Statistical modelling of the relationship between intimate partner violence and HIV infection among women in Zimbabwe

By

Isobella Chimatira (212561708@stu.ukzn.ac.za)

University of KwaZulu-Natal

Supervisor: Dr. Thomas N. O. Achia

Co-Supervisor: Prof. Henry G. Mwambi

University of KwaZulu-Natal, South Africa

Submitted in partial fulfilment of a Master of Science in Biostatistics

UNIVERSITY OF
KWAZULU-NATAL

INYUVESI
YAKWAZULU-NATALI

PIETERMARITZBURG CAMPUS, SOUTH AFRICA

10 October 2014

Disclaimer

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

Isobella Chimatira, 30 September 2014

Declaration

The research work is the original work done by the author (Isobella Chimatira) and it is not a duplicate of some of the research work done by other authors. All references that were used to refer to are duly acknowledged.

Isobella Chimatira

Signature:_____ Date:_____

Dr. Thomas N. O. Achia

Signature:_____ Date:_____

Professor Henry G. Mwambi

Signature:_____ Date:_____

Abstract

Zimbabwean women between the ages of $15-49$ years are among the women most affected by HIV and Intimate Partner Violence in the world. The high rates of HIV infection among women have raised an alarm and stimulated research on the problem of violence against women. Intimate Partner Violence (IPV) is a well-known violation of human rights and is a problem in public health. It usually overlaps with the HIV/AIDS epidemic and has been reported to be a determinant of women's risk for HIV. The present study explored relevant statistical methods in modelling the relationship between Intimate Partner Violence (IPV) and HIV in Zimbabwe. The data used in the current research is from a Demographic and Health Survey (DHS) conducted in Zimbabwe for year $2005-06$.

The study aimed at analysing the relationship between IPV and HIV using the following explanatory variables: age; marital status; religion; education; wealth index; region; decision making; media exposure; STI; physical and sexual violence. Principal Component Analysis was used to create indices of IPV, media exposure and decision making among women in the age group $15-49$. Survey Logistic Regression models accounting for multi-stage survey design was also used to adjust for socio-demographic and socio-economic factors. In order to explore the relationship between IPV and HIV prevalence among women, a generalised linear mixed model was adapted, controlling for socio-demographic variables and treating DHS survey clusters as random effects.

Since IPV takes up more than two categories, Multinomial Logit Modelling was used to analyse the relationship of IPV with socio demographic and socio-economic variables. The results from the survey logistic regression modelling were as follows: unadjusted odds ratios (OR) for sexual or physical IPV ranged from $0.91-1.09$ and 95% confidence intervals (CI) were $(0.72, 1.14)$ for sexual and $(0.92, 1.28)$ for physical violence. The adjusted odds ratios for sexual violence 0.82 $[95\%CI: 0.63, 1.06]$ and physical violence 1.12 $[95\%CI: 0.97, 1.36]$. Both survey logistic regression models and generalised linear mixed models found no association between HIV and IPV among women in Zimbabwe. This study provides further evidence that IPV and HIV are not associated. In addition, the analysis revealed that the covariates which were associated with HIV and IPV were age, education, marital status, STI, religion and wealth index.

As a result the study recommends that more research is required to find the situations or circumstances under which IPV is associated with HIV prevalence.

Dedication

I would like to thank the Lord Almighty for giving me the strength and determination to conduct and finish this thesis under the most challenging conditions of my life.

This research is dedicated to my husband Raymond, daughter Joan Tariro, my mother Charity Faith Mberengwa and my late aunt Rosemary Khatso.

Acknowledgements

I am grateful and indebted to all the people whom I have had pleasure to work with during this research project. I owe my deepest gratitude to the support, patience, encouragement, advice and guidance of Dr Thomas Achia and Professor Henry Mwambi. Their knowledge and commitments to the highest standards has inspired and motivated me and enabled me to complete this research.

I also wish to thank Dr Farai Chirove and Abdul Karim who were always ready to assist me when I had problems with my thesis. Most importantly, I wish to thank my husband Raymond who was my inspiration and all the contributions he made so that I could complete this research.

Contents

List of Figures

Chapter 1

Introduction

Chapter 1 gives a summary of the definitions of what Intimate Partner Violence and HIV are. It also describes and illustrates what previous studies have revealed about the relationship of Intimate Partner Violence and HIV. Finally redundant the problem statement and the objectives of the study are delineated.

1.1 Background

Human Immunodeficiency Virus (HIV) and Intimate Partner Violence (IPV) affect millions of women worldwide (Shi, Kouyoumdjian & Dushoff 2013, Hove & Gwazane 2011, Jewkes, Dunkle, Nduna & Shai 2010). Global figures show that 93% of HIV infections have occurred in the developing world. Maman, Campbell, Sweat & Gielen (2000) identified that 60% of the HIV infections worldwide are found in Africa; a continent that comprises 11% of the world's population (Abramsky, Watts, Garcia-Moreno, Devries, Kiss, Ellsberg, Jansen & Heise 2011*b*). Statistics show that the number of women with HIV infection and AIDS has increased worldwide from 50.94% in 2001 to 51.62% in 2009. In addition, data and statistics from United Nations Entity for Gender Equality and the Empowerment of Women estimates that globally, more than 50% of people living with HIV are women and girls, while in sub-Saharan Africa women over 15 years of age account for 59% of people living with HIV (Campbell, Baty, Ghandour, Stockman, Francisco & Wagman 2008). The sub-Saharan African (SSA) region remains the

hardest hit by the HIV pandemic, with some SSA countries having a very high proportion of women infected with HIV, for example in Botswana 61% of adults living with HIV are women and in Kenya its 64% (Campbell et al. 2008). In South Africa (SA), women between the ages of 15 - 24 years comprise 90% of new HIV infections, while the infection rate of HIV is six times higher among women as compared to men (Campbell et al. 2008). Zimbabwe is also experiencing the HIV and IPV epidemic, resulting in social, economic and public health hardships (Hove & Gwazane 2011).

There is an increased risk of being infected with HIV as a result of gender based violence and gender inequality (Jewkes et al. 2010, Silverman, Decker, Saggurti, Balaiah & Raj 2008). Several authors state that IPV against women is psychological, physical, and sexual abuse directed at a spouse (Abeya, Afework & Yalew 2011, Campbell et al. 2008). These physical and sexual threats are usually between married, romantically involved partners or former partners. Violence against women can either occur in public or private life and has been identified as a significant human rights concern and public health issue, in both the industrialised and less developed countries (Abeya et al. 2011, Osinde, Kaye & Kakaire 2011). Globally, it has been noted that any form of psychological, physical or sexual abuse can cause harm to the health of women (Osinde et al. 2011). An estimated 15% women in Ethiopia and 71% of women in Japan have experienced IPV at a certain point of their lives as evidenced by the WHO Multi-Country study on Women's Health and Domestic Violence (Abramsky et al. 2011*b*).

The consequences of IPV among women are many and varied. They include unwanted pregnancy, low birth weights induced abortions, death from homicide, physical injury, disability, depression, post-traumatic stress syndrome, pre-term birth and poor acceptance of services to prevent vertical transmission of HIV (Osinde et al. 2011). Researchers have reported evidence of an association between IPV and irregular use of condoms, as well as sexual coercion, leading to sexually transmitted infections (STIs) and HIV (Osinde et al. 2011). The risk of HIV infection among women who are abused by their partners increases especially if their partners have multiple sexual partners (Osinde et al. 2011).

There is increasing evidence of the impact of Intimate Partner Violence against women from studies conducted in several countries, such as the United States of America (USA), Kenya, India, Rwanda, Tanzania and South Africa. Studies in Kenya showed that 40% of women who were exposed to IPV acquired HIV infection from their

partners. A survey carried out in the USA in 2005 on violence against women revealed that 64% of women above the age of 18 years had experienced rape or physical abuse by a partner, ex husband, current husband or just a boyfriend. Centers for Disease Control and Prevention (CDC) in the United States give statistics that one sixth of women have experienced either an attempted or complete rape defined as forced or threatened vaginal, oral or anal penetration (Campbell et al. 2008). A survey in 10 developing countries of physical violence by male partners among women showed that between 13% to 61% of women have experienced physical violence, with some reporting results as high as 30% - 50% (Harling, Msisha & Subramanian 2010). Japanese surveys have reported 6% prevalence of sexual violence as compared to Ethiopia which reported 59%. The prevalence of sexual violence reported in Namibia and Tanzania was 17% and 31% respectively (Campbell et al. 2008). These high rates suggest that women who are abused face difficulties in protecting themselves from HIV infections.
Studies done in the USA comparing HIV-positive and HIV-negative women found no significant difference among the women regarding rates of physical, sexual and emotional abuse by an intimate partner. Some international studies found that the risk of IPV is higher among HIV positive women when they analysed data using multivariate logistic regression. Research done in South Africa, Tanzania and Kenya revealed that HIV-positive women experienced more partner violence in their life as compared to HIV-negative women. The highest rate was reported in Tanzania with a comparison of 52% vs 29%. These studies were community-based except for SA which used cluster sampling of seventy South African communities (Campbell et al. 2008).

A study done investigating abuse in relation to the time HIV is diagnosed showed, that 13% of women had experienced IPV after they were diagnosed as HIV positive. The behaviour of abused and non-abused women in terms of HIV is different. Studies have shown that women who have faced IPV are unlikely to have an HIV test compared to non-abused women. If abused women tested positive for HIV they were more likely to report that they were HIV positive. In addition to reporting being HIV positive, abused women also reported having had a Sexually Transmitted Infection (STI). The risk of HIV transmission is high in the presence of STI (World Health Organisation 2004).

1.2 IPV and HIV: Mutual risk factors

There are a number of mechanisms that explain how IPV increases a woman's risk of acquiring STI's or HIV. Maman et al. (2000) hypothesized three ways that show how exposure can increase a woman's risk for HIV infection. These are:

(1) Through forced sex with an infected partner.

(2) Through limited or compromised negotiation of safer sex practices.

(3) Through increased sexual risk-taking behaviour.

The rates of HIV infection among women equals and may surpass that of men as a result of biological and social risk factors. Evidence suggests that women are more vulnerable to getting HIV as compared to men due to their biological make up. (Jewkes et al. 2010, World Health Organisation 2004) suggest that the biological make up of women made it easier for them to be exposed to HIV transmission if they are raped by an infected man. This is because in women there is a large surface area of mucous membrane exposed while they have sex, thus high volumes of fluid are transferred from men to women. In addition, HIV infected men have a high viral content in their sexual fluids. Research has shown that 40% to 45% of people in physically intimate relationships have experienced forced sex. As a result of forced sex, women are at risk of STIs by 2 to 10 times more than for physical abuse alone (Campbell et al. 2008). Forced sex leads to genital injuries like vaginal or rectal lacerations that facilitate the transmission of infections. Other mechanisms used by abusive partners to transmit HIV infection are forced injection of drugs and not disclosing their known HIV positive serostatus to intimate partners.

Furthermore, there is some evidence from studies conducted in North America and South Africa, that women who have been abused as minors are more likely to engage in risky sexual behaviour, which may expose them to HIV. Research from different parts of the world has revealed that men who rape women or are physically violent to their partners have many sexual partners and have intercourse frequently. Such men are likely to be infected with HIV and STIs that are spread to their sexual partners. Qualitative research has shown that men who control their women and want to show their strength and toughness are involved in gender based violence(GBV). Research done in India and South Africa has shown evidence that men involved in GBV are highly infected with HIV (Campbell

et al. 2008, Jewkes et al. 2010, World Health Organisation 2004).

1.3 Sexual decision making and intimate partner violence

IPV makes communication between partners difficult, which results in abused women fearing a violent reaction when negotiating safe sex practices such as condom use. Fear of violence negatively impacts on some women's ability to make decisions in areas such as how they can prevent HIV, or whether they can utilise voluntary counselling and testing services. Studies in South Africa and the USA have revealed that relationship status or power has an impact on sexual health practices (Campbell et al. 2008). The researchers noted that women who have been exposed to sexual abuse and IPV have a greater chance of engaging in risky sexual behaviour. Women with low relationship powers were unlikely to practice safe sex methods as compared to their counterparts with high levels of power. Furthermore, the women with less relationship power have less power in negotiating condom use, limited ability to negotiate HIV preventive behaviours as well as refusing to have sex. As a result of fear of violence, some HIV-positive women end up not seeking care, treatment or support when needed, in addition to keeping their HIV-positive status a secret. Other consequences of IPV are increased prevalence of stress, depression and chronic anxiety.

There is evidence that males who abuse their partners in intimate relationships, usually put them at a greater risk for HIV infection. This has been suggested by self-reported information by women who have been abused who reported greater high-risk behaviour than from non-abused females. Abraham, Jewkes, Hoffmans, and Laubsher in Campbell et al. (2008) explored problematic alcohol use and having more than one current partner as HIV risk behaviour associated with male sexual IPV. Similar studies done in South Africa and sub-Saharan Africa found that men with many female partners were perpetrators of violence. Some literature suggests that women who have been tested for HIV and decided to share their status with their partners faced an increased risk of violence. Campbell et al. (2008) concludes that both adult women and adolescents are at risk of getting HIV/AIDS as a result of IPV. To prevent HIV, efforts should be focused on the reduction of males using violence against women, the reduction of concurrent and multiple partners, as well as male circumcision. Figure 1.1 shows the pathways through which GBV and power inequity may place women at risk of HIV infection.

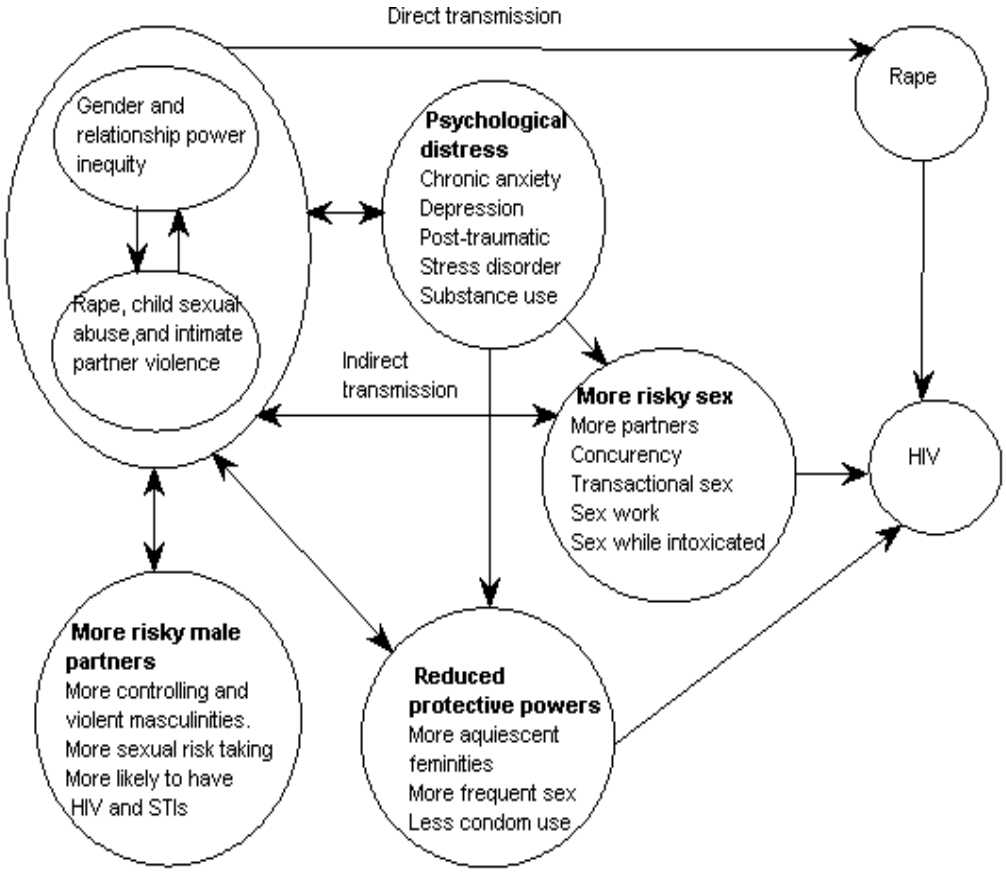

Figure 1.1: Pathways through which gender-based violence, and gender and relationship power inequities might place women at risk of HIV infection: **Adapted from Rachel K Jewkes, Kristin Dunkle, Mzikazi Nduna, Nwabisa Shai (2010)**

1.4 Geography and economy of Zimbabwe

Figure 1.2: Zimbabwe Demographic and Health Survey 2005-06

Zimbabwe is a landlocked Southern African country, which is found between the Limpopo and Zambezi rivers, just north of the Tropic of Capricorn (Zimbabwe & Inc. 2007). It is bordered by Mozambique, South Africa, Botswana and Zambia.The country's major foreign currency earning sectors are agriculture and mining.
The population of Zimbabwe, as estimated from the 2002 census is, 11.6 million people (Zimbabwe & Inc. 2007).

1.5 Statement of the problem

Global statistics, as stated in (World Health Organisation 2004, Garcia-Moreno & Watts 2000), show that:

(1) Between 10% and 69% of women report physical abuse by an intimate partner at least once in their lives.

(2) Between 6% and 47% women worldwide report being sexually assaulted by an intimate partner in their life.

(3) Between 7% and 48% of girls and young women between the age of $10-24$ years report their first sexual encounter as coerced.

The Zimbabwe Demographic Health Survey (DHS) 2005-06 reports that 18% of adults from Zimbabwe aged 15-49 years are infected with HIV. Of these, 21% are women and 15% men (Zimbabwe & Inc. 2007). In Zimbabwe, domestic violence is common across all socio-economic and cultural backgrounds (Zimbabwe & Inc. 2007). This is supported by evidence from a study that identified that 95% of domestic violence cases were against women (Hove & Gwazane 2011). In Zimbabwe, 36% of all women have experienced physical violence as early as 15 years of age and 17% of women experienced violence 12 months before the survey (Hove & Gwazane 2011). The DHS report states that 47% of the women who experienced violence from the age of 15 were violated by their present husband or partner, and 18% reported that the perpetrators of violence were ex-partners or husbands. Information in the DHS report reveals that 25% of women had experienced sexual violence in their lives; 21% of women reported that their first sexual encounter was forced and not agreed to (Hove & Gwazane 2011). The $2005-06$ DHS reports that 65% of women claimed that their current/ex-husband, partner or boyfriend committed sexual violence. The Zimbabwe Demographic Health Survey (ZDHS) (1999) shows that 51% of women in Zimbabwe believe that men have a right to beat them (Hove & Gwazane 2011). This poses a challenge for women to negotiate safe sex with men who dominate them.

As a result the high rate of HIV infection among women as outlined above, there is a need to focus on analysing the relationship between IPV and HIV.

1.6 Objectives

The objectives of this study are:

1. To develop an index of IPV using Principal Component Analysis (PCA).
2. To review statistical properties of the logistic regression model.
3. To use logistic regression to link HIV to IPV.
4. To assess the impact of physical and sexual violence on HIV.
5. To review the statistical properties of the multinomial logit regression model.
6. To review the statistical properties of the generalised linear mixed models.

1.7 Structure of the thesis

This study aims to advance the understanding of the relationship between Intimate Partner Violence (IPV) and HIV among women in the age range $15-49$, years in Zimbabwe. In Chapter 2 of the thesis we focus on exploratory data, using cross-tabulations chi-square statistics to check which variables are significantly associated with HIV status. The Principal Component Analysis (PCA) is used to create indices. Chapter 3 of this thesis focuses on using logistic regression to develop a best fit model for the variables associated with a woman getting HIV. In Chapter 4 a Multinomial Logit model is developed to see the variables associated with IPV. Finally in Chapter 5 a generalized linear mixed model with random effects is applied to the HIV status data. The data is analysed using SPSS PASW 18 (Norusis et al. 2010) and STATA (StataCorp 2009).

Chapter 2

Data description, exploratory analysis and indicator development

This chapter describes the data used in this study. It also discusses the methodology used to create the indicators of IPV, media exposure and decision making.

2.1 Data Source

The dataset for this study was obtained from the Measure Demographic and Health Survey (DHS) website (www.measuredhs.com). More than 90 countries have participated in the DHS surveys which include both men and women. This study was based on data provided by women in Zimbabwe who were asked about their experiences of IPV and also tested for HIV during the $2005/2006$ survey. The target population included women in the age range of 15 to 49 years.

2.1.1 Study design

This $2005-06$ Zimbabwe Demographic Health Survey (ZDHS) was the fourth comprehensive survey conducted in Zimbabwe and was intended to furnish programme managers and policy makers with detailed information on levels and trends in fertility; nuptiality; sexual activity; fertility preferences; awareness and use of family planning methods; breastfeeding practices; nutritional status of mothers and young children; early childhood mortality and maternal mortality; maternal and child health; and awareness and behaviour regarding HIV. It was the first survey to collect information on domestic violence and to provide data on HIV prevalence in Zimbabwe (Zimbabwe & Inc. 2007).

Zimbabwe has ten provinces including the two metropoles of Harare and Bulawayo. For the purposes of the DHS one urban stratum was formed for each of Harare and Bulawayo, while each of the eight provinces was stratified into four strata according to land use. Thus a total of 34 strata were formed. A sample of 10 800 households were selected for the survey. This was done by first selecting 1200 enumeration areas (EAs) as stage one, and sampling of household units in stage two. Systematic random sampling was used to select the enumeration areas for the 34 strata. These were further divided into 3 replicates of 400 EAs from the 1200 EAs. Households were selected by random sampling. The listing of households did not include people living in barracks, hospitals, police camps, boarding schools or any institutional households (Zimbabwe & Inc. 2007).

2.1.2 Study population and sample size

The study population consisted of women aged 15-49 years (n=9 870). HIV testing and IPV related questions were requested from two independent and randomly selected subsets of the population (Zimbabwe & Inc. 2007). The number of women who were tested for HIV are 7 494 and 2 376 declined the test or their test result was not available. IPV and HIV questions were given to 4 082 women, but 17 had missing information on the covariates. The final analytic sample size was 4 065. The domestic violence module was only asked of women who were currently married then or had been married. The analysis therefore will focus on these women (Zimbabwe & Inc. 2007). Table 2.1 shows the sampling characteristics of the Zimbabwe population.

Table 2.1: Sampling characteristics of the Zimbabwe population

	Residence		
	Urban	Rural	Total
Household interviews			
Households selected	3 455	7 297	10 752
Households occupied	3 248	6 530	9 778
Households interviewed	3 056	6 229	9 285
Household response rate	94.1	95.4	95.0
Interviews with women			
Number of eligible women	3 763	6 107	9 870
Number of eligible women interviewed	3 203	5 704	8 907
Eligible women response rate	85.1	93.4	90.2
HIV Tests			
Households selected for HIV test	3 455	7 297	10 752
Women accepting tests and results available	2 450	5 044	7 494
HIV test response rate	65.1	82.6	75.9
Domestic Violence module			
Households selected for DV module	3 455	7 297	10 752
Women eligible for DV module			6 351
DV respondents			4 854
DV response rate			76.4
Women with valid DV and HIV responses			4 082
Women missing Covariate information			17
Final analytic sample			4 065

2.1.3 Selection criteria

The people who were eligible to be selected in the sample for the survey were permanent residents or visitors present on the night before the survey in the households in the $2005-06$ ZDHS. These included all women aged 15-49 years and all men aged 15-54 years in each household. All men and women who were eligible, consented to be tested for HIV. Eligible women were asked questions on domestic violence (Zimbabwe & Inc. 2007).
A questionnaire was used in the 2005-06 ZDHS to collect information from all women aged 15-49 years. These questions included awareness and behaviour regarding HIV/AIDS and other sexually transmitted infections (STIs) and domestic violence.

2.1.4 The response variables

The primary response variable when analysing the relationship between IPV and HIV was the dichotomous variable representing the HIV status of women. Women who volunteered to be tested provided five drops of blood. The collection of the blood specimen and analysis was anonymous and linked to the protocol developed for MEASURE DHS. The HIV results were then merged to the socio-demographic data collected in the questionnaire for individuals. HIV results were identified by a bar code in the spreadsheet (Zimbabwe & Inc. 2007).

To identify the determinants of IPV, a response variable taking 4 values:

- $102 =$ both low physical and sexual violence,
- $103 =$ high physical and low sexual violence,
- $202 =$ high sexual and low physical violence and
- $203 =$ both high sexual and physical violence

was used.

2.1.5 The independent variables

The variables used in this study are listed in the Table 2.2 below: **Media exposure**, IPV and decision making

Table 2.2: Independent variables used in study

Variable: Levels
Socio-demographic
Age: $1 = 15-19, 2 = 20-24, 3 = 25-29, 4 = 30-34, 5 = 35-39, 6 = 40-44, 7 = 45-49$
Education: 1=primary, 2=secondary, 3=higher
Religion: 1=Traditional, 2=Roman Catholic
Marital Status: 1=Married/living together, 2=separated
Type of Residence: 1=Urban, 2=Rural
Socio-Economic
Wealth index: 1=poorest, 2=poorer, 3=middle, 4=richer, 5=richest
STI: 1=None, 2=Yes
Other variables
Sexual violence: 1=low, 2=high
Physical violence: 1=low, 2=high
Media exposure: 1=low, 2=medium, 3=high
Decision making: 1=independent, 2=consults, 3=subservient

variables will be computed in chapter 5 and 6. Media exposure index will be based on the responses to questions posed on the frequency of watching TV, the frequency of listening to radio and the frequency of reading newspapers. The **decision making index** was based on the responses to questions on final say on health care, final say on making large household purchases, final say on making household purchases for daily needs, final say on visits to family or relatives, final say on food to be cooked each day and final say on deciding what to do with money husband earns.

2.2 Managing the data

The $2005-06$ ZDHS comprised the IPV data and HIV data on women in the range 15 to 49 years. Variables from the HIV data were made the same as the HIV variables in the IPV data which was the bigger data set. The variables were cluster number, household number and respondent's line number. The HIV data were then imported into the bigger data set with IPV variables. Data were then sorted and merged. This was done by

deleting unselected cases, removing the gaps of the women not tested as well. The data was cleaned in SPSS by choosing the best variables on IPV and HIV, these were:

(1) Demographic/Residential Characteristics that is, age group, religion, urban.

(2) Biological Characteristics that is, circumcised, had STI.

(3) Behavioural Characteristics that is, ever used a condom, number of partners in the past year, frequency of alcohol use in the past year and frequency of travel in the past year.

(4) Social characteristics such as: education, wealth index, marital status, age at first sex, religion,perceived risk of getting HIV.

2.3 Index Development

2.3.1 Principal component analysis

In this section we use Principal Component Analysis (PCA) to compute indices for IPV (sexual and physical violence), media exposure and decision making, using variables in the Zimbabwe demographic and Health survey $2005-06$ for women under survey. PCA is a multivariate statistical analysis procedure that was put forward by Karl Pearson in the early 1900's. Early applications of PCA were in educational testing, but the procedure has now been used in psychological, biological and medical applications. The procedure is used when a simpler representation is required for a set of intercorrelated variables (Afifi, Clark & May 2003). PCA reduces the number of original variables by forming small sets of uncorrelated principal components from a set of p correlated variables, without losing much information from the original data and with the new variables explaining almost all of the variation from the original variables (Afifi et al. 2003).

The principal components are linear combinations of the original variables. The amount of information conveyed by each principal component is contained in its variance. Principal components are then ordered in terms of increasing variance. The most informative principal component is the first as it explains most of the variability in

the original data set, while the last contains the least information. Selecting the first few components to represent the original data can usually be done without losing too much information.

2.4 PCA model description

As mentioned above, PCA develops new variables which are linear combinations of the original data/variables. Suppose that we have p variables $X_1 \dots X_P$ measured in n households, the observation vector $X'_{1\times p} = [X_1, X_2 \dots, X_p]$ with mean μ and covariance matrix $\sum$ of full rank p. Further let $\vec{X} = (X_1, \cdots, X_p)^1$ be a p-dimensional random vector with mean μ and covariance matrix $\sum$. The primary aim of PCA is to form a set of uncorrelated variables $Z_1, Z_2, \cdots, Z_p$ that are linear combinations of the original variables $X_1, \cdots, X_p$ that is

$$\begin{aligned} Z_1 =& a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ Z_2 =& a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\vdots \\ Z_p =& a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p. \end{aligned}$$

The equations given above can be expressed as **z** = **Ax**, where **z** $= (Z_1, Z_2 \cdots, Z_p)$ and **x** $= (X_1, \cdots, X_p)$ and **A** is the matrix of coefficients (Achia, Wangombe & Khadioli 2010, Timm 2002). The j^{th} principal component can be expressed as $Z_j = \vec{a}_j^1 \vec{X}$ where $\vec{a}_J = (a_{ji}, \cdots, a_{jp})^1$ which is the optimal weight vector and $\vec{X}$ is the random vector as described previously. Taking the first principal component,the coefficients $a_{11}, \dots, a_{1p}$ are chosen in such a way that the variance of Z_1 is maximised subject to the constraint that $a_{11}^2 + \dots + a_{1p}^2 = 1$ The variance of Z_1 is given by:

$$\begin{aligned} Var(Z_1) &= Var(a_1'X), \\ &= a_1' Var(X) a_1, \\ &= a_1' \sum a_1. \end{aligned} \tag{2.4.1}$$

Maximum $a_1' \sum a_1$ subject to $a_1' a_1 = 1$

a_1' is the optimal weight vector, $a_1' a_1 = 1$ is the charecteristic vector associated with the largest eigenequation $|\sum - \lambda I| = 0$ λ is the variance matrix of the principal component

Therefore the largest variance of the first component is denoted by the largest root λ_1 which is the largest eigenvalue (a corresponding scalar value for each eigenvector of a linear transformation) of **A**. The linear combination of the second principal component is given by $Z_2 = a_2'X$. There is no correlation between the first component and second principal component. The variance of the second linear component λ_2 which is the second biggest eigenvalue of **A**, can be shown if covariance between Z_1 and Z_2 are zero.

Thus $\sum a_1 = a_1\lambda_1$ so that:

$$Cov(Z_2, Z_1) = a_2 \sum a_1 \tag{2.4.2}$$

$$= a_2' a_1 \lambda_1 \tag{2.4.3}$$

$$= 0. \tag{2.4.4}$$

Thus $a_2'a_1 = 0$ (Afifi et al. 2003, Timm 2002)

If a_2 is the second largest eigenvector (the coefficients of the original variables used to construct factors) of the eigenequation $\sum P_2 = P_2\lambda_2$. $Var(Z_2) = a_2' \sum a_2 = \lambda_2$ where $\lambda_1 \geq \lambda_2$.

The second principal component explains more but fewer variation in the original variable than the first component due to the same constraint. Principal components up to the maximum of p are all defined in the same way. $Z_1 \ldots Z_p$ components are uncorrelated and the squares of their coefficients add to 1. PCA determines the eigenvalues and eigenvectors of the correlation matrix (Dunteman 1989).

In the case of p variables $X_1, X_2 \ldots X_p$. Each principal component is a linear combination of the X variables. Coefficients of linear combinations should satisfy the following 3 requirements:

(1) $VarZ_1 \geq VarZ_2 \geq \ldots \geq VarZ_p$

(2) The values of any 2 principal components are uncorrelated.

(3) For any principal component the sum of the squares of the coefficients is 1.

A rule to follow is that a sufficient number of principal components should be kept to explain a certain percentage of the total variance. Components with small variances can be discarded. Another method to choose the principal components is to use scree plots. One looks at the cut-off point at the lines joining the steep left of the cut-off

point and the flat right of the cut-off point. The cut-off point is sufficient to explain the original variables. This method can sometimes fail if the change on the slope is not clear (Afifi et al. 2003, Timm 2002, Dunteman 1989).

2.4.1 Transforming coefficients to correlations

The coefficient a_{11} can be transformed into a correlation between X_1 and Z_1. The correlation between the ith principal component and the jth variable is

$$r_{ij} = \frac{a_{ij}(VarZ_i)^{1/2}}{(VarX_j)^{1/2}} \tag{2.4.5}$$

where a_{ij} is the coefficient of X_j for the principal component. If the first component is positively correlated with all the original variables we expect high and positive correlations.

2.4.2 Using standardised variables

Before performing PCA it is better to standardise the X variables by dividing each variable with its sample standard deviation. Standardisation is the same as analysing the correlation matrix instead of the covariance matrix. Principal components derived from correlation matrix are interpreted in the following ways:

(1) The total variance is simply the number of variables p, and the proportion explained by each principal component is the corresponding eigenvalue divided by p.

(2) The correlation between the ith principal component and Z_i and the jth variable X_j is $r_{ij} = a_{ij}(VarZ_i)$. Some computer programmes call the correlation of Z_i and a_{ij} Factor Loading.

2.4.3 Application of PCA

SPSS was used to develop indices for IPV, media and decision making using PCA. Data for IPV variables was classified into an index of control, and emotional, physical and sexual violence that women are exposed to. This

section mainly focuses on indices of physical and sexual violence.The indices of media exposure and decision making, will be used in the Multinomial Logit regression analysis to identify the factors that are associated with IPV. Table 2.3 shows the variables used to construct the indices.

Table 2.3: List of variables used and the responses used to generate indices

Variables	Response
Physical Violence	(Yes/No)
Sexual Violence	(Yes/No)
Media exposure	(Not at all/ Less than once a week/ At least once a week/ Almost everyday)
Decision making	(Respondent/ Husband)

2.5 Results from SPSS of the principal components

In Table 2.4 we present the principal components of sexual violence, physical violence, media exposure and decision making. (Afifi et al. 2003) The results of the PCA show that the first principal component for physical violence

Table 2.4: Indices and corresponding component score for all constructs

CONSTRUCT	VARIABLES	COMP 1	COMP 2
PHYSICAL VIOLENCE	Spouse ever pushed, shook or threw something	0.76	
	Spouse ever slapped	0.69	
	Spouse ever punched with fist or something harmful	0.77	
	Spouse ever tried to strangle or burn	0.78	
	Spouse ever threatened with knife/gun or other weapon	0.54	
	Spouse ever attacked with knife/ gun or other weapon	0.48	
PERCENT VARIANCE EXPLAINED		46.27%	
SEXUAL VIOLENCE	Spouse ever physically forced sex when not wanted	0.92	
	Spouse ever forced other sexual acts when not wanted	0.92	
PERCENT VARIANCE EXPLAINED		84.63%	
MEDIA EXPOSURE	Frequency of watching television	0.86	
	Frequency of listening to radio	0.81	
	Frequency of reading newspapers	0.78	
PERCENT VARIANCE EXPLAINED		67.00%	
DECISION MAKING	Who decides how to spend money	0.39	0.55
	Final say on own health	0.60	−0.09
	Final say on making large household purchases	0.66	−0.30
	Final say on making household purchase for daily needs	0.69	−0.03
	Final say on visits to family or relatives	0.61	−0.25
	Final say on deciding what to do with money husband earns	(0.28	0.76
PERCENT VARIANCE EXPLAINED		30.98%	17.38%

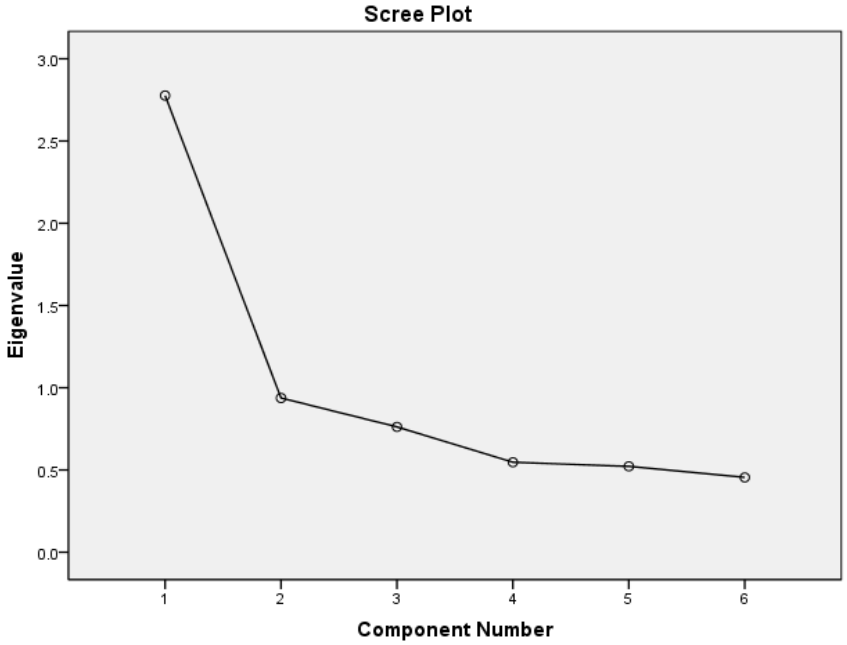

Figure 2.1: Screeplot for physical violence variables

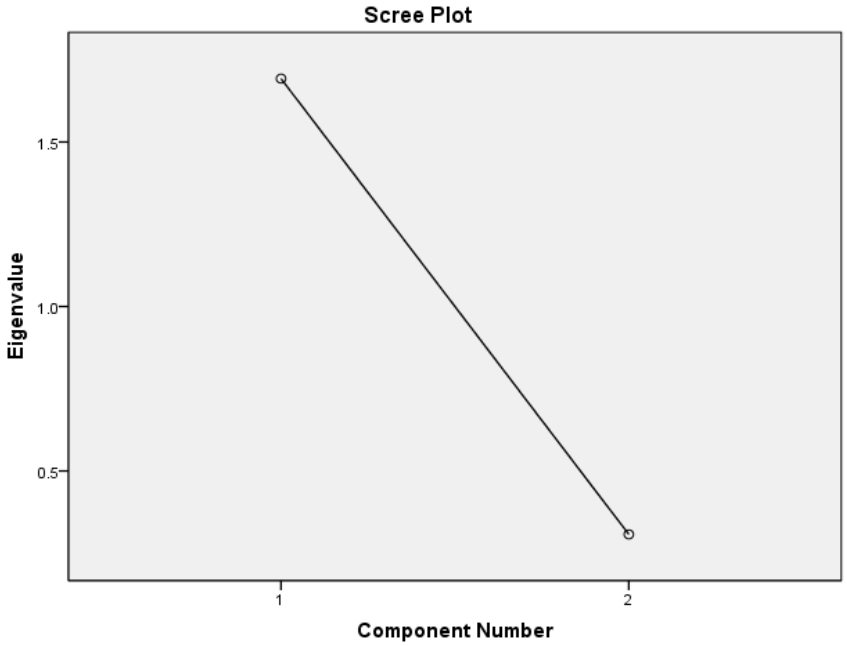

Figure 2.2: Screeplot for sexual violence variables

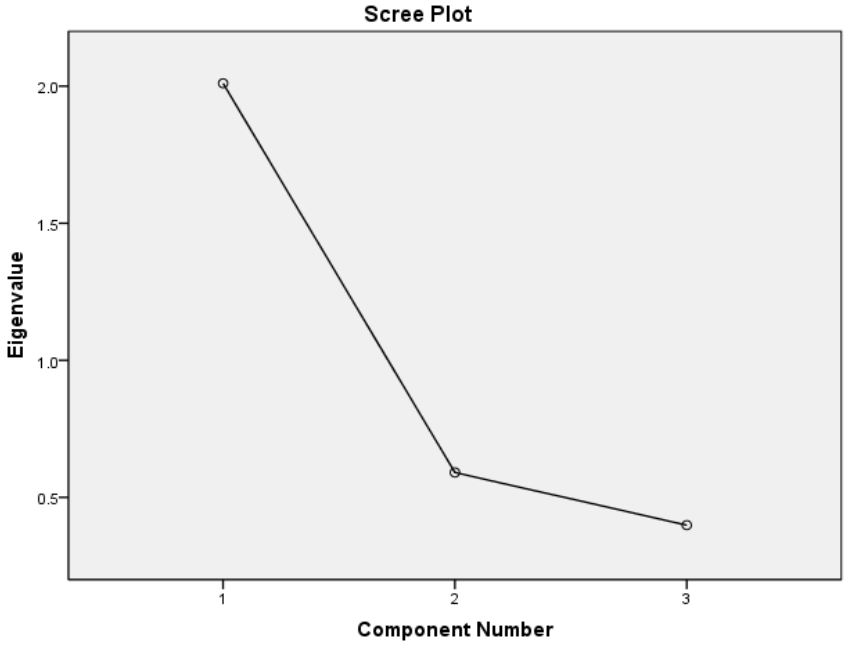

Figure 2.3: Screeplot for media variables

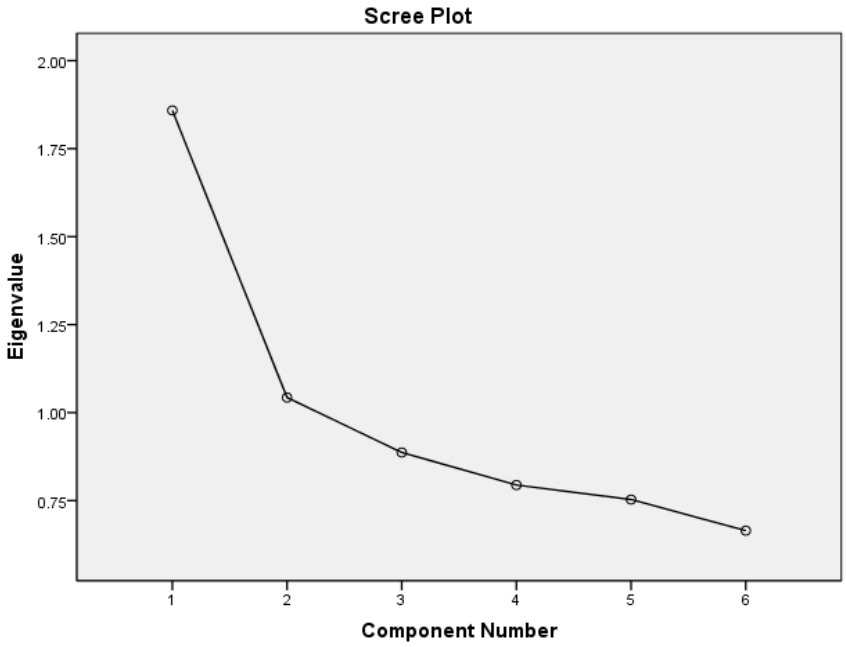

Figure 2.4: Screeplot for decision making variables

explains 46.27% of the variation of the original variables and each subsequent component explains a decreasing proportion of variance as shown in 2.4. For example the second component shows 15.63% and the third 12.70%

and so on.

The proportion of variance explained by the scree plot indicates that the first 4 components would give enough information to explain the original variables. The factor score, that is the eigenvectors of the first principal component, will be used to construct the physical violence index.

The variables of the first component have high coefficients indicating high correlation between the variables and the principal component. The correlations of all the variables with the first component are positive. Most women experienced physical violence, such as when their spouses pushed, shook, or threw something at them. Spouses also punched or used something, used harmful, as well as trying to strangle or burn them.

The sexual violence variables only formed one component which was explained by 84.63% variation of the original variables. The second component only shows 15.37%. The scree plot only shows that the one component gives enough information to explain the original variables.

The media exposure variables show that the first principal component explains 67% variation of the original variables. The second and third component explains a decreasing proportion of variance of 19.70% and 13.29% respectively.

The decision making variables show that 30.98% of the first component explain the variation of the original variables, followed by 17.38% of the second component and so on. The decision making index variables are not highly correlated as the coefficients are low and this signifies a low correlation between the variable and the principal component as shown in table 2.4. The loadings for decision making are also negative and less than 0.3 on the second principal component thus making it a weak index (Afifi et al. 2003).

2.6 Summary statistics

Table 2.5 shows the demographic variables of how the respondents were distributed by each covariate. Table 2.6

Table 2.5: Distribution of respondents by covariates

Variable	Category	N(%)
age 5-year groups	15-19	253(6)
	20-24	911(23)
	25-29	912(22)
	30-34	800(19)
	35-39	544(13)
	40-44	431(10)
	45-49	347(8)
Total		4,198(100)
religion	Traditional	135(3)
	Roman,Pentecos,Protes,Other church	2,245(53)
	Apostolic,Muslim,None, Other	1,818(43)
Total		4,198(100)
marital status	married	3,434(82)
	separate	764(18)
Total		4,198(100)
education	no education	224(5)
	primary	1,653(39)
	secondary	2,212(53)
	higher	109(3)
Total		4,198(100)
wealth index	poorest	948(23)
	poorer	895(21)
	middle	780(19)
	richer	946(23)
	richest	629(15)
Total		4,198(100)
STI	no	3,983(95)
	yes	195(5)
Total		4,198(100)

shows the cross tabulation of HIV status with each of the covariates.

Table 2.6: Sociodemographic breakdown of HIV prevalence and p-values

Variable	Category	HIV Positive (n) %	p-value
age 5-year groups	15-19	28(11)	
	20-24	146(16)	
	25-29	246(27)	
	30-34	260(33)	
	35-39	184(34)	
	40-44	117(27)	
	45-49	70(20)	
Total		1,051(25)	0
religion	Traditional	20(15)	
	Roman,Pentecos,Protes,Other church	578(26)	
	Apostolic,Muslim,None, Other	453(25)	
Total		1,051(25)	0.02
marital status	married	680(20)	
	separate	371(49)	
Total		1,051(25)	0
education	no educa	48(21)	
	primary	412(25)	
	secondar	570(26)	
	higher	21(19)	
Total		1,051(25)	0.25
wealth index	poorest	211(22)	
	poorer	207(23)	
	middle	209(27)	
	richer	279(29)	
	richest	145(23)	
Total		1,051(25)	0
STI	no	943(24)	
	yes	101(52)	
Total		1,051(25)	0

2.7 Summary of cross-classifying HIV

Table 2.6 shows that the following socio-demographic covariates are associated with HIV or have an effect on a woman getting HIV. These are age, religion, marital status, wealth and the presence of STI in a woman. Mostly women who are between the age of $25 - 44$ are HIV positive, with the majority falling in the range of $35 - 39$. We also noted that those who follow the following religions: Roman Catholic, Pentecost, Protestant, Apostolic, and Muslim have more women who are HIV positive than those who are involved in the traditional religion.

There is a high percentage of women who are HIV positive among those women who are separated, than among

those who are married. The results also reveal the proportionality that there are more women who are HIV positive among those who are rich or middle class than among those who are poor. More women who have an STI are HIV positive, than those who do not have an STI.

Education has no influence on a woman being HIV positive. As a result of what we have just analysed there is a need for further analysis using statistical methods such as logistic regression modelling and generalised linear mixed models to see if the above mentioned covariates have an association with HIV or not. These methods will be discussed in chapter 3 and chapter 5.

Chapter 3

Survey logistic regression analysis of IPV

After developing indices of physical and sexual violence in Chapter 2, we now apply survey logistic regression analysis to assess the relationship between these indices and HIV serostatus.

3.1 Introduction

Logistic Regression is a common mathematical modelling procedure used in the analysis of epidemiological data (Archer, Lemeshow & Hosmer 2007, Kleinbaum, Dietz & Krickeberg 1994). Thus logistic regression analyses the association between a categorical response variable and a set of explanatory variables. In research we usually want to know what the relationship is of one or more exposure (or study) variables to a disease or illness outcome (Kleinbaum et al. 1994). In this study, we evaluate the extent to which IPV is associated with HIV among women, controlling for additional variables such as the age, physical, emotional, psychological and sexual violence on women, among other covariates.

3.2 Statistical modelling

3.2.1 The exponential family of distributions

A number of distributions belong to the exponential family. These include discrete distributions like the Bernoulli or Poisson distribution and continuous as the Gaussian (normal) or Gamma distribution. Suppose $Z_i(i = 1, \cdots, n)$ is a set of independent random response variables. Z_i belongs to the exponential family if its probability (density) function can be written as:

$$f(z_i; \xi_i) = r(z_i)s(\xi_i)exp[t(z_i)u(\xi_i)] = exp[t(z_i)u(\xi_i) + v(z_i) + w(\xi_i)]$$

ξ_i is the location parameter (Lindsey 1997).

Further, let $y = t(z)$ and $\theta = \mu(\xi)$ then we obtain the canonical form and the model becomes:

$$f(y_i; \theta_i) = exp[y_i\theta_i - b(\theta_i) + c(y_i)],$$

where $b(\theta_i)$ is the normalising constant of the distribution and $Y_i(i = 1, \dot{,} n)$ is a set of independent random response variables with mean (μ_i). Therefore $y_i = \mu_i + \epsilon_i$. The exponential family can be generalised by letting ϕ be a (constant) scale parameter such that:

$$f(y_i, \theta_i, \phi) = exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right].$$

where θ_i is the natural parameter or canonical form of the location parameter and some function of the mean μ_i (Lindsey 1997, McCullagh & Nelder 1989). Also $a_i(\theta)$ has the form $a_i(\theta) = \frac{\phi}{w_i}$ for known weight w_i, where ϕ is the dispersion parameter. When y_i is a mean of n_i independent readings the $w_i = n_i$. The dispersion parameter ϕ is a nuisance parameter which can be used in exponential family distributions such as the normal or gamma but is not required for one parameter families, such as the binomial and Poisson (Lindsey 1997).

If a response Y has a distribution in the exponential family then there is a special relationship between the mean and variance. The relationship between the mean and variance is given for any likelihood function $L(\theta_i, \phi; y_i) = f(y_i, \theta_i, \phi)$. The first derivative of its logarithm for one observation is given by:

$$U_i = \frac{\partial log[L(\theta_i, \phi; y_i)]}{\partial\theta}$$

This is the score function. Setting the score function to zero, the resulting score equations yield the maximum likelihood estimates (Lindsey 1997). The standard inference theory shows that $E[U_i] = 0$ and $var[U_i] = E[U_i^2] = E[-\frac{\partial U_i}{\partial \theta_i}]$. For the exponential dispersion family,

$$log[L(\theta_i, \phi, y_i)] = exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right]$$

Then forθ_i, $U_i = \frac{y_i - \frac{\partial b(\theta_i)}{\partial \theta_i}}{a_i(\phi}$ so that

$$E[Y_i] = \frac{\partial b(\theta_i)}{\partial \theta_i},$$

$$= \mu_i.$$

Let $U = \frac{\partial l}{\partial \theta}$ then

$$U_i' = -\frac{\frac{\partial^2 b(\theta_i)}{\partial \theta_i^2}}{a_i(\phi)}$$

Thus we have variance:

$$var[U_i] = \frac{var[Y_i]}{a_i^2(\phi)} = \frac{\frac{\partial^2 b(\theta_i)}{\partial \theta_i^2}}{a_i(\phi)}(Lindsey\ 1997).$$

The three components of the Generalized Linear Model as outlined by McCullagh & Nelder (1989) are:

- The random component describes the conditional distribution of the response Y with explanatory variables, which is a member of exponential family such as normal, poisson, gamma and binomial.
- The systematic component involving the explanatory variables $x_1, x_2, \cdots, x_p$ used as a linear predictor
- The third component is the link function g, that links the predictor to the natural mean of the response variable Y.

Linear predictor

A set of $p + 1$ (usually) unknown parameters,$\beta_i (i = 0, 1, 2, \cdots, p)$ and the design matrix of known explanatory variables $\boldsymbol{X}_{n\times(p+1)}$ define a linear predictor η given by:

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

where $\mathbf{X}\beta$ is the linear structure. The $i'th$ row of $\mathbf{X}$ is given by $x_i = (1, x_{i1}, \cdots, x_{ip})'$ with x_{ij}, $i = 1, \cdots, n$; equal to the value of the $j'th$ predictor or explanatory variables x_j, $i = 1, \cdots, p$ and $\beta' = (\beta_0, \beta_1, \cdots, \beta_p)$ is a vector of regression coefficients including the constant β_0 corresponding to $X_0 = 1$ (Lindsey 1997).

Link function

The link function that is $g_i(.)$, gives the relationship between the mean of the $i'th$ observation and its linear predictor so that:

$$\eta_i = g_i(\mu_i) = \mathbf{X}'_{\mathbf{i}}\beta,$$

The link function must be monotonic and differentiable. The canonical link function is that function which makes the linear predictor η_i the same as the canonical parameter θ_i, from the exponential family member. With the canonical link function, all unknown parameters in β have sufficient statistics if the response distribution is a member of the exponential family with known scale parameter (Lindsey 1997).

3.3 The logistic regression model

3.3.1 Modelling of binomial data

As previously stated the data used in this thesis is binary and binomial data obtained from the DHS. In this section we describe what binary and binomial data is and how it is used in logistic regression modelling. The responses for HIV status for women in Zimbabwe in the year $2005 - 2006$ is binary and follows a binomial distribution.

3.3.2 Binary data and responses

The response variable Y can take two possible outcomes that is either a 'success' or 'failure' denoted by 1 or 0 respectively. Let π_i and $1-\pi_i$ be the probabilities of success and failure respectively the on the i'th $(i = 1, \cdots, N)$ observational units then, $\mathrm{P}r(Y_i = 1) = \pi_i$ and $\mathrm{P}r(Y_i = 0) = 1 - \pi_i$. These are the probabilities of 'success' and 'failure' respectively. In statistics our objective is to investigate the relationship between the response probability

$\pi = \pi(x)$ and the explanatory variables $x = (x_i, \cdots, x_p)$. Binary data are ungrouped data which lists observations by individual experimental units (McCullagh & Nelder 1989).

3.3.3 Covariate classes

A covariate class is formed if the ith combination of experimental conditions are characterised by the p-dimensional vector $(x_{i1}, \cdots, x_{ip})$. The observations are available on m_i individuals. $N = m_1 + m_2 + \cdots + m_n$ are the individuals under study and m_i share the covariate vector $(x_{i1}, \cdots, x_{ip})$.

The responses of binary data grouped by a covariate class have the form $(y_1|m_1, \cdots, y_n|m_n)$ where $0 \leq y_i \leq m_i$ is the number of successes out of the m_i subjects in the ith covariate class. The binomial index vector or binomial denominator vector is a vector of covariate class sizes $\mathbf{m} = (m_1, \cdots, m_n)$. $m_1 = \cdots = m_n = 1$ is a special case for ungrouped data or data listed by individual subjects.

McCullagh and Nelder (1989) state two reasons for the distinction between grouped (binomial) and ungrouped (binary) data. These are:

(1) Not all methods of analysis applicable to binary data are appropriate for binomial data especially involving normal approximation.

(2) The asymptotic approximation appropriate for models of binary data is $N \to \infty$. Asymptotic approximations for models applied to binomial data can take either $m \to \infty$ or $N \to \infty$.

3.3.4 Bernoulli distribution

If $m = 1$ that is the response of a single experimental unit then we get the Bernoulli distribution given by $\mathrm{P}r(Y_i = 1) = \pi_i$ and $\mathrm{P}r(Y_i = 0) = 1 - \pi_i$ (McCullagh & Nelder 1989).

3.3.5 Binomial distribution

If Y has positive counts which are bound by fixed values then this gives rise to a binomial distribution. A binomial distribution can arise from a set of Poisson or Bernoulli distributions. In the case of Poisson distribution, let Y_1, Y_2

be independent Poisson random variables with means μ_1 and μ_2. The sum $Y_1 + Y_2$ has the Poisson distribution with $\mu_1 + \mu_2$. Then the the conditional distribution of Y_1 given $Y_1 + Y_2 = m$ is given by:

$$\Pr(Y_1 = y | Y_1 + Y_2 = m) = \binom{m}{y} \pi^y (1-\pi)^{m-y} \quad y = 0, 1, \cdots, m \tag{3.3.1}$$

where $\pi = \mu_1 | \mu_1 + \mu_2$. This implies that the conditional distribution depends only on the ratio of the Poisson means and not on $\mu_1 + \mu_2$. Thus $Y \sim B(m, \pi)$ means that Y has the binomial distribution (3.28) with index m and parameter π.

The sum of independent homogeneous bernoulli trials grouped by covariate class give rise to the binomial distribution that is:

$$\Pr(Y_i = y_i) = \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{m_i - y_i} \quad y = 0, 1, \cdots, m \tag{3.3.2}$$

where π_i is the true probability of success in the i'th covariate class (McCullagh & Nelder 1989).

3.3.6 Fitting the linear logistic model to binomial data

Suppose that Y_i, i $= 1, \cdots, n$ are independent binomial random variables, where $Y_i \sim BIN(n_i, \pi_i)$. It follows that the logistic regression model is given by:

$$\begin{aligned} g(\pi_i) &= \log\left(\frac{\pi_i}{1-\pi_i}\right), \\ &= \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}, \\ &= X_i' \boldsymbol{\beta} \end{aligned} \tag{3.3.3}$$

where the $i'th$ row of $\mathbf{X}$ is given by $x_i = (1, x_{1i}, \cdots, x_{pi})'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)'$.

The binomial distribution belongs to the exponential family. The binomial probability distribution function in (3.29) is equivalent to:

$$f_Y(y_i; \pi_i) = exp\left(y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) + \log\binom{n_i}{y_i}\right)$$

Where y_i is the natuarl parameter and π_i is the scale parameter. Since the link function gives the relationship between the mean and the $i'th$ observation and its linear predictor theng $(\pi_i) = g(\mu_i)$. Applying the Newton-Raphson procedure with:

$$g(\mu_i) = \log\left(\frac{n_i \pi_i}{n_i - n_i \pi_i}\right) = \log\left(\frac{\mu_i}{n_i - \mu_i}\right) = \eta_i. \tag{3.3.4}$$

Using $w_i = \frac{\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2}{var(Y_i)} = \frac{1}{\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2 var(Y_i)}$, and taking partial derivatives with respect to μ_i in equation 3.3.4 we have,

$$\frac{\partial \eta_i}{\partial \mu_i} = \frac{\partial}{\partial \mu_i} \log\left(\frac{\mu_i}{n_i - \mu_i}\right) = \frac{\frac{n_i}{(n_i-\mu_i)^2}}{\frac{\mu_i}{(n_i-\mu_i)}} = \frac{n_i}{\mu_i(n_i-\mu_i)} = \frac{1}{n_i\pi_i(1-\pi_i)}. \tag{3.3.5}$$

Thus $\frac{\partial \mu_i}{\partial \eta_i} = n_i\pi_i(1-\pi_i)$ and $w_i = \frac{[n_i\pi_i(1-\pi_i)]^2}{n_i\pi_i(1-\pi_i)} = n_i\pi_i(1-\pi_i)$

Since $var(Y_i) = n_i\pi_i(1-\pi_i)$ and the Fisher scoring equation for using least squares to fit a linear model given by:

$$\beta^{(k+1)} = \boldsymbol{X'W}^{(k)}\mathbf{X})^{-\mathbf{1}}\mathbf{X'W}^{(\mathbf{k})}\mathbf{z}^{(\mathbf{k})}$$

which is used iteratively to find the maximum likelihood estimates of the parameter $\boldsymbol{\beta}$. $\mathbf{W}^{(\mathbf{k})}$ is $\mathbf{W}$ evaluated at the k'th iteration. The i-th element of $\mathbf{z}^k$ is:

$$z_i^{(k)} = \log\left(\frac{\pi_i^{(k)}}{1-\pi_i^{(k)}}\right) + \frac{y_i - n_i\pi_i^{(k)}}{n_i\pi_i^{(k)}(1-\pi_i^{(k)})}.$$

The multiple logistic regression model is written as

$$\text{logit}(\pi_i) = \beta_0 + X_{i1}\beta_1 + \cdots + X_{ik}\beta_k = \mathbf{X_i'}\boldsymbol{\beta} \tag{3.3.6}$$

where $\mathbf{X_i'} = (\mathbf{X_{i1}}, \cdots, \mathbf{X_{ik}})'$ is a vector of independent variables corresponding to the i-th individual and $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_k)$ is a vector of unknown parameters. Further

$$\pi_i = P(Y_i = 1|X_i')$$

The linear predictor π in the equation 3.3.6 can be expressed as $\eta_i = \mathbf{X}_i\boldsymbol{\beta}$, (Lindsey 1997) . The mean of the i-th observations and the linear predictor is given by the link function

$$\eta_i = g_i(\mu_i) = \mathbf{X_i'}\boldsymbol{\beta},$$

where

$$\mu_i = g_i^{-1}(\eta_i) = G(\eta_i).$$

G is the inverse link function. The properties of a link function are that it must be monotonic and differentiable. The link function is called a canonical link function when the linear predictor η_i is the same as the canonical

parameter θ_i from the exponential family (Lindsey 1997).That is:

$$\mathbf{X}_i'\beta = \eta_\mathbf{i} = \theta_\mathbf{i}$$

The canonical link function of a binomial distribution is given by

$$\eta_i = \log\left[\frac{\pi_i}{1-\pi_i}\right] = \log\left[\frac{\mu_i}{n_i-\mu_i}\right].$$

Given a response distribution which is a member of the exponential dispersion family, and the scale parameter is known when using the canonical link function then all unknown parameters of the linear structure have sufficient statistics in β (Lindsey 1997).

3.4 Stages in building the logistic regression model

There are four stages in building a logistic regression model of survey data. These are:

(1) Model specification.

(2) Estimation of model parameters and their standard errors.

(3) Model evaluation and diagnostics.

(4) Interpretation of results and inference based on the final model.

3.4.1 Model specification

The best logistic regression model for the survey data is formed by identifying the predictors and evaluating them individually and in the multivariate context with other relevant explanatory variables. To specify the initial model the Hosmer and Lemeshow's (2000) incremental process is recommended then the final logistic model is predicted by using the following method:

(1) Initially perform a bivariate analysis of the relationship of y (outcome) to individual predictor variable candidates.

(2) Using the significance $p < 0.25$ as candidates for main effects select the predictors that have a bivariate association.

(3) Using the Wald test evaluate the contribution of each predictor to the multivariate model.

(4) Check the assumption as to whether there is a linear relationship for the continuous variables.

(5) The interactions among the predictors should be justified scientifically.

There is a final step recommended by Hosmer and Lemeshow (2000) of applying the polynomial functions and smoothing splines to test whether the logistic model is really linear in the logit.

3.4.2 The likelihood function

The likelihood function L, is the function of the unknown parameters denoted as $L(\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ denotes the vector of unknown parameters being estimated in the model. The joint probability, or likelihood of observing the data that have been collected is given as $L = L(\boldsymbol{\beta})$ and logit$\pi(x) = \mathbf{X_i'}\boldsymbol{\beta}$ where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \cdots, \beta_k)$.

3.4.3 Estimation of model parameters and standard errors

After specifying our model the second step is to estimate the model's parameters and standard errors. Maximum Likelihood (ML) estimation is one of the methods used to estimate the parameters in a mathematical model. Another method is the least squares (LS) estimation, mainly used in estimating the parameters in a classified straight line or multiple linear regression model.The variables can be nominal, ordinal and interval when using ML estimation.

The two ML approaches used are unconditional and conditional methods. When choosing the ML method the number of parameters relative to the total number of subjects in one's model play an important role. Unconditional ML estimation is preferred if the number of parameters in the model is small, relative to the number of subjects. Conditional ML prefers large numbers of parameters that is more than 2 clusters or groups like in a survey of a country relative to the subjects (Heeringa, West & Berglund 2010).

For example if we have a logistic regression of the form $logit(\pi(x)) = \beta_0 + X_{i1}\beta_1 + \cdots + X_{ik}\beta_k$. If the sample

is a simple random sample (SRS) the parameters $\beta_0, \beta_1, \cdots, \beta_k$ are estimated using the method of Maximum Likelihood (ML). The likelihood function for a SRS with n observations on a binary variable y with possible values 0 and 1 is based on the binomial distribution given:

$$L(\boldsymbol{\beta}|x) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}, \tag{3.4.1}$$

where

$$\pi(x_i) = \frac{e^{(x_i\boldsymbol{\beta})}}{1+e^{(x_i\boldsymbol{\beta}).}} \tag{3.4.2}$$

Maximum Likelihood Estimation is not possible when survey data has been collected under a complex sample design. In this case we require sampling weights to estimate the logistic regression parameters.The weighted least squares (WLS) estimation and the pseudo-maximum likelihood estimation (PMLE) are the two methods used to estimate model parameters for logistic models of complex survey data. If we have a finite population the regression parameters are the values that maximise a likelihood equation for $i = 1, \cdots, N$ elements in the population under survey.

The population likelihood for a binary dependent variable y is given by:

Let $\boldsymbol{\beta}$ be the finite population model parameters.

$$L(\boldsymbol{\beta}|x) = \prod_{i=1}^{N} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \tag{3.4.3}$$

where $\pi(x_i) = \frac{e^{(x_i\boldsymbol{\beta})}}{1+e^{(x_i\boldsymbol{\beta})}}$

If we maximise the estimates of the population likelihood which is a weighted function of the sample data and the $\pi(x_i)$ values, we get the estimates of the finite population regression parameters as shown below:

$$PL(\boldsymbol{\beta}|x) = \prod_{i=1}^{n} \{\pi(x_i)^{y_i}.[1-\pi(x_i)]^{1-y_i}\}^{w_i} \tag{3.4.4}$$

where $\pi(x_i) = \frac{e^{(x_i\boldsymbol{\beta})}}{1+e^{(x_i\boldsymbol{\beta})}}$

The iterative Newton-Raphson method is used to maximise the weighted pseudo-likelihood function. This method will be discussed in the next section.

When analysing logistic regression models for complex survey data it is important to estimate the sampling variances and covariances of the parameter estimates. This is achieved by applying the multivariate version of Taylor series linearisation (TSL). As a result we get the **sandwich-type variance estimator** which is:

$$var(\hat{B}) = (\mathbf{J}^{-1})var[S(\hat{B})](\mathbf{J}^{-1}) \tag{3.4.5}$$

If we apply the natural log function to the likelihood defined in (3.38) we get $\hat{B_J}$ the pseudo-log-likelihood. **J** is a matrix of second derivatives with respect to $\hat{B_J}$. $var(S(\hat{\mathbf{B}}))$ is the variance-covariance matrix and $var(\hat{B})$ is the default variance.

The Newton-Raphson method

One of the methods for solving nonlinear equations is the Newton-Raphson Method. The examples are likelihood equations that show where a function is maximised. The Newton-Raphson method determines the value $\hat{\beta}$ of β that maximises a function $\tau(\beta)$. Let $(g\prime = (\frac{\partial \tau}{\partial \beta_1}, \frac{\partial \tau}{\partial \beta_2}, \cdots)$, and let the Hessian **H** and $\mathbf{H} = h_{xy}$ denote the matrix of the second derivative that is $\mathbf{H} = \frac{\partial^2 \tau}{\partial \beta_x \partial \beta_y}$.

Let $g^{(m)}$ and $\mathbf{H}^{(m)}$ be the values evaluated at $\beta^{(m)}$, the mth guess for $\hat{\beta}$. The mth iteration process $(m = 0, 1, 2, \cdots)$, $\tau(\beta)$ is approximated near $\beta^{(m)}$ by the terms up to second order in its Taylor series expansion

$$Q^{(m)}(\beta) = \tau(\beta^{(m)}) + g^{(m)\prime}(\beta - \beta^{(m)}) + (\frac{1}{2})(\beta - \beta^{(m)})\prime \mathbf{H}^{(m)}(\beta - \beta^{(m)}), \tag{3.4.6}$$

then

$$\frac{\partial Q^{(m)}}{\partial \beta} = g^{(m)} + \mathbf{H}^{(m)}(\beta - \beta^{(m)}) = 0 \tag{3.4.7}$$

after solving for β in equation 3.4.7 we thus get

$$\beta^{(m+1)} = \beta^{(m)} - (\mathbf{H}^{(m)})^{(-1)} g^{(m)} \tag{3.4.8}$$

assuming $\mathbf{H}^{(m)}$ is non-singular. $\beta^{(0)}$ denotes the first estimate of $\hat{\beta}$. Then at each iteration $\beta^{(m)}$ is used to obtain $H^{(m)}$ and $g^{(m)}$ which are then used to estimate $\beta^{(m+1)}$ which is used to obtain $\beta^{(m+2)}$ and the process continues until convergence.

3.5 Evaluation of the fitted Model

3.5.1 Statistical inference for logistic regression

The estimates obtained using the ML estimation are used to make statistical inferences concerning the exposure-disease relationships.In this stage we are testing hypotheses and obtaining confidence intervals for parameters

in the model. Statistical inference is achieved by using the maximised likelihood value, that is the numerical value of the likelihood function L. When the ML estimates $\hat{\beta}$ that is $L(\hat{\beta})$ The second value is the estimated variance-covariance matrix denoted $\hat{V}\hat{\beta}$. The values on the diagonal of the matrix are the estimated variances and those off the diagonals are the covariances of ML estimates. The more the parameters a model has the better it fits the data and the law of parsimony can be used to choose the best parameters that fit the data (Heeringa et al. 2010). In the law of parsimony statistical methods like least squares and maximum likelihoods are used to identify the parameters to fit the model

The higher the $\hat{L}$ the better the model fit. If $\hat{L}_1 \leq \hat{L}_2 \leq \hat{L}_3$ then $\ln \hat{L}_1 \leq \ln \hat{L}_2 \leq \ln \hat{\leq} L_3$ multiply each log by -2 $-2 \ln \hat{L}_3 \leq -2 \ln \hat{L}_2 \leq -2 \ln \hat{L}_1$. The statistic $-2 \ln \hat{L}_i$ is the log likelihood statistic. The statistic can be used to test the hypothesis about parameters in the model using likelihood Ratio test.

Wald test

The Wald Test is used in logistic regression to carry out hypothesis testing. It is done when there is only one parameter to be tested. For large N Wald statistic is given by $Z = \frac{\hat{\beta}}{s_{\hat{\beta}}}$ and is approximately $N(0,1)$ or Z^2 is approximately chi-square statistic with one degree of freedom. The information for the Wald test is contained in the output where we get the variables in the model and the ML coefficients and standard errors. The Wald test is best suited for large samples. The statistical significance of one or more logistic regression parameters can be evaluated using a likelihood ratio test. According to Heeringa et al. (2010)The null hypothesis $H_0 : \beta_j = 0$ (single parameter) or $H_0 : \beta_q = 0$ (with q parameters) the test statistic G follows a χ^2 distribution with 1(for a single parameter) or qd degrees of freedom.

$$G = -2 \ln \left[\frac{L(\hat{\beta}_{MLE})_{reduced}}{L(\beta_{MLE})_{full}} \right] \tag{3.5.1}$$

Where $L(\hat{\beta}_{MLE} =$ the likelihood under the model evaluated at the maximum likelihood estimates of β

Wald tests are used to assess the significance of each regression coefficient β_J in the model. $H_0 : C\beta = 0$, where C is the constant matrix of the hypothesis being tested, is the null hypothesis of the Wald test. Wald test statistic is given by:

$$\chi^2_{wald} = (C\hat{\beta})'[Cvar(\hat{\beta})C']^{-1}(C\hat{\beta}) \tag{3.5.2}$$

$var(\hat{\beta})$ is a design -consistent estimate of the variance-covariance matrix of the estimated logistic regression coefficients. The square roots of the diagonal elements of the inverse information matrix are the standard errors of the regression coefficients. The wald statistic follows an approximate standard normal distribution $N(0,1)$.To reject the null hypothesis $H_0 : \beta_q = 0$ the $\chi^2_{wald} > \chi^2_{\alpha,1}$ and conclude that the predictor variable associated with β_q is significant to the model.

Confidence intervals for individual regression coefficients can be constructed using Wald tests. A confidence interval of the q'th regression coefficient can be estimated by

$$\hat{\beta}_q - \chi^2_{wald} s.e.(\hat{\beta}_q) < \beta_q < \hat{\beta}_q + \chi^2_{wald} s.e.(\hat{\beta}_q) \tag{3.5.3}$$

Score test

This is an alternative of the Wald test. A score test corresponding to the q'th regression coefficient is the derivative of the log likelihood with respect to β_q.This is given by:

$$V_q = \frac{\partial l}{\partial \beta_q}. \tag{3.5.4}$$

The score test statistic is:

$$Z_s = \frac{\hat{V}_q}{s.e.(\hat{V}_q)}, \tag{3.5.5}$$

The square root of the diagonal elements of the information matrix are the standard errors of the score statistics and the Z_s follows a standard normal distribution.

3.5.2 Goodness of fit and logistic regression diagnostics

The deviance

A measure of discrepancy or goodness of fit, called the deviance, was introduced by Nelder and Wedderburn (1972). Let μ denote the mean value parameter, θ the canonical parameter and ϕ some dispersion parameter. Let $l(\hat{\mu}, \phi, y)$ be the log-likelihood maximised over the vector of parameters β for a fixed value of ϕ and $l(\hat{\mu}, \phi, y)$ be the maximum log-likelihood achievable in the saturated model. The scaled deviance is given by:

$$D^* = -2\frac{[\log(\hat{\mu}, \phi, y) - \log(y, \phi, y)]}{\phi}. \tag{3.5.6}$$

The scaled deviance D^* is the deviance expressed as a multiple of the dispersion parameter ϕ. If $\phi = 1$ then the deviance is defined as

$$D = -2[\log(\hat{\mu}, \phi, y) - \log(y, \phi, y)] \tag{3.5.7}$$

If $\phi = 1$ or $\phi \neq 1$ (but known) then we can measure the closeness of the fit of a model to the data. With a small deviance with the log-likelihood close to $l(y, \phi, y)$ then the model describes the data well. A large deviance shows that the data is not fitted well by the model. D has an approximate χ^2 distribution with $n - p - 1$ degrees of freedom. P is the number of explanatory variables in the linear predictor. The fitted model is said to be adequate if $D \leq \chi^2_{\alpha, n-p-1}$. If the value of dividing the deviance by its degrees of freedom $n - p - 1$ is close to 1 then we can conclude that the fitted model is adequate. A large value means that the model is incorrectly specified.

Another important measure of discrepancy is the generalised Pearson χ^2 statistic. This is given by:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \tag{3.5.8}$$

where ($\hat{\mu}_i$ is the estimated variance function for the distribution concerned. For normal-theory linear models, the deviance and the generalised Pearson χ^2 statistic have exact χ^2 distributions with $n - p - 1$ degrees of freedom. Usually the Deviance and Pearson χ^2 statistic has an asymptotic distribution. The deviance as a measure of discrepancy has an advantage that it is an additive for nested sets of models if maximum likelihood estimates are used (McCullagh & Nelder 1989).

There are many approaches that can be used to test whether the logistic regression model fits the data. A number of approaches use the idea of comparing the observed number of individuals with the expected if the fitted model is the valid one. The observed (O) and the (E) numbers are combined to form a χ^2 goodness-of-fit statistic. Large values of the test statistic show that the model is a poor fit as with small p values.

There are two different approaches of the goodness-of-fit test, the classical approach and the Lemershow and Hosmer (1982). Both tests require large sample sizes. The goodness-of-fit test statistic of the classis approach is given by:

$$\chi^2 = \sum 2O \left(\ln \frac{O}{E} \right) \tag{3.5.9}$$

The goodness-of-fit test statistic of the Lemeshow and Hosmer (1982) is given by:

$$\chi^2 = \sum \frac{(O - E)^2}{E} \tag{3.5.10}$$

The summary techniques to measure goodness-of-fit are:

(1) Two test statistic based on the Pearson and deviance residuals for the fitted model.

(2) The Hosmer-Lemeshow goodness-of-fit test.

(3) Classification tables comparing observed value of y with discrete classifications formed from the models' predicted values, $\hat{\pi}(x)$.

(4) The area under the receiver operating characteristic (ROC) curve.

(5) Several $pseudo-R^2$ measures have also been proposed as summary measures of the fit of a logistic regression model but should not be used in scientific papers. The standard Hosmer and Lemeshow goodness-of-fit test is applied to complex sample survey data.

The following is recommended in doing a goodness-of-fit evaluation:

(1) Applying the Archer and Lemeshow goodness-of-fit test and other available summary goodness of fit measures in the software.

(2) If the logistic regression programme of complex survey data in the chosen software system does not provide capabilities to generate goodness-of-fit measures, re-estimate the model using the sampling weights in the system's standard logistic regression programme.The weighted estimates of parameters and predicted probabilities will be identical and any serious lack of fit should be quantifiable.

3.5.3 Interpretation and inference

Wald χ^2 test and Confidence intervals are used in Logistic regression modelling to make inferences about the significance and importance of individual predictor variables. They also give information on the strength and uncertainty related to the estimated effects of individual predictor variables.

CI for the logistic regression model is computed as

$$CI_{1-\alpha}(B_j) = \hat{B}_j \pm t_{df,1-\frac{\alpha}{2}}.se(\hat{B}_j) \tag{3.5.11}$$

$\alpha = 0,05$ and the df are based on the design and this is a 95% confidence interval for the parameter which include the true population value of B_j. A logistic regression model with a single predictor, x_1 can give an estimate of the unadjusted odds ratio. This is given by:

$$\hat{\psi} = exp(\hat{B}_1) \tag{3.5.12}$$

If a logistic regression has multiple predictors then the result is an adjusted odds ratio:

$$\hat{\psi}_j|\hat{B}_{k\pm j} = exp(\hat{B}_j) \tag{3.5.13}$$

CI limits for adjusted odds ratio are given by:

$$CI_{\psi_j} = exp(\hat{B}_j \pm t_{df,1-\frac{\alpha}{2}}.se(\hat{B}_j)) \tag{3.5.14}$$

Categorical variables, ordinal variables and continuous variables all can be estimated using adjusted odds ratio and CI.

3.6 Data analysis

The results of the survey logistic regression analysis are presented in this section.

Table 3.1 reveals that the overall prevalence of HIV among women is 24.1%, with 48.6% of women surveyed reporting having experienced some form of either physical or sexual IPV in their most recent sexual relationship; 30.2% of women having experienced physical IPV, and 25.2% who were HIV positive also experiencing physical violence with their recent partners. Those who experienced sexual violence were 13.9% and 22.6% who experienced sexual violence were HIV positive. 6.5% of women reported having experienced both physical and sexual violence.

3.7 Application to Zimbabwe DHS data on HIV and intimate partner violence

Linear logistic models for binomially distributed variables were fitted to the Zimbabwe DHS data $(2005-2006)$. Principal Component Analysis was used to develop physical and sexual indices used to create the variables for logistic regression.

The variables

The response variable or dependent variable for each subject was the HIV status of the women under investigation. HIV status was coded by 1 if the woman was HIV positive, and 0 if HIV negative. The response variable is therefore the probability that the woman is HIV positive if exposed to either sexual or physical intimate partner violence. The other explanatory variables of the data are age, sexual violence, physical violence, media exposure, education, marital status, religion, wealth index, STI status and self efficacy.

Grouped data

The subjects are grouped by covariate class. y_i are the responses of the number of HIV positive women out of the n_i women in the i'th covariate class. The response probability $\frac{y_i}{n_i}$ is the estimated proportion of HIV positive subjects in the i'th covariate class. This is the estimated prevalence rate for the covariate class. Let $\hat{\pi}_i = \frac{y_i}{n_i}$, $i = 1, \cdots, k$, where k is the number of covariate classes.

3.8 Model specification for the Zimbabwe DHS data

3.8.1 Analysis

The aim of the analysis is to predict whether IPV influences the HIV status of a woman in Zimbabwe. Socio-demographic and socio-economic factors will also be considered. The analysis for the data begins by specifying the complex design features of the data sample in the Stata svyset command.

We specified the sampling unit as cluster number $v001$, strata as the type of residence $v025$, and the sampling weights as the lab number $HIV02$. There are 2 strata and 398 sampling units and 396 design based degrees of freedom. The introduction of clustering increases all the standard errors for the fitted coefficients in the model. Logistic regression models were used to estimate the Odds Ratio (OR) and Confidence Intervals (CI) of the IPV outcomes. Results were presented using 95% CI and statistical precision was ascertained using two-tailed Wald tests.

We estimated unadjusted bivariate associations between reported physical violence, sexual violence, age, religion, marital status, education, wealth index, media and STI with HIV. This was followed by re-estimating the adjusted association including the other covariates which were believed to have affected the women reporting IPV and being HIV positive. A comparison was made of women reporting either high physical or sexual violence to those reporting low physical or sexual violence. A logistic/ multinomial model was formed to compare women reporting both physical and sexual violence. Interactions of different variables was also compared to see which made the best fit model. We also looked at the effects of married women and separated women to determine the IPV and HIV relationship in these populations.

Exploring the data

The model building starts by examining bivariate associations of HIV with each of the potential predictor variables. The predictors are categorical variables, the bivariate relationship of each predictor with HIV is analysed in stata by using the svy: logistic $HIV03$ and predictor command and requesting odds ratio of the analysis. We get the odds ratios of the predictors and the CI and see if the predictors have significant bivariate associations.
Based on these tests the physical violence, sexual violence, region, type of place of residence, partners' education level are not significant, and thus do not have an association with HIV. The predictors which are significant are age, education, marital status, STI, religion and wealth index.

3.8.2 Results for the best variables to fit in the model

Age is presented as seven grouped categories ranging from $15-49$.The range with age $15-19$ is the reference. The unadjusted odds ratio for HIV infection was strongest when comparing women in the age group $30-39$, that is an odds ratio ranging from $3.37-3.56$, and was statistically significant $(p=0.00)$.Women who are $15-24$ and $40-49$ are at a lower risk of getting HIV. $p<0.1$ and confidence intervals show that the test is significant therefore age is associated with a woman getting HIV.Thus, middle-aged respondents have significantly higher odds of getting HIV relative to younger respondents and older respondents.
Relative to women who are married, respondents who were separated had a high odds ratio of 3.60.This is 3 times higher than women who are married or living with a partner. The CI was $2.93, 4.40$ and $p=0.00$ thus making marital status statistically significant with HIV. The presence of an STI had an influence of HIV status on women three times greater than that of women without STI, with an odds ratio of 3.20, $p=0.00$ and CI of $2.21, 4.64$ and so was statistically significant. Women who experienced high physical violence were more prone to HIV infection as compared to those who experienced sexual violence. The unadjusted odds ratio for women with high physical violence was (OR $1.09,\ 95\%$ CI $0.92,\ 1.28$) and $p=0.32$ and was not statistically significant.
Sexual violence had a lower odds outcome as compared to physical violence 0.91.The CI was $(0.72, 1.14)$ and $p=0.41$ and was not statistically significant implying there is no association of sexual violence and HIV. The unadjusted odds ratio for HIV was stronger when comparing women who had a primary and secondary education

as compared to those who had a higher educational level. The reference was women who had no education. The test was not significant since the $p > 0.05$ for all the levels of education. Middle class women and those richer on the wealth index had a stronger odds ratio and the test was statistically significant since $p = 0.05$. The odds ratio of those who attended the Apostolic church was almost twice that of those who followed a traditional religion $p = 0.04$ making the test statistically significant. Further, media did not affect the odds of outcome for the OR was 1.00. The results presented in the table of the analysis of test of association of the predictor variables with HIV show that the the following predictors have significant bivariate associations with HIV; age, religion, marital status, education, wealth index and STI. These variables generate significant chi-square and Rao-Scott F-test statistics with the covariates age, religion, marital status, wealth index and STI. Education is "marginally" significant and so will be included in the model. The table also shows adjustments for odds ratios, CI, chi-square and Rao-Scott F-test of associations for covariates and this confirms the variables which are strong and need to be included in the model.

Since our interest is IPV and HIV in generating the best fit model, we kept physical violence and sexual violence fixed and thus included all the other variables which were associated with HIV.

3.8.3 Model estimation

The Hosmer-Lemeshow (2000) model building suggests that the initial multivariate model can be formed with all eight predictors. The following logistic model was fitted:

$$g(\pi_i) = \log\left[\frac{\pi_i}{1-\pi_i}\right]. \tag{3.8.1}$$

$$= \beta_0 + \beta_1 \text{phy} + \beta_2 \text{sex} + \beta_3 \text{age} + \beta_4 \text{mstatus} + \beta_5 \text{rel} + \beta_6 \text{ edu} + \beta_7 \text{STI} + \beta_8 \text{wi}. \tag{3.8.2}$$

Stata was used to generate the estimated logistic regression model for the HIV outcome.The output in the table below was generated by using the svy: logit command.

3.8.4 Model evaluation

The adjusted Wald tests in Stata for the age, religion, education, and wealth index categorical predictors in the model, were generated using the test command. Note that we did not do a multi parameter Wald test for the physical, sexual, STI and marital status variables because they are represented by single indicator variables in the regression model.The overall Wald test for each predictor is equivalent to the t-test reported for the single estimated parameter for that predictor(Heeringa et al. 2010).The results of table 3.1 show the design-adjusted F-versions of the Wald test statistic and associated p-values. Two of the four design-adjusted Wald tests are significant at the 0.1 level. The exception is the Wald tests for education $F(3, 394) = 1.93, p = 0.12$ and religion $F(2, 395) = 1.82, p = 0.16$ which suggests that the parameters associated with education and religion in this logistic regression model are not significantly different from zero. Education and religion are not important predictors of getting HIV when adjusting for the relationships of the other predictor variables with the outcome. Since education and religion are marginally significant we will retain them in the model.

3.8.5 Model interpretation/inference

Based on the estimated logistic regression model and the design-adjusted Wald Tests for the parameters associated with the categorical predictors, each of the covariates in the model has a significant or marginally significant relationship with the probability of HIV after adjusting for the other variables. The resulting design adjusted F-statistic reported in the Stata results is $F_{(16,381)} = 14.75$, with a p-value of 0. As the model fits the data well we do not reject the null hypothesis, and therefore accept the fit of this model.

We considered testing some scientifically relevant two-way interactions between the predictor variables. We supposed the two-way interaction of physical violence and sexual violence with the other six covariates with no interaction. After fitting the model in Stata we found that the two-way interactions were not making a significant contribution to the fit of the model to the Zimbabwe DHS data. The design-adjusted Wald tests on the categorical parameters suggested that the parameter for religion was not significant but the rest of the parameters were significant. Further, an interaction of marital status with physical and sexual violence suggested that these two-way interactions were actually not making a significant contribution to the fit of the model.

3.9 Summary

After using the indices developed for IPV, that is physical and sexual violence, to see if they are associated with HIV, we concluded that they were not significantly associated with HIV. One would have assumed that women who have experienced sexual or physical violence were more likely to have HIV and sexually transmitted diseases. The analysis, using survey logistic regression, only revealed an association with STIs. In Chapter 4 we will analyse the covariates which are associated with IPV using a multinomial logit regression model.

Table 3.1: HIV Prevalence in Zimbabwe by sample variance by values of independent variables

Variable	N	PLHIV	Perc PLHIV
Complete case analytical sample	4 065	980	24.1
Age			
15-19	250	27	10.8
20-24	885	139	15.7
25-29	895	234	26.1
30-34	781	250	32.0
35-39	514	163	31.7
40-44	411	105	25.5
45-49	329	62	18.8
Marital status			
Currently	3 388	699	19.7
Formerly	677	311	45.9
Urbanity			
Urban	1 109	287	25.9
Rural	2 956	693	23.4
Wealth quantiles			
Poorest	924	199	21.5
2nd poorest	866	194	22.4
Middle	755	194	25.7
2nd Richest	908	256	28.2
Richest	612	137	22.4
Education			
None	212	43	20.3
Primary	1 593	382	24.0
Secondary and above	2 260	555	24.6

	N	PLHIV	Perc PLHIV
Occupation			
Not employed	2 195	512	23.3
Manual	219	52	23.7
Agricultural	830	186	22.4
Non manual, non-agricultural	821	230	28.0
Religion			
Christian	3 507	845	24.1
Muslim	26	6	23.1
Hindu			
Other/none	532	129	24.2
Lifetime no. of partners			
Zero or one	2 660	466	17.5
Two or more	1 405	514	36.6
Intimate partner violence			
No physical or sexual violence	1 732	295	17.0
Any physical or sexual violence	1 636	308	18.8
Any physical violence	1 226	309	25.2
Any sexual violence	566	128	22.6
Physical and sexual violence	359	88	24.5

Table 3.2: HIV Results

HIV03	
Number of Strata	2
Number of Observations	5289
Number of PSUs	398
Population size	3612458
Design df	396

Table 3.3: HIV Prevalence in Zimbabwe by type of intimate partner violence exposure

	Sample size	**HIV-positive**	%	**Exposed**	%	% **exposed HIV-positive**
Total Sample	4 065	980	24.1			
Any Physical IPV				1 226	30.2	25.2
Physical or sexual IPV				1 636	48.6	18.8
Any Sexual IPV				566	13.9	22.6
Physical and sexual IPV				359	8.8	24.5

Table 3.4: Estimates of unadjusted odds ratios [95% confidence intervals], χ^2 statistic, p-value, Rao-Scott F-test, p-value for the association between HIV prevalence and various variables (*Analysis based on the Zimbabwe DHS* $2005-2006$ *data*).

Predictor	Category	Unadajusted OR (95% CI)	Chi2 Statistic	Pr	Rao-Scott F-test	p-value
Physical violence	low	1.00				
	High	1.09(0.92-1.28)	3.78	0.052	F(1,396)=0.99	0.32
Sexual violence	low	1.00				
	High	0.91(0.72-1.14)	0.19	0.67	F(1,396)=0.70	0.4
Age(v013)	1	1.00				
	20-24	1.41(0.92-2.16)				
	25-29	2.64(1.70-4.11)				
	30-34	3.37(2.14-5.32)				
	35-39	3.56(2.24-5.64)				
	40-44	2.50(1.52-4.09)				
	45-49	1.92(1.14-3.22)	119.05	0	F(5.8,2306.93)=13.04	0
Religion(V13)	Traditional	1.00				
	Rom,Pen,Ap	1.76(1.02-3.04)				
	Other	1.81(1.04-3.15)	8.13	0.02	F(1.97,779.98)=2.05	0.13
Marital status(v501)	married,living together	1.00				
	separated	3.60(2.93-4.40)	275.4	0	F(1,396)=166.73	0
Education(v106)	none	1.00				
	primary	1.28(0.85-1.94)				
	secondary	1.32(0.87-1.98)				
	higher	0.73(0.35-1.50)	4.13	0.25	F(2.97,1175.94)=1.77	0.15
Wealth index(v190)	poorest	1.00				
	poorer	1.10(0.85-1.41)				
	middle	1.13(0.85-1.51)				
	richer	1.31(1.01-1.72)				
	richest	0.89(0.67-1.19)	18.25	0	F(3.83,1518.29)=2.18	0.07
Media	low	1.00				
	medium	1.03(0.83-1.29)				
	high	1.000.82-1.22)				
STI(v763a)	No	1.00				
	Yes	3.20(2.21-4.64)	78.41	0	F(1,396)=41.63	0

Table 3.5: Estimates of adjusted odds ratios and [95% confidence intervals] for the association between HIV prevalence and various variables (*Analysis based on the Zimbabwe DHS* $2005-2006$ *data*).

Predictor	Category	Adjusted OR(95%CI)
Physical	Low	1.00
	High	1.07(0.89-1.29)
Sexual	Low	1.00
	High	0.81(0.62-1.07)
Age(v013)	15-19	1.00
	20-24	1.29(0.83-2.02)
	25-29	2.61(1.64-4.15)
	30-34	3.23(2.01-5.19)
	35-39	2.96(1.87-4.71)
	40-44	2.08(1.26-3.45)
	45-49	1.25(0.74-2.10)
Education(v106)	none	1.00
	Primary	1.36(0.88-2.10)
	Secondary	1.48(0.93-2.35)
	Higher	0.82(0.38-1.77)
Religion	Traditional	1.00
	Rom,Pen,Ap	1.62(0.88-2.97)
	Other	1.74(0.96-3.16)
STI (v763a)	no	1.00
	Yes	2.70(1.86-3.93)
Wealth Index(v190)	Poorest	1.00
	Poorer	1.05(0.80-1.38)
	Middle	1.05(0.78-1.42)
	Richer	1.13(0.85-1.50)
	Richest	0.75(0.54-1.03)
Mstatus	Married	1.00
	separated	3.66(2.96-4.52)

Table 3.6: Estimated logistic regression model for the HIV outcome (Output generated by using the svy: logit command) (*Analysis based on the Zimbabwe DHS* $2005-2006$ *data*).

Predictor	Category	Coef.	Std. Err.	p-value	95% CI
Physical	low	0			
	high	0.07	0.09	0.49	(-0.12-0.25)
Sexual	low	0			
	high	-0.21	0.14	0.14	(-0.49-0.07)
Marital Status	Married/Living together	0			
	Separated	1.30	0.11	0.00	(1.08-1.51)
Age	15-19	0			
	20-24	0.26	0.23	0.26	(-0.19-0.71)
	25-29	0.96	0.24	0.00	(0.49-1.42)
	30-34	1.17	0.24	0.00	(0.70-1.65)
	35-39	1.09	0.24	0.00	(0.62-1.55)
	40-44	0.73	0.26	0.00	(0.23-1.24)
	45-49	0.22	0.26	0.40	(-0.30-0.74)
Religion	Traditional	0			
	Rom,Pen,Ap	0.48	0.31	0.12	(-0.13-1.09)
	Other	0.55	0.30	0.07	(-0.04-1.15)
Education	None	0			
	Prim	0.31	0.22	0.16	(-0.12-0.74)
	Sec	0.39	0.23	0.09	(-0.07-0.85)
	Higher	-0.20	0.39	0.61	(-0.97-0.57)
STI	No	0			
	Yes	0.99	0.19	0.00	(0.62-1.37)
Wealth Index	Poorest	0			
	Poorer	0.05	0.14	0.73	(-0.23-0.32)
	Middle	0.05	0.15	0.73	(-0.25-0.35)
	Richer	0.12	0.14	0.39	(-0.16-0.40)
	Richest	-0.29	0.16	0.08	(-0.61-0.03)
Intercept	Constant	-3.02	0.39	0.00	(-3.79–2.26)

Table 3.7: Design-adjusted Wald tests for the parameters associated with the categorical predictors in the initial HIV logistic regression model (*Analysis based on the Zimbabwe DHS* $2005-2006$ *data*).

Categorical Predictor	F-Test Statistic	P-value
Age	$F_{(6,391)} = 12.78$	$0 < 0.1$
Religion	$F_{(2,395)} = 1.82$	0.16
Education	$F_{(3,394)} = 1.93$	0.12
Wealth Index	$F_{(4,393)} = 2.24$	$0.06 < 0.1$

Table 3.8: Estimated OR and CI for the HIV outcome, including interactions of physical and sexual variables (*Analysis based on the Zimbabwe DHS* $2005-2006$ *data*).

Predictor	Category /Interaction	OR(95%CI)
Physical	Low	1.00
	High	1.08(0.90-1.30)
Sexual	Low	1.00
	High	0.76(0.52-1.10)
Interaction	phyXsex(high)	1.20(0.74-1.92)
Age	15-19	1.00
	20-24	1.38(0.89-2.15)
	25-29	2.81(1.83-4.32)
	30-34	3.55 (2.30-5.46)
	35-39	3.22(2.06-5.04)
	40-44	2.33(1.46-3.73)
	45-49	1.30(0.78-2.16)
Religion	Traditional	1.00
	Rom,Pen,Ap	1.60(0.94-2.72)
	Other	1.66(0.98-2.82)
Marital Status	Married	1.00
	Separated	3.77(3.15-4.52)
Education	None	1.00
	Primary	1.26(0.86-1.84)
	Secondary	1.29(0.86-1.93)
	Higher	0.76(0.39-1.46)
STI	None	1.00
	Yes	2.93(2.14-4.01)
Wealth Index	Poorest	1.00
	Poorer	1.03(0.82-1.31)
	Middle	1.19(0.93-1.51)
	Richer	1.33(1.05-1.68)
	Richest	0.89(0.67-1.19)

Chapter 4

The multinomial logit regression model

Since there was no relationship between a woman experiencing IPV and her getting an HIV test, Chapter 4 is used to analyse the relationship of IPV and socio-demagraphic and socio economic variables using the Multinomial Logit Regression Model.

4.1 Introduction

If we have survey response variables with 2 or more distinct or unordered categories we use the multinomial logistic regression model (MLRM). Thus MLRM is an extension of the binary logistic model which is relevant for survey variables with nominal response categories. Examples of questions with nominal responses are: the places one goes to when one is sick. Here the response could be hospital, doctor,clinic etc. Another question is one which involves the current employment status of a person on a particular day, with responses being employed, unemployed, retired, self-employed etc (Heeringa et al. 2010).

In a survey when we have numerous categorical detailed responses which are coded, the multiple categories can be recoded into a smaller useful set of nominal groupings. The MLRM is good for multivariate analysis of dependent variables that can be recoded. MLRM can be used in application to survey variables of Likert-type scales. Examples of a Likert-type scale are; ($1 =$ strongly agree to $5 =$ strongly disagree), ordered categorical

response scales for example self (rated health status $1 = excellent$ to $5 = poor$). Another efficient method for modelling ordinal data is the cumulative logit regression model (Heeringa et al. 2010, Dolgun 2012).

4.1.1 The model

The model can be fitted by simultaneously estimating binary logistic regression models for all possible comparisons of the outcome category with a baseline category. MLRM is sometimes called the "baseline" category logistic regression model. Taking Labour force as an example, we can recode variables for labour force status with 3 nominal categories: 1 =Employed; 2 =Unemployed; 3 =in the labour force;(NLF). The "trinomial" dependent variable of labour force can be fitted into, 2 generalized logits given by (Heeringa et al. 2010);

$$logit(\pi(\text{"UN"}|X)) = \text{logit}(\pi_2) = \ln\left[\frac{\pi(y=2|x)}{\pi(y=1|x)}\right] = \beta_{2:0} + \beta_{2:1}X_1 + \cdots + \beta_{2:p}X_p$$

and

$$logit(\pi(\text{"NLF"}|X)) = \text{logit}(\pi_3) = \ln\left[\frac{\pi(y=3|x)}{\pi(y=1|x)}\right] = \beta_{3:0} + \beta_{3:1}X_1 + \cdots + \beta_{3:p}X_p..$$

Note: It is not possible to estimate the multinomial logit regression model as a series of binary logistic regression models that consider only the response data for 2 categories. Software systems like stata that support analysis of complex survey data allows for the simultaneous estimation of the multinomial logit regression model. The general formula of the MLRM to the kth logit is given by;

$$logit(\pi_k(x)) = \ln\left[\frac{\pi(Y=K|\mathbf{x})}{\pi(Y=0|\mathbf{x})}\right] = \beta_{0k} + \mathbf{x}\prime\boldsymbol{\beta}_{\mathbf{k}},\ \mathbf{k} = \mathbf{1}, \mathbf{2}, \cdots, \mathbf{K}. \tag{4.1.1}$$

In equation 4.1.1, β_{0k} is the constant term and $\boldsymbol{\beta}_k$ is the regression coefficient vector in the kth logit. MLRM has an advantage that it models the odds of each category relative to a baseline as a function of covariates (Heeringa et al. 2010, Dolgun 2012).

4.2 Stages of multinomial logistic regression model

4.2.1 Specification stage

This stage is the same as that for specifying the logistic regression model described in the previous chapter. The 2 aspects which are different from logistic regression are:

(i) Choice of the baseline category. The analyst can choose which category of the k-categories they prefer to be the baseline. This choice does not affect the fit of the MLRM or overall test for the significance of the parameters associated with predictors in the model. The interpretation of the parameters will depend on the baseline category. When using stata the lowest numbered category is chosen as the baseline category to estimate the logit and corresponding odds ratio. The base outcome can be used to choose a different category as the baseline to represent the value of the desired baseline category. The most common category (or model) of the nominal dependent variable is usually used when a choice of baseline category is not clear.

(ii) Parsimony. The $K-1$ logits include the design vector of covariates, $\mathbf{x} = \mathbf{1}, \mathbf{x_1}, \cdots, \mathbf{x_p}$. Each of the estimated logit will have $\boldsymbol{\beta}_k = \beta_{k:0}, \beta_{k:1}, \cdots, \beta_{k:p}$ parameters. Thus the total number of parameter estimates will be $(K-1) \times (p+1)$. The best model should reduce the number of predictors that are either not significant or have a high correlation with other significant covariates. The test used to determine the best predictors across the $K-1$ logit functions is the design-adjusted multiparameter Wald tests (Heeringa et al. 2010).

4.2.2 Estimation stage

The pseudo-likelihood function is used to estimate parameters for the data in MLRM and is based on the multinomial distribution unlike the binomial distribution in the simple binary logit model. The difference of the 2 models is that the number of parameters and standard errors to be estimated are from $(p+1)$ in the logistic model and $(K-1) \times (p-1)$ for the MLRM.

To estimate the variance-covariance matrix of the MLRM parameters, a multinomial version of Binder's (1983)

Taylor series linearisation (TSL) estimator is used by current software. This takes the sandwich form:

$$\hat{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{J}^{-1})var[S(\hat{\boldsymbol{\beta}})](\mathbf{J}^{-1}).$$

The matrices **J** and $var[S(\hat{\boldsymbol{\beta}})]$ are $(K-1)X(p+1)$ symmetric matrices, which reflect the full dimension of the parameter vector for the multinomial logit regression model (Heeringa et al. 2010).

4.2.3 Evaluation stage

The stage starts with the Wald Tests of hypotheses concerning the model parameters. Since the parameter estimates are $(K-1)X(p+1)$, the number of possible hypothesis tests are vast. Standard t-tests are used for single parameters and Wald tests for multiple parameters to evaluate the significance of the covariate effects in individual logits that is $H_0 : \beta_{k:j} = 0$ and across all estimated logits that is $H_0 : \beta_{2:j} = \cdots = \beta_{k:j} = 0$ (Heeringa et al. 2010).

4.2.4 Interpretation stage

If we exponentiate the parameter estimate result in an adjusted odds ratio, which corresponds to the multiplicative impact of a one unit increase in the predictor variable, x_j on the odds that the response is equal to K relative to the odds of a response in the baseline category. This is given by:

$$\hat{\psi}_{k:j} = exp(\hat{\beta}k : j).$$

The CI is given by:

$$CI(\hat{\psi}_{k:j}) = exp[\hat{\beta}_{k:j} \pm t_{df,1-\frac{\alpha}{2}}.se(\hat{\beta}_{k:j})],$$

where $\hat{\beta}_{k:j}$ = the parameter estimate corresponding to the predictor in logit equation K. In case we are interested in the impact of a one unit increase in predictor x_j, on the odds that it belongs to one of 2 non baseline categories, then the following odds ratio estimates the multiplicative effect of a one unit change in x_j on the odds of being in category K compared with category K';

$$\hat{\psi}_{k,k':j} = exp(\hat{\beta}_{k:j} - \hat{\beta}_{k':j}),$$

and the CI is given by;

$$CI(\hat{\psi}_{k,k':j}) = exp[(\hat{\beta}_{k:j} - \hat{\beta}_{k':j}) \pm t_{df,1-\frac{\alpha}{2}}.se(\hat{\beta}_{k:j} - \hat{\beta}_{k':j})],$$

where $\hat{\beta}_{k:j}, \hat{\beta}_{k'j} =$ the parameter estimates corresponding to predictor j in logit equations k and k'.

4.3 Fitting a multinomial logistic regression model to Zimbabwe survey data

In this analysis we are fitting the multinomial logistic regression model to Zimbabwe survey data response variable intimate partner violence, which includes physical and sexual violence variables. We are looking at the socio-demographic and socio-economic factors which affect IPV. The response variable takes on 4 values/categories which are $102 =$ both low physical and sexual violence, $103 =$ high physical and low sexual violence, $202 =$ high sexual and low physical violence and $203 =$ both high sexual and physical violence.

The predictor variables considered in this analysis are occupation, education, age, region, type of residence, marital status, wealth index, religion, media and decision making. The results of the preliminary analysis of the IPV variable and each of the ten categorical predictors considered in the initial model are given in the table below. The results of Wald F-test of the preliminary analysis of the bivariate associations of the IPV categorical data and each of the ten categorical predictors suggest that nine of the predictors have significant bivariate associations with IPV, with p values less than 0.1. Decision making has a weak association and will be dropped from the model. In order to find out if the remaining nine predictors would remain significant when controlling for other predictors, we fitted a multinomial regression logit model taking the complex design of the data. We identified the appropriate sampling weight using svyset. The following command was used to fit the model; svy: mlogit ipv i.v013 i.v106 i.region i.religion i.v190 i.ocupation i.mstatus i.v102 i.media and the estimated odds ratio and CI were fitted by the command; svy: mlogit,rrr. The lowest category (low physical and low sexual violence) was the default baseline category for the multinomial logit regression model.

Table 4.1: Initial bivariate design-based tests assessing potential predictors of IPV for the Zimbabwe data (*Analysis based on the Zimbabwe DHS* $2005-2006$ *data*).

Categorical Predictor	F-test Statistic	p-value
Age	$F_{(16.73,6625.25)}=$ 1.95	0.01
Education	$F_{(8.55,3385.38)}=$ 2.71	0
Region	$F_{(23.40,9265.20)}=$ 3.83	0
Religion	$F_{(5.57,2206.72)}=$ 6.2347	0
Wealth Index	$F_{(11.55,4574.23)}=$ 3.9737	0
Occupation	$F_{(10.84,4293.96)}=$ 3.45	0
Decision Making	$F_{(5.80,2255.98)}=$ 1.71	0.11
Marital Status	$F_{(2.96,1172.92)}=$ 4.69	0
Type of Residence	$F_{(2.95,1167.36)}=$ 8.20	0
Media	$F_{(5.79,2292.24)}=$ 5.84	0

4.3.1 Interpretation of results

Table 4.3 provides the stata output for the estimated model, including the coefficient estimates, linearized standard errors, p-values of each of the three generalized logits. Table 4.2 provides the odds ratio and 95% CI for the odds ratio. The fitted model was evaluated using the multiparameter Wald tests for the overall significance of each of the predictors. The Wald tests were testing the null hypothesis that all parameters associated with each individual variable in the 3 logits were not significantly different from zero.

Table 4.6 presents the Wald F-tests results and the p-values associated with the tests of individual effects. The results show that eight of the covariates are strongly significant determinants of the relative odds of IPV. The predictor type of residence a woman stays in does not have a significant effect on IPV.

With referrence to the t-tests of each individual logit, with age the odds of being physically abused is high and sexually abused low, as compared to both high sexual and physical abuse as well as high sexual and low physical abuse. Higher education is significantly associated with all categories of IPV. Region is highly significant with

the first two logits except for high sexual and low physical violence. Media and Wealth index have a significant effect on both high physical and sexual violence as well as high physical and low sexual violence. Marital status is significant with the last two categories, but is not significant with high physical violence and low sexual violence. The type of residence a woman stays in is not significant with any category of IPV.

We considered another test for the hypothesis that the three educational parameters in the three logits were equivalent that is:

(i) high physical, low sexual = high physical, high sexual

(ii) high physical, low sexual = high sexual, low physical

(iii) high physical, high sexual = high sexual, low physical

The F-test statistics and associated p-values for these Wald tests revealed that (i) we rejected the null hypothesis and concluded that the pairs were not equal to each other since $F_{3,394} = 4,08$ and $p = 0.01$. The test for (ii) had the values for F and p as 0.76 and 0.52 respectively. For this test we failed to reject the null hypothesis and concluded that the pairs of coefficients were equal to each other. The Wald test for (iii) suggested that the pairs of coefficients were equal to each other as F and p were 1.97 and 0.12 respectively.

The magnitude and direction of significance effects of the predictors and IPV was also interpreted using the odds ratios and CI for each logit. The odds ratios of being high physical, high sexual violence are significantly high for age groups, with the age group of $35-39$ having no association with any of the physical violence. The reference category for age is $15-19$. The odds ratio of education is high, with high physical, low sexual violence. In this category the odds ratio of being highly educated is five times that of no education which is 5.72.

The odds ratio of region increases from 1.55 to 2.05 on the logit high physical violence and high sexual violence. Women in Mashonaland experience high IPV as compared to Manicaland which is the reference category. The odds ratio of occupation range from 0.87 to 2.83 with unskilled women facing high physical, low sexual violence. The reference category for occupation is no work.

Table 4.2: Estimates of adjusted odds ratios [95% confidence intervals], p-value for the IPV outcome (*Analysis based on the Zimbabwe DHS* $2005-2006$ *data*).

IPV		low physical,low sexual	(base outcome)				
		high physical, low sexual		high physical, high sexual		high sexual, low physical	
Predictor	Category	RRR(95%CI)	p-value	RRR(95%CI)	p-value	RRR(95%CI)	p-value
Age	15-19	1.00		1.00			
	20-24	0.42(0.22,0.81)	0.01	1.57(1.05,2.35)	0.03	1.03(0.64,1.66)	0.91
	25-29	0.45(0.23,0.87)	0.02	1.29(0.85,1.94)	0.23	0.75(0.46,1.23)	0.26
	30-34	0.62(0.33,1.17)	0.14	1.21(0.80,1.85)	0.37	0.68(0.40,1.17)	0.16
	35-39	0.38(0.17,0.85)	0.02	1.00(0.66,1.52)	0.99	0.55(0.32,0.96)	0.04
	40-44	0.61(0.30,1.23)	0.17	1.30(0.79,2.14)	0.29	0.54(0.27,1.05)	0.07
	45-49	0.53(0.24,1.18)	0.12	1.38(0.84,2.27)	0.20	0.67(0.34,1.31)	0.24
Education	no education	1.00		1.00			
	primary	1.96(0.84,4.53)	0.12	1.09(0.67,1.77)	0.72	1.07(0.61,1.87)	0.82
	secondary	1.73(0.71,4.19)	0.23	1.01(0.60,1.70)	0.96	0.98(0.54,1.78)	0.95
	higher	5.72(1.54,21.28)	0.01	0.37(0.13,1.07)	0.07	1.80(0.54,5.95)	0.34
Region	Manicaland	1.00		1.00			
	Mashonaland	0.57(0.37,0.87)	0.01	2.05(1.48,2.83)	0.00	1.22(0.83,1.79)	0.31
	Matebeleland	0.40(0.21,0.76)	0.01	1.55(1.09,2.22)	0.02	0.46(0.28,0.77)	0.00
	Midlands	0.63(0.40,1.00)	0.05	1.55(1.08,2.21)	0.02	1.00(0.66,1.53)	0.99
Religion	Traditional	1.00		1.00			
	Attend a church	0.85(0.31,2.30)	0.74	0.64(0.38,1.07)	0.09	0.39(0.20,0.77)	0.01
	No religion	0.91(0.34,2.45)	0.85	0.81(0.48,1.35)	0.41	0.51(0.27,0.96)	0.04
Wealth Index	Poorest	1.00		1.00			
	poorer	1.02(0.63,1.66)	0.92	0.94(0.72,1.23)	0.65	0.96(0.68,1.37)	0.84
	middle	0.73(0.42,1.26)	0.26	0.70(0.51,0.95)	0.02	0.82(0.56,1.19)	0.29
	richer	0.54(0.27,1.05)	0.07	0.60(0.40,0.91)	0.02	0.86(0.52,1.44)	0.57
	richest	0.31(0.11,0.84)	0.02	0.37(0.21,0.65)	0.00	0.52(0.22,1.26)	0.15
Occupation	No work	1.00		1.00			
	professional	0.87(0.25,2.97)	0.82	0.96(0.47,1.98)	0.92	0.85(0.29,2.46)	0.76
	skilled	2.13(1.45,3.14)	0.00	1.11(0.90,1.36)	0.33	1.50(1.16,1.95)	0.00
	unskilled	2.83(1.60,4.99)	0.00	1.09(0.77,1.56)	0.63	1.39(0.86,2.26)	0.18
	don't know	0.00(0.00,0.00)	0.00	1.08(0.28,4.22)	0.91	2.77(0.58,13.29)	0.20
Marital Status	Married	1.00		1.00			
	Separated	0.93(0.57,1.52)	0.78	1.37(1.08,1.74)	0.01	1.83(1.37,2.46)	0.00
Type of Residence	Urban	1.00		1.00			
	Rural	0.92(0.44,1.93)	0.83	0.83(0.56,1.22)	0.34	1.42(0.77,2.64)	0.26
Media	low exposure	1.00		1.00			
	medium exposure	1.46(0.98,2.16)	0.06	1.61(1.24,2.08)	0.00	1.34(0.99,1.81)	0.06
	high exposure	1.04(0.58,1.88)	0.89	1.39(1.01,1.91)	0.04	1.12(0.69,1.82)	0.64

Table 4.3: Estimated multinomial logit regression model for IPV, adjusted Wald F-tests for all parameters: $F_{78,319} = 66.71$, $p = 0$ (*Analysis based on the Zimbabwe DHS* $2005 - 2006$ *data*).

IPV		low physical vs low sexual	(base outcome)		
		logit 2: high physical vs low sexual			
Predictor	Category	Coef.	Std Err	t	P-value
Age	15-19	Reference			
	20-24	-0.87	0.34	-2.58	0.01
	25-29	-0.80	0.34	-2.37	0.02
	30-34	-0.48	0.32	-1.49	0.14
	35-39	-0.96	0.41	-2.37	0.02
	40-44	-0.50	0.36	-1.39	0.17
	45-49	-0.63	0.40	-1.55	0.12
Education	no education	Reference			
	primary	0.67	0.43	1.57	0.12
	secondary	0.55	0.45	1.22	0.23
	higher	1.74	0.67	2.61	0.01
Region	Manicaland	Reference			
	Mashonaland	-0.57	0.22	-2.62	0.01
	Matebeleland	-0.92	0.32	-2.83	0.01
	Midlands	-0.46	0.23	-1.96	0.05
Religion	Traditional	Reference			
	Attend a church	-0.17	0.51	-0.33	0.74
	No religion	-0.10	0.51	-0.19	0.85
Wealth Index	poorest	Reference			
	poorer	0.02	0.25	0.1	0.92
	middle	-0.32	0.28	-1.14	0.26
	richer	-0.63	0.34	-1.83	0.07
	richest	-1.19	0.52	-2.3	0.02
Occupation	no work	Reference			
	professional	-0.14	0.62	-0.22	0.82
	skilled	0.76	0.20	3.84	0.00
	unskilled	1.04	0.29	3.6	0.00
	don't know	-19.57	0.44	-44.99	0.00
Marital Status	married	Reference			
	Separated	-0.07	0.25	-0.28	0.78
Type of Residence	Urban	Reference			
	Rural	-0.08	0.38	-0.22	0.83
Media	low exposure	Reference			
	medium exposure	0.38	0.20	1.88	0.06
	high exposure	0.04	0.30	0.14	0.89
	Constant	-2.05	0.73	-2.8	0.01

Table 4.4: Estimated multinomial logit regression model for IPV, adjusted Wald F-tests for all parameters:

		logit 3: high physical vs high sexual			
Predictor	Category	Coef.	Std Err	t	P-value
Age	15-19	Reference			
	20-24	0.45	0.21	2.18	0.03
	25-29	0.25	0.21	1.21	0.23
	30-34	0.19	0.21	0.9	0.37
	35-39	0.00	0.21	0.02	0.99
	40-44	0.27	0.25	1.05	0.29
	45-49	0.32	0.25	1.28	0.20
Education	no education	Reference			
	primary	0.09	0.25	0.36	0.72
	secondary	0.01	0.26	0.05	0.96
	higher	-0.99	0.54	-1.85	0.07
Region	Manicaland	Reference			
	Mashonaland	0.72	0.16	4.37	0.00
	Matebeleland	0.44	0.18	2.42	0.02
	Midlands	0.44	0.18	2.41	0.02
Religion	Traditional	Reference			
	Attend a church	-0.45	0.26	-1.72	0.09
	No religion	-0.21	0.26	-0.82	0.41
Wealth Index	poorest	Reference			
	poorer	-0.06	0.13	-0.45	0.65
	middle	-0.36	0.16	-2.26	0.02
	richer	-0.50	0.21	-2.43	0.02
	richest	-0.99	0.29	-3.47	0.00
Occupation	no work	Reference			
	professional	-0.04	0.37	-0.1	0.92
	skilled	0.10	0.10	0.98	0.33
	unskilled	0.09	0.18	0.49	0.63
	don't know	0.08	0.69	0.11	0.91
Marital Status	married	Reference			
	Separated	0.32	0.12	2.57	0.01
Type of Residence	urban	Reference			
	Rural	-0.19	0.20	-0.97	0.34
Media	low exposure	Reference			
	medium exposure	0.48	0.13	3.64	0.00
	high exposure	0.33	0.16	2.05	0.04
	Constant	-1.45	0.45	-3.22	0.00

Table 4.5: Estimated multinomial logit regression model for IPV, adjusted Wald F-tests for all parameters:

		logit 4: high sexual vs low physical			
Predictor	Category	Coef.	Std Err	t	P-value
Age	15-19	Reference			
	20-24	0.03	0.24	0.12	0.91
	25-29	-0.28	0.25	-1.13	0.26
	30-34	-0.38	0.27	-1.4	0.16
	35-39	-0.59	0.28	-2.12	0.04
	40-44	-0.63	0.34	-1.83	0.07
	45-49	-0.40	0.34	-1.17	0.24
Education	no education	Reference			
	primary	0.07	0.28	0.23	0.82
	secondary	-0.02	0.30	-0.06	0.95
	higher	0.59	0.61	0.96	0.34
Region	Manicaland	Reference			
	Mashonaland	0.20	0.20	1.01	0.31
	Matebeleland	-0.78	0.26	-2.99	0.00
	Midlands	0.00	0.22	0.01	0.99
Religion	Traditional	Reference			
	Attend a church	-0.93	0.35	-2.71	0.01
	No religion	-0.68	0.32	-2.08	0.04
Wealth Index	poorest	Reference			
	poorer	-0.04	0.18	-0.2	0.84
	middle	-0.20	0.19	-1.05	0.29
	richer	-0.15	0.26	-0.57	0.57
	richest	-0.65	0.44	-1.45	0.15
Ocupation	no work	Reference			
	professional	-0.17	0.54	-0.3	0.76
	skilled	0.41	0.13	3.07	0.00
	unskilled	0.33	0.25	1.34	0.18
	don't know	1.02	0.80	1.28	0.20
Marital Status	Married	Reference			
	Separated	0.61	0.15	4.05	0.00
Type of Residence	Urban	Reference			
	Rural	0.35	0.31	1.12	0.26
Media	low exposure	Reference			
	medium exposure	0.29	0.15	1.88	0.06
	high exposure	0.12	0.25	0.47	0.64
	Constant	-1.44	0.58	-2.49	0.01

Table 4.6: Overall Wald tests for the predictors in the multinomial model for IPV (*Analysis based on the Zimbabwe DHS* $2005-2006$ *data*).

Categorical Predictor	F-test Statistic	p-value
Age	$F_{(18,379)} = 1.63$	0.05
Education	$F_{(9,388)} = 1.63$	0.1
Region	$F_{(9,388)} = 5.11$	0
Religion	$F_{(6,391)} = 2.18$	0.04
Wealth Index	$F_{(12,385)} = 2.09$	0.02
Occupation	$F_{(12,385)} = 301.10$	0
Marital Status	$F_{(3,394)} = 6.94$	0
Type of Residence	$F_{(3,394)} = 0.77$	0.51
Media	$F_{(6,391)} = 2.63$	0.02

4.4 Summary

Chapter 4 focused on the multinomial logit regression model using physical and sexual violence variables as categories to analyse the association of IPV and various covariates. The results revealed significant association between IPV and the explanatory variables (age, education, region, religion, wealth index, occupation, marital status, type of residence and media) The t-tests of the individual logits showed that some of the covariates were strongly associated with specific categories of IPV, for example age affects women who are highly physically abused and experiencing low sexual abuse as compared to women who experience both high physical and high sexual abuse. We concluded that not all covariates were associated significantly with all categories of IPV presented in this chapter. In the following chapter 5 we use the generalised linear mixed models to examine the association between IPV and HIV infection. This is because the GLMM gives us the opportunity to model the binary response variable HIV test result with covariates of interest and to take random effects of survey cluster into account.

Chapter 5

Generalised linear mixed models with random effects

Another statistical method, the generalised linear mixed models (GLMM) with random effects, is used to compare the results obtained in chapter 4 and to confirm that HIV and Intimate Partner Violence are not associated. In this chapter we will also decide which is the preferred model between the logistic regression model and GLMMs. Recommendations are also given in chapter 5.

5.1 Introduction

In order to analyse longitudinal and clustered data, statistical models with random effects are frequently used. This is because the responses of the same cluster are likely to be correlated. Therefore, we must take into account dependence by assessing the relationship of the responses Y with explanatory variables X. If the responses are Gausian, random effects are used to account for the dependence within a cluster. The models help to derive predicted values of the random effects (McCulloch & Neuhaus 2011, Zeger & Karim 1991). Non-normal data with random effects can be analysed using Generalized Linear Mixed Models (GLMM) (Bolker, Brooks, Clark, Geange, Poulsen, Stevens & White 2009). Random effects serve as a purpose of quantifying the variation among

units. The two statistical frameworks which are used in GLMM are a combination of linear mixed models with random effects and generalised linear models which use non-normal data with link functions and exponential family (Bolker et al. 2009). Therefore, to analyse GLMM one has to specify the distribution, link function and structure of the random effects.

5.2 A generalised linear mixed model

In this section we consider a generalised linear mixed model (GLMM) for clustered data with random cluster specific terms, $\mathbf{b}_i$. We define a generalised linear model with random effects as follows: Let $y = [y_1, \cdots, y_n]'$ be a vector of *n* observations which can be written as $y = \mu + e$ where *e* is a vector of random errors with zero expectation and covariance matrix V (McCulloch & Neuhaus 2011, Schall 1991). Furthermore, let Y_{it} represent the t-th observation $t = 1, \cdots, n_i$ within cluster i $(i = 1, \cdots, m)$. Given the condition of random effects then we assume that Y_{it} are independent and that $g(.)$ is a monotonic function, the link. Therefore, $g(\mu)$ can be written as a linear model in the general form:

$$Y_{it}|\mathbf{b}_i \sim \text{independent } F_Y \quad i = 1, \cdots, \text{m}; t = 1, \cdots, n_i$$

$$g(Y_{it}|\mathbf{b}_i) = g(\mu) = \eta = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{it}\mathbf{b}_i + e_{it}$$

$$\mathbf{b}_i \sim i.i.d\ F_b,$$

$$E[\mathbf{b}_i] = 0\ and\ var(\mathbf{b}_i) = \sum b$$

where $\mathbf{b}_i$s are independently distributed and identically distributed (i.i.d) and they have zero mean and they all have the same variance σ_b^2. $\boldsymbol{\beta}$ is the parameter vector for fixed effects,Y_{it} is the vector of responses, $\mathbf{b}_i$ is the random effects, $\mathbf{z}_{it}$ links the random effects to the observations and $\mathbf{x}_{it}$ is a vector of covariates for cluster *i* at time *t* (McCulloch & Neuhaus 2011). e_{it} is the vector of errors such that: $\binom{\mathbf{b}}{e} \sim \text{N}\left(0, \begin{smallmatrix}\mathbf{G} & 0 \\ 0 & \sigma_e^2\mathbf{I}_n\end{smallmatrix}\right)$ (McCulloch & Neuhaus 2011, Schall 1991, Zeger & Karim 1991).

We do not directly estimate random effects. They are characterised by elements of **G** known as variance components. Random effects can be predicted by fitting a mixed model and estimating $\boldsymbol{\beta}$, σ_e^2 and the variance components in **G**.

If conditionally on μ the components of y are independently distributed and y is a member of the exponential func-

tion, then the equation above defines a generalised linear mixed model with random effects. Let $q_1 + \cdots + q_m = \mathbf{q}$, $\mathbf{z} = [z_1, \cdots, z_m]$ and $\mathbf{b} = [b_1', \cdots, b_m']'$ The random vectors are assumed to be uncorrelated with an expectation of zero. The random effects are also said to be uncorrelated with *e*. Moreover, $Cov(b_i) = \sigma_i^2 \mathbf{I}_{qi}$ $(i = 1, \cdots, m)$ so that $cov(b) = \mathbf{D} = diag(\sigma_1^2 I_1, \cdots, \sigma_m^2 I_m)$, where $I_1, \cdots, I_m$ are identity matrices of orders $q_1 \times q_1, \cdots, q_m \times q_m$ In the random effects model, $E(Y_{it}) = \mathbf{X}\boldsymbol{\beta}$, $cov(b) = \mathbf{D}$, $cov(e) = \mathbf{V} = \sigma^2 \mathbf{I}_n$ therefore, $cov(Y_{it}) = V + ZDZ'$. If both the distribution of the random effects and the conditional distribution of y are assumed to be normal, then parameters of β can be estimated by maximum likelihood (Schall 1991). Maximum likelihood estimation is also used in logistic and probit models where the random effects follow a normal distribution and the conditional distribution of y is binomial. In generalised linear models with random effects an algorithm for estimation is proposed. This is because it is difficult to justify a certain distribution or class of distribution of random effects. Maximum likelihood estimation based on the marginal distribution of the observations involves the "integrating out" of the random effects.Numerical integration is usually not practical when random effects are not nested but crossed. Random effects can be nested, crossed, even occur in regression and analysis of covariance models (Schall 1991). We note that a generalised linear mixed model includes a linear predictor, η and a link or inverse link function. Additionally the conditional mean μ depends on the linear predictor through an inverse link function, $h(.)$ and the covariance matrix, R, depends on μ through a variance function.

5.2.1 Inverse link function

In order to map the value of the linear predictor for observations *i*, η_i to the conditional mean of observation *i*, μ_i we need to use the link function. Both μ_i and η_i are scalars for the inverse link function on a one to one basis. The binomial distribution inverse link function is given by:

$$h(\eta) = \frac{e^{\eta}}{1 + e^{\eta}}.$$

Mostly univariate link functions, link and inverse link functions are increasing monotonic functions. Increasing the linear predictor results in an increase in the conditional mean but it's not at a constant rate. The selection of an inverse link function is based on the error distribution (Kachman 2000).

Mean of Y

The mean of Y is given by: $E[y_i] = E[E[y_i|b]] = E[\mu_i] = E[g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}+\mathbf{z}_i'\mathbf{b})]$. Since $g^{-1}(.)$ is a non linear function we cannot simplify the mean. For a certain g(.), if we have a log link $g(\mu) = \log\mu$ and $g^{-1}(x) = expx$. Then:

$$E[y_i] = E[\exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{b})] = \exp(\mathbf{x}_i'\boldsymbol{\beta})E[\exp(\mathbf{z}_i'\mathbf{b})] = \exp(\mathbf{x}_i'\boldsymbol{\beta})M_u(\mathbf{z}_i), \tag{5.2.1}$$

where $M_u(\mathbf{z}_i)$ is the moment generating function of **b** evaluated at $\mathbf{z}_i$ (McCulloch, Searle & John 2008). Suppose $\mu_i \sim N(0, \sigma_\mu^2)$ and that each of the rows of **Z** only has a single non zero entry equal to 1. Then $M_u(\mathbf{z}_i) = \exp\frac{\sigma_\mu^2}{2}$ and $E[y_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta})\exp(\frac{\sigma_\mu^2}{2})$ or $\log E[y_i] = \mathbf{x}_i'\boldsymbol{\beta} + \frac{\sigma_\mu^2}{2}$ (McCulloch et al. 2008)

Variances

The marginal variance of Y is derived by:

$$\begin{aligned} var(y_i) &= var(E[y_i|b]) + E[var(y_i|b)], \\ &= var(\mu_i) + E[\tau^2 v(\mu_i)], \\ &= var(g^{-1}[\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{b}]) + E[\tau^2 v(g^{-1}[\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{b})]. \end{aligned}$$

This cannot be simplified. We have to make assumptions about $g(.)$ or the conditional distribution of y. Therefore we assume that we have a log link and that the elements of **y** are conditionally independent given **b** with a Poisson distribution. The conditional variance of y_i given **b** is $\tau^2 v(\mu_i) = \mu_i$. Applying these to 5.2.1 we get:

$$\begin{aligned} var(y_i) &= var(\mu_i) + E[\mu_i], \\ &= var[exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{b})] + E[exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{b})], \\ &= E[(exp(2(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{b})))] - [E(exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{b}))]^2 + E[exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{b})], \\ &= exp(2\mathbf{x}_i'\boldsymbol{\beta})(M_u(2\mathbf{z}_i)) - [M_u(\mathbf{z}_i)]^2 + exp(-\mathbf{x}_i'\boldsymbol{\beta})M_u(\mathbf{z}_i)). \end{aligned}$$

Furthermore if we assume that $\mu_i \sim N(0, \sigma_\mu^2)$ and that each row of **Z** has only a single non zero entry equal to 1 then:

$$\begin{aligned} var(y_i) &= exp(2\mathbf{x}_i'\boldsymbol{\beta})(exp(2\sigma_\mu^2) - exp(\sigma_\mu^2)) + exp(\mathbf{x}_i'\boldsymbol{\beta})exp(\frac{\sigma_\mu^2}{2}), \\ &= E[y_i](exp(\mathbf{x}_i'\boldsymbol{\beta})[exp(\frac{3\sigma_\mu^2}{2}) - (exp(\frac{\sigma_\mu^2}{2})] + 1). \end{aligned}$$

The value of $var(y_i)$ is greater than 1. Thus the variance is larger than the mean. The marginal distribution will always be over-dispersed compared to the conditional distribution of y_i given **b** which is Poisson. Thus random effects are a way to model over-dispersion to a certain source (McCulloch et al. 2008).

Variance function

A function that is used to model a non-systematic variability is the variance function. The two sources in which residual variability arises from are the sampling distribution or over-dispersion as described in the previous section. An example of sampling distribution is a Poisson random variable with mean μ and variance μ(Kachman 2000). The variance function of a binomial distribution is $\frac{\mu(1-\mu)}{n}$.

An approach used to model the over-dispersion is to scale the residual variability as $var(y_i|\boldsymbol{\mu}) = \phi v(\mu_i)$, where ϕ is an over-dispersion parameter. Secondly, we can add a random effect that is $e_i \sim N(0, \phi)$, to the linear predictor for each observation. In a third approach, we can select another distribution, that is a two parameter (μ, ϕ) negative binomial distribution, which can be used instead of a one parameter μ Poisson distribution for count data.These approaches all use the estimation of an additional parameter ϕ(Kachman 2000).

In summary, the variance function $v(\mu, \phi)$ is used to model the residual variability. The selection of the variance function is influenced by the error distribution which was chosen. There is a need to account for an over-dispersion parameter since the observed residual variability is usually greater than that which is expected due to sampling (Kachman 2000).

Covariances and correlations

The use of random effects bring in a correlation among observations which have any random effect in common. If we assume that the elements of y are conditionally independent then:

$$\begin{aligned} cov(y_i, y_j) &= cov(E[y_i|b], E[y_j|b]) + E[cov(y_i, y_j|\mathbf{b})], \\ &= cov(\mu_i, \mu_j) + E[0], \\ &= cov(g^{-1}[\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{b}], g^{-1}[\mathbf{x}_j'\boldsymbol{\beta} + \mathbf{z}_j'\mathbf{b}]). \end{aligned}$$

Introducing the log link we evaluate the equation as:

$$\begin{aligned} cov(y_i, y_j) &= cov(exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{b}), exp(\mathbf{x}_j'\boldsymbol{\beta} + \mathbf{z}_j'\mathbf{b})), \\ &= exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{x}_j'\boldsymbol{\beta}) cov(exp(\mathbf{z}_i'\mathbf{b}), exp(\mathbf{z}_j'\mathbf{b})), \\ &= exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{x}_j'\boldsymbol{\beta})[M_u(\mathbf{z}_i + \mathbf{z}_j) - M_u(\mathbf{z}_i)M_u(\mathbf{z}_j)]. \end{aligned}$$

Assuming that $\mathbf{b} \sim N(0, \mathbf{I}\sigma_\mu^2)$ and that each row of $\mathbf{Z}$ has a single non zero entry equal to 1. Then, $cov(y_i, y_j) = exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{x}_j'\boldsymbol{\beta})[exp(\sigma_\mu^2)(exp(\mathbf{z}_i'\mathbf{z}_j\sigma_\mu^2) - 1)] cov(y_i, y_j) = 0$ if $\mathbf{z}_i'\mathbf{z}_j = 0$. This is only possible if the two observations do not share a random effect and is positive otherwise that is $\mathbf{z}_i'\mathbf{z}_j = 1$ (McCulloch et al. 2008).

When $\mathbf{z}_i'\mathbf{z}_j = 1$ we can calculate the correlation. We need to cancel $exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{x}_j'\boldsymbol{\beta})$ in the numerator and denominator so that:

$$\begin{aligned} corr(y_i, y_j) &= \frac{e^{2\sigma_\mu^2} - e^{\sigma_\mu^2}}{\sqrt{\left(e^{2\sigma_\mu^2} - e^{\sigma_\mu^2} + e^{-\mathbf{x}_i'\boldsymbol{\beta} + \frac{\sigma_\mu^2}{2}}\right)\left(e^{2\sigma_\mu^2} - e^{\sigma_\mu^2} + e^{-\mathbf{x}_j'\boldsymbol{\beta} + \frac{\sigma_\mu^2}{2}}\right)}}, \\ &= \frac{1}{\sqrt{\left(1 + \eta e^{-\mathbf{x}_i'\boldsymbol{\beta}}\right)\left(1 + \eta e^{-\mathbf{x}_j'\boldsymbol{\beta}}\right)}}, \end{aligned}$$

where η is given by $\frac{1}{\left(e^{3\frac{\sigma_\mu^2}{2}} - e^{\frac{\sigma_\mu^2}{2}}\right)}$ (McCulloch et al. 2008).

5.2.2 Estimation and prediction

The estimating equations for a generalised linear mixed model can be derived using a Bayesian approach which gives solutions which are posterior mode predictors. A Laplacian approximation of the likelihood can be used to estimate the equations. The fixed and random effects estimating equations are: $\begin{pmatrix} \mathbf{X}'\mathbf{H}'\mathbf{R}^{-1}\mathbf{H}\mathbf{X} & \mathbf{X}'\mathbf{H}'\mathbf{R}^{-1}\mathbf{H}\mathbf{Z} \\ \mathbf{Z}'\mathbf{H}'\mathbf{R}^{-1}\mathbf{H}\mathbf{X} & \mathbf{Z}'\mathbf{H}'\mathbf{R}^{-1}\mathbf{H}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\mu} \end{pmatrix}$

$= \begin{pmatrix} \mathbf{X}'\mathbf{H}'\mathbf{R}^{-1}\mathbf{y}* \\ \mathbf{Z}'\mathbf{H}'\mathbf{R}^{-1}\mathbf{y}* \end{pmatrix}$, where

$$\mathbf{H} = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}'},$$

$$\mathbf{R} = var(\mathbf{y}|\boldsymbol{\mu}),$$

$$\mathbf{y}* = \mathbf{y} - \boldsymbol{\mu} + \mathbf{H}\boldsymbol{\eta}.$$

If we rewrite the equation in the form: $\begin{pmatrix} \mathbf{X'WX} & \mathbf{X'WZ} \\ \mathbf{Z'WX} & \mathbf{Z'WZ} + \mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\mu} \end{pmatrix} = \begin{pmatrix} \mathbf{X'hry} \\ \mathbf{Z'hry} \end{pmatrix}$ where $\mathbf{W} = \mathbf{H'R}^{-1}\mathbf{H}$ and $\mathbf{hry} = \mathbf{H'R}^{-1}\mathbf{y}*$ then we see the similarity of GLMM and LMM. We note that the estimating equations of a generalized mixed model is solved iteratively (Kachman 2000).

Likelihood equations for the fixed effects parameters

(McCulloch et al. 2008) states that we can write the likelihood equations in the form:

$$l = \log \int f_{\mathbf{Y}|\mathbf{b}}(\mathbf{Y}|\mathbf{b}) f_{\mathbf{B}}(\mathbf{b}) d\mathbf{b} = \log f_{\mathbf{Y}}(\mathbf{y})$$

then

$$\begin{aligned} \frac{\partial l}{\partial \beta} &= \frac{\partial}{\partial \beta} \int \frac{f_{\mathbf{Y}|\mathbf{b}}(\mathbf{Y}|\mathbf{b}) f_{\mathbf{B}}(\mathbf{b}) d\mathbf{b}}{f_{\mathbf{Y}}(\mathbf{y})}, \\ &= \int \frac{\left[\frac{\partial}{\partial \beta} f_{\mathbf{Y}|\mathbf{b}}(\mathbf{Y}|\mathbf{b})\right] f_{\mathbf{B}}(\mathbf{b}) d\mathbf{b}}{f_{\mathbf{Y}}(\mathbf{y})}. \end{aligned} \tag{5.2.2}$$

Since $f_{\mathbf{B}}(\mathbf{b}$ does not involve β we note that:

$$\begin{aligned} \frac{\partial}{\partial \beta} f_{\mathbf{Y}|\mathbf{b}}(\mathbf{Y}|\mathbf{b}) &= \left(\frac{1}{f_{\mathbf{Y}|\mathbf{b}}(\mathbf{Y}|\mathbf{b})} \frac{\partial f_{\mathbf{Y}|\mathbf{b}}(\mathbf{Y}|\mathbf{b})}{\partial \beta} \right) f_{\mathbf{Y}|\mathbf{b}}(\mathbf{Y}|\mathbf{b}), \\ &= \frac{\partial \log f_{\mathbf{Y}|\mathbf{b}}(\mathbf{Y}|\mathbf{b})}{\partial \beta} f_{\mathbf{Y}|\mathbf{b}}(\mathbf{Y}|\mathbf{b}). \end{aligned}$$

Rewriting 5.2.2:

$$\begin{aligned} \frac{\partial l}{\partial \beta} &= \int \frac{\partial \log f_{\mathbf{Y}|\mathbf{b}}(\mathbf{Y}|\mathbf{b})}{\partial \beta} f_{\mathbf{Y}|\mathbf{b}}(\mathbf{Y}|\mathbf{b}) f_{\mathbf{B}}(\mathbf{b}) d\mathbf{b} / f_{\mathbf{Y}}(\mathbf{y}), \\ &= \int \frac{\partial \log f_{\mathbf{Y}|\mathbf{b}}(\mathbf{Y}|\mathbf{b})}{\partial \beta} f_{\mathbf{B}|\mathbf{y}}(\mathbf{b}|\mathbf{y}) d\mathbf{b}. \end{aligned} \tag{5.2.3}$$

Using the matrix notation $\frac{\partial l}{\partial \beta} = \frac{1}{\tau^2}\mathbf{X'W\Delta}(\mathbf{y} - \boldsymbol{\mu})$ which gives the derivative of the log likelihood for a generalised linear model in 5.2.3 we get:

$$\begin{aligned} \frac{\partial l}{\partial \beta} &= \int \mathbf{X'W}^*(\mathbf{y} - \boldsymbol{\mu}) f_{\mathbf{B}|\mathbf{y}}(\mathbf{b}|\mathbf{y}) d\mathbf{b}, \\ &= \mathbf{X'E}[\mathbf{W}^*|(\mathbf{y}]\mathbf{y} - \mathbf{X'E}[\mathbf{W}^*\boldsymbol{\mu}|\mathbf{y}]. \end{aligned}$$

$\mathbf{W}^* = \{_d[\tau v(\mu_i) g_\mu(\mu_i)]^{-1}\}$ The likelihood equation for β is therefore:

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

Note that $\mathbf{W}^*$ and $\mathbf{W}^*\boldsymbol{\mu}$ are replaced by their conditional expected values given y. If y follows a Poisson distribution then $\mathbf{W}^* = \mathbf{I}$. The equations simplify to $\mathbf{X'y} = \mathbf{X'E}[\boldsymbol{\mu}|\mathbf{y}]$ (McCulloch et al. 2008).

Likelihood equations for the random effects parameters

Given the distribution of $f_{\mathbf{b}}(\mathbf{b})$ we can derive an equation similar to 5.2.3 for the ML equations for the parameters. Let φ denote the parameters then:

$$\begin{aligned}\frac{\partial l}{\partial \Phi} &= \int \frac{\partial \log f_{\mathbf{B}}(\mathbf{b})}{\partial \varphi} f_{\mathbf{B}|\mathbf{y}}(\mathbf{b}|\mathbf{y}) d\mathbf{b}, \\ &= \mathbf{E}\left[\frac{\partial \log f_{\mathbf{B}}(\mathbf{b})}{\partial \varphi}|\mathbf{y}\right].\end{aligned}$$

To simplify the equation further we need to specify a form for the random effects (McCulloch et al. 2008).

5.2.3 Variance component estimation

The estimation of the variance component is based on its association with the random effects and the error distribution.

Residual maximum likelihood estimation (REML) quasi-likelihood estimation

The variance components of a GLMM can be estimated using the approximate REML quasi-likelihood given by:

$$ql(\beta, \sigma) = -\frac{1}{2}\ln|V| - \frac{1}{2}\ln|\mathbf{X}'\mathbf{H}'\mathbf{V}^{-1}\mathbf{H}\mathbf{X}| - \frac{1}{2}(\mathbf{y}^* - \mathbf{H}\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}^* - \mathbf{H}\mathbf{X}\boldsymbol{\beta}),$$

where σ is the vector of variance component and $\mathbf{V} = \mathbf{R} + \mathbf{H}\mathbf{Z}\mathbf{G}\mathbf{Z}'\mathbf{H}'$. When the variance component is associated with the random effects in $\mathbf{G}$, the estimating equations remain the same. $\hat{\mu}$ and $\mathbf{c}$ are obtained using 5.2.2. The quadratic form of the variance components associated with the error distribution is given by $(\mathbf{y} - \hat{\boldsymbol{\mu}})'\mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial \phi}\mathbf{R}^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}})$. The functions of the left hand sides of ϕ are

$$f_{00}(\mathbf{C}) = [tr(\boldsymbol{\Phi}) - 2tr(\boldsymbol{\Omega}\boldsymbol{\Phi})] + tr(\boldsymbol{\Omega}\boldsymbol{\Phi}\boldsymbol{\Omega}\boldsymbol{\Phi})],$$

$$f_{i0}(\mathbf{C}) = tr(\mathbf{C}^i \begin{pmatrix} \mathbf{X}'\mathbf{H}'\boldsymbol{\Phi}\mathbf{H}\mathbf{X} & \mathbf{X}'\mathbf{H}'\boldsymbol{\Phi}\mathbf{H}\mathbf{Z} \\ \mathbf{Z}'\mathbf{H}'\boldsymbol{\Phi}\mathbf{H}\mathbf{X} & \mathbf{Z}'\mathbf{H}'\boldsymbol{\Phi}\mathbf{H}\mathbf{Z} \end{pmatrix} \mathbf{C}^{i'}),$$

where $\mathbf{C}^i = (\mathbf{C}^{i0}\ \mathbf{C}^{i1} \cdots \mathbf{C}^{ir}), \Phi = \mathbf{R}^{-1}\frac{\partial \mathbf{R}}{\partial \phi}\mathbf{R}^{-1}$ and $\boldsymbol{\Omega} = (\mathbf{H}\mathbf{X}\ \mathbf{H}\mathbf{Z})\mathbf{C}\begin{pmatrix} \mathbf{X}'\mathbf{H}' \\ \mathbf{Z}'\mathbf{H}'. \end{pmatrix}$ (Kachman 2000).

5.2.4 The random-intercept logistic regression model

We define the Random-Intercept Logistic Regression Model as a model with p covariates for the dichotomous response Y_{ij} of subject $i(i = 1, \cdots, N)$ within cluster j $(j = 1, \cdots, n_i)$: Therefore the model can be written as

$$
\begin{aligned}
Y_{ij}|\Pi_{ij} &\sim Binomial(1, \Pi_{ij}), \\
\Pi_{ij} &= Pr(Y_{ij} = 1|X_{2j}, \cdots, X_{nj}, U_j), \\
\log\left[\frac{Pr(Y_{ij}=1)}{1-Pr(Y_{ij}=1)}\right] &= \text{logit}(\Pi_{ij}), \\
logit(\Pi_{ij}|U_i) &= \beta_0 + \beta_1 X_{1j} + \cdots + \beta_i X_{ij} + U_j + \epsilon_{ij} \\
&= \mathbf{X'_{ij}}\boldsymbol{\beta} + \mathbf{U_j} + \epsilon_{\mathbf{ij}}
\end{aligned}
$$

where:

- $Y_{ij} =$ dichotomous response of subject i within cluster j
- $X_{ij} = (p+1)$x1vector of covariates
- $\beta_0 =$ log odds of response for a typical subject with $\mathbf{X} = 0$ and $U_j = 0$
- $\boldsymbol{\beta} = (p+1)$x1vector of regression coefficients
- $U_j =$ random subject effects distributed $NID(0, \sigma_u^2)$
- $\sigma_u^2 =$ degree of heterogeneity across subjects in the probability of response not attributable to $\mathbf{X}$
- $\epsilon_{ij} \sim$ std logistic(mean 0, variance $\frac{\pi^2}{3}$)

We note that:

(1) Every subject has its own propensity for response U_j.

(2) The influence of covariates $\mathbf{X}$ is determined controlling (or adjusting) for the subject effect.

(3) The covariance structure, or dependency, of the repeated observations is explicitly modelled.

(4) It is most useful when the objective is to make inference about subjects rather than the population average.

(5) The interest is in the heterogeneity of subjects.

5.2.5 Logistic regression as a latent variable model

If $logit(\Pi_{ij}|U_i) = \beta_0 + \beta_1 X_{1j} + \cdots + \beta_i X_{ij} + (U_j + \epsilon_{ij})$, then $logit(\Pi_{ij}) = 1 \leftrightarrow logit(\Pi_{ij}) > 0$. Let $\xi = (U_j + \epsilon_{ij})$ then the variance is $var(\xi) = \sigma_u^2 + \frac{\pi^2}{3}$ The interclass correlation coefficient is given by $\rho = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\pi^2}{3}}$ A correlation between any two observations from the same subject is formed after allowing for subject to subject variability through a random effect. A model with a random intercept can form a compound correlation matrix given by:

$$\begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

5.3 Application of generalised linear mixed models to the Zimbabwe data

We explored the relationship between IPV and HIV prevalence in Zimbabwe by applying the generalised linear mixed model. This was done by controlling for socio-demographic factors and using the DHS survey clusters, as random effects.

GLMM give us an opportunity to model a binary response variable which in our case is an HIV test result, and to take random effects such as a survey cluster into account. In order to account for relationships between predictors we assessed the socio-demographic variables used in the logistic regression model together with the HIV response in a GLMM. This was done to see whether they affected the magnitude of association between IPV and HIV and if so, how the magnitude was affected.

5.3.1 Model fitting

The following model was fitted:

$$logit(\Pi_{ij}|U_j) = \beta_0 + \beta_1 phyj + \beta_2 sexj + \beta_3 agej + \beta_4 mstatusj + \beta_5 relj + \beta_6 eduj + \beta_7 stij + \beta_8 wij + U_j + \epsilon_{ij}$$

where $sex = sexual$, $phy = physical$, $mstatus = maritalstatus$, $rel = religion$, $edu = education$, $wi = wealth\ index$, $U_j = v001 = cluster\ number$. The following command was used to generate the GLMM model,(xtmelogit hiv03 i.sexual i.physical i.v013 i.religion i.mstatus i.v763a i.v130 i.v190, v001:, covariance (independent) or).

Table 5.1: Estimates of adjusted odds ratios [95% confidence intervals], p-value for the HIV outcome and the random effects parameter using GLMM (*Analysis based on the Zimbabwe DHS* $2005 - 2006$ *data*).

	Predictor	Categories	OR(95%CI)	Std. Err.	z	P¿z
Fixed Effects	Sexual	low	1			
		High	0.85(0.67,1.07)	0.10	-1.36	0.17
	Physical	low	1.00			
		high	1.11(0.94,1.32)	0.10	1.22	0.22
	Age	15-19	1.00			
		20-24	1.39(0.89,2.17)	0.32	1.46	0.14
		25-29	2.84(1.84,4.38)	0.63	4.72	0
		30-34	3.58(2.32,5.54)	0.80	5.75	0
		35-39	3.27(2.09,5.13)	0.75	5.16	0
		40-44	2.38(1.48,3.82)	0.58	3.58	0
		45-49	1.32(0.79,2.19)	0.34	1.06	0.29
	Education	None	1.00			
		Prim	1.20(0.82,1.76)	0.23	0.95	0.34
		Sec	1.23(0.82,1.85)	0.25	1.01	0.31
		Higher	0.74(0.38,1.44)	0.25	-0.89	0.38
	Religion	Traditional	1.00			
		Roman	1.56(0.91,2.68)	0.43	1.63	0.10
		Other	1.64(0.96,2.80)	0.45	1.83	0.07
	Marital Status	Married/Living together	1.00			
		Separated	3.84(3.20,4.61)	0.36	14.42	0
	STI	None	1.00			
		Yes	2.95(2.14,4.05)	0.48	6.67	0
		Present1	1.32(0.33,5.28)	0.93	0.4	0.80
		Present2	1.42(0.30,6.66)	1.12	0.44	0.66
	Wlth Index	Poorest	1.00			
		Poorer	1.04(0.82,1.31)	0.13	0.28	0.78
		Middle	1.19(0.93,1.52)	0.15	1.37	0.17
		Richer	1.33(1.04,1.69)	0.16	2.27	0.02
		Richest	0.88(0.66,1.18)	0.13	-0.84	0.40
Random effects	Cluster(cons)	v001	sd-0.24,CI(0.10,0.55)	0.10		

5.4 Results

The results in table 5.1 for the generated GLMM show whether sexual and physical violence affect a woman's chance of contracting HIV. The odds ratio for HIV risk of a woman experiencing high physical violence was 1.11, which was higher than that of high sexual violence which was 0.85. The results show that neither sexual violence nor physical violence is significantly associated with HIV as $p > 0.01$. Of the socio-demographic variables age, marital status and the presence of an STI were significant predictors of HIV in the multivariate model. Women in the age group 25 to 44 years were likely to be HIV positive (OR were between 2-3.73 with $p = 0$). The OR for marital status was 3.84 with CI(3.20,4.61) and $p = 0$. The estimated OR for a person with STI was 2.95 ($p = 0$). The religion, education and Wealth Index variables did not give a statistically significant response. The random effects test was that $H_0 = \sigma_u^2 = 0$ and $H_1 = \sigma_u^2 > 0$. The random effect on the model gave the result LR test vs. logistic regression:$chibar2(01) = 1.60$ $Prob \geq chibar2 = 0.10$. Therefore, we reject the null hypothesis and conclude that the random effects model provides a better fit of the GLMM (it has the lowest log likelihood).

Similarly a GLMM with IPV as an index after combining indices of physical and sexual violence revealed that IPV was not statistically significant with HIV. The variables age, marital status and STI showed an association with HIV. Using standard deviation 0.24 for the cluster, the intraclass correlation coefficient is:

$$\begin{aligned}\rho &= \frac{\sigma_u^2}{\sigma_u^2 + \frac{\pi^2}{3}} \\ &= \frac{0.0576}{0.0576 + 3.29} \\ &= 0.0172.\end{aligned}$$

If we multiply by 100 then it is approximately equal to two percent. The intraclass correlation further suggests that the GLMM is a better model than the logistic regression model, and that the observations in the same cluster are similar and related.

Chapter 6

Conclusion and recommendations

6.1 Discussion

The results of the research suggest that there was no consistent association between reported covariates and HIV. Our focus was to find out how physical and sexual violence are associated with a woman getting HIV. Unadjusted odds Ratio using Survey Logistic Regression analysis for physical and sexual IPV revealed that women who experienced high physical and sexual violence had a high prevalence of HIV. Adjusted associational ORs were between $0.81 - 1.07$ for physical and sexual violence. These relationships showed no significant association with HIV when we considered the p-values and chi-square significance tests. This suggested that there was no association between HIV and IPV.

The same findings that HIV and a PCA-based index of IPV confirmed the same results that HIV and IPV were not associated using the GLMM. This was after controlling for the socio-demographic variables. Taking the survey cluster as random effects in the GLMM provided a better fit of the model, compared to the survey logistic regression model.

These results support the research done by (Harling et al. 2010) which concluded that there was no association between HIV and IPV among women in 10 developing countries and Zimbabwe was one of them. Shi et al. (2013)

conducted a study on seven nations in east and southern Africa and found no significant association between IPV and HIV sero-conversion in discordant couples. However this study found an association between IPV and the prevalence of HIV infection (Shi et al. 2013). When considering the association of any variables, it is important to consider the cross-sectional nature of the data. Most of the time we believe that IPV is a consequence of a woman being HIV-positive. The DHS data was strong in that women who tested HIV-positive were unaware of their status before the interview, which reduced the connection between sero-positivity and partner awareness. When conducting a survey we want to rule out selection bias. However the data suggests a possibility of selection bias if those women who participated in HIV testing, or the domestic violence module, were different from those who did not. Information on HIV test acceptance was not available (Harling et al. 2010). When testing someone for HIV it is always advisable to get the consent of the person before a test is done to show that a person was not forced to do so.

It follows that the measures of IPV gathered by DHS are not perfect. Sometimes it is difficult to interpret the association of IPV and HIV as the physical violence questions asked related to a woman's last "husband/partner". There were no questions on the number of times IPV was experienced except where in the categories of "often" or "sometimes" (Harling et al. 2010). A woman without a partner may interpret the question differently. The analysis done in survey logistic regression when women were stratified by marital status, suggested that those who were separated experienced high physical and sexual violence. Marital status had an association with HIV and we concluded that women who were separated were at high risk of getting HIV. As a result of separation, a woman might tend to be involved in sexual risk behaviour by having many partners some of whom might be HIV positive. Age is significantly associated with HIV. Women between the age of $20-24$ were at high risk of acquiring HIV. This could be as a result of many women in that age group being very sexually active, as compared to those in the age group of $15-19$ and $45-49$. Women in the age group of $15-19$ might still be under the care of parents, and those in the age group $45-49$ are approaching menopause. Results based on GLMM revealed that the variables age, marital status and the presence of STI were significantly associated with HIV.

Interestingly, similar studies done on Indian, South African and Kenyan women of the same age group reported a contrast in their findings. These studies suggested an association of IPV and HIV (Harling et al. 2010, Shi et al. 2013). In Kenya, different studies had different findings. In a bivariate analysis of people attending an STI

clinic in Nairobi an association was found were as other Kenyan studies found no association (Shi et al. 2013). This calls for further studies in the same country.

In this paper we explored data of the female population in Zimbabwe whereas other studies used samples that did not include the whole female population, but used samples from health facilities. However, the study in India using the same data analysis with the same year of survey, reported a relationship between IPV and HIV. Harling et al. (2010) concludes that although there was an association between IPV types and HIV in India, the association was not significant at 95% confidence interval. Considering all the previous studies until 2010, the strongest evidence for a relationship of IPV and HIV till was found in South Africa.

STI has an impact on HIV. In Zimbabwe, women with STIs were at a high risk of HIV and thus there was an association of HIV and STI. Harling et al. (2010) found similar results in their study in Kenya and Rwanda that there was a positive association of HIV and STI. Various results in different countries show conflicting results on the relationship of HIV and IPV because data might not be stratified by country which could result in a geographical variation in the strength of the association between IPV and HIV (Shi et al. 2013). Authors have suggested that further research is required to determine whether the relationship between IPV and HIV in various countries is affected by whether the study population is clinic based or is a national sample. Another factor to consider is geographical effect-modification (Harling et al. 2010). There should be interventions to see which settings prove to have a stronger relationship between HIV and IPV.

Studies done in India reported that men who indulge in IPV also commit high gendered HIV risk behaviours. These include sexual infidelity, coercive condom practices and transactional sex. In cases where we have Most-at-Risk-Populations (MARP), including sex workers, and partners of IPV perpetrators there is an increased risk of HIV as a result of these risky behaviours (Harling et al. 2010). Some authors argue that although countries like Haiti have high HIV prevalence, with high infection among MARPS, its social dislocation and economic stagnation looms large. In Haiti, the factors which put women at risk of IPV actually reduce the risk of the partner being HIV positive, reducing the woman's risk of getting HIV. If HIV is present in a partner's sexual network then changes in the HIV risk behaviour only affects the HIV risk. What can help us understand the reasons for HIV and IPV being associated at times, is to determine the risk factors for the presence of HIV in a social network (Harling et al. 2010).

Individuals are likely to be HIV-positive if they are exposed to the following risk factors: increased risk of violent intercourse; lower decision-making powers and partner's high-risk sexual behaviour. Studies suggest that gender inequity and sexual risk can increase IPV rates (Harling et al. 2010). If gender inequity does not change the sexual networks of an individual, such that HIV is likely to be present in either partner, then this might not have an effect on the HIV status of an individual. Gender inequity is not consistently associated with IPV in a society with populations experiencing both IPV and non IPV. There are settings with more inequity that are connected to high HIV rates (Harling et al. 2010).

Besides focusing on the relationship of HIV with IPV, we also explored the relationship between IPV and socio-economic and socio-demographic factors using the multinomial logit regression model. We found IPV to be associated with quite a number of factors such as age, education, region, religion, marital status, wealth index, media exposure and occupation. A woman's ability to make a decision in a relationship, and the type of residence she stayed in did not have any effect of IPV. This study presents evidence that IPV and HIV are not consistently associated among ever-married women in Zimbabwe and worldwide.

6.2 Conclusion

The main objective of this study was to use appropriate statistical models to find out if there is a relationship between intimate partner violence and HIV infection among women in Zimbabwe. A generalized linear model in the form of survey logistic regression and a generalized linear mixed model were used in this study. Research has shown that HIV prevalence is higher in women than in men, suggesting that HIV intervention programmes can benefit from a gender perspective (Harling et al. 2010). However, these results have a limitation on policy recommendations since policies have to be made to protect women who experience IPV are vulnerable and are considered to be at high risk of contracting HIV. Instead of funding for HIV prevention and adult IPV, prevention funds can be diverted somewhere and victims of IPV fail to benefit from this assistance. Our results do not give insights into reducing IPV, therefore more studies are recommended to focus on addressing childhood IPV. However, it is still important to advocate for the reduction of IPV as a public health goal and a basic human right. Further research or investigation is also required to understand under which situations HIV and IPV are

related, and the methods which might be used to determine the strengths of the relationship.

References

Abeya, S. G., Afework, M. F. & Yalew, A. W. (2011), Intimate partner violence against women in western ethiopia: prevalence, patterns, and associated factors, *BMC Public Health* **11**(1), 913.

Abramsky, T., Watts, C., Garcia-Moreno, C., Devries, K., Kiss, L., Ellsberg, M., Jansen, H. & Heise, L. (2011*a*), What factors are associated with recent intimate partner violence? findings from the who multi-country study on women's health and domestic violence, *BMC Public Health* **11**(1), 109.

Abramsky, T., Watts, C. H., Garcia-Moreno, C., Devries, K., Kiss, L., Ellsberg, M., Jansen, H. A. & Heise, L. (2011*b*), What factors are associated with recent intimate partner violence? findings from the who multi-country study on women's health and domestic violence, *BMC Public Health* **11**(1), 109.

Achia, T. N., Wangombe, A. & Khadioli, N. (2010), A logistic regression model to identify key determinants of poverty using demographic and health survey data, *European Journal of Social Science* **13(1)**, 38–45.

Afifi, A., Clark, V. A. & May, S. (2003), *Computer-aided multivariate analysis*, Vol. 62, Chapman & Hall/CRC.

Akyüz, A., Yavan, T., Şahiner, G. & Kiliç, A. (2012), Domestic violence and woman's reproductive health: A review of the literature, *Aggression and Violent Behavior* .

Archer, K. J., Lemeshow, S. & Hosmer, D. W. (2007), Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design, *Computational Statistics & Data Analysis* **51**(9), 4450–4464.

Barros, C., Schraiber, L. B. & França-Junior, I. (2011), Association between intimate partner violence against women and hiv infection, *Revista de Saúde Pública* **45**(2), 365–372.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H. & White, J.-S. S. (2009), Generalized linear mixed models: a practical guide for ecology and evolution, *Trends in ecology & evolution* **24**(3), 127–135.

Campbell, J. C., Baty, M., Ghandour, R. M., Stockman, J. K., Francisco, L. & Wagman, J. (2008), The intersection of intimate partner violence against women and hiv/aids: a review, *International journal of injury control and safety promotion* **15**(4), 221–231.

Campbell, J. et al. (2002), Health consequences of intimate partner violence, *Lancet* **359**(9314), 1331–1336.

Dolgun, A. (2012), Multivariate Analysis: Logistic Regression, PhD thesis, Hacettepe University.

Dunkle, K. L., Jewkes, R. K., Brown, H. C., Gray, G. E., McIntryre, J. A. & Harlow, S. D. (2004), Gender-based violence, relationship power, and risk of hiv infection in women attending antenatal clinics in south africa., *The lancet* .

Dunteman, G. H. (1989), *Principal components analysis*, number 69, Sage.

Garcia-Moreno, C. & Watts, C. (2000), Violence against women: its importance for hiv/aids, *Aids* **14**.

Hardle, W. K. & Simar, L. (2012), *Applied multivariate statistical analysis*, Springer.

Harling, G., Msisha, W. & Subramanian, S. (2010), No association between hiv and intimate partner violence among women in 10 developing countries, *PLoS One* **5**(12), e14257.

Heeringa, S. G., West, B. T. & Berglund, P. A. (2010), Applied survey data analysis.

Hosmer Jr, D. W., Lemeshow, S. & Sturdivant, R. X. (2013), *Applied logistic regression*, Wiley. com.

Hove, K. & Gwazane, M. (2011), A study to determine factors associated with domestic violence among concordant and discordant couples in zimbabwe, *International Journal of Humanities and Social Sciences Vol. 1 No. 7 [Special issue–June 2011]* .

Jewkes, R. (2002), Intimate partner violence: causes and prevention, *The Lancet* **359**(9315), 1423–1429.

Jewkes, R., Dunkle, K., Nduna, M., Levin, J., Jama, N., Khuzwayo, N., Koss, M., Puren, A. & Duvvury, N. (2006), Factors associated with hiv sero-status in young rural south african women: connections between intimate partner violence and hiv, *International Journal of Epidemiology* **35**(6), 1461–1468.

Jewkes, R. K., Dunkle, K., Nduna, M. & Shai, N. (2010), Intimate partner violence, relationship power inequity, and incidence of hiv infection in young women in south africa: a cohort study, *The Lancet* **376**(9734), 41–48.

Kachman, S. D. (2000), An introduction to generalized linear mixed models, *in* 'Proceedings of a symposium at the organizational meeting for a NCR coordinating committee on Implementation Strategies for National Beef Cattle Evaluation, Athens', pp. 59–73.

Kleinbaum, D., Dietz, K. & Krickeberg, K. (1994), Statistics in the health sciences: logistic regression, *Statistics in the health sciences: logistic regression* .

Kleinbaum, D. G., Klein, M. & Pryor, E. R. (2002), *Logistic regression: a self-learning text*, Springer Verlag.

Lange, K. (2002), *Mathematical and statistical methods for genetic analysis*, Springer.

Lindsey, J. K. (1997), *Applying generalized linear models*, Springer Verlag.

Maman, S., Campbell, J., Sweat, M. D. & Gielen, A. C. (2000), The intersections of hiv and violence: directions for future research and interventions, *Social science & medicine* **50**(4), 459–478.

Martin, S. L. & Curtis, S. (2004), Gender-based violence and hiv/aids: recognising links and acting on evidence, *The Lancet* **363**(9419), 1410–1411.

McCullagh, P. & Nelder, J. A. (1989), *Generalized linear model*, Vol. 37, Chapman & Hall/CRC.

McCulloch, C. E. & Neuhaus, J. M. (2011), Prediction of random effects in linear and generalized linear models under model misspecification, *Biometrics* **67**(1), 270–279.

McCulloch, C. E., Searle, S. R. & John, M. (2008), *Generalized, Linear, and Mixed Models*, A John Wiley & Sons, Inc., Hoboken, New Jersey.

Miner, S., Ferrer, L., Cianelli, R., Bernales, M. & Cabieses, B. (2011), Intimate partner violence and hiv risk behaviors among socially disadvantaged chilean women, *Violence Against Women* **17**(4), 517–531.

Norusis, M. J. et al. (2010), Pasw statistics 18 advanced statistical procedures.

Nyamayemombe, C., Mishra, V., Rusakaniko, S., Benedikt, C. & Gwazane, M. (2010), The association between violence against women and hiv: Evidence from a national population-based survey in zimbabwe., Technical report.

Organisation, W. H. (2005), Who multi-country study on womens health and domestic violence against women: summary report of initial results on prevalence, health outcomes and womens responses, Technical report, World Health Organisation.

Osinde, M. O., Kaye, D. K. & Kakaire, O. (2011), Intimate partner violence among women with hiv infection in rural uganda: critical implications for policy and practice, *BMC women's health* **11**(1), 50.

Schall, R. (1991), Estimation in generalized linear models with random effects, *Biometrika* **78**(4), 719–727.

Sharma, S. (1995), *Applied multivariate techniques*, John Wiley & Sons, Inc.

Shi, C.-F., Kouyoumdjian, F. G. & Dushoff, J. (2013), Intimate partner violence is associated with hiv infection in women in kenya: A cross-sectional analysis, *BMC public health* **13**(1), 512.

Silverman, J. G., Decker, M. R., Saggurti, N., Balaiah, D. & Raj, A. (2008), Intimate partner violence and hiv infection among married indian women, *JAMA: The Journal of the American Medical Association* **300**(6), 703–710.

StataCorp, L. (2009), Stata 11 base reference manual, *College Station, TX: Stata Corporation* .

Timm, N. H. (2002), Principal component, canonical correlation, and exploratory factor analysis, *in* 'Applied Multivariate Analysis', Springer New York.

Van der Straten, A., King, R., Grinstead, O., Vittinghoff, E., Serufilira, A. & Allen, S. (1998), Sexual coercion, physical violence, and hiv infection among women in steady relationships in kigali, rwanda, *AIDS and Behavior* **2**(1), 61–73.

World Health Organisation, L. S. o. H. & Tropical Medicine, S. A. M. R. C. (2013), Global and regional estimates of violence against women: Prevalence and health effects of intimate partner violence and non partner sexual violence., Technical report, London school of hygiene and tropical medicine.

World Health Organisation, U. (2004), Violence against women and hiv/aids: Critical intersections (intimate partner violence and hiv/aids), Technical report, World Health Organisation, UNAIDS.

Zeger, S. L. & Karim, M. R. (1991), Generalized linear models with random effects; a gibbs sampling approach, *Journal of the American statistical association* **86**(413), 79–86.

Zimbabwe, C. S. O. C. & Inc., M. I. (2007), *Zimbabwe Demographic and Health Survey 2005-2006*, Central Statistics Office Harare, Zimbabwe and Macro International Inc. Calverton Maryland, USA.