, the results give the household level and the cluster level variabilities. In particular the household to household variability accounts for greater variation in HIV as compared to cluster level for both approaches. This is indicated by larger values for the variance components for the household level, ($\sigma_u^2 = 1.375$), than for the cluster level ($\sigma_v^2 = 0.206$) displayed in Table 5.3 above. The corresponding estimated intra-household correlation coefficient is $(1.375 + 0.206)/(1.375 + 0.206 + 0.461) = 0.774$ whereas the intra-cluster correlation coefficient is $0.206/(1.375 + 0.206 + 0.461) = 0.101$. This implies that the outcomes or responses within a household are more strongly correlated than those from two different households. This in a way makes sense since in Africa HIV transmission tends to be (spatially) homogeneous. Thus if one of the couple member living together is infected there is a high probability that the other member is also infected. Both the intra-household and intra-cluster correlation coefficients are nonzero which justifies the use of the multilevel approach to the analysis.

5.4 Conclusion

The analysis of survey data that depict a hierarchical structure need to account for the variability induced by the multi-layering and the clustering of the data that results from the prominent features of the underlying population. We computed hierarchical or multilevel logistic regression models for HIV on demographic and socio-economic variables using a complete case analysis and multiple imputation analysis. The multi-level modelling approach provides a framework for accounting for the different sources of variability in the response that are nested within the different layers of the hierarchy. It was established that HIV prevalence depends on an individual's age, gender,

marital status, place of residence and age-gender and gender-marital status interaction effects. The results also showed a substantial household-to-household variability and a relatively small variability in HIV prevalence hence a high intra-household correlation coefficient.

Chapter 6

A Bayesian logistic regression for estimating risk of HIV using population-based survey data

Abstract

Statistical models that incorporate prior known information about the unknown model parameters are vital in scientific and health research especially in studies where replicative experimental investigations are not possible. The Bayesian statistical paradigm is designed to allow for combining prior knowledge about model parameters with the appropriate likelihood of the observed data to obtain a posterior distribution. Under the Bayesian framework, likelihood based methods are often used for parameter estimation whilst statistical inference is carried out based on the posterior distribution. Computer-intensive simulation-based algorithms such as the Markov chain Monte Carlo (MCMC) methods are then used to draw samples from the posterior distribution to be used for the statistical inference purposes. Diagnostics in the form of trace plots and Geweke plots together with the Hiedelberger test for stationarity are used to assess convergence,

which is a necessary requirement of the Markov chains.

There has been a host of prior knowledge about HIV/AIDS that can be combined with the likelihood of the observed data to enhance explaining the variation of HIV prevalence. Use of population-based survey data also facilitates linking of HIV to demographic, socio-economic and behavioural factors of the respondents. A Bayesian logistic regression model is fitted from a generalized linear modelling (GLM) perspective for HIV on demographic and socio-economic factors using the 2010-11 Zimbabwe Demographic and Health Surveys (2010-11 ZDHS) data. A non-informative t -family Cauchy prior distribution was utilized for the unknown model parameters. It was established that HIV prevalence is dependent on one's gender, age, marital status, place of residence, literacy level and the age by gender and gender by marital status interaction effects.

6.1 Introduction

Statistical models that incorporate prior known information about the unknown parameters are vital in scientific and health research. Incorporation of such prior knowledge into a statistical analysis of HIV has the potential to enhance the quality of the statistical results. The Bayesian approach to statistical analysis allows the incorporation of prior knowledge about the parameters often expressed as distributions. Specifically the Bayesian framework works by combining prior information about unknown model parameters with the appropriate likelihood of the observed data to give a posterior distribution.

The fundamental ideas as described in Bolstad (2007), Press (1989), Raiffa & Schlaifer (1961), Bernardo & Smith (1994), Lesaffre & Lawson (2012) and Gill (2009) that form the basis of the Bayesian analysis framework are:

- The unknown model parameters are considered to be random variables and hence

are specified by prior distributions.

- Probability statements are interpreted as measures of ‘degree of belief’.
- The Bayes’ theorem, that underlies the Bayesian analysis, is used to revise the beliefs about the parameters in light of the observed data to obtain the posterior distribution. The posterior distribution gives the relative weights to each parameter value after analyzing the sample data.

The relationship between the prior distribution, the observed data and the posterior distribution is expressed as $posterior = prior \times likelihood$, see Gill (2009), Bolstad (2007) and Press (1989). The Bayesian approach has a number of attractive features particularly as compared to the frequentist for statistical analysis. These features include the that the approach allows a consistent way to modify one’s belief about the parameters given the data that actually occurred, implying that inference is based on actual data not on all possible data sets that might have occurred (as in the frequentist approach) as presented by Rao (2011), Raiffa & Schlaifer (1961) and Bolstad (2007). Allowing the parameter to be a random variable enables one to make probability statements about it (the parameter) *a posteriori* to observing the data. However, many practical analyses utilize both approaches in a complementary way in which design-based inferences can be derived from the Bayesian perspective, using frequentist models with non-informative (as defined below) prior distributions.

The debates and disagreements around the merits and/or the demerits of the frequentist and the Bayesian approach to statistical analysis emanate mainly from the differing fundamental interpretations of probability. As described in Bolstad (2007), frequentists define probability as long-run tendencies of events that eventually converge on some true population proportion whereas Bayesians interpret probability as “degree of belief”. The Bayesian philosophy implies that prior distributions are descriptions of relative likelihoods of events based on past experience, personal intuition or expert opin-

ion, and posterior distributions are those prior distributions updated by conditioning on new observed data.

Central to the Bayesian paradigm is the choice of priors (see for example Press (1989) and Lesaffre & Lawson (2012)) with an option of choosing between conjugate, informative and non-informative. However the subjectivity surrounding the choice of a model and the prior, and the problems stemming from model mis-specification provide a basis for most of the critics for Bayes methods, Raiffa & Schlaifer (1961) and de Finetti (1937).

The current chapter focuses on a logistic regression model for HIV fitted from a GLM perspective utilizing the Bayesian statistical analysis framework. Essentially we constructed the model with the assumption that the parameters are random variables that have to be assigned a probability distribution, that is, the prior distribution and obtain the likelihood of the observed data. The approach enables utilization of the knowledge about HIV obtained from past studies as prior information together with the likelihood of the observed data.

It is argued that the Bayesian analysis paradigm works well over a frequentist approach in biological, health and social science research, see for example Bolstad (2007) and Press (1989). This is mainly due to the availability of immense prior data information on most of the phenomena in these fields. For instance, information such as susceptibility variations to HIV among males and females, between urban and rural residents, between the literate and the non-literate and across different marital statuses. Furthermore, most phenomena do not allow the replicative experimental nature of the randomization process responsible for the stochastic data generating mechanism, which is the cornerstone of the frequentist approach.

The chapter is organized as follows. Section 6.2 presents the fundamental theory underlying the Bayesian statistical analysis framework in general and the Bayesian logistic regression in particular. Descriptions of the data and statistical computing

resources used are also given in this section. Section 6.3 gives the results and discussion of the results. Concluding remarks are given in Section 6.4.

6.2 Methods

6.2.1 An overview of the Bayesian methods

We present a brief outline of the fundamental principles underpinning the Bayesian statistical analysis. Suppose that we have an observable random vector \mathbf{y} with probability mass (or density for continuous) function $f(\mathbf{y}|\theta)$, where θ denotes an unobservable parameter. The Bayes' theorem as given by Press (1989), asserts that the probability function of θ , for a given value of \mathbf{y} is expressed as

$$p(\theta|\mathbf{y}) = \begin{cases} \frac{f(\mathbf{y}|\theta)g(\theta)}{\sum_{\theta} f(\mathbf{y}|\theta)g(\theta)} & \text{for a discrete parameter} \\ \frac{f(\mathbf{y}|\theta)g(\theta)}{\int f(\mathbf{y}|\theta)g(\theta)d\theta} & \text{for a continuous parameter} \end{cases} \quad (6.1)$$

Here $p(\theta|\mathbf{y})$ is called the posterior probability function of θ given the observed data, and $g(\theta)$ is the prior probability function of θ . Since the denominators of Equation 6.1 depends only on the \mathbf{y}' s, we can write

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta) g(\theta)$$

where \propto denotes proportionality and $p(\mathbf{y}|\theta)$ denotes the likelihood function.

6.2.1.1 The prior distributions

Incorporating prior information in Bayesian analysis is argued to make the approach more attractive to most empirical research, Press (1989) and Lesaffre & Lawson (2012). The three main classes of priors are the conjugate, the informative and the non-

informative. Priors are often expressed probabilistically or using a distribution by which their parameters are called hyper-parameters.

A prior distribution and a posterior distribution (as given in Sub-subsection 6.2.1.3 below) are said to be conjugate distributions if they both come from the same family of distributions, Raiffa & Schlaifer (1961), Lindley (1965) and Bernardo & Smith (1994). For instance most known distributions belong to the exponential family of distributions, see McCullagh & Nelder (1989), Dobson & Barnett (2008) and Lesaffre & Lawson (2012), and have priors that come from the same family.

An informative prior is one that summarizes the evidence about the parameters concerned from various sources and usually has a significant impact on the results. It provides specific and definite information about a parameter. This usually comes in the form of historical data mainly from past studies and priors based on expert knowledge see Lesaffre & Lawson (2012), Spiegelhalter et al. (2003), Kass & Wasserman (1996) and Vail et al. (2001). On the other hand, a non-informative prior, also referred to as non-subjective, objective or reference prior, as defined by Box & Tiao (1973) and Bernardo & Smith (1994) is one that provides little information relative to the experiment or gives minimal effect relative to the data. A non-informative prior is often regarded as formal representation of ignorance. Non-informative priors that are not a distribution or that have an infinite area under the curve are called improper priors, Lesaffre & Lawson (2012), Press (1989) and Bernardo & Smith (1994).

6.2.1.2 The likelihood

The concept of the likelihood, as first introduced by Fisher (1922), expresses the plausibility of the observed data given as a function of the parameters of a stochastic model. Essentially, the likelihood contains information provided by the observed sample.

Suppose that \mathbf{y} , (y_1, \dots, y_n) denotes the observed data and θ denotes an unobservable parameter, the likelihood function, denoted by $p(\mathbf{y}|\theta)$ is given by

$$p(\mathbf{y}|\theta) = \prod_{i=1}^n p(y_i|\theta)$$

The likelihood function can be viewed as representing the plausibility of θ in light of the data and the value of θ that maximizes $p(\mathbf{y}|\theta)$ is called the maximum likelihood estimate (MLE). The observed data \mathbf{y} affect the posterior distribution through $p(\mathbf{y}|\theta)$. Of importance to note is that Bayesian inferences are based on the probabilities assigned due to the observed data, not due to other imaginary data (as in the frequentist approach) that might have been observed. It is argued that adherence to the likelihood principle implies that inferences are conditional on observed data since the likelihood function is parameterized by the data.

6.2.1.3 The posterior distributions

Under the Bayesian approach, the posterior distribution contains all the information of interest, as it combines the prior information and the likelihood of the data. The posterior distribution is usually presented as (a) summary measures for location and variability; (b) interval estimators for the parameters of interest, and (c) the posterior predictive distribution (PPD) used to predict future observations.

The three most commonly used measures of location are the posterior mode, the posterior mean and the posterior median. The posterior mode is defined by $\hat{\theta}_M = \arg \max_{\theta} p(\theta|\mathbf{y})$, and gives the value of θ for which $p(\theta|\mathbf{y})$ is maximal. The posterior mean is defined by $\bar{\theta} = \int \theta p(\theta|\mathbf{y}) d\theta$, which minimizes the squared loss, that is $\int (\theta - \hat{\theta})^2 p(\theta|\mathbf{y}) d\theta$. The posterior median is the solution to the equation $0.5 = \int_{\hat{\theta}_M} p(\theta|\mathbf{y}) d\theta$. A measure of variability which determines the shape of the distribution is the posterior variance, $\bar{\sigma}^2$ (together with the posterior standard deviation $\bar{\sigma}$) defined as $\bar{\sigma}^2 = \int (\theta - \bar{\theta})^2 p(\theta|\mathbf{y})$. A range of plausible parameter values of θ , termed the credibility interval, with probability $1 - \alpha$ can be obtained from the posterior distribution. Formally, an interval $[a, b]$ is a $100(1 - \alpha)\%$ credibility interval for θ if

$$P(a \leq \theta \leq b|\mathbf{y}) = 1 - \alpha.$$

A PPD is the distribution of unobserved observations conditional on the observed data. Suppose $p(y|\theta)$ be the distribution of y and assume an i.i.d. sample $\mathbf{y} \equiv \{y_1, \dots, y_n\}$ is available and suppose we wish to predict future observations \tilde{y} or sets of observations $\tilde{\mathbf{y}}$, that is, we wish to obtain the distribution of \tilde{y} that belongs to the same population as the observed sample. Then the distribution of \tilde{y} is the PPD and is given by

$$p(\tilde{y}|\mathbf{y}) = \int p(\tilde{y}|\theta) p(\theta|\mathbf{y}) d\theta.$$

Lesaffre & Lawson (2012) gave examples of posterior distributions and PPD under such distributions as binomial for binary data and Poisson for count data.

6.2.1.4 Determining the posterior distribution

Under the Bayesian analysis, proper determination of the posterior distribution is key for valid statistical inference. The two most popular techniques available are the numerical integration and sampling from the posterior distribution.

There are several techniques available for approximating integrals numerically. In the Bayesian framework, as described by Bauwens et al. (2000), the aim is to obtain posterior densities that can be summarized by posterior expectations, variances and graphs of marginals. Essentially, the idea is to evaluate integrals that correspond to moments of the posterior density. Suppose $g(\theta)$ is a function of θ having the density $p(\theta|y)$, interest is in computing

$$E[g(\theta)] = \int g(\theta) p(\theta|y) d\theta. \tag{6.2}$$

In most practical complex problems the integral in Equation 6.2 has no known

analytical solution and it is not analytically tractable, hence the use of numerical integration giving an approximation of the integral. Suppose that we can write Equation 6.2 as the integral of $h(\theta) = g(\theta)p(\theta|y)$, numerical integration rules approximate the integral of h by a weighted average of values of h given by

$$\int h(\theta) d\theta \approx \sum_{j=1}^n w_j h(\theta_j) \quad j = 1, \dots, n, \quad (6.3)$$

where w_j are positive weights summing to 1. Taking cognisance of the fact that an integral is a measure of area, the fundamental idea behind Equation 6.3 is to split the integration space into small parts in order to evaluate the area of each part as $w_j h(\theta_j)$ and to sum the areas of the small parts. Different methods of numerical integration are determined by the rules used to choose the points to split the area and the weights. Common rules fall into two main categories; deterministic and stochastic (Monte Carlo methods). Under deterministic rules, points are chosen systematically in order to cover the whole space with a grid, whereas under the stochastic, points are chosen randomly, according to some probability distribution.

Under the stochastic rules, the points are chosen in areas where the integrand varies the most, in most cases resulting in fewer points than the deterministic rules. The Monte Carlo methods are simulation-based and use random numbers generated from some probability distribution, to generate samples from the posterior see for example Robert & Casella (1999). Examples of the Monte Carlo methods are the Gibbs sampling and Metropolis-Hastings sampling which are based on simulating dependent samples of a Markov chain type resulting in what is called Markov chain Monte Carlo (MCMC) sampling. MCMCs are algorithm-based and do not use integration.

The Gibbs sampling, as first introduced by Geman & Geman (1984), is a randomized algorithm for obtaining a sequence of observations which are approximated from a specified probability distribution, Robert & Casella (1999), Hastings (1970) and Dey et al. (2000). It generates a Markov chain of samples, with neighbouring samples being

correlated. To illustrate the Gibbs sampling procedure, suppose that we are interested in sampling from the posterior $p(\boldsymbol{\theta}|\mathbf{y})$, where $\boldsymbol{\theta}$ is a vector of two parameters, θ_1 and θ_2 . The sampling procedure is initiated by a starting value for the parameters, say $\boldsymbol{\theta}^{(0)}$, for $\theta_1^{(0)}$ and $\theta_2^{(0)}$ and then explore the posterior distribution by generating θ_1^k and θ_2^k , where $k = 1, 2, \dots$ in a sequential order. Basically, given θ_1^k and θ_2^k at iteration k , the $(k + 1)$ th value for each of the parameters is generated according to the following iterative scheme:

- Draw $\theta_1^{(k+1)}$ from $p(\theta_1|\theta_2^k, \mathbf{y})$;
- Draw $\theta_2^{(k+1)}$ from $p(\theta_2|\theta_1^{(k+1)}, \mathbf{y})$.

Then the Gibbs sampler produces a sequence of values $\boldsymbol{\theta}^k = (\theta_1^k, \theta_2^k)^T$, $k = 1, 2, \dots$ which are dependent and create a chain. Summary measures such as the mode or the mean from the chain estimate the true posterior measures.

The Metropolis-Hastings algorithm, as first introduced by Metropolis et al. (1953) and further extended by Hastings (1970) is an MCMC procedure, that, unlike the Gibbs sampler, does not require the full conditionals. For the Metropolis-Hastings algorithm, suppose that a Markov chain is at $\boldsymbol{\theta}^k$ at the k th iteration when exploring the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$, and denote the next position in the chain by $\tilde{\boldsymbol{\theta}}$. The algorithm uses an acceptance rejection criterion to make iterations. A new position is accepted if it is in an area of higher posterior mass see Lesaffre & Lawson (2012), otherwise it is accepted with a certain probability. A proposal density evaluated for $\tilde{\boldsymbol{\theta}}$ at iteration k is denoted as $q(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}^k)$. If $\tilde{\boldsymbol{\theta}}$ is accepted, that is if $\boldsymbol{\theta}^{k+1} = \tilde{\boldsymbol{\theta}}$, then a next move to $\tilde{\boldsymbol{\theta}}$ is made, however if $\tilde{\boldsymbol{\theta}}$ is rejected the process stays at $\boldsymbol{\theta}^k$. The probability of accepting a proposed value depends on the posterior distribution. When the candidate position lies in an area where the posterior distribution has a higher value, that is $p(\tilde{\boldsymbol{\theta}}|\mathbf{y})/p(\boldsymbol{\theta}^k|\mathbf{y}) > 1$, then the move will always be made, whereas when the candidate position lies in an area where the posterior distribution has a lower value, that is $p(\tilde{\boldsymbol{\theta}}|\mathbf{y})/p(\boldsymbol{\theta}^k|\mathbf{y}) < 1$, then

the move will be made with probability $r = p(\tilde{\theta}|\mathbf{y})/p(\theta^k|\mathbf{y})$. The process continues until convergence.

6.2.1.5 Convergence diagnostics for a Markov chain

Use of Markov chain techniques are based on the property that the generated chain ultimately provides a sample from the posterior distribution and that the summary measures computed consistently estimate the corresponding true posterior summary measures, Robert & Casella (1999) and Lesaffre & Lawson (2012). The ideal is for the chain to converge to a stationary distribution, that is the target posterior distribution. Convergence techniques in an MCMC algorithm are aimed at checking how close the process is to the true posterior distribution.

The convergence in an MCMC algorithm is an asymptotic property which implies that the distribution of θ^k , that is $p_k(\theta)$ converges to the target distribution $p(\theta|\mathbf{y})$ as $k \rightarrow \infty$, where k are the number of iterations, see Lesaffre & Lawson (2012), Brooks (1998) and Brooks & Roberts (1998). Evaluating convergence of a chain involves assessing convergence of the marginal posterior distributions by checking how well the chain is mixing, or moving around the parameter space.

Several tests, both graphical and statistical that can be used to check convergence are available via convergence diagnostics. Basically the diagnostics are used to check for stationarity of the chain and verify the accuracy of the posterior summary measures. Various convergence diagnostics are available and for the current research, we utilize the trace plots, the Geweke plots and the Heidelberger-Welch test as described by Lesaffre & Lawson (2012), Brooks & Roberts (1998) and Brooks (1998).

The trace plots are plots of iteration number against the value of the draw of the parameter at each iteration. Trace plots are produced for each parameter separately and the evaluations are done univariately. A chain that is stationary forms the informal “thick pen” as explained by Gelfand et al. (1990). A trace plot that depicts dependence

of the chain on its initial state by revealing an upward or downward trend is indicative of gross deviations from stationarity.

The Geweke diagnostic suggested by Geweke (1992), is based on comparing the means of an early and a late part of the chain using a significant test. Suppose that there are n values θ^k assumed to be i.i.d. and that they are split into two parts: the early part (A) with n_A elements and the late part (B) with n_B elements with posterior means $\bar{\theta}_A$ and $\bar{\theta}_B$ respectively. The means can be compared using a Z – test based on

$$Z = \frac{\bar{\theta}_A - \bar{\theta}_B}{\sqrt{s_A^2/n_A + s_B^2/n_B}}, \quad (6.4)$$

where s_A^2 and s_B^2 are the classical estimates of the respective variances. However, elements of a Markov chain are dependent, thus another estimator of variances is needed. A spectral density concept based on a time series approach is utilized. To ensure that $\bar{\theta}_A$ and $\bar{\theta}_B$ are independent Geweke (1992) suggested taking for A the initial 10% of the iterations, $n_A = n/10$, and for B the last 50%, $n_B = n/2$ to create a distance between the two parts. Then, if the ratios n_A/n and n_B/n are fixed, with $(n_A + n_B)/n < 1$, then it is known that, Lesaffre & Lawson (2012) and Brooks & Roberts (1998),

$$Z = \frac{\bar{\theta}_A - \bar{\theta}_B}{\sqrt{s_A^2/n_A + s_B^2/n_B}} \xrightarrow{d} N(0, 1) \quad \text{as } n \rightarrow \infty \quad (6.5)$$

The result is used to test the null hypothesis of equal location. The null hypothesis is rejected if $|Z|$ is large, indicating that the chain has not converged by iteration k .

The Heidelberger-Welch (HW) diagnostic proposed by Heidelberger & Welch (1983), is an automated test for checking the stationarity of the chain and further evaluate whether the length of the chain is sufficient to ensure desired accuracy for the posterior means of the parameters. The test is based on the Cramer-von Mises test statistic to decide to either accept or reject the null hypothesis that the chain is from a stationary distribution. The test consists of two steps: checking stationarity and determining

accuracy.

Under the first step, based on say N , iterations, a test statistic is calculated and the null hypothesis of stationarity is either rejected or accepted. If rejected, the first 10% of the chain is discarded and another test statistic is calculated on the remaining 90%. This process continues until the null hypothesis is accepted or if 50% of the chain is discarded at which stage the chain fails the test and needs to be run longer. For the second step, the part of the chain not discarded is considered and half-width of the $(1 - \alpha)\%$ credible interval around the mean calculated. A threshold value ϵ , say, is determined and if the ratio of the half-width and the mean is lower than ϵ , then the chain passes the test, otherwise it must be run longer.

6.2.2 Bayesian logistic regression

We consider a Bayesian logistic regression modelling from a generalized linear modelling (GLM) framework as described by Dey et al. (2000). In general, a GLM approach as first introduced by Nelder & Wedderburn (1972) and modified by McCullagh & Nelder (1989) provides a flexible and unified approach to analyzing both normal and non-normal data. Initially, application of the GLM often take a classical approach, however the availability of complex and high speed software routines have witnessed a rapid growth in Bayesian analyses carried out through GLMs. The fundamental idea of a GLM is based on the assumption that the underlying distribution of responses belong to the exponential family of distributions, and a link function transformation of its expectation is modelled as a linear function of observed covariates. Furthermore, it is assumed that the variance of the response is a specified function of its mean. Formally, let $\mathbf{y} = y_1, \dots, y_n$ denotes a vector of observed data for a random variable, \mathbf{X} denotes a design matrix of covariates (x_1, \dots, x_n) , under the classical notation a distribution that

belongs to the exponential family of distribution can be expressed in the form

$$f(\mathbf{y}; \theta, \phi) = \exp \left[\frac{\mathbf{y}\theta - b(\theta)}{a(\phi)} + c(\mathbf{y}; \phi) \right], \quad (6.6)$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions and θ and ϕ are a specific set of unknown parameters. In most practical applications, $a(\phi) = \phi/w$, where w is a prior weight. The classical estimation procedure is the maximum likelihood. With reference to Equation 6.6, the log-likelihood function is given by

$$l = l(\theta, \phi; \mathbf{y}) = \log f(\mathbf{y}; \theta, \phi). \quad (6.7)$$

Under a GLM approach, the mean μ is related to the covariates via a monotone differentiable (link) function $g(\mu)$ given by

$$g(\mu) = \eta = \sum_{j=1}^p \mathbf{x}_j \beta_j = \mathbf{X}^T \boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is a vector of parameters. The maximum likelihood estimates are obtained as iterative solutions of the log-likelihood equations

$$\frac{\partial l}{\partial \beta_j} = 0.$$

see McCullagh & Nelder (1989), Dobson & Barnett (2008) and McCulloch & Searle (2001).

For the logistic regression, let Y be a binary response variable taking on values $[0,1]$, and let X_1, \dots, X_p be a set of explanatory variables. Suppose that $\pi(x_i) = P(Y = 1|X = x_i)$ denotes the conditional probability that $Y = 1$ given the explanatory variables for subject i .

Then the logistic regression is given as a probability of success as

$$\begin{aligned}\pi(x_i) &= g^{-1}(\mathbf{x}'_i \boldsymbol{\beta}) \\ &= \frac{\exp\{\mathbf{x}'_i \boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}'_i \boldsymbol{\beta}\}}.\end{aligned}$$

The likelihood contribution from the i th subject is

$$L_i = \left(\frac{\exp\{\mathbf{x}'_i \boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}'_i \boldsymbol{\beta}\}} \right)^{y_i} \left(1 - \frac{\exp\{\mathbf{x}'_i \boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}'_i \boldsymbol{\beta}\}} \right)^{(1-y_i)}.$$

Under the ordinary logistic regression, subjects are independent hence the likelihood function over all say, n subjects is given by

$$L = \prod_{i=1}^n \left[\left(\frac{\exp\{\mathbf{x}'_i \boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}'_i \boldsymbol{\beta}\}} \right)^{y_i} \left(1 - \frac{\exp\{\mathbf{x}'_i \boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}'_i \boldsymbol{\beta}\}} \right)^{(1-y_i)} \right]. \quad (6.8)$$

Under a Bayesian framework, the prior distributions are the respective distributions of the set of parameters $\beta_0, \beta_1, \dots, \beta_p$. Options for priors depend on available information. The most common priors are of the form

$$\beta_j \sim N(\mu_j, \sigma_j^2), \quad \sigma_j^2 \sim \text{inv} - \chi^2(v_j, s_j^2), \quad (6.9)$$

where μ_j is often taken to be zero and σ is usually chosen to be large enough in order for the prior to be non-informative, v and s denote degrees of freedom and scale for the t – distribution respectively.

The posterior distribution is obtained by combining the full likelihood function (Equation 6.8) and the prior (Equation 6.9) to obtain

$$\begin{aligned} \text{posterior} &= \prod_{i=1}^n \left[\left(\frac{\exp\{\mathbf{x}'_i \boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}'_i \boldsymbol{\beta}\}} \right)^{y_i} \left(1 - \frac{\exp\{\mathbf{x}'_i \boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}'_i \boldsymbol{\beta}\}} \right)^{(1-y_i)} \right] \\ &\times \prod_{j=0}^n \frac{1}{\sqrt{2\pi\sigma_j}} \exp \left\{ -\frac{1}{2} \left(\frac{\beta_j - \mu_j}{\sigma_j} \right)^2 \right\}. \end{aligned}$$

The posterior distribution above does not have a closed form expression, that is it is not analytically tractable, thus the sampling algorithm explained in Section 6.2.1.4 above, are utilized.

6.2.3 Statistical computing

All the analyses for this chapter were done in **R** packages, Team (2013). In particular, the package **arm** (applied regression and multilevel modelling) by Gelman et al. (2010) was used to compute the Bayesian logistic regression via the function **bayesglm**. The function allows specification of independent prior distributions for the parameter in the t family of distributions. Specifically, we considered the non-informative Student- t family of prior distributions described by Gelman et al. (2008), that focuses on the Cauchy with center 0, degrees of freedom v and scale s . The scale, s is chosen to provide minimal prior information in order to constrain the coefficients to lie within a reasonable range. Furthermore, the choice of the prior was informed by the need to accommodate what Gelman et al. (2008) termed a longer-tailed version of the distribution by assuming one-half additional success and one-half additional failure in a logistic regression. The prior distribution is incorporated in the estimation by altering the weighted least squares via augmenting the approximate likelihood with the prior distribution.

For statistical inference, the empirical distribution of the simulate values was obtained based on iterative draws from the posterior distribution using the package **MCMCpack** by Martin et al. (2013). Summary measures (the mean and median) were computed from the MCMC simulates and were used for the inference. Assessing

convergence was done using package **mcmcplots** in **R** by Curtis et al. (2015). In particular, trace plots and Geweke plots were constructed for the parameters and selected ones are displayed. Furthermore, a Heidelberger-Welch test was also performed in the **MCMCpack** package.

6.3 Results and discussion

For the current research, we considered 14 491 respondents who had complete information regarding the relevant variables discarding all the units with missing values. Here missing data were assumed missing completely at random (MCAR) and the observed complete cases were regarded as a random sample of the full sample and possibly the target population, Pigott (2001) and Rubin (1976). The HIV test results indicated that 12 103 (83.52%) were HIV negative and 2 388 (16.48%) were HIV positive. Basic plots, univariate summary statistics and design-consistent tests for association (results not shown here) were used to explore relationships between the response and the predictor variables and also to select significant predictor variables. The significant predictor variables were age, gender, marital status, literacy level and place of residence. The literacy level was measured in terms of one's ability to read and write as: the non-literate were classified as those who cannot read nor write; the partially literate were those who could read or write part of a sentence; and the literate were those who could read and write a full sentence. We also included interaction terms for age and gender and age and marital status. In the interpretation of the results, for the categorical variables, we adopted the reference cell approach.

Table 6.1 gives parameter estimates, standard errors, p-value for the estimates for the posterior distribution. Independent t -distributions with conditional means that corresponds to the parameter estimates obtained from an ordinary logistic regression model and scales 10 for intercept and 2.5 for the other coefficients were used.

Table 6.1: Parameter estimates, standard errors and p-values for the posterior distribution for the observed data using a non-informative prior distribution

Parameter	Estimate	S. E.	p-value
Intercept	-3.257	0.153	< 0.001
Gender			
Male	-0.247	0.174	0.156
Age group			
20 – 24	0.938	0.144	< 0.001
25 – 29	1.596	0.145	< 0.001
30 – 34	1.951	0.149	< 0.001
35 – 39	1.922	0.153	< 0.001
40 – 44	1.456	0.166	< 0.001
45 – 49	1.328	0.174	< 0.001
50 – 54	0.640	0.772	0.718
Marital status			
Married	-0.060	0.108	0.574
Divorced	0.726	0.132	< 0.001
Widowed	1.594	0.142	< 0.001
Place of residence			
Urban	0.189	0.052	< 0.001
Literacy			
Partially	0.450	0.128	< 0.001
Literate	0.177	0.105	0.094
Gender*Age group			
20 – 24 : Male	-0.776	0.240	0.001
25 – 29 : Male	-0.674	0.241	0.005
30 – 34 : Male	-0.631	0.251	0.012
35 – 39 : Male	-0.246	0.256	0.337
40 – 44 : Male	0.487	0.268	0.069
45 – 49 : Male	0.678	0.282	0.016
Gender*Marital status			
Married : Male	0.419	0.179	0.019
Divorced : Male	0.555	0.232	0.017
Widowed : Male	0.637	0.319	0.046

Convergence diagnostic tests as explained in section 6.2.1.5 above to assess the convergence of the Markov chain before obtaining the descriptive summary statistics for the parameters. The posterior distribution was obtained after two thousand iterations performed gradually and assessing convergence at every stage. Figure 6.1 gives the trace plots for a few of the parameters of the posterior distribution obtained by the

MCMC algorithm as explained in section 6.2.1.5 above. All the trace plots do not display any significant upward or downward trend along the iterations and the density plots also show almost symmetrical distributions. In particular the trace plots exhibit the so called “thick pen” as described by Gelfand et al. (1990). This is indicative of insignificant deviations from stationarity and the MCMC algorithm can be considered to have converged.

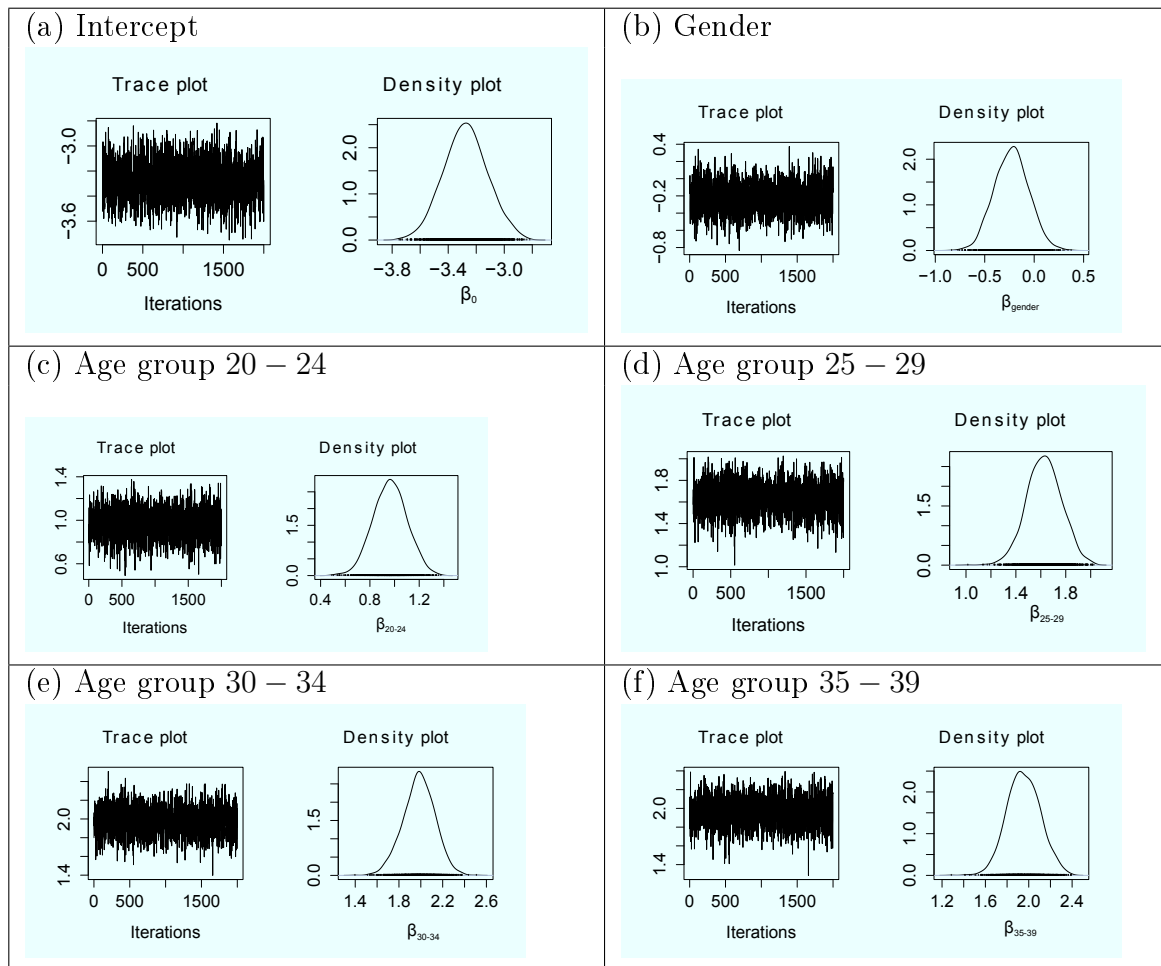


Figure 6.1: Trace plots and density plots for the first six coefficients from the posterior distribution

Figure 6.2 gives the Geweke plots for selected parameters. As a rule of thumb, a significant proportion of Z -scores outside the two-standard deviation bands is indicative of a chain that has not converged by iteration k . The results in Figure 6.2 show that all

the Z -scores fall within the two-standard deviation bands for the parameters gender, and age groups 20 – 14, 25 – 29 and 30 – 34 whereas there is a negligible proportion for the Z -scores under the intercept and age group 35 – 39 that are outside the bands. This is a strong indication of a chain that has converged by iteration 2000.

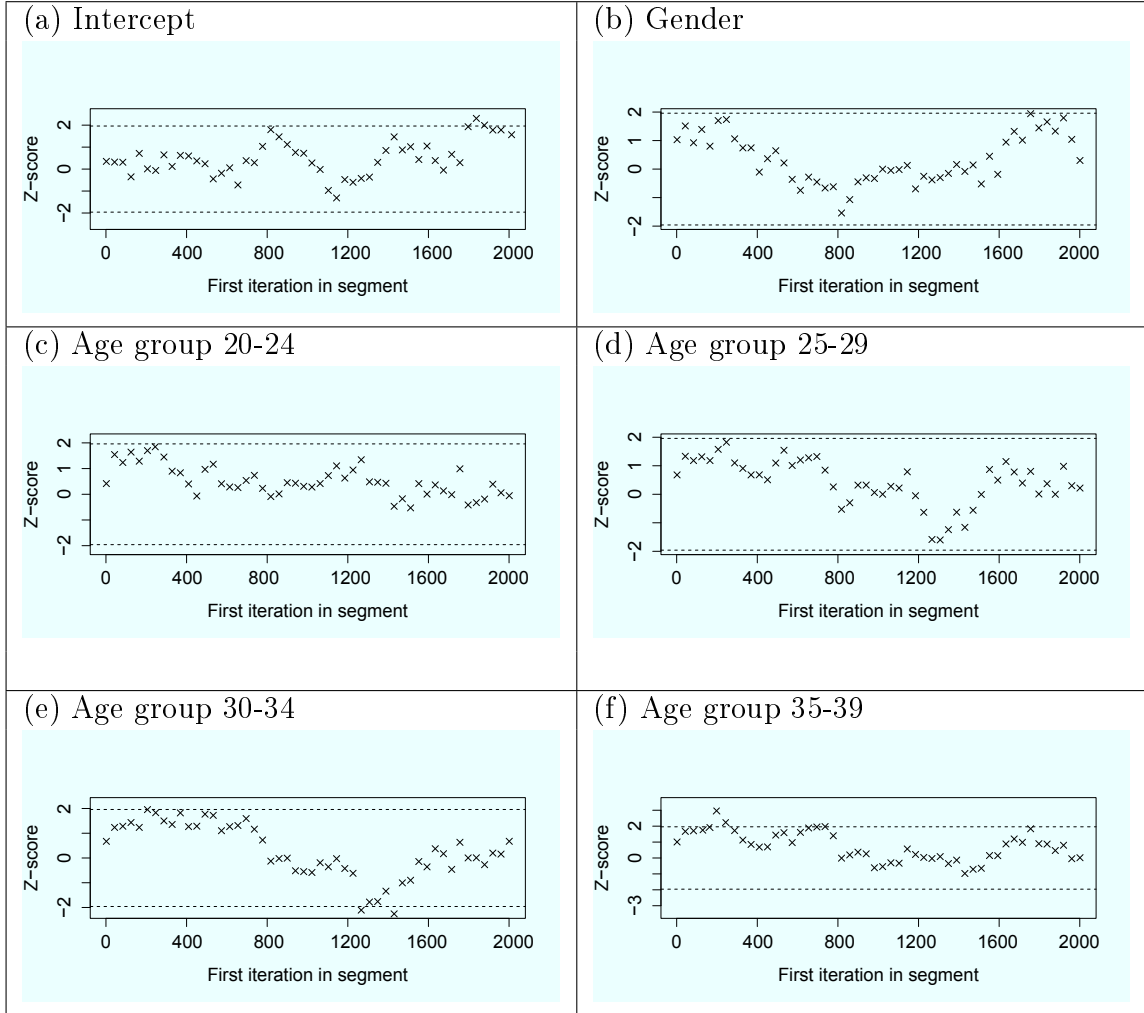


Figure 6.2: Geweke plots for the first six coefficients from the posterior distribution

We also considered the Heidelberg-Welch diagnostic test. The results of a test with $\epsilon = 0.1$ show that most of the parameters have passed the stationarity test except for male, age group 50 – 54, the married level for marital status, the literate level for literacy and the gender by age group interaction terms 35 – 39 :M and 40 – 44 :M. All the parameters passed the half-width test indicating that the chain was run sufficiently

long. Table 6.2 gives the summary measures, in the form of means, standard deviations and the 95% credibility intervals from the empirical distribution. These are used to perform statistical inference about the individual parameters.

Table 6.2: Summary measures (mean, median, standard deviation and credibility intervals) for the coefficients obtained from the posterior distribution

Parameter	Mean	Median	St. Dev.	95% Credibility interval
Intercept	-3.261	-3.259	0.149	(-3.509, -2.932)
Gender				
Male	-0.243	-0.252	0.173	(-0.593, 0.063)
Age group				
20 – 24	0.938	0.921	0.145	(0.621, 1.242)
25 – 29	1.627	1.619	0.141	(1.344, 1.898)
30 – 34	1.973	1.974	0.152	(1.683, 2.242)
35 – 39	1.940	1.925	0.161	(1.645, 2.255)
40 – 44	1.460	1.459	0.159	(1.177, 1.768)
45 – 49	1.352	1.353	0.164	(1.061, 1.659)
50 – 54	0.972	0.885	0.933	(-3.009, 4.346)
Marital status				
Married	-0.078	-0.073	0.108	(-0.329, 0.146)
Divorced	0.725	0.723	0.136	(0.476, 1.014)
Widowed	1.585	1.602	0.163	(1.281, 1.864)
Place of residence				
Urban	0.183	0.180	0.053	(0.093, 0.288)
Literacy				
Partially	0.440	0.444	0.128	(0.219, 0.673)
Literate	0.174	0.175	0.109	(-0.030, 0.364)
Gender*Age group				
20 – 24 : Male	0.773	-0.745	0.226	(-1.163, -0.393)
25 – 29 : Male	-0.714	-0.700	0.252	(-1.162, -0.354)
30 – 34 : Male	-0.642	-0.616	0.288	(-1.285, -0.107)
35 – 39 : Male	-0.251	-0.252	0.288	(-0.777, 0.265)
40 – 44 : Male	0.493	0.515	0.292	(-0.051, 1.004)
45 – 49 : Male	0.670	0.692	0.306	(0.083, 1.202)
Gender*Marital status				
Married : Male	0.435	0.410	0.189	(0.097, 0.784)
Divorced : Male	0.536	0.539	0.237	(0.121, 0.975)
Widowed : Male	0.653	0.613	0.350	(-0.148, 1.250)

The results in Table 6.1 show that HIV status is dependent on one's age, gender, marital status, literacy level and place of residence (rural or urban). The interaction

terms for age and gender, and age and marital status were also included in the model as they were found to be significant. In the Bayesian paradigm, the interpretation of the credibility interval is that, there is a 95% probability that the true parameter (for respective interval) is included in the given interval. For instance, for the the covariate gender, there is a 95% probability that the true parameter falls in the interval $(-0.243, 0.063)$. It is important to know that the covariate gender was included, although it is non-significant, because its interaction with age and with marital status are significant. We computed the odds ratios for the means that are presented in Table 6.2 in order to facilitate the interpretation of the model parameters. The results are displayed in Table 6.3.

Table 6.3: Odds ratios and credibility intervals for parameter estimates obtained from simulating from the posterior distribution

Parameter	OR	95% Credibility intervals
Intercept	0.039	(0.029 , 0.052)
Gender		
Male	0.781	(0.557 , 1.096)
Age group		
20 – 24	2.555	(1.923 , 3.395)
25 – 29	4.933	(3.742 , 6.504)
30 – 34	7.036	(5.223 , 9.477)
35 – 39	6.835	(4.985 , 9.370)
40 – 44	4.289	(3.140 , 5.857)
45 – 49	3.773	(2.736 , 5.204)
50 – 54	1.896	(0.305 , 11.807)
Marital status		
Married	0.942	(0.762 , 1.164)
Divorced	2.067	(1.583 , 2.698)
Widowed	4.923	(3.577 , 6.777)
Place of residence		
Urban	1.208	(1.089 , 1.340)
Literacy		
Partially	1.568	(1.220 , 2.016)
Literate	1.194	(0.964 , 1.478)
Gender*Age group		
20 – 24 : Male	0.460	(0.296 , 0.717)
25 – 29 : Male	0.510	(0.311 , 0.835)
30 – 34 : Male	0.532	(0.303 , 0.935)
35 – 39 : Male	0.782	(0.445 , 1.375)
40 – 44 : Male	1.627	(0.918 , 2.884)
45 – 49 : Male	1.970	(1.081 , 3.589)
Gender*Marital status		
Married : Male	1.520	(1.050 , 2.202)
Divorced : Male	1.742	(1.095 , 2.772)
Widowed : Male	1.891	(0.952 , 3.755)

The results show that the males have slightly lower odds of HIV (OR = 0.781, 95% CI = 0.557 – 1.096) than the females. The difference in the odds of HIV between the males and the females is not statistically significant as the confidence interval includes a 1. There are several possible explanations for the observed differences in the risk of and susceptibility to HIV between males and females. Studies have revealed that biologically

females are more likely to become infected with HIV through unprotected heterosexual intercourse than males. In addition women, especially in many sub-Saharan African countries, are less likely to be able to negotiate condom use and are more likely to be subjected to non-consensual sex. The risk of transmitting HIV from men to women is much higher than from women to men because women are exposed to considerable amounts of seminal fluid during vaginal sexual intercourse. Low economic status tend to force females into transactional sex increasing their risk of HIV.

There is a general increase in the odds of HIV with age peaking at the age group 30 – 34 (OR = 7.036, 95% CI = 5.223 – 9.477), and gradually falls thereafter. This implies that those who are in the 30 – 34 years age group have over seven times higher odds of HIV as compared to the 15 – 19 year age group. The higher odds of HIV associated with those in the age groups 25 – 45 is possibly due to the increased sexual activities that characterize individuals of these ages. The odds of HIV vary considerably across different categories of marital status. Specifically, the married have slightly lower odds of HIV (OR = 0.942, 95% CI = 0.762 – 1.164), the divorced have over twice odds of HIV (OR = 2.067, 95% CI = 1.583 – 2.698) and the widowed have almost eight times higher odds of HIV (OR = 7.923, 95% CI = 3.577 – 6.777) as compared to the single/never married. The urban residents have slightly higher odds of HIV (OR = 1.208, 95% CI = 1.089 – 1.340) as compared to their rural counterparts. This disproportionate odds of HIV is possibly because the urban residents are mainly middle aged and well of, often associated with high rates of risky sexual behaviours. It is argued that in sub-Saharan African urban places, the wealthier individuals tend to attract multiple and concurrent sexual partners. Urban areas are also synonymous with increased commercial sex activities which is argued to be responsible for the rapid spread of HIV. In addition, urban residents are predominantly male, in some instances whose wives stay in the rural areas. This has been argued to facilitate extramarital sexual relationships.

The gender by age group interaction gives the additional effect of gender on age group on the odds of HIV. The results show that the younger age groups show lower odds of being HIV positive among the males as compared to females whereas the older age groups show higher odds of being HIV positive as compared to females. This agrees well the general belief regarding the disparities in the odds of HIV among males and females of different age groups. It is believed that younger females engage in sexual relationships older males. The gender by marital status interaction shows the additional effect of gender on marital status in determining the odds of HIV. In particular, it is more likely for a married male individual (OR = 1.52, 95% CI = 1.050 – 2.202), for a divorced male person (OR = 1.742, 95% CI = 1.095 – 2.772) and for the widowed male individual (OR = 1.891, 95% CI = 0.952 – 3.755) to be HIV positive as compared to single/never married females.

6.4 Conclusion

We computed a Bayesian logistic regression model from a GLM perspective for HIV on demographic and socio-economic variables. Non-informative prior probability distributions for the demographic and socio-economic variables were used in the building of the model. We conclude that HIV is related to an individual's gender, age, marital status, place of residence, literacy level and age by gender and gender by marital status effect. Females have slightly higher odds of being HIV positive than males whereas urban dwellers are at more risk of HIV than their rural counterparts. As compared to the single/never married, the married individuals are less likely to be HIV positive whereas the divorced and the widowed have a higher likelihood of being HIV positive. There is a gender by age interaction effect that is significant in the logistic regression model for predicting the probability of being HIV positive.

Chapter 7

Bayesian hierarchical logistic regression for estimating risk of HIV using population-based survey data

This chapter combines the fundamental concepts of hierarchical models presented in Chapter 5 and those of Bayesian analysis given in Chapter 6 to develop a logistic regression model for HIV. Essentially the Bayesian approach is imposed on the GLMM in order to incorporate prior information in the modelling process. In addition the numerical integration explained in Chapter 6 is also utilized to mitigate the evaluation of the high-dimensional integrals common in the estimation of the GLMM via maximum likelihood.

Abstract

Most practical complex survey data exhibit some multilevel or hierarchical structural form brought about by the prominent features of the sampling design and the underlying target population. These data are often obtained using stratified multistage clustered sampling designs which give rise to a ‘clustered’ or ‘nested’ data with a multi-

layered structure that usually induce intra-class correlations of units within clusters. Appropriate statistical inference and conclusions based on such data require methods of analysis that take account of these features of the data. We utilized the generalized linear mixed modeling framework to obtain a hierarchical logistic regression model for HIV as a function of demographic and socio-economic variables. The hierarchical models are capable of capturing the layered structure of the data and determine how different layers interact and impact a response variable. A standard technique for fitting the models involves maximum likelihood computations based on the assumption of normality of the parameter estimates. However, in practical instances, the underlying process is often not Gaussian, especially when the data are sparse. In addition, generalizing models to non-Gaussian data is practically difficult since integrating over the random effects is intractable. This necessitates the use of external information about model parameters for example using a Bayesian statistical analysis paradigm.

We impose a Bayesian approach to the computation of the hierarchical logistic model. In particular, we utilize the strength of the Bayesian framework for evaluating the intractable high-dimensional integrals encountered in most likelihood based statistical analysis. Thus we combine the fundamental concepts of the Bayesian analysis and hierarchical modelling to explain the variation in HIV using demographic, socio-economic and behavioural variables.

The research used the 2010-11 Zimbabwe Demographic and Health Surveys (2010-11ZDHS) data. The results show that HIV prevalence is dependent on one's gender, age, marital status, literacy level, sex of head of the household, wealth index, place of residence and interaction effects of gender by age and gender by marital status. Furthermore, the results show a substantial cluster to cluster and household to household variability.

7.1 Introduction

Most complex survey data encountered in practice, especially those in scientific and social investigations, often depict some structural form brought about by the prominent features of the underlying target population. The complex sampling design, that usually involves stratification, multistage clustering and application of unequal selection probabilities to the sampling units, gives rise to data that show a multilevel or hierarchical and nested or clustered. The clustering induces intra-class correlations among units sharing the same cluster. This renders standard statistical methods, that are reliant on the assumption of independence, inappropriate. The multilevel structure of the data gives rise to multiple sources of variation representing randomness introduced at different levels of the data structure.

Research has indicated stark geographical variation in HIV prevalence. The spatial variation highlights a localized clustering in HIV transmission within micro-epidemics of varying scales and intensity. For example Tanser et al. (2009) and Cuadros et al. (2013) identified spatial clusters with high and low numbers of HIV in sub-Saharan African countries and measure the strength of clustering using a Kulldorff spatial scan statistics analysis under randomness. For instance if one partner in a sexual relationship is infected with HIV and if no intervention is done, the chances are high that the HIV free partner will also be infected. MTCT of HIV before and after birth results in what is termed vertical transmission of HIV results in clustered HIV patterns, Cout-soudis et al. (1999), Bobat et al. (1997) and Dunn et al. (1992). Social sexual mixing at the community level such as in mining communities, at growth point, along national roads and in border towns enhances 'hot spots' for cases of HIV resulting in clustered effect of HIV, Tanser et al. (2009). Therefore survey data for phenomena such HIV exhibit considerable individual to individual, household to household and community to community heterogeneity. An effective analysis approach of the variation in a response variable depicting such multi-layering and clustering should reflect these multiple

sources of variabilities. In addition, these spatial variations and multi-layering poses questions about the drivers of such heterogeneities. Key drivers responsible for these heterogeneities in prevalence are mainly socio-economic, demographic and behavioural factors. The spatial structure of HIV prevalence can have a considerable impact on the dynamics of the epidemic, its progression, persistence and the nature and success of the interventions.

The basic statistical approach for explaining a process with multiple sources of variation is via hierarchical models built around mixed effects models. The mixed effects models include both fixed and random effects, McCulloch & Searle (2001). For a unified version that accounts for both normal and non-normal data generalized linear mixed models (GLMM), an extension of the generalized linear models (GLM) as first introduced by Nelder & Wedderburn (1972) and further expanded by McCullagh & Nelder (1989) is often utilized. Use of hierarchical models allows different levels of variation to be characterized by ascribing causes (of the variability) via the use of covariates. A standard technique for fitting the models involves maximum likelihood computations based on the assumption of normality of the parameter estimates. However, in practical instances, the underlying process is often not Gaussian, especially when the data are sparse. In addition, generalizing models to non-Gaussian data is practically difficult since integrating over the random effects is intractable, Hadfield (2010). In such cases, external information about model parameters is required to inform the observed data. The Bayesian approach to statistical analysis can offer an appropriate framework to incorporate such external information. The use of techniques such as the Markov chain Monte Carlo (MCMC) methods, see Breslow & Clayton (1993), via marginalizing the random effects is commonly adopted.

The Bayesian paradigm is an approach to statistical analysis based on the Bayes rule in which model parameters are regarded as random variables in the sense that knowledge of them is incomplete, Johnson (2010). Prior beliefs about the model parameters,

represented by a probability distribution, describe the degree of uncertainty with which these parameters are known. The beliefs about the parameters are then combined, using the Bayes rule, with the likelihood of the data leading to a posterior distribution, see Press (1989), Raiffa & Schlaifer (1961), Johnson (2010) and Lesaffre & Lawson (2012). A Bayesian version of the hierarchical model, as first introduced by Lindley & Smith (1972) can be computed by providing prior distribution for both the fixed and the random effects and combining them with the likelihood of the data.

Application of Bayesian hierarchical logistic regression has received considerable attention in a variety of fields. For instance Rouder & Lu (2005) used Bayesian hierarchical models to explain variation in signal detection in the presence of participant and item variability. Gopal et al. (2012) employed a Bayesian approach to explain the dependencies in parent-child relationships in a hierarchical structured setting under hierarchical classification problem.

7.2 Methods

7.2.1 Hierarchical Bayesian modelling

We consider a GLMM given as in Chapter 5 as

$$\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} \quad (7.1)$$

formulated in the Bayesian framework, where the prior distributions for $\boldsymbol{\beta}$ and \boldsymbol{b} are required. Under the Bayesian standpoint, as given by Clayton (1996), there is no need to partition the vector of explanatory variables as $(\boldsymbol{X}, \boldsymbol{Z})$ with a corresponding partition of the parameter vector as fixed and random effects $(\boldsymbol{\beta}, \boldsymbol{b})$ as all the parameters are regarded as random variables. Thus Equation 7.1 can simply be expressed as

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}. \quad (7.2)$$

The difference between the fixed and random effects comes in the specification of the prior distributions of β which is usually assumed to be multivariate normal with variance-covariance matrix Λ , a precision matrix, as explained in Bernardo & Smith (2000). In particular, for the fixed effects, the elements of this matrix are known constants expressing subjective prior knowledge whereas for the random effects they depend on unknown hyper-parameters denoted by θ which are estimated from the data. Hence only a hyper-prior distribution for θ is required to complete the Bayesian formulation. For simplification, Clayton (1996) suggested adopting improper uniform priors for fixed effects resulting in a partitioning of the prior precision matrix of the form

$$\Lambda = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Lambda_1(\theta) \end{bmatrix} \quad (7.3)$$

A gamma hyper-prior distribution for θ with a scale parameter α and a shape parameter of v was used. A combination of the Gibbs sampling algorithm by Geman & Geman (1984) and the Metropolis-Hastings algorithm by Metropolis et al. (1953) and Hastings (1970) was used to sample the hyper-parameters from their full conditional. Breslow & Clayton (1993) discussed a wide variety of application of these iterative methods to GLMM.

To compare the models having different sets of parameters for a given set of priors, similar to the use of the Akaike information criterion (AIC) for nested models under GLM, we used a generalized measure of the AIC known as the deviance information criterion (DIC). The deviance information criterion is calculated from the deviance (D) which is expressed as

$$D = -2\log(P(y|\Omega)) \quad (7.4)$$

where Ω is some parameter set of the model. Under the MCMC algorithm, the deviance is calculated at each iteration, and hence a mean deviance \bar{D} can be calculated over all

iterations. As explained by Hadfield (2010), the deviance is calculated at each mean estimate of the parameters ($D(\bar{\Omega})$) and hence the deviance information criterion is given by

$$DIC = 2\bar{D} - D(\bar{\Omega}) \quad (7.5)$$

7.2.2 Statistical computing

We used the package **MCMCglmm** by Hadfield (2014) in **R** to compute the posterior distribution of the parameters using an MCMC algorithm. The algorithm is iterative and is based on the proposal, at each step, of a new value for a given parameter as a function of the other parameters in the model. In **MCMCglmm** package, the prior argument takes a list of the three elements specifying the priors for fixed effects, the random effects as well as the residuals. For the fixed effects, a multivariate normal prior distribution was specified with mean vector μ and a covariance matrix V . For the prior distribution of the random effects, we used an non-informative inverse-gamma (a special case of the inverse Wishart distribution) parameterized by v and V as suggested by Hadfield (2010).

In particular we used the function **MCMCglmm** in the **MCMCglmm** package to determine the marginal posterior distribution by drawing random samples from the joint distribution of the prior and the likelihood of the data. To enhance movement of the chain through the parameter space the package uses a combination of the Gibbs sampling and the Metropolis-Hastings updates, see Hadfield (2010). We calculated the deviance information criterion using the probability of the data given the parameters $(\boldsymbol{\beta}, \mathbf{b})$.

7.3 Results and discussion

Table 7.1 presents the estimates of the parameters for the marginal posterior distribution obtained by random draws of samples from the joint distribution together with the 95% credibility intervals and the the p-values for each parameter estimate. The results were obtained from running 25 000 iterations with a burn-in phase of 1 000 and a thinning interval of 100. The deviance information criterion for the model was 11 249.64. Table 7.2 gives the variance components for the random effects for cluster, household and the residuals.

Table 7.1: The marginal posterior distribution for the fixed effects

Coefficient	Posterior Mean	95% Cred. Int.	p-value
Intercept	-3.587	(-3.887, -3.301)	< 0.001
Gender			
Male	-0.221	(-0.559, 0.056)	0.164
Age group			
20 – 24	1.021	(0.719, 1.349)	< 0.001
25 – 29	1.686	(1.389, 2.041)	< 0.001
30 – 34	2.058	(1.760, 2.398)	< 0.001
35 – 39	2.024	(1.706, 2.383)	< 0.001
40 – 44	1.566	(1.218, 1.920)	< 0.001
45 – 49	1.421	(1.026, 1.795)	< 0.001
50 – 54	1.469	(0.991, 1.935)	< 0.001
Marital status			
Married	-0.022	(-0.267, 0.240)	0.826
Divorced	0.723	(0.450, 1.024)	< 0.001
Widowed	1.659	(1.346, 1.956)	< 0.001
Literacy level			
Partially	0.441	(0.193, 0.658)	< 0.001
Literate	0.267	(0.0922, 0.470)	0.009
Place of residence			
Urban	0.520	(0.321, 0.725)	< 0.001
Sex of household			
Female	0.159	(0.051, 0.265)	0.013
Wealth index			
Poorer	-0.013	(-0.170, 0.140)	0.868
Middle	-0.001	(-0.153, 0.192)	0.937
Richer	-0.241	(-0.439, -0.025)	0.027
Richest	-0.576	(-0.803, -0.335)	< 0.001
Gender*Age group			
M:20 – 24	-0.742	(-1.194, -0.307)	< 0.001
M:25 – 29	-0.659	(-1.130, -0.178)	0.003
M:30 – 34	-0.625	(-1.168, -0.135)	0.007
M:35 – 39	-0.210	(-0.730, 0.294)	0.470
M:40 – 44	0.555	(-0.020, 1.159)	0.060
M:45 – 49	0.730	(0.144, 1.232)	0.008
Gender*Marital status			
M:Married	0.413	(0.066, 0.825)	0.043
M:Divorced	0.515	(0.017, 1.007)	0.049
M:Widowed	0.788	(0.044, 1.461)	0.034

Table 7.2: Variance components for the random effects

Coefficient	Posterior Mean	95% Credibility Interval
Cluster	0.233	(0.168, 0.311)
Household	0.009	(0.002, 0.020)
Residual	0.204	(0.090, 0.329)

The results displayed in Table 7.1 show that HIV prevalence is dependent on one's age, marital status, literacy level, place of residence, sex of head of household, socio-economic status (measured by wealth index) and age by gender and gender by marital status interaction effects. It is evident that males have lower odds of HIV (OR = 0.802, 95% CI = 0.572 – 1.058) than the females. This is mainly due to biological factors that make females to be more at risk to HIV than males especially during vaginal sexual intercourse. In addition, there are traditional and social norms such as polygamy that make females more vulnerable to HIV than males. The results also show that HIV prevalence increase with age peaking at the 30 – 34 year age group. Relative to the 15 – 19 year old individuals, the 20 – 24 year olds have more than two times higher odds of HIV, OR = 2.776, 95% CI = 2.502 – 3.854, whereas the 30 – 34 (with highest odds) have over seven times higher odds of HIV, (OR = 7.830, 95% CI = 5.812 – 11.001).

The results further show that the odds of HIV are slightly lower (OR = 0.978, 95% CI = 0.766 – 1.271), for married, more than twice higher, (OR = 2.061, 95% CI = 1.568 – 2.784), for the divorced and more than five times higher, OR = 5.254, 95% CI = 3.842 – 7.071, for the widowed as compared to the single/never married. The lower of HIV among the married is possibly because the married are often in more stable sexual relationships as compared to the single/never married who are more likely to be involved in multiple sexual partnerships. The divorced are possibly more likely to be involved in multiple-sexual partners whereas the widowed might have lost their partners due to AIDS related illnesses hence the relatively higher risk of HIV than the single/never married. The same interpretation can be applied to the literacy level, place of residence, sex of head of household and wealth index. The gender by age group interaction shows

the additional effect of gender on the odds of HIV with age. The results show that the odds of HIV increase faster among males than females with age. This is possibly due to what is generally believed that females have sexual debut at a younger age than males. In addition, it is also believed that young females often engage in sexual relationships with older males. The gender by marital status interaction gives the additional gender effect on the risk of HIV by marital status. It is evident that the risk of HIV varies greatly between males and females across different categories of marital status. In particular, the odds are generally higher in males than in females in the category married, divorced and widowed with corresponding OR = 1.511, 1.674 and 2.199 respectively.

The results displayed in Table 7.2 show that there is a substantially high cluster to cluster variability, $\sigma_{\text{cluster}}^2 = 0.233$, 95% CI = 0.168 – 0.311, and individual to individual variability, $\sigma_{\text{residual}}^2 = 0.204$, 95% CI = 0.090 – 0.329 in the risk of HIV whereas there is low household to household, $\sigma_{\text{household}}^2 = 0.009$, 95% CI = 0.002 – 0.020. This in turn gives a high intra-cluster correlation, ICC = 0.522 which is indicative of strong correlation among individuals within the same cluster whereas there is low intra-household correlation, ICC = 0.020 showing relatively weak correlation among individuals from the same household.

7.4 Conclusions

The analysis of data that exhibit multi-layered structure requires methods that are capable of accounting for the underlying processes that are responsible for the observed data structure. The hierarchical models are suitable for explaining sources of variability in a response variable across different levels in multi-layered and clustered data. We adopted a Bayesian approach to computing a hierarchical logistic regression model from a GLMM perspective to explain the variation in HIV prevalence. Under the Bayesian paradigm, maximum likelihood estimation methods of parameter estimation are not

tractable, hence we used iterative methods based on MCMC simulations. Posterior summary measures in the form means and credibility intervals were computed from drawing random samples from the posterior distribution. The prior distributions for the fixed effects were assumed to be multivariate normal parameterized with mean μ and variance-covariance matrix V , whereas prior distributions for the random effects were assumed to be non-informative inverse-gamma. It was established that HIV prevalence is dependent on one's gender, age, marital status, literacy level, sex of head of the household, wealth index, place of residence and interaction effects of gender by age and gender by marital status. Furthermore, the results show a substantially spatial heterogeneity in HIV prevalence as evidenced by high cluster to cluster and household to household variability. Combining hierarchical modelling and Bayesian methodology enhanced the identification of risk factors for HIV as well as the estimation of prevalence.

Chapter 8

A predictive model for HIV using a semi-parametric spline approach

Abstract

The generalized additive models (GAMs), extensions of the generalized linear models (GLMs), enable exploring the non-linear dependence of a response variable on predictor(s) variable(s) in a non-parametric or a semi-parametric way. GAMs are often used when there is no *a priori* reason for determining a particular response function in a regression setting and allow the data to “speak for themselves”. This is achieved via the use of smoothing functions.

A semi-parametric logistic GAM for HIV on demographic, socio-economic and behavioural variables using population-based 2010-11 Zimbabwe Demographic and Health Surveys (2010-11ZDHS) data is fitted. The dependence of HIV on the non-parametric smooth function of the variable age, as a continuous covariate, and parametrically on the other demographic and socio-economic factors was investigated. The results were compared with the results of an equivalent ordinary logistic GLM from a likelihood perspective.

8.1 Introduction

The knowledge and accurate accounting of the nature of the relationship or dependence between a response variable and covariates is essential in a modelling approach to statistical analysis of data. Methods that are flexible in exploring the dependence of the response to covariates by allowing the observed data to influence (in a non-parametric fashion without imposing rigid assumptions about the form) the nature of the relationships are useful especially in many biological, health and social research.

The formulation of the generalized linear models (GLMs) that were first introduced by Nelder & Wedderburn (1972) and expanded by McCullagh & Nelder (1989) assumes that the dependence of the response to the covariates is linear, however in practice this is not always the case. Instead nonlinear relationships exist. For instance the relationship between HIV prevalence and age as indicated in Figure 8.1. It is clear that, HIV prevalence depends non-linearly on age and any modelling approach that does not account for the non-linear nature of the relationship becomes potentially inadequate.

We consider the generalized additive models (GAMs) by Hastie & Tibshirani (1986) and Hastie & Tibshirani (1990) as extensions of GLMs in which the linear predictor involves a sum of smooth function of the covariates. Essentially, a GAM is a non-parametric or semi-parametric regression technique not restricted to linear relationships and are flexible with regards to the statistical distribution of the data, Swartman et al. (1995). The models provide a flexible specification of the dependence of a response on the covariates by expressing the model in terms of smooth functions rather than giving detailed parametric relationships, Wood (2006). The fundamental idea behind the smoothing function, as given by Hastie & Tibshirani (1990) is to “let the data show us the appropriate functional form” instead of imposing rigid parametric assumption regarding the dependence. The beauty of the GAM approach is that, it can help prevent model mis-specification and that they enhance revealing of information about the exact relationship between the predictors and the response variable that are not possible with

standard modelling techniques. In addition, because of their non-parametric or semi-parametric nature, GAMs are usually less restrictive and are more robust against model assumption violations.

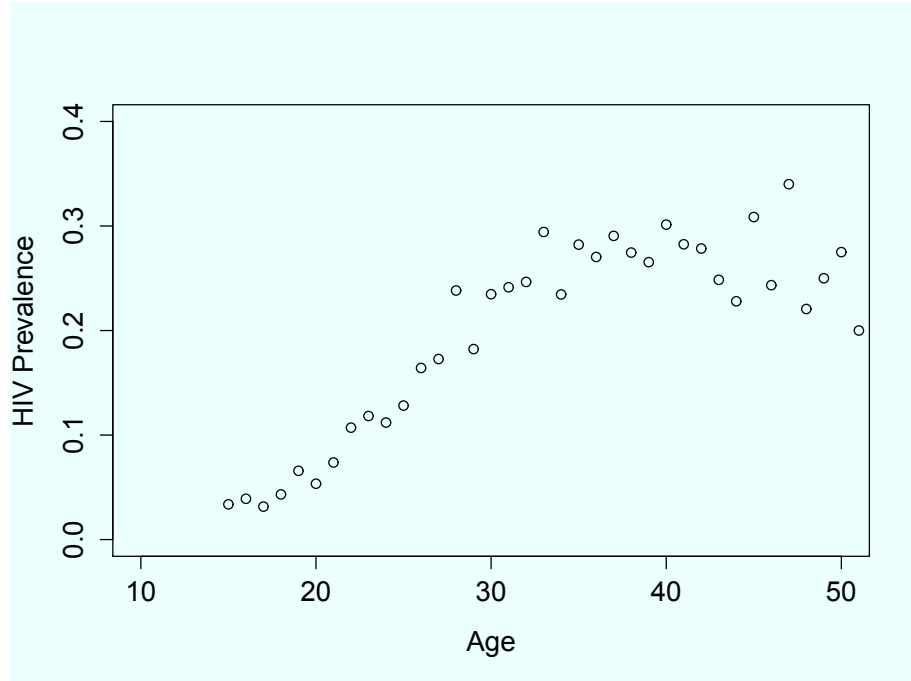


Figure 8.1: Plot showing the relationship between HIV prevalence and age

GAMs have been widely applied (non- or semi-parametrically) in many fields of research where a non-linear relationship between the response and the covariates is evident. For instance Shiboski (1998) used GAMs to explain the variations in breastfeeding practices in developing countries and epidemiological studies of HIV transmission using current status data from a semi-parametric perspective. More recently, Shen (2011) applied GAMs (and their generalized additive mixed models (GAMM) extensions) in a non-parametric and semi-parametric regression way to model the complex non-linear relationships between medication adherence (to antiretroviral therapy (ART)), viral load change over time and other factors such as medication regimen type and medication naive versus experienced at enrollment. In biological studies, GAMs have been applied to explore relationship between environmental factors and the spatial distri-

bution of fish, see for example Murase et al. (2009) and Swartman et al. (1995) and in the dependence of the spatial distribution of plant species on climatic variables, see for example Yee & Mitchell (1991). Hastie & Tibshirani (1987) discussed a variety of applications of the GAMs in an analysis of covariance and a logistic regression formulation. The current chapter develops a model from a GAM framework to explain the variation in HIV on demographic and socio-economic factors of respondents using population-based survey data. In particular, the 2010-11 Zimbabwe demographic and health surveys (2010-11ZDHS) data were used for the analysis. Section 8.2 gives the theory of the GAMs and detailed descriptions of the statistical computing resources and the data used. The results of the analysis and the discussion of the results are presented in Section 8.3. Section 8.4 presents the concluding remarks.

8.2 Methods

8.2.1 Generalized additive models

Noting the non-linear nature of the relationship between HIV and age, we fitted a GAM following Hastie & Tibshirani (1986) and Hastie & Tibshirani (1990) for modelling the dependence of a response variable Y (for example for the current study HIV prevalence) on covariates X_i , for $i = 1, \dots, p$ (the demographic, socio-economic and behavioral factors) in a non-parametric way. In particular, a GAM is an extension of the GLMs that were first discovered by Nelder & Wedderburn (1972) and later expanded by McCullagh & Nelder (1989). Under the GAM, the linear predictor depends linearly on some unknown smooth functions of the predictor variables. The non-parametric form of the dependence makes the model more flexible and more robust.

Under the GLM framework, following McCullagh & Nelder (1989) the response Y

is assumed to have exponential density given by

$$f(y, \theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}, \quad (8.1)$$

where θ is called the natural parameter, ϕ is called the dispersion parameter and $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions. It is assumed that the $E(Y) = \mu$ is related to X_i via a link function $g(\mu) = \eta$ where $\eta = \alpha + \sum_j X_j \beta_j$, whereas μ is related to the natural parameter by $\mu = b'(\theta)$. Here η is called the linear predictor. The difference between a GLM and a GAM comes about in that, under the GAM the mean $\mu = E(Y)$ is linked to the predictors via

$$g(\mu) = \alpha + \sum_{j=1}^p f_j(X_j), \quad (8.2)$$

where $f_j(X_j)$ are smooth functions of the covariates to be estimated from the data. The linear predictor in Equation 8.2 can also be expressed as a mixture of smoothed functions and linear functions of other predictors.

For the estimation of the smooth functions, $f_j(\cdot)$'s we employed the forward stepwise local scoring algorithm utilizing the scatter-plot smoothers as a generalizations of the Newton Raphson and the Fisher scoring procedure used to compute the least squares and the maximum likelihood estimates under the GLM approach, Hastie & Tibshirani (1986). Alternatively, the smooth functions can be estimated via the local likelihood procedure that assumes that, locally $f_j(\cdot)$ is linear and fits a line in the neighbourhood around each X value. Use of regression splines is commonly used for these two (the forward stepwise local scoring algorithm and the local likelihood procedure), see for example Wood (2006). The degree of smoothness in the functions can be determined and a “wiggleness” penalty is added to the fitting process to account of the respective level of smoothness. This results in penalized versions of the least squares or maximum likelihood or iterative reweighted least squares (IRLS) methods.

Suppose $\mathbf{X} = (X_1, \dots, X_p)$ denotes the p -dimensional vector of covariates for a

response Y , the general model specifies $E(Y|\mathbf{X}) = \mu$ and $g(\mu) = \eta(\mathbf{X})$ where η is a function of the p variables, as by Hastie & Tibshirani (1986), we assumed that

$$Y = \eta(\mathbf{X}) + \epsilon, \quad (8.3)$$

where $\eta(\mathbf{X}) = E(Y|\mathbf{X})$, $\text{Var}(Y|\mathbf{X}) = \sigma^2$ and the errors are independent of \mathbf{X} . The estimation of $\eta(\mathbf{X})$ under the local scoring is done via the least squares criterion $E(Y - \eta(\mathbf{X}))^2$ with the use of scatter-plot smoothers. However, as described by Hastie & Tibshirani (1986) in higher dimensions, near neighbours cease to be local and the scatter-plot smoothing becomes inadequate, which turns out to be enough motivation for the use of additive models. Thus we consider an additive regression model given by

$$E(Y|\mathbf{X}) = \alpha + \sum_{j=1}^p f_j(X_j),$$

where $E[f_j(X_j)] = 0$ for $\forall j$. Under the local scoring algorithm, following Hastie & Tibshirani (1986) suppose that

$$Y = \alpha + \sum_{j=1}^p f_j(X_j) + \epsilon, \quad (8.4)$$

is correct, and we assume that $\alpha, f_1(\cdot), \dots, f_p(\cdot)$ are known. If the partial residuals, R_j can be defined as

$$R_j = Y - \alpha - \sum_{k \neq j} f_k(X_k), \quad (8.5)$$

then $E(R_j|X_j) = f_j(X_j)$ thus minimizing $E\left(Y - \alpha - \sum_{k=1}^p f_k(X_k)\right)^2$. However, $f_k(\cdot)$'s are unknown but they provide a way for estimating $\hat{f}_j(\cdot)$ given estimates of $\{\hat{f}_i(\cdot), i \neq j\}$. This leads to an iterative procedure known as back-fitting algorithm proposed by Friedman & Stuetle (1981). Convergence of the iteration is confirmed if the residual sum of

squares (RSS) does not change, where

$$\text{RSS} = E \left(Y - \alpha - \sum_{j=1}^p f_j^m(X_j) \right)^2.$$

Here $f_j^m(\cdot)$ denotes the estimates of $f_j(\cdot)$ at the m th iteration.

Assessment of model fit is likelihood-based or deviance-based hence fitted models are directly comparable with GLMs using likelihood techniques such as the AIC or the classical tests based on model deviance such as the χ^2 or F tests.

8.2.2 Logistic generalized additive regression

Under the GLM, the logistic regression that relates the mean of a binary response $\pi(\mathbf{x}_i) = P(Y = 1)$ is given as

$$\text{logit}(\pi(\mathbf{x}_i)) = \log \left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right) = \mathbf{x}_i' \beta \quad (8.6)$$

We computed an additive logistic regression model in which the linear term in Equation 8.6 is replaced by its additive equivalence of the functional form given by

$$\text{logit}(\pi(\mathbf{x}_i)) = \log \left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right) = \alpha + \sum_{j=1}^p f_j X_j \quad (8.7)$$

where the f_j 's are smooth functions and $E[f_j(\mathbf{x}_i)] = 0$ as explained in Subsection 8.2.1 above. The logistic regression model represented in Equation 8.7 is additive on the logit scale but not on the probability scale.

Estimation of the smooth functions follows the theory given in Subsection 8.2.1 above. In a semi-parametric logistic additive model, the parametric terms can be estimated using the ordinary GLM methods such as the iterative re-weighted least squares (IRLS) and maximum likelihood due to Nelder & Wedderburn (1972).

8.2.3 Statistical computing

The computations of the model were done using **mgcv** package by Wood (2006) in **R**. The package uses a local scoring algorithm as explained in Section 8.2, to iteratively fit weighted additive models through the back-fitting procedure. Handling the smoothness section in the package is based on a Bayesian smoothing model, which enables simulations from the posterior distribution of the model coefficients and provide credible intervals. The **gam** function in the **mgcv** package allows inclusion of both smooth functions, $s(\cdot)$ with options for controlling the smoothness, and other covariates as parametric linear. Each predictor in the model is considered individually then split into sections delimited by ‘knots’ and then fit polynomial functions to each section separately. All functions are based on the theory of GAM by Hastie & Tibshirani (1986) and Hastie & Tibshirani (1990). The **glm** functions is used to compute an ordinary GLM. Both the **glm** and the **gam** functions enable fitting the models allowing for specification of the link that is dependent on the error structure.

8.2.4 The data

This section is considered together with the general data description given in section 2.2. For administration purposes, Zimbabwe is divided into ten provinces. During the 2002 population census (which was used as the sampling frame in the 2010-11 ZDHS) each province was subdivided into districts and each district is made up of wards and the wards consist of a number of Enumeration Areas (EAs). For the current research the response variable is HIV status, a binary variable indicating whether a respondent is HIV positive or negative. The study investigates the relationship between HIV and socio-economic and demographic factors of the population using a logistic GAM. The socio-economic and the demographic variables (that were used as the predictors) are selected as those factors thought to influence HIV infection as informed by the proximate determinants conceptual framework as described by Boerma et al. (2003). These factors

include age, gender, marital status, education level, economic status (household wealth), religion, province and place of residence (whether rural or urban). Because of the non-linear dependence of HIV on age, a non-parametric smooth function of age and parametric effects of the other factors were included. The sample consists of 17 434 respondents, 14 491 with non-missing value and an additional 2 943 with missing values. The current chapter assumed a complete case analysis.

8.3 Results and discussion

We computed a logistic GAM using the **gam** and the **glm** functions in **mgcv** package and utilizing the theory presented in Section 8.2 above. In particular we constructed a logistic GAM with a non-parametric smooth function of age and parametric effects of the other predictors. For comparison purposes, results of an ordinary logistic GLM were presented alongside the logistic GAM results. Table 8.1 displays the results (parameter estimates, standard errors and p-values) for the best GAM based on the AIC and the percentage of deviance explained, whereas Table 8.2 displays the parameter estimates, standard errors and p-values for the GLM.

Table 8.1: Results of a generalized additive model (GAM) with (a) the parametric terms and (b) a smooth term of age

<i>(a) Parametric coefficients</i>			
Coefficients	Estimate	S. E.	p-value
Intercept	-2.081	0.134	< 0.001
Gender			
Male	-0.655	0.121	< 0.001
Marital status:			
Married	-0.100	0.104	0.340
Divorced	0.688	0.130	< 0.001
Widowed	1.366	0.136	< 0.001
Place of residence			
Urban	0.177	0.052	0.001
Literacy			
Partially	0.479	0.128	< 0.001
Literate	0.224	0.105	0.034
Gender*Marital status			
Male : Married	0.566	0.136	< 0.001
Male : Divorced	0.597	0.205	0.004
Male : Widowed	1.132	0.291	< 0.001
<i>(b) Approximate significance of smooth terms</i>			
Parameter	estimated df	χ^2 - value	p-value
s (age)	3.675	233.7	< 0.001

$AIC = 11527.73$ Deviance explained=11.03% Residual deviance = 11498

Table 8.2: Parameter estimates, standard errors and p-values for an ordinary logistic GLM

Coefficients	Estimate	S. E.	p-value
Intercept	-3.507	0.148	< 0.001
Age	0.036	0.003	< 0.001
Gender			
Male	-0.559	0.119	< 0.001
Marital status:			
Married	0.462	0.095	< 0.001
Divorced	1.245	0.122	< 0.001
Widowed	1.910	0.131	< 0.001
Place of residence			
Urban	0.209	0.051	< 0.001
Literacy			
Partially	0.489	0.128	< 0.001
Literate	0.308	0.105	0.003
Gender*Marital status			
Male : Married	0.460	0.134	0.001
Male : Divorced	0.536	0.203	0.008
Male : Widowed	0.850	0.289	0.003

$AIC = 11740$ Deviance explained=9.67% Residual deviance=11716

Comparisons of the GAM and the GLM are based on the likelihood techniques. Table 8.3 gives the AICs for the two models. Although the GAM is more parameterized and faces a stiffer penalty due to more degrees of freedom, its AIC is substantially lower. This is indicative of a better fit for the GAM than the GLM.

Table 8.3: Analysis of the AICs for the GAM and the GLM

Model	d.f	AIC
GLM	12.000	11739.650
GAM	14.675	11527.730

Table 8.4 gives the χ^2 – test based analysis of the deviances for the GAM and GLM models. The GAM also shows superiority as its residual deviance is significantly lower than the GLM (with p-value < 0.001). In addition, the results also show that the GAM explains more percentage of deviance, 11.03%, than the GLM, 9.67%. These results are consistent with results from other studies in the region, see for example Mara et al.

(2015), Liang & Weiss (2007) and Ngesa et al. (2014).

Table 8.4: Analysis of the deviances for the GAM and the GLM

Model	Res. d.f.	Res. Dev.	d.f.	Deviance	p-value
GLM	14479	11716			
GAM	14476	11498	2.675	217.27	< 0.001

We further explored the results of the GAM. The parametric estimates (Table 8.1) are presented with tests of significance against a null of zero. The approximate significant results for the smooth function term is a test of whether the smoothed function significantly reduces model deviance. The results indicate that the smooth function of age does indeed reduce the model deviance (p-value < 0.001). The plots of the smooth function for age together with the effect of each level of the other factors (gender, marital status, place of residence and literacy) and their respective 95% confidence intervals are displayed in Figure 8.2. The first levels for each factor have an effect of zero because they are the reference levels.

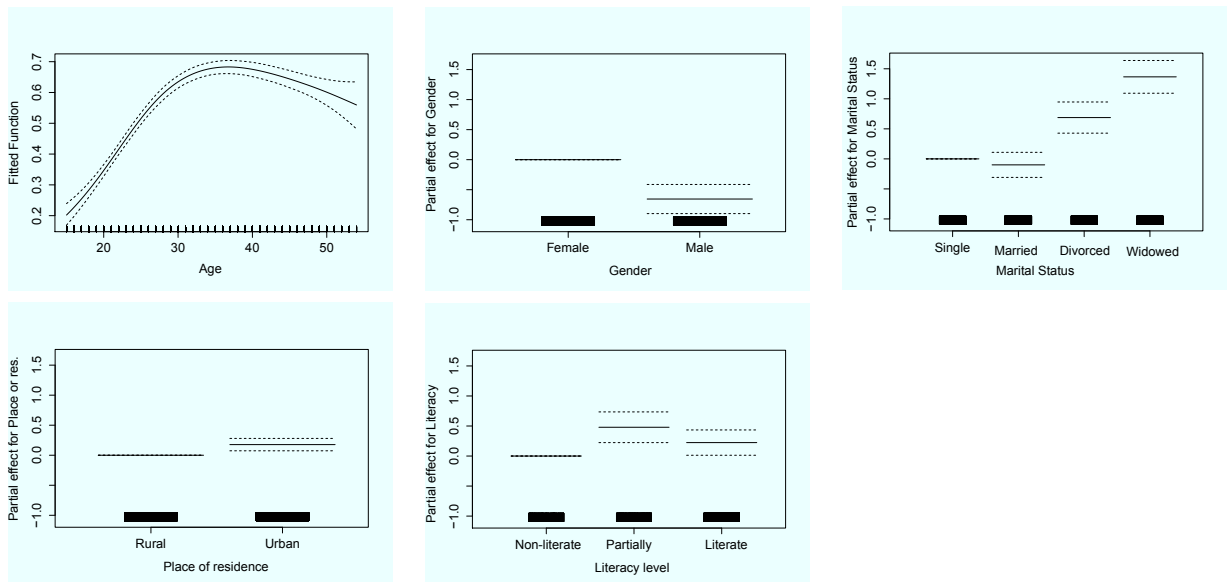


Figure 8.2: Plot of the final logistic GAM, a semi-parametric model of HIV with a smooth function, $\hat{f}(\text{age})$, of age together with the factors of gender, marital status, place of residence and literacy levels and their respective 95% confidence intervals.

Interpretation of the parametric estimates is facilitated by expressing the estimates

on an odds ratios (OR) scale. The results, as adjusted ORs, are presented in Table 8.5. The interpretation of the ORs assumes a reference level approach. Thus the OR for gender indicates the likelihood of a male individual being HIV positive as compared to a female individual controlling for the effect of the other factors in the model. Thus the odds of HIV are lower (OR = 0.519, 95% CI = 0.410 – 0.658), for a male person as compared to a female person controlling for the effects of the other factors. For the marital status factor, with reference to the single/never married, the odds of HIV are slightly lower (OR = 0.905, 95% CI = 0.738 – 1.109), for a married individual, almost twice higher, (OR = 1.99, 95% CI = 1.542 – 2.567), for a divorced person and close to four times higher, (OR = 3.92, 95% CI = 3.002 – 5.117), for a widowed person, after adjusting for the effects of the other factors. The urban residents have lower odds of HIV, (OR = 1.194, 95% CI = 1.078 – 1.322), than their rural counterparts. The gender by marital status interaction effects show the additional gender effect on marital status. Thus, with reference to single/never married females, the married and divorced males have odds of HIV almost twice (OR = 1.761, 95% CI = 1.349 – 2.299, and OR = 1.817, 95% CI = 1.216 – 2.715, respectively) and over three times higher (OR = 3.102, 95% CI = 1.754 – 5.487).

Table 8.5: Odds ratios and their corresponding 95% confidence intervals for the parametric effects

Coefficients	Odds ratios	95% Confidence intervals
Intercept	0.125	(0.096, 0.162)
Gender		
Male	0.519	(0.410, 0.658)
Marital status:		
Married	0.905	(0.738, 1.109)
Divorced	1.990	(1.542, 2.567)
Widowed	3.920	(3.002, 5.117)
Place of residence		
Urban	1.194	(1.078, 1.322)
Literacy		
Partially	1.614	(1.256, 2.075)
Literate	1.251	(1.018, 1.537)
Gender*Marital status		
Male : Married	1.761	(1.349, 2.299)
Male : Divorced	1.817	(1.216, 2.715)
Male : Widowed	3.102	(1.754, 5.487)

8.4 Conclusion

The assumption of linear dependence of a response variable to covariates that is often made under GLMs does not always hold in many practical investigations. There are a variety of functional forms of the relationship that exist between a response and the covariate(s). Thus approaches that are flexible to explore non-parametric smooth functions of the relationships are often applied. We considered a GAM to explore the dependence of HIV on demographic and socio-economic factors. In particular, a semi-parametric logistic GAM with a non-parametric smooth function for age and parametric effects for gender, marital status, place of residence and literacy was computed. For comparative purposes, an equivalent GLM that assumed parametric effects of all the covariates was also constructed. Based on likelihood measures (AIC and deviances) the results show that the GAM is superior to the GLM in terms explaining the variation of HIV using the demographic and socio-economic variables.

Chapter 9

Conclusions and future research

9.1 Introduction

This chapter presents a summary of all the conclusions from all the preceding chapters. Furthermore, potential strengths and shortcomings of the study and directions for future research are also given.

9.2 Summary of conclusions

The study focused on estimating HIV prevalence using population-based survey data. Methods of analyzing survey data that take account of the complex sampling schemes commonly encountered in practice were explored and applied. Models built both from a frequentist and a Bayesian perspective were computed. In particular, different forms of logistic regression models that account for the complex survey design, the hierarchical and clustering nature of the data, the prior beliefs about model parameters and the non-linear (in a non- and semi-parametric) relationship of the response and the covariates were fitted to explain the variation in HIV on demographic, socio-economic and behavioural factors. In addition, design-consistent national and domain level (crude and adjusted) estimates of HIV prevalence were computed. The domains (which can be

regarded as risk factors) considered were based on administrative provinces, gender, age, marital status, place of residence, literacy levels and wealth index. Additional complexities brought about by missing data due to non-response, which are not only pervasive but also inevitable in survey research were also considered. A plausible method for handling missing data via multiple imputations accounting for the structure of the underlying population was considered.

The multiple sources of variability in a response that results from the hierarchical and multilevel structure of most practical survey data that is brought about by the prominent features of the target population were accounted for in the modelling process. In addition, the dependencies of units that are induced by the complex sampling designs, that involve stratification and multi-stage clustering that render most classical statistical methods based on the assumption of independence inappropriate, were also considered. The data structures and the dependencies among units that share the same cluster were correctly accounted for by the use of hierarchical models built from a GLMM framework.

The data used for the analyses were obtained from the 2010-11 Zimbabwe Demographic and Health Surveys (2010-11ZDHS). As with most practical survey data, the 2010-11ZDHS data were characterized by missing data, and hence the study explored the available methods of handling missing data in surveys. In particular, a procedure based on multiple imputation of the missing observations, and simultaneously account for the variability due to the missing values was used to ‘fill in’ the missing data. For comparative purposes, results from a complete case analysis (based on a list-wise deletion of cases with missing values) and the multiple imputations were presented for the three modelling approaches.

The overall design-consistent estimate for HIV prevalence for the entire population was found to be 15.35% with a 95% confidence interval of (14.72%, 15.97%). For more detailed analysis of the prevalence of HIV across different subgroups of the popula-

tion, domain level estimates were also computed and presented in Chapter 3 from the complete case analysis, and in Chapter 4 from the multiple imputations. The study established that HIV prevalence varies greatly within the respective domains.

The results show that HIV prevalence is dependent on one's gender, age, marital status, place of residence and the gender by age and gender by marital status interaction effects. Under the hierarchical modelling approach in Chapter 5, the results show significant household to household and cluster to cluster (enumeration area) variabilities effects on the HIV. Chapter 6 brings in the Bayesian paradigm that utilises the prior probability distributions of variables. The non-linear relationship between HIV and age was captured in a semi-parametric fashion via the use of GAMs.

9.3 Strengths and limitations of the study and future research

9.3.1 Strengths

The study draws its main strength from the appropriate application of sound statistical methods coupled with the utilization of advanced statistical and computational tools that are available in statistical software packages such as **R** in estimating HIV prevalence. The multiple imputation technique has, not only the strength to 'fill in' missing observations (common in HIV research), but also to account for the uncertainty due to the imputation process itself. In addition, noting that missing data are inevitable (especially in survey data), pervasive and have severe consequences if not properly handled, use of sound statistical methods and computing resources to estimate disease measures of interest and appropriate measures of variability (that account for both the sampling mechanism and the imputation process) can enhance the validity of the statistical interpretations and inferences. This study has demonstrated that statistical methods for

handling missing data have the potential to enhance estimation of HIV and avoid loss of statistical information that come with, for example, deleting cases with missing values. The use of DHS data also brings in an additional advantage in that the data are collected by highly trained statisticians with excellent expertise in survey methodology. Furthermore, incorporation of prior knowledge about the parameters, the cornerstone of Bayesian statistical analysis, in the modelling process gives the study an edge in that it allows findings obtained in previous studies to be used in order to improve the results. The use of GAMs also provides flexibility and robustness in the modelling process by allowing the data to determine the form of the relationship that exist between variables without imposing rigid assumption on the relationships.

9.3.2 Limitations

Potential drawbacks of the current research come from the use of secondary data which often leaves the data analyst with limited control over the data collection process. In addition, and particularly for the current research, a major drawback of using secondary is the limited knowledge about the reasons for the missing value. However this is not to downplay the importance of DHSs which are carefully designed, by a team of highly trained statisticians with excellent expertise in survey methodology, to collect population level information which is very important for public health policies. Some of the **R** packages such as **mi**, although very powerful and flexible comes with their own limitations that they cannot allow users to alter the prior distributions for the conditional imputation models used under the Bayesian paradigm. Therefore further methodological and software developments research is necessary in order to make the approach even more flexible. Further work on the problem as a future extension is possible with inclusion of methods that allow for MNAR assumption by means of sensitivity analysis.

In in addition, the cross-sectional nature of the study, that does not allow the time dependence of HIV prevalence does not allow analysis of the results over time. It would

be interesting to compare results from different surveys once HIV data become available in future population-based surveys. The study also does not allow multiple or repeated measurements on respondents that would aid in tracking a participant's status in follow-up visits and response to intervention over time as in a longitudinal analysis see for example Diggle et al. (2002). Although the study allows one to provide for the likelihood of a randomly selected individual being infected with HIV, it does not enable one to determine how long someone has been infected. In addition, incorporation of spatial components can allow identification of HIV hot spots in the population adding more relevance to the study especially as a tool for devising targeted intervention programs.

9.3.3 Direction for future research

As directions for future research, incorporating, time dependency and longitudinal components can enhance analysis of HIV over time. With the use of geographic information systems (GIS), spatial analysis can allow capturing of the spatial autocorrelations and distributional (in place) issues regarding HIV prevalence, utilizing methods as given by Moore & Carpenter (1999). The authors reviewed spatial analytic methods commonly used in biological and health research, such techniques as disease mapping, clustering techniques and diffusion studies. Complementing the cross-sectional approach with the longitudinal methods and spatial analysis has potential to enhance a good understanding of HIV prevalence. In addition, the multiple imputation technique can be strengthened and also incorporate sensitivity analysis.

Bibliography

- Adler, N. E. (2006). Overview of Health Disparities. In *Examining the health disparities research plan of the National Institute of Health: Unfinished business*. Washington: National Academic Press.
- Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. Wiley Series in Probability and Statistics.
- Anderson, R., Anker, M., Asamoah-Odei, E., Berkeley, S., Carael, M., & Castilho, E. (1999). *Trends in HIV Incidence and Prevalence*. UNAIDS, Geneva, Switzerland.
- Baraldi, A. N. & Enders, C. K. (2010). An Introduction to Modern Missing Data Analysis. *Journal of School Psychology*, 48, 5–37.
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *Linear mixed-effects models using Eigen and S4*. U package version 1.1-7.
- Bauwens, L., Lubrano, M., & Richard, J. F. (2000). *Bayesian Inference in Dynamic Econometric Models*. Oxford University Press.
- Bedrick, E. J. (1983). Adjusted Chi-squared Tests for Cross-classified Tables of Survey Data. *Biometrika*, 70, 591–595.
- Berger, Y. & Skinner, C. (2004). *Variance estimation for unequal probability designs*. Technical report, DACSEIS deliverables.

- Bernardo, J. & Smith, A. (1994). *Bayesian Theory*. John Wiley and Sons, Ltd.
- Bernardo, J. M. & Smith, A. F. M. (2000). *Bayesian Theory*. Wiley Series in Probability and Statistics.
- Bertozzi, S., Padian, N. S., Wegbreit, J., de Maria, L. M., Feldman, B., Gayle, H., Gold, J., Grant, R., & Isbell, M. T. (2006). *HIV/AIDS Prevention and Treatment*, chapter 18, (pp. 331–369). World Bank.
- Binder, D. A. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review*, 51, 279–292.
- Bobat, R., Moodley, D., Coutsooudis, A., & Coovadia, H. M. (1997). Breastfeeding by HIV-1 infected women and outcome in their infants: a cohort study from Durban, South Africa. *AIDS*, 11, 1627–1633.
- Boerma, J. T., Ghys, P. D., & Walker, N. (2003). Estimates of HIV-1 Prevalence from National Population-based Surveys as a New Gold Standard. *The LANCET*, 362, 1929 – 1931.
- Bolstad, W. M. (2007). *Introduction to Bayesian Statistics*. John Wiley and Sons, Ltd.
- Box, G. E. P. & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. John Wiley and Sons, Ltd.
- Breslow, N. E. & Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 88, 9–25.
- Breslow, N. E. & Holubkov, R. (1997). Weighted Likelihood, Pseudo-likelihood and Maximum Likelihood Methods for Logistic Regression Analysis of Two-stage Data. *Statistics in Medicine*, 16, 103–116.
- Brick, J. M. & Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215–238.

- Brooks, S. P. (1998). Quantitative Convergence Assessment for Markov Chain Monte Carlo via cusums. *Statistics and Computing*, 8, 267–274.
- Brooks, S. P. & Roberts, G. O. (1998). Convergence Assessment Techniques for Markov Chain Monte Carlo. *Statistics and Computing*, 8, 319–335.
- Brus, D. J. & de Gruijter, J. J. (1997). Random Sampling or Geostatistical Modelling? choosing between design-based and model-based sampling strategies for soil. *Geoderma*, 80, 1–44.
- Buseh, A. G., Glass, L. K., & McElmurry, B. J. (2002). Cultural and Gender Issues Related to HIV/AIDS Prevention in Rural Swaziland: A Focus Group Analysis. *Health Care for Women*, 23, 173–184.
- Bwayo, J. J., Mutari, A. N., Omari, M. A., Kreiss, J. K., Jaoko, W., & Sekkade-Kigundu, C. (1991). Long distance truck drivers 2: Knowledge and Attitudes concerning Sexually Transmitted Diseases and Sexual Behaviour. *East African Medical Journal*, 68, 714–719.
- Calcagno, V. & de Mazancourt, C. (2010). glmulti: An R Package for Easy Automated Model Selection with Generalized Linear Models. *Journal of Statistical Software*, 34, 1–29.
- Chambers, R. L. & Skinner, C. J. (2003). *Analysis of Survey Data*. Wiley Series in Survey Methodology.
- Champredon, D., Bellan, S., & Dushoff, J. (2013). HIV Sexual Transmission Is Predominantly Driven by Single Individual Rather than Discordant Couples: A Model-Based Approach. *Plos one*, 8, 1–10.
- Clayton, D. G. (1996). *Generalized Linear Mixed Models*, chapter 16, (pp. 275–301). Chapman and Hall.

- Cochran, W. G. (1977). *Sampling Techniques*. Wiley Series in Probability and Mathematical Statistics.
- Coombs, R. W., Reichelderfer, P., & Landay, A. L. (2003). Recent Observations on HIV-type 1 Infection in the Genital Tract of Men and Women. *AIDS*, 4, 455–480.
- Coutsoudis, A., Pillay, K., Spooner, E., & Coovadia, H. M. (1999). Influence of Infant Feeding Patterns on Early Mother-To-Child Transmission MTCT of HIV-1 in Durban, South Africa: A Prospective Cohort Study. *The LANCET*, 354, 471–476.
- Cuadros, D. F., Awad, S. F., & Abu-Raddad, L. J. (2013). Mapping HIV Clustering: A Strategy for Identifying Populations at High Risk of HIV Infection in sub-Saharan Africa. *International Journal of Health Geographics*, 12, 1–9.
- Curtis, S. M., Goldin, I., & Evangelou, E. (2015). *mcmcplots*. <http://cran.r-project.org//package-mcmcplots>.
- de Finetti, B. (1937). *Foresight: Its logical laws, its subjective sources*, chapter 4, (pp. 55–187). John Wiley and Sons, Ltd.
- Degenholtz, H. B. & Bhatnagar, M. (2009). Introduction to Hierarchical Modeling. *Journal of Palliative Medicine*, 12, 631–638.
- Dempster, A. P. (1968). A Generalization of Bayesian Inference. *Journal of the American Statistical Society*, 30, 205–247.
- Dey, D. K., Ghosh, S. K., & Mallick, B. K. (2000). *Generalized Linear Models: A Bayesian Perspective*. Marcel Dekker, Inc.
- Diggle, P. J., Heagerty, P. J., Liang, K. Y., & Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford University Press.
- Dobson, A. J. & Barnett, A. G. (2008). *An Introduction to Generalized Linear Models*. Chapman and Hall.

- Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the Multilevel Rasch Model: With the lme4 Package. *Journal of Statistical Software*, 20, 1–18.
- Doucet, A., Godsill, S., & Andrieu, C. (2000). On Sequential Monte Carlo Sampling Methods for Bayesian Filtering. *Statistics and Computing*, 10, 197–208.
- Dunn, D. T., Newell, M. L., & S., P. C. (1992). Risk of HIV type 1 Transmission through Breastfeeding. *The LANCET*, 340, 585–588.
- Fabiani, M., Accorsi, S., Aleni, R., Rizzardini, G., Nattabi, B., Gabrielli, A., Opira, C., & Declich, S. (2003). Estimating HIV Prevalence and the impact of HIV/Aids on a Ugandan Hospital by Combining Serosurvey Data and Hospital Discharge Records. *AIDS*, 34, 62–66.
- Feinstein, J. S. (1993). The Relationship between Socio-economic Status and Health: A Review of the Literature. *The Milbank Quarterly*, 71, 279–322.
- Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. *The Royal Society*, 222, 309–368.
- Freeman, E. E. & Glynn, J. R. (2004). Factors affecting HIV Concordancy in Married Couples in Four African Cities. *AIDS*, 18, 1715–1721.
- Friedman, J. H. & Stuetle, W. (1981). Projection Pursuit Regression. *Journal of American Statistical Association*, 76, 829–836.
- Fuller, W. A. (1975). Regression Analysis for Sample Surveys. *Survey Methodology*, 10, 97–118.
- Gabler, S., Hader, S., & Lynn, P. (2006). Design Effects for Multiple Design Surveys. *Survey Methodology*, 32, 115–120.
- Gamerman, D. (1997). Sampling from the Posterior Distribution in Generalized Linear Mixed Models. *Statistics and Computing*, 7, 57–68.

- Gelfand, A. E., Hills, S. E. Racine-Poon, A., & Smith, A. F. M. (1990). Illustration of Bayesian Inference in Normal data Models using Gibbs Sampling. *Journal of the American Statistical Association*, 85, 972–985.
- Gelman, A., Hill, J., Su, Y., Yajima, M., & Pittau, M. G. (2015). *Missing Data Imputation and Model Checking in R*.
- Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A Weakly Informative Default Prior Distribution for Logistic and other Regression Models. *The Annals of Applied Statistics*, 2, 1360–1383.
- Gelman, A., Su, Y., Yajima, M., Hill, J., Pittau, M. G., Kerman, J., & Zheng, T. (2010). *Package arm*. Available at: <http://cran.r-project.org/web/packages/arm>.
- Geman, S. & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.
- Geweke (1992). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, chapter 5, (pp. 169–193). Oxford University Press.
- Gill, J. (2009). *Bayesian Methods: A Social and Behavioral Sciences Approach*. Chapman and Hall.
- Goldstein, H. (1986). Multilevel Mixed Linear Model Analysis using Iterative Generalized Least Squares. *Biometrika*, 73, 43–56.
- Goldstein, H. (1991). Nonlinear Multilevel Models, with an Application to Discrete Response Data. *Biometrika*, 78, 45–51.
- Goldstein, H. (2011). *Multilevel Statistical Models*. Wiley Series in Probability and Statistics.

- Goldstein, H. & McDonald, R. P. (1988). A General Model for the Analysis of Multilevel Data. *Psychometrika*, 53, 455–467.
- Gonese, E., Dzangare, J., Gregson, S., Jonga, N., Mugurungi, O., & Mishra, V. (2010). Comparison of hiv prevalence estimates for zimbabwe from antenatal clinic surveillance (2006) and the 2005-06 zimbabwe demographic and health survey. *Plos one*, 5, e13819. doi:10.1371.
- Gopal, S., Yang, Y., Bai, B., & Niculescu-mizil, A. (2012). Bayesian models for large-scale hierarchical classification. In *Neural Information Processing Systems*.
- Gregoire, T. G. (1998). Design-based and Model-based Inference in Survey Sampling: Appreciating the Differences. *Canadian Journal of Forestry Research*, 28, 1427–1447.
- Gregson, S., Garnett, G. P., Nyamukapa, C. A., Hallett, T. B., Lewis, J. J. C., Mason, P. R., Chandiwana, S. K., & Anderson, R. M. (2006). HIV Decline with Behavior Change in Eastern Zimbabwe. *AAAS/Science*, 311, 664–666.
- Gregson, S., Nyamukapa, C. A., Garnett, G. P., Mason, P. R., Zhuwau, T., Carael, M., Chandiwana, S. K., & Anderson, R. M. (2002). Sexual mixing patterns and sex-differentials in teenage exposure to HIV infection in rural Zimbabwe. *Lancet Infectious Disease*, 359, 1896–1903.
- Guo, G. & Zhao, H. (2000). Multilevel Modeling for Binary Data. *Annual Review of Sociology*, 26, 441–462.
- Hadfield, J. (2010). MCMC methods for Multi-response Generalized Linear Mixed Models: The MCMC R Project. *Journal of Statistical Software*, 23, 1–22.
- Hadfield, J. (2014). *MCMC Generalized Linear Mixed Models (MCMCglmm)*. Technical report, Cran.r-project.

- Hallett, T. B., Aberle-Grasse, J., Bello, G., Boulos, L. M., Cayemittes, M. P. A., Cheluget, B., & at al. (2006). Declines in HIV Prevalence can be Associated with Changing Sexual Behaviour in Uganda, urban Kenya, Zimbabwe and urban Haiti. *Sexually Transmitted Infections*, 82(suppl), i1–i8.
- Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample Survey Methods and Theory*. Chapman and Hall, London.
- Hansen, M. H., Madow, W. G., & Tepping, B. J. (1983). An Evaluation of Model-dependent and Probability-sampling Inferences in Sample Surveys. *Journal of American Statistical Association*, 78, 776–807.
- Hastie, T. & Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1, 297–310.
- Hastie, T. & Tibshirani, R. (1987). Generalized additive models: Some applications. *Journal Of the American Statistical Association*, 82, 371–386.
- Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall.
- Hastings, W. (1970). Monte Carlo Sampling Methods using Markov Chains and their Applications. *Biometrika*, 57, 97–109.
- Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied Survey Data Analysis*. Chapman and Hall/CRC Press.
- Heidelberger, P. & Welch, P. (1983). Simulation Run-length Control in the Presence of an Initial Transient. *Operations Research*, 31, 1109–1144.
- Holt, D., Scott, A. J., & Ewings, P. D. (1980). Ch-squared tests with survey data. *Royal Statistical Association*, 143, 303–320.

- Hosmer, D. W. & Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, A10, 1043–1069.
- Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression*. Wiley Series in Probability and Statistics.
- Humphrey, J. H., Marinda, E., Mutasa, K., Moulton, L. H., Ilif, P. J., Ntozini, R., Chidavanyika, H., J., N. K., Tavengwa, N., Jenkins, A., Piwoz, E. G., van de Pere, P., & Ward, B. J. (2010). Mother to child transmission of HIV among Zimbabwean women who seroconverted postnatally: prospective cohort study. *BioMedical Journal*, 341, c6580.
- Jiang, J. & Lahiri, P. (2006). Mixed Model Prediction and Small Area Estimation. *Test*, 15, 1–96.
- Johnson, M. L. (2010). *Essential Numerical Computer Methods*. Elsevier Inc.
- Kalton, G. & Brick, J. M. (1996). Handling Missing Data in Survey Research. *Statistical Methods in Medical Research*, 5, 215–238.
- Kalton, G. & Kasprzyk, D. (1986). The Treatment of Missing Survey Data. *Survey Methodology*, 12, 1–16.
- Kass, R. & Wasserman, L. (1996). The Selection of Prior Distributions by Formal Rules. *Journal of American Statistical Association*, 91, 1343–1370.
- Kedir, A. M. (2001). *Rural poverty report: the challenge of ending rural poverty*. Technical report, International Fund for Agriculture Development (IFAD).
- Khan, M. H. R. & Shaw, J. E. H. (2011). Multilevel Logistic Regression Analysis Applied to Binary Contraceptive Prevalence Data. *Journal of Data Science*, 9, 93–110.

- Kish, L. (1965). *Survey Sampling*. Wiley Classics Library.
- Kish, L. & Frankel, M. R. (1974). Inference from Complex Samples. *Journal of the Royal Statistical Society, Series B.*, 36, 1–37.
- Lee, E. & Forthofer, R. (2006). *Analyzing Complex Survey Data*. Sage Publications, London, UK.
- Lee, Y. & Nelder, J. A. (1996). Hierarchical Generalized Linear Models. *The Royal Statistical Society*, 58, 619–678.
- Lehtonen, R. & Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*. John Wiley and Sons, Ltd.
- Lele, S. R., Keim, J. L., & Solymos, P. (2015). *Resource Selection (Probability) Functions for Use-Availability Data*. <http://cran.r-project.org/package=ResourceSelection>.
- Lepkowski, J. M., Mosher, W. D., Davis, K. E., Groves, R. M., & van Hoewyk, J. (2006). *National Survey of Family Growth, Cycle 6: Sample Design, Weighting, Imputation and Variance Estimation*. Technical report, National Center for Health Statistics Series 2, Number 142.
- Lesaffre, E. & Lawson, A. B. (2012). *Bayesian Biostatistics*. John Wiley and Sons, Ltd.
- Liang, L. J. & Weiss, R. E. (2007). A Hierarchical Semiparametric Regression Model for Combining HIV-1 Phylogenetic Analysis Using Reweighting Algorithm. *Biometrics*.
- Lindley, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 1: Probability*. Cambridge University Press.
- Lindley, D. V. & Smith, A. F. M. (1972). Bayes Estimates for the Linear Model. *Journal of the Royal Statistical Society*, 34, 1–41.

- Little, R. J. (1988). A Treatment of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*, 83, 1198–1202.
- Little, R. J. (2003). *The Bayesian Approach to Sample Survey Inference*, in *Analysis of Survey Data*, chapter 4, (pp. 49–57). Wiley Series in Survey Methodology, UK.
- Little, R. J. & Rubin, D. B. (1987a). *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics.
- Little, R. J. & Rubin, D. B. (1987b). Statistical Analysis with Missing Data. *Journal of Education*, 16, 150–155.
- Lohr, S. L. (2010). *Sampling: Design and Analysis, 2nd Edition*. Cengage Learning.
- Lumley, T. (2004). Analysis of Complex Survey Samples. *Journal of Statistical Software*, 9, 1–19.
- Lumley, T. (2010). *Complex Surveys: A guide to Analysis Using R*. John Wiley and Sons Inc., Washington, USA.
- Lyerla, R., Gouws, E., Garcia-Calleja, J. M., & Zaniewski, E. (2006). The 2005 Workbook: An Improved Tool for Estimating HIV Prevalence in Countries with Low Level and Concentrated Epidemics. *Sexually Transmitted Infections*, 82(Suppl III), iii41–iii44.
- Maheswaran, H. & Bland, R. (2009). Preventing Mother-To-Child Transmission of HIV in Resource-limited Settings. *Future Virology*, 4, 165–175.
- Mahomva, A., Greby, S., Dube, S., Mugurungi, O., Hargrove, J., Rosen, D., Dehne, K., Gregson, S., St Louis, M., & Hader, S. (2006). HIV Prevalence and Trends from Data in Zimbabwe. *Sexually Transmitted Infections*, 82, i42–i47.

- Makondo, L. & Makondo, O. (2014). Commercial Sex Work and HIV and AIDS: An Onomastic Perspective. *Journal of Social Science*, 38, 53–61.
- Mara, G., Rosalba, R., Till, B., Wood, S. N., & McGovern, M. E. (2015). A Unified Modelling Approach to Estimating HIV Prevalence in Sub-Saharan Africa. Research Report number 324, Department of Statistical Science, University College London.
- Martin, A. D., Quinn, K. M., & Park, J. H. (2013). *Markov chain Monte Carlo (MCMC) Package*. <http://mcmcpack.wustl.edu>.
- Mastro, T. D. & de Vincenzi, I. (2006). Probabilities of Sexual HIV-1 Transmission. *AIDS*, 10 (supplement), s75–82.
- Mathers, C. D. & Loncar, D. (2006). Projections of Global Mortality and Burden of Disease from 2002 to 2030. *Plos one Med* 3(11): DOI:10.1371/journal.pmed.0030442.
- Mbirimtengerenji, N. D. (2007). Is HIV/AIDS Epidemic Outcome of Poverty in sub-Saharan Africa? *Croat Medical Journal*, 48, 605–617.
- McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall.
- McCulloch, C. E. & Searle, S. R. (2001). *Generalized, Linear and Mixed Models*. John.
- Meer, J., Miller, D. L., & Rosen, H. S. (2003). Exploring the Health-wealth Nexus. *The Journal of Health Economics*, 22.
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953). Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21, 1087–1092.
- Moore, D. A. & Carpenter, T. E. (1999). Spatial Analytic Methods and Geographic Information Systems: Use in Health Research and Epidemiology. *Epidemiologic Reviews*, 21, 143–161.

- Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmosa, S., Tomba, G. S., Wallinga, J., Heilne, J., Sadkowska-Todys, M., Rosinska, M., & Edmunds, W. J. (2008). Social Contacts and Sexual Mixing Patterns Relevant to the Spread of Infectious Diseases. *Public Library of Science Med*, 5, 0381–0391.
- Murase, H., Nagashima, H., Yonezaki, S., Matsukura, R., & Kitakado, T. (2009). Application of a Generalized Additive Model (gam) to Reveal Relationships Between Environmental Factors and Distributions of Pelagic Fish and Krill: A Case Study in Sendai Bay, Japan. *International Council for the Exploration of the Sea*, 66, 1417–1424.
- Mutasa, D. (2012). *Zimbabwe DHS*. Technical report, Zimbabwe National Statistics Agency.
- Myer, L., Kuhn, L., Stein, Z. A., Wright, T. C., & L., D. (2003). Intravaginal Practices, Bacterial Vaginosis, and Women’s Susceptibility to HIV Infections: Epidemiological Evidence and Biological Mechanisms. *Lancet Infectious Disease*, 12, 786–794.
- Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society*, 135, 370–384.
- Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97, 558–625.
- Ngesa, O., Mwambi, H., & Achia, T. (2014). Bayesian Spatial Semi-parametric Modelling of HIV Variation in Kenya. *PLOS One*.
- Park, M. Y. & Hastie, T. (2010). *L2 Penalized Logistic Regression with Stepwise Variable Selection*.
- Perez, F., Orne-Gliemann, J., Mukotekwa, T., Miller, A., Glenshow, M., Mahomva, A., & Dabis, F. (2004). Prevention of Mother-To-Child Transmission of HIV: Evaluation

- of A Pilot Programme in A District Hospital in Rural Zimbabwe. *British Medical Journal*, 329, 1147–1150.
- Perry, M. J. (1998). Gender, Race and Economic Perspective on the Social Epidemiology of HIV Infection: Implications for Prevention. *The Journal of Primary Prevention*., 19, 97–104.
- Pettifor, A., Rees, H., Kleinschmidt, I., Steffenson, A., MacPhail, C., Hlongwe-Madikizela, L., Vermaak, K., & Padian, N. (2005). Young people’s sexual health in South Africa: HIV prevalence and sexual behaviours from a nationally representative household survey. *AIDS*, 19, 1525–1534.
- Pfeffermann, D. (1993). The Role Of Sampling Weights When Modeling Survey Data. *International Statistical Review*., 61, 317–337.
- Pfeffermann, D. (1996). The Use of Sampling Weights for Survey Data Analysis. *Statistical Methods in Medical Research*., 5, 139–261.
- Pigott, T. D. (2001). A Review of Methods for Missing Data. *Education Research and Evaluation*, 7, 353–383.
- Press, S. J. (1989). *Bayesian Statistics*. John Wiley and Sons, Ltd.
- Raghunathan, T. (2004). What do we do with missing data? Some options for analysis of incomplete data. *Annual Review of Public Health*, 25, 99–117.
- Raghunathan, T. (2010). Survey Inference with Incomplete Data. In *National Conference on Health Statistics*.
- Raiffa, H. & Schlaifer, R. (1961). *Applied Statistical Decision Theory*. The M.I.T. Press.
- Ramjee, G. & Eleanor, G. (2002). Prevalence of HIV among Truck Drivers Visiting Sex Workers in KwaZulu-Natal, South Africa. *Sexually Transmitted Diseases*, 29, 44–49.

- Rao, J. N. K. (2003). *Small area estimation*. Wiley, New York.
- Rao, J. N. K. (2011). Impact of Frequentist and Bayesian Methods on Survey Sampling Practice: A Selective Appraisal. *Statistical Science*, 26, 240–256.
- Rao, J. N. K. & Scott, A. J. (1981). The Analysis of Categorical Data From Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. *Journal of American Statistical Association*, 76, 221–230.
- Rao, J. N. K. & Thomas, D. R. (2003). *Analysis of Categorical Response Data from Complex Surveys: an Appraisal and Update*, chapter 7, (pp. 85–108). John Wiley and Sons, Ltd.
- Robert, C. P. & Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer Series in Statistics, New York.
- Rouder, J. N. & Lu, J. (2005). An Introduction to Bayesian Hierarchical Models with an Application in the Theory of Signal Detection. *Psychonomic Bulletin and Review*, 12, 573–604.
- Royall, R. (2003). *Interpreting a Sample as Evidence about a Finite Population*, chapter 5, (pp. 59 – 71). Wiley Series in Survey Methodology, UK.
- Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63, 581–592.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, Ltd.
- Rust, K. (1985). Variance Estimation for Complex Estimators in Sample Surveys. *Journal of Official Statistics*, 1, 381–397.
- Rust, K. F. & Rao, J. N. K. (1996). Variance Estimation for Complex Surveys using Replication Techniques. *Statistical Methods in Medical Research.*, 5, 283–310.

- Rutstein, S. O. & Kiersten, J. (2004). *The DHS Wealth Index. DHS Comparative Report*. Technical report, U. S. Agency for International Development, Maryland: ORC Macro.
- Sarndal, C. E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Series in Statistics, New York.
- Sarndal, C. E., Thomsen, I., Hoem, J. M., Lindley, D. V., Barndorff-Nielsen, O., & Dalenius, T. (1978). Design-Based and Model-Based Inference in Survey Sampling. *Scandinavian Journal Of Statistics*, 5, 27–52.
- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall.
- Schafer, J. L. (1999). Multiple Imputation: A Primer. *Statistical Methods in Medical Research*, 8, 3–15.
- Schafer, J. L. & Olsen, M. K. (1998). Multiple Imputation for Multivariate Missing Data Problems: A Data Analyst's Perspective. *Multivariate Behavioral Research*, 33, 545–571.
- Scott, D. & Holmberg, M. D. (1996). The Estimated Prevalence and Incidence of HIV in 96 Large US Metropolitan Areas. *American Journal Of Public Health*, 86, 642–654.
- Sedransk, J. (2008). Assessing the Value of Bayesian Methods for Inference about Finite Population Quantities. *Journal of Official Statistics*, 24, 495–506.
- Shao, J. & Chen, Y. (1998). Balanced Repeated Replication for Stratified Survey Data under Imputation. *Journal of American Statistical Association*, 93, 819–831.
- Shen, J. (2011). *Additive Mixed Modeling of HIV Patient Outcome Across Multiple Studies*. PhD thesis, Department of Statistics, University of California, LA.
- Shiboski, S. C. (1998). Generalized Additive Models for Current Status Data. *Lifetime Data Analysis*, 4, 29–50.

- Skinner, C., John, H. D., & Smith, T. M. F. (1989). *Analysis of Complex Surveys*. Wiley Series in Survey Methodology, UK.
- Smith, T. M. F. (1976). The Foundations of Survey Sampling: A Review. *Journal of the Royal Statistical Society*, 139, 183–204.
- Snijders, T. A. B. & Bosker, R. J. (1999). *Multilevel Analysis*. Sage Publications, London, UK.
- Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2003). *Bayesian Approaches to Clinical Trials and Health Care Evaluation*. John Wiley and Sons, Ltd.
- Spratt, M., Carpenter, J. R., Sterne, J. A., Carlin, J. B., Heron, J., Henderson, J., & Tilling, K. (2010). Strategies for Multiple Imputation in Longitudinal Studies. *American Journal Of Epidemiology*, 172, 478–487.
- Steenbergen, M. R. & Jones, B. S. (2002). Modeling Multilevel Data Structures. *American Journal of Political Science*, 46, 218–237.
- Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple Imputation for Missing Data in Epidemiology and Clinical Research: Potential and Pitfalls. *BioMedical Journal*, 338b, 2393.
- Su, Y. S., Gelman, A., Hill, J., & Yajima, M. (2011). Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box. *Journal of Statistical Software*, 45, 1–31.
- Susser, E., Valencia, E., & Conover, S. (1993). Prevention of HIV Infection among Psychiatric Patients in New York City Men’s Shelter. *American Journal Of Public Health.*, 83, 568–570.

- Swartman, G., Silverman, E., & Williamson, N. (1995). Relating Trends in Walleye Pollock (*Theragra chalcogramma*) Abundance in the Bering Sea to Environmental Factors. *Canadian Journal of Fisheries and Aquatic Sciences*, 52, 369–380.
- Sweeting, T. (1981). Scale parameters: A Bayesian Treatment. *Journal of the Royal Statistical Society*, 43, Series B, 333–338.
- Szwarcwald, C. L., Junior, a. B., Borges de Souza-Junior, P. R., Valente de Lemos, K. R., Germano de Frias, P., Luhm, K. R., Holcman, M. M., & Esteves, M. A. (2008). HIV Testing during Pregnancy: Use of Secondary Data to Estimate 2006 Test Coverage and Prevalence in Brazil. *Brazilian Journal of Infectious Diseases and Context*, 12, 167–172.
- Tanser, F., Barnighausen, T., Cooke, G. S., & Newell, M. L. (2009). Localized Spatial Clustering of HIV Infections in a Widely Disseminated Rural South African Epidemic. *International Journal of Epidemiology*, 38, 1008–1016.
- Team, R. C. (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Ugen, K. E., Goedert, J. J., Boyer, J., Refaeli, Y., Frank, I., Williams, W. V., Willoughby, A., Landesman, S., Mendez, H., & Rubinstein, A. (1992). Vertical Transmission of Human Immunodeficiency Virus (HIV) infection. Reactivity of Maternal Sera with Glycoprotein 120 and 41 Peptides from HIV Type 1. *The Journal of Clinical Investigation*, 89, 1923–1930.
- Vail, A., Hornbuckle, J., Spiegelhalter, D. J., & Thornton, J. G. (2001). Prospective application of Bayesian Monitoring and Analysis in an Open Randomized Clinical Trials. *Statistics in Medicine*, 20, 3777–3787.
- van Buuren, S. & Groothuis-Oudshoorn, K. (2011). Multiple Imputation by Chained Equations in R. *Journal of Statistical Software*, 45, 1–67.

- von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gotsche, P. C., & Vandembroucke, J. P. (2007). The strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. *Bulletin of the World Health Organization*, 85, 867–872.
- Wang, R., Sedransk, J., & Jinn, J. H. (1992). Secsecond Data Analysis when there are Missing Observations. *Journal of American Statistical Association*, 87, 952–961.
- Welz, T., Hosegood, V., Jaffar, S., Batzing-Feigenbaum, J., Herbst, K., & Newell, M. L. (2007). Continued very high prevalence of HIV infection in rural KwaZulu-Natal, South Africa: A population based longitudinal study. *AIDS*, 21, 1467–1472.
- Wiegert, K., Dinh, T. H., Mushavi, A., Mugurungi, O., & Kilmarx, P. H. (2014). Integration of Prevention of Mother-To-Child Transmission of HIV (PMTCT) Postpartum Services with Other HIV Care and Treatment Services within the Maternal and Child Health Setting in Zimbabwe, 2012. *Plos one*, 9, 1–6.
- Winship, C. & Radbill, L. (1994). Sampling Weights and Regression Anlysis. *Sociological Methods and Research*, 23, 230–257.
- Wolter, K. (1985). *Introduction to Variance Estimation*. Springer Series in Statistics, New York.
- Wong, G. Y. & Mason, W. M. (1985). The Hierarchical Logistic Regression Model for Multilevel Analysis. *American Statistical Association*, 80, 513–523.
- Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall.
- Woodruff, R. S. (1971). A Simple Method for Approximating the Variance of a Complicated Estimate. *Journal of American Statistical Association*, 66, 411–414.

Yee, T. W. & Mitchell, N. D. (1991). Generalized Additive Models in Plant Ecology.
Journal of Vegetation Science, 2, 587–602.

Appendix A

R code for Chapter 3

This is an R code that we used to specify the complex sampling design, to compute the design-consistent crude national and sub-group level estimates of the HIV prevalence. Also included is the code for estimating the survey logistic regression model as well as for performing the goodness of fit test.

```
>library(survey)
>survey.design<-svydesign(id=~identifier,strata=~province,fpc=~newfpc,
  weights=~rhivweight,data=phddata,nest=TRUE,variance="HT",pps=FALSE)
>hiv.overall<-svymean(~hivresult,survey.design,deff=TRUE)
>confint(hiv.overall)
>bygender<-svyby(~hivresult,~gender,survey.design,svymean)
>confint(bygender)
>byagegroup<-svyby(~hivresult,~agegroup,survey.design,svymean)
>confint(byagegroup)
>bymaritalstatus<-svyby(~hivresult,~maritalstatus,survey.design,svymean)
>confint(bymaritalstatus)
>byliteracy<-svyby(~hivresult,~literacy,survey.design,svymean)
>confint(byliteracy)
```

```

>byplaceofresidence<-svyby(~hivresult,~placeofresidence,
    survey.design,svymean)
>confint(byplaceofresidence)
>bywealthindex<-svyby(~hivresult,~wealthindex,survey.design,svymean)
>confint(bywealthindex)
>byemployment<-svyby(~hivresult,~employment,survey.design,svymean)
>confint(byemployment)
>byeducation<-svyby(~hivresult,~education,survey.design,svymean)
>confint(byeducation)
>byreligion<-svyby(~hivresult,~religion,survey.design,svymean)
>confint(byreligion)

>svychisq(~hivresult+gender,design=survey.design,statistic="F")
>svychisq(~hivresult+agegroup,design=survey.design,statistic="F")
>svychisq(~hivresult+maritalstatus,design=survey.design,statistic="F")
>svychisq(~hivresult+literacy,design=survey.design,statistic="F")
>svychisq(~hivresult+placeofresidence,design=survey.design,statistic="F")
>svychisq(~hivresult+religion,design=survey.design,statistic="F")
>svychisq(~hivresult+education,design=survey.design,statistic="F")
>svychisq(~hivresult+wealthindex,design=survey.design,statistic="F")

>svy.model<-svyglm(hivresult~factor(agegroup)+factor(gender)+
    factor(maritalstatuscode)+factor(literacycode)+factor(placeofresidence)+
    factor(agegroup):factor(gender)+factor(gender):factor(maritalstatus),
    family=quasibinomial,design=survey.design)

>hoslem<-hoslem.test(svy.model$model$hivresult,fitted(svy.model),g=10)

```

Appendix B

R code for Chapter 4

We present the R code that we used to carry out the multiple imputation procedure. Specifically the functions that were used for the set up and for obtaining the multiple data sets.

```
>library(mi)
>info<-mi.info(phddata.missing)
>new.data<-mi.preprocess(phddata.missing)
>imp<-mi(new.data,n.imp=5,n.iter=30,R.hat=1.1,max.minutes=30,
  rand.imp.method="bootstap",run.past.convergence=FALSE,
  seed=NA,check.coef.convergence=FALSE,
  add.noise=noise.control(post.run.iter=10))
```

Appendix C

R code for Chapter 5

We present the R code that was used to compute the hierarchical models as generalized linear mixed effects models.

```
>library(lme4)
>hierarchical.model<-glmer(hivresult~1+factor(gender)+factor(agegroup)+
  factor(maritalstatuscode)+factor(placeofresidence)+
  factor(gender)*factor(agegroup)+factor(gender)*factor(maritalstatuscode)+
  (1|codedhh)+(1|cluster),data=clean,family=binomial)
>summary(hierarchical.model)
```


Appendix D

R code for Chapter 6

Presented here is the R code used to compute the Bayesian logistic regression model. Included is the code for the MCMC used to make the iterative draws from the posterior distribution.

```
>.library(arm)
>library(MCMCpack)
>bayesglm.model3.b<-bayesglm(hivresult~factor(gender)+factor(agegroup)+
  factor(maritalstatuscode)+factor(placeofresidence)+factor(literacycode)+
  factor(gender)*factor(agegroup)+factor(gender)*factor(maritalstatuscode),
  family=binomial(link="logit"),
  prior.mean=c(-.22859,0.9663,1.6257,1.98141,1.95297,1.48609,
    1.35851,1.27691,
    -0.07197,0.71503,1.5824,0.18828,0.45203,0.1782,-0.80913,
    -0.70844,-0.66751,-0.28284,0.45272,0.64413,0,0.43532,
    0.57216,0.65905),n.iter=50000,data=clean)
>posterior.intercept<-simulates[,1]
>mcmc.intercept<-mcmc(posterior.intercept)
>geweke.intercept<-geweke.plot(mcmc.intercept,
```

```
      frac1=0.1,frac2=0.5,nbins=50,pvalue=0.05,auto.layout=TRUE,main="")
>posterior.gender<-simulates[,2]
>mcmc.gender<-mcmc(posterior.gender)
>geweke.gender<-geweke.plot(mcmc.gender,frac1=0.1,
      frac2=0.5,nbins=50,pvalue=0.05,auto.layout=TRUE,main="")
```

Appendix E

R code for Chapter 7

```
>bayes.gllm<-(MCMCglmm(hivresut~factor(gender)+factor(agegroup)+
  factor(literacy)+factor(maritalstatus)+factor(placeofresidence)+facto(sexofhead)+
  factor(wealthindex)+factor(gender):factor(agegroup)+
  factor(gender):factor(maritalstatus),
  random=~houhold+cluster,data=clean,
  verbose=FALSE,prior=prior,nitt=10000,
  burnin=100,thin=10,family="categorical")
>prior = list(R = list(V = 2, nu = 0.004), G = list(G1 = list(V = 2, nu = 0.004),
  G2 = list(V = 2, nu = 0.004)))
>plot(bayes.gllm25)
```

Appendix F

R code for Chapter 8

We present the R code for computing the generalized additive model in Chapter 7.

```
>library(gam)
>gam.model3<-gam(hivresult~s(age,bs="cr")+factor(gender)+
  factor(maritalstatuscode)+factor(placeofresidence)+
  factor(literacycode)+factor(gender):factor(maritalstatuscode),
  family=binomial,data=clean)
```