Estimating Risk Determinants of HIV and TB in South Africa

A thesis presented to

The University of KwaZulu Natal

in fulfilment of the requirement for the degree

of

Master of Science in Statistics

by

Thembile Mzolo

School of Statistics and Actuarial Sciences

UNIVERSITY OF
KWAZULU-NATAL

University of KwaZulu Natal

Abstract

Where HIV/AIDS has had its greatest adverse impact is on TB. People with TB that are infected with HIV are at increased risk of dying from TB than HIV. TB is the leading cause of death in HIV individuals in South Africa. HIV is the driving factor that increases the risk of progression from latent TB to active TB. In South Africa no coherent analysis of the risk determinants of HIV and TB has been done at the national level this study seeks to mend that gab.

This study is about estimating risk determinants of HIV and TB. This will be done using the national household survey conducted by Human Sciences Research Council in 2005. Since individuals from the same household and enumerator area more likely to be more alike in terms of risk of disease or correlated among each other, the GEEs will be used to correct for this potential intraclass correlation. Disease occurrence and distribution is highly heterogeneous at the population, household and the individual level. In recognition of this fact we propose to model this heterogeneity at community level through GLMMs and Bayesian hierarchical modelling approaches with enumerator area indicating the community effect.

The results showed that HIV is driven by sex, age, race, education, health and condom use at sexual debut. Factors associated with TB are HIV status, sex, education, income and health. Factors that are common to both diseases are sex, education and health. The results showed that ignoring the intraclass correlation can results to biased estimates. Inference drawn from GLMMs and Bayesian approach provides some degree of confidence in the results. The positive correlation found at an enumerator area level for both HIV and TB indicates that interventions should be aimed at an area level rather than at the individual level.

Declaration

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

December 2008.

Ms Thembile Mzolo (203512217) Date

Dr Henry Mwambi Date

Dr Khangelani Zuma Date

Notes

A number of papers have been produced from this thesis. The abstracts of the papers appear in the Appendix. The full references of the papers are as follows:

1. Estimating Risk Determinants of HIV and TB in the South African Population. *Presented at the Faculty of Science and Agriculture Postgraduate Research Day, University of KwaZulu Natal, Durban, South Africa, 28 September 2007.*

2. Estimating Risk Determinants of HIV and TB in South Africa. *Presented at the 1st Africa Conference of Young Statisticians 2008, Pretoria, South Africa, 1-3 July 2008.*

3. Bayesian approach in Estimating Risk determinants of infectious diseases. *Presented at the 51st Annual South African Statistical Association (SASA) Conference, Pretoria, South Africa, October 2008.*

The third paper was also presented at the Young Statisticians Seminar Series on the 14 November 2008 organized by Statistics South Africa in Pretoria.

Acknowledgements

This study has been a long journey with its own ups and downs and I would like to thank those who helped me to survive until now. First I would to pass my gratitude to my supervisor Dr Henry Mwambi for the time he spent in making sure that this thesis is successful. Your support is very much duly acknowledged. I would like to thank him for the comments, advise and everything he did for me. I would also like to thank my co-supervisor Dr Khangelani Zuma without you this would have not been what it is today. I want to say thank you to Professor Scott Zeger for his constructive comments.

Dr Manda helped me to get through Winbugs software and I am very grateful for that. This study was made possible by the bursary I received from the DST/NRF Centre for Epidemiological Modelling and Analysis (SACEMA) to them I would like to say thank you so much. I would like to thank HSRC for allowing me to use their rich dataset. My attendance to the Young Statistician Seminar Series, Young Statistician and SASA conferences was made possible by Statistics South Africa therefore, Statistics South Africa thank you very much.

To my family, parents, brother, late brothers, nephews and niece thank you for supporting me even in difficult times but you were my pillar of strength. I would also like to thank my wonderful friends and my fellow postgraduates for their support and wise advises they gave me.

Contents

Abstract ii

Declaration iii

Notes iv

Acknowledgements v

List of Figures ix

List of Tables x

List of Abbreviations x

1 Introduction **1**

1.1 HIV and AIDS in the World and South Africa 1

1.2 Tuberculosis in the World and South Africa 7

1.3 Relationship between HIV and TB 10

1.4 Objectives of this Research 12

1.4.1 Research hypothesis 12

1.5 The Data 13

1.5.1 Sample Weights 15

1.6 Statistical Methods 17

1.6.1 Generalized Linear Models 17

1.6.2 Binary Response 19

1.7 Thesis overview 22

2 Exploratory Data Analysis **22**

2.1 Introduction 22

2.2 Descriptive Analysis of HIV 23

2.3 Descriptive analysis of TB 29

2.3.1 Interaction of TB and HIV 31

2.4 Logistic Regression Modelling 32

2.4.1 Model results for HIV data 34

2.4.2 Model results for TB 36

2.5 Discussion 37

2.5.1 Human Immune Virus 37

2.5.2 Tuberculosis 40

3 **Generalized Estimating Equations and Generalized Linear Mixed Models** **44**

3.1 Generalized Estimating Equations 44

3.1.1 Introduction 44

3.1.2 Modelling Correlated Binary Data 45

3.1.3 Results of GEEs 48

3.1.3.1 HIV Results from GEEs 48

3.1.3.2 TB Results from GEEs 48

3.2 Generalized Linear Mixed Models 49

3.2.1 Introduction 49

3.2.2 Modelling binary response using GLMMs 49

3.2.2.1 Quasi-likelihood 50

3.2.3 Results of GLMMs 57

3.2.3.1 GLMMs HIV Results 57

3.2.3.2 GLMMs TB results 58

3.3 Discussion 59

4 **Bayesian Markov Chain Monte Carlo methods** **64**

4.1 Introduction 64

4.2 Direct Sampling Methods 66

4.2.1 Acceptance-Rejection Sampling 67

4.3 Markov Chain Monte Carlo 67

4.3.1 The Metropolis-Hastings Algorithm 69

4.3.1.1 Convergence Diagnostics 75

4.3.1.2 A Metropolis-Hastings Acceptance-Rejection Algorithm 78

4.3.2 The Gibbs Sampler 81

4.3.2.1 Gibbs sampler as a special case of the M-H algorithm 85

4.3.2.2 Convergence Diagnostics 86

4.3.3 Graphical Modelling 87

4.4 Analysis of the data 89

4.5 Discussion 96

5 **Discussion and Conclusions** **99**

5.1 Human Immune Virus 99

5.2 Tuberculosis 103

5.3 Intervention Implications 105

5.4 Different Statistical approaches 106

5.5 Conclusions 109
5.6 Future Work 110

Appendices **110**

References **117**

List of Figures

1.1 Projected HIV prevalence in South Africa from 1987 to 2015 (*Source:WHO)* 4

1.2 HIV prevalence in KwaZulu Natal, Western Cape and South Africa from 2001 to 2006 6

1.3 HIV prevalence in South African adults and TB notification rate per 10 000 000 population since early 1980s 11

1.4 The 1000 enumerator areas (EAs) used in the 2005 survey 15

4.1 History: Complete trace plots for HIV fixed effects parameters 92

4.2 Density diagrams for HIV fixed effects parameters 93

4.3 History: Complete trace plots for TB fixed effects parameters 94

4.4 Density diagrams for TB fixed effects parameters 95

5.1 GR statistic for HIV model 115

5.2 GR statistic for TB model 116

List of Tables

1.1 Incidence estimates of TB in South Africa in 2004 9

2.1 Socio-demographic factors 25

2.2 Sexual behavioral factors 26

2.3 Biomedical factors 27

2.4 Substance-use factors 28

2.5 Interaction of HIV and TB 31

2.6 Parameter estimates for HIV 42

2.7 Parameter estimates for TB 43

3.1 GEEs HIV model: Parameter Estimates (Standard Error) 60

3.2 GEEs TB model: Parameter Estimates (Standard Error) 61

3.3 GLMMs HIV model: Parameter Estimates (Standard Error) 62

3.4 GLMMs TB model: Parameter Estimates (Standard Error) 63

4.1 HIV model: Parameter Estimates and Standard Errors 97

4.2 TB model: Parameter Estimates and Standard Error 98

List of Abbreviations

STI	Sexually Transmitted Infection
HIV	Human Immunodeficiency Virus
AIDS	Acquired Immunodeficiency Syndrome
TB	Tuberculosis
MDR-TB	Multi-Drug Resistant TB
XDR-TB	Extreme-Drug Resistant TB
CFR	Case Fatality Rate
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
GEE	Generalized Estimating Equation
LMM	Linear Mixed Model
ML	Maximum Likelihood
IRLS	Iterative Re-weighted Least Squares
OR	Odds Ratio
MCMC	Markov Chain Monte Carlo
M-H	Metropolis-Hastings
ARS	Acceptance-Rejection Sampling
Stats SA	Statistics South Africa
DOH	Department of Health

List of Abbreviations *cont.*

SIR	Sampling Importance Resampling
AdRS	Adaptive-Rejection Sampling
DAG	Directed Acyclic Graph
EA	Enumerator Area
HSRC	Human Sciences Research Council
VP	Visiting Point
CDC	Center for Disease Control
DBS	Dry Blood Specimen

Chapter 1

Introduction

1.1 HIV and AIDS in the World and South Africa

The first case of Acquired Immune Deficiency Syndrome (AIDS) in the USA was identified in 1981, by the Centre for Disease Control (CDC) in 1981. In South Africa the first case of AIDS was reported in 1983 (Ras *et al*. 1983). A Rwandan study carried out in 1984 reported that urban environment, low income and heterosexual promiscuity are the risk factors for AIDS in Africa (van de Perre *et al*. 1984). A study carried out in the Democratic Republic of Congo (former Zaire) in 1984 then reported that infection is only due to heterosexual transmission (Piot *et al*. 1984). Infection levels since the occurrence of these early cases have continued to vary from country to country and even within a country such as South Africa infection levels are highly variable between provinces. There is further high heterogeneity in infection levels even within provinces. The southern part of Africa is the worst affected with HIV than other parts.

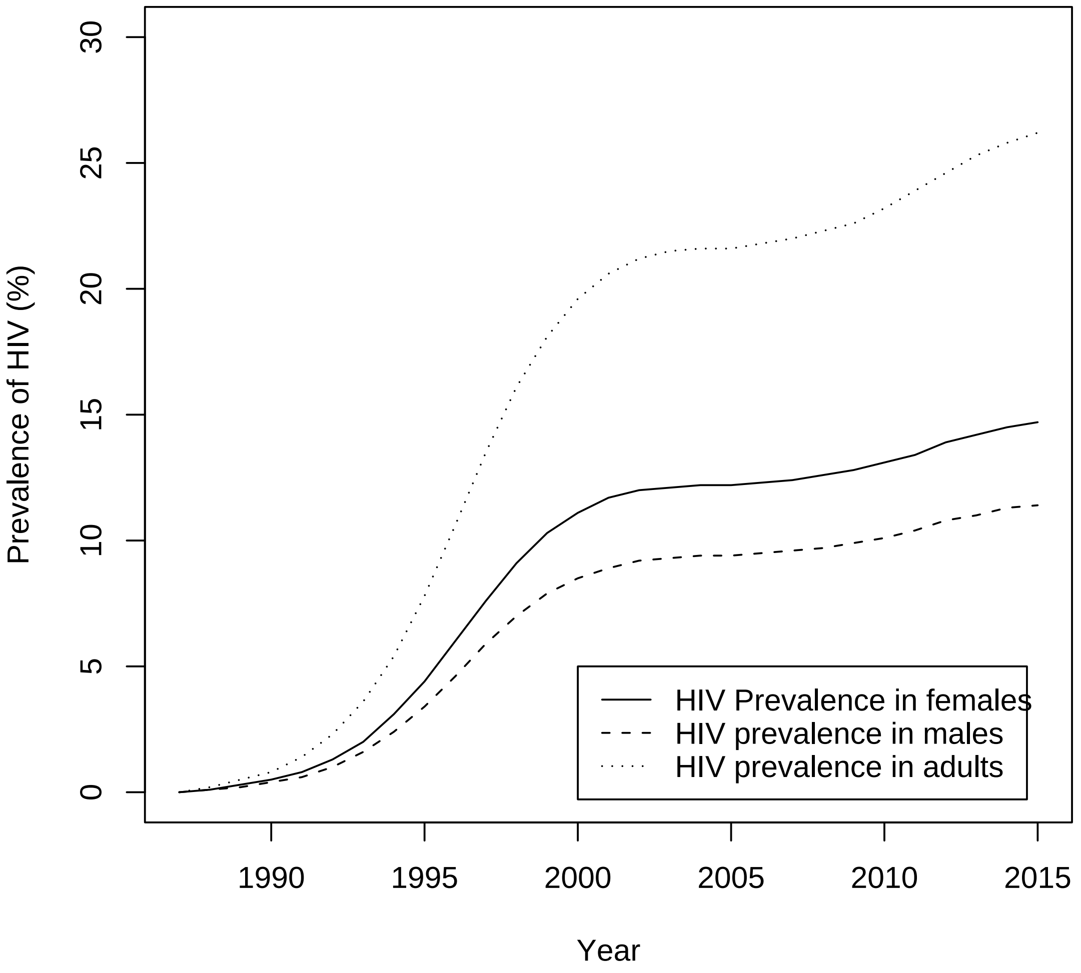
To date South Africa is the country with the highest number of people living with HIV and AIDS in the world (UNAIDS, 2007). The life expectancy of South Africans is estimated to have dropped to 49 years for males and 53 years for females (Dorrington *et al*. 2006). The biggest contributing factor to this drop is HIV and AIDS.

In the 1980s little attention was given to AIDS. Much attention was on apartheid regime which was a huge social and political problem then. The first national antenatal survey was carried out in 1990. Results of the survey found that 0.8% of pregnant women were HIV positive (Department of Health, 1991). Antenatal HIV and syphilis prevalence surveys have been used to monitor HIV prevalence in South Africa. In South Africa only two national household based population surveys have been conducted so far, one in 2002 and the second in 2005 by the Human Sciences Research Council (HSRC). However, sub-national surveys have been conducted, in order to monitor the progress of the epidemic.

In 1991 there was no difference in the prevalence of HIV for heterosexual and homosexual transmission. The Department of Health reported that HIV prevalence had increased by 60% between 1991 and 1992. HIV infection relentlessly spread out within urban areas and then rural areas in early 90s. The rising trend of HIV continued till 2006 (Department of Health, 2007). Prevalence of HIV among women was higher than that for men in the 15-24 age group while it was higher for men than women in the over 45 years age bracket, reaching peaks at ages 25-29 (32.5%) for women and 30-34 for men (26.5%) in 2006, (Dorrington *et al*. 2006). Figure 1.1 shows how HIV varies among adults (15 years and above), males and females in South Africa. Females have a higher probability of being HIV positive than men, Figure 1.1. This trend has been like this since the first case of HIV. This may be explained by many factors both biological and socio-demographic factors related to HIV transmission dynamics which put women at higher risk of infection.

Estimates used to draw Figure 1.1 were taken from the WHO website (accessed on the 12 March 2008) where they also have the similar graph. The estimates are in a form of percentages and there are no estimates for the population groups i.e female and male population. It is not clearly stated on how these estimates where calculated. Therefore the graphs are plotted based on the prevalence against the years. In this way the graphs compare the prevalence for males and females in the same year, say 2005 and the overall prevalence for adults.

Figure 1.1: Projected HIV prevalence in South Africa from 1987 to 2015 (*Source:WHO)*

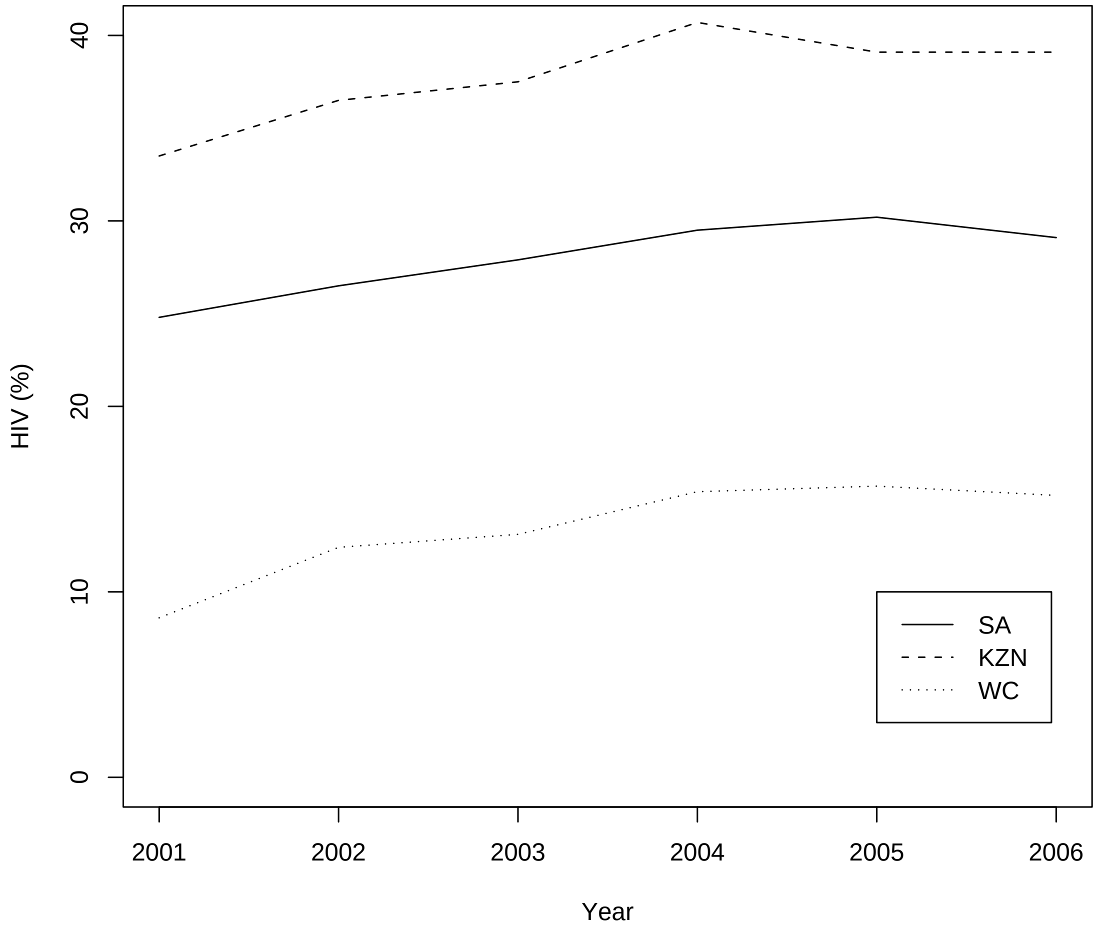


In southern Africa Botswana was the first southern Africa country to begin providing antiretroviral treatment through public sector in 2002 while the government of South Africa made treatment available to the public sector in 2004 after some long criticism and pressure by AIDS activists and the international community[1]. Statistics South Africa (Stats SA) estimated that 4.5 million people were infected with HIV in 2005, whilst the Department of Health estimated the number of people infected as 6.3 million. Only 33% of infected individuals were receiving antiretroviral drugs in 2006. KwaZulu Natal is the province with the highest burden of HIV. The prevalence in the province increased from 1.6% in 1990 to 40.7% in 2004 while Western Cape is the province with the lowest prevalence of HIV. In this province prevalence increased from 0.1% in 1990 to 15.4% in 2004 based on the antenatal data (Department of Health, 2005).

The pattern of HIV in South Africa remained the same even in later years (Department of Health, 2006). This report was based on data from the antenatal clinic attendees. Figure 1.2 shows that prevalence of KwaZulu Natal is consistently higher than the national prevalence in the period 2001 to 2006. There are major causes and determinants of HIV in South Africa. These factors include biological, individual and social factors. Different modes of HIV transmission exist and these include mother-to-child HIV transmission, blood transfusion, exposure to blood and injecting drug use. Some of these like exposure to blood and injecting drug use are not common in South Africa or they could be under-researched and detected.

[1] http://www.aids.about.com/od/clinicaltrials/a/safrica.htm

Figure 1.2: HIV prevalence in KwaZulu Natal, Western Cape and South Africa from 2001 to 2006



1.2 Tuberculosis in the World and South Africa

Tuberculosis (TB) is a contagious airborne disease killing around 1.6 million people each year worldwide. TB is caused by an organism called mycobacterium tuberculosis. These bacteria attack any part of the body but they commonly attack lungs. TB is a common disease in developing countries. Within developing countries TB is more prevalent in the poorer communities. There are two types of TB, namely, latent TB also known as inactive TB and active TB. Latent TB is the term used to describe the cases when the immune system of the host individual is able to stop TB bacteria from causing illness. This type of TB does not induce any symptoms but when the immune system is weakened, factors such as co-infection with HIV and poor nutrition, this may enable the bacteria to begin to multiply and thus lead to active TB (Achmat *et al*. 2005; Santosh, 2007).

The reproductive rate of TB is very high. An individual with active TB can infect an estimated number of 10 to 15 people every year. TB is diagnosed by injecting a protein found in TB bacteria into the skin of an individual's arm which makes the skin to swell when the person is infected with TB. This test is not accurate at diagnosing active TB in people living in areas where TB is very prevalent since it detects both active and latent TB. Active TB can be cured with a combination of antibiotics. A strain of TB that is resistant to two or more first line antibiotic drugs is called multi-drug resistant TB, abbreviated as MDR-TB. A strain of TB that is resistant to three or more second line antibiotics in addition to resisting first line drugs is called extreme drug resistant TB, abbreviated as XDR-TB.

The rural and crowded habitats or areas such as the Karroo towns have the highest incidence of TB in South Africa. These areas are characterized by low income, inadequate housing and inadequate or no medical facilities. The discovery of gold and diamonds in South Africa resulted in high immigration of people from other parts of Africa and Europe who were possibly infected with TB. Conditions in the mining areas favoured the rapid spread of TB. These miners (particularly from Southern Africa) during holidays spread TB within their households and the surrounding neighbourhoods. Urbanisation is another cause for the spread of TB because it is a social phenomenon associated with growing of individuals, high human mobility and density, all of which are conducive for TB transmission from one person to another.

In South Africa all forms of TB are detectable, except the case of XDR-TB which is associated with high case fatality rate (CFR) because of short incubation period making it difficult to contain and mitigate. This deadly strain of TB was recently detected in Tugela Ferry, KwaZulu Natal in 2006 where 53 people were reported to have XDR-TB. It was reported that 52 out of 53 cases died within 25 days of TB being diagnosed and the majority of these were HIV infected[2]. Clearly the high fatality rate could easily be attributable to HIV infection and possibly lack of quick medical intervention. In 2002 Statistics South Africa (StatsSA) reported that Western Cape had the highest incidence of TB. KwaZulu Natal, Eastern Cape, Western Cape and Northern Cape showed incidences of TB that were higher than the national average incidence. Walzl *et al.* (2005) reported that Cape Town, in Western Cape province had one of the highest incidence rates of TB in the world and between 1997 and 2002 the number of new TB cases increased by 22%.

In 2006, World Health Organization (WHO) ranked South Africa fifth among the world's 22 high TB-burden countries. The intermittent and unsuccessful completion of treatment necessitated the initiation of a supervised therapy approach called directly observed treatment short course (known as DOTS). This treatment method cures TB in 85% of the cases and was adopted in South Africa in 1996. This approach was introduced because some people on treatment do not complete their treatment schedule successfully. In 2002 the Medical Research Council (MRC)

[2] http://library.thinkquest.org/tb_in_the world.htm

reported that 1.6% of new TB cases and 6.7% of re-treated cases had MDR-TB[3]. Table 1.1 shows incidence estimates and burden of TB in the nine provinces of South Africa and for the entire country[4] (Weyer *et al*. 2004). Table 1.1 indicates that those provinces with high HIV prevalence tend to have high TB incident rates. This emphasizes the importance of understanding the determinants of co-infection with both HIV and TB. However it is interesting to note that KwaZulu Natal which had the highest percentage of HIV cases had the lowest percentage of new TB cases. This might sound contradictory but since TB status was self reported the results of Table 1.1 need to be interpreted with caution.

Table 1.1: Incidence estimates of TB in South Africa in 2004

Province	**Incidence**	**Total cases**	**% of TB cases**	**% of HIV cases**
Western Cape	1333	58577	2.28	50.4
Northern Cape	822	8033	10.4	52.0
Eastern Cape	1307	102152	1.3	58.8
KwaZulu Natal	1696	173944	0.9	83.4
Limpopo	647	41108	1.6	55.1
Mpumalanga	1052	35977	2.9	77.9
Free State	871	29790	2.9	70.5
Gauteng	1034	85855	1.2	63.6
North West	754	29472	2.6	64.3
SOUTH AFRICA	1084	529320	2.9	66.4

The totals of South Africa are calculated from incidence rounded to nearest full digit, therefore total differ from sum of provincial totals.

[3] http://www.usaid.gov

[4] http://www.health_e.org.za/resources/stats_tuberculosis

1.3 Relationship between HIV and TB

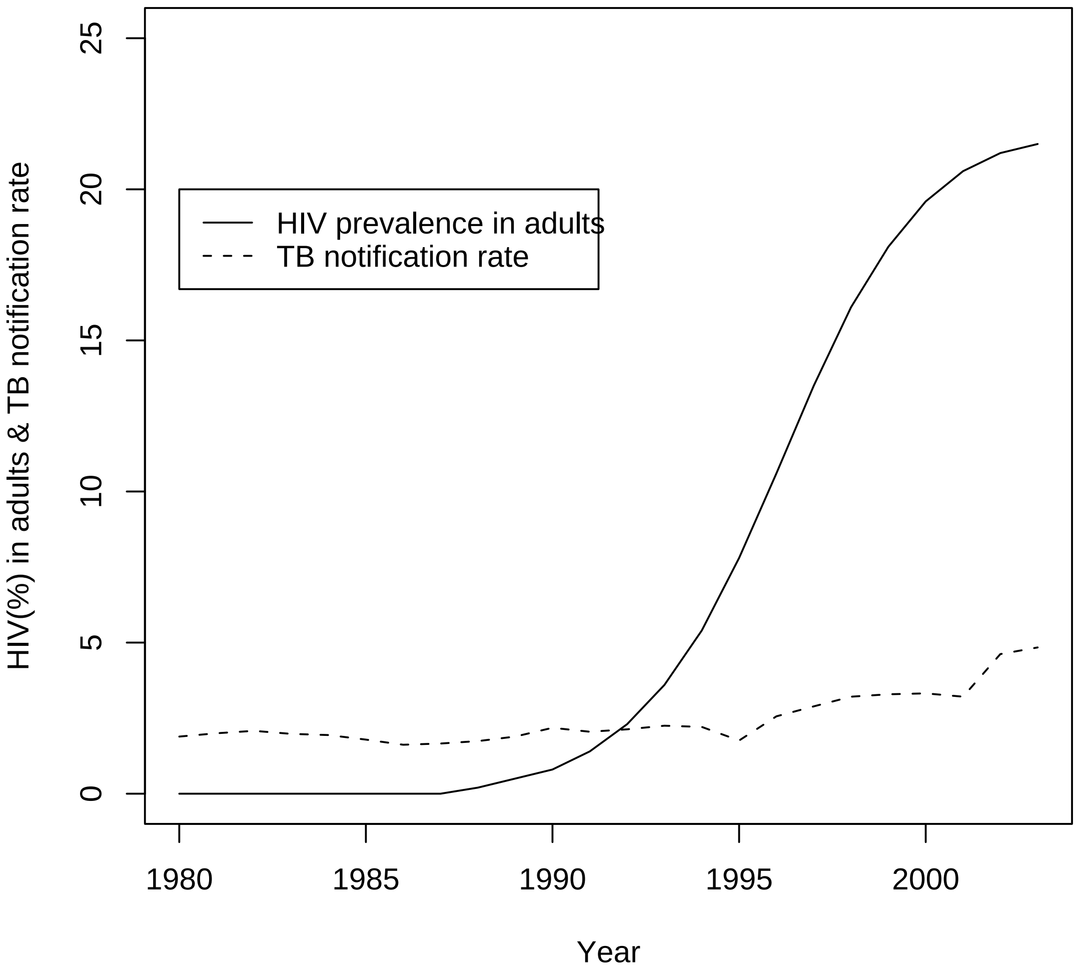
Where HIV and AIDS has had its greatest adverse impact is probably on TB. Although, TB is curable and preventable, it has again emerged as the main cause of morbidity in developing countries. Moreover, people with TB that are infected with HIV are at increased risk of dying from TB (Bucher *et al.* 1999; Corbett *et al.* 2003). The burden of TB in countries with high rates of HIV has increased rapidly over the past decade, especially in the severely affected countries of eastern and southern Africa (Cantwell *et al.* 1996; Wilkinson *et al.* 1997; Churchyard *et al.* 1999; Kenyon *et al.* 1999). Corbett *et al.* (2003) provides a comprehensive review of the relationship between HIV and TB.

TB is the leading cause of death in HIV infected individuals in South Africa. Increase of active TB among people infected with both HIV and TB results in more rapid transmission of TB bacteria, leading to an outbreak of TB *(www.avert.org.za/tuberculosis)*. HIV infection causes immune system impairment which leads to increased risk of reactivation of latent TB infection. Thus HIV is a driving factor that increases the risk of progression from latent TB to active TB (Dlodlo *et al.* 2005). The immune system initially defends the body against TB but co-infection with HIV leads to a weakened immune system which fails to prevent the growth and spread of MDR-TB. Eventually the bacteria starts affecting other areas other than lungs hence increasing the likelihood of extra pulmonary TB (WHO, 2006) and finally death.

Diagnosing TB in HIV infected individuals is very challenging yet it is likely to be fatal if left undetected and thus untreated. Viral load and replication of HIV increase at high rate among those people who also have TB, thus worsening the course of HIV related immunodeficiency (Achmat *et al.* 2005). The incidence of TB is high in countries where HIV is severe. Within South Africa, KwaZulu Natal has the highest cases of patients with TB who are co-infected with HIV, where about

80% of TB patients are also infected with HIV (Gandhi *et al.* 2006). Figure 1.3 show how TB and HIV have been increasing since the early 80s. It is evident that as HIV prevalence began to increase TB notifications also started to increase drastically. An individual infected with HIV may take about 5 years before he becomes infected with TB.

Figure 1.3: HIV prevalence in South African adults and TB notification rate per 10 000 000 population since early 1980s



1.4 Objectives of this Research

In South Africa, no coherent analysis of the determinants of HIV and TB has been conducted at a national level and thus this thesis seeks to mend this gab. The objective of this research is to investigate the risk determinants associated with HIV and those associated with TB with the aim of providing valuable information to aid the design of control and intervention strategies. The project will therefore provide a detailed and better understanding of the risk factors responsible for the spread of the two epidemics in South Africa, viz HIV and TB. The specific objectives of the study are:

- Determine the risk determinants of HIV.
- Determine the risk determinants of TB.
- Identify risk factors that are common to both epidemics.
- Analyse the data in the study to assess the significance of the various potential risk determinants by means of a series of statistical methods in increasing complexity.
- Account for correlated data within a cluster by means of generalized estimating equations.
- Incorporate heterogeneity between geographic areas in modelling by means of random effects models and Bayesian hierarchical modelling approaches.

1.4.1 Research hypothesis

The specific hypotheses that the work will be able to analyse and test are that:

- Socio-economic factors explain most of the spatial variation between the two infections,

- Communities that are predominantly African and disadvantaged socially, have a higher proportion of unemployment, and/or a higher migratory population hence have disproportionately higher prevalence of HIV and TB burdens.
- High population density or poor areas are associated with high burden of HIV and TB infections.

1.5 The Data

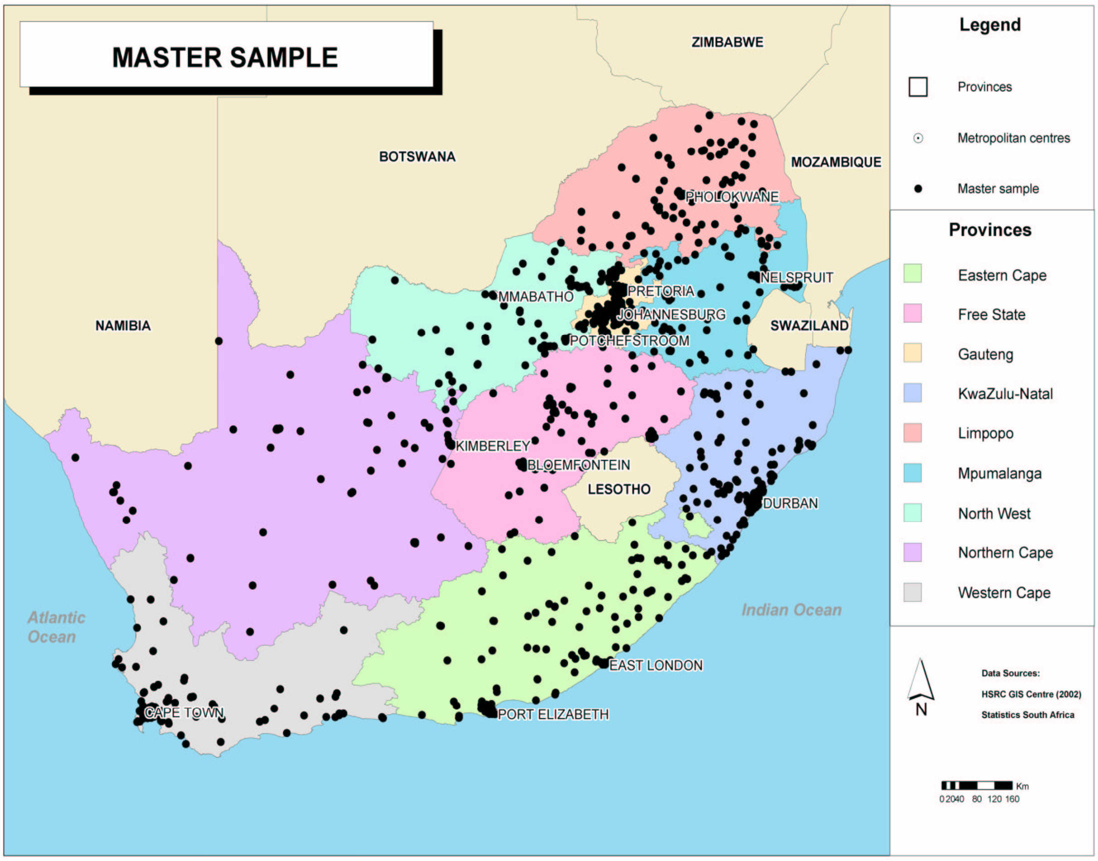
The data that is to be used in this research is a household based second-generation surveillance survey of HIV in South Africa conducted by HSRC in 2005 (Shisana *et al.* 2005). The survey design applied a multi-stage disproportionate, stratified sampling approach. The sampling frame for the 2005 survey as shown in Figure 1.4 was based on the master sample of 1000 enumerator areas used by Stats SA for the 2001 census. An enumerator area (EA) is the spatial area that is used by Stats SA to collect information on the South African population. An EA consists of approximately 180 households in an urban area and 80-120 households in a deep rural area. An EA is considered to be of small enough size for one person to collect census information for Stats SA.

The sample was stratified by province and locality type of the EAs. The locality types were urban formal, urban informal, rural formal (including commercial farms) and rural informal. Formal means those places where houses are built in an ordered manner, whilst informal means the opposite of formal. In urban formal areas, race was also used as a third stratification variable. The master sample allowed for reporting of results at the level of province, type of locality, age and race group.

The data consists of 13 422 households that were visited. Out of these households 10584 households participated in the study. From the participated households, 23 275 people aged two and above completed the interviews and 15 581 were tested for HIV. The thesis will focus on individuals who are 15 years and above totalling 16398. The primary sampling unit was the EA, the secondary sampling unit was the visiting point or household and the ultimate sampling unit was the individual eligible to be selected for the survey. At most, three persons in each household could potentially be selected, with only one from each of the following age groups: 2-14 years, 15-24 years and 25 years and older. Some of the individuals agreed to be tested for HIV. Individuals were asked if they then had any of the following illnesses: TB, hypertension/high blood pressure, diabetes, pneumonia, cancer, malaria and STIs.

Some of the individuals have high or lower chance of selection than others. Sample weights were introduced to correct for the potential bias at the EA, household and individual levels and also to adjust for non-response. The weighting process was as follows: the data file of drawn EAs contained the selection probabilities and sampling weights of these EAs. These weights reflected the disproportionate allocation of EAs according to the stratification variables, i.e. race, locality type and province. The visiting points (VPs) sampling weight was then calculated. This weight was the counted number of VPs in the EA, proportionally corrected for invalid VPs and divided by the number of VPs participating in the survey. The final VP sampling weight was the product of the EA sampling weight and the VP sampling weight.

Figure 1.4: The 1000 enumerator areas (EAs) used in the 2005 survey



1.5.1 Sample Weights

Demographic and HIV testing information on all individuals in all households in all responding EAs was then assembled in order to calculate individual sample weights. In each of the three age groups, the individual weight was the total number of individuals in that age group. Individual sample weights were benchmarked using the mid-year population estimates for 2004 provided by Stats SA. These individual sample weights were also adjusted for HIV testing non-response. In the final step, the information at the individual level was integrated and the final sampling weight for each data record was calculated. This weight was equal to the final VP sampling weights multiplied by the selected individual's sampling weight per VP per age group. This process produced a final sample representative of the population in

South Africa for gender, age, race, locality type and province.

Figure 1.4 shows enumerator areas used in the survey. However, the primary objective of this thesis is based on estimating the risk determinants of HIV and TB. Thus, our analyses will be primarily at the individual level rather than at the aggregated level (community). This is in recognition of the fact that aggregation can lead to loss of information and ecological fallacies, where the relationship between prevalence of disease and community risk factors may be different from the relationship between the same risk factors at the individual level.

1.6 Statistical Methods

Presence and absence of a disease is binary response data. Therefore specific statistical models capable to handle such type of data namely generalized linear models and their extensions if necessary will be used to analyse the data. Since individuals from the same cluster (e.g. households and enumerator area) are likely to be more alike in terms of risk of disease or correlated among each other, generalized estimating equations will be used to correct for this potential intraclass correlation. The modelling work will entail investigating risk factors associated with individual disease occurrence and transmission as well as those that are responsible for the apparent association between the diseases.

Disease occurrence and distribution is highly heterogeneous at the population, household and the individual level. In recognition of this fact we propose to model this heterogeneity at community level through generalized linear mixed models and Bayesian modelling approaches with enumerator area capturing the community effect. Generalized estimating equations will also be used to take into account of the correlation between outcomes within a cluster provided fixed effects represent population averaged effects. Thus in effect the work will end up (comparing) applying both cluster specific and population averaged models but each has its merits and demerits to be discussed in the analysis.

1.6.1 Generalized Linear Models

Generalized Linear Models (GLMs) is an extension of the classical linear models, so that the latter, form a suitable starting point. The standard array of GLMs was initially constructed by Wedderburn and Nelder in the mid 1970s. McCullagh and Nelder published a very detailed book on Generalized Linear Models in 1983. A revised version of their book appeared in 1989 which is now the most referenced book on GLMs. GLMs are based on the exponential family distributions which include the Normal or Gaussian, Inverse Gaussian, Binomial, Poisson and Negative Binomial distributions among others. The log-likelihood form of an exponential family type of distribution is written generally as:

$$\ell = \sum \Big[\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\Big] \tag{1.1}$$

where y is the response, θ is the canonical parameter and ϕ is the dispersion parameter. For these distributions the mean and variance are related and given by

$$\mu = b'(\theta) \tag{1.2}$$

and

$$\mathrm{var}(y) = \phi b''(\theta). \tag{1.3}$$

Since the mean depends only on θ, in a standard GLM the term $c(y, \phi)$ in equation 1.1 can be left unspecified without affecting the likelihood-based estimation of the regression parameters. The variance consists of two terms, the first of which depends only on ϕ, the dispersion parameter and the second is a function of θ, the canonical parameter and hence implicitly the variance also depends on μ. Thus the second term of the variance of y can be expressed as a function of μ, denoted by $V(\mu)$ and it is commonly called the variance function. The variance function defines a distribution in the GLM class of families, if one exists and the function $b(\theta)$ is the cumulant-generating function. The generalization of the systematic part allows the linear predictor to be a monotone function of the mean. One simplifying assumption in the model is often that some function of the mean response varies in a linear way as conditions change in a similar manner as the linear regression model. With n independent units observed, this can be written as a linear predictor. In the simplest case, the canonical parameter is equated to a linear function of other parameters, of the form

$$\theta(\mu) = \mathbf{X}\beta \tag{1.4}$$

where β is the vector of $p<n$ (usually) unknown parameters. The matrix $\mathbf{X}_{n\times p}$ is a set of known explanatory variables called the design or model matrix and $\mathbf{X}\beta$ is the

linear structure. This strictly linear model can be further generalized by allowing other smooth functions of the mean

$$\eta = g(\mu) \tag{1.5}$$

where $g(\mu)$ is called the link function. The mean μ is the inverse link since it can be written as μ=$g^{-1}(\eta)$. Generalized linear models allow two extensions. First the distribution may come from an exponential family and secondly the link function may come from any monotonic differentiable function (McCullagh and Nelder, 1983; Lindsey, 1997; Lee *et al.* 2006). If η=θ, we have the canonical link. Canonical links give rise to simple sufficient statistics, but there is often no reason why they should be particularly appropriate in forming models. Some prefer to use the linear model with normal errors after first transforming the response variable. With GLMs the identification of the mean-variance relationship and the choice of the scale on which the effects are to be measured can be done separately, thus overcoming the shortcomings of the data-transformation approach. GLMs transform the parameters to achieve the linear additivity. In Poisson GLMs for count data, there is no need to transform the data to log y, which is not defined for $y = 0$. In GLMs log(μ)=$\mathbf{X}\beta$ is used which causes no problem when $y = 0$ (Lee *et al.* 2006).

There are many possible choices of link functions, such as the logit, probit, complementary log-log, log, identity, quadratic inverse, exponent, square root and the reciprocal links depending on the type of response. The generalization of the GLMs is that it allows data from non-normal distributions to be modelled in a natural way.

1.6.2 Binary Response

When the response variable is binary, then the interpretation of estimates is different from the normal linear regression where one varies one parameter and keep other parameter(s) constant. For example, the logit link is written as $\ln\left(\frac{p}{1-p}\right)$, where

$\left(\frac{p}{1-p}\right)$ is called the odds of a given event. Therefore the logit is the log odds. The logit link is preferred over other link functions when the response is binary. It is the canonical link for binary data therefore sufficient statistics can be found and conditional likelihoods constructed to analyze the data. Interpretation of parameters in the logistic model is applicable regardless of whether the data were collected in a cohort study or case-control study (Pendergast *et al*. 1996). In the logistic regression approach with one predictor or explanatory variable the log odds has the following linear relationship

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x \tag{1.6}$$

where β_0 is an intercept, x is an explanatory variable and β_1 is the regression parameter corresponding to x. If the sign of β_1 is positive or negative this shows that logit(p) is increasing or decreasing as x increases, respectively. Solving the above equation for p will give us the following logistic function,

$$p = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{1.7}$$

expressing the probability of success as a function of the covariate X. From equation (1.6) we can write

$$\frac{p}{1-p} = \exp(\beta_0 + \beta_1 x) = e^{\beta_0}(e^{\beta_1})^x \tag{1.8}$$

which provides a basis for the interpretation of β_1, (Agresti 1990). Equation (1.8) implies that the odds of an event increases multiplicatively by e^{β_1} for every unit increase in x. Thus β_1 is a measure of the linear effect of x on the log odds of the event of interest. The confidence intervals for odds ratios can be used to assess the significance of the parameters of interest. If confidence intervals contain one this shows that there is no significant association between the event of interest and the explanatory variable x. Therefore one is the null value in hypothesis testing. Equation (1.6) can be generalized to p covariates as

$$\begin{aligned} \text{logit}(p) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \\ &= x'\beta \end{aligned} \tag{1.9}$$

where $x' = (1 \quad x_1 \quad x_2 \ldots x_p)$ and $\beta = (\beta_0 \quad \beta_1 \ldots \beta_p)'$. If we have a random sample $Y_1, Y_2, \cdots, Y_n$ representing n individuals each contributing $x_i' = (1 \quad x_{i1} \quad x_{i2} \ldots x_{ip})$ vector of covariates the equation (1.9) can be generalized to

$$\text{logit}(p_i) = \beta_0 + \beta_{i1} x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{ip} x_{ip} \tag{1.10}$$

$i = 1, 2, \cdots, n$ which can be written more compactly as

$$\text{logit}(p_i)_{n \times 1} = X\beta \tag{1.11}$$

where the ith row of X is x_i' and the matrix X is known as the design matrix for the GLM. Note that Y_i's are assumed to be independent Bernoulli (p_i). For inferences the Wald, likelihood ratio and score tests can be used. These methods give similar results for large samples. The likelihood ratio test uses more information since it incorporates the log-likelihood at the null hypothesis as well as at the estimate of β, so that is why it is more preferred than the Wald test. The Wald test is not as powerful as the likelihood ratio test because it can show deviating behavior when the absolute value of β is large, (Agresti 2002).

The model can be generalized by further considering n clusters each contributing n_i responses $Y_{i1}, Y_{i2}, \ldots, Y_{in_i} \sim Bernoulli(P_i)$ or in the case of longitudinal data where each individual is observed n_i times. In this case it makes sense to account for the correlations structure of observations within the same cluster or observations from the same individuals. Generalized Estimating Equations (GEEs) were introduced by Liang and Zeger (1986) and Zeger and Liang (1986) to account for correlation in GLMs, which introduces the working assumptions of the correlation structure. This is a technique of restricting the analyses to the first moment only. A detailed review of marginal models and other approaches including random effects models and transition models was provided by Diggle *et al.* (1994) and recently by Diggle *et al.* (2002).

1.7 Thesis overview

The thesis is structured as follows: The first part of chapter 2 has the descriptive results and the in second part, the model build up is done where the logistic regression is applied to get the important determinants of HIV and those of TB with the aim of obtaining ball-part estimates. The discussion follows after the model build up. In chapter 3, two models are applied that account for the intra-class correlation expected at an enumerator area level. The methods are the generalized estimating equations and the generalized linear mixed models. In chapter 4 we apply the Bayesian approach in the data. The results obtained are compared with those obtained in the generalized linear mixed models. The results are discussed thoroughly in chapter 5 followed by the conclusions and future research work in chapter 6.

Chapter 2

Exploratory Data Analysis

2.1 Introduction

The initial approach to understanding any data is to do some exploratory analysis using tools such as cross tabulations and graphical displays. Results from such analyses will aid in understanding variables more clearly and their association with the response variable. Our primary objective is to identify potential risk determinants of HIV and TB. We first start by exploring association of HIV status with a list of potential determinants. A similar approach is done for TB. The problem of interaction between HIV and TB will briefly be discussed.

Risk factors of interest have been classified into specific categories. Descriptive results are presented in a tabular form, namely Table 2.1 presents socio-demographic factors, Table 2.2 presents sexual behavioral determinants, Table 2.3 presents biomedical factors and Table 2.4 presents substance use factors.

2.2 Descriptive Analysis of HIV

From this dataset the overall prevalence of HIV is 13.94%. This HIV prevalence varies according to different variables as shown in this section. Results in Table 2.1 show that females are more likely of being HIV positive than males. HIV prevalence estimates for males and females are 11% and 17% respectively. Probability of being infected with HIV also differs by race. Africans are more likely to be infected with HIV than any other race group. Prevalence among Africans is 18%, 0.5% for Whites, 3% for Coloureds and 1.4% for Indians. Whites are the least affected by HIV than any other race. The risk of HIV also varies by age. The prevalence of HIV is highest among those in age group 25 to 34 with prevalence of 24% followed by those in age group 35 to 44 with prevalence of 19%. Those who are in age group 55 years and above are less likely to be infected with HIV.

Individuals were asked to comment on their health status. Individuals who indicated that they are in good health have a lower probability of being HIV positive than those who stated that they are in poor health. The prevalence among those in better health status is 13% whilst the prevalence among those in poor health conditions is 17%. Individuals who reside in urban formal areas have a lower prevalence of HIV compared to other locality types. The prevalence for those who reside in urban informal areas is 24% and is the highest compared to other geotypes. Individuals with no education or primary education have a higher prevalence of HIV than those with secondary and higher education levels. HIV prevalence among those with no or primary education is 15% while the prevalence for those who have tertiary education is only 6%. Those who are unemployed are more likely to be infected with HIV than those who are employed. The prevalence was estimated to be 20% among the unemployed. Exploratory analysis shows that those who spent a night or more away from home are more likely to be infected with HIV than those who did not frequently spend nights away.

On condom use the preliminary exploratory analysis does suggests that individuals who used a condom during their sexual debut are less likely to be HIV positive than those who do not. On the contentious question on whether circumcision is protective against HIV, our descriptive analysis indicates otherwise. On the relationship with STIs the analysis shows that those who had STIs are more likely to be HIV positive than those who did not have STIs. Individuals who had a penetrative sexual intercourse in the past 12 months are likely to be infected with HIV. The prevalence for this group is 15.3% compared to those who practice secondary abstinence (12%).

The prevalence figures for those with and without non-regular partnerships seems fairly close. For the latter the prevalence is 17% and for the former 18%. This was not statistically significant. The data also shows that monthly frequency of sexual intercourse acts may contribute to infection rates. This may not be a surprise because the dominant transmission mode of HIV in our case is heterosexual. The prevalence for those who reported to have one, two and greater than two acts monthly is 17%, 17.3% and 20% respectively. This indicates that the number of acts increases the prevalence of HIV increases also.

Table 2.1: Socio-demographic factors

Parameter	Total	HIV+(%)	TB+(%)	Parameter	Total	HIV+(%)	TB+(%)
Sex of respondent				*Geotype*			
male	6338	10.72	2.55	urban formal	9523	11.51	1.83
female	10057	16.67	1.66	urban informal	1734	23.72	2.55
Race				rural formal	3710	15.25	2.31
African	9664	17.67	2.43	rural informal	1431	12.29	1.93
White	1913	0.52	0.38	*Education*			
Coloured	3013	2.84	1.35	primary	4592	14.54	3.58
Indian	1772	1.35	0.97	secondary	10027	14.85	1.58
Age				tertiary	1585	5.53	0.26
15 to 24	5708	10.28	1.15	*Income*			
25 to 34	2688	23.96	2.14	unemployed	5828	19.91	2.74
35 to 44	2928	18.57	2.70	employed	5168	12.82	1.09
45 to 54	2375	10.06	2.80	other	5148	7.00	2.25
55+	2696	4.03	2.44				
Health							
good	12998	13.22	1.03				
poor	3086	16.56	6.70				
Marital status							
never	7540	16.18	2.00				
ever	8571	11.67	2.13				
Migration							
yes	1563	15.29	2.67				
no	14625	13.62	2.00				

Table 2.2: Sexual behavioral factors

Parameter	Total	HIV+(%)	TB+(%)	Parameter	Total	HIV+(%)	TB+(%)
Condom at sex-debut				*Age sex-debut*			
yes	2552	11.18	1.53	10 to 20	8528	16.02	2.05
no	10470	16.21	2.26	21+	2655	12.48	1.55
Monthly sex				*age difference*			
once	6864	16.74	1.72	0 to 5	8433	14.24	1.91
twice	652	17.30	0.27	6+	1314	15.32	1.36
>2	81	19.68	1.48				
Sex past 12mnths				*Rape*			
virgin	2608	4.30	1.39	yes	323	14.20	4.11
abstinence	3769	12.02	3.28	no	12746	15.44	2.09
sexually active	9313	16.70	1.79				
non-regular partners							
none	8953	16.76	1.76				
one	156	17.67	3.45				
Partners past 12mnths							
one partner	8474	16.24	1.68				
two partners	446	21.33	3.58				
>2 partners	225	19.16	1.66				
Ever had sex							
yes	13206	15.32	2.16				
no	2608	4.30	1.39				

Table 2.3: Biomedical factors

Parameter	Total	HIV+(%)	TB+(%)
STI			
yes	117	29.40	50.34
no	15703	13.79	1.67
Ulcers last 3 months			
yes	268	27.95	2.12
no	13802	14.43	2.00
no resp	1355	7.89	1.99
Circumcision			
yes	228	15.65	3.08
no	13058	13.03	2.15

Table 2.4: Substance-use factors

Parameter	Total	HIV+(%)	TB+(%)
Smoking			
no	15774	13.77	2.06
yes	340	15.97	2.59
Alcohol			
never	1029	12.54	3.29
ever	4068	10.91	2.01
drug injection			
never	15436	13.62	2.10
ever	692	18.54	1.23

2.3 Descriptive analysis of TB

In this section we discuss the prevalence of TB with respect to a number of key factors thought to be associated with the risk of TB. The discussion and analysis presented in this section is based on a cross-tabulation analysis presented in Table 2.1. The overall prevalence of TB in SA is 2.07%. Results show that males are at higher risk of TB infection than females. Prevalence among Africans is 2.4%, 0.4% for Whites, 1.4% for Coloureds and 0.9% for Indians. Risk of TB also differs by age. The older generation is at higher risk of contracting TB compared to the younger people. Individuals in good health are at lower risk of contracting TB than those in poor health conditions. Individuals who never married are at lower risk of TB infection, their TB prevalence is 2% compared to 2.1% for those who have ever been married. Some of these observed associations are not directly discernable but could as well be indirect effects.

Individuals in urban formal settlements are at lower risk of TB infection than those living in other areas such as urban informal, rural informal and rural formal. Those in urban informal settlements are at increased risk of contracting TB. This could be attributed to the crowded living conditions in such locality types. Overcrowding is conducive environment for the spread of airborne diseases. The prevalence for this group is 2.6%, whilst lowest in urban formal areas with prevalence of 1.8%. Individuals with lower level of education are at higher risk of contracting TB bacteria. Prevalence for those with none/primary education is 3.6%, 1.6% for those with secondary education and 0.3% for those with tertiary education. Those who are unemployed are at higher risk of contracting TB. Their prevalence is 2.7% compared to 1.1% for those who are employed. It should be noted however that the category of individuals with somewhat higher prevalence of TB fall in the poverty stricken class thus generalizing the findings in later analyses may be done with caution.

Those who have spend some time away from home are at higher risk of being infected with TB than those who have not. Individuals who are smoking are at higher risk of being infected with TB compared to those who are not smoking. However, these results of different classification of individuals aught to be subjected to a rigorous multivariate statistical analyses in order to properly discern and confirm these associations. This is the subject of the current work.

2.3.1 Interaction of TB and HIV

Table 2.5 shows that the risk of TB infection is higher among individuals infected with HIV compared to those who are HIV negative. This is an important empirical evidence showing the synergy between the two infections. TB prevalence for those who are HIV negative is 1.5% compared to 6.2% for those infected with HIV. On the other hand, risk of HIV infection is higher among individuals infected with TB than among those not infected. HIV prevalence for those who are infected with TB is 39.9% and 13.3% for those who are not infected with TB as shown in Table 2.5.

The causal effect from HIV to TB is well understood because people with HIV have an immune compromised system therefore making it easier for latent TB to re-activate or get infected with new TB. However, the causal effect from TB to HIV is not very well documented and studied. Such a finding may require a cohort type design where individuals with TB and without TB are followed and HIV infections compared among the two groups of individuals. Our descriptive data analysis currently shows that the prevalence of HIV infection is higher among individuals infected with TB than those not infected with TB. Thus these results aught to be taken with caution because also in our case TB status was self reported. Therefore, this may be an underestimate due to latent TB.

Table 2.5: Interaction of HIV and TB

Parameter	Total	Prevalence of TB (%)	Parameter	Total	Prevalence of HIV (%)
HIV status			*TB status*		
negative	10681	1.51	negative	15577	13.34
positive	1351	6.16	positive	283	39.85

2.4 Logistic Regression Modelling

In this section we apply logistic regression modelling to HIV and TB status data separately. A Joint modelling approach is beyond the scope of the current study. We fit a univariate model where the Bernoulli distribution is assumed for individual responses. The logistic regression is among a broad set of models under Generalized Linear Models (GLMs) due to McCullagh and Nelder (1983, 1989). Hosmer and Lemeshow (1989) present an overview of these models. In this case we model observed effects only.

In the current data there are two types of response variables namely an individual's HIV and TB status. Both are binary responses therefore the Bernoulli distribution will be used. Let $y = (y_1, \cdot, y_N)$ and $p = (p_1, \cdot, p_N$ where N is the number of individuals in the study. Then the log-likelihood is given by the following general expression

$$\ell(y_i, p_i) = \sum_{i=1}^{N} \Big\{ y_i \ln\Big(\frac{p_i}{1-p_i}\Big) + n_i \ln(1-p_i) \Big\} \tag{2.1}$$

where $\theta = \ln\left(\frac{p_i}{1-p_i}\right)$, the log prevalence odds of an event of either HIV or TB. The mean and variance of the response variables are given by

$$\mu_i = n_i \frac{e^{\theta_i}}{1+e^{\theta_i}} = n_i p_i \tag{2.2}$$

$$\text{var}(y_i) = n_i \frac{e^{\theta_i}}{1+e^{\theta_i}} \frac{1}{1+e^{\theta_i}} = n_i p_i (1-p_i) \tag{2.3}$$

In this case, n_i=1 because we are modelling individual responses. For the binomial distribution the dispersion parameter is one. Therefore, the variance function is the same as the variance. The link function is the logit link therefore to include covariate dependence the model is given by

$$\text{logit}(p_i) = X\beta \tag{2.4}$$

where the left hand side is a vector of log odds of being HIV or TB positive. The matrix X is given by

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

and vector of parameters β is given by

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

where $n = 16398$ and p is the number of variables (fixed effects) in the model. Univariate models were fitted in order to get variables that fit the model. Statistically significant variables were included in the model. Stepwise model development approach was used. The predictor variables used in HIV status model are listed in Section 2.4.1. The optimization method used to obtain estimates of the parameters is the Fisher's scoring method. These estimates are then used to calculate the odds ratios which are obtained by exponentiating the parameter estimates of the fixed effects. Since parameter estimates are accompanied by standard errors, confidence intervals for the odds ratios were also easily obtained. Multivariate results are given together with univariate results for comparison.

2.4.1 Model results for HIV data

Logistic regression modelling approach was applied in the data. Stratification and other design variables were incorporated using SAS approach SURVEYLOGISTIC to take into account the multistage unequal cluster sample design. The final model for HIV status has the following fixed effects variables:

$$\mathbf{X} = \left\{ \begin{array}{ll} SEX & \text{male, female} \\ RACE & \text{African, White, Coloured, Indian} \\ AGE & \text{15 to 24yrs, 25 to 34yrs, 35 to 44yrs} \\ & \text{45 to 54yrs, 55+ yrs} \\ EDUCATION & \text{primary, secondary, tertiary} \\ HEALTH & \text{good, poor} \\ CONDOM & \text{yes, no} \end{array} \right.$$

ORs are presented in Table 2.6. Sex of the respondent is highly associated with HIV status. The odds of being infected with HIV if you are male are 0.66 times less than the odds of being infected with HIV if you are female. Africans are associated with high prevalence of HIV compared to Indians, OR=3.07. Whites are at lower odds of being infected than Indians. But this is not statistically significant, Table 2.6. Results showed Coloureds are at lower risk of infection than Indians. Race by education interaction effect is also significantly associated with HIV infection. This means that the effect for race on the prevalence of HIV differs at different levels of education. In summary it should be noted that some of the confidence intervals are quite wide hence making the point estimate less reliable. Particularly the race-education interaction odds ratios for HIV seem to be quite wide. The estimates should therefore be interpreted with caution.

The younger generation has higher odds of infection. This is evident from the results since the OR=8.65 [95% CI: 5.23, 8.32] for 25 to 34 compared to older persons. Equivalently, this result implies that individuals in this age group are nine times more likely to be HIV positive compared to those in age group 55+. The odds ratio of HIV infection for those in age group 35 to 44 compared to 55+ are OR=6.13 [95% CI: 3.73, 10.10]. The odds of infection for those in age group 45 to 54 is lower than that for age group 15 to 24, since the odds ratios are respectively OR=2.79 [95% CI: 1.61, 4.86] and OR=4.74 [2.69, 8.32]. Those who are in good health are at lower risk of HIV infection compared to those who are in poor health status, OR=0.678 [95% CI: 0.53, 0.86]. Individuals who used a condom at their sexual debut are at lower risk of HIV infection than those who did not, OR=0.69 [95% CI: 0.51, 0.95].

2.4.2 Model results for TB

The same model development approach as for HIV data was applied to TB data. The final model had the following determinants:

$$\mathbf{X} = \begin{cases} HIVstatus & \text{positive, negative} \\ SEX & \text{male, female} \\ EDUCATION & \text{primary, secondary, tertiary} \\ INCOME & \text{unemployed, employed, other} \\ HEALTH & \text{good, poor} \end{cases}$$

The model was fitted where the response variable was TB status. It should however be noted that for TB, individuals were not tested for but instead TB status was self reported. Thus the results should be interpreted with caution due to possible under-reporting. Preliminary results are tabulated in Table 2.7. Results show that individuals who are not infected with HIV are at lesser risk of being infected with TB than those who have HIV, OR=0.23 [95% CI: 0.15, 0.34]. Results also showed that males are at higher odds of TB infection than females, OR=2.40 [95% CI: 1.66, 3.48]. Risk of TB infection is highly associated with highest achieved education levels where those with higher education qualification are at lower risk of being infected with TB than those with lower education. The risk of TB infection for those with secondary education is higher compared to those with tertiary education, but this is not statistically significant, OR=5.56 [95% CI: 0.87, 3.50].

Unemployed individuals are at higher odds of being infected with TB than those that are employed. This was no statistically significant. Individuals who are in good health are at lower odds of contracting TB bacteria than those with poor health, OR=0.20 [95% CI: 0.13, 0.31].

2.5 Discussion

2.5.1 Human Immune Virus

The descriptive analysis carried out in this chapter strongly indicated that HIV is more prevalent among females than males. There are many possible factors behind this difference. One possible reason is that females are coerced into risky sexual behaviours which expose them to HIV infection. Such behaviour is of course bound to be age dependent. There are other factors which are biological in nature and related to the transmission of the disease which may increase the susceptibility of women to infection more than males. A possible biological reason is that more semen move from a male to a female than otherwise. Furthermore, females have a larger cervical area which makes it easier for HIV to establish itself in females than in males. There are also socio-dynamic forces such as poverty which may entice females to enter into such risky activities, such as sex for money. The increased prevalence in young individuals could be due to the fact that they are more sexually active and inexperienced thus tend to use condoms inconsistently which put them at higher risk of HIV infection. There are a number of reasons attributable to the observation that prevalence is higher among younger than older ages. In Africa, HIV is mostly heterosexually transmitted therefore those in age group 15 to 24 have just entered their sexual active stage and secondly they may tend to use condoms more than older age groups.

Africans are mostly found in rural areas and informal settlements and these places are characterized by low levels of income and poverty leading to mostly young females to look for alternative sources of income such as such as sex for money and also such areas have highly connected social networks making it easier for the disease to spread. Urban formal settlements have lower HIV prevalence than urban informal and other locality types. A possible reason for this observation is that in urban areas there are more awareness campaigns about the scourge. Furthermore it is these people who live in urban formal areas who are more educated hence they tend to be more cautious in their sexual behaviour than those less educated.

Individuals with poor health have a higher HIV prevalence than those with good health. This supports the fact that HIV like any other illness is directly associated with poor health particularly where good nutrition and health care are not accessible. This could be totally different in a situation where the level of care for HIV patients is much more enhanced and individual targeted to ensure maximum benefit. In addition, individuals with HIV are considered to be in poorer health by their infection. Those who most often spent some time away from home (mostly men) have high HIV prevalence than those who do not. The possible reason for this is that such people tend to engage in risky sexual contacts, for example with sex workers who themselves are high risk group. This aspect of migration has been studied by, among others, (Lurie *et al.* 2003; Grosskurth *et al.* 1995, Zuma *et al.* 2003). This has a compounding effect in that the partner left behind might also get infected by the returning partner or get involved with other partners in the absence of her/his regular partner hence increasing the risk of HIV (Lurie *et al.* 2003).

People who are sexually active with multiple partners are more likely to be infected with HIV, though some of them use condoms to minimize the risk of infection. In these results, it is evident that virgins are also at high risk of HIV infection. However, this could be the artefacts of a survey. In Africa, HIV transmission is mostly through heterosexual transmission. Individuals engaged in multiple partnerships are at higher risk of HIV. Inconsistence of condom use will result in an increase on the number of new infections. Individuals who used a condom during their sexual debut are more aware of STIs and thus, tend to protect themselves from being infected with HIV and other STIs. This could be because some of these people have their sexual debut when they are well-matured and know the consequences of unprotected sex. These people may use condoms not to protect themselves from any STI directly but as a contraceptive tool.

Being involved with non-regular partners is an indication of being careless. These non-regular partners include commercial sex workers. These people are most likely to get infected by "their secret lovers". Data show that individuals who experienced sexual abuse and rape have a lower prevalence than those who did not, yet we would have expected otherwise. The reason for this may be attributable to recall bias because of the sensitive nature of the attribute. Exploratory analysis shows that individuals who were infected with STIs had a higher prevalence than those who were not. HIV and STIs are both sexually transmitted and they are known to share same risk factors. Having a partner who has STIs is an indication that the partner is at higher risk of HIV.

Those who reported a penile discharge or ulceration in their genital organs recently, have a higher probability of being infected than those who did not. Again this is because such discharges or ulcerations are signs that an individual had an STI infection which could have enhanced ones susceptibility to HIV. Male circumcision is another proposed intervention strategy that is expected to minimize female to male transmission of HIV. However, this may lead to belief that those who are circumcised will never contract, HIV which is untrue. Thus circumcised men may tend to have multiple sexual partnerships in false belief that they will not become infected because they are circumcised (Peltzer *et al.* 2007). This may thus lead to an adverse impact on HIV transmission. Our results also showed that circumcised men are at higher risk of HIV infection. However, this result may be misleading because in the sample there were very few individuals who were circumcised to be representative of the circumcised population. Furthermore, it is not known whether they were circumcised after infection.

Irresponsible behavior that is caused by alcohol and drug abuse results in high risk of HIV infection.

2.5.2 Tuberculosis

Preliminary results indicate that TB prevalence is higher among males than females. Males tend to work in more TB prone environments than females. For example in the mines of South Africa. This environment where overcrowding is also a common feature is an ideal environment for the spread of TB and other air-borne diseases. Hostels where men often stay are also overcrowded, shafts in mines are poorly ventilated and therefore facilitating easy spread of TB bacteria. Dust in the mine shafts also expose miners to lung diseases which possibly enhances their susceptibility to TB. Migrant mine workers carry the bacteria back home during holidays and spread it to their households and surrounding areas (Trapido *et al.* 1998). TB was previously more prevalent in older individuals but now TB is increasingly becoming prevalent among younger individuals. This is possibly due to the fact that younger individuals are increasingly becoming more susceptible due to co-infections with HIV. TB is one of the opportunistic infections among HIV infected individuals.

People with low level of education are mostly in low income bracket hence they live in rural areas or informal settlements and work in poor working environments such as in factories and mines where there is a lot of pollution. These people also tend to live in crowded environments which are conducive to the spread of TB. Crowded living environments are also the effect of urbanisation where people move to cities in search of work and most of these end up living in crowded informal settlements. This may explain why the prevalence of TB is high in rural, urban informal and in hostels which are an example of crowded environments. TB also spreads rapidly in low income areas with poor health facilities. One reason for this could be that infected individuals harbour the bacteria for longer durations because of lack of treatment therefore infecting more people during their infectious period. Exploratory analysis also suggests that alcohol indulgence is also a risk factor. People who indulge in alcohol tend to be smokers, the analysis also suggests that those who smoke have a higher TB prevalence than those who do not. There are currently few statistical and mathematical models addressing the link between STIs and TB via HIV. This is possibly one area that needs further research.

The association between TB, HIV and other STIs can be explained as follows. STIs are known to put one at high risk of HIV which lowers one's immune response considerably. The low immune response will consequently put one at high risk of endogeneous or exogeneous TB. Thus the association between STIs and TB can be said to be confounded by HIV, while the association between HIV and TB is rather more direct.

The weakness of logistic model is that it ignores potential correlation at an EA level and hence other methods that can handle such correlation are going to be explored in the next chapters. However, logistic regression provided ball point figures for the data.

Table 2.6: Parameter estimates for HIV

	Univariate	Multivariate
Parameter	OR (95%CI)	OR (95% CI)
Sex of respondent		
male	0.596 [0.491-0.722]	0.663 [0.527-0.834]
female	1	1
Age in years		
15 to 24	2.730 [1.695-4.396]	4.739 [2.699-8.321]
25 to 34	7.506 [4.846-11.626]	8.649 [5.229-14.305]
35 to 44	5.432 [3.494-8.446]	6.132 [3.725-10.096]
45 to 54	2.664 [1.612-4.402]	2.797 [1.610-4.858]
55+	1	1
Race group		
African	15.909 [5.884-43.018]	21.432 [2.774-165.505]
White	0.391 [0.115-1.332]	0.782 [0.0623-9.806]
Coloured	2.198 [0.761-6.347]	6×10^{-5} [7.76×10^{-6}-4.7×10^{-4}]
Indian	1	1
Education status		
none/primary	2.908 [1.905-4.438]	7.159 [0.633-80.964]
secondary/matric	3.174 [2.096-4.804]	3.239 [0.3317-31.627]
tertiary	1	1
Health status		
good	0.768 [0.627-0.940]	0.677 [0.531-0.863]
poor	1	1
Condom at sexual debut		
yes	0.651 [0.497-0.851]	0.694 [0.509-0.946]
no	1	1

Table 2.7: Parameter estimates for TB

	Univariate	Multivariate
Parameter	OR (95% CI)	OR (95% CI)
HIV status		
negative	0.233 [0.161-0.333]	0.226 [0.148-0.344]
positive	1	1
Sex of respondent		
male	1.546 [1.117-2.139]	2.402 [1.656-3.484]
female	1	1
Education status		
primary	14.484 [4.291-48.891]	9.233 [1.448-58.854]
secondary	6.242 [1.811-21.522]	5.548 [0.867-35.504]
tertiary	1	1
Income status		
unemployed	1.225 [0.840-1.787]	1.102 [0.645-1.881]
employed	0.477 [0.272-0.837]	0.519 [0.278-0.969]
other	1	1
Health condition		
good	0.145 [0.100-0.209]	0.202 [0.130-0.314]
poor	1	1

Chapter 3

Generalized Estimating Equations and Generalized Linear Mixed Models

3.1 Generalized Estimating Equations

3.1.1 Introduction

Most epidemiological studies deal with non-normal data which necessitates the use of generalized linear models for analysis. In addition these datasets may be highly correlated and standard generalized linear models such as the classical logistic regression models do not account for this intracluster correlation. An extension of GLMs that takes into account this correlation have been developed and one such extension is the generalized estimating equations (GEEs) due to Liang and Zeger (1986). In Section 3.1.2 the theory of the GEEs is briefly discussed.

Since individuals from the same cluster (EA) are likely to be alike in terms of risk of disease, we aim to model HIV and TB status, including correction for such correlation that, if ignored, can lead to incorrect inference. We postulate that observations from the same EA are more correlated than those from a different EA.

The aim is to capture this correlation structure for observations within a cluster in the estimation of parameters of interest in the model. Ignoring this correlation can lead to misleading results or underestimate of the standard errors.

3.1.2 Modelling Correlated Binary Data

Generalized Estimating Equations (GEEs) were developed to extend GLM algorithm to accommodate correlated data that would have otherwise been modelled under independent assumptions in GLMs (Zeger and Liang, 1986). The term GEE indicates that an estimating equation is not the result of a likelihood-based derivation perse, but that it is obtained by generalizing the score equation to a general estimating equation. They are sometimes called marginal models since the random variable is integrated out to get the marginal model for fixed effects only. In GEEs, it is assumed that the mean and variance are characterized as in the GLM and the distribution of the response is not assumed to follow the exponential distribution. The logit link model for the GEE is still the same as for the GLMs but the variance specification is modified.

The score equation and the marginal covariance matrix are given by equation (3.1) and equation (3.2), respectively.

$$S(\hat{\beta}) = \sum_{i=1}^{N} X_i'(A_i^{1/2} R_i(\alpha) A_i^{1/2})^{-1}(y_i - \mu_i) = 0 \tag{3.1}$$

$$V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2} \tag{3.2}$$

where A_i is the main diagonal matrix of (marginal) variance functions ν_{ij}, and $R_i(\alpha)$ is the (marginal) working correlation matrix of the observation vector Y_i indexed by a vector of parameters α. The working correlation matrix is introduced in the variance. In the score equations a small modification is made that accounts for the correlation among variables. The correlation is accounted for by the use of empirical (sandwich/robust) variance estimator. Under mild conditions, the marginal estimates can be found by solving equation (3.1) when the marginal mean is correctly specified. There are different types of working correlation structures that can be used (Verbeke and Molenberghs 2000). Independence correlation structure is used when observations are assumed to be uncorrelated. An exchangeable correlation structure is used when there is no coherent ordering of observations within a cluster, e.g survey data. In this case the correlation between any two observations from the same cluster is constant. It is also called the uniform or compound symmetry covariance structure.

Horton *et al.* (1999) suggested that an exchangeable correlation matrix is preferred when there is no coherent ordering for observations within a cluster such as in survey data. Zeger *et al.* (1986) used an exchangeable working correlation matrix to analyze longitudinal data of mothers' stress on children's morbidity. Orelien *et al.* (2001) suggested that an exchangeable correlation structure should be used for toxicity studies because of clustering that occurs in these studies. Unstructured correlation matrix is preferred when the number of observations per cluster is small in a balanced and complete design (Horton *et al.* 1999). For datasets that have time dependent measurements, namely, longitudinal data, it is advisable to use the M-dependent or autoregressive correlation matrix where the correlation is a function of time between observations. Instead of using a pre-specified working correlation such as the exchangeable, unstructured, auto-regressive or M-independent, etc, one can opt to use the empirical variance which is data driven and gives the so called robust standard errors of estimation.

In GEEs it is required that the data be missing completely at random (MCAR)[1] to give valid inferences (Beunckens *et al.* 2007). If data is missing at random (MAR)[2] it is not advisable to use GEEs instead Weighted GEEs are suitable for this type

[1] MCAR - probability of an observation being missing does not depend on both the observed and the unobserved response

[2] MAR - data is said to be missing at random if the missingness process does not depend on the unobserved responses but may depend on the observed responses

of data. This requires the specification of a dropout model in terms of observed outcomes or covariates (Jansen *et al*. 2006). Each observation is weighted by the inverse probability that an observation drops out or be missing at the same time the participant dropped out. Weighted GEEs are also applicable in the analysis of clustered data where the cluster size is informative. An iterative procedure for estimating parameter estimates is given in Liang *et al*. (1986).

In this work two models are fitted, one for HIV and another for TB. This is done by taking into account the correlation between outcomes within a cluster (EA) assuming an exchangeable correlation structure. We weight the contributions of information from different clusters. In other words we assume the cluster sizes were informative. We used variables that were determined by simple models in Chapter 2.

3.1.3 Results of GEEs

3.1.3.1 HIV Results from GEEs

HIV model was fitted using PROC GENMOD a procedure in SAS. The results show that males are at lower risk of HIV infection than females, Table 3.1. Those in age group 25 to 34 are more likely to be infected with HIV than any other age groups. Africans are at higher risk of HIV compared to other race groups. The results show that those who are in good health are at lower risk of HIV infection than those who are in poor health. Individuals who used a condom during their sexual debut are less likely to be infected with HIV.

3.1.3.2 TB Results from GEEs

The model for TB analysis was also fitted using PROC GENMOD. Table 3.2 shows that individuals who are not infected with HIV are at lower risk of being infected with TB. Results show that males are at higher risk of being infected with TB compared to their females. Individuals with low level of education are more likely to be infected with TB compared to those with tertiary education. Those who are unemployed are more likely to be infected with TB. Individuals who are in good health are at lower risk of being infected with TB.

3.2 Generalized Linear Mixed Models

3.2.1 Introduction

Disease occurrence is highly heterogenous at the population level, household level and the individual level which is why we need to make use of generalized linear mixed models (GLMMs). By controlling for both the observed (fixed) and unobserved (random) risk factors, we will be able to quantify any excess association of HIV and TB at an EA level. The GLMMs are an extension of the GLMs. GLMMs are designed to take account for extra variability in the data which cannot be accounted for by the observed covariates.

3.2.2 Modelling binary response using GLMMs

Estimation in classical GLMs is likelihood based while the GLMMs estimation is based on quasi-likelihood approximation. These models incorporate a random component to account for correlation within cluster or units, heterogeneity and overdispersion from the longitudinal, hierarchical or clustered data structures. A GLMM is a member of a class of statistical models that combine GLMs and ideas from the Linear Mixed Model (LMM) with normal random effects. GLMMs assume the data is from an exponential family of distributions. Linear Models (LMs) characterize fixed effects only apart from model errors. In the GLMM, as in the LMM, the linear predictor can contain random effects such that

$$\eta = X\beta + Z\mathbf{u} \tag{3.3}$$

where β is a vector of fixed effects while $\mathbf{u}$ is the vector of random effects, $\mathbf{X}$ and $\mathbf{Z}$ are design matrices for the fixed effects and random effects, respectively (Wolfinger *et al*. 1993). The conditional mean $\mu|\mathbf{u}$, relates to the linear predictor through a link function

$$g(\mu|\mathbf{u}) = \eta \tag{3.4}$$

The expected value of the random vector Y conditional on $\mathbf{u}$ is given by

$$E(Y|\mathbf{u}) = \mu \tag{3.5}$$

where Y = $(Y_{i1}, \cdots, Y_{in_i})$ is from cluster i and the variance is given by

$$\text{var}(Y|\mathbf{u}) = \phi V(\mu) \tag{3.6}$$

The distribution of the random effects $\mathbf{u}$, is assumed to be $\sim \mathbf{N}(0, G)$. The conditional distribution of the data (given $\mathbf{u}$) is a member of an exponential family. The random effects are commonly assumed to be normally distributed. Estimation of parameters is not that different from that of the GLMs (fixed effects only) except that now the linear predictor includes an extra term representing the random effects. As usual the solution of the model parameters is numerically intensive hence there is a need for better modified methods that can be used to evaluate the solutions.

Overdispersion occurs when data appear to be more dispersed than expected under some reference model. This may also be as a result of the data being grouped in some manner which is not accounted for by the model. Other possible causes may be due to wrong distributional assumptions, missing covariates and/or random effects and positive correlation between observations. Since the mean and variance of the normal data are not related, normal data is not overdispersed as compared to non-normal data. To account for overdispersion, for example in the case of binomial data, we multiply the variance function with a constant, such that, instead of var(y) being $\mu(1-\mu)$ it becomes $\mu(1-\mu)\phi$, where ϕ is the dispersion parameters.

3.2.2.1 Quasi-likelihood

Quasi-likelihood estimation was introduced by Wedderburn (1974). McCullagh and Nelder (1989) discuss quasi-likelihood in detail. Quasi-likelihood can be applied to any response variable whose mean and variance functions can be specified. The quasi-likelihood function is written as $q(\mu_i, y_i)$ where μ_i is a mean for observation

i and y_i is the response variable for observation i. The quasi-likelihood function in vector form $Q(\mu, \mathbf{y})$ satisfies

$$\frac{\partial \mathrm{Q}(\mu, \mathbf{y})}{\partial \mu} = \frac{\mathbf{y} - \mu}{\mathrm{V}(\mu)}. \tag{3.7}$$

where $\mathrm{V}(\mu)$ is the variance function. The estimating equations are determined by

$$\mathrm{Q}(\mu, \mathbf{y}) = \frac{\mathbf{y}\theta - b(\theta)}{a(\phi)} \tag{3.8}$$

where θ is a canonical (natural) parameter, $\mathrm{b}(\theta)$ is the cumulant-generating function and $\mathrm{a}(\phi)$ is the scale parameter, as explained before. The main objective of the quasi-likelihood is that you can apply the theory and methods of GLMs provided $\mathrm{Q}(\mu, \mathbf{y})$ is known. To use the quasi-likelihood function you must know:

- the mean (μ).
- the mean function $\theta(\mu)$ used in the generalized linear model.
- the variance function $\mathrm{Var}(\mu)$.
- the scale parameter $a(\phi)$.

Quasi-likelihood extends to the distributions that are not members of an exponential family. The joint quasi-likelihood in matrix form can be written as, (Littell *et al.* 1996)

$$\mathbf{Q}(\mu, \mathbf{u}; \mathbf{y}) = [\mathbf{y}'\mathbf{A}^{-1}\theta - (\mathbf{b}_\theta^{1/2})'\mathbf{A}^{-1}\mathbf{b}_\theta^{1/2}] + \frac{\mathbf{1}}{\mathbf{2}}\mathbf{u}'\mathbf{G}^{-1}\mathbf{u} \tag{3.9}$$

where

$\mathbf{A}$ is the matrix of $\mathrm{a}(\phi_i)$'s

θ is the vector of $\theta(\mu_i)$'s

$\mathbf{b}_\theta$ is the vector of $\mathrm{b}(\theta_i)$'s

There are two different variance estimators that can be estimated using the quasi-likelihood approach. The first variance estimate is a well known formula using the Hessian matrix. When the mean-variance relationship is specified correctly, this yields efficient estimates. The second variance estimate is from the sandwich formula using the method of moments and this yields robust estimates when the assumption of the correctness of mean-variance relationship is relaxed. The method of quasi-likelihood may be used to analyze overdispersed data without making any distributional assumptions, but only using the specification of the mean and variance. This is similar to using the method of least squares since in both cases the likelihood function is not fully specified.

Method of quasi-likelihood may be conveniently implemented by suitably adapting the Iterated Reweighted Least Squares (IRWLS) procedure used to fit GLMs (Morgan, 1992 p.275). An alternative approach to dealing with overdispersion, which may be suitable for some applications, is to include random effects which extends GLMs to GLMMs (Morgan, 2000). Multiplying the variance function with a constant in GLMs changes the standard errors of the parameter estimates by a proportional amount and does not affect the fixed effects estimates. In GLMMs this changes the fixed effects estimates, random effects solutions, and the covariance parameter estimates and also makes likelihood estimation impossible.

In linear models, correlation between individual responses, is accounted for by the addition of suitable random effects. For some data it is useful to have several independent random effects. These random effects have variances which are called variance components. In GLMs, the linear predictor term η is written as η=$\mathbf{X}\beta$. In GLMMs, the linear predictor is written as shown in equation (3.3), but now with a vector of random effects, the contribution of which is determined by the parameter vector $\mathbf{u}$ (Lee *et al.* 2006). Conditional upon a set of random effects in the vector $\mathbf{u}$, we may write the log-likelihood as $\ell(\beta, \sigma; \mathbf{y}|\mathbf{u})$, where σ is a vector containing variance components. If the random variable $\mathbf{u}$ has the multivariate probability distribution function (pdf) $f(\mathbf{u})$, which is usually assumed to be multivariate normal, then in order to obtain maximum likelihood (ML) estimates for β and σ, it is necessary to formulate the unconditional log-likelihood obtained by integrating out the random effects as,

$$\ell(\beta, \sigma; y) = \int \ldots \int \ell(\beta, \sigma; y|\mathbf{u}) f(\mathbf{u}) d\mathbf{u} \tag{3.10}$$

In some cases it is possible to evaluate this integral explicitly. If the distribution of $\mathbf{Y}|\mathbf{u}$ is not normal, solving the integral for marginal distribution is not readily achievable. This can be solved numerically using PROC NLMIXED, a procedure in SAS. A pseudo-likelihood approach can be used when applying a linear mixed model estimation. This approach is included in PROC GLIMMIX procedure in SAS (Littell *et al.* 2006). Jansen (1993) showed how numerical analysis may be used effectively for approximating the integral above. However, for complex problems the numerical analysis approach is not a practical proposition. For LMMs, parameters are estimated by a combination of Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML). REML is a procedure that yields unbiased estimates of the variance components. Iterated Reweighted Restricted Maximum Likelihood (IRREML) shares the robustness property of quasi-likelihood, in that it does not fully specify distributions (Morgan, 2000). Large covariate effects are obtained under the random-effects model in comparison to the marginal model (Jansen *et al.* 2006).

Solutions for β and $\mathbf{u}$ can be obtained by iteratively solving

$$\begin{bmatrix} X'WX & X'WZ \\ Z'WX & Z'WZ + G^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} X'Wy^* \\ Z'Wy^* \end{bmatrix}$$

where,

$$\begin{aligned}
&W = D^{-1}(R_{\mu}^{1/2} A R_{\mu}^{1/2})^{-1} D \\
&D = \left[\tfrac{\partial \mu}{\partial \eta}\right] = \mathrm{diag}[\phi_{ij}(1-\phi_{ij})] \\
&R_{\mu} = \mathrm{diag}[\pi_{ij}(1-\pi_{ij})] \\
&A = \mathrm{diag}\left[\tfrac{1}{n_{ij}}\right] \qquad (3.11) \\
&y^* = \hat{\eta} + (y - \hat{\mu}) D^{-1} \\
&\hat{\eta} = X\hat{\beta} + Z\hat{\mu} \\
&\hat{\phi} = \tfrac{\exp(X\hat{\beta}+Z\hat{\mu})}{1+\exp(X\hat{\beta}+Z\hat{\mu})}.
\end{aligned}$$

(Breslow and Clayton, 1993). A is the diagonal matrix whose i^{th} diagonal element is $a(\phi)$ the scale parameter, G is the variance-covariance matrix of random effects, i.e. $\mathrm{var}(\mathbf{u}) = G$. The vector y^* is a working response vector variable known as pseudo-data. This pseudo-data is motivated by a Taylor series expansion of the mean in GLMs. Taylor series and pseudo-data methods play an important role in extending the class of GLMs by including random effects. There are other estimation techniques that can be used to estimate the model parameters, such as

- RSPL - Residual (Restricted) Subject-specific Pseudo-likelihood. This is the default method in PROC GLIMMIX.
- RMPL - Residual Marginal Pseudo-likelihood.
- MSPL - Maximum Subject-specific Pseudo-likelihood.
- MMPL-Maximum Marginal Pseudo-likelihood.

The RSPL equals the Restricted Pseudo-likelihood (REPL) and MSPL equals Pseudo-likelihood (PL) if the dispersion parameter estimate (ϕ) is estimated (Wolfinger and O'Connell 1993). MMPL equals Maximum Quasi-likelihood (MQL) if ϕ is not estimated (Breslow and Clayton 1993). Also there are various optimization techniques that can be used in optimizing an objective function. Some of these techniques or algorithms require first-order (gradient evaluation) or second-order (Hessian evaluation) or both derivatives. These techniques are:

- TRUG - True Region Optimization. Requires both first-order and second-order derivatives
- NEWRAP - Newton-Raphson Optimization with Line Search. Requires both first-order and second-order derivatives
- NRRIDG - Newton-Raphson Ridge Optimization.Requires both first-order and second-order derivatives
- QUANEW - Quasi-Newton Optimization. Requires first-order derivative
- DBLDOG - Double Dogleg Optimization. Requires first-order derivative
- CONGRA - Conjugate Gradient Optimization. Requires first-order derivative
- NMSIMP - Nelder-Mead Simplex Optimization. Does not use any derivatives

The primary objective in GLMMs is the predictable function given by $\mathbf{K}'\beta + \mathbf{M}'\mu$. The prediction error variance of this function is

$$Var[\mathbf{K}'\hat{\beta} + \mathbf{M}'(\hat{\mu} - \mathbf{u})] = \mathbf{L}'\mathbf{C}\mathbf{L} \tag{3.12}$$

where

$$\begin{aligned} \mathbf{L}' &= [\mathbf{K}'\mathbf{M}'] \\ \mathbf{C} &= \begin{bmatrix} \mathbf{X}'\mathbf{W}\mathbf{X} & \mathbf{X}'\mathbf{W}\mathbf{Z} \\ \mathbf{Z}'\mathbf{W}\mathbf{X} & \mathbf{Z}'\mathbf{W}\mathbf{Z} + \mathbf{G^{-1}} \end{bmatrix}^{-1} \end{aligned} \tag{3.13}$$

Little research has been done on the small-sample properties of inference. The Wald statistics can be used for test statistics (Littell *et al.* 1996). The standard errors are computed on the logit scale and the output in SAS gives the estimate converted to

μ. This can be converted to standard error using the delta method which involves a Taylor series approximation of g(η). It is given by

$$s.e.[g(\eta)] = \sqrt{\left[\frac{\partial \mathbf{g^{-1}}(\eta)}{\partial \eta}\right]^2 \mathbf{K}'(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{K}} \tag{3.14}$$

when $(X'WX)^{-}$ is a generalized inverse. The standard error for the logit link is

$$\pi(1-\pi)\sqrt{\mathbf{K}'(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{K}} \tag{3.15}$$

(Littell *et al*. 1996, 2006). The GLMMs are cluster specific therefore the interpretation of the results is different from interpreting GEEs estimates of which are population averaged estimates. Estimates are calculated by applying the GLMM solutions to the predictable functions. These estimates are defined on a linked scale. Inverse link (mean) is used to express estimate on the original scale. Several methods can be used to determine the degrees of freedom. In this case, the Satterthwaite method is used to estimate the degrees of freedom which are used to calculate the t-test statistics. Confidence limits for odds ratios apply to the linked scale.

3.2.3 Results of GLMMs

3.2.3.1 GLMMs HIV Results

The variables used are those that were determined by the simple logistic regression model, that is sex, age, race, education, health and condom use during sexual debut. The model was fitted in SAS using PROC GLIMMIX. When this model was fitted it appeared to be over-parameterized, but after excluding the interaction term which is race by education the model appeared to be much more stable. An estimated model constant is the log odds of being infected with HIV for someone in reference categories of other variables. This variable is -5.813 and it gives OR=0.0029 and applying the inverse link gives the probability 0.50 of being infected with HIV in that category. Other variables will be interpreted in relation to the reference category while other variables are kept fixed. The interpretation is within a community level (EA).

In the results tables parameter estimates are given rather than odds ratios and also standard errors which can be used to calculate the confidence interval limits, Table 3.3. These parameter estimates are exponentiated to get the odds ratios. Results showed that males are at lower risk of HIV infection than females with OR=0.71. Those in age group 24 to 34 are at higher risk of HIV infection compared to other age groups within a community, OR=10.80. Also, those in age group 15 to 24, 35 to 44 and 45 to 54 are at higher risk of infection than those in age group 55 years and above, OR=6.2, OR=6.98, OR=2.80, respectively.

Africans are at higher risk of HIV infection compared to Indians within a community, Table 3.3. Whites are at higher risk of HIV infection than Indians. But this is not statistically significant. The odds of infection for Coloureds shows an indication that they are at higher risk of HIV infection compared to Indians, OR=2.45. Individuals with lower level of education are more likely to be infected with HIV than those with tertiary education, OR=2.11. Those who have secondary or matric education are more likely OR=2.044 to be infected with HIV than those with tertiary education. Those in good are less likely to be infected with HIV. Those who used a condom during their sexual debut are at lower risk of being infected with HIV.

3.2.3.2 GLMMs TB results

This model was fitted using variables from the logistic regression model, that is HIV, sex, education, income, and health status. These results are tabulated in Table 3.4. In this table it is evident that those who are HIV positive are most likely to be infected with TB compared to those who are HIV negative, Table 3.4. Males are at higher risk of being infected with TB than females, OR=2.37. Those with primary or no education are at higher risk of being infected with TB, OR=7.84 compared to those with tertiary education. Individuals with secondary or matric education are more likely to be infected with TB, OR=4.01. But this is not statistically significant, since the 95% confidence interval [0.85, 18.83] includes one. Those who are unemployed are at higher risk of being infected with TB compared to those who are employed. Individuals who are in good health are less likely to be infected with TB (OR=0.19).

3.3 Discussion

The two modelling approaches carry different underlying assumptions thus results do not necessarily have to be same. GEEs are population-averaged or marginal models while GLMMs are cluster specific models. In the GEEs the random effects are integrated out to remain we a marginal model for fixed effects only. In GLMMS the fixed effects and random effects are both included in the model. This is because in GEEs we incorporate the correlation that exists among individuals, while in GLMMs we model both the observed and unobserved effects in order to take into account of the heterogeneity that exists within an EA. However, though parameter estimates values are different but conclusions are comparable for both approaches. For GEEs the interpretation is at the population level while for GLMMs the interpretation is cluster specific. It should be noted that the standard errors under GLMMs are smaller than those under GEEs. However, the standard errors from the multivariate GLMM model are generally higher than the simple logistic model since the GLMM accounts for more variability in the data.

Table 3.1: GEEs HIV model: Parameter Estimates (Standard Error)

	Univariate	Multivariate	GEEs
Parameter	Estimate (SE)	Estimate (SE)	Estimate(SE)
Sex of the respondent			
Male	-0.518 (0.098)	-0.411 (0.117)	-0.417 (0.118)
Female	0	0	0
Age of the respondent			
15 to 24	1.004 (0.243)	1.556 (0.287)	1.544 (0.279)
25 to 34	2.016 (0.223)	2.157 (0.257)	2.135 (0.254)
35 to 44	1.692 (0.225)	1.814 (0.254)	1.802 (0.252)
45 to 54	0.979 (0.256)	1.029 (0.282)	1.017 (0.270)
55+	0	0	0
Race			
African	2.753 (0.451)	3.065 (1.043)	2.592 (0.525)
White	-0.958 (0.571)	-0.246 (1.290)	-0.587 (0.640)
Coloured	0.759 (0.486)	-9.714 (1.047)	0.650 (0.555)
Indian	0	0	0
Education			
None/primary	1.066 (0.210)	1.968 (1.237)	0.795 (0.236)
Secondary/matric	1.090 (0.209)	1.175 (1.163)	0.795 (0.231)
Tertiary	0	0	0
Health status			
Good	-0.265 (0.103)	-0.390 (0.124)	-0.389 (0.123)
Poor	0	0	0
Condom use at sexual debut			
yes	-0.429 (0.137)	-0.366 (0.158)	-0.363 (0.163)
no	0	0	0

Table 3.2: GEEs TB model: Parameter Estimates (Standard Error)

	Univariate	Multivariate	GEEs
Parameter	Estimate (SE)	Estimate (SE)	Estimate(SE)
HIV status			
Negative	-1.457 (0.189)	-1.488 (0.216)	-1.485 (0.209)
Positive	0	0	0
Sex of the respondent			
Male	0.436 (0.166)	0.876 (0.189)	0.866 (0.203)
Female	0	0	0
Education			
None/primary	2.673 (0.621)	2.223 (0.945)	2.153 (0.899)
Secondary/matric	1.831 (0.632)	1.714 (0.947)	1.646 (0.897)
Tertiary	0	0	0
Income			
Unemployed	0.203 (0.193)	0.097 (0.273)	0.108 (0.284)
Employed	-0.739 (0.286)	-0.656 (0.319)	-0.656 (0.328)
Other	0	0	0
Health status			
Good	-0.934 (0.188)	-1.597 (0.225)	-1.594 (0.224)
Poor	0	0	0

Table 3.3: GLMMs HIV model: Parameter Estimates (Standard Error)

Parameter	Estimate	Standard Error
Intercept	-5.813	0.404
Sex of the respondent		
Male	-0.346	0.074
Female	0	
Age of respondent		
15 to 24	1.825	0.176
25 to 34	2.379	0.168
35 to 44	1.943	0.165
45 to 54	1.030	0.179
55+	0	
Race		
African	2.063	0.342
White	0.312*	0.468
Coloured	0.894	0.363
Indian	0	
Education		
None/primary	0.749	0.175
Secondary/matric	0.715	0.168
Tertiary	0	
Health status		
Good	-0.645	0.0859
Poor	0	
Condom use at sexual debut		
Yes	-0.359	0.104
No	0	

* p-value>0.05

Table 3.4: GLMMs TB model: Parameter Estimates (Standard Error)

Parameter	Estimate	Standard Error
Intercept	-3.797	0.814
HIV status		
Negative	-1.445	0.172
Positive	0	
Sex of the respondent		
Male	0.863	0.162
Female	0	
Education		
None/primary	2.0593	0.789
Secondary/matric	1.388*	0.789
Tertiary	0	
Income		
Unemployed	0.188*	0.181
Employed	-0.571	0.243
Health status		
Good	-1.678	0.167
Poor	0	

* p-value>0.05

Chapter 4

Bayesian Markov Chain Monte Carlo methods

4.1 Introduction

In previous chapters a frequentist approach has been applied in modelling HIV and TB. These methods include the logistic regression, GEEs and the GLMMs. In this chapter the focus is on Bayesian approach. Unlike frequentist approaches Bayesian methods require the prior information to estimate the posterior distribution. The main challenge with Bayesian approaches is on estimating the posterior distribution which involves integrating high-dimensional functions. Several approaches have been proposed to remedy this problem. These approaches involve simulating realizations from the joint posterior distribution. In this chapter the focus will be on the Markov Chain Monte Carlo (MCMC) methods of simulating data.

In the MCMC approaches one uses the previous sample values to randomly generate the next sample value, generating a Markov Chain. The Gibbs sampler is a MCMC method that is widely applicable to a broad class of Bayesian problems and it has sparked a major increase in the application of Bayesian analysis. The roots of the MCMC methods come from the Metropolis Algorithm (Metropolis and Ulam

1949; Metropolis *et al.* 1953) attempted by physicists to compute complex integrals by expressing them as expectations for some distribution and then estimate this expectation by drawing samples from that distribution. The Gibbs sampler (Geman and Geman 1984) has its origins in image processing.

The Markov Chains are defined as follows: Let X_t denote the value of a random variable at time t and let the state space refer to the range of possible X values. The random variable is a Markov process if the transition probabilities between different values in the state space depend only on the random variable's current state, that is

$$\Pr(X_{t+1} = s_j | X_0 = s_k, \cdots, X_t = s_i) = \Pr(X_{t+1} = s_j | X_t = s_i). \tag{4.1}$$

The only information needed to predict the future state is the current state. A Markov chain refers to a sequence of random variables $(X_0, \cdots, X_n)$ generated by a Markov chain process. A particular chain is defined most particulary by its transition probabilities, which is the probability that a process at state space s_i moves to state s_j in a single step,

$$P(i, j) = P(i \to j) = \Pr(X_{t+1} = s_j | X_t = s_i). \tag{4.2}$$

Let $\Pi(t)$ denote the row vector of the state space probabilities at step t. The probability that the chain has state value s_i at time t+1 is given by the Chapman-Kolomogorov equation, which sums over the probability of being in a particular state at the current step and the transition probability from that state into state s_i. Successive iteration of the Chapman-Kolomogorov equation describes the evolution of the chain. Defining the n-step transition probability $p_{ij}^{(n)}$ as the probability that the process is in state j given that it started in state i for n steps ago, that is,

$$p_{ij}^{(n)} = \Pr(X_{t+n} = s_j | X_t = s_i). \tag{4.3}$$

A Markov chain is said to be irreducible if there exists a positive integer such that $p_{ij}^{(n_{ij})} > 0$ for all i, j. That is, all states communicate with each other, as one can always go from any state to any other state. Also, a chain is aperiodic when the number of steps required to move between two states is not required to be multiple of some integer. A Markov chain may reach a stationary distribution π^*, where the vector of probabilities of being in any particular given state is independent of the initial condition. A sufficient condition for a unique stationary distribution is that the detailed balanced equation holds, that is

$$P(i,j)\pi_i^* = P(j,i)\pi_j^* \tag{4.4}$$

(Walsh *et al.* 2004). If equation (4.4) holds for all (i,j), the Markov Chain is said to be reversible, thus equation (4.4) is called the reversibility condition. This is the same as stating that at equilibrium, the stationary (invariant) distribution satisfies the following expression

$$\pi^*(dy) = \int \pi^*(x)P(x,dy)dy \tag{4.5}$$

4.2 Direct Sampling Methods

Direct sampling methods are suitable for low-dimensional parameter space and they become less efficient as the dimensional space starts to increase. Direct sampling methods that are usually used by many researchers include; Acceptance-Rejection Sampling, Adaptive-Rejection Sampling and Sampling Importance Resampling. These methods draw random samples from a candidate density and reshaping it to accept some of the values into the final sample. The direct non-iterative sampling techniques involve two steps of obtaining a sample from a posterior distribution which may only be known up to its unscaled form. The first step samples random variables from a candidate distribution and in the second step the methods adjust the sample to approximate the posterior distribution. Samples that are generated by these methods are statistically independent. But this is sometimes not true when the variance is introduced as a reduction tool (Chib *et al.* 1995).

4.2.1 Acceptance-Rejection Sampling

An Acceptance-Rejection (A-R)Sampling method (Tierney 1994) is one of the very important direct sampling methods. The objective is to generate samples from a truly continuous target density $\pi(x) = f(x)/K$ where $x \epsilon \Re^d$, $f(x)$ is the unnormalized density and K is the (possibly unknown) normalizing constant. Let $h(x)$ be a density that can be simulated by some known method and suppose there is a known constant c such that $f(x) \leq ch(x)$ for all x. Now, to obtain a random variate from $\pi(\cdot)$,

1. Generate a candidate z from $h(\cdot)$ and a value u from $U(0,1)$, the uniform distribution on (0,1).
2. If $u \leq f(z)/ch(z)$, return $z = y$.
3. Else go to Step 1.

The accepted value y is a random variate from $\pi(\cdot)$. For this method to be efficient, c must be carefully selected. Since the anticipated number of iterations of steps 1 and 2 to obtain a draw is given by c^{-1}, the rejection method is optimized by setting

$$c = \sup_x \tfrac{f(x)}{h(x)}.$$

Even this choice may result in an undesirably large number of rejections.

4.3 Markov Chain Monte Carlo

As stated above, the direct methods are inefficient in high dimensional parameter space. MCMC methods act as a remedial tool for high dimensional parameter space. In the MCMC methods, the stationary density is known and is given by $\pi(\cdot)$ which is the target density from which samples are desired and the target kernel (posterior distribution) is unknown. The samples are generated from $\pi(\cdot)$. MCMC methods find and utilize a transition kernel $P(x, y)$ whose nth iterate converges to $\pi(\cdot)$ for large n. The process starts at an arbitrary x and is iterated for a number of times.

After this large number the distribution of the observations generated from the simulation is approximately the target distribution. The most important part is to find $P(x,y)$. Suppose that the transition kernel for some function $p(x,y)$ can be expressed as,

$$P(x,y) = p(x,y)dy + r(x)\delta_x(dy) \tag{4.6}$$

where $p(x,x)$=0, $\delta_x(dy)$=1 if $x\epsilon dy$ and 0 otherwise. $r(x) = 1 - \int_{\Re^d} p(x,y)dy$ is the probability that the chain remains at x. The integral of $p(x,y)$ over y is not necessarily 1 since $r(x) \neq 0$. If $p(x,y)$ in equation (4.6) satisfies the reversibility condition, that is

$$\pi(x)p(x,y) = \pi(y)p(y,x) \tag{4.7}$$

then $\pi(\cdot)$ is the invariant density of $P(x,\cdot)$ (Tierney 1994). To verify this we evaluate the right-hand side of equation (4.5):

$$\begin{aligned}
\smallint P(x,\cdot)\pi(x)dx &= \int \Big[\int_A p(x,y)dy \Big] \pi(x)dx + \int r(x)\delta_x(A)\pi(x)dx \\
&= \int_A \Big[\int p(x,y)\pi(x)dx \Big] dy + \int_A r(x)\pi(x)dx \\
&= \int_A \Big[\int p(y,x)\pi(y)dx \Big] dy + \int_A r(x)\pi(x)dx \\
&= \int_A (1 - r(y))\pi(y)dy + \int_A r(x)\pi(x)dx \\
&= \int_A \pi(y)dy
\end{aligned} \tag{4.8}$$

The left-hand side of the reversibility condition is the unconditional probability of moving from x to y, where x is generated from $\pi(\cdot)$. The right-hand side is the unconditional probability of moving from y to x, where y is generated from $\pi(\cdot)$. The reversibility condition states that the two sides are equal. This result gives us a sufficient condition that must be satisfied by $p(x,y)$. A Metropolis-Hastings uses this condition to find $p(x, y)$ which is discussed in the next section (Walsh *et al.* 2004).

4.3.1 The Metropolis-Hastings Algorithm

The common difficulty that arises with the Monte Carlo integration is that of obtaining samples from some complex probability distribution $p(x)$. Attempts to solve this problem are the roots of MCMC methods. These methods trace attempts by mathematical physicists to integrate very complex functions by random sampling and the resulting Metropolis-Hastings algorithm. Suppose the goal is to draw samples from some distribution $p(\theta)$ where $p(\theta) = f(\theta)/K$, where the normalizing constant K may not be known and very difficult to compute (Walsh *et al.* 2004). A sequence of draws is generated from this distribution using the *Metropolis algorithm* (Metropolis and Ulam 1949, Metropolis *et al.* 1953) as follows:

1. Start with any initial value θ_0 satisfying $f(\theta_0) > 0$.

2. Using current θ, sample a *candidate point* θ^* from a distribution which is a probability of returning a value of θ_2 given a previous value of θ_1. This distribution is also kown as the *jumping distribution* denoted by $q(\theta_1, \theta_2)$. It is also referred to as the *proposal* or *candidate-generating distribution.* The candidate-generating distribution in the Metropolis algorithm is symmetric, i.e $q(\theta_1, \theta_2) = q(\theta_2, \theta_1)$.

3. Given the candidate point θ^*, calculate the ratio of the density at the candidate (θ^*) and current (θ_{t-1}) points,

$$\alpha = \frac{p(\theta^*)}{p(\theta_{t-1})} = \frac{f(\theta^*)}{f(\theta_{t-1})}$$

4. If the jump increases the density ($\alpha > 1$), accept the candidate point, i.e set $\theta_t = \theta^*$ and return to step 2. If $\alpha < 1$, then with probability α accept the candidate point, else reject it and return to step 2.

The Metropolis sampling can be summarized as follows. First compute

$$\alpha = \min\left(\frac{f(\theta^*)}{f(\theta_{t-1})}, 1\right)$$

and then accept a candidate point with probability α, i.e the probability of a move. This generates a Markov chain $(\theta_0, \theta_1, \ldots, \theta_k, \ldots)$. The transition probabilities from θ_t to θ_{t+1}, depends only on θ_t and not $(\theta_0, \ldots, \theta_{t-1})$. Following a *burn-in period*, the chain approaches its stationary distribution and samples from the vector $(\theta_{k+1}, \ldots, \theta_{k+n})$ are samples from $p(x)$. This Metropolis algorithm is generalized by using an arbitrary transition probability function $q(\theta_1, \theta_2) = Pr(\theta_1 \to \theta_2)$ and setting the acceptance probability for a candidate point as

$$\alpha = \min\left(\frac{f(\theta^*)q(\theta^*, \theta_{t-1})}{f(\theta_{t-1})q(\theta_{t-1}, \theta^*)}, 1\right) \tag{4.9}$$

This is the *Metropolis-Hastings algorithm*, when assuming that the proposal distribution is symmetric recovers the original Metropolis algorithm.

The Metropolis-Hastings (M-H) updating scheme was first described by Hastings (1970) as a generalization of the Metropolis algorithm. Given a partition of the state vector into components, $x = (x_1, \ldots, x_k)$ and that we wish to update the ith component, the M-H update proceeds as follows: We begin with a candidate-generating distribution, $q(x, y)$, which is used to generate candidate observations y such that $y_i = x_i$. In this case we use x instead of θ_1 and y instead of θ_2 as we did in the Metropolis algorithm definition. Now having generated a new state $y = (x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_k)$, from density $q(x, y)$, we then accept this point as the new state of the chain with probability $\alpha(x, y)$ given by

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\} \tag{4.10}$$

But if the proposed point is rejected, the chain remains in the current state (Brooks *et al* 1998). There may be many acceptance functions which provide a chain with the desired properties as the form of the acceptance probability is not unique. This form is optimal in that suitable candidates are rejected least often and so statistical efficiency is maximized (Peskun, 1973). The resulting transition kernel $P(x,\cdot)$ which denotes the probability that the next state of the chain lies within some set A, given that the chain is currently in state x, is given by

$$P(x,\cdot) = \int_A K(x,y)dy + r(x)I_A(x) \tag{4.11}$$

where $K(x,y) = q(x,y)\alpha(x,y)$ is the density associated with selecting a point which is accepted and $r(x) = 1 - \int_E q(x,y)\alpha(x,y)dy$ is the size of the point mass associated with a rejection. I_A is the indicator function and K satisfies the reversibility condition which also implies that the kernel P also preserves detailed balance for π (Brooks *et. al.* 1998). The function $\pi(\cdot)$ only enters through α and the ratio $\pi(y)/\pi(x)$, therefore knowledge up to a constant of proportionality is sufficient for implementation. In the case where the candidate generating function is symmetric, that is $q(x,y) = q(y,x)$ the acceptance function reduces to

$$\alpha(x,y) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\} \tag{4.12}$$

This is the special case of the original Metropolis update.

The Metropolis-Hastings (M-H) algorithm is specified by its candidate-generating density $q(x,y)$. Secondly if the candidate value is rejected the current value is taken as the next item in the sequence. Thirdly, the calculation of $\alpha(x,y)$ does not require knowledge of the normalizing constant of $\pi(\cdot)$ because it appears both in the numerator and denominator. If a candidate-generating density is symmetric, an important special case, $q(x,y) = q(y,x)$ and the probability of move reduces to $\pi(y)/\pi(x)$ hence if $\pi(y) \geq \pi(x)$ the chain moves to y; otherwise, it moves with probability given by $\pi(y)/\pi(x)$. If the jump goes uphill it is always accepted; if downhill it is accepted with nonzero probability.

In summary the M-H algorithm proceeds as follows, initialized with an arbitrary value $x^{(0)}$:

1. Repeat for $j = 1, 2, \ldots, N$.

2. Generate y from $q(x^{(j)}, \cdot)$ and u from $U(0, 1)$.

3. If $u \leq \alpha(x^{(j)}, y)$, set $x^{(j+1)} = y$.

4. Else, set $x^{(j+1)} = x^{(j)}$.

5. Return the values $\{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$.

The draws are treated as a sample from the target density $\pi(x)$ only after the chain has passed the transient stage and the effect of the fixed starting value has become redundant that it can be ignored. This happens only if irreducibility and aperiodicity conditions are met. This means that if x and y are in domain of $\pi(\cdot)$, it must be possible to move from x to y in a finite number of iterations with nonzero probability and the number of moves required to move from x and y is not required to be a multiple of some integer.

It is necessary to show that the M-H sampling generates a Markov chain whose equilibrium density is the candidate density $p(x)$. Therefore, it is sufficient to show that the M-H kernel satisfy the detailed balance equation (4.4) with $p(x)$. Under the M-H algorithm, we sample from $q(x, y) = \Pr(x \rightarrow y|q)$ and accept the move with probability $\alpha(x, y)$, so that the transition probability kernel is given by

$$\Pr(x \rightarrow y) = q(x, y)\alpha(x, y) = q(x, y) \times \min\left[\frac{p(y)q(y, x)}{p(x)q(x, y)}, 1\right] \tag{4.13}$$

If the M-H kernel satisfies $P(x \rightarrow y)p(x) = P(y \rightarrow x)p(y)$ then the stationary distribution from this kernel corresponds to draws from the target distribution. We show that the balance equation is indeed satisfied with this kernel by considering the three possible cases of any particular x, y pair.

1. $q(x, y)p(x) = q(y, x)p(y)$. Here $\alpha(x, y) = \alpha(y, x) = 1$,
   $\Rightarrow P(x, y)p(x) = q(x, y)p(x)$ and $P(y, x)p(y) = q(y, x)p(y)$
   hence $P(x, y)p(x) = P(y, x)p(y)$, showing that equation (4.4) holds.

2. $q(x, y)p(x) > q(y, x)p(y)$, in which case $\alpha(x, y) = \frac{p(y)q(y,x)}{p(x)q(x,y)}$ and $\alpha(y, x) = 1$.
   Hence

$$\begin{aligned} P(x,y)p(x) &= q(x,y)\alpha(x,y)p(x) \\ &= q(x,y)\frac{p(y)q(y,x)}{p(x)q(x,y)}p(x) \\ &= q(y,x)p(y) = q(y,x)\alpha(y,x)p(y) \\ &= P(y,x)p(y) \end{aligned}$$

3. $q(x,y)p(x) < q(y,x)p(y)$. Here $\alpha(x,y) = 1$ and $\alpha(y,x) = \frac{q(x,y)p(x)}{q(y,x)p(y)}$.

   Hence

$$\begin{aligned} P(y,x)p(y) &= q(y,x)\alpha(y,x)p(y) \\ &= q(y,x)\frac{q(x,y)p(x)}{q(y,x)p(y)}p(y) \\ &= q(x,y)p(x) = q(x,y)\alpha(x,y)p(x) \\ &= P(x,y)p(x) \end{aligned}$$

A key issue in the successful implementation of M-H or any other MCMC sampler is the number of runs until the chain approaches stationarity, i.e the length of the burn-in period. Typically the first 1000 to 5000 realizations are thrown out and one of the various convergence tests is used to assess whether stationarity has been reached. A poor choice of starting values and/or proposal distribution can greatly increase the required burn-in-time. The area of much importance is whether an optimal starting point and proposal distribution can be found. One suggestion for a starting value is to start the chain as close to the centre as possible, for example taking a value close to the distribution's mode, such as using an approximate Maximum Likelihood Estimate as the starting value.

A chain is said to be *poorly mixing* if it stays in small regions of the parameter space for long period of time, as opposed to a *well mixing* chain that seems to happily explore the space. A poorly mixing chain can arise because the target distribution is multimodal and the choice of starting values traps us near one of the modes. Such multimodal posteriors can arise if we have a strong prior in conflict with the observed data. Two approaches have been suggested for situations where the target distribution may have multiple peaks. The most straightforward is to use multiple highly dispersed initial values to start several different chains (Gelman and Rubin 1992). A less obvious approach is to use *simulated annealing* (Kirkpatrick *et al.* 1983, Cerny *et al.* 1985) on a single-chain.

The Metropolis sampler works with any symmetric distribution and the M-H algorithm is more general and now we discuss our best options for the candidate-generating distributions. Different approaches have been proposed to ease the selection of this distribution. These approaches are random walks and independent chain sampling. Based on the random walk chain using the candidate-generating distribution, the new value y equals the current value x plus a random variable w. In this case, $q(x, y) = g(y - x) = g(w)$, the density associated with the random variable w. If $g(w) = g(-w)$, i.e the density for the random variable w is symmetric (as it occurs with a normal or multivariate normal with mean zero or a uniform centered around zero), then we can use Metropolis sampling as $\frac{q(x,y)}{q(y,x)} = \frac{g(w)}{g(-w)} = 1$. The variance of the proposal distribution can be thought of as a *tuning parameter* that can be adjusted to get better mixing.

Using an independent chain, the probability of jumping to point y is independent of the current position, x, of the chain $g(x, y) = g(y)$. Therefore, the candidate value is simply drawn from a distribution of interest, independent of the current value. Also, any number of standard distributions can be used for $g(y)$. Also very important, in this case the candidate-generating distribution is generally not symmetric, as $g(x)$ is not necessarily equal to $g(y)$ and the M-H sampling must be used. As mentioned the candidate-generating distribution can be tuned to adjust the mixing and in particular the acceptance probability of the chain. This is generally done by adjusting the standard deviation of the candidate-generating distribution.

An acceptance probability can be increased by decreasing the standard deviation of the candidate-generating distribution (Draper 2000). This can be further explained that, if the standard deviation is too large, moves are large which is an advantage, but are not accepted often which is a disadvantage. This also leads to high autocorrelation and very poor mixing which requires much longer chains. If the proposal standard deviation is too small, moves are generally accepted therefore there is high acceptance probability, but they are also small, again generating high autocorrelations and poor mixing.

4.3.1.1 Convergence Diagnostics

Given that samples from M-H algorithm are most likely to be correlated, the question will be how does this affect use of the sequence for estimating parameters of interest from the distribution? It is expected that adjacent samples from a M-H sequence to be positively correlated and the nature of this correlation can be quantified by using an autocorrelation function. Consider a sequence $(\theta_1, \dots, \theta_n)$ of length n. Correlations can occur between adjacent sampled values. The kth order autocorrelation ρ_k can be estimated by

$$\begin{aligned} \widehat{\rho}_k &= \frac{Cov(\theta_t, \theta_{t+k})}{Var(\theta_t)} \\ &= \frac{\sum_{t=1}^{n-k}(\theta_t - \bar{\theta})(\theta_{t+k} - \bar{\theta})}{\sum_{t=1}^{n-k}(\theta_t - \bar{\theta})^2} \end{aligned} \tag{4.14}$$

where,

$$\bar{\theta} = \frac{1}{n}\sum_{t=1}^{n}\theta_t.$$

An important result from the theory of time series analysis is that if θ_t are from a stationary and correlated process, correlated draws still provide an unbiased picture of the distribution provided the sample size is sufficiently large. Some indication of the required sample size comes from the theory of a first-order autoregressive process (AR_1), which is of the form

$$\theta_t = \mu + \alpha(\theta_{t-1} - \mu) + \epsilon \tag{4.15}$$

where ϵ is *white noise* which is normally distributed with mean zero and variance σ^2. In this case $\rho_1 = \alpha$ and the kth order autocorrelation is given by $\rho_k = \rho_1^k$. Under this process, $E(\bar{\theta}) = \mu$ with standard error

$$SE(\bar{\theta}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{1+\rho}{1-\rho}}. \tag{4.16}$$

In the above equation the first ratio is the standard error for the white noise and the second ratio is the *sample size inflation factor* which shows how autocorrelation inflates the sampling variance. Suppose, for ρ=0.5, 0.75, 0.9, 0.95 and 0.99, the associated sample size inflation factor are 3, 7, 19, 39 and 199, respectively. Therefore with an autocorrelation of 0.95 which is common in a M-H sequence, roughly forty times as many points are required for the same precision as with an uncorrelated sequence. A possible modification for reducing autocorrelation is *thinning* the output storing only every mth realization after the burn-in period. Suppose a M-H sequence follows an AR_1 model with $\rho_1 = 0.99$. In this case, sampling every 50, 100 and 500 points gives the correlation between the thinned samples as $0.605 (= 0.99^{50})$, 0.366 and 0.007, respectively. In addition to reducing autocorrelation, thinning the sequence also saves computer memory. (Walsh *et al.* 2004).

On testing convergence one should always look at the time series trace, the plot of the random variable(s) being generated versus the number of iterations. In addition to showing poor mixing, such traces can also suggest a minimum burn-in period for some starting value. Suppose the trace moves slowly away from the initial value to a rather different value, say 5000 iterations, around which appears to settle down. In fact, the burn-in period is at least 5000 in this case. The actual time may be far longer than that suggested by the trace. But, the trace often indicates that the burn-in is not complete. Also a plot of α_k against k, i.e kth order autocorrelation against the lag, should show a geometric decay if the sampler series closely follows an AR_1. A plot of partial autocorrelations as a function of lag is very useful. The kth partial autocorrelation is the excess correlation not accounted for by a k-1 order autoregressive model (AR_{k-1}). If the first order model fits, the second order partial autocorrelation is zero, as the lagged autocorrelations are completed accounted for the AR_1 model. Both of these autocorrelation plots may indicate underlying correlation structure in the series not obvious from the time series trace.

There are different formal tests that are available to test for stationarity of the sampler. These include those studied by Geyer (1992), Gelman and Rubin (1992), Raftery and Lewis (1992b), Geweke (1992), Ritter and Tanner (1992), Garren and Smith (1993), Johnson (1994), Zellner and Min (1995) and Roberts (1995). The Geweke test (Geweke 1992) splits the sample after removing a burn-in period into two parts, say the first 10% and the last 50%. If the chain is at stationarity, the means of the two samples should be equal. A modified z-test can be used to compare the two subsamples and the resulting test statistic is often called the Geweke z-score. A value larger than 2 indicates that the mean of the series is still drifting and a longer burn-in is required before monitoring of the chain can begin. A more informative method is the Raftery-Lewis test (Raftery and Lewis 1992a). In this approach, one specifies a particular quantile q of the distribution of interest, typically 2.5% and 97.5%, to give a 95% confidence interval, an accuracy ϵ of the quantile and power (1-β) for achieving this accuracy on the specified quantile. With these parameters set, the Raftery-Lewis test breaks the chain into a $(1, 0)$ sequence, that is 1 if $\theta_t \leq q$, 0 otherwise. This generates a two-state Markov chain and the Raftery-Lewis test uses the sequence to estimate the transition probabilities. With these probabilities in hand, one can estimate the number of additional burn-ins required to approach stationarity, the thinning ratio, i.e how many points should be discarded for each sampled point, and the total chain length required to achieve the preset level of accuracy (Walsh *et al*. 2004).

One can either use a single long chain (Geyer 1992, Raftery and Lewis 1992b) or multiple chains each starting from different initial values (Gelman and Rubin 1992). An advantage about using parallel processing machines is that multiple chains may be computationally more efficient than a single long chain. Others argue that using a single longer chain is the best approach (Geyer 1992). In a case where long burn-in periods are required or if the chains have higher autocorrelations, using a number of smaller chains may result in each not being long enough to be of any value. Applying the diagnostic tests given here can resolve some of these issues for any particular sampler.

4.3.1.2 A Metropolis-Hastings Acceptance-Rejection Algorithm

Recall that in the A-R method described earlier, a constant c and a density $h(x)$ are needed such that $ch(x)$ dominates or blankets the possibly unnormalized target density $f(x)$. In some applications it may be difficult to find a c that does the trick. If $f(x)$ depends on parameters that are revised during an iterative cycle, finding a new value of c for each new set of the parameters may significantly slow the computations. For these reasons it is worthwhile to have an A-R method that does not require a blanketing function. Tierney's (1994) remarkable algorithm does this by using an A-R step to generate candidates for an M-H algorithm. This algorithm, which seems complicated at first, can be derived rather easily using the intuition developed for the M-H algorithm.

Interest is in sampling the target density $\pi(x)$, $\pi(x) = f(x)/K$, where K may be unknown and a pdf $h(\cdot)$ is available for sampling. Suppose $c > 0$ is a known constant, but $f(x)$ is not necessarily less than $ch(x)$ $\forall x$, i.e $ch(x)$ does not necessarily dominate $f(x)$. It is convenient to define the C where domination occurs as;

$$C = \{x : f(x) < ch(x)\}$$

In this algorithm, given $x^{(n)} = x$, the next value $x^{(n+1)}$ is obtained as follows: First a candidate value z is obtained, independent of the current value x, by applying the A-R algorithm with $ch(\cdot)$ as the dominating density. The A-R step is implemented through steps 1 and 2 of the A-R algorithm. We have,

$$\begin{aligned} q(y) &= P(y|u \leq f(z)/ch(z)) \\ &= \frac{P(u \leq f(z)/ch(z)|z = y) \times h(y)}{P(u \leq f(z)/ch(z))} \end{aligned}$$

But since $P(u \leq f(z)/ch(z)|z = y) = \min\{f(y)/ch(y), 1\}$, it follows that

$$q(y) = \min \frac{\{f(y)/ch(y), 1\} \times h(y)}{d}$$

where $d \equiv P(u \leq f(z)/ch(z))$. By simplifying the numerator of this density a more useful representation for the candidate-generating density can be obtained

$$\begin{aligned} q(y) &= f(y)/cd \quad \text{if} \quad y \epsilon C \\ &= h(y)/d \quad \text{if} \quad y \notin C \end{aligned} \tag{4.17}$$

There is no need to write $q(x, y)$ for this density since the candidate y is drawn independent of x. Because $ch(y)$ does not dominate the target density in C^c, it follows that the target density is not adequately sampled there. This can be corrected with an M-H step applied to the y values that come through the A-R step. Since x and y can each be in C or C^c, there are four possible cases: (a) $x\epsilon C$, $y\epsilon C$; (b) and (c) $x \notin C$, $y\epsilon C$ or $x\epsilon C$, $y \notin C$ and (d) $x \notin C$, $y \notin C$. Now the objective is to find the M-H moving probability $\alpha(x, y)$ such that $q(y)\alpha(x, y)$ satisfies reversibility. It is required to derive $\alpha(x, y)$ in each of the four possible cases. We consider $\pi(x)q(y)$ and $\pi(y)q(x)$ or equivalently $f(x)q(y)$ and $f(y)q(x)$ to see how the probability of moves should be defined to ensure reversibility. That is, we need to find $\alpha(x, y)$ and $\alpha(y, x)$ such that

$$f(x)q(y)\alpha(x, y) = f(y)q(x)\alpha(y, x)$$

in each of the cases $(a) - (d)$, where $q(y)$ is chose from equation (4.17).

Case (a): $x\epsilon C$, $y\epsilon C$. In this case it is easy to verify that $f(x)q(y) \equiv f(x)f(y)/cd$ is equal to $f(y)q(x)$. Accordingly, setting $\alpha(x, y) = \alpha(y, x) = 1$ satisfies reversibility.

Case (b) and (c): $x \notin C$, $y\epsilon C$ or $x\epsilon C$. In the first case $f(x) > ch(x)$, or $h(x) < f(x)/c$, which implies (on multiplying both sides by $f(y)/d$), that

$$\frac{f(y)h(x)}{d} < \frac{f(y)f(x)}{cd}$$

or from equation (4.17), $f(y)q(x) < f(x)q(y)$. We now see that there are relatively too few transmissions from y to x and too many in the opposite direction. By setting $\alpha(x,y) = 1$ the first problem is alleviated and then $\alpha(x,y)$ is determined from

$$\frac{f(y)h(x)}{d} = \alpha(x,y)\frac{f(x)f(y)}{cd}$$

which gives $\alpha(x,y) = ch(x)/f(x)$. If $x\epsilon C$, $y \notin C$, reverse the roles of x nd y above to find that $\alpha(x,u) = 1$ and $\alpha(y,x) = ch(y)/f(y)$.

Case (d): $x \notin C$, $y \notin C$. In this case we have $f(x)q(y) = f(x)h(y)/d$ and $f(y)q(x) = f(y)h(x)/d$ and there are two possibilities. There are too few transitions from y to x to satisfy reversibility if

$$\frac{f(x)h(y)}{d} > f(y)q(x).$$

In that case set $\alpha(y,x) = 1$ and determine $\alpha(x,y)$ from

$$\alpha(x,y)\frac{f(x)h(y)}{d} = \frac{f(y)h(x)}{d}$$

which implies that,

$$\alpha(x,y) = \min\left\{\frac{f(y)h(x)}{f(x)h(y)}, 1\right\}$$

If there are fewer transitions from x to y, just interchange x and y in the above.

It is evident that in two of the cases where $x\epsilon C$, the probability of the move to y is 1, regardless of where y lies. In summary the probability of move to the candidates y that are produced from the A-R step are derived as follows:

- Let $C1 = \{f(x) < ch(x)\}$ and $C2 = \{f(y) < ch(y)\}$.

- Generate u from $U(0,1)$ and
  - if $C1 = 1$, then let $\alpha = 1$;
  - if $C1 = 0$ and $C2 = 1$, then let $\alpha = (ch(x)/f(x))$;
  - if $C1 = 0$ and $C2 = 0$, then let $\alpha = \min\{f(y)h(x)/f(x)h(y), 1\}$
- If $u \leq \alpha$
  - return y.
- Else,
  - return x.

(Chib *et al.* 1995)

4.3.2 The Gibbs Sampler

The Gibbs sampler is a special case of the Metropolis-Hastings algorithm. Gibbs sampler has been found to be very useful in many multidimensional applications. The Gibbs sampler is also known as an *alternating conditional sampling* which is defined in terms of subvectors of θ, i.e. $\theta = (\theta_1, \ldots, \theta_d)$ (Gelman *et al.* 1998). In the Gibbs sampler one needs only to consider the univariate conditional distributions, that is, the distribution when all of the random variables but one are assigned fixed values. These conditional distributions have simple forms and are easier to simulate than complex joint distributions. One simulates n random variables sequentially from the n univariate conditionals rather than generating a single n-dimensional vector in a single pass using the full joint distribution (Walsh *et al.* 2004).

Consider a bivariate random variable (x, y) and suppose we wish to compute one or both marginals, $p(x)$ and $p(y)$. The idea behind the sampler is that it is far easier to consider a sequence of conditional distributions $p(x|y)$ and $p(y|x)$ than it is to obtain the marginals by integration of the joint density $p(x, y)$. The sampler starts with some initial value y_0 for y and x_0 for x by generating a random variable from the conditional distribution $p(x|y = y_0)$. The sampler uses x_0 to generate a new value of y_1, drawing from the conditional distribution based on the value x_0, $p(y|x = x_0)$. The sampler proceeds as follows

$$x_i \sim p(x|y = y_{i-1})$$
$$y_i \sim p(y|x = x_i)$$

This process is repeated k times, generating a Gibbs sampler of length k, where the subset of points (x_j, y_j) for $1 \leq j \leq m < k$ are taken as the simulated draws from the full joint distribution. One iteration of all the univariate distributions is often called a 'scan' of the sampler. In order to get the desired total m sample points, one samples the chain

- after a sufficient burn-in to remove the effects of the initial sampling values.
- at a set time points (say, every n samples) following the burn-in.

The Gibbs sampler sequence converges to a stationary distribution that is independent of the starting values and by construction this stationary distribution is the target distribution we are trying to simulate from (Tierney 1994).

Each iteration of the Gibbs sampler cycles through the subvectors of θ, drawing each subset conditional on the value of all the others. There are thus d steps in iteration t. At each iteration t, each θ_j^t is sampled from the conditional distribution given all the other components of θ:

$$p(\theta_j|\theta_{-j}^{t-1}, y)$$

where θ_{-j}^{t-1} represents all the components of θ, except for θ_j, at their current values:

$$\theta_{-j}^{t-1} = (\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_d^{t-1}).$$

Therefore, each subvector θ_j is updated conditional on the latest value of θ for the other components, which are the iteration t values for the components already updated and the iteration $t-1$ values for the others (Gelman *et al.* 1998). When more than two variables are involved, the sampler is extended in the obvious fashion. In particular, the value of the kth variable is drawn from the distribution

$$\theta_i^k \sim p(\theta^{(k)}|\theta^{(1)} = \theta_i^{(1)}, \ldots, \theta^{(k-1)} = \theta_i^{(k-1)}, \theta^{(k+1)} = \theta_{i-1}^{(k+1)}, \ldots, \theta^{(n)} = \theta_{i-1}^{(n)})$$

Suppose there are five variables (s, t, u, v, w), then the sampler becomes,

$$\begin{aligned}
s_i &\sim p(s|t = t_{i-1}, u = u_{i-1}, v = v_{i-1}, w = w_{i-1}) \\
t_i &\sim p(t|s = s_i, u = u_{i-1}, v = v_{i-1}, w = w_{i-1}) \\
u_i &\sim p(u|s = s_i, t = t_i, v = v_{i-1}, w = w_{i-1}) \\
v_i &\sim p(v|s = s_i, t = t_i, u = u_i, w = w_{i-1}) \\
w_i &\sim p(w|s = s_i, t = t_i, u = u_i).
\end{aligned}$$

Power of the Gibbs sampler to address a wide variety of statistical issues have been studied (Gelfand and Smith 1990). The Gibbs sampler can be thought of as a stochastic analog to the Expectation-Maximization (Hartley 1958; Dempster *et al.* 1977; McLachlan and Krishnan 1997) approaches used to obtain likelihood functions when missing data are present. In the sampler, random sampling replaces the expectation and maximization steps. Any feature of interest for the marginals can be computed from the m realizations of the Gibbs sequence. Suppose that, the expectation of any function f of the random variable x is approximated by

$$E[f(x)]_m = \frac{1}{m}\sum_{i=1}^{m} f(x_i). \tag{4.18}$$

This is the *Monte-Carlo (MC)* estimate of $f(x)$, as $m \to \infty$ $E[f(x)]_m \to E[f(x)]$. Likewise, the MC estimate for any function of n variables $(\theta^{(1)}, \ldots, \theta^{(n)})$ is given by

$$E[f(\theta^{(1)}, \ldots, \theta^{(n)})]_m = \frac{1}{m}\sum_{i=1}^{m} f(\theta_i^{(1)}, \ldots, \theta_i^{(n)}). \tag{4.19}$$

Computing the actual shape of the marginal density is slightly cumbersome as compared to computing the MC estimate of any moment using the sampler. Suppose one uses the Gibbs sequence of x_i values, say, to give a rough approximation of the marginal distribution of x, this turns out to be inefficient, especially for obtaining the tails of the distribution. A more useful approach is to use the average of the conditional densities $p(x|y = y_i)$, as the function form of the conditional density contains more information about the shape of the entire distribution than the sequence of individual realizations x_i (Gelfand and Smith 1990, Liu *et al.* 1991). Since

$$p(x) = \int p(x|y)p(y)dy = E_y[p(x|y)]. \tag{4.20}$$

The marginal density can be approximated using

$$\widehat{p}_m(x) = \frac{1}{m}\sum_{i=1}^{m} p(x|y = y_i). \tag{4.21}$$

Suppose we are interested in using an appropriately thinned and burned-in Gibbs sequence $\theta_1, \ldots, \theta_n$ to estimate some function $h(\theta)$ of the target distribution, such as mean, variance, or specific quantiles like a cumulative probability value. As we are drawing random variables, there is some sampling variance associated with MC estimate

$$\widehat{h} = \frac{1}{n}\sum_{i=1}^{n} h(\theta_i). \tag{4.22}$$

When the length of the chain is increased ($n \uparrow$) the sampling of the variance of $\widehat{h}$ can decrease, but it will be nice to have some estimate of the size of this variance. A less complicated approach is to run several chains and use the between-chain variance in $\widehat{h}$. If $\widehat{h}_j$ denotes the estimate for chain $j(1 \leq j \leq m)$ where each of the m chains has the same length, then the estimated variance of the MC estimate is

$$Var(\widehat{h}) = \frac{1}{m-1}\sum_{j=1}^{m}\Big(\widehat{h}_j - \frac{1}{m}\sum_{j=1}^{m}\widehat{h}_j\Big)^2. \tag{4.23}$$

An alternative approach is to use results from the theory of time series using only a single chain. Estimate the lag-k autocovariance associated with h by

$$\widehat{\gamma}(k) = \frac{1}{n}\sum_{i=1}^{n-k}\Big[\Big(h(\theta_i) - \widehat{h}\Big)\Big(h(\theta_{i+k}) - \widehat{h}\Big)\Big]. \tag{4.24}$$

This is the natural generalization of the k-th order autocorrelation to the random variable generated by $h(\theta_i)$. The resulting estimate of the MC variance is

$$Var(\widehat{h}) = \frac{1}{n}\Big(\widehat{\gamma}(0) + 2\sum_{i=1}^{2\delta+1}\widehat{\gamma}(i)\Big) \tag{4.25}$$

δ is the smallest positive integer satisfying $\widehat{\gamma}(2\delta)+\widehat{\gamma}(2\delta+1)>0$, meaning that the higher order (lag) autocovariances are zero. One measure of the effects of autocorrelation between elements in the sampler is the *effective chain size*,

$$\widehat{n}=\frac{\widehat{\gamma}(0)}{Var(\widehat{h})}. \tag{4.26}$$

In the absence of autocorrelation between members, $\widehat{n}=n$.

4.3.2.1 Gibbs sampler as a special case of the M-H algorithm

Gibbs sampling can be viewed as a special case of the M-H algorithm with this jumping distribution which at iteration (j,t) only jumps along the jth subvector with the conditional posterior density of $\theta_j|\theta_{-j}^{t-1}$;

$$J_{j,t}^{Gibbs}(\theta^*|\theta^{t-1})=\begin{cases} p(\theta_j^*|\theta_{-j}^{t-1},y) & \text{if} \quad \theta_{-j}^*=\theta_{-j}^{t-1} \\ 0 & \text{otherwise}\end{cases}$$

where $J(\cdot)_t$ is the jumping or candidate-generating distribution. The only possible jumps are to parameter vectors θ^* that match θ^{t-1} on all components other than the jth. Under this jumping distribution, the ratio of importance ratios at the jth step of the iteration t is

$$\begin{aligned} r &= \frac{p(\theta_j^*|y)/J_{j,t}^{Gibbs}(\theta^*|\theta^{t-1})}{p(\theta_j^{t-1}|y)/J_{j,t}^{Gibbs}(\theta^{t-1}|\theta^*)} \\ &= \frac{p(\theta^*|y/p(\theta_j^*|\theta_{-j}^{t-1},y)}{p(\theta^{t-1}|y/p(\theta_j^{t-1}|\theta_{-j}^{t-1},y)} \\ &= \frac{p(\theta_{-j}^{t-1}|y)}{p(\theta_{-j}^{t-1}|y)} \\ &= 1 \end{aligned} \tag{4.27}$$

and therefore every jump is accepted. The second step above follows from the first since, under this jumping rule, θ^* differs from θ^{t-1} only in the jth component. The third step follows from the second by applying the rules of conditional probability to $\theta=(\theta_j,\theta_{-j})$ and noting that $\theta_{-j}^*=\theta_{-j}^{t-1}$. Usually, one iteration of the Gibbs sampler is defined as, to include all d Metropolis steps corresponding to the d components of θ, thereby updating all of θ at each iteration. It is possible to define Gibbs sampling without the restriction that each component be updated in each iteration, as long as each component is updated periodically.

For some problems sampling from some or all of the conditional distributions $p(\theta_j|\theta_{-j}, y)$ is impossible, but one can construct approximations, denoted by $g(\theta_j|\theta_{-j})$, from which samples are possible. The general form of the M-H algorithm can be used to compensate for the approximation. As in the Gibbs sampler, an ordering for altering the d elements of θ can be chosen, the jumping distribution at the jth Metropolis step at iteration t is then

$$J_{j,t}(\theta^*|\theta^{t-1}) = \begin{cases} g(\theta_j^*|\theta^{t-1}) & \text{if} \quad \theta_{-j}^* = \theta_{-j}^{t-1} \\ 0 & \text{otherwise} \end{cases} \tag{4.28}$$

and the ratio of importance ratios, r, must be computed and the acceptance or rejection of θ^* decided.

4.3.2.2 Convergence Diagnostics

Since the Gibbs sampler is a special case of the M-H algorithm, therefore the convergence diagnostics discussed for the M-H algorithm also apply to the Gibbs sampler. It was noted that the Gibbs sampler usually produces chains with smaller autocorrelations than other MCMC samplers (Draper 2000). Tanner (1996) discussed an approach for monitoring approach to convergence based on the Gibbs stopper, in which weights based on comparing the Gibbs sampler and the target distribution are computed and plotted as a function of the sampler iteration number. It is expected that the distribution of the weights to spike as the sampler approaches stationarity (Tanner 1996).

4.3.3 Graphical Modelling

In Bayesian perspective, the model is presented as a joint density for the observations. Thus, the construction of a directed graphical model to express the joint distribution of the model is more convenient. Directed graphs provide a helpful representation of the joint distribution of several random variables. Relationship between variables in a model can be presented by letting nodes in a graph represent those variable and edges between nodes represent the presence or otherwise of a direct relationship defined in terms of conditional independence between them. Such graphs are usually presented as a hierarchical structure with those variables which exert the most influence on the data placed closest to the bottom and those of lesser influence placed in decreasing order up the graph. These graphs give a simplified implementation of MCMC algorithms by indicating which other variables feature in the full conditional distribution of any given variable (Brooks *et al.* 1998).

The basic concept of a directed acyclic graph (DAG) is as follows. It consists of nodes and directed edges where the direction of an edge between two nodes represents the direction of the relationship between the two corresponding variables. Nodes can be represented in two ways, either a circle denoting that the value of the corresponding variable is unknown and hence subject to estimation or by a square in which case the value of that variable is known. Hence, observed data and prior or hyperprior parameters are often represented by square nodes.

Arrows run between nodes from their direct influence (parents) to descendants to indicate the conditional independence assumption of the model. The graph represents the assumption that, given its parent nodes, each node is independent of all other nodes in the graph except descendants of the node. Directed links may be of two types: a solid arrow indicates a stochastic dependence while a dashed arrow indicates a logical function. To obtain the full joint probability distribution of all the quantities, only the conditional parent-child distributions need to be specified and express the joint distribution by the product of these distributions due to conditional independence assumptions.

In general, we have that for any particular node ν

$$\nu \perp \text{ non-descendants of } \nu | \text{parents of } \nu.$$

This implies that the joint distribution of both the parameters in the model, $\nu \epsilon V$ say, and the data can be factorized as

$$p(V) = \textstyle\prod_{\nu \epsilon V} p(\nu | \text{parents of } \nu).$$

Therefore, the DAG is equivalent to the factorization assumption that allows us to decompose the full joint distribution into smaller components. The conditional independence graph is obtained by moralizing the DAG. This involves marrying parents by introducing edge between the parents of each node ν in the graph. We then drop the directions of all edges to obtain the conditional independence graph. This graph provides us with all information regarding the relationships between the parameters in the model. For example, given a particular node ν, the full conditional distribution of that node, given the value of all the nodes can be expressed as the product

$$p(\nu | \text{rest}) \propto (\nu | \text{parents of } \nu) \textstyle\prod_{u \epsilon C_\nu} p(u | \text{parents of } u)$$

where C_ν denoted the set of children of node ν (Brooks *et al*. 1998).

Therefore, the conditional independence graph makes it easy to see which other parameters are required in the specification of the full conditional distributions for each parameter and thus simplifies the implementation of MCMC algorithms for such models. More explanatory examples can be found in Albert *et al*. (1998) and Brooks *et al*. (1998).

4.4 Analysis of the data

In this chapter Bayesian methods, specifically Gibbs sampler is used to model HIV and TB data. Potential determinants that are used are those that were identified as important using the simple logistic regression models. These variables were also used in Chapter 3 in the application of GEEs and GLMMs. In this chapter, WinBugs software is used to get the results. The appropriateness of the sampled realizations was assessed using the Gelman-Rubin statistic which is modified by Brooks and Gelman (1998). The diagrams for these statistics are shown in Appendix B. In each graph there are three lines that are shown, the green line (the width of the central 80% interval of the pooled runs), blue line (the average width of the 80% interval within the individual runs) and the red line (the ratio of of the pooled over the within calculated in bins of length 50). For convergence diagnostic the red line converges to 1 where the other two lines converge to stability. It is evident from the diagrams in Appendix B that convergence is achieved.

The HIV model was implemented with the following variables used: sex of the respondent, age of the respondent, race group, educational qualification, health status and condom use at sexual debut. The dataset has 16398 observations. However, these were also incomplete cases. Since WinBugs is highly sensitive to incomplete or missing data, all those records with missing cases were removed from the data. Thereafter only 9412 observations remained.

Priors for fixed effects were assumed multivariate normal centered at zero. Specifically the prior values were set at $\mu = 0.0$ and $\sigma^2 = 1.0 \times 10^{-6}$. Priors for the random effects were assumed to follow a normal distribution with parameters zero and τ. The ancillary parameter τ is assumed to follow a gamma distribution with parameters α and β. The variance is a reciprocal of τ (precision) i.e variance=$1/\tau$. As stated above τ follows a gamma distribution and the prior values were α=0.001 and β=0.001. The model was initialized and 1000 iterations were simulated. It was noted that coefficient parameters for education levels have not converged yet based on the trace plot. In addition, a further 1000 iterations were simulated. The trace plot for education showed convergence. Therefore, 2000 iterations were taken as the burn-in period. Thereafter, 10 000 iterations were simulated with every 20^{th} value was taken (i.e thinning value=20) which resulted with the final sample of 1000 realizations. For HIV model, the results are presented in Table 4.1 and Table 4.2 for TB model. In the case of the TB model the same procedure as that, of HIV was followed. The variables that were fitted for TB model are: HIV status, sex of the respondent, educational qualification, income status and health status.

In the HIV model the variance was estimated to be 0.1659 with standard error 0.04718. The results confirms that males are less likely to be infected with HIV as compared to females. Individuals in age group 25 to 34 are more likely to be infected with HIV than any other age group. As the age group increases the chances of being infected with HIV decrease. This is evident in Table 4.1. The odds of being infected also varies according to racial groupings. In this analysis it clear that the odds of being infected with HIV are higher among Africans than any other race group. The odds of being infected with HIV for Whites are 1.8 times higher than that for Indians. But, this was not statistically significant. It is also evident that Coloureds are more likely to be infected with HIV than Indians.

The rate of HIV infection also varies according to educational qualification. The odds of being infected with HIV for those individuals with no or primary education are 2.23 times higher than that for those with tertiary education. Those with secondary or matric education level are 2 times more likely to be infected with HIV than those with tertiary education. Health status is also an important indicator of HIV infection. In this study, individuals who are in good health are less likely to be infected with HIV than those with poor health. Using a condom at the sexual debut indicated that one is well-informed of the risk of HIV and in the analysis those who used a condom during their sexual debut are less likely to be infected with HIV than those who did not use a condom during their sexual debut.

TB status was modeled using variables mentioned above. An estimated variance is 0.4413 with standard error 0.2021. Individuals who are infected with HIV are more likely to be infected with TB than those who are HIV negative. Results indicate that the odds of being infected with TB are higher for males than they are for females. The results show that individuals with no or primary education are more likely to be infected with TB than those with tertiary education. The more educated one gets, the lesser the chances of contracting TB, Table 4.2. Individuals who are unemployed are more likely to be infected with TB than those who get income from other sources (grants, donations, pensions, etc). However, this might be confounded by the fact that these people may be unemployed due to TB. The statistical test shows that this is not statistically significant. The odds of being infected with TB for those who are employed are 0.5 times less than that for those who receive income from other sources and this is statistically significant. Individuals who are in good health are less likely to be infected with TB than those who are in poor health.

The estimated intraclass correlations for HIV and TB are respectively given by ρ_{HIV}=0.169 and ρ_{TB}=0.249. These positive correlations indicate that people in the same EA are more likely to be correlated in their risk of HIV (ρ_{HIV}) as well as their risk of infection with TB (ρ_{TB}). Diagnostic diagrams are presented from Figure 4.1 to Figure 4.4. The two chains are plotted in the trace plots together with the credible intervals.

Figure 4.1: History: Complete trace plots for HIV fixed effects parameters

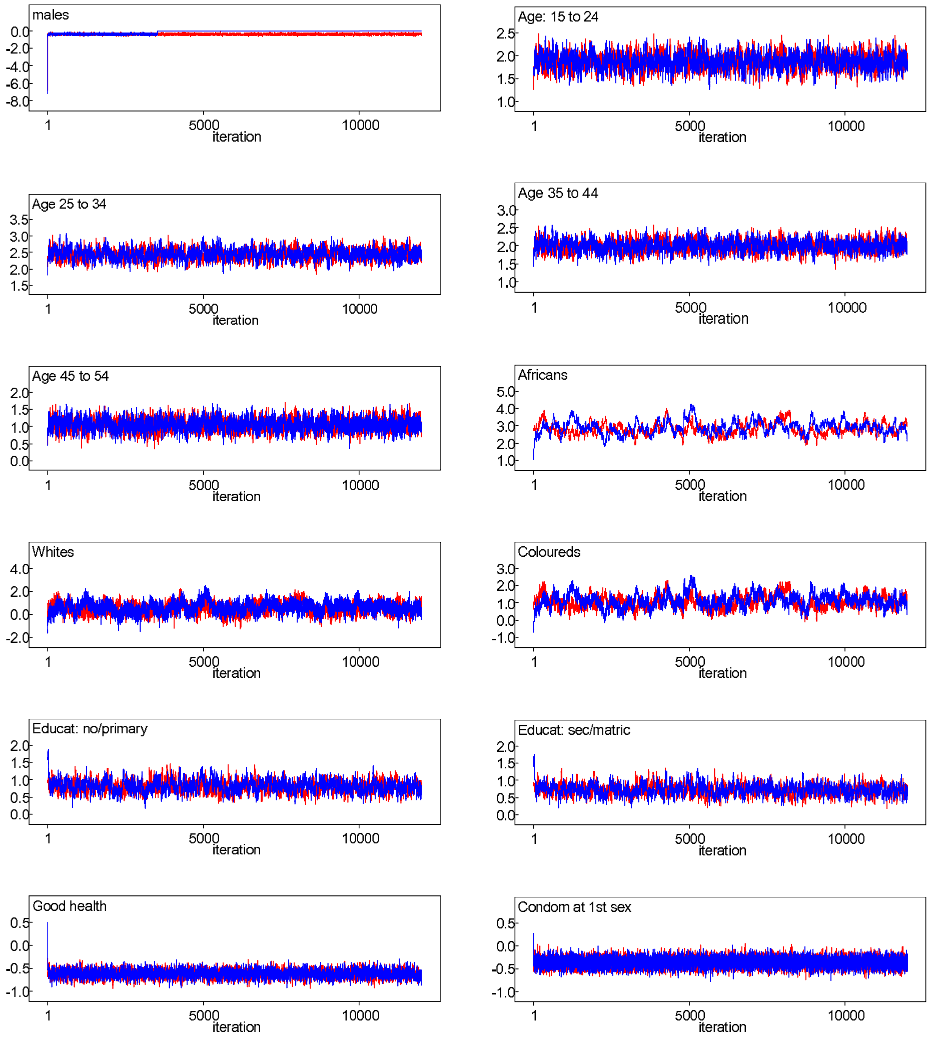


Figure 4.2: Density diagrams for HIV fixed effects parameters

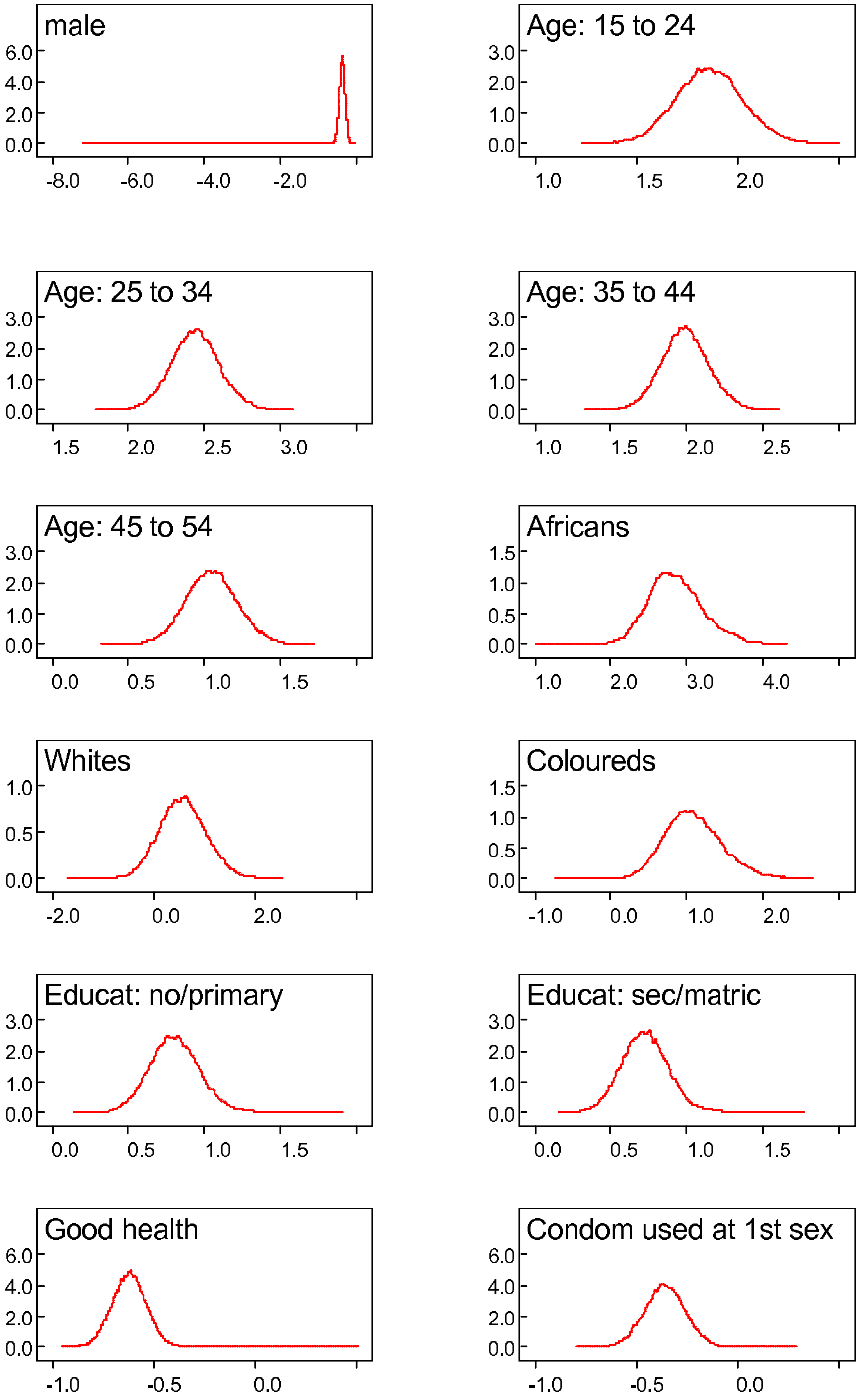


Figure 4.3: History: Complete trace plots for TB fixed effects parameters

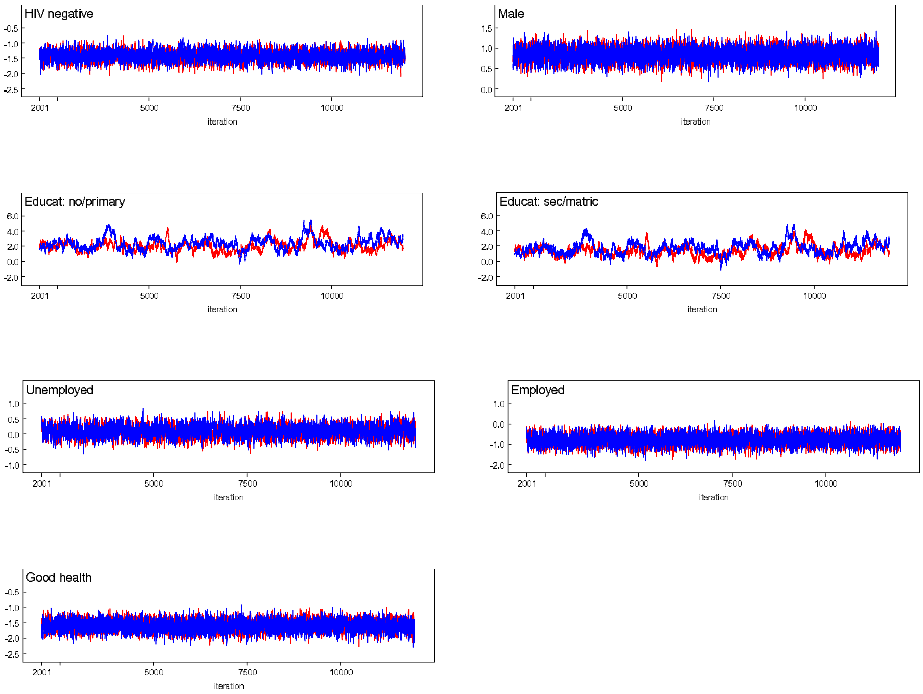
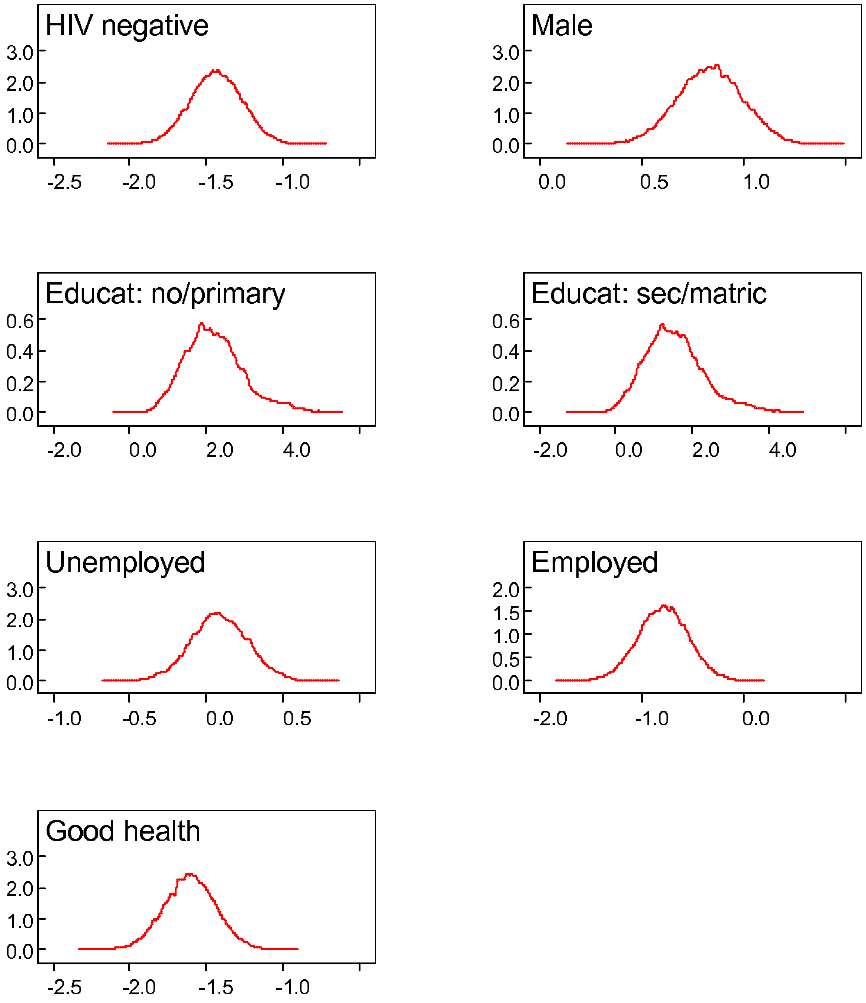


Figure 4.4: Density diagrams for TB fixed effects parameters



4.5 Discussion

In this section a Bayesian approach was applied in the data where we model the random effects as we did with the GLMMs under the frequentist approach. Therefore it is worth comparing the two results from these different approaches. The distinction between the GLMMs and the Bayesian GLMMs is that in Bayesian all variables are random draws from a multivariate normal distribution with mean zero (Gilks, 1998). The Bayesian results are presented together with the GLMMs results. From these results one can see that the estimates are quite comparable. Looking at the standard errors especially those for HIV model, it is clear that the standard errors estimated from the Bayesian hierarchical model are lesser than those estimated by the GLMMs. When looking at standard errors for the TB model one can see that they are comparable. In the GLMMs the random effects can only be normally distributed while in Bayesian model this assumption is relaxed, where the random effects can take any distribution. This results in a model which uses the Gibbs sampling. The comparability of the results between the two models is re-assuring. Thus, inference drawn from the two modelling approaches provides some degree of confidence in the results.

The positive correlation observed at an EA level for both HIV and TB indicates that interventions should consider the area level effect rather than only the individuals. Studies that intervene at community level, for example Mwanza Trial (Grosskurth, 1995) should be encouraged to fight epidemics of diseases such as HIV and TB.

Table 4.1: HIV model: Parameter Estimates and Standard Errors

	Bayesian results	GLMMs results
Parameter	Estimate (SE)	Estimate (SE)
Intercept	-6.177 (0.404)	-5.813 (0.404)
Sex of the respondent		
Male	-0.376 (0.072)	-0.346 (0.074)
Female	0	0
Age of respondent		
15 to 24	1.861 (0.160)	1.825 (0.176)
25 to 34	2.438 (0.157)	2.379 (0.168)
35 to 44	1.987 (0.151)	1.943 (0.165)
45 to 54	1.046 (0.164)	1.030 (0.179)
55+	0	0
Race		
African	2.857 (0.360)	2.063 (0.342)
White	0.589* (0.457)	0.312* (0.468)
Coloured	1.080 (0.374)	0.894 (0.363)
Indian	0	0
Education		
None/primary	0.803 (0.164)	0.749 (0.175)
Secondary/matric	0.721 (0.156)	0.715 (0.168)
Tertiary	0	0
Health status		
Good	-0.623 (0.083)	-0.645 (0.086)
Poor	0	0
Condom use at sexual debut		
Yes	-0.358 (0.099)	-0.359 (0.104)
No	0	0

* p-value>0.05

Table 4.2: TB model: Parameter Estimates and Standard Error

	Bayesian results	GLMMs results
Parameter	Estimate (SE)	Estimate (SE)
Intercept	-4.098 (0.822)	-3.797 (0.814)
HIV status		
Negative	-1.430 (0.167)	-1.445 (0.172)
Positive	0	0
Sex of the respondent		
Male	0.837 (0.161)	0.863 (0.162)
Female	0	0
Education		
None/primary	2.205 (0.797)	2.0593 (0.789)
Secondary/matric	1.534 (0.796)	1.388* (0.789)
Tertiary	0	0
Income		
Unemployed	0.084* (0.192)	0.188* (0.181)
Employed	-0.789 (0.251)	-0.571 (0.243)
other	0	0
Health status		
Good	-1.598 (0.163)	-1.678 (0.167)
Poor	0	0

* p-value>0.05

Chapter 5

Discussion and Conclusions

5.1 Human Immune Virus

The data in this project is based on a multi-stage disproportionate stratified sampling approach based on a master sample of 1000 enumerator areas (EAs). or first stage clusters. Thus, the modelling approach used in this thesis incorporated the first stage clusters in the design in order to account for between cluster variability. The master sample allowed for reporting of results at the level of province, type of locality, age, gender and race group. Gender inequality is a huge problem that South Africa is facing since women are less empowered to negotiate safe sex with their partners (Worth *et al.* 1990; Stein *et al.* 1990). In relationships where there is an imbalance of power young women's ability to practice safer sex is oppressed by their partners' demands (Eaton *et al.* 2003). Women are then exposed to sexual abuse, rape and commercial sex activities for survival. So far it has been reported that HIV is higher among women than men based on antenatal clinic attendee studies and other studies (Department of Health 2007; Shisana *et al.* 2003).

Young girls engage in risky sexual activities mostly to better them lives. They tend to be sexually involved with men of older age than theirs who might be at high risk of HIV. Prostitution is a way of getting income but, most often clients coerce prostitutes to have unprotected sex (Swart-Kruger *et al.* 1997). Some of the young girls engage in transactional sex where the only benefit is getting a better life and luxuries. The impact of transactional sex is that these girls have no say in taking preventive measures. Thus, this mostly results in unprotected sexual act which put young ladies at higher risk of HIV as their ability to use condoms is compromised (Pettifor *et al.* 2005). Individuals in age group 15 to 24 are just entering their active sexual life therefore some of them may make extensive use of condoms but may be inconsistent. The results are in agreement with other analyses regarding heterosexually transmitted HIV dynamics (Anderson and May, 1991). That is, the prevalence of HIV for this age group (15 to 24) is less than that for the age group 25 to 34, the more sexually active age group. Individuals in this age group engage in highly connected sexual networks with possibly inconsistence use of condoms which results in high rate of new infections.

In South Africa more than 50% of the youth becomes sexually active by the age of 16, and more than 80% are sexually active by age of 20 (Eaton *et al.* 2003). Africans are more likely to become sexually active earlier than other race groups (Eaton *et al.* 2003). Eaton also mentions that South African youth (14-34 years) are at higher risk of HIV infection due to unprotected sex which starts at their teenage ages. Youth in urban areas are better informed about HIV than the youth from rural areas. South African youth is highly affected by problems that are associated with poverty (Eaton *et al.* 2003). Individuals who are older are not as sexually active as the youth and thus their risk of being infected with HIV is lower. Especially those who are above 54 years where some of them are no longer sexually active.

Many studies have shown that Africans are at higher risk of HIV infection than any other race groups (Eaton *et al.* 2003). Our results also confirm this. Mostly Africans live in very crowded places like informal settlements. Often these places are associated with poverty, overcrowding and it is difficult for people living in these areas to access some basic needs, for example, clinics and cannot afford better medical care. On the other hand, individuals living in urban formal areas are more equipped with information regarding HIV. A big percentage of the White and Indian people live in such places. Thus, these people are more aware of HIV and their risk of being infected with HIV is very low.

Level of education is important in explaining the risk of HIV infection. Individuals with lower education level tend to be less informed about the risks of HIV. Most likely individuals who have no education or have less education level are found in poor communities. Thus, poor/disadvantaged communities are associated with high degree of HIV risks of infection (Kalichman *et al*. 2006). Lack of education also results in the increase of behavioral risk factors for HIV infection (Kalichman *et al*. 2006). Low levels of education, poverty, overcrowding and unemployment are much associated with the increase of adolescents sexual activity and less knowledge about HIV/AIDS (Preston-Whyte *et al*. 1991; Du Plessis *et al*. 1993; Wood *et al*. 1997b).

Individuals from poor rural communities do not have access to the media and they need more information about HIV and they are more dependent on outside experts for their information (Kelly, 2000). It appeared that those with lower education level are at higher risk of HIV. Individuals with low level of education are less knowledgeable about the transmission of the disease and tend to engage in more risky sexual activities than those who have higher education levels. They also tend to live in crowded environment with less health facilities. Furthermore, low level of education implies one is not exposed to factual knowledge about the epidemic hence being at higher risk.

In this study the analysis shows that education is associated with race (only in the simple logistic regression model). Results showed that it is mostly Africans that are at higher risk of HIV in at most all different educational levels when compared to Indians. Poor health care infrastructure is caused by poverty which is in turn linked to HIV infection (Kalichman *et al.* 2006). A strong immune system is capable of suppressing heterosexually transmitted infections such as ghonorrhea and syphillis. Thus people with a weakened immune system are more likely to be infected with STIs. Infection with HIV is also a significant risk factor of STIs possibly due to reduced immunity (Zuma *et al.* 2005).

Condom use is very crucial because it offers dual protection, against not only unneccessary pregnancies but also from sexually transmitted infections which include HIV. The analysis has shown that people who used condoms during their sexual debut are at lower risk of HIV. This shows that these people are more informed about the risk of HIV and thus, they make sure that they protect themselves from being infected. In a study done by Magnani *et al.* 2005, condom use at sexual debut was increasing in females, Africans and younger youth. It was noted that condom use at sexual debut among Africans was much lower than that for the youth from other racial groups (Magnani *et al.* 2005).

However, it appeared that condom use is consistently increasing among Africans and younger youth unlike the condom use at sexual debut among males (it is important to note that Magnani's study was based on students who are attending school in age group 14-24 years). The majority of sexually active individuals use condoms inconsistently or not at all (Eaton *et al.* 2003). Other individuals do not use condoms because of the myths associated with them, for example that a condom can disappear inside a woman which could be dangerous, associated with promiscuity and also less sexual satisfaction/pleasure and people rather prefer *skin on skin* sexual intercourse (Eaton *et al.* 2003).

5.2 Tuberculosis

Tuberculosis (TB) is one of the leading opportunistic infections for HIV infected individuals. Individuals who have TB and also HIV positive are more likely to die from TB than any other infections including HIV even though TB is preventable and curable (Bucher *et al.* 1999; Corbett *et al.* 2003). HIV is a powerful catalyst in reactivating latent TB. This study showed that individuals who are infected with HIV are at higher risk of contracting TB. This is in agreement with other studies done by other researchers (Zwang *et al.* 2007; Dlodlo *et al.* 2005; Corbett *et al.* 2003). TB has emerged as the major source of the increased morbidity and mortality among HIV infected individuals especially in the sub-Saharan countries. Nonetheless studies show that TB prevalence is declining in the sub-Saharan Africa (Dlodlo *et al.* 2005) but the compounding effect of HIV is making it hard to realize this success.

The problem is often two-way because individuals who are infected with TB are likely to be infected with HIV, because for such individuals their immune system is at a much weaker state to fight the deadly virus on entry. While those with a strong immune system are at lower risk of being infected with TB. The mortality of HIV and TB is higher among adults who are 20 years and above (Zwang *et al.* 2007). Individuals infected with HIV are at increased risk of TB infection as soon as they get HIV infection (Sonnenberg *et al.* 2002). Increase in HIV prevalence results in an increase in active tuberculosis in HIV positive individuals because of the immune system being compromised which increases the risk of tuberculosis (Sonnenberg *et al.* 2004).

The results showed that males are more likely to be infected with TB than females. Males are likely to be exposed to poor working environments that put them at increased risk of TB than females. Zwang *et al.* (20070 stated that the co-infection of TB and HIV affects more males at an earlier age than females and the mortality rate is higher among males who are co-infected with both HIV and TB than females for all ages. In addition the mortality rate of TB only is higher among males than females.

Educated individuals are likely to be employed and thus have enough income to take good care of their health. The more educated are at lesser risk of being infected with TB. This is in agreement with the study by Harling *et al.* (2008) which showed that there is a 10% reduction in the odds of being infected with TB for an additional year of education completed. Individuals who have higher education qualification have better paying jobs and they reside in highly resourceful areas where they can afford better health facilities and they live in less crowded places, thus, the odds of infection for them are small.

This study showed that income status is an important risk determinant of TB where individuals who are unemployed are more likely to be infected with TB. The risk of TB for unemployed individuals is not different from that for those individuals who receive income from other sources, like donations, grants and pensions. TB is mostly taken as a disease of those in low social status (Harling *et al.* 2008). Factors such as income inequality, low levels of education, high levels of poverty and high social deprivation are highly associated with TB (Krieger *et al.* 2003; Parslow *et al.* 2001; Spence *et al.* 1993; Tocque *et al.* 1999). Harling *et al.* (2008) reports that there is a 40% reduction in the odds of being infected with TB for those individuals who were employed in the past year.

Individuals with a low socioeconomic status usually suffer from the compromised immune system. Though in this study no interaction of factors were identified but it still stands that individuals who have a weakened immune system are usually the ones who are unemployed, have low education level, living in crowded areas with inadequate health facilities. Also, bearing in mind that TB is a communicable infectious disease and thus it spreads rapidly in crowded areas. Some studies have shown that income inequality is associated with different health outcomes more especially in high inequality settings (Subramanian *et al.* 2004; Kawachi *et al.* 2000).

5.3 Intervention Implications

The impact of HIV in increasing the risks of TB infection needs thorough considerations. Intervention strategies or policy makers should focus more on HIV and TB co-infected individuals. The policy makers should consider gender inequities when initiating intervention programmes as it appears that females are still at higher risk of being infected with HIV than males. In addition to that intervention strategies should be directed to the youth from all places not only those who reside in formal settlements. Africans are at higher risk of HIV infection compared to other race groups. Therefore policy makers should implement strategies that will also suit everyone including those who are less fortunate in terms of socioeconomic status.

People who live in poor communities usually do not have access to media facilities such as televisions, radios and billboards. These media facilities are the ones that are used most often for intervention campaigns and people who are likely to have access to them are those who are not at higher risk of HIV infection. The GLMMs and Bayesian results showed positive correlations of HIV and TB among individuals from the same enumerator area which is an indication that individuals from the same enumerator area are more likely to be correlated in their risk of infection with the two epidemics. Therefore intervention strategies should rather aim at a community level rather than at an individual level as those who are not yet infected are not aware of the epidemic and they will only know more about it once they are infected.

The importance of education should be emphasized thoroughly. As the result showed that individuals who have tertiary education are at lower risk of being infected with HIV and TB. Thus, this implies that if everyone can attain higher education standards there will be a significant decrease in the risk of infection. Emphasizing the condom use at sexual debut and the consistent use of condoms at every sexual contact will have a significant reduction in the risk of HIV.

The relationship between HIV and TB is very complicated as there are many factors supporting this. If individuals can be diagnosed early for TB and treated this will reduce the mortality rate of individuals who are HIV infected but die due to TB. Individuals do not know their TB status until when they discover that they are HIV infected. Treating TB in individuals who are HIV infected is difficult as some of them are already also taking the antiretroviral drugs and this could result in complications in adherence to the treatment. Intervention strategies should highlight the importance of testing for TB at any time.

The government and those responsible should develop job opportunities to those who have no jobs as it is evident in this study that individuals who are unemployed are at higher risk of becoming infected with TB. The unemployed individuals end up living in crowded areas as they cannot afford better accommodation and these crowded places are the catalyst of the TB infection.

5.4 Different Statistical approaches

In this work three statistical approaches are applied in the data. At the beginning the GLMs were used to determine the potential risk determinants for the two diseases. GLMs were extended in order to model the intracluster correlation within an EA, that is the GEEs. The GLMMs were also applied to model the heterogeneity that exists in a community level. Lastly, the Bayesian hierarchical approach was also applied in the data where we also model the heterogeneity and this approach most likely gives the same results as the GLMMs. The GEEs are marginal or population-averaged models. They model the marginal expectation of the dependent variable as a function of covariates (Zorn *et al.* 2001). Cluster-specific approaches model the probability distribution of the dependent variable as a function of covariates and a parameter specific to each cluster. The cluster-specific models are so called random effects models. Marginal models assume that the relationship between the outcome and the covariates is the same for all subjects while the random effects model allows this relationship to differ between subjects (Carriere *et al.* 2002). In the cluster-specific models, interpretation of fixed effects parameters is conditional on a constant level of the random effects parameter.

Marginal models make no use of within cluster comparisons for cluster varying covariates and they are thus not useful to assess within-subjects effects (Neuhaus *et al.* 1991). Marginal models have many attractive properties including robustness to the specification of the variance-covariance structure (Crouchley *et al.* 2001). From a substantive perspective this is often a model of limited interest as it does not distinguish the behavioural dynamics of the process. Though it can be argued that GEEs are potentially more flexible than the random effects approaches in the different variance-covariance structures readily incorporated into the approach. But it is not clear how one can select empirically between different structures within the GEEs framework because there is no measure of goodness of fit (Crouchley *et al.* 2001).

Marginal models are easy to implement and represent a first solution, but the random models, although more complex use all available data and they are more suitable for explicative studies (Carriere *et al.* 2002). The GEE parameters and the random-effects parameters cannot be compared directly as the variability of the random effect is highly significant (Jansen *et al.* 2006). The MCMC methods are more flexible in handling multiple clustering levels but they are more complex to handle (Zeger *et al.* 1991). GLMMs and the Bayesian hierarchical models are not different from each other. In GLMMs the fixed and random effects are partitioned in terms of **X** and **Z**. In the Bayesian hierarchical models there is no need to partition the variables in terms of **X** and **Z**. In Bayesian all variables are random variables drawn from a multivariate normal distribution with mean zero (Gilks *et al.* 1998). The distinction between the two effects may be dropped and the linear model may simply be specified as for a simple GLM. The difference between the two sorts of effect lies in the specification of the precision matrix in the prior model.

The reason behind the extensive target on the GLM is that the GLM is used as a foundation of the whole thesis. This method is used to determine the potential risk determinants of HIV and TB. These estimates are then used on the more complex statistical methods in such way that we will be able to compare or critique the results. Thus, more about these methodologies is discussed above.

In Chapter 2 under the GLM section we firstly have some background theory about the GLM. We then proceed by explaining how the models are formulated and estimated. This notation is only used in this chapter only as in next chapters we are going to use only the determinants identified using these GLM models. Whereas in the GEE and GLMM section we only highlight the theory of the two statistical methodologies without going into deep details on how the models are formulated. Thus the notation is not that different as we work step-by-step until the last chapter of the thesis. Almost all the variables there were identified using the GLM method are used in the GEE and GLMM sections except for the interaction term education by race. The reason behind this is that when this term was includes in the models, it appeared that the models were over-parameterized in a way that convergence was difficult to achieve.

In chapter 2 the GLM was used to determine the potential determinants of both infectious diseases. In doing so it is important to note that the sampling design was incorporated in the models. These variables were identified using surveylogistic in SAS where there is an option for stratification variable(s). The survey involved two stratification variables which is the province and the geographical type of an EA. Therefore, in the analysis we had geographical type of an EA nested within a province, i.e province(geotype). In addition we adjusted for weights in order the minimise the potential bias. Therefore the determinants identified by the GLM do capture the impact of the design. Including the weights in the GEE and GLMM methods resulted in wide confidence intervals. Thus, it was suggested or rather obvious that adjusting for weights at these levels was rather redundant.

5.5 Conclusions

The analysis showed potential risk determinants of HIV and those of TB. Also the common risk factors can easily be identified from the two models. As stated in the previous chapters that individuals within a cluster tend to be more alike in terms of risk of the disease, the results showed that incorporating the intraclass correlation results in more accurate results. In the HIV model, the intraclass correlation was ignored the results seem to be overestimated, for example, SE for Africans in the multivariate model is 1.043 whereas in the GEE model is 0.525. This highlight the important of modelling this correlation. Furthermore, this indicates that race plays a role in capturing the intraclass correlation since historically, the country has been racially divided for years. There are unknown effects that are driving the rapid spread of a disease. These unknown effects are taken as random and they are modelled using the generalized linear mixed models and also the Bayesian hierarchical models. As the disease occurrence is highly heterogeneous at the community level, therefore, in this study the community level was taken to be an enumerator area.

Factors that drive the two epidemics are most socio-economic factors, for example, income, education, health status, race to mention the few. This is an indication that HIV and TB infection are highly influence by the socio-economic factors. It was evident in the analysis that those with the poor socio-economic status are the ones that are seriously at higher risk of infection. The analysis showed that individuals who are unemployed are more likely to be infected with the disease.

5.6 Future Work

In this study our focus was on the South African results, those who are 15 years and above. This means that those individuals who are younger than 15 years are not sexually active or they are not exposed to HIV or TB and this may not be true. In South Africa there is a high rate of teenage pregnancy which shows that young kids engage in sexual encounter without using any preventive measures. Short comings that we mostly experience in South Africa is the lack of biologically verified TB infection. This may lead to an underestimate of TB since only those who were previously diagnosed with TB are aware. Getting such dataset will help a lot in improving the results. For example one can compare the results from the self-reported TB and those from the tested TB so that those involved in policy making can implement necessary policies.

The GLMMS and Bayesian results showed a positive correlation of HIV and TB at an enumerator area level, thus spatial modelling of these data can help map areas that are prone to HIV and or TB. One may be interested in doing a joint modelling of the two diseases in getting a clear indication of the causal effect.

In the TB model it is shown that health status is a significant risk determinant of TB. One may argue that while this may be true also TB can be significant determinant of being in a poor health status. Using the same data the chi-square test showed (results not shown) that indeed TB is a potential determinant of being at poor health state. Therefore, this dataset can be used in estimating factors that affect the health status of the South Africans.

Appendix A

Abstract of papers from the thesis

Estimating Risk Determinants of HIV and TB in the South African Population

Abstract

Where HIV/AIDS has had its greatest adverse impact is probably on its effect on tuberculosis (TB). People with TB that are infected with HIV are at higher risk of dying from TB. The burden of TB in countries with high rates of HIV has increased rapidly over the past decade, especially in the severely affected countries of eastern and southern Africa. Similarly the relationship between HIV and STIs is complicated. This is because HIV is also sexually transmitted and therefore shares the same behavioural risk factors. The epidemiological importance of STIs has acquired greater significance as it became apparent that they enhance transmission of HIV and are important co-factors driving the HIV epidemic. However, infection with HIV is also a significant risk factor of STIs possibly due to reduced immunity. Lack of reliable, up-to-date data on HIV, TB and STIs has hampered efforts to inform sound policies on intervention and control strategies of these seemingly highly dependent diseases particularly the HIV/AIDS pandemic. In South Africa, no coherent analysis of determinants of HIV and TB has been done at a national level and thus this study seeks to mend this gab using the second generation national

survey data. The objectives of the study are to investigate the risk determinants of HIV and those of TB therefore preliminary analysis will be performed where the primary interest is on the main effects only. The results will therefore provide a better understanding of the spread of the two epidemics in South Africa, viz HIV and TB. Presence and absence of disease is non-normal data therefore specific statistical models capable to handle such type of data will be used. Since individuals from teh household are likely to be correlated among each other therefore the generalized estimating equations will be used to correct for this potential clustering. Results will be compared with those found when ignoring clustering effect, i.e using logistic regression modelling.

Estimating Risk Determinants of HIV and TB in South Africa

Abstract

HIV/AIDS has had greatest adverse impact on tuberculosis (TB). People coinfected with TB and HIV are at higher risk of dying from TB. Similarly the relationship between HIV and STIs is complex as they are both sexually transmitted. In South Africa, no coherent analysis of determinants of HIV and TB has been done at a national level and thus this study seeks to mend this gab. The objectives of the study are to investigate the risk determinants of HIV and those of TB also identifying factors that are common in both epidemics. The results will therefore provide a better understanding of the spread of these two epidemics in South Africa. In this work GEEs are used to capture potential intraclass correlation at an area level. The dataset used is from a population based second-generation surveillance survey conducted by HSRC in 2005. The survey was conducted at households and collected information on socio-demographic, sexual, behavioral and biomedical factors in 15 000 households from 1 000 enumerator areas. The results show that sex or gender, race, age, education and health are important determinants of HIV and TB. Furthermore, the spread of HIV is clustered within an enumerator area. Thus interventions should be aimed at an enumerator level for their maximum effects.

Bayesian approach in Estimating Risk determinants of infectious diseases

Abstract

In South Africa, no coherent analysis of determinants of HIV and TB has been done at a national level and thus this study seeks to mend that gab. Results will therefore provide a better understanding of the spread of these two epidemics. The dataset used is from a population based second-generation surveillance survey conducted by HSRC in 2005. The survey was conducted in 15 000 households from 1000 enumerator areas which applied a stratified random sampling. In recognition of the heterogeneity that exists at the community level we propose to model this by applying the Bayesian modelling approaches and the frequentist approaches. The results show that sex, race, age, education and health are important determinants of HIV and TB. The inference drawn from the two modelling approaches provides some degree of confidence in the results. Thus, interventions should target enumerator area level rather than individual level for their maximum effects.

Appendix B

In figure 5.1 and 5.2 the red line for each variable converges to 1 whereas the pooled and the within interval widths converge to stability indicating the significant convergence.

Figure 5.1: GR statistic for HIV model

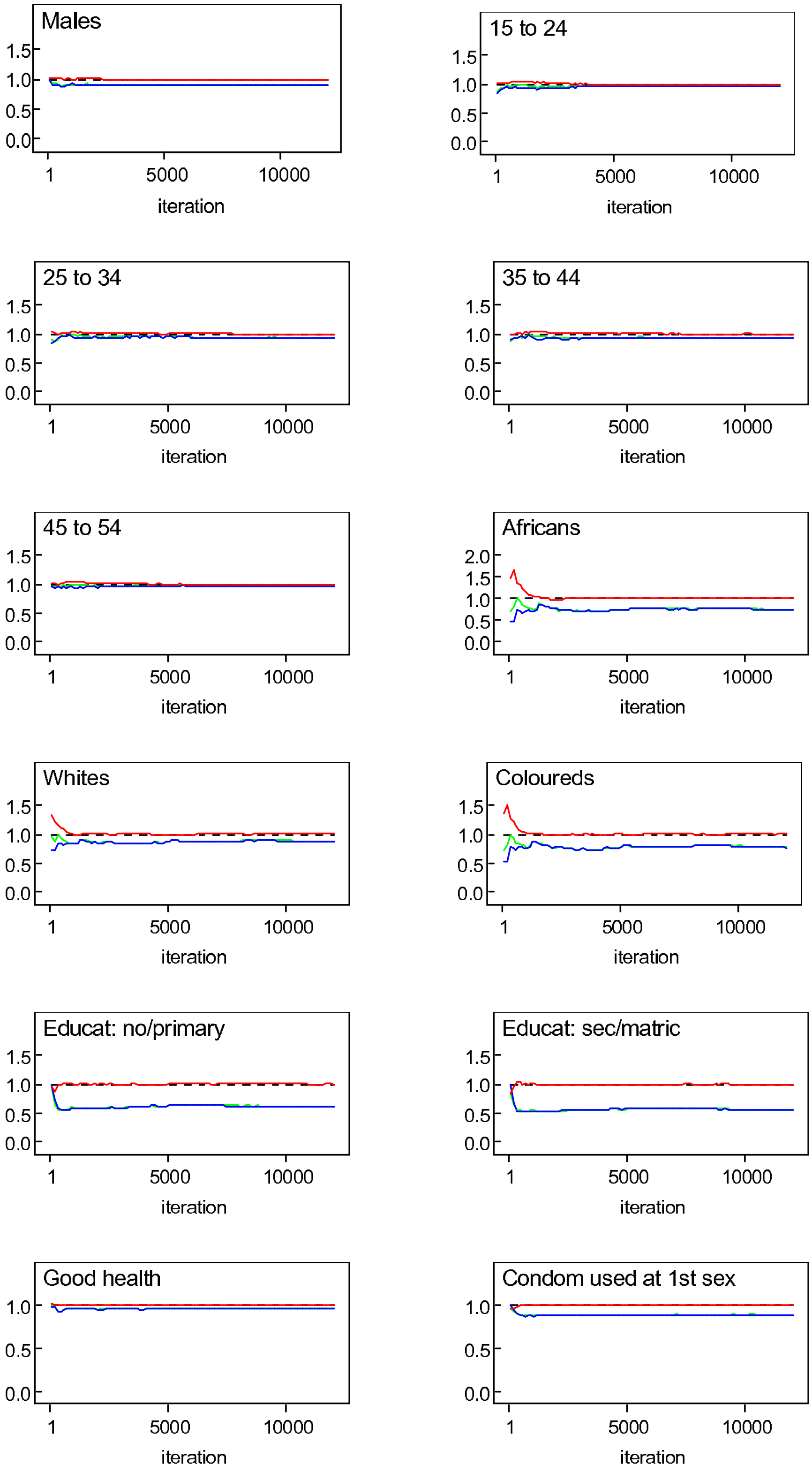
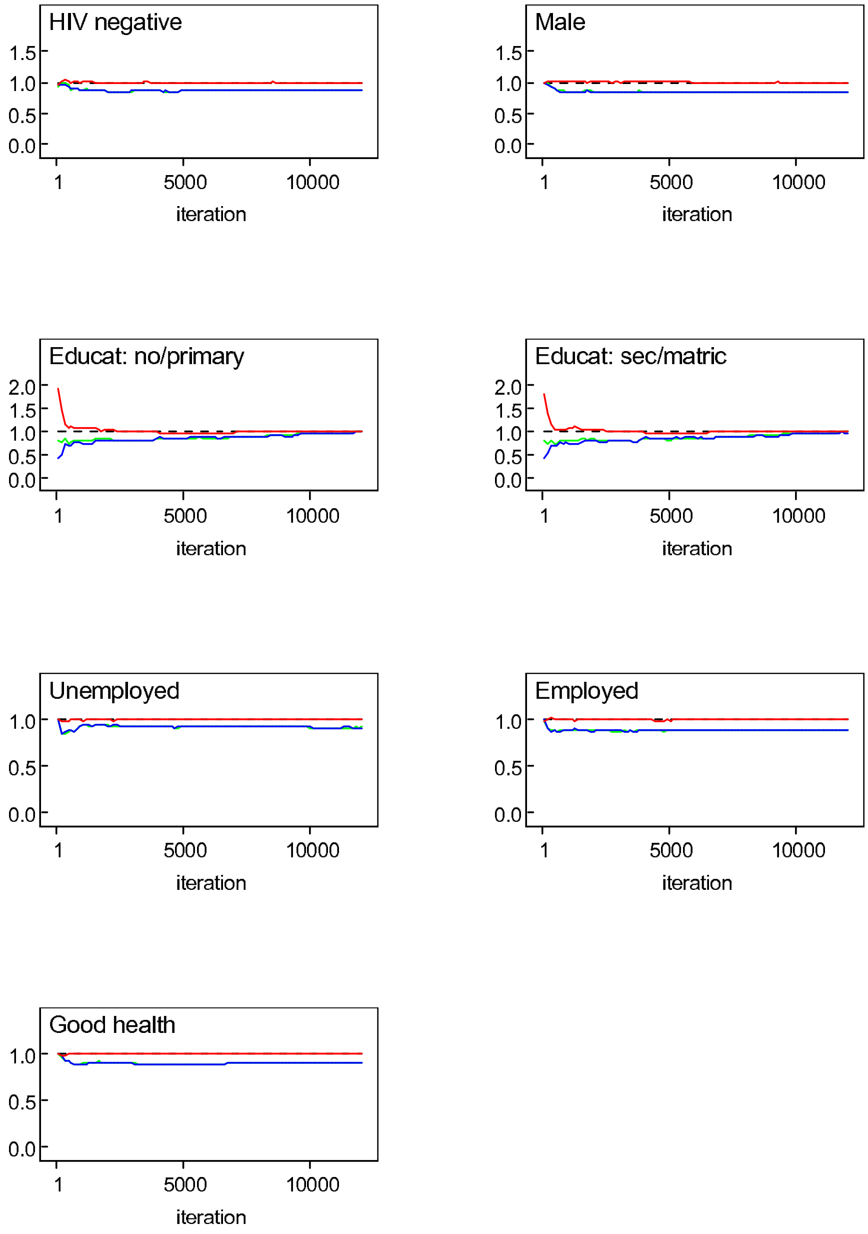


Figure 5.2: GR statistic for TB model



Bibliography

[1] Achmat, Z. and Roberts, R. A. Steering the Storm: TB and HIV in South Africa A policy paper of the Treatment Action Campaign Report June 2005.

[2] Aerts, M., Geys, H., Molenberghs, G. and Ryan, L. M. (2002). Topics in Modelling of Clustered Data. Chapman and Hall.

[3] Agresti, A. and Finlay, B. (1997). Statistical Methods for the Social Sciences. Prentice Hall International.

[4] Agresti, A. (1990). Categorical Data Analysis. John Wiley and Sons.

[5] Agresti, A. (2002). Categorical Data Analysis. John Wiley and Sons.

[6] Anderson, R. M. and May, R. M. (1991). Infectious Diseases of Humans: Dynamics and Control, Oxford University Press, Oxford, UK.

[7] Albert, I. and Jais, J. (1998). Gibbs sampler for the logistic model in the analysis of longitudinal binary data. *Statistics in Medicine*, **17**, 2905-2921.

[8] Auvert, B., Taljaard, D., Lagarde, E., Sobngwi-Tambekou, J., Sitta, R. and Puren, A. (2005). Randomized, Cobtrolled Intervention Trial of Male Circumcision for Reduction of HIV Infection Risk: The ANRS 1265 Trial. *Plos Medicine* Vol. 2, Issue 11, e298, doi:10.1371/journal.pmed.0020298

[9] Bamford, L., Loveday, M. and Verkuijl, S. Tuberculosis, 15. Health Systems Trust

[10] Besag, J., Green, P. J., Higdon, D. and Mengersen, K. L. M. (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science*, **10**, 3-66.

[11] Beunckens, C., Sotto, C. and Molenberghs, G. (2007). A simulation study comparing weighted estimating equation with multiple imputation based estimating equation for longitudinal binary data. *Comput. Statis. Data Anal.*, doi:10.1016/j.csda.2007.04.020

[12] Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalizeed linear mixed models. *J. Amer. Statist. Assoc.*, **88**, 9-25.

[13] Brooks, S. P. (1998). Markov chain Monte Carlo method and its application. *The Statistician*, **47**, 69-100.

[14] Bucher, H. C., Griffith, L. E. and Guyatt, G. H. (1999). Isoniazid prophylaxis for tuberculosis in HIV infection: a meta analysis of randomized controlled clinical trials. *AIDS*, **13**, 501-507.

[15] Cantwell, M.F., and Binkin, N. J. (1996). Tuberculosis in sub-Saharan Africa: a regional assessment if the impact of the human immunodeficiency virus and National Tuberculosis Control quality. *Tuber Lung Dis*, **77**, 220-225.

[16] Casella, G., and George, E. I. (1992). Explaining the Gibbs sampler. *Am. Stat.*, **46**, 167-174.

[17] Carriere, I. and Bouyer, J. (2002). Choosing marginal or random-effects models for longitudinal binary responses: application to self-reported disability among older persons. *BMC Medical Research Methodology*, **2**, 15.

[18] Cerny, V. (1985). A thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm. *Journal of Optimization Theory and Applications*,**45**, 41-51.

[19] Chib, S. and Greenberg, E. (1995). Understanding the Metropolic-Hastings Algorithm. *The American Statistician Association*, **49** No.4, 327-335.

[20] Churchyard, G. J., Kleinschmidt, I. and Corbett, E. L. (1999). Mycobacterial diseases in South Africa gold miners in the era of HIV infection. *Int J Tuberc Lung Dis*, **3**, 791-798.

[21] Corbett, E. L., Watt, C. J., Walker, N., Maher, D., Williams, B. G., Raviglione, M. C. and Dye, C. (2003). The Growing Burden of Tuberculosis: Global Trends and Interactions With the HIV Epidemic. *Arch Intern Med*, **163**, 1009-1021.

[22] Cowles, M. K and Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A comparative Review. *Journal of the American Statistical Association*, **91**, 883-904.

[23] Crouchley, R. and Davies, R. B. (2001). A comparison of GEE and random effects models for distinguishing heterogeneity, nonstationarity and state dependence in a collection of short binsrty event series. *Statistical Modelling*, **1**, 271-285.

[24] Diggle, P., Liang, K. and Zeger, S. (1994). Analysis of Longitudinal Data. *Claredon Press*, Oxford.

[25] Diggle, P. J., Heagerty, P., Liang, K. and Zeger, S. L. (2002). Analysis of longitudinal data. *Oxford University Press*, Second Edition.

[26] Dlodlo, R. A., Fujiwara, P. I. and Enarson, D. A. (2005). Should tuberculosis treatment and control be addressed differently in HIV-infected and -uninfected individuals. *Eur Respir J*, **25**, 751-757.

[27] Dorrington, R. E., Johnson, L. F., Bradshaw, D. and Daniel, T. The Demographic Impact of HIV/AIDS in South Africa, National and Provincial Indicators for 2006. Cape Town: The Centre for Actuarial Research, South African Medical Research Council and Actuarial Society of South Africa.

[28] Draper, D. (2000). Bayesian Hierarchical Modeling.

[29] Du Plessis, G. E., Meyer-Weitz, A. J. and Steyn, M. (1993). Study of knowledge, attitudes, perceptions and beliefs regarding HIV and AIDS (KABP) among the general publiic. Pretoria: Human Sciences Researrch Council.

[30] Eaton, L., Flisher, A. J. and Aaro, L. E. (2003). Unsafe sexaul behaviour in South African youth. *Social Science & Medicine*, **56**, 149-165.

[31] Gouws, E. Incidence of HIV infection in rural KwaZulu Natal, PHD thesis submitted, University of KwaZulu Natal, 2006.

[32] Fleming, D. T. and Wasserheit, J. N. (1999). From an epidemiological synergy to public health policy and practice: the contribution of other sexually transmitted diseases to sexaul transmission of HIV infection. *Sexually Transmitted Infections*, **75**, 3-17.

[33] Frothingham, R., Stout, J. E., Hamilton, C. D. (2005). Current issues in global tuberculosis control. *International Journal of Infectious Diseases*, **9**, 297-311.

[34] Gandhi, N., Moll, A., Sturm, A. W., Pawinski, R., Govender, T., Lalloo, U., Zeller, K., Andrews, J. and Friedland, G. (2006). Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis amd HIV in a rural area of South Africa. *Lancet*, **368**, 1575-80.

[35] Gelfand, A. E., and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Stat.*, **85**, 398-409.

[36] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1998). Bayesian Data Analysis. *Chapman and Hall.*

[37] Gelman, A. and Rubin, D. B. (1992). Inferences from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457-511.

[38] Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution and Bayesian restoration of images. *IEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721-741.

[39] Geweke, J. (1992). Evaluating the accuracy of sampling-based appraoches to the calculation of posterior moments. *In, Bayesian Statistics*, **4**, eds. Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M., 169-193.

[40] Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion). *Stat. Sci.*, **7**, 473-511

[41] Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1998). Markov Chain Monte Carlo In Practice. Chapman and Hall/CRC.

[42] Grosskurth, H., Todd, J., Mwijarubi, E., Klokke, A., Senkoro, K., Mayaub, P., Changalucha, J., Nicoll, A. and ka-Gina, G. (1995). Impact of improved treatment of sexually transmitted diseases on HIV infection in rural Tanzania: randomised controlled trial. *Lancet*, **346(8974)**, 530-536.

[43] Grosskurth, H., Mosha, F., Todd, J., Senkoro, K., Newell, J., Klokke, A., Changalucha, J., West, B., Mayaub, P. and Gavyole, A. (1995). A community trial of the impact of improved sexually transmitted disease treatment on the HIV epidemic in rural Tanzania: 2. Baseline survey results. *AIDS*, **9(8)**, 927-34.

[44] Hardin, J. W. and Hilbe, J. M. (2003). Generalized Estimating Equations. Chapman and Hall.

[45] Harling, G., Ehrlich, R. and Myer, L. (2008). The social epidemiology of tuberculosis in South Africa: A multilevel analysis. *Social Science & Medicine*, **66**, 492-505.

[46] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, **57**, 97-109.

[47] Horton, N. J. and Lipsitz, S. R. (1999). Review of Software to Fit Generalized Estimating Equation Regression Models. *The American Statistical Association*, **53**, 160-169.

[48] Hosmer, D. W. and Lemeshow, S. (1989). Applied Logistic Regression. John Wiley and Sons.

[49] IZINDABA (October 2006). Living the TB resistance Nightmare. *SAMJ*, Vol. **96**, No.10.

[50] Jansen, I., Beunckens, C., Molenberghs, G., Verbeke, G. and Mallinckrodt, C. (2006). Analyzing Incomplete Discrete Longitudinal Clinical Trial Data. *Statistical Science*, **21**, No 1, 52-69.

[51] John-arne, R., Cameron, W. and Garnett, G. P. (2001). A systematic review of epidemilogical interactions between classic sexually transmitted diseases and HIV: How much really is known? *Sexually Transmitted Infections*, **28**, 579-597.

[52] Kalichman, S. C., Simbayi, L., Kagee, A., Toefy, Y., Jooste, S., Cain, D. and Cherry, C. (2006). Association of poverty, substance use, and HIV transmission risk behaviors in three South African communities. *Social Science & Medicine*, **62**, 1641-1649.

[53] Kawachi, I. (2000). Income inequality and health. In L. F. Berkman & I. Kawachi (Eds). *Social Epidemiology.* New York: Oxford University Press.

[54] Kelly, K. (2000). Communiting for action: A context evaluation of youth responses to HIV/AIDS. Pretoria: Department of Health (Sentinel Site Monitoring and Evaluation Project for the Beyond Awareness Campaign).

[55] Kenyon, T. A., Mwasekaga, M. J. and Huebner, R. (1999). Slow levels of drug resistance amidst rapidly increasing tuberculosis and human immunodeficiency virus co-epidemics in Botswana. *Int J Tuberc Lung Dis.*, **3**, 4-11.

[56] Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science*, **220**, 671-680.

[57] Krieger, N., Waterman, P. D., Chen, J. T., Soobader, M. J. & Subramanian, S. V. (2003). Monitoring socioeconomic inequalities in sexually transmitted infections, tuberculosis and violence: Geocoding and choice of area-related socioeconomic measures. *Public Health Reports*, **118**, 240-260.

[58] Lee, Y., Nelder, J. A. (1996). Hierarchical Generalized Linear Models. *Royal Statistical Society*, **58**, No. 4, 619-678.

[59] Lee, Y., Nelder, J. A. and Pawitan, Y. (2006). Generalized Linear Models with Random Effects. Chapman and Hall/CRC

[60] Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

[61] Lindsey, J. K. (1997). Applying Generalized Linear Models. Springer texts in Statistics.

[62] Littell, R. C., Milliken, G. A., Stroup, W. W. and Wolfinger, R. D. (1996). SAS System for Mixed Models. Cary, NC: SAS Institute Inc.

[63] Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D. and Schabenberger, O. (2006). SAS System for Mixed Models, Second Edition. Cary, NC: SAS Institute Inc.

[64] Lurie, M. N., Williams, B. G., Zuma, K., Mkaya-Mwamburi, D., Garnett, G. P., Sturm, A. W., Sweat, M. D., Gittelsohn, J. and Abdool Karim, S. S. (2002). The Impact of Migration on HIV-1 Transmission in South Africa: A Study of Migrant and Nonmigrant Men and Their Partners. *Sexually Transmitted Diseases*, Vol. 30, No.2.

[65] Lurie, M. N., Williams, B. G., Zuma, K., Mkaya-Mwamburi, D., Garnett, G. P., Sweat, M. D., Gittelsohn, J. and Abdool Karim, S. S. (2003). Who infects whom? HIV-1 concordance and discordance among migrant and nonmigrant couples in South Africa. *AIDS*, **17**, 2245-2252.

[66] Magnani, R., MacIntyre, K., Karim, A. M., Brown, L., and Hutchinson, P. (2005). The impact of life skills education on adolescent sexual risk behaviors in KwaZulu-Natal, South Africa. *Journal of Adolescent Health*, **36**, 289-304.

[67] Martinson, N., Hausler, H., Churchyard, G. J., Lawn, S. D. (September 2005). Dealing with the dual epidemics of HIV and TB. *The Southern African Journal of HIV Medicine.*

[68] McCullagh, P. and Nelder, J. A. (1983). Generalized Linear Models. Chapman and Hall.

[69] McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models. Chapman and Hall.

[70] Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *J. Amer. Statist. Assoc.*, **44**, 335-341.

[71] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. and Teller, H. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087-1091.

[72] Mertens, T. E., Hayes, R. J. and Smith, P. G. (1990). Epidemiological methods to study the interaction between HIV infection and other sexually transmitted diseases. *AIDS*, **4**, 57-65.

[73] Molenberghs, G., Verbeke, G. (2005). Models for Discrete Longitudinal Data. Springer Science and Business Media, Inc.

[74] Monteiro, E. F., Lacey, C. J. N. and Merrick, D. (2005). The interrelation of demographic and geospatial risk factors between four common sexually transmitted diseases. *Sexually Transmitted Infections*, **81**, 41-46.

[75] Morgan, B. J. T. (2000). Applied Stochastic Modelling. Arnold texts in Statistics.

[76] Muller, M. (2004). Generalized Linear Models. Fraunhofer Institute for Industrial Mathematics. (Prepared for J. Gentle, W. Hardle, Y. Mori (eds): *Handbook of Computational Statistics (Volume I). Concepts abd Fundamentals,* Springer-Verlag, Heidelberg, 2004)

[77] Myer, L., Wilkinson, D., Zuma, K., Rotchford, K. and Abdool Karim, S. S. (2003). Impact of on-site testing for maternal syphilis on treatment delays, treatment rates, and perinatal mortality in rural South Africa: a randomised controlled trial. *Sex. Transm Infect*, **79**, 208-213.

[78] Neuhaus, J. M., Kalbfleisch, J.D. and Hauck, W. W. (1991). A Comparison of Cluster-Specific and Population-Averaged Approaches for Analyzing Correlated Binary Data. *International Statistical Review*, **59**, 25-35.

[79] Orelien, J. G. Model Fitting in PROC GENMOD (2001). *Statistics, Data Analysis and Data Mining*, 264-26.

[80] Parslow, R., El-Shimy, N. A., Cundall, D. B. and McKinney, P. A. (2001). Tuberculosis, deprivation, and ethnicity in Leeds, UK, 1982-1997. *Archives of Disease in Childhood*, **84**, 109-113.

[81] Peltzer, K., Niang, C. I., Muula, A. S., Bowa, K., Okeke, L., Boiro, H. and Chimbwete, C. (2007). Editorial Review: Male circumcision, gender and HIV prevention in sub-Saharan Afruca: a (social science) research agenda. *Journal of Social Aspects of HIV/AIDS*, *4*, 658-667.

[82] Pendergast, J., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M. and Fisher, M. R.(1996). A Survey of Methods for Analyzing Clustered Binay Response Data. *International Statistical Review*, **64**, 89-118.

[83] Pettifor, A. E., Rees, H. V., Kleinschmidt, I., Steffenson, A. E., MacPhail, C., Hlongwa-Madikizela, L., Vermaak, K. and Padian, N. S. (2005a). Young people's sexual health in South Africa: HIV prevalence and sexual behaviours from a nationally representative household survey, *AIDS*, **19**, 1525-1534.

[84] Pettifor, A. E., Kleinschmidt, I., Levin, J., Rees, H. V., MacPhail, C., Hlongwa-Madikizela, L., Vermaak, K., Napier, G., Stevens, W. and Padian, N. S. (2005b). A community-based study to examine the effect of a youth HIV prevention intervention on young people aged 15-24 in South Africa; results of the baseline survey. *Tropical Medicine and International Health*, **10**, 971-980.

[85] Piot, P., Taelman, H., Minlangu, K. B., Mbendi, N., Ndangi, K., Kalambayi, K., Bridts, C., Quinn, T. C., Feinsod, F. M., Wobin, O., Mazebo, P., Stevens, W., Mitchell, S. and McCormick, J. B. (1984). Acquired Immunodefinciency Syndrome in a Heterosexual Population in Zaire. *The Lancet*, **324**, 65-69.

[86] Preston-Whyte, E. and Zondi, M. (1991). Adolescent sexuality and its implications for teenage pregnancy and AIDS. *Continuing Medical Education*, **9**, 1389-1394.

[87] Ras, G. J., Simson, I. W., Prozesky, O. W. and Hamersma, T. (1983). Acquired immunodeficiency syndrome. A report of 2 South African cases. *S Afr Med J*, **64**, 140-2.

[88] Raftery, A. E, and Lewis, S. (1992a). How many iterations in the Gibbs sampler? *In Bayesian Statistics*, **4**, eds: Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. f. M., 763-773.

[89] Raftery, A. E, and Lewis, S. (1992b). Comment: One long run with diagnostics: Implementation strategies for Markov Chain Monte Carlo, *Stat. Sc.*, **7**, 493-497.

[90] Robert, C. P. and Casella, G. (1999). Monte Carlo Statistical Methods. *Springer Verlag.*

[91] Schabenberger O, Senior Research Statistician, SAS Institute Inc. Kansas State University Conference on Applied Statistics in Agriculture, April 30, 2006.

[92] Shisana, O. (2002). South African National HIV Prevalence, Behavioural Risks and Mass Media Household Survey 2002. Cape Town: Human Sciences Research Council.

[93] Shisana, O., Zungu-Dirwayi, N., Simbayi, L. C., Malik, S. and Zuma, K. (2004). Marital status and risk of HIV infection in South Africa. *SAMJ*, Vol. 94, No. 7.

[94] Shisana, O. (2005). South African National Prevalence, HIV Incidence, Behaviour and Communication Survey 2005. Cape Town: Human Sciences Research Council.

[95] Singh, J. A., Upshur, R., Padayatchi, N. (2007). XDR-TB in South Africa: No time for denial or complacency. *PLoS Med* 4(1):e50.doi:10.1371/journal.pmed.0040050.

[96] Smith, A. F. M. (1991). Bayesian computational methods. *Phil. Trans. R. Soc. Lond.*, **337**, 369-386.

[97] Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte-Carlo methods (with discussion). *J. Roy. Stat. Soc. Series*, **55**, 3-23.

[98] Sonnenberg, P., Glynn, J. R., Fielding, K., Murray, J., Godfrey-Fausett, P. and Shearer, S. (2002). How soon after HIV infection does the risk of TB start to rise? A retrospective cohort study in South African gold miners. In: *XIVth International AIDS Conference.* Barcelona, Spain [Abstract MoOrC1102].

[99] Sonnenberg, P., Glynn, J. R., Fielding, K., Murray, J., Godfrey-Fausett, P. and Shearer, S. (2004). HIV and pulmonary tuberculosis: the impact goes beyond those infected with HIV. *AIDS*, **18**, 657-662.

[100] Spence, D. P., Hotchkiss, J., Williams, C. S. and Davies, P. D. (1993). Tuberculosis and poverty. *British Meducal Journal*, **307**, 759-761.

[101] Stein, Z. A. (1990). HIV prevention: the need for methods women can use. *Am J Public Health*, **80**, 460-462.

[102] Subramanian, S. V. and Kawachi, I. (2004). Income inequality and health: What have we learned so far? *Epidemiologic Reviews*, **26**, 78-91.

[103] Swart-Kruger, J. and Richter, L. M. (1997). AIDS-related knowledge, attitudes and behaviour among South African street youth: Reflections on power, sexuality and the autonomous self. *Social Science & Medicine*, **45**, 957-966.

[104] Tanner, M. A. (1996). Tools for Statistical inference. Third Edition. *Springer Verlag.*

[105] Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.*, **22**, 1701-1762.

[106] Tocque, K., Regan, M. Remmington, T., Beeching, N. J., Jamieson, I. and Syed, Q. (1999). Social factors associated with increases in tuberculosis notifications. *European Respiratory Journal*, **13**, 541-545.

[107] Trapido, A. S., Mqoqi, N. P., Williams, B. G., White, N. W., Solomon, A., Goode, R. H., Macheke, C. M., Davies, A. J. and Panter, C. (1998). Prevalence of Occupational Lung Disease in a Random Sample of Former Mineworkers, Libode District, Eastern Cape Province, South Africa. *American Journal of Industrial Medicine*, **34**, 305-313.

[108] van de Perre, P., Rouvroy, D., Lepage, P., Bogaerts, J., Kestelyn, P., Kayihigi, J., Hekker, A. C., Butzler, J. P. and Clumeck, N. (1984). Acquired immunodeficiency syndrome in Rwanda. *Lancet*, **2(8394)**, 62-5.

[109] Verbeke, G. and Molenberghs, G. (2003-2004). Correlated and multivariate data.

[110] Verbeke, G. and Molenberghs, G. (2000). Linear Mixed Models for Longitudinal Data. Springer Series in Statistics.

[111] Walsh, B. (2004). Markov Chani Monte Carlo and Gibbs sampling. Lecture Notes.

[112] Walzl, H., Beyers, N. and van Helden, P. (2005). TB: A partnership for the benefit of research and community. *Transaction of the Royal Society of Tropical Medicine and Hygiene*, **99**, 15-19.

[113] Wasserheit, J. N. (1992). Epidemiological synergy: Interrelationships between human immunodeficiency virus infection and other sexually transmitted diseases. *Sex Transm Dis.*, **19(2)**, 61-77.

[114] Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized models and the Gauss-Newton method. *Biometrika*, **61**, 439-447.

[115] Weyer, K. and Fourie, P. B. Epidemiology of tuberculosis in South Africa and anticipated impact of HIV. In: Tuberculosis Control in South Africa-Joint Programme Review in 2004.

[116] World Health Organization. HIV and Tuberculosis Fact Sheet; September 2006

[117] Wilkinson, D. and Davies, G. R. (1997). The incresing burden of tuberculosis in rural South Africa: impact of HIV epidemic. *South African Medical Journal*, **87**, 447-450.

[118] Wilson, D. (June 2005). Diagnosing HIV-Associated Tuberculosis. *The Southern African Journal of HIV Medicine.*

[119] Wolfinger, R. and O'Connell, M.(1993). Generalized Linear Mixed Models: A Pseudo-Likelihood Approach. *J. Stat. Compt. Simul.*, **48**, 233-243.

[120] Worth, D. (1990). Sexual decision making and AIDS: why condom promotion among vulnerable women is likely to fail. *Stud Family Plann.*, **20**, 297-307.

[121] Wood, K., Maepa, J. and Jewkes, R. (1997b). Adolescent sex and contraceptive experiences: Perspectives of teenagers and clinic nurses in the Northern Province. Pretoria: Centre for Epidemiological Research in South Africa (Women's Health).

[122] Zeger, S. L. and Liang, K. (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics*, **42**, 121-130.

[123] Zeger, S. L., Liang, K., and Albert, P. S. (1988). Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics*, **44**, 1049-1060.

[124] Zeger, S. L. and Karim , M. R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *J Am Stat Assoc*, **86**, 79-86.

[125] Zorn, C. (2001). Generalized Estimating Equation Models for Correlated Data: A review with Applications. *American Journal of Political Science*, **45**, 470-490.

[126] Zuma, K., Gouws, E., Williams, B. and Lurie, M. (2003). Risk factors for HIV infection among women in Carletonville, South Africa: migration, demography and sexually transmitted diseases. *International Journal of STD and AIDS*, **14**, 814-817.

[127] Zuma, K., Lurie, M. N., Williams, B. G., Mkaya-Mwamburi, D., Garnett, G. P. and Sturm, A. W. (2005). Risk factors of sexually transmitted infections among migrant and non-migrant sexual partnerships from rural South Africa. *Epidemiol. Infect.*, **133**, 421-428.

[128] Zwang, J., Garenne, M., Kahn, K., Collinson, M. and Tollman, S. M. (2007). Trends in mortality from pulmonary tuberculosis and HIV/AIDS coinfection in rural South Africa (Agincourt). *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **101**, 893-898.