UNIVERISTY OF KWAZULU-NATAL, PIETERMARITZBURG

SCHOOL OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE

A COMPLEX SURVEY DATA ANALYSIS OF TB AND HIV MORTALITY IN SOUTH AFRICA

By

JOIE LEA MURORUNKWERE

STUDENT NUMBER: 210511980

A thesis submitted in fulfilment of the academic requirements for the degree of

MASTER OF SCIENCE

In

APPLIED STATISTICS

2012

DECLARATION

I, the undersigned, hereby declare that the work contained in this thesis is my original work, and that any work done by others or by myself previously has been acknowledged and referenced.

November, 2012.

Ms JOIE LEA MURORUNKWERE (210511980) Date

Supervisor: Prof. HENRY MWAMBI Date

Co-Supervisor: Dr. ACHIA THOMAS Date

DEDICATION

Almighty God

To my family

To all my friends

ACKNOWLEDGEMENTS

The greatest of all thanks goes to the Almighty God who has sustained me throughout all these years. The knowledge and wisdom all come from him.

This work could not be done without the help of many people whom, we hereby acknowledge. My special thanks go to all who have contributed to my education to date.

In a special way, I would like to thank my supervisor Professor Henry Mwambi for his tremendous hard work with me to complete this thesis successfully. Thank you for your invaluable guidance, unfathomable support, encouragement, and patience all the time.

I am grateful to my co-supervisor, Dr. Achia Thomas without you this would have not been what is today. Thank you for your fatherly role and excellent assistance for the successful submission of my thesis.

My sincere gratitude is also expressed to the staff of Statistics of The University of KwaZulu-Natal. I would like to thank Statistics South Africa for allowing me to use their rich dataset in this thesis. I am highly indebted to the ACTs PMB members for the assistance and cooperation I received from them. Thanks to my fellow Statistics Postgraduates students who, willingly came forward with elaborate suggestions during this undertaking.

To my family, late parents, cousins, brothers and sisters: thank you for supporting me even in difficult times, you were my pillar of strength.

I would like to extend my deepest and most heartfelt gratitude to Father Incimatata Oreste for his fatherly role, love and support that you have given me since what seems like forever. My sincere gratitude goes to Carmel Sisters especially Sister Liberatha.

I would also like to thank my wonderful friends, Mukandoli Caritas, Nadine Ineza, Egidia Uwizeyemariya and Muna William for their support, wise advises, prayers and love.

I would like to acknowledge all those who contributed in one way or another to the completion of this thesis.

ABSTRACT

Many countries in the world record annual summary statistics such as economic indicators like Gross Domestic Product (GDP) and vital statistics for example the number of births and deaths. In this thesis we focus on mortality data from various causes including Tuberculosis (TB) and HIV. TB is an infectious disease caused by bacteria called Mycobacterium tuberculosis. It is the main cause of death in the world among all infectious diseases. An additional complexity is that HIV/AIDS acts as a catalyst to the occurrence of TB. Vaidyanathan and Singh revealed that people infected with mycobacterium tuberculosis alone have an approximately 10% life time risk of developing active TB, compared to 60% or more in persons co-infected with HIV and mycobacterium tuberculosis. South Africa was ranked seventh highest by the World Health Organization among the 22 TB high burden countries in the world and fourth highest in Africa.

The research work in this thesis uses the 2007 Statistics South Africa (STATSSA) data on TB and HIV as the primary cause of death to build statistical models that can be used to investigate factors associated with death due to TB. Logistic regression, Survey Logistic regression and generalized linear models (GLM) will be used to assess the effect of risk factors or predictors to the probability of deaths associated with TB and HIV. This study will be guided by a theoretical approach to understanding factors associated with TB and HIV deaths. Bayesian modeling using WINBUGS will be used to assess spatial modeling of relative risk and spatial prior distributions for disease mapping models. Of the 615312 deceased, 546917 (89%) died from natural death, 14179 (2%) were stillborn and 54216 (9%) from non-natural death possibly accidents, murder, suicide. Among those who died from natural death and disease, 65052 (12%) died of TB and 13718 (2%) died of HIV. The results of the analysis revealed risk factors associated with TB and HIV mortality.

LIST OF SYMBOLS AND NOTATIONS

$\chi^2_{\alpha,n}$: Chi –square distribution with n degree of freedom

$\hat{\mu}$: Estimator of μ

$\hat{\sigma}^2$: Estimator of population variance

$\mathrm{E}(x)$: Expectation of x

$\mathrm{E}(y)$: Expectation of y

$F_{\alpha,m,n}$: F –distribution with m and n degrees of freedom

μ: Population mean

σ^2: Population variance

θ: Proportional parameter

$\hat{\alpha}$: Estimator of regression coefficient

$\hat{\beta}$: Estimator of regression coefficient

$\overline{X}$: Sample mean

α: Significance level of a test or probability of Type I error

$t_{\alpha,n}$: t- distribution with n degree of freedom

$\hat{\theta}$: Unbiased estimator of the parameter θ

Z_α: $pr(\mathrm{Z} > \mathrm{Z}_\alpha) = \alpha$

$\mathrm{E}(x/y)$: Expectation of x given y

$Var(x)$: Variance of x

$f(x_1, x_2, ..., x_n; \theta)$: Joint probability distribution

n: Sample size

LIST OF ABREVIATIONS

AID: Acquired Immune Deficiency Syndrome

DOTS: Directly Observed Treatment Shot course

GDP: Gross Domestic Product

GLM: Generalized Linear Models

GLMM: Generalized Linear Mixed Models

HIV: Human Immunodeficiency Virus

OR: Odds ratio

SLR: Survey Logistic Regression

STATA: Statistic Analysis

STATSSA: Statistics South Africa

TB: Tuberculosis

WHO: World Health Organization

Table of Contents

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION AND LITERATURE REVIEW

1.1 Background

Many countries in the world record annual summary statistics for economic indicators (such as Gross Domestic Product (GDP) and unemployment rate under Millennium Development Goals (MDGs) and vital statistics (such as the number of births and deaths). In particular, Statistics South Africa (STATSSA) collects annual data on nationwide number of deaths and associated causes. Tuberculosis (tubercle bacillus- TB) is an infectious disease caused by bacteria called Mycobacterium tuberculosis. These bacteria attack mainly the lungs (pulmonary TB), but also at lower extent other parts of the body such as the central nervous system, circulatory system, and the skeletal system (Khaled, 2008).

TB is the main cause of death in the world among all infectious diseases (Herchline and Amorosa, 2010). TB is classified as latent when it is not yet causing illness or active when illness has already been developed. Details can be found in Mzolo (2009). Despite advances in TB treatments which dramatically reduced TB cases up to the 1980s, the appearance of HIV/AIDS during the 1980s led to a rapid increase of TB, especially in the poorest parts of the world, mainly in Africa (Williams and Dye, 2003).

HIV/AIDS acts as catalyst to the occurrence of TB; hence it can dramatically increase the proportion of active TB cases. A study done in India by Vaidyanathan and Singh (2003) revealed that people infected with mycobacterium tuberculosis alone have an approximately 10% life time risk of developing active TB, compared to 60% or more in persons co-mortality with HIV and mycobacterium tuberculosis. In other words, regions with high rates of HIV/AIDS cases have

also high rates of active TB. A short report of WHO (2009) provides the following frightening data:

> "In 2008, there were an estimated 8.9–9.9 million incident cases of TB, 9.6–13.3 million prevalent cases of TB, 1.1–1.7 million deaths from TB among HIV-negative people and an additional 0.45–0.62 million TB deaths among HIV-cause of death people (classified as HIV deaths in the International Statistical Classification of Diseases), with best estimates of 9.4 million, 11.1 million, 1.3 million and 0.52 million, respectively".

Lawn (2010) states that because of HIV and TB co-infection, the WHO DOTS (Directly Observed Treatment Short course) program has failed to control TB in Sub-Saharan Africa, even in countries with good model of TB control such as Tanzania and Malawi.

The research work in this thesis uses the 2007 Statistics South Africa (STATSSA) data on TB and HIV as the primary causes of death to build statistical models that can be used to investigate factors associated with death due to TB and HIV.

1.2 Literature Review

1.2.1 The Problem

According to Singer (1997) the battle with TB in South Africa poses immense challenge to the government. The annual number of new TB cases in South Africa averages at 377 per 100,000 members of the population. Comparatively in other hard-hit parts of the world, the average is only about 200 per 100,000. Right now, approximately 10,000 people die of TB in South Africa every year. Singer (1997) argues further that in South Africa TB tends to affect the poorer populations, who have historically suffered a low standard of health care. But poverty is not the only contributing factor. Nearly two-thirds of the population of the country is infected with the

TB germ, thus approximately 160,000 South Africans from all walks of life become ill with TB every year. In 2006, South Africa was ranked seventh highest by the WHO among the 22 TB high burden countries in the world and fourth highest in Africa. In general, ensuring that patients adhere and complete their TB treatment has presented major challenges; treatment takes six to eight months, and patients often discontinue treatment before they are cured. The primary goal of the new national TB control programme is to ensure a high cure rate of infectious TB patients the first time around by insuring that they complete their treatment. In the strategic priorities for the National Health System set by the Department of Health for 2004-2009, the TB control programme is cited as achieving limited success, given its synergistic relationship with Human Immunodeficiency Virus (HIV) (SA, DoH, 2003). In South Africa responsibility for public health care is devolved to provinces among which the quality of TB control varies greatly. TB Treatment success remains low compared with other African countries with a higher prevalence of HIV and with considerably fewer resources (SA, DoH, 2003).

1.2.2 TB and HIV Interaction

According to the report published by Williams and Dye in 2003, HIV/AIDS has dramatically increased the incidence of TB in Sub-Saharan Africa where up to 60% of TB patients are co-infected with HIV and each year 200,000 TB deaths are attributed to HIV co-infection. In their report, they also indicate that antiretroviral (ARV) drugs can prevent TB by preserving immunity and that early therapy, plus high levels of coverage and compliance, will be needed to avert a significant fraction of TB cases. However, they assert that ARVs could enhance the treatment of TB while TB programmes provide an important entry point for the treatment of HIV/AIDS.

Corbett et al (2003) considers the decades leading up to 1980 when TB was in the decline throughout the world. However, as published by World Health Organization (WHO), in their

report on *Global Tuberculosis Control: Surveillance, Planning, Financing,* in 2003, 30% of people in Sub-Saharan Africa are latently infected with *Mycobacterium Tuberculosis* and the rapid spread of HIV during the 1980s and 1990s led to a similarly rapid increase in the incidence of TB, with notification rates in some countries increasing by more than five times in ten years. The report by UNAIDS in 2003 on the global HIV/AIDS epidemic presents the fact that HIV/AIDS control strategies have not substantially reduced the mortality of HIV in the Sub-Saharan Africa. Raviglione and Pio, (2002) suggest that the decline in immunity in people co-infected with HIV and TB has meant that even good TB control programmes based on short-course chemotherapy have not been sufficient to contain the rising incidence of TB (De Cock and Chaisson, 1999). Cohen (2002) argued that the development of new classes of ARV drugs, the availability of cheap generic equivalence, and the increasing commitment of international donors to making ARV drugs widely available in poor countries should all help to reduce HIV-related illness and deaths over the subsequent years (Tan, Upshur, and Ford, 2003). Whether ARVs have a significant impact on TB depends on their efficacy in preventing disease progression and prolonging life on population coverage and patient compliance. The impact also depends on the synergy between the treatment of TB patients and the provision of ARV therapy to those patients who are HIV infected.

As the TB and HIV pandemic continue to collide in sub-Saharan Africa resulting in increased incidence and mortality, the relative contribution to disease specific mortality of AIDS-related Smear-Negative Pulmonary TB (SNPTB) as result of increased incidence, under-recognition and diagnosis, and poor management practices is unknown (Getahun, Harrington, and Nunn, 2007).

In 2007 the World health Organization (WHO) published revised recommendations for the diagnosis of SNPTB to address the diagnostic and treatment challenges of HIV-associated TB in

resource-constrained settings. In South Africa, more than 16% of the population is infected with HIV, and 1000 people die from AIDS-related diseases each day, and two-thirds of those with HIV also suffer from TB, because of their weakened immune systems (AMREF, 2008). In 2004 estimates exceeded the 50% mark for TB patients living with HIV in South Africa (Dye, 2006). According to Bekker and Wood (2010), South Africa is believed to have the most people (approximately 1 million persons) living with both TB disease and HIV mortality. When the HIV epidemic set in, existing rates of latent TB mortality (LTBI) were extremely high in many communities, with over two-thirds of adults in poor South African communities for example, being infected. In those with HIV co-infection, subsequent risk of developing TB through reactivation of latent TB was extremely high, with overall rates reaching as high as 20-30% per year in those with the most advanced immunodeficiency.

Lawn (2010) state that DOTS does not reduce the very high susceptibility of HIV-infected individuals to develop rapidly progressive disease following exposure, thus although DOTS reduces transmission risk in the community, this may be out-weighed many-fold by the greatly increased risk of rapidly progressive disease in HIV-infected. Major increases in incidence rates of TB may further contribute to transmission, although this is off-set to some extent by the fact that HIV-associated TB cases are generally less infectious than disease cases in HIV-uninfected people.

Furthermore, the result that co-mortality with HIV significantly increases the risk of developing TB was established by Raviglione et al. (1997). However, as published by World Health Organization (WHO), on their report in (2000), the TB and HIV co-epidemic is increasing and will continue to fuel the TB epidemic.

1.2.3 Theoretical Framework

This study will be guided by a theoretical approach to understanding factors associated with TB death. Such factors can be grouped into specific categories as shown in the figure below:

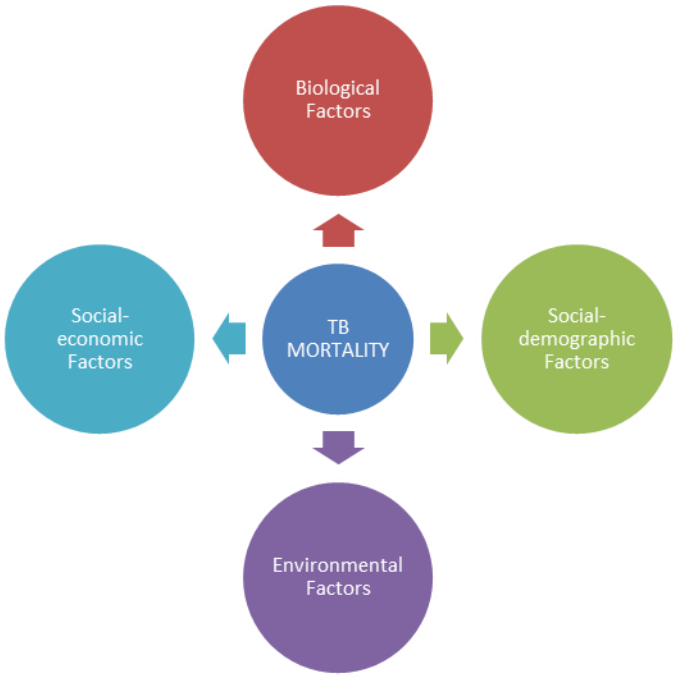


Figure 1: Factors associated with TB mortality.

1.2.3.1 Biological factors

Tuberculosis (TB) is one of the leading causes of death among individuals living with AIDS, not only because they are more susceptible to TB, but also because TB can increase the rate at which the AIDS virus replicate. One of the first indications of HIV mortality may be the sudden start of TB often in a site outside the lungs (extra-pulmonary TB). Individuals who have TB and also HIV infected are more likely to die from TB than any other deaths. TB can occur at any point in the course of progression of the HIV disease. The risk of developing TB rises sharply with decline in immune status. HIV promotes the rapid progression of latent TB death (LTBI) to active disease and is the most powerful known risk factor for the activation of latent TB (Uriz, Reparaz, and Sola, 2007).

1.2.3.2 Socioeconomic factors

Many studies have shown that factors that drive the TB epidemic are mostly socio-economic factors. Examples include education, occupation, and health status just to point out a few. TB is also associated with poverty. The majority of the poor in the world are likely to contract TB as a result of contributing factors such as lack of basic health services, poor nutrition and inadequate living conditions. It is evident that those who are exposed to conditions such as unemployment and living in crowded areas are more likely to be infected with the disease. The higher rates in poorer sectors of society are due not only to the poor housing and overcrowding brought about by urbanization and population increase, but also attributable to poor diets which lower resistance to the disease (Collins, 1981).

1.2.3.3 Environmental factors

Poor working environments may increases the risk of tuberculosis. For example working in the mines where shafts are poorly ventilated may facilitate easy spread of the TB bacteria. Incidence rates in prisons and homeless shelters are higher than that in general population. TB incidence is generally higher in urban than in rural areas. The tendencies for the burden of TB to be higher in urban than in rural areas may be due to high population density, crowded living and working conditions as well as life style changes associated with urban living. TB bacteria also can establish in nursing homes because older adults often have immune systems weakened by illness.

1.2.3.4 Socio-demographic factors

Demographic factors include age (expressed as a grouped variable), gender and marital status have been linked to TB infection. The TB epidemic in rural and urban areas is most severe for a variety of reasons including population dynamics where migrant mine and factories workers

carry the bacteria, back home during holidays and spread it to their households and surrounding areas (Zuma et al.,2005). Crowded living environments are also the effect of urbanization where people move to cities in search of work and most of these end up living in crowded informal settlements. In addition Zwang et al. (2007) stated that the co-infection of TB and HIV affects more males at an earlier age who are likely to be exposed to poor working environments that put them at risk of TB than females.

1.3 Methodology used in the study

1.3.1 Collection of data

The data used in this study is registration and records survey data on deaths from various causes gathered by Statistics South Africa in 2007. Our main interest is on deaths due to TB and HIV.

1.3.2 Statistical Analysis and Statistical Software

Exploratory analysis is performed using graphical displays and some basic summary statistics such as mean and the three quartiles as well as associated dispersion statistics in the form of tables.

Logistic regression, as a special case of the Generalized Linear Models (GLM), was used to assess the effect of risk factors or predictors to the probability of deaths associated with TB and HIV. Statistical modeling and analysis was done using STATA software.

1.4 Objectives of this thesis

1.4.1 General objective

The study aims to identify factors that can be used to explain TB and HIV mortality in South Africa. The work will also be concerned with statistical methods that can be best used to model

these associations, to identify factors associated with TB and HIV mortality in South Africa during the year 2007.

1.4.2 Specific objectives

1. To investigate and identify factors associated with TB and HIV death in South Africa using mortality gathered by STATS SA in 2007.
2. To apply logistic regression, a special case of generalized linear regression modeling, to relate a binary outcome namely death due to TB (HIV) to a number of predictor variables including the effect of HIV (TB) co-mortality.
3. To extend the regression model in Objective No. 2 to account for correlated data using survey logistic regression.
4. To extend the univariate modelling approach to a joint modeling of the two binary outcomes in one model as possible future study.
5. Suggest a spatial modelling approach to study the distribution of risk due to TB and HIV in South Africa.

1.5 Overview of the thesis

In addition to Chapter one which contains the introduction and literature review, Chapter two presents exploratory data analysis. Chapter three gives a brief review of generalized linear models, discusses important statistical issues in binary logistic regression modeling and the estimation of parameters involved. These models will also be used to model TB and HIV mortality and associated causes to achieve the research objectives. Chapter four will focus on data analysis and the interpretation of results. In Chapter five, we apply the Bayesian modeling and mapping using WinBUGS version 1.4. Finally, Chapter six will provide conclusions, implications, and avenues for future research work.

CHAPTER TWO

EXPLORATORY DATA ANALYSIS

2.1 Introduction

The data was sourced from Statistics South Africa consist of 615312 deaths from various causes in the year 2007. As a preliminary exploratory analysis, the use of tools such as cross tabulations and graphical displays will guide in understanding of important relationships.

Results from such an exploratory analysis will assist in building a more formal statistical model to understand the relationship between key predictor variables and the response variable. Our interest in the current work is death due to tuberculosis (TB) and HIV. The synergy between TB and HIV has attracted a huge interest in recent times.

However, in this study, the author most importantly considered only four variables among fourteen, namely those which have potential significant effect on TB death and HIV death defined as the presence or absence of the disease.

The four used variables are: age group, sex, death province and death Institution. It should be noted however that we cannot include the other ten variables in analysis such as death type, marital status, province of birth, province of residence, smoking status, pregnancy, HIV cause-related (self-reported), education level, occupation, and type of industry or business of work because there was a high rate of missing data in these variables.

2.2 Data Description

Table 2.0.Shows a description of the factor variables to be used and the codes assigned to the levels of each variable.

Table 2.0: Data description

Variable Name	Description
TB cause-related	Yes=1, No=0
HIV cause-related	Yes=1, No=0
Age group	0-15=1, 16-30=2, 31-45=3, 46-60=4, 61-75=5, 76-90=6, >90=7
Sex	Male=1, Female=2
Death Institution	Hospital (in-patient)=1, Emergency room/out-patient)=2, Death on arrival=3, Nursing home=4, Home=5,Other=6
Death province	Western Cape=1, Eastern Cape=2, Northern Cape=3, Free State=4, KwaZulu-Natal=5, North West=6, Gauteng=7, Mpumalanga=8, Limpopo=9, Outside South Africa=10

Note that Other=Unknown, not applicable and unspecified. Yes=TB-cause of death and HIV- cause of death; No=Non TB and HIV negative. A similar description of variables is used for HIV Data.

2.3 Exploratory Analysis of TB Deaths Data

In this section, an exploratory analysis of the TB data is presented. The interpretation and analysis presented in this section is based on a cross-tabulation analysis presented in Table 2.1. Of the 615312 deceased people, 546917 (88%) died from natural death and disease, among the deceased 65052 (12%) died of TB. The percentage of TB deaths among males is 11.18%, P <0.001 (see Table 2.1). The percentage of deaths among 0-15 years old is 2.29%, 16.49% for 16-30 years old, 19.05% for 31-45, 12% for 46-60 years old, 4.28% for 61-75 years old, 1.47% for 76-90 years, to late 50's old and 2.41% for above 90 years old. It shows that death due to TB appear to be in younger age groups (16-30 years, 31-45 years, 46-60 years) than older people, P < 0.001 (see Table 2.1).

Western Cape Province, and Gauteng to some extent, has, in general, lower risk of TB deaths than the other regions of South Africa while KwaZulu-Natal, Mpumalanga and Eastern Cape have higher risk of TB death. Risk of TB death also differs by death Institution. The results

indicate that risk of death was highest in hospital (in-patient) followed by emergency room (out-patient) and home, with rates of 14.53%, 9.1% and 8.55% respectively than other death Institutions.

It is observed from Figure 2.1 that overall 32% of TB Deaths occurred in KwaZulu-Natal, followed by Eastern Cape (16%) and Gauteng (13%). The lowest percentage of deaths occurred in Northern Cape (2%). Less than 1% of deaths registered were outside South Africa. It is important to note that the distribution of deaths by province of occurrence is largely similar to the distribution of the South African population by province.

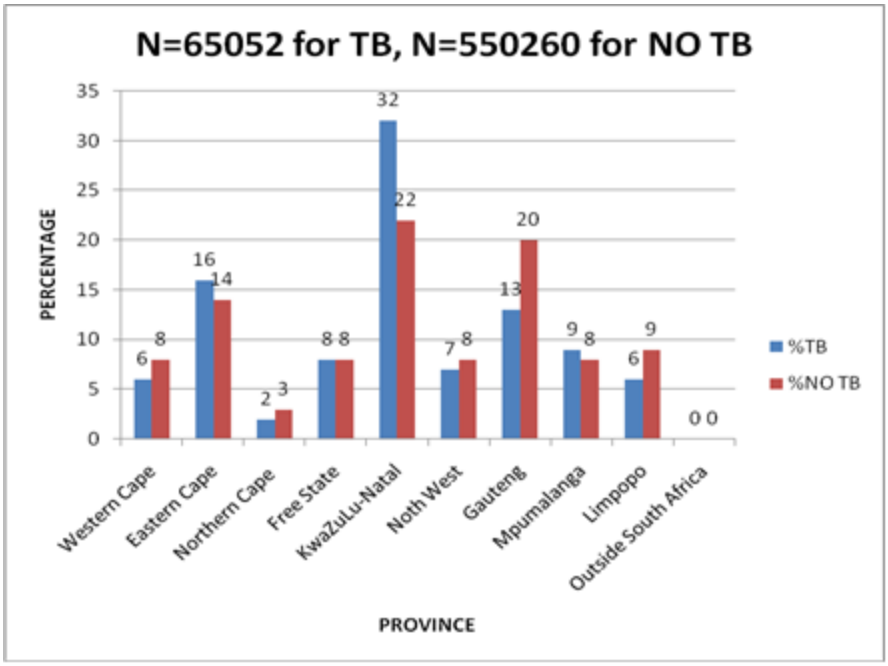


Figure 2 .1: **Percentage distribution of TB deaths by province of death occurrence, 2007**

Table 2.1: Percentage of TB and NON TB deaths, With P-values for Chi-Square test, According to selected Demographic, Social, Health status and life style

	NON TB	TB	N		NON TB	TB	N
Demographic/Provincial Characteristics				**Death province**			P<0.001
Age group			P<0.001	Western Cape	92.07	7.93	48091
				Eastern Cape	88.5	11.5	88200
0-15	97.71	2.29	86111	Northern Cape	91.2	8.8	15466
16-30	83.51	16.49	85939	Free State	89.73	10.27	52341
31-45	80.95	19.05	157694	KwaZuLu-Natal	85.42	14.58	142861
46-60	88.1	11.9	114469	Noth West	89.91	10.09	46331
61-75	95.72	4.28	93158	Gauteng	92.59	7.41	118449
76-90	98.53	1.47	66109	Mpumalanga	88.1	11.9	49168
>90	97.59	2.41	11832	Limpopo	92.24	7.76	53826
				Outside South Africa	90.5	9.5	579
Sex			P<0.001				
				Health status and life style			
Male	88.82	11.18	314138				
Female	90.05	9.95	299933	**Death Institution**			P<0.001
				Hospital (in-patient)	85.47	14.53	263962
				Emergency room/out-patient)	90.9	9.1	10672
				Death on arrival	93.63	6.37	15254
				Nursing home	94.81	5.19	12630
				Home	91.45	8.55	193850
				Other	93.66	6.34	118944

2.4 Exploratory Analysis of HIV Death Data

The data provided by Statistics South Africa consist of 615312 deaths from various causes in the year 2007. Of these deaths, the proportion of non-HIV cause related deaths was 97.77% and the proportion of HIV cause of death was 2.23%. HIV cause-related deaths varied according to different factors as described in this Section. Table 2.2 present the distribution of the number of HIV cause-related deaths by each variable. The mortality rate due to HIV was 2.23%, with the risk of HIV death also varying by age. The risk of HIV cause related death was highest among those in age group 31 to 45 with a rate of 4.32% followed by those in age group 16 to 30 with a

rate of 3.88%. Those who are in age group 76 to 90 and above were less likely to be infected with HIV. Thus the risk of HIV cause related death in this group is much lower. Results in Table 2.2 show that females are more likely to die of HIV than males. The HIV cause-related death rate for males and females were 1.95% and 2.53% respectively.

The analysis shows that in 2007 death due to HIV was highest in the Western Cape and KwaZulu-Natal with rates of 3.25% and 3.17% respectively. These were followed by Northern Cape with 2.32%. Limpopo province had the lowest risk of death due to HIV.

Examination of results in Table 2.2 indicates that the number of death by HIV is higher for hospitalized people 3.49%.

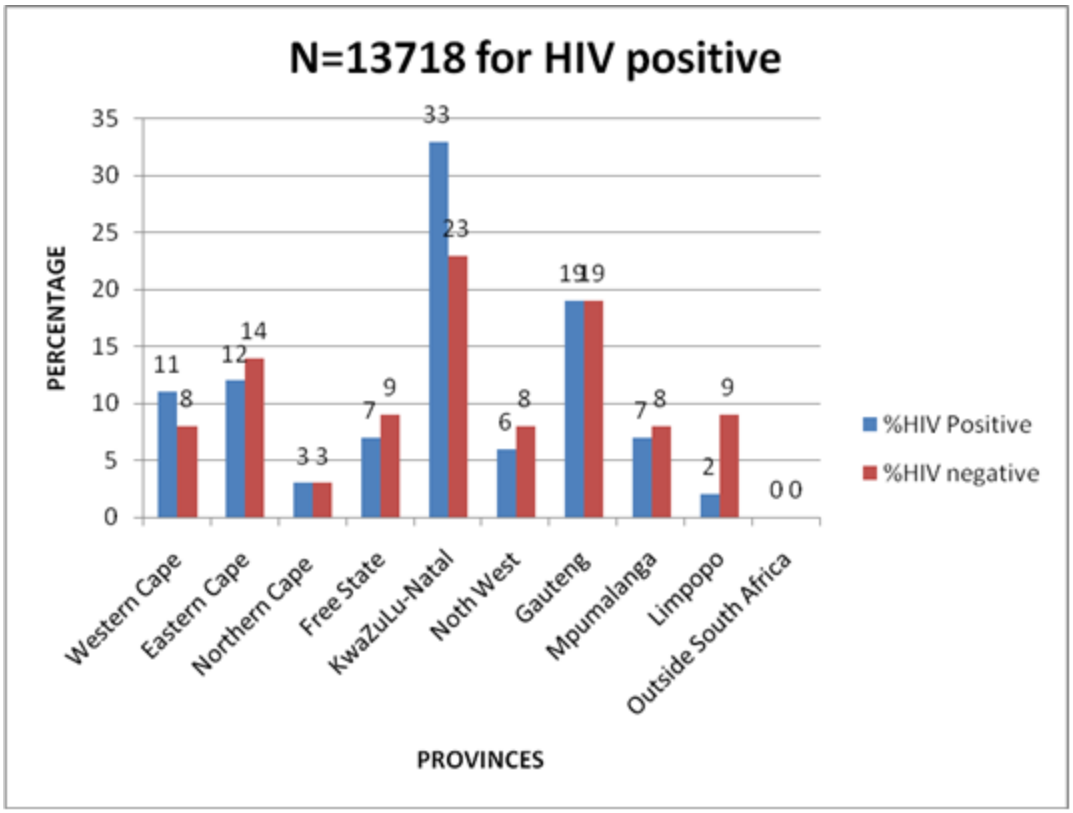


Figure 2.2: **The percentage distribution of HIV deaths by province of death occurrence, 2007.**

The Overall is 615312, for HIV negative N=601594 and N=13718 for HIV cause of death.

Table 2.2: **Percentage of Non HIV death and HIV cause of death, With P-values for Chi-Square test, According to selected Demographic, Social, Health status and life style**

	Non HIV death	HIV death	N
Demographic/Provincial Characteristics			
Age group			P<0.001
0-15	98.84	1.16	86111
16-30	96.12	3.88	85939
31-45	95.68	4.32	157694
46-60	98.06	1.94	114469
61-75	99.72	0.28	93158
76-90	99.95	0.05	66109
>90	99.52	0.48	11832
Sex			P<0.001
Male	98.05	1.95	314138
Female	97.47	2.53	299933
Death province			P<0.001
Western Cape	96.75	3.25	48091
Eastern Cape	98.16	1.84	88200
Northern Cape	97.68	2.32	15466
Free State	98.03	1.91	52341
KwaZuLu-Natal	96.83	3.17	142861
North West	98.36	1.64	46331
Gauteng	97.77	2.23	118449
Mpumalanga	98.05	1.95	49168
Limpopo	99.48	0.52	53826
			P<0.001
Death Institution			
Hospital (in-patient)	96.51	3.49	263962
Emergency room/out-patient)	97.03	2.97	10672
Death on arrival	98.55	1.45	15254
Nursing home	98.36	1.64	12630
Home	99.03	0.97	193850
Other	98.42	1.58	118944

2.5 Interaction of TB and HIV

Table 2.3 shows that the risk of TB death is higher among individuals infected with HIV compared to those who are HIV negative.

People who died of HIV related causes are at higher risk of TB death as the primary cause. The observed probability of dying of TB given HIV cause of death is 24% compared to 10% for non HIV related causes.

Table 2.3: **Two-way table showing the joint distribution of TB deaths by HIV deaths**

Variable	Category	TB	No TB	Total
HIV cause-related	**HIV negative**	61734 (10)	539860 (90)	601594
	HIV cause of death	3318 (24)	10400 (76)	13718
	Total	65052	550260	615312

The table shows that 24% were reported to have died due to co-mortality while 10% died of TB but not with HIV. The results shows that individuals who died of other causes of death (non TB) but with HIV related causes is 76% while those who died of TB alone with no HIV related cause was 10%.

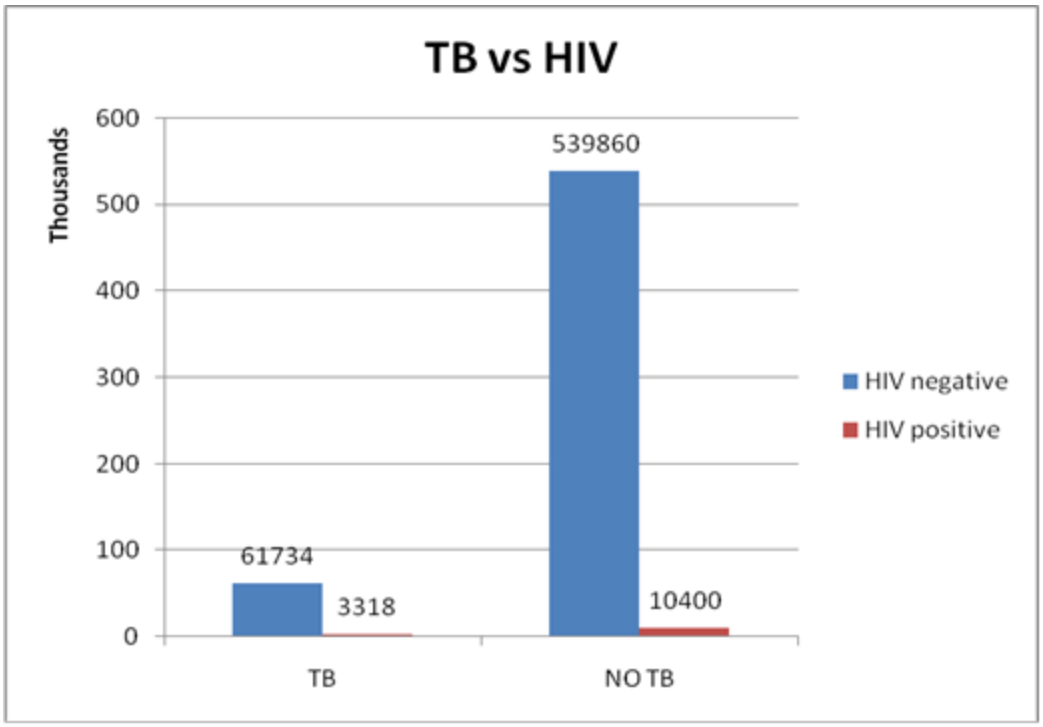


Figure 2.3: **The joint distribution of TB deaths by HIV deaths**

Figure 2.3 shows the effect of the joint dynamics of HIV and TB. From a disease modelling standpoint modelling co-mortality can present formidable mathematical challenges due to the fact that the models of transmission are quite intertwined. Furthermore, the fact that HIV activates TB an individual who died of TB could have been co-infected with HIV and vice-versa. Here the risk of TB and HIV mortality is give 24% corresponding to 3318 cases within TB cause related deaths.

2.6 Summary

2.6.1 Tuberculosis

The exploratory analysis carried out in this chapter indicates that the risk of TB death is higher among males than females. The possible reason is that males tend to work in environments that increase the risk of TB infection. One possible working environment is that males work in mines

more than females where shafts in mines are poorly ventilated and therefore facilitating very easy spread of TB bacteria. Returning migrant mine workers carry the bacteria back home during holidays and may possibly spread it to their surrounding areas. The preliminary results on TB death data indicate that TB cause-related seems to be higher among younger individuals. The reason is possibly due to the fact that younger individuals are more vulnerable to co-infection with HIV. The fact that TB is an opportunistic infection among HIV infected individuals may explain this correlation.

Individuals who live in informal settlements, or work in crowded environments, such as factories where there is a lot of pollution, or in crowded households tend to be at higher risk of contracting and dying of TB than other living and working condition.

2.6.2 Human Immune Virus

The exploratory analysis carried out in this chapter indicates that HIV cause related death is high among females than males. Possible reasons for this include the fact that women are exposed to sexual abuse, rape and commercial sex activities for survival which expose them to HIV. A possible biological reason for a high HIV transmission rate in females is that females have a larger cervical area which makes it easier for HIV to establish itself in females than in males. The cause of death in young individuals could be due to the fact that they are more sexually active and inexperienced which lead them to be at higher risk of HIV infection hence high HIV mortality. Low levels of education, poverty, overcrowding and unemployment are much associated with the less knowledge about HIV/AIDS.

CHAPTER THREE

GENERALIZED LINEAR MODELS

3.1 The Exponential Family

Generalized Linear Models (GLMs) are an extension of the classical linear models and are used to model observations on random variables having a distribution belonging to the exponential family of distributions. If the probability density function (p.d.f.) of the i-th observation from a random sample of size n from a random variable Y is given by

$$f(y_i;\theta_i,\phi)=\exp\left\{\frac{y_i\theta_i-b(\theta_i)}{a\ (\phi)}+c(y_i,\phi)\right\}, \tag{3.1}$$

where a, b and c are known functions, then *f(.)* is said to belong to the exponential family. The parameter θ_i is called the "natural location parameter" whilst ϕ is the "dispersion parameter". Many known distributions belong to the exponential family (e.g. normal, binomial and Poisson distributions). The mean and variance of Y_i are respectively given by

$$\mu = E(Y_i) = b'(\theta_i), \tag{3.2}$$

and

$$\sigma^2 = \mathrm{Var}(Y_i) = a(\phi)b''(\theta_i). \tag{3.3}$$

McCullagh and Nelder (1989) and Myers, Montgomery and Vining (2002, pp.157-160) provide a detailed theoretical background of these models.

In particular, the publication by McCullagh and Nelder (1989) is the most referenced book on generalized linear models (GLMs). The idea was first developed by Nelder and Wedderburn

(1972) and extended later by Dobson (1990), with discussion on the theory and application of such models, to numerous application areas.

In order to discuss the use of GLMs to regression problems, let us consider n independent observations $y_1, y_2, \cdots, y_n$ of a random variable Y.

Let $\mu_i = \mathrm{E}(Y_i)$ and suppose each of the y_i depend on a set of predictor variables or explanatory variables $x_1, x_2, \ldots, x_p$, also called covariates in application areas such as medical research. We aim to estimate or fit the model of the form

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Then formally, the random variable Y is said to conform to a generalized linear model (GLM) if it has the following three conditions hold:

(1) Each realization y_i of Y belongs to an exponential family of distributions with p.d.f. of the form (1) for which the natural parameter is $\theta_i, i = 1,2,\cdots,n.$ θ_i is considered to be a function of $\beta = (\beta_0, \beta_1, \cdots, \beta_p)^T$ with $p < n.$

(2) A linear predictor $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$ which is a linear combination of values of the explanatory variables.

(3) A monotonic function called link function $g(\mu_i) = \eta_i$ between the mean response $\mu_i = E(Y_i)$ and the linear predictor η_i for $i = 1,2,\cdots,n.$

If $\theta_i = \eta_i = g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$, then g is called a canonical link. (McCullagh and Nelder, 1989; Myers, Montgomery and Vining, 2002, p.161).

To see how the idea of a canonical link arises let $Y_1, Y_2, \ldots, Y_n$ be independently distributed observations such that Y_i is distributed as $\beta(1, \mathrm{P}_i)$ then,

$$f(y_i; p_i) = p_i^{y_i}(1-p_i)^{1-y_i} \quad i = 1,2,\ldots n.$$

Clearly the p.d.f above can be re-written as an exponential family because,

$$f(y_i; p_i) = \exp\left\{ y_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i)\right\}$$

It follows that

$$\theta_i = \log\left(\frac{p_i}{1-p_i}\right)$$

therefore

$$p_i = \frac{e^{\theta_i}}{\left(1+e^{\theta_i}\right)}$$

and

$$bi(\theta_i) = \log\left(\frac{1}{1-p_i}\right) = \log\left(1+e^{\theta_i}\right)$$

Note that since in this case

$$\mu_i = \mathrm{E}(\mathrm{Y}_i) = \eta_i$$

then

$$\theta_i = \log\left(\frac{\mu_i}{1-\mu_i}\right)$$

The function $g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$ is called the canonical link function. Therefore, if

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip},$$

it implies that

$$p_i = \mu_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}}.$$

Since also $\text{var}(Y_i) = p_i(1 - p_i)$ it can easily be shown that

$$b'(\theta_i) = p_i$$

and

$$b''(\theta_i) = p_i(1 - p_i).$$

Some well-known distributions and their associated canonical link functions are tabulated below.

Table 3: **Common distributions with corresponding link functions for constructing generalized linear models**

Distribution	**Link function**
Normal	Identity link: $\eta_i = \mu_i$
Binomial	Probit link: $\eta_i = \Phi^{-1}(\mu_i)$, where Φ is the cumulative function of the standard normal distribution
	Logit link: $\eta_i = \ln \frac{p_i}{1 - p_i}$
	Complementary log-log link: $\eta_i = \ln(-\ln(1 - \mu_i))$
	Power link: $\eta_i = \begin{cases} \mu_i^{\lambda}, \lambda \neq 0 \\ \ln(\mu_i), \lambda = 0 \end{cases}$
Poisson	Log link: $\eta_i = \ln(\mu_i)$
Gamma	Reciprocal link: $\eta_i = \frac{1}{\mu_i}$

Source: Myers, Montgomery and Vining (2002, p.162).

As shown in the Table 3, it is assumed the link function (denoted by g) is a monotonic and differentiable function which links the mean response $\mu_i = \mathrm{E}(y_i)$ and the linear predictor $\eta_i = x_i'\beta$. If θ_i equals η_i, the link function is called a canonical link function. If $g(\mu_i)$ corresponding to $\theta_i = \theta_i(\mu_i)$. Thus each member of the exponential family of distributions has a unique canonical link function. For example, the canonical link function for Binomial (or Binary) data is the logit link given by

$$\theta_i = \theta_i(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$$

where

$$\mu_i = p_i.$$

The generalized linear model for independent Bernoulli observations with logit link is referred to as the logistic regression model. 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.

3.2 Estimation of Parameters

Parameter estimation for generalized linear models is done using the method of maximum likelihood. It follows from equation (3.2) that the log-likelihood of a generalized linear model can be written as

$$l = \frac{1}{a(\phi)}\sum_{i=1}^{n}[(y_i\theta_i - b(\theta_i)) + c(y_i,\phi)] \qquad (3.4)$$

(Myers, Montgomery and vining, 2002, p.163).

Consider the case of a GLM with canonical link function of the form

$$\theta_i = \eta_i = g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\,.$$

Estimates of the parameters $\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_p)$ are computed by differentiating the log-likelihood function given by equation (3.4) with respect to β and then solving the system of equations $\frac{\partial l}{\partial \beta} = 0$. This leads to the score equations given by

$\sum_{i=1}^{n} (y_i - \mu_i)\mathrm{x}_{\mathrm{ij}} = 0$ for $j = 0,1,\ldots, p$ and $x_{i0} = 1 \;\forall\; i$, denoting the first column of X.

Thus this system of $p+1$ equations can be written in matrix form as

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & \vdots & & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} y_1 - \mu_1 \\ \vdots \\ y_n - \mu_n \end{pmatrix}$$

$$X^T (y - \mu) = 0 \qquad (3.5)$$

where X is a $n \times (p+1)$ design matrix, $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ is an $n \times 1$ vector of observations and $\mu = (\mu_1, \mu_1, \cdots, \mu_n)^T$ is the $n \times 1$ vector of expected mean responses.

The simultaneous systems of equations (3.5) are solved iteratively using for example the Taylor approximation. After convergence the asymptotic variance-covariance matrix of $\hat{\beta}$ is given by

$$\mathrm{Var}(\hat{\beta}) = (\mathrm{X}^{\mathrm{T}} W X)^{-1} \qquad (3.6)$$

Where W is the $n \times n$ diagonal matrix with $(i,i)th$ element given by

$$w_{ii} = \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \qquad (3.7)$$

3.3 Model Checking

3.3.1Goodness-of-fit Test

The log-likelihood-ratio (deviance) and the Pearson's chi-square statistics are the main tools used for assessing the goodness-of–fit of the fitted generalized linear model (Agresti, 2002). They measure the discrepancy of fit between the maximum log-likelihood achievable and the achieved log-likelihood by the fitted model. The most commonly used measure in GLMs called deviance, is defined as

$$D(y,\hat{\mu}) = 2\{\ell(y;y) - \ell(\hat{\mu};y)\} \tag{3.8}$$

where $\ell(\hat{\mu};y)$ is the log-likelihood under the model of interest and $\ell(y;y)$ is the log-likelihood under the maximum achievable (saturated) model (Agresti, 2002, p.118). Under the hypothesis that the model is correct, the deviance (3.8) has a chi-square distribution with $n-p$ degrees of freedom where n is the number of observations and p is the number of model predictor variables (Myers, Montgomery and Vining, 2002, p.134). For a binomial model such as the one we are dealing with defined by

$$P(Y=y) = \frac{n!}{y!(n-y)!}\, p^{y}(1-p)^{n-y},\ y = 0,1,2,\cdots n \tag{3.9}$$

the deviance (3.8) for binomial data is given by

$$D = 2\left\{\sum_{i=1}^{n} n_i y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + \sum_{i=1}^{n} n_i (1-y_i)\ln\left(\frac{1-y_i}{1-\hat{\mu}_i}\right)\right\} = 2\sum \text{observed} \times \ln(\text{observed/fitted})$$

(Agresti, 2002, pp.140-141).

3.4 Model Selection and Diagnostics

3.4.1 Model Selection

There can be a number of models in the family of generalized linear models that describe a given data set. Therefore, it is necessary to select the simplest rational model that sufficiently describes the particular data (Agresti, 1990). As in most applications including the current one there can be many variables under consideration. In this case the stepwise selection procedure is mostly preferred because it has an advantage of minimizing the chances of keeping redundant variables and leaving out important variable in the model. In all the procedures, a variable that leads to a significant change in the deviance (given by equation 3.8) when added to or dropped from the model is retained, otherwise it is dropped. This method of model selection is referred to as deviance analysis and is used to test the model for the goodness-of –fit.

3.5 Logistic Regression Model (LRM)

The logistic regression model (LRM) is a member of generalized linear models used to model binary data and its main properties will be discussed because it will be the main application tool in analysis of the mortality data in the thesis. Consider n independent observations y_i of a binary random variable Y_i taking values 1 for a success and 0 for a failure. Each realization y_i of Y_i is said to follow a Bernoulli distribution with probability density function given by

$$f(y_i) = p^{y_i}(1-p)^{1-y_i}, y = 0,1,$$

where p is the probability of success, i.e. $p = P(Y_i = 1)$. For n independent Bernoulli trials, the number of successes $Y = \sum_{i=1}^{n} Y_i$ follows a binomial distribution with probability density function given by

$$\binom{n}{y} p^y (1-p)^{n-y};\ y = 0,1,2,\ldots,n.$$

Thus let us consider a binomial random variable Y with parameters n and p. Given a set of explanatory variables $x_1, x_2, \cdots, x_p$ assumed to have an effect on the response y, the probability of response $p = P(Y = y \mid x_1, x_2, \cdots, x_p)$ is said to follows a logistic distribution if

$$p(\mathbf{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)} \tag{3.10}$$

or in terms of the logit function as

$$\text{logit}(p(x)) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{3.11}$$

where $\beta_0, \beta_1, \beta_2, \cdots \beta_p$ are unknown model parameters to be estimated (Agresti, 2002, p.182). The predictor variables $x_1, x_2, \cdots, x_p$ can be continuous (example, age) or categorical (example, sex, marital status). The parameters $\beta_0, \beta_1, \beta_2, \cdots \beta_p$ are interpreted as log odds ratios with respect to the reference level of the factor variable under consideration.

Thus, parameter estimates and associated variance-covariance matrix are calculated using equations (3.5), (3.6) and (3.7) earlier stated.

3.5.1 Fitting a logistic regression model

The fitting of a logistic regression model is exactly the same as for any Generalized Linear Models for binomial but with n fixed at $n=1$. Therefore the details of its fitting process will not be repeated here but for interested readers the book by Agresti (2002) is recommended. As already stated in Section **3.2,** the estimating equations for a GLM particularly the case of the logistic regression model are readily solved using iterative methods generally installed in statistical packages such as SAS, Genstat, and SPSS. As in (3.6), the variance-covariance matrix of the vector $\underline{\hat{\beta}} = \left(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p\right)^T$ is given by

$$\hat{V}(\underline{\hat{\beta}}) = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1},$$

where **W** is the $k \times k$ diagonal matrix with diagonal elements $n_i \hat{p}_i (1-\hat{p}_i)$ for $i = 1,2,\cdots,k$ (Agresti, 2002). Hence, the standard error of $\hat{\beta}_j$ is $\sqrt{n_i \hat{p}_i (1-\hat{p}_i)}$. As a consequence, a $(1-\alpha)\times 100\%$ confidence interval of β_j is $\hat{\beta}_j \pm t_{\frac{\alpha}{2}, \nu}\, se(\hat{\beta}_j)$ where $se(\hat{\beta}_j) = \sqrt{n_i \hat{p}_i (1-\hat{p}_i)}$, and $t_{\frac{\alpha}{2}, \nu}$ is the value of the t-distribution on ν =k-1 degrees of freedom at the left of which the area under the curve or distribution is $1-\frac{\alpha}{2}$

3.5.2 Odds ratios

For interpretation of regression parameters in the logistic regression model given by (3.11), many researchers prefer reporting odds ratios than the direct model parameter $\hat{\beta}_j, j = 1,2,\ldots,p$. In general, in the case of a binomial distribution with probability of success p, the odds of a success is defined as

$$O = \frac{\text{prob of success}}{\text{prob of failure}} = \frac{p}{1-p}.$$

For two probabilities of success p_1 and p_2, the ratio of the associated odds O_1 and O_2 is called odds ratio and is given by

$$\psi = \frac{O_1}{O_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}.$$

Clearly, the logistic regression defined in terms of logit (3.11) is a log (odds).

To explain the dependence of the odds ratio on covariates, consider the special case of one categorical explanatory variable x, for example exposure status with value $x = 0$ for unexposed and $x = 1$ for exposed. Then, from equation (3.10) assuming $p(x)$ is the probability of infection by a disease, we have

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{3.12}$$

or equivalently

$$\log it(p(x)) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x. \tag{3.13}$$

The odds of disease for those exposed $(x = 1)$ is $O_1 = \frac{p_1}{1-p_1} = \exp(\beta_0 + \beta_1)$.

Likewise the odds of disease for those unexposed $(x = 0)$ is $O_2 = \frac{p_2}{1-p_2} = \exp(\beta_0)$.

Finally the odds ratio of exposed relative to unexposed is now given by

$$\psi = \frac{O_1}{O_2} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1).$$

Hence, the odds ratio comparing the two odds of disease is the exponential of the slope parameter β_1 in model (3.13) or likewise β_1 is the log odds ratio, $\log(\psi)$.

The calculations of odds ratio in the case of a binary explanatory variable such as exposure status (exposed versus unexposed) can be generalized to the case of a categorical variable with l levels where $l \geq 3$. In such a situation one level is taken as the reference, and model (3.13) can be extended to the case of multiple linear logistic regression model with $l-1$ dummy variables $x_1, x_2, \cdots, x_{l-1}$ where $x_i = 1$ if level i is considered, otherwise $x_i = 0$ for $i = 1,2,\cdots,l-1$. Model (3.13) becomes

$$\log it(p(x)) = \ln \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{l-1} x_{l-1}$$

(see Agresti, 2002, p.178). Note that term model is used for $p(x)$ because the equations describe the probability of success $p(x)$ in terms of the covariates. The odds ratio associated with level i relative to the reference is calculated in the same way as for the case of a one variable with two levels except here it has to be interpreted conditional on the other variables held constant.

Now, consider the case when x is a continuous variable, we can calculate the odds of an event When x increases by one unit relative to the odds when x remains unchanged. The two odds at $x+1$ and x are respectively

$$O_1 = \frac{p_1}{1-p_1} = \frac{\dfrac{\exp[\beta_0 + \beta_1(x+1)]}{1+\exp[\beta_0 + \beta_1(x+1)]}}{\dfrac{1}{1+\exp[\beta_0 + \beta_1(x+1)]}} = \exp[\beta_0 + \beta_1(x+1)].$$

and similarly

$$O_2 = \frac{p_2}{1-p_2} = \exp(\beta_0 + \beta_1 x).$$

It follows that the odds ratio is then given by

$$\psi = \frac{O_1}{O_2} = \frac{\exp[\beta_0 + \beta_1(x+1)]}{\exp(\beta_0 + \beta_1 x)} = \exp(\beta_1).$$

Here, the odds ratio is again the exponential of the slope parameter, and can be interpreted as the ratio of the odds when x increases by one unit. It immediately follows from $\psi = \exp(\beta_1)$ that $\beta_1 = \ln(\psi)$, i.e. the slope parameter β_1 can be interpreted as the natural logarithm of the odds ratio ψ. Hence, if $(1-\alpha)\times 100\%$ confidence interval of β_1 is $\hat{\beta}_1 \pm t_{\frac{\alpha}{2}, v}\, se(\hat{\beta}_1)$, then a $(1-\alpha)\times 100\%$ confidence interval of the odds ratio ψ is $\{\exp[\hat{\beta}_1 \pm t_{\frac{\alpha}{2}, v}\, se(\hat{\beta}_1)]\}$ where $\hat{\beta}_1$ is the estimate of β_1, α is the significance level (often taken as 0.05), $se(\hat{\beta}_1)$ is the standard error of $\hat{\beta}_1$, $t_{(\frac{\alpha}{2}, v)}$ is appropriately read or derived from the t-distribution quantiles. Now, consider the multiple logistic regression model (3.10) or equivalently model (3.11).

The interpretation of the slope parameter β_1 in the case of the one-variable binary logistic regression model (3.12) can be extended to the case of multiple logistic regression model (3.10). If an explanatory variable X_j is continuous, the parameter β_j in model (10) is the increase in natural logarithm of odds ratio at $X_j = x_j + 1$ relative to $X_j = x_j$ when the other $p-1$ variables are maintained unchanged or constant. If an explanatory variable X_j is categorical with l levels, the parameter β_k can be interpreted as the increase in natural logarithm of odds ratio at level k relative to the reference level of X_j with $k = 1,2,\cdots,l-1$ and $j = 1,2,\cdots,p$. Confidence intervals of parameters and odds ratios are calculated in similar way as for the case of one explanatory variable.

3.6 Model Selection and Diagnostics for LRM

The same procedures discussed in section 3.4 for model selection apply here. For ungrouped binary data, the deviance statistic D (or D^*) is used only to select variables and not as a measure

of goodness-of-fit. Hosmer and Lemeshow (1989) proposed and discussed the goodness-of-fit as explained in the next section. In the section inappropriateness of the deviance statistic as a measure of goodness-of-fit test will also be discussed.

3.7 Model checking

3.7.1 Goodness-of-fit Test

Recall that the deviance is given by (3.8) as

$$D(y,\hat{\mu}) = 2\{\ell(y;y) - \ell(\hat{\mu};y)\} \tag{3.14}$$

Where $\ell(\hat{\mu};y)$ is the log-likelihood under the current model and $\ell(y;y)$ is the log-likelihood under the maximum achievable (saturated) model. We consider the typical case of grouped data where the ith group has m_i observations in it instead of the case ungrouped binary data. Suppose generally $Y_i \sim \text{Bin}(m_i, \pi_i)$, then $E(Y_i) = m_i\pi_i = \mu_i$. The likelihood function is

$$\prod_{i=1}^{n} f(y_i;\pi_i) = \prod_{i=1}^{n} \frac{m_i!}{y_i!(m_i-y_i)!}\pi_i^{y_i}(1-\pi_i)^{m_i-y_i}.$$

The log-likelihood is

$$\ell(\mu;y) = \ln\prod_{i=1}^{n}\frac{m_i!}{y_i!(m_i-y_i)!} + \sum_{i=1}^{n} y_i\ln(\pi_i) + \sum_{i=1}^{n}(m_i-y_i)\ln(1-\pi_i)$$

$$= \ln\prod_{i=1}^{n}\frac{m_i!}{y_i!(m_i-y_i)!} + \sum_{i=1}^{n} y_i\ln\left(\frac{m_i\pi_i}{m_i}\right) + \sum_{i=1}^{n}(m_i-y_i)\ln\left[\frac{m_i-m_i\pi_i}{m_i}\right]$$

$$= \ln\prod_{i=1}^{n}\frac{m_i!}{y_i!(m_i-y_i)!} + \sum_{i=1}^{n} y_i\ln\left(\frac{\mu_i}{m_i}\right) + \sum_{i=1}^{n}(m_i-y_i)\ln\left[\frac{m_i-\mu_i}{m_i}\right].$$

Therefore, the log-likelihood for the fitted model is

$$\ell(\hat{\mu}; y) = \ln \prod_{i=1}^{n} \frac{m_i!}{y_i!(m_i - y_i)!} + \sum_{i=1}^{n} y_i \ln\left(\frac{\hat{\mu}_i}{m_i}\right) + \sum_{i=1}^{n} (m_i - y_i) \ln\left[\frac{m_i - \hat{\mu}_i}{m_i}\right]$$

The log-likelihood for the maximal (saturated) model ($\hat{\mu}_i = y_i$) is

$$\ell(y; y) = \ln \prod_{i=1}^{n} \frac{m_i!}{y_i!(m_i - y_i)!} + \sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{m_i}\right) + \sum_{i=1}^{n} (m_i - y_i) \ln\left[\frac{m_i - y_i}{m_i}\right]$$

Substituting $\ell(\hat{\mu}; y)$ and $\ell(y; y)$ in equation (3.14) gives

$$\begin{aligned} D &= -2\left[\ln \prod_{i=1}^{n} \frac{m_i!}{y_i!(m_i - y_i)!} + \sum_{i=1}^{n} y_i \ln\left(\frac{\hat{\mu}_i}{m_i}\right) + \sum_{i=1}^{n} (m_i - y_i) \ln\left(\frac{m_i - \hat{\mu}_i}{m_i}\right),\right. \\ &= -2\left\{\ln \prod_{i=1}^{n} \frac{m_i!}{y_i(m_i - y_i)!} + \sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{m_i}\right) + \sum_{i=1}^{n} (m_i - y_i) \ln\left(\frac{m_i - y_i}{m_i}\right)\right\}. \end{aligned}$$

With rearrangement of terms, D becomes

$$\begin{aligned} D &= -2\left[\sum_{i=1}^{n} y_i \ln\left[\frac{\hat{\mu}_i}{m_i} \times \frac{m_i}{y_i}\right] + \sum_{i=1}^{n} (m_i - y_i) \ln\left[\frac{m_i - \hat{\mu}_i}{m_i} \times \frac{m_i}{m_i - y_i}\right]\right] \\ &= -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\mu}_i}{y_i}\right) + (m_i - y_i) \ln\left(\frac{m_i - \hat{\mu}_i}{m_i - y_i}\right)\right], \\ &= -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i - y_i) \ln\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right)\right] \end{aligned}$$

Recall in the case of ungrouped binary outcomes $m_i = 1$, for all i, therefore D becomes

$$\begin{aligned} D &= -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1 - y_i) \ln\left(\frac{1 - \hat{\pi}_i}{1 - y_i}\right)\right] \\ &= -2\sum_{i=1}^{n}\left[y_i \ln(\hat{\pi}_i) + (1 - y_i) \ln(1 - \hat{\pi}_i)\right] \end{aligned}$$

because $y_i \ln y_i = 0$ and $(1 - y_i)\ln(1 - y_i) = 0$ if $y_i = 0$ or 1. After rearrangement of terms D becomes

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i}\right) + \ln(1 - \hat{\pi}_i)\right] \qquad (3.15)$$

Since

$$\sum_{i=1}^{n} y_i \ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \sum_{i=1}^{n} \hat{\pi}_i \ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right)$$

the equation (3.17) can be written as

$$D = -2\sum_{i=1}^{n}\left[\ \hat{\pi}_i \ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) + \ln\left(1-\hat{\pi}_i\right)\ \right] \qquad (3.16)$$

A detailed discussion about the deviance can be found in Collett (2003) and in McCullagh and Nelder (1989). It is pointed out that the deviance cannot be used as a measure of goodness-of-fit of the model for ungrouped binary data (see Myers, Montgomery and vining, 2002). Note that in all the model construction and formulation we evaluate the outcome or response at the population and not at the individual level which is not the focus of the current thesis.

3.8 Cluster Survey Logistic Regression Model (CSLRM)

3.8.1 Introduction

Logistic regression models used to analyze data from the complex surveys is referred to in the literature as Cluster Survey Logistic Regression Models (CSLRMs) to distinguish them from the ordinary logistic regression models discussed above. Survey logistic regression models follow the same theory as ordinary logistic regression models. The exception is that they account for the complexity of survey designs. When data are from simple random sampling, the survey logistic regression model and the ordinary logistic regression model are identical. In addition to that, Cluster survey logistic regression models obtain more reliable estimates compared to simple logistic regression which fail to take into account features of population leading to inference results that are statistically unreliable particularly if within cluster correlation is large.

3.8.2 The Model (CSLRM)

In order to concisely define the models consider the problem of disease prevalence in epidemiology.

Let $\pi_{ijh} = p(y_{ijh} = 1)$ be the probability that the disease is present and $1 - \pi_{ijh} = p(y_{ijh} = 0)$ that it is not present in the i^{th} observation or individual within the j^{th} primary sampling unit (PSU) nested within the h^{th} stratum $(i = 1.2,\ldots, m_{hj};\ j = 1,2,\ldots, n_h; h = 1,2,\ldots, \mathrm{H})$. In this case the log-likelihood function is given by

$$l(\beta; y) = \sum_{h=1}^{H}\sum_{j=1}^{n_h}\sum_{i=1}^{m_{hj}}\left\{ y_{ijh}\log\left(\frac{\pi_{ijh}}{1-\pi_{ijh}}\right) - \log\left(\frac{1}{1-\pi_{ijh}}\right)\right\} \tag{3.17}$$

Thus in general the survey logistic regression model is given by

$$\text{logit}(\pi_{ijh}) = \mathrm{X}'_{ijh}\beta, \qquad i = 1,2,\ldots, m_{hj};\ j = 1,2,\ldots, n_h; h = 1,2,\ldots, \mathrm{H} \tag{3.18}$$

where X_{ijh} is the row of the design matrix corresponding to the characteristics of the i^{th} observation in the j^{th} PSU within h^{th} stratum, and β is a vector of unknown parameters of the model.

3.8.3 Estimation of parameters for CSLRM

We refer to sections 3.2 and 3.3 for discussion of the method of maximum likelihood estimation used to estimate parameters of the model. Calculation of the standard errors of the parameter estimates, which are used to perform appropriate statistical tests and construct confidence intervals for the parameters, when data come from complex design, is complicated. The covariance matrix of parameter estimates under the CLRM can be obtained through the Taylor expansion approximation procedure (Vittinghoff et al., 2005). This technique estimates variance taking account of variation among clusters and computes the overall variance estimate by

pooling stratum variance estimates together. There are other methods of variance estimation for complex survey data other than the Taylor expansion approximation (also known as linearization method). These methods are based sample re-use principle. Key among them are the jackknife and the bootstrap methods (see Vittinghoff, 2005; Lehtonen and Pahkinen, 1995; and Skinner, Holt, and Smith, 1989). The jackknife and bootstrap methods are illustrated with examples in Lehtonen and Pahkinen (1995). Currently only the Taylor expansion approximation will be used. The degrees of freedom for the t-test statistics used for testing the significance of the parameters equals the number of clusters minus the number of strata in the sample survey design. This statistic can then be used to construct confidence intervals for the parameters, especially if n (the overall sample size) is small. When n is large, as is the case with our data, the sampling distribution of the parameter estimators are approximated by a normal distribution. Hence, the Wald chi-square statistic can also be used to test for the significance of the parameters and to construct confidence intervals (which are also called normal confidence intervals) given by

$$\hat{\beta}_j \pm Z_{\frac{\alpha}{2}} \sqrt{v_{jj}}, \tag{3.19}$$

where $z_{\frac{\alpha}{2}}$ is the $100\left(1-\frac{\alpha}{2}\right)$th percentile of the standard normal distribution, and v_{jj} is the variance of $\hat{\beta}_j$ given by diagonal elements of variance-covariance matrix of $\hat{\beta}$ after model convergence. Note again that these intervals are on a logit scale, if the canonical link function is used.

Fortunately, the trouble of calculating estimates and their variance has been circumvented by implementation of the procedures in statistical packages that appropriately account for the complexity of survey designs. This procedure is implemented in packages such as STATA 11, under survey logistic regression (SLR). It was developed basically for fitting a linear logistic

regression model for discrete response variables to survey data. When the data are from the simple random sampling method, SLR is identical to the standard logistic regression. Maximum likelihood estimation method and the Taylor expansion approximations procedure will be used to fit the model in chapter 4. For SLR to be used it requires that there be at least two or more clusters per stratum, otherwise the stratum will not make any contribution in the estimation.

CHAPTER FOUR

APPLICATION OF THE LOGISTIC REGRESSION MODELS

The model discussed in Chapter 3 under Section 3.5, was fitted using TB and HIV as the binary response Y with the four variables discussed and presented in Chapter 2 as predictor variables. The survey logistic regression (SLR) model was fitted with province of death as the clustering variable.

The STATA commands or code used are listed in appendix A.1

4.1 Interpretation of Model results for TB Death Data

Table 4.1 contains the odds ratios, standard errors, 95% confidence intervals and their p-value of the incidence of the disease. Cluster and stratification design variables were built-in using STATA SURVEYLOGISTIC procedure. Note that survey logistic regression models have the same theory as ordinary logistic regression models, with the exception that they account for the complexity of survey designs. When data are from simple random sampling, the survey logistic regression model and the ordinary logistic regression model are identical (Heeringa, West, and Berglund, 2010).

Results for both simple and cluster survey logistic regression models, with TB mortality was the response variable, are presented in Table 4.1. The results show that the odds ratio of death due to TB for females to males is 0.8784 with a 95% confidence interval of 0.8642- 0.8928, (P <0.001). This means that females are at lower risk of TB death than males. The corresponding estimates under survey logistic regression are OR = 0.8784, 95% CI 0.8449- 0.9132, P < 0.001. The parameter estimates for the ORs is the same in both approaches but the standard errors are different with the one from survey logistic regression larger than that from the simple logistic

regression. Thus the danger of using ordinary logistic regression with clustered data is that Type I error may be committed more often than when survey logistic regression is used.

The age group 31-45years has the highest odds of TB death. The results from both simple and survey logistic regression models are (OR=10.0222, 95% CI 9.5683- 10.4975, P <0.001), and (OR=10.0222, 95% CI 7.9326-12.6622, P<0.001). For age groups 16-30 the respective results are (OR=8.4066, 95% CI 8.0118-8.8209, P <0.001) and (OR=8.4066, 95%CI 6.8986-10.2442, P <0.001). The age group with least odds ratio is age groups 76 to 90 whose OR point estimate 0.636 with a 95% confidence interval of 0.5886-0.6872 and under the simple logistic regression (P < 0.001) and (OR=0.636, 95% CI 0.481-0.841, P < 0.005) for cluster survey logistic.

Examination of the odds ratios in Table 4.1 indicate that the proportion of death by TB is higher for hospitalized people. The simple interpretation is that TB infected seek medical care than just remaining at home. In summary it should be noted that results from tables 4.1 shows that the odds ratios are the same for simple logistic regression and for clustering survey logistic regression, but the standard error and confidence intervals are significantly different. The standard errors are higher under cluster logistic regression.

Table 4.1: Logistic Regression for TB death data

	Simple Logistic Regression				Cluster Survey Logistic Regression			
Variable	**OR (Std.Err)**	**95% CI**		**p-value**	**OR (Std.Err)**	**95% CI**		**p-value**
Age group								
0-15	REF							
16-30	8.406(0.206)	8.0118	8.8209	< 0.001	8.406(0.735)	6.8986	10.244	<0.001
31-45	10.02(0.237)	9.5683	10.497	< 0.001	10.022(1.035)	7.9326	12.662	<0.001
46-60	5.749(0.141)	5.4796	6.0324	< 0.001	5.749(0.632)	4.4839	7.372	<0.001
61-75	1.902(0.053)	1.8009	2.0093	< 0.001	1.902(0.231)	1.4449	2.5042	<0.001
76-90	0.64(0.025)	0.5886	0.6872	< 0.001	0.636(0.078)	0.481	0.841	0.005
>90	1.051(0.67)	0.9268	1.1917	0.439	1.051(0.179)	0.7148	1.545	0.777
Sex								
Male	REF							
Female	0.878(0.007)	0.8642	0.8928	<0.001	0.878(0.015)	0.8449	0.9132	<0.001
Death Institution								
Hospital (in-patient)	REF							
Emergency room/out-patient)	0.5888(0.02)	0.5507	0.6295	<0.001	0.5888(0.024)	0.5375	0.6450	<0.001
Death on arrival	0.3999(0.0131)	0.3744	0.4272	<0.001	0.3999(0.062)	0.2810	0.5691	<0.001
Nursing home	0.3218(0.0132)	0.2972	0.3483	<0.001	0.3218(0.057)	0.2152	0.4811	<0.001
Home	0.5497(0.0051)	0.5392	0.5604	<0.001	0.5497(0.046)	0.4545	0.6648	<0.001
Other	0.3979(0.0052)	0.3878	0.4082	<0.001	0.3979(0.039)	0.3171	0.4993	<0.001

4.2 Interpretation of Model results for HIV Death Data

The same model as that for TB death was fitted for HIV cause-related at death. The HIV model was also fitted using survey logistic regression a procedure in STATA. The results in Table 4.2 contain the odds ratios, Standard error, 95% confidence intervals and p-values of HIV cause-related at death. The odds of death with HIV if you are a female is 1.3075 times that of males with a 95% confidence interval of 1.2637-1.3529, and P <0.001 and the results under CSLR are (OR = 1.3075, 95% CI 1.1877-1.44, P < 0.001). This means that the HIV related death burden is more in females than in males. Overall point estimates of ORs are the same but the standard errors are larger and confidence intervals wider under both the CSLR.

Individual in age group 31-45 have higher odds of dying with HIV than any other age group. This result is confirmed by logistic regression with estimates (OR =3.8399, 95% CI, 3.5914-4.1054, P <0.001), and for cluster survey logistic regression OR=3.8399, 95% CI 3.1847-4.63, P<0.001). For age group 16-30 the results are (OR=3.4349, 95% CI 3.1986-3.6886, P <0.001) under the simple logistic regression and OR=3.4349, 95%CI 2.9854-3.952, P <0.001) under cluster survey logistic. The odds rations of the age groups 61-75, 76-90 and > 90 are all estimated as less than one in both the simple and cluster logistic regression models confirming that these age groups die less with HIV related causes than the younger age groups where the impact of HIV infection is most felt.

Table 4.2: Simple and Survey Logistic regression for HIV death data

	Simple logistic regression				Survey logistic regression			
Variable	**OR(Std.Err)**	**95% CI**		**p-value**	**OR (Std.Err)**	**95% CI**		**p-value**
Age group								
0-15	REF							
16-30	3.4349(0.125)	3.1986	3.6886	<0.001	3.43499(0.213)	2.9854	3.952	<0.001
31-45	3.8399(0.131)	3.5914	4.1054	<0.001	3.8399(0.317)	3.1847	4.63	<0.001
46-60	1.6816(0.064)	1.5598	1.8128	<0.001	1.6816(0.184)	1.3121	2.155	0.001
61-75	0.237(0.016)	0.2067	0.2718	<0.001	0.237(0.040)	0.1614	0.348	<0.001
76-90	0.0386(0.007)	0.0268	0.0555	<0.001	0.0386(0.011)	0.0198	0.075	<0.001
>90	0.4116(0.056)	0.3149	0.5378	<0.001	0.4116(0.174)	0.1578	1.073	0.066
Sex								
Male	REF							
Female	1.3075(0.023)	1.2637	1.3529	<0.001	1.3075(0.556)	1.1877	1.44	<0.001
Death Institution								
Hospital (in-patient)	REF							
Emergency room/out-patient)	0.8457(0.049)	0.7548	0.9474	0.004	0.8457(0.101)	0.6468	1.1059	0.191
Death on arrival	0.4061(0.028)	0.355	0.4645	<0.001	0.4061(0.097)	0.2354	0.7007	0.005
Nursing home	0.4603(0.033)	0.4006	0.5289	<0.001	0.4603(0.079)	0.3107	0.6819	0.002
Home	0.2699(0.007)	0.2568	0.2838	<0.001	0.2699(0.045)	0.1845	0.395	<0.001
Other	0.4427(0.011)	0.4211	0.4654	<0.001	0.4427(0.157)	0.1974	0.9925	0.048

4.3 Interpretation of Multiple Logistic Regression Model for TB Death

Tables 4.3 contain the odds ratios, p-value and their confidence intervals of the odds of TB death using the multiple logistic regression models. The tables were constructed from the estimated model in Appendix A1. It is important to note that both simple and multiple logistic regression assess the association between the independent variables or sometimes called predictor variables, but the simple logistic regression is mostly used as an exploratory tool for associations between one (dichotomous) outcome and one (categorical) exposure variable while multiple logistic regression is used to explore associations between one (dichotomous) outcome variable and two or more exposure variables (which may be continuous, ordinal or categorical). The purpose of multiple logistic regressions is to let the user to isolate the relationship between the outcome variable and the effects of one or more variables conditional on other factors being present.

Since the odds ratios produced by Multiple Logistic regression are the same as those given by survey logistic regression, interpretation given in Section 4.1 also apply here. The only difference is the confidence intervals, standard errors and p-value.

It can be seen that multiple logistic regressions (Table 4.3) confirm this difference. We note that the odds ratio of death due to TB for females to males is 0.867 with a 95% confidence interval of 0.8513- 0.8830, with P <0.001 for the multiple logistic regression and OR = 0.867, 95% CI 0.8472- 0.8873, P < 0.001 for survey multiple logistic regression. The odds of TB death is highest in the age group 31-45 years old followed by those in age group 16- 30 years old. The smallest OR is that for individuals in the age groups 76 - 90 and >90 years old. This result is also confirmed by the multiple logistic regression results for the age group 31-45 years (OR = 9.9218, 95% CI 9.3628- 10.5143, P <0.001), and for the survey multiple logistic regression model the point estimate is OR =9.9218, 95% CI 7.4085-13.2878, and P<0.001.

Table 4.3 shows that the odds of death with TB is higher in hospitals than any other health institution. This should be interpreted to mean that TB infected individuals do go to hospitals more than any other institution seeking for treatment. This result is in line with what one would expect compared to results from the univariate analysis. In general confidence interval are wider under survey multiple logistic regression than multiple logistic regression. Thus again it is clear the survey multiple logistic regression is more protective to type I error than the multiple logistic regression. One notable difference between the univariate analysis compared to the multivariate analysis is that the latter assess significance of predictor variables accounting for the presence of other covariates in the models therefore leading to more reliable results and conclusions.

Table 4.3: Multiple Logistic regression for TB death

	Multiple logistic regression				survey : Multiple Logistic regression			
Variable	OR(Std.Err)	95% CI		p-value	OR(Std.Err)	95% CI		p-value
Age group								
0-15	REF							
16-30	8.369(0.25)	7.8844	8.8844	<0.001	8.36(0.92)	6.5207	10.7424	<0.001
31-45	9.92(0.29)	9.3628	10.5143	<0.001	9.92(1.28)	7.4085	13.2878	<0.001
46-60	6.20(0.18)	5.8439	6.5789	<0.001	6.20(0.76)	4.6905	8.1968	<0.001
61-75	2.248(0.07)	2.1075	2.3995	<0.001	2.25(0.29)	1.6664	3.0347	<0.001
76-90	0.82(0.03)	0.76	0.9019	<0.001	0.82(0.12)	0.5951	1.1518	0.228
>90	1.46(0.09)	1.2876	1.6689	<0.001	1.46(0.27)	0.9645	2.2284	0.069
Sex								
Male	REF							
Female	0.86(0.008)	0.8513	0.883	<0.001	0.86(0.01)	0.8472	0.8873	<0.001
Death Institution								
Hospital (in-patient)	REF							
Emergency room/out-patient)	0.57(0.02)	0.5325	0.6111	<0.001	0.57(0.04)	0.4926	0.6604	<0.001
Death on arrival	0.43(0.01)	0.4003	0.4588	<0.001	0.43(0.05)	0.3297	0.5569	<0.001
Nursing home	0.64(0.03)	0.5942	0.7018	<0.001	0.64(0.09)	0.4691	0.8891	0.013
Home	0.59(0.01)	0.5877	0.6121	<0.001	0.59(0.05)	0.5000	0.7195	<0.001
Other	0.38(0.01)	0.3741	0.3955	<0.001	0.38(0.03)	0.3235	0.4574	<0.001
Province of death								
Western Cape	REF							
Eastern Cape	0.74(0.03)	0.6871	0.8027	<0.001	0.74(0.04)	0.6640	0.8305	<0.001
Northern Cape	0.75(0.05)	0.6502	0.8633	<0.001	0.75(0.11)	0.5334	1.0524	0.087
Free State	0.89(0.05)	0.8012	0.9991	0.048	0.89(0.09)	0.6981	1.1468	0.337
KwaZuLu-Natal	0.96(0.03)	0.8961	1.0318	0.276	0.96(0.03)	0.8859	1.0437	0.307
North West	1.04(0.04)	0.9546	1.1327	0.37	1.04(0.04)	0.9438	1.1456	0.385
Gauteng	0.72(0.03)	0.6689	0.7785	<0.001	0.72(0.04)	0.6282	0.8290	<0.001
Mpumalanga	0.98(0.04)	0.8990	1.0742	0.701	0.98(0.06)	0.8618	1.1206	0.771
Limpopo	0.83(0.04)	0.7555	0.9136	<0.001	0.83(0.07)	0.6908	0.9993	0.049
Outside South Africa	1.07(0.16)	0.8022	1.4438	0.624	1.07(0.04)	0.9873	1.1731	0.086

4.4 Interpretation of Multiple Logistic Regression Model for HIV Death

In the results in Table 4.4 below, similar interpretation procedure as for TB is done. The results show that among age groups, those who are in 31 to 45 years old are more likely to die due to HIV related causes than those in the age group 76 to 90 years old. The odds ratio for the 31-45 age group is 3.8596, 95% CI 3.488-4.2704, P-value < 0.001 for multiple logistic and (OR=3.8596, 95%CI 3.1960-4.6610, P-value < 0.001) for the survey multiple logistic. The odds ratios of HIV related mortality in both types of multivariate logistic regression models for those in age groups 61-75 years, 76-90 years and > 90 are all less than one confirming the univariate analysis result that HIV related mortality is higher in the younger age groups since this is where the impact of HIV infection is most experience.

The odds of HIV related mortality for females 1.4154 times of males with a 95% confidence intervals (1.364-1.469, with P <0.001) the multiple logistic regression and (1.2818-1.563, P < 0.001) for survey multiple logistic. This again is a confirmation of the result found under the univariate analysis that females experience a higher burden of HIV related mortality than males.

Table 4.4: MULTIPLE AND SURVEY LOGISTIC REGRESSION FOR HIV

	MULTIPLE LOGISTIC REGRESSION				SURVEY : MULTIPLE LOGISTIC REGRESSION			
Variable	OR(Std.Err)	95% CI		p-value	OR(Std.Err)	95% CI		p-value
Age group								
0-15	REF							
16-30	3.43(0.18)	3.096	3.8159	<0.001	3.43(0.27)	2.8681	4.1192	<0.001
31-45	3.85(0.19)	3.488	4.2704	<0.001	3.85(0.32)	3.196	4.661	<0.001
46-60	1.80(0.09)	1.62	2.0038	<0.001	1.80(0.24)	1.3247	2.4509	0.002
61-75	0.27(0.02)	0.233	0.3183	<0.001	0.27(0.07)	0.1501	0.4936	0.001
76-90	0.04(0.01)	0.031	0.0658	<0.001	0.04(0.02)	0.0203	0.1009	<0.001
>90	0.47(0.07)	0.358	0.6255	<0.001	0.47(0.22)	0.1653	1.3544	0.142
Sex								
Male	REF							
Female	1.41(0.03)	1.364	1.469	<0.001	1.41(0.06)	1.2818	1.563	<0.001
Death Institution								
Hospital (in-patient)	REF							
Emergency room/out-patient)	0.82(0.05)	0.737	0.9287	0.001	0.82(0.07)	0.6771	1.0104	0.061
Death on arrival	0.39(0.03)	0.348	0.4583	<0.001	0.39(0.09)	0.2409	0.6622	0.003
Nursing home	0.96(0.07)	0.834	1.112	0.604	0.96(0.18)	0.6257	1.4811	0.846
Home	0.34(0.008)	0.325	0.3599	<0.001	0.34(0.07)	0.2184	0.5356	<0.001
Other	0.45(0.01)	0.43	0.4791	<0.001	0.45(0.16)	0.2057	1.0018	0.05
Death province								
Western Cape	REF							
Eastern Cape	0.63(0.04)	0.553	0.7283	<0.001	0.63(0.05)	0.5277	0.7637	<0.001
Northern Cape	0.51(0.07)	0.39	0.6745	<0.001	0.51(0.12)	0.2989	0.8802	0.021
Free State	0.43(0.05)	0.352	0.5425	<0.001	0.43(0.09)	0.2669	0.7153	0.004
KwaZuLu-Natal	0.61(0.04)	0.542	0.6871	<0.001	0.61(0.06)	0.4846	0.7684	0.001
North West	0.53(0.04)	0.459	0.6309	<0.001	0.53(0.13)	0.3083	0.9396	0.033
Gauteng	0.62(0.04)	0.552	0.7129	<0.001	0.62(0.03)	0.5606	0.7013	<0.001
Mpumalanga	0.42(0.04)	0.358	0.5069	<0.001	0.42(0.08)	0.2741	0.6615	0.002
Limpopo	0.27(0.03)	0.224	0.3481	<0.001	0.27(0.02)	0.2416	0.3219	<0.001
Outside South Africa	0.39(0.15)	0.185	0.8452	0.017	0.39(0.04)	0.3202	0.4886	<0.001

4.5 TB death and HIV cause of death Co-mortality

As mentioned previously, TB and HIV are closely interlinked. TB is a leading cause of HIV-related morbidity and mortality. HIV is the most important factor fuelling the TB plague in populations with a high HIV death.

Table 4.5 below shows the distribution of TB deaths and death with HIV according to different levels of key important factor variable. The table shows the synchronization between these two causes of morbidity and mortality. The distribution by age shows that death due to TB and death with HIV is high in the same age group of 31-45 year old with rates of 16.49 % and 3.88 % respectively. The distribution according to sex shows that males are at higher risk of TB death (11.18%) compared to females (9.95 %), but for HIV it is females who are at higher risk of death with HIV with rates of (2.53%) for females and (1.95%) for males.

Table 4.5: TB and HIV Co-mortality

Parameter	Total(N=65052)	TB+(%)	Total (N=13718)	HIV+(%)
Age group				
0-15	1976	2.29	1001	1.16
16-30	14170	16.49	3337	3.88
31-45	30046	19.05	6814	4.32
46-60	13618	11.90	2220	1.94
61-75	3984	4.28	259	0.28
76-90	973	1.47	30	0.05
>90	285	2.41	57	0.48
Sex				
Male	35115	11.18	6111	1.95
Female	29856	9.95	7584	2.53
Death province				
Western Cape	3813	7.93	1563	3.25
Eastern Cape	10146	11.50	1622	1.84
Northern Cape	1361	8.80	359	2.32
Free State	5373	10.27	998	1.91
KwaZulu-Natal	20832	14.58	4528	3.17
North West	4674	10.09	762	1.64
Gauteng	8773	7.41	2643	2.23
Mpumalanga	5850	11.90	958	1.95
Limpopo	4175	7.76	278	0.52
Outside South Africa	55	9.50	7	1.21
Death Institution				
Hospital (in-patient)	38352	14.53	9221	3.49
Emergency room/out-patient)	971	9.10	317	2.97
Death on arrival	971	6.37	221	1.45
Nursing home	655	5.19	207	1.64
Home	16567	8.55	1876	0.97
Other	7536	6.34	1876	1.58

The pattern presented under death province shows that KwaZulu-Natal has highest deaths due to TB followed by Mpumalanga, Eastern Cape, Free State and North West. On the other hand Western Cape seems to have high death with HIV.

TB Death and HIV Death results

In this section we discuss results for fitting TB and HIV jointly. In this analysis if TB and HIV were both reported as causes of death either as primary cause or secondary cause that observation was given a value of 1 and all other pairs given a value of 0. This binary variables was then used as the response variables as the analysis.

The predictor variables used are those that were used in the simple and multiple logistic regression models that is age, sex, and death institution. The model was fitted in STATA using multivariate logistic regression and survey multivariate logistic regression. The results for both types of adjusted logistic regressions are given in Table 4.6. The estimates of odds ratios are the same in both models but the cluster survey multivariate logistic regression model estimates have higher standard errors as expected. The risk of death due to co-mortality is highest in the age group 31-45 years (OR=10.25, 95% CI, 7.63-13.77 under the survey multivariate logistic regression). The odds ratios and the 95% confidence intervals for the age groups 16-30, 46-60, 61-75, 76-90, > 90 are respectively 8.52 (95% CI: 6.62, 10.98), 6.27 (95% CI: 4.73, 8.29), 2.28 (95%CI: 1.68, 3.10), 0.85 (95% CI: 0.61,1.18) and 1.42 (95% CI:0.95, 2.11). There seems to be a decreasing odds death due to co-mortality as we move to higher age groups from the 31-45 years age group. The odds of death due to co-mortality is lower for females compared to males (OR=0.87, 95% CI: 0.85-0.88).In the results tables parameter estimates and their standard errors are given rather than odds. The standard errors can be used to calculate the confidence interval limits for the odds ratios by first finding confidence limits of the parameter estimates which represent the log odds ratios.

Table 4.6: **MULTIVARIATE LOGISTIC REGRESSION AND SURVEY MULTIVARIATE LOGISTIC REGRESSION FOR TB DEATH AND HIV DEATH CO-MORTALITY**

	Multivariate logistic regression				Survey multivariate logistic regression			
Parameter	OR(Std. Err)	95% CI		p-value	OR(Std.Err)	95% CI		p-value
Age group								
0-15	REF							
16-30	8.52(0.27)	8.02	9.06	<0.001	8.52(0.95)	6.62	10.98	<0.001
31-45	10.25(0.31)	9.65	10.87	<0.001	10.25(1.34)	7.63	13.77	<0.001
46-60	6.27(0.19)	5.89	6.66	<0.001	6.27(0.78)	4.73	8.29	<0.001
61-75	2.28(0.08)	2.14	2.44	<0.001	2.28(0.31)	1.68	3.1	<0.001
76-90	0.85(0.04)	0.78	0.92	<0.001	0.85(0.12)	0.61	1.18	0.283
>90	1.42(0.09)	1.24	1.62	<0.001	1.42(0.25)	0.95	2.11	0.079
Sex								
Male	REF							
Female	0.87(0.008)	0.85	0.88	<0.001	0.87(0.01)	0.85	0.89	<0.001

CHAPTER FIVE

BAYESIAN MODELLING AND MAPPING USING WINBUGS

5.1 Introduction

In previous Chapters a frequentist likelihood approach has been applied in modeling TB and HIV mortality data. These methods include the logistic regression under the generalized linear model (GLM) and Survey logistic regression (SLR). In this chapter we focus on the Bayesian spatial disease approach which requires spatial prior distributions for model parameters and information or data to estimate the posterior distributions. Prior distributions and data likelihood provide two sources of information about any problem. The likelihood informs the model about the parameter via the data, while the prior distribution informs the model via prior beliefs or assumptions about the model parameters. Let θ denote the vector of parameters of interest in the likelihood and let y denote the data. The product of the likelihood and the prior distributions is called the posterior distribution defined as:

$$p(\theta \mid y) \propto L(y \mid \theta) g(\theta),$$

where $g(\theta)$ is the prior distribution of the parameter vector θ.

When the sample size is large and the data is informative about cause specific mortality; the likelihood will contribute more to the relative risk estimation (Lawson et al., 2003, p.28). Disease and mortality mapping studies aim to summarize spatial variation in disease risk or generally in the risk of an event, in order to asses and quantify the amount of true spatial heterogeneity and the associated patterns, to help infer about areas of elevated or lowered risk. Much more work has been done in disease mapping therefore the area is relatively more

developed compared to other application areas. In this work it was assumed that the observed deaths (O_i) for each census area $(i = 1,\ldots,n)$ follow a Poisson distribution with mean $\mu_i = E_i\theta_i$, where E_i are the expected cases for each census area obtained by indirect standardization, and θ_i is the relative risk (RR) for each specific area. The expected number of cases, E_i was computed via

$$E_i = rN_i,$$

where $r = \frac{\sum_{i=1}^{9} O_i}{\sum_{i=1}^{9} N_i}$ was the overall risk of death in the population and N_i was the population at risk in the i-th province. We first calculated the RR through conventional Poisson models with TB deaths and HIV deaths as individual categorical variable.

In order to take into account both extra-Poisson variability and spatial correlation, smoothed RR estimators were also obtained in an entirely Bayesian approach, using the Poisson generalized linear mixed models with two random effects. The structured spatial component is modeled by including a neighborhood (adjacent) structure that reflects the effect of factors with a greater action scope than the spatial unit that smoothly varies with provinces. In this chapter the focus will be on spatial models for smoothing area data.

5.2 Spatial models for smoothing area data

Given the observed number of deaths in an area, O_i, and the expected number of deaths, E_i, it is assumed that the observed cases O_i follow a Poisson distribution that is:

$$O_i \sim Poisson(\theta_i E_i), \quad i = 1,2,\ldots,n \qquad (5.1)$$

therefore the likelihood of the relative risk θ_i is

$$L(\mathrm{O}_i \mid \theta_i) = \frac{e^{-\theta_i \mathrm{E}_i} (\theta_i \mathrm{E}_i)^{\mathrm{O}_i}}{\mathrm{O}_i!} \tag{5.2}$$

It follows that the maximum likelihood estimate of the relative risk of mortality in area i, θ_i, is given by:

$$\hat{\theta}_i = \frac{\mathrm{O}_i}{\mathrm{E}_i} \tag{5.3}$$

with the estimated standard error of $\hat{\theta}_i$ given by

$$\hat{\delta}_i = \frac{\sqrt{\hat{\theta}_i}}{\mathrm{E}_i} \tag{5.4}$$

5.3 Model fitting and interpretation of results

In this chapter we will fit four models namely the Poisson gamma, Poisson generalized linear mixed model, Spatial model and Spatial convolution model in order to compare which one is the best. In what follows, we consider first the likelihood models in equation (5.2) for case event data which allows the application of Poisson-process models in this analysis. The probability model given by equation (5.1) is the classic model assumed in many disease mapping studies involving counts and similarly to equation (5.3) which gives its likelihood. The log-likelihood associated with this model given by:

$$l = \sum_{i=1}^{n} \mathrm{O}_i \ln(\mathrm{E}_i \theta_i) - \sum_{i=1}^{n} \mathrm{E}_i \theta_i \tag{5.5}$$

By differentiating and equating the derivative to zero we get the maximum likelihood estimator of θ_i as just $\frac{\mathrm{O}_i}{\mathrm{E}_i}$, which is the SMR (Lawson et al., 2003, p.19), where SMR mean the standardized mortality ratio.

5.3.1 Poisson-Gamma model

When the data likelihood is Poisson it is assumed that there is a common relative risk parameter which follows a single gamma prior and hence the posterior distribution is given by:

$$p(\theta \mid y) \propto L(y \mid \theta) g(\theta),$$

where $g(\theta)$ is gamma distributed with parameters $\alpha, \beta,$ or $Gamma(\alpha, \beta)$, and therefore ignoring terms which do not depend on θ the data likelihood is given by

$$L(\mathrm{O} \mid \theta) = \prod_{i=1}^{n} \left\{ (\mathrm{E}_i \theta)^{\mathrm{O}_i} \exp(\mathrm{E}_i \theta) \right\}$$

dependent only on a single parameter θ with a gamma prior. A Bayesian specification for this model is:

$$\mathrm{O}_i \mid \theta \sim Poisson(\mathrm{E}_i \theta),$$
$$\theta \sim Gamma(\alpha, \beta).$$

Also note that in the notation of multilevel models, it is common to use $\lambda_i = \mathrm{E}_i \theta_i$.

Consider $\mathrm{Y}_i \sim Poisson(\lambda), i = 1, \ldots, n.$ The likelihood is given by

$$L = f(y \mid \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} = \frac{e^{-n\lambda} \lambda^{n\bar{y}}}{\prod_{i=1}^{n} y_i!}$$

To account for heterogeneity in the Poisson rates we assume a Gamma(α, β) prior distribution for λ. That is, $\lambda \sim Gamma(\alpha, \beta)$, so that the probability density function of λ is

$$f(\lambda) = \frac{1}{\Gamma(\alpha)\beta} e^{\lambda/\beta} \lambda^{\alpha - 1}$$

The posterior distribution then follows immediately as

$$f(\lambda \mid y) \propto f(y \mid \lambda) f(\lambda)$$
$$\propto e^{\lambda\left(n + \frac{1}{\beta}\right)} \lambda^{n\bar{y} + \alpha - 1}$$

The Gamma distribution is conjugate to the Poisson distribution, therefore posterior mean is then

$$E(\lambda \mid y) = \frac{n\bar{y} + \alpha}{n + \frac{1}{\beta}}$$

and likewise the posterior variance is given by

$$\text{var}(\lambda \mid y) = \frac{n\bar{y} + \alpha}{\left(n + \frac{1}{\beta}\right)^2}.$$

Thus in effect

$$\lambda \mid y \sim Gamma\left(n\bar{y} + \alpha, \left(n + \frac{1}{\beta}\right)^{-1}\right).$$

5.3.2 Poisson-Gamma with hyper parameters for α and β

In the Poisson-gamma model, θ has a $Gamma(\alpha, \beta)$ distribution at the first level of the hierarchy. The parameters α and β will have a hyper prior distributions h_α and h_β, respectively, at the second level of the hierarchy. This hierarchical structure can then be written as:

$$y_i \mid \theta \sim Poisson(\mathrm{E}_i\theta),$$

$$\theta \mid \alpha, \beta \sim Gamma(\alpha, \beta),$$
$$\alpha \mid \nu \sim h_\alpha(\nu).$$
$$\beta \mid \rho \sim h_\beta(\rho).$$

At this point the parameters are assumed to be fixed. For example in this model if we assume α, β fixed then the gamma prior would be fixed. However, by allowing a higher level of variation, it means hyper-prior for $\alpha, \beta,$ we can fix the value of ν and ρ without heavily influencing the lower-level variation. In this case, the relative risks have posterior distribution given by

$$\theta_i \sim Gamma(y_i + \alpha, \mathrm{E}_i + \beta)$$

and the posterior expectation of θ_i based on a single count observation is

$(y_i + \alpha) \big/ (\mathrm{E}_i + \beta)$.

5.3.3 Poisson-Gamma spatial moving average (convolution) model

A conjugate Poisson-gamma spatial moving average distribution can be specified for non-negative counts defined on a spatial lattice using the distribution Pois.conv in WinBUGS version 1.4.

Spatial moving average models have been developed primarily for continuous spatial processes and are carried out in WinBUGS 1.4. Suppose we have a set of area-specific spatially correlated Poisson count data (or random effects) O_i, $i = 1, \ldots, n$ (where n is the number of areas in the study region). It assumed that the counts O_i are conditionally independent (given area mean λ_i):

$$\mathrm{O}_i \sim \text{Poisson}\ (\lambda_i\ )$$

The model for each is constructed by specifying an arbitrary grid of latent i-th gamma random variables $\gamma_j, j = 1, \ldots, J$ (where J is the total number of grid cells defining the latent process) covering the study region. These are then convolved with a kernel matrix whose elements, k_{ij}, represent the relative contribution of the latent variable in grid cell j to the Poisson mean in area i:

$$\lambda_i = \sum_j \gamma_j * k_{ij}.$$

One interpretation of Poisson-gamma moving average model is to view the gamma random variables as representing the location and magnitude of unmeasured risk factors, and the area-specific Poisson means λ_i as representing the cumulative effect of these risk factors in each area, weighted by their distance from the area according to the kernel 'weights' k_{ij}.

5.3.4 Poisson –Gamma with spatial conditional autoregressive

Conditional autoregressive (CAR) modelling has found considerable application for the analysis of spatial data. CAR variables used in conjunction with the likelihood to implement the Gibbs sampling updating implies a pairwise difference joint specification and hence improper joint posterior distribution. The main idea for a conditional autoregressive model is that the probability estimated at any given location, say i, are conditional on the level of neighbouring value. The standard or "proper'' CAR models for expectation of a specific observation, y_i, is often of the form

$$\mathrm{E}(y_i \mid y_{-i}) = \lambda_i + \rho \sum_{j \neq i} w_{ij} (y_j - \lambda_j)$$

where λ_i is the expected mean value at location i and ρ is a spatial autocorrelation parameter. The spatial correlation parameter, ρ determines the size and nature of the spatial neighbourhood effect. The notation y_{-i} means observations from all other locations except i

5.4 MCMC methods

To generate random samples from $f(\theta \mid y)$, we use a Markov chain satisfying the following:

$f(\theta_{t+1} \mid \theta_t)$ should be easy to generate from, the equilibrium distribution of the selected Markov Chain which must be the posterior distribution, $f(\theta \mid y)$ of interest.

In the MCMC approaches one uses the previous sample value to randomly generate the next sample value, from the sample generating a Markov Chain. The process proceeds as follows:

(1) Select an initial θ_0

(2) Monitor convergence using convergence diagnostics. If no convergence occurs, generate more observations

(3) Cut off the first *B* observations called the burn in period, and

(4) Plot the posterior distribution

(5) Obtain summaries of the posterior distribution (mean, median, standard deviation, quantiles, correlations, 95% confidence intervals)

In this approach, we specified a particular quantile of the distribution of interest, typically 2.5% and 97.5%, to give a 95% confidence interval. The maps below consist of smoothed relative risk and standard mortality rate.

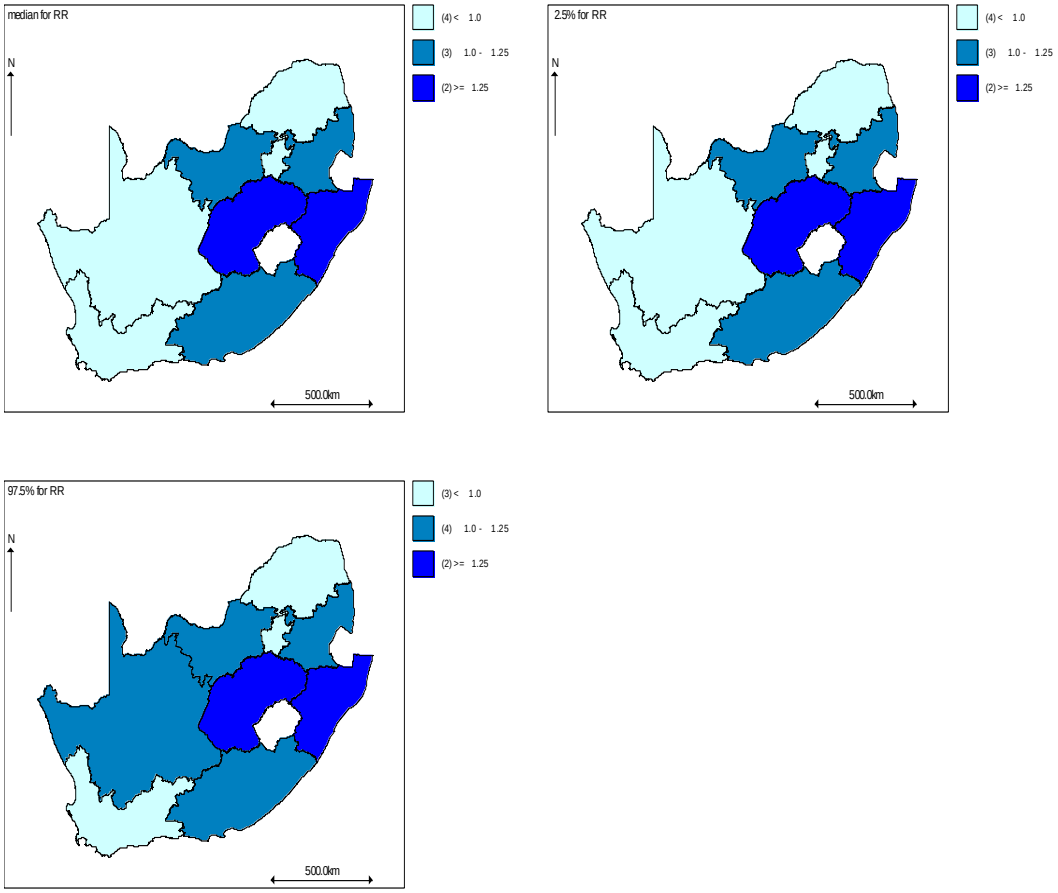


Figure 5: RR of TB Death mapped in 9 Province of South Africa: (top left) RR; (top right) 2.5% lower limit for the RR; (bottom left) 97.5% upper limit for the RR.

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 Bayesian analysis. This is an algorithm that generates a sequence of random variables from a joint distribution of two or more random variables. These sequences are required to approximate the joint distribution or to compute the summary statistics such as the expectation of the distribution.

5.4.1 Sampling the Hyper-parameter α

As mentioned already, from Gibbs sampler approach, it can be shown that the full conditional density for the hyper parameter α is given by

$$f(\alpha|...) \propto \frac{\alpha^{Ja+\varepsilon-1}}{\Gamma(\alpha)^J} \mathrm{H}^{\alpha-1} \exp(-\alpha S)$$

where

$$\mathrm{H} = \prod_{j=1}^{J} \mathrm{X}_j ,$$

$$S = \prod_{j=1}^{J} X_j + \varepsilon .$$

using a Gibbs sampler from the gamma distribution, where X_j are the latent variables. For the hyper-parameter α one assumes a gamma random variable distributed as

$$\alpha \sim Gamma(\varepsilon, \varepsilon)$$

for small $\varepsilon < 0$ in an attempt not to be informative.

5.5 Application and Interpretation of Model result for TB Death Data

Consider the following data from annual report on deaths from various causes gathered by Statistics South Africa in 2007.

Table 5: TB and HIV Deaths Data

Id	Provinces	Census 2007	TB observed	TB expected	HIV observed	HIV expected
1	KwaZulu-Natal	10259230	20832	13748.26	4528	2900.17
2	Free State	2773059	5373	3716.14	998	783.91
3	Eastern Cape	6527747	10146	8747.75	1622	1845.32
4	Mpumalanga	3643435	5850	4882.52	958	1029.96
5	Limpopo	5238286	4175	7019.76	278	1480.81
6	Northern Cape	1058060	1361	1417.89	359	299.10
7	North West	3271948	4674	4384.70	762	924.94
8	Western Cape	5278585	3813	7073.76	1563	1492.19
9	Gauteng	10451713	8773	14006.21	2643	2954.58

The models were fitted using WINBUGS. We fitted the four models defined above for TB death using 100 000 iteration with a burn in period of 10 000 iterations and thinning of 10. On testing convergence one should always look at the time series history, the plot of the random variable being generated versus the number of iterations. In addition to showing poor mixing, such a history can also suggest a minimum burn-in period for some starting value. The model scripts and the history plots are in appendix C.

Table 5.1: Relative risk parameter estimates of TB death from four models

Provinces	Poisson-Gamma	Poisson-GLMM	Spatial CAR	Convolution Model
	Median(95%Credible interval)	Median(95%Credible interval)	Median(95%Credible interval)	Median(95%Credible interval)
1	1.52(1.49,1.54)	1.52(1.50,1.54)	0.48(0.45,0.48)	14.52(-5.63,27.51)
2	1.45(1.41,1.48)	1.45(1.41,1.48)	0.42(0.39,0.45)	-1.95(-11.34,3.98)
3	1.16(1.14,1.18)	1.16(1.14,1.18)	0.20(0.18,0.22)	-1.22(-3.62,0.35)
4	1.20(1.17,1.23)	1.20(1.17,1.23)	0.23(0.21,0.26)	-1.39(-5.91,-0.02)
5	0.60(0.58,0.61)	0.59(0.58,0.61)	-0.47(-0.49,-0.44)	-0.48(-7.3,1.83)
6	0.96(0.91,1.01)	0.96(0.91,1.01)	0.01(-0.04,0.06)	-0.24(-6.92,4.9)
7	1.07(1.04,1.10)	1.07(1.03,1.09)	0.12(0.09,0.14)	-0.38(-2.34,1.92)
8	0.54(0.52,0.56)	0.54(0.52,0.56)	-0.57(-0.59,-0.54)	-0.38(-3.85,4.51)
9	0.63(0.61,0.64)	0.63(0.61,0.64)	-0.42(-0.44,-0.39)	-3.15(-4.79,-0.52)
beta	6.87(2.13,16.2)	-0.05(-0.33,0.22)	-0.052(-0.06,-0.04)	-5.27-8.96,-0.68)
sigma		0.38(0.25,0.68)	0.66(0.44,1.17)	10.14 (2.29,26.69)
Sigma v				0.05(0.01,1.003)

Provinces keys: 1-KwaZuLu-Natal, 2-Free State, 3-Eastern Cape, 4-Mpumalanga, 5-Limpopo, 6-Northern Cape, 7-North West, 8-Western Cape, 9-Gauteng

Table 5.2: **DIC Values**

	TB Observed			
Model	**Dbar**	**Dhat**	**Pd**	**DIC**
Poisson-Gamma	103.283	94.311	8.972	112.255
Poisson-GLMM	103.302	94.313	8.989	112.291
Spatial CAR	107.128	94.316	12.812	119.94
Convolution Model	118.879	94.303	24.576	143.455

5.6 Discussion of TB Model Results

We first explored variations in TB mortality by administrative province as per 2007 census as shown in table 5. The map of standard deviations of TB death rates showed spatial variations in South Africa with KwaZulu-Natal, Free State, Mpumalanga and Eastern Cape provinces showing the highest variations as shown in Table 5.1, and it shows that TB mortality is higher in KwaZulu-Natal followed by Free State and it shows that these mortality levels are significant with p-value =0.005 and the standard deviation was 0.0104. In this chapter, a Bayesian approach was applied in the data using WinBugs software where we model the data using the Poisson-Gamma, Poisson GLMMs, Spatial Car and Spatial convolution priors. As stated above θ follows a gamma distribution and the prior values were $\alpha = 0.001$ and $\beta = 0.001$. As defined above in section 5.4.1 note that the Gibbs sampler usually produces chains with smaller autocorrelations than other samplers reason why from the Table 5.1 sigma was also used to get results, it is expected that the distribution of the weight to spike as the sampler approaches stationary. For all the models, 100 000 iterations were carried out, discarding the first 10 000 samples and storing every tenth sample. These were then summarized to get the relevant estimates and it was noted that coefficient parameters had converged based on the history plot. The corresponding 2.5 and 97.5 percentiles were mapped as estimates of an approximate 95% credible interval for the posterior mean coverage. Therefore it is worth comparing the four results from these different approaches. From these results one can see that the estimates are quite comparable, see Tables

5.1 for more details. Thus, inference drawn from the four modeling approaches provides some degree of confidence in the results. When looking at standard deviations, means, medians, DIC and confidence interval for the TB model it is clear that they are comparable and significant, but the best model is the Poisson-Gamma based on the results given by Table 5.2 where the DIC is 112.25 followed by Poisson-GLMMs with a DIC value of 112.3.

5.7 Application and Interpretation of Model result for HIV Death Data

In the case of the HIV model the same procedure as that of TB was followed. The data set was presented in Table 5. In the HIV model, the map of standard deviation showed that spatial variations in HIV deaths rates is highest in KwaZulu-Natal followed by Free State, then Northern Cape and Western Cape. In Table 5.3 below, the DIC shows that Poisson-Gamma model was the best model with a DIC value of 96.9. The model was initialized and 100 000 iterations were simulated with a burn in of 10,000 and thinning of 10. These were then summarized in Table 5.4 to get the relevant estimates of mortality risk and it was noted that coefficient parameters had converged based on the history plot but sigma was not considered as it doesn't monitor the convergence and we then drop the directions of all edges to obtain the conditional independence graph. The corresponding 2.5 and 97.5 percentiles were mapped as estimates of an approximate 95% credible interval for the posterior mean coverage. It was noted that coefficient parameters converged and the history plot showed convergence (see appendix D for more detail).

When we plotted the two chains history plots together we observed and hence concluded that there is a positive correlation of HIV and TB risk in the same province. The Scripts of the selected model and history of complete trace plots for HIV fixed effects parameters can be found in appendix D.

Table 5.3: **DIC VALUE FOR HIV DEATH**

Model	Dbar	Dhat	pD	DIC
Poisson-Gamma	88.5	79.4	8.8	96.9
Poisson-GLMM	88.6	79.6	8.9	97.5
Spatial CAR	92.7	79.6	13.1	105.8
Convolution Model	107.1	79.5	27.5	134.7

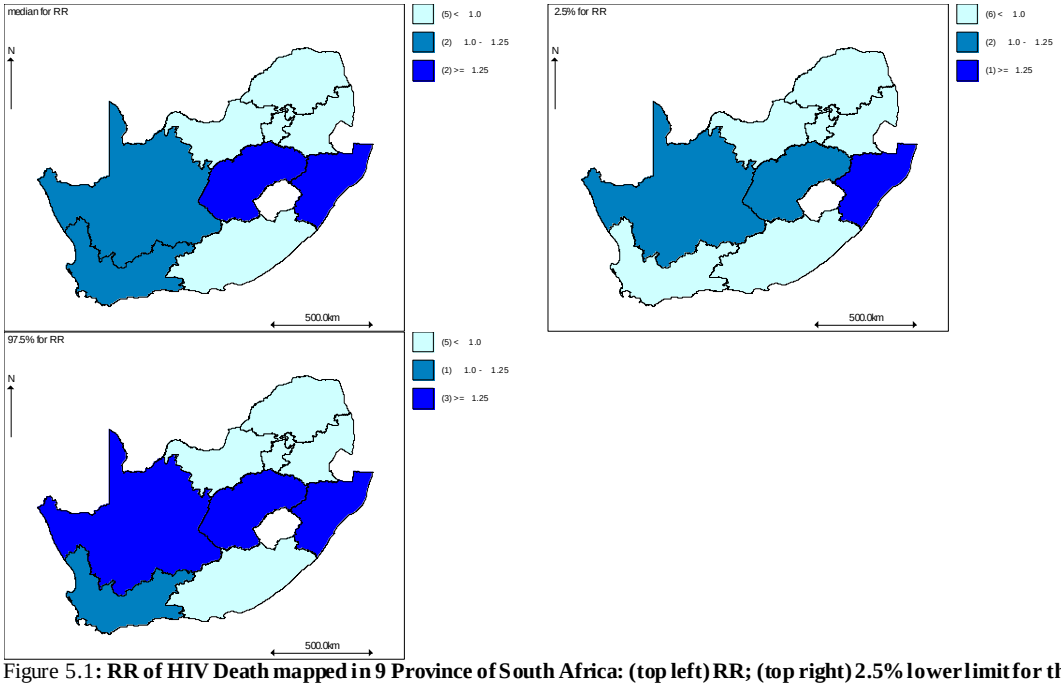


Figure 5.1**: RR of HIV Death mapped in 9 Province of South Africa: (top left) RR; (top right) 2.5% lower limit for the RR; (bottom left) 97.5% upper limit for the RR.**

Table 5.4: **Parameter estimate of HIV Death**

Provinces	Poison-Gamma	Poisson-GLMM	Spatial CAR	Convolution Model
	Median(95%Credible interval)	**Median(95%Credible interval)**	**Median(95%Credible interval)**	**Median(95%Credible interval)**
1	1.56(1.52,1.61)	1.6(1.52,1.61)	1.6(1.50,1.64)	1.56(1.44,1.68)
2	1.27(1.2,1.4)	1.3(1.19,1.35)	1.3(1.18,1.36)	1.27(1.15,1.39)
3	0.88(0.84,0.92)	0.9(0.84,0.92)	0.87(0.83,0.93)	0.87(0.81,0.96)
4	0.93(0.87,0.98)	0.93(0.87,0.99)	0.92(0.86,0.99)	0.93(0.84,1.02)
5	0.19(0.16,0.21)	0.2(0.17,0.21)	0.19(0.17,0.21)	0.18(0.16,0.21)
6	1.2(1.1,1.3)	1.19(1.07,1.32)	1.19(1.1,1.3)	1.19(1.05,1.35)
7	0.8(0.76,0.88)	0.82(0.77,0.88)	0.82(0.76,0.89)	0.82(0.74,0.91)
8	1.05(0.9,1.1)	1.04(0.99,1.1)	1.05(0.98,1.11)	1.046(0.96,1.14)
9	0.9(0.86,0.93)	0.89(0.86,0.93)	0.894(0.8,0.9)	0.89(0.82,0.97)
Beta	3.9(1.18,9.4)	-0.14(-0.58,0.29)	-0.14(-0.17,-0.11)	-015(-0.65,0.26)
sigma		1.013(0.67,1.79)	1.013(0.67,1.79)	0.05(0.01,1.005)
Sigma v				0.51(0.33,096)

Provinces keys: 1-KwaZuLu-Natal, 2-Free State, 3-Eastern Cape, 4-Mpumalanga, 5-Limpopo, 6-Northern Cape, 7-North West, 8-Western Cape, 9-Gauteng.

CHAPTER SIX

DISCUSSION AND CONCLUSION

The objective of this study is to understand factors that can be used to explain mortality due to TB and co-mortality with HIV in South Africa. The analysis was based on mortality data from STATS SA collected for the year 2007. The study is particularly concerned with statistical methods that can best be used to model these associations, and to identify factors affecting TB and HIV mortality in South Africa during the year 2007. The study aims to explore the association of risk factors associated with TB and HIV mortality which can include demographic, environment, biological and social factors. But because of the problem of a high rate of missing values the study focused mainly on demographic type of risk factors. The study was also extended to attempt and explain the spatial distribution of risk of mortality due to these two conditions.

6.1 Tuberculosis

Tuberculosis (TB) is the main cause of death in the world among all infectious diseases (Herchline and Amorosa, 2010). TB is one of the leading causes of death in HIV individuals in South Africa. HIV can dramatically fuel the rate of TB mortality, because HIV compromises the immune system.

The exploratory analysis carried out in this study indicates that TB incidence is higher among males than females. This is because males tend to work in more TB prone environments than females. One possible such working environment is that males work in mines more than females where shafts in mines are poorly ventilated and therefore facilitating very easy spread of TB bacteria. Environments where overcrowding is a common feature are ideal conditions for the spread of TB and air-borne diseases. Migrant mine workers carry the bacteria back home during

holidays and spread it to their surrounding areas. Previous studies on TB prevalence indicate that TB prevalence seems to be higher among younger individuals. This can be attributed to the fact that younger individuals are increasingly becoming more vulnerable due to infection with HIV. Given TB is one of the opportunistic infections among HIV infected- individuals may explain this correlation. It also suggests that people with low level of education are more TB infected because they are unemployed which lead them to live in high crowded places and high levels of poverty. Those who live in informal settlements and those who work in crowded environments such as factories where there is a lot of pollution tend to die of TB than other living and working condition.

6.2 Human Immune Virus (HIV)

Globally the estimated number of people living with HIV in 2007 was 33.2 million and 22.5 million persons were living with HIV in Sub-Saharan Africa (UNAIDS, 2007). HIV represents one of the most serious challenges to health and society in general. In South Africa, 16 % of the population is infected with HIV, and 1000 people die from AIDS-related causes each day, and two-thirds of those with HIV also suffer from TB, because of their weakened immune systems (AMREF, 2008).

The exploratory analysis carried out in this study indicates that HIV is more prevalent among females than males. This is largely because females are exposed to sexual abuse, rape and commercial sex activities for survival which expose them to HIV mortality. A possible biological reason is that females have a larger cervical area which makes it easier for HIV to establish itself in females than in males. The high prevalence in young individuals could be due to the fact that they are more sexually active and inexperienced which puts them at higher risk of HIV mortality. Individuals with lower education levels tend to be less informed about the risks of HIV; thus low

levels of education, poverty, overcrowding and unemployment are much associated with the less knowledge about HIV/AIDS.

6.3 Conclusion

This work has investigated factors associated with TB mortality and HIV deaths in South Africa. A number of models were used to aid key the estimation of effects of important factors associated with the risk of death due to the two infections. Generalized linear models, survey logistic regression models and Bayesian spatial disease mapping were used to identity these factors. The survey logistic regression model that accounts for more variability in the data helped to produce more reliable standard errors of parameter estimates. The results from it given in Table 4.1 lead to the same conclusions as the ones given by simple logistic regression model in the same table for TB deaths data but with better standard errors. Similar results for HIV mortality data are shown in Table 4.2. To ensure that the estimates of effects were adjusted for other factors in the model multiple logistic regression and survey logistic regression models.

The analyses identified factors affecting deaths with TB and HIV in South Africa during the year 2007, and the identified factors may be used to guide policy and decision making to speed up the provision of a better life for all. A statistical model was fitted and parameters to assess significance of a number of factors estimated. The analysis seems to indicate that TB deaths are highly associated with demographic factors. These factors are such as age and sex.

The researcher analyzed a complex survey data of TB mortality and HIV deaths in South Africa as presented by STATSSA. The most intriguing fact is the variations of results depending on different provinces. For example, KwaZulu-Natal is leading in the number of TB and HIV deaths as compared to other provinces. We have also noticed that TB is probably a major contributor of deaths among individuals infected with HIV.

In chapter 5 under the Bayesian modeling and mapping using WinBUGS section we gave some background theory, and then proceed without going into deep details on how the models are formulated and estimated. The Tables 5, 5.1, 5.2, 5.3, and 5.4 showed a positive correlation of TB and HIV at enumerator area level, therefore the spatial modeling helped a lot in mapping areas that are prone to TB and to HIV. We were able to evaluate the proportion of deaths associated with TB in South Africa in the year 2007, and reviewed regression modeling for relating a binary outcome to a number of predictor variables. Finally, this study was able to quantify factors related to TB and co-mortality with HIV and such results will help to guide decisions on how to mitigate the problem.

The major limitation of the study is the data which is very large and could not allow analysis at the level of individual members; therefore policy makers and further researchers should focus more on individual level, for example TB and HIV both as individual causes of death and co-mortality. In addition to that, analysis at the individual level might give more insight into the disease than analysis at the general level. One major limitation was the high rate of missing values in most categories of risk factors which made it difficult to use such information in the analysis. There are avenues for further work on the subject. In this study our focus was on TB and HIV in South African as well as those who died with both diseases. The results of the study can be used in a number of ways as regards public health policies. In terms of provision of public health services priority should be directed to provinces with high TB and HIV burden such as KwaZulu-Natal. It might be necessary to conduct a detailed province and district within province assessments of TB and HIV control programs, setting targets for each district and tracking the progress. This may help to understand the root causes of high HIV and TB infection rates hence

put in place control measures in order to reduce mortality due these two highly synergistic diseases.

It may also be necessary to conduct research and surveillance about- other TB relevant variables such as TB and HIV age categories and gender, number of death by sex and district municipality. Also conduct further research to explore the associations between TB variables such as environmental, demographic, and other socioeconomics variables not used in this study. The GLM and Survey logistic regression models results showed a positive correlation of TB and HIV at an enumerator area level, therefore spatial modeling of these data helped a lot in improving results and in mapping areas that are prone to TB and HIV.

BIBLIOGRAPHY

Africa Medical and Research Foundation (AMREF). (2008, August 24). *TB and HIV Control in South*. Retrieved February 18, 2011, from amref: .http://www.amref.org/whatwedo/tb and hiv control in South Africa.

Agresti, A. (2002). *Introduction to categorical data analysis.* John Wiley.

Agresti, A. (1990). *Categorical data analysis.* John Wiley and Sons.

Bekker, L. G., & Wood, R. (2010). The changing natural history of tuberculosis and HIV coinfection in an urban area of hyperendemicity. *Clin Infect Dis.50 Suppl 3* , S208-S214.

Besag J, Mollie A. (1989). Bayesian mapping of mortality rates. *Bulletin of the International Statistical Institute, 47th Session*, 53:127-28.

Cohen, J. (2002). *Science*. Retrieved MAY 20, 2011, from Therapies:Confronting the limits of success: http://aidscience.org/science/2320.html

Collet, D. (2003). *Modeling Binary data.2nd edition.* New York.: Chapman& Hall/CRC.

Collins, T. F. (1981). Applied epidemiology and logic in tuberculosis control. *South Africa Medical journal* , **61**, 566-9.

Corbett, E. L., Watt, C. J., Walker, N., Maher, D., Williams, B. G., Raviglione, M. C., et al. (2003). The growing burden of tuberculosis: Global trends and interactions with the HIV epidemic. *Archives of Internal Medicine **163*** , 1009-1021.

Decock, & Chaisson, R. E. (1999). *International Journal of Tuberculosis and Lung Disease* **, 3**, 457.

Dobson, A. J. (2002). *An Introduction to generalized Linear Models second edition.* New York: Chapman and Hall/CRC.

Dobson, A. J. (1990). *An Introduction to generalized Linear Models.* T.J. Press, cornwall.

Dye, C. (2006). Global epidemiology of tuberculosis. *Lancet* , **367** (3): 938-940.

Getahun, H., Harrington, M., O'Brien, R., & 16, P. N. (2007, Junuary). Diagnosis of smear-negative pulmonary tuberculosis in people with HIV infection or AIDS in resource-constrained settings. *informing urgent policy changes* , 2042-2049.

Herchline, T., & Amorosa, J. K. (2010, October 4). *Tuberculosis.* Retrieved August 2, 2010, from emedicine: http://emedicine.medscape.com/article/230802-print

HOSMER, D., & LEMESHOW, S. (1989). *Applied Logistic Regression.* John Wiley and Sons.

Khaled, K. M. (2008). *Tuberculosis (TB) Progress Toward Millennium Development Goals (MDGS) and DOTS in Who Eastern Mediterranean Region (EMR), MPH Thesis.* Atlanta: Georgia State University.

Lawson, A., Browne, W., & Rodeiro, C. V. (2003). *Disease Mapping with WinBUGS and MLwiN*. Chichester, West Sussex: John Wiley & Sons, Ltd.

Lawson, A. B., & William, F. (2001). *An introductory guide to disease mapping.* Chichester John Wiley & Sons, Ltd.

Lawn, S. (2010). *The challenge of the HIV –associated tuberculosis epidemic in sub-Saharan Africa: will antiretroviral therapy help?* Retrieved August 6, 2010, from http://www.sacemaquarterly.com

Lehtonen, R., & Pahkinen, E. J. (1995). *Practical Methods for Design and Analysis of Complex Surveys.* Chichester: John Wiley& Sons.

Leyland AH. Spatial analysis. In Leyland AH, Goldstein H eds. *Multilevel modelling of health statistics.* Chichester: John Wiley & Sons, Ltd, 2001

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models(second edition).* Chapman and Hall.

McCullagh, P., & Nelder, J. (1983). *Generalized linear models.* Chapman and Hall.

Myers, Montgomery, & vining. (2002). *Generalized linear models (with Applications in Engineering and the Sciences).*

Mzolo, T. (2009). *Estimating Risk Determinants of HIV and TB in South Africa.* Pietermaritzburg: University of KwaZulu-Natal.

Nelder, J., & Wedderburn, R. M. (1972). Quasi-likelihood functions,generalized linear models and the Gauss-Newton method. *Biometrika* ***61*** , 439-447.

Pfeffermann, D. (1993). The role of Sampling Weights When Modeling Survey Data. *International Statistical Review* , **64 (2)**, 317-37.

Raviglione, M. C., & Pio, A. (2002). Evolution of WHO policies for tuberculosis control,1948-2001. *Science direct* , 775-780.

Raviglione, M. C., Harries, A. D., Msiska, R., Wilkinson, D., & Nunn, P. (1997). Tuberculosis and HIV:Current status in Africa. *AIDS Supplement 11* , S115-S123.

S. Geman & D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

Singer, C. (1997). *TB in South Africa: The People's Plague.* Pretoria: National Department of Health.

Skinner, C. J., Holt, D., & Smith, T. M. (1989). *Analysis of Complex Survey*. Chichester: John Wiley& Sons.

South Africa. Department of Health (2007). *National HIV and syphilis prevalence survey in South Africa 2006*. Pretoria: Department of Health.

South Africa. Department of Health (1998). *TB in South Africa: The People's Plague*. Pretoria: Department of Health.

Steven, G.Heeringa, Brady T. West & Patricia , A. Berglund (2010). *Applied Survey Data Analysis (second edition).* Chapman and Hall.

Tan, D., Upshur, R. E., & Ford, N. (Apr 1 2003). *BMC International Health and Human Rights* ***3****, 2*.

UNAIDS. (2003). *Report on the global HIV/AIDS epidemic Tech. Report No. UNAIDS/02.26E* . UNAIDS.

UNAIDS/WHO. (2007). *2007 AIDS epidemic update*. Retrieved November 27, 2007, from Joint United Nations Programme on HIV/AIDS, 2007: http://data.unaids.org/pub/EpiSlides/2007/2007_EpiUpdate_en.pdf

UNAIDS. (2008). *Sub-Saharan Africa AIDS epidemic update*: *Regional Summary*. Switzerland: World Health Organization.

UNAIDS. (2008). *Report on the global HIV/AIDS epidemic.*Geneva.

Uriz, J., Reparaz, J., Castiello, J., & Sola, J. (2007). Tuberculosis in patients with HIV infection. *An sist sanit Navar* , 30 Suppl 2:131-142.

Vaidyanathan, P. S., & Singh, S. (2003). TB-HIV co-infection in India. *NTI Bulletin* , 39, 3&4, 11-18.

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch, C. E. (2005). *Regression Methods in Biostatistics: Linear, Logistic, Survival and Repeated Measures Models.* New York: Springer.

Williams, B. G., & C Dye, (2003). *Antiretroviral Drugs for Tuberculosis Control in the Era of HIV/AIDS.* Geneva: Sciencexpress.

World Health Organization. (2003). *"Global Tuberculosis Control: Surveillance, planning financing (World Health Report) " Tech. Report No. WHO/CDS/TB/2001.287.* Geneva: World Health Organization.

World Health Organization. (2000). *Anti-tuberculosis Drug Resistance in the World, World Health Organization Report no. 2, WHO/CDS/TB/ 2000.278.* Geneva: WHO.

World Health Organization. (2005). *WHO Report 2005*: *Global Tuberculosis Control.* Geneva: World Health Organization.

World Health Organization (2008). *Tuberculosis detection rate under DOTS*. http://www.who.int/whosis/indicators/compendium/2008/4tdr/en/index.html

World Health Organization (2009). Global tuberculosis control: a short update to the 2009 report. Available on HYPERLINK "http://www.who.int/tb/publications/global_report/2009/update/en/index.html"

Zuma, K., Lurie, M. N., Williams, B. G., Mkaya-Mwamburi, D., Garnett, G. P., & 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.

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

APPENDIXES

Appendix A

Procedures for the Generalized Linear Models

A.1 STATA Procedures

The STATA system was used to fit the logistic regression model discussed in Chapter 3 and fitted in Chapter 2 and 4. LOGISTIC REGRESSION was used to fit the model. The stepwise procedure implemented in LOGISTIC REGRESSION was applied.

A.1.1 Model Selection Using STATA Code

The following stepwise selection procedure was used by including the following statements:

```
Set memory 400m
Set more off
log using "D:\dataset.log", replace
describe
summarize
gen tb=0
replace tb =1 if CauseA=="A16" | CauseA=="A17" | CauseA=="A18" | CauseA=="A19"
label var tb "TB Cause-related"
label values tb tbv
label define tbv 0 "No TB" 1 "TB"
gen hiv=0
replace hiv=1 if CauseA=="B20" | CauseA=="B21" | CauseA=="B22" | CauseA=="B23" |
CauseA=="B24"
label var hiv "HIV cause-related"
label values hiv hv
```

```
label define hv 0 "HIV negative" 1 "HIV cause-related "
//HIVB
gen hiv2=0
replace hiv2=1 if CauseB=="B20" | CauseB=="B21" | CauseB=="B22" | CauseB=="B23" |
CauseB=="B24"
label var hiv2 "HIV cause-related"
label values hiv2 hv2
label define hv2 0 "HIV negative" 1 "HIV cause-related "
//HIVC
gen hiv3=0
replace hiv3=1 if CauseC=="B20" | CauseC=="B21" | CauseC=="B22" | CauseC=="B23" |
CauseC=="B24"
label var hiv3 "HIV cause-related "
label values hiv3 hv3
label define hv3 0 "HIV negative" 1 "HIV cause-related "
tab tb, miss
tab hiv, miss
tab hiv2, miss
tab hiv3, miss
//HIVD
gen hiv4=0
replace hiv4=1 if CauseD=="B20" | CauseD=="B21" | CauseD=="B22" | CauseD=="B23" |
CauseD=="B24"
label var hiv4 "HIV cause-related "
label values hiv4 hv4
label define hv4 0 "HIV negative" 1 "HIV cause-related "
//HIV5
gen hiv5=0
replace hiv5=1 if OtherCause=="B20" | OtherCause=="B21" | OtherCause=="B22" |
OtherCause=="B23" | OtherCause=="B24"
label var hiv5 "HIV cause-related "
```

```
label values hiv5 hv5
label define hv5 0 "HIV negative" 1 "HIV cause-related "
//Combination of cases of HIV
gen hivc=0
replace hivc=1 if (hiv==1 | hiv2==1 | hiv3==1 |hiv4==1 | hiv5==1)
label var hivc "HIV cause-related "
label values hivc hvc
label define hvc 0 "HIV negative" 1 "HIV cause-related "
tab hivc
gen EduCodeg=1 if EduCode==0
replace EduCodeg=2 if EduCode>0 & EduCode<=9
replace EduCodeg=3 if EduCode>9 & EduCode<=12
replace EduCodeg=4 if EduCode==13
replace EduCodeg=5 if EduCode>=97 & EduCode<=99
label var EduCodeg "Education group"
label values EduCodeg edv
label define edv 1 "None" 2 "Primary Education" 3 "Secondary Education" 4 "University" 5
"Other"
tab EduCodeg
gen ageg=1 if Age<=15
replace ageg=2 if Age>15 & Age<=30
replace ageg=3 if Age>30 & Age<=45
replace ageg=4 if Age>45 & Age<=60
replace ageg=5 if Age>60 & Age<=75
replace ageg=6 if Age>75 & Age<=90
replace ageg=7 if Age>90 & Age<.
label var ageg "Age group"
label values ageg agev
label define agev 1 "0-15" 2 "16-30" 3 "31-45" 4 "46-60" ///
5 "61-75" 6 "76-90" 7 ">90"
//New variable for Sex
```

```
replace Sex=3 if Sex==8 | Sex==9
label variable Sex "Gender"
label values Sex sex
label define sex 1 "Male" 2 "Female" 3 "Other"
//New variable for Marital status
replace MStatus=8 if MStatus==9
label var MStatus "Marital Status"
label values MStatus mst
label define mst 1 "Single" 2 "Civil marriage" 3 "Living as married" 4 "Widowed" ///
5 "Religious law marriage" 6 "Divorced" 7 "Customary marriage" 8 "Other"
//New variable for Province of birth
replace Birth_Prov=10 if Birth_Prov==98
replace Birth_Prov=11 if Birth_Prov==97 | Birth_Prov==99
label var Birth_Prov "Province of birth"
label values Birth_Prov bpr
label define bpr 1 "Western Cape" 2 "Eastern Cape" 3 "Northern Cape" 4 "Free State" ///
5 "KwaZuLu-Natal" 6 "North West" 7 "Gauteng" 8 "Mpumalanga" 9 "Limpopo" 10 "Outside
South Africa" 11 "Other"
tab Birth_Prov
//New variable for Death province
replace Death_Prov=10 if Death_Prov==98
label var Death_Prov "Province of death"
label values Death_Prov dpr
label define dpr 1 "Western Cape" 2 "Eastern Cape" 3 "Northern Cape" 4 "Free State" ///
5 "KwaZuLu-Natal" 6 "North West" 7 "Gauteng" 8 "Mpumalanga" 9 "Limpopo" 10 "Outside
South Africa"
tab Death_Prov
//New variable for Place of death
replace DeathInst=6 if DeathInst==8 | DeathInst==9
label var DeathInst "Place of death"
label values DeathInst dinst
```

```
label define dinst 1 "Hospital(in-patient)" 2 "Emergency room(out-patient)" 3 "Death on arrival"
///
4 "Nursing home" 5 "Home" 6 "Other"
tab DeathInst
//New variable for Province of residence
replace Res_Prov=10 if Res_Prov==98
replace Res_Prov=11 if Res_Prov==97 | Res_Prov==99
label var Res_Prov "Province of residence"
label values Res_Prov rpr
label define rpr 1 "Western Cape" 2 "Eastern Cape" 3 "Northern Cape" 4 "Free State" ///
5 "KwaZuLu-Natal" 6 "North West" 7 "Gauteng" 8 "Mpumalanga" 9 "Limpopo" 10 "Outside
South Africa" 11 "Other"
tab Res_Prov
//New variable for Occupation of deceased
label var Occupation "Occupation of deceased"
label values Occupation occ
label define occ 0 "Armed forces" 1 "Legislators,Senior officer" 2 "Professionals" 3
"Technicians and Associated" ///
4 "Clerks" 5 "Searvice workers,shop" 6 "Skilled agricultural" 7 "Craft and related trader" ///
8 "Plant and machine operator" 9 "Elementary occupation"
tab Occupation
//New variable for Type of Industry/Business
replace Industry=10 if Industry==97 | Industry==98 | Industry==99
label var Industry "Type of Industry"
label values Industry ind
label define ind 0 "Private households,etc" 1 "Agriculture,hunting,etc" 2 "Mining and quarrying"
3 "Manufacturing" ///
4 "Electricity,gas and etc" 5 "Construction" 6 "Wholesale and retail" 7 "Transport,storage and
etc" ///
8 "Financial intermediate" 9 "Community, social and etc" 10 "Other"
tab Industry
```

```
//New variable for Smoking status
replace Smoker=3 if Smoker==4 | Smoker==8 | Smoker==9
label var Smoker "Smoking status"
label values Smoker sms
label define sms 1 "Yes" 2 "No" 3 "Other"
tab Smoker
//New variable for Pregnancy status
replace Pregnancy=3 if Pregnancy==8 | Pregnancy==9
label var Pregnancy "Pregnancy status"
label values Pregnancy preg
label define preg 1 "Yes" 2 "No" 3 "Other"
tab Pregnancy
//Demographic variables
//Age
tab ageg, miss
tab ageg tb,miss row chi2
//Sex
tab Sex, miss
tab Sex tb, miss row chi2 exact
//Marital status
tab MStatus, miss
tab MStatus tb, miss row chi2
//Province of birth
tab Birth_Prov
tab Birth_Prov tb, miss row chi2
//Death province
tab Death_Prov
tab Death_Prov tb, miss row chi2
//Place of death
tab DeathInst
tab DeathInst tb, miss row chi2
```

```
//Province of residence
tab Res_Prov
tab Res_Prov tb, miss row chi2
//Education level
tab EduCodeg
tab EduCodeg tb, miss row chi2
//Occupation
tab Occupation
tab Occupation tb, miss row chi2
//Type of industry
tab Industry
tab Industry tb, miss row chi2
//Smoking status
tab Smoker
tab Smoker tb, miss row chi2
//Pregnancy
Tab Pregnancy
tab Pregnancy Sex if Age<=15, cell
tab Pregnancy tb if Sex==2 & Age>15, miss row chi2
//Underlying causes of death
tab NaturalUnnatural tb, miss row chi2 exact
tab NaturalUnnatural
//HIV
tab hivc
tab tb hivc, row chi2 exact
```

A.1.2 Model Fitting Using STATA Statements

The following codes were used to fit Simple and Multiple Logistic Regression, by including the following statements:

```
//
//Logistic regression
```

```
//
//Univariable
//
//TB versus Age
xi: logistic tb i.ageg if ageg<.
//TB versus Sex
xi: logistic tb i.Sex if Sex<8
//TB versus Marital status
xi: logistic tb i.MStatus
//TB versus Province of birth
xi: logistic tb i.Birth_Prov
//TB versus Province of death
xi: logistic tb i.Death_Prov
//TB versus death Institution
xi: logistic tb i.DeathInst
//TB versus Province of residence
xi: logistic tb i.Res_Prov
//TB versus Education level
xi: logistic tb i.EduCodeg
//TB versus Occupation
xi: logistic tb i.Occupation
//TB versus type of industry
xi: logistic tb i.Industry
//TB versus smoking status
xi: logistic tb i.Smoker
//TB versus HIV
xi: logistic tb i.hivc
//
//Logistic regression
//
//Multivariable
```

//

xi: logistic tb Age i.Sex i.MStatus i.Birth_Prov i.Death_Prov i.DeathInst i.Res_Prov ///
i.EduCodeg i.Occupation i.Industry i.Smoker i.hivc if (Age>90 & Age<. & Sex<3)

Appendix B

STATA SURVEYLOGISTIC

This procedure was used to fit a survey logistic regression model discussed and fitted in Chapter 4. The same variables selected by LOGISTIC REGRESSION were used to fit the survey logistic regression model. On the other hand, for complex survey data this means data presented by strata and clusters. In our case we choose province of death as cluster so that we have ten clusters (provinces) and one stratum (South Africa).

```
//
//Clustering
//
svyset Death_Prov {to define variable which is cluster}

//
//Logistic regression
//
//Univariable
//
//TB versus Age
xi: svy: logistic tb i.ageg if ageg<.
//TB versus Sex
xi: svy:logistic tb i.Sex if Sex<8
//TB versus Marital status
xi:svy: logistic tb i.MStatus
//TB versus Province of birth
```

```
xi: svy: logistic tb i.Birth_Prov
//TB versus death Institution
xi: svy: logistic tb i.DeathInst
//TB versus Province of residence
xi: svy: logistic tb i.Res_Prov
//TB versus Education level
xi: svy: logistic tb i.EduCodeg
//TB versus Occupation
xi: svy: logistic tb i.Occupation
//TB versus type of industry
xi: svy: logistic tb i.Industry
//TB versus smoking status
xi: svy: logistic tb i.Smoker
//TB versus HIV
xi: svy: logistic tb i.hivc
//
//Logistic regression
//
//Multivariable
//
xi: svy:logistic tb Age i.Sex i.MStatus i.Birth_Prov i.DeathInst i.Res_Prov ///
i.EduCodeg i.Occupation i.Industry i.Smoker i.hivc if (Age>90 & Age<. & Sex<3)
//
```

APPENDIX C

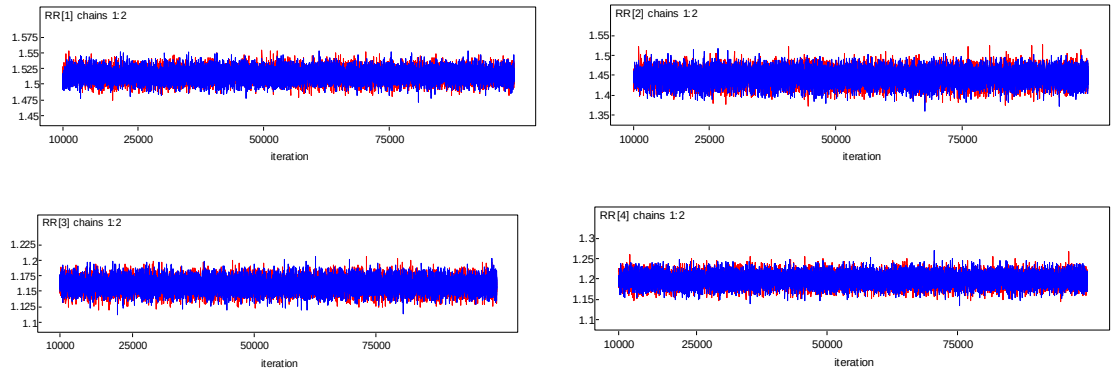
TB DEATH MODELING CODES AND HISTORY PLOTS

C.1 Poisson-gamma/Test Code

```
# The model
model
{
```

```
# The likelihood
        for (i in 1:N) {
                Tbobserved[i]   ~ dpois(mu[i])
                mu[i]<- Tbexpected[i]*theta[i]
                RR[i] <- theta[i]                                   # Area-specific relative risk (for
maps)
                SMR[i]<-Tbobserved[i]/Tbexpected[i]
                theta[i]~dgamma(alpha,beta)
        }
        # Other priors:
        alpha~dgamma(0.1,0.0001)
        beta~dgamma(0.1,0.0001)

}


Data
list(N=9)

Tbobserved[]  Tbexpected[]
20832 13748.26412
5373 3716.141225
10146 8747.751034
5850 4882.521073
4175 7019.760688
1361 1417.892798
4674 4384.696052
3813 7073.764867
8773 14006.20815
END

# Initialization
list(theta=c(1,1,1,1,1,1,1,1,1),alpha=1,beta=1)
list(theta=c(1,1,1,1,1,1,1,1,1),alpha=1.5,beta=1)
```

Figure 5.2: History: Complete trace plots for TB fixed effects parameter fitting Poisson-Gamma model



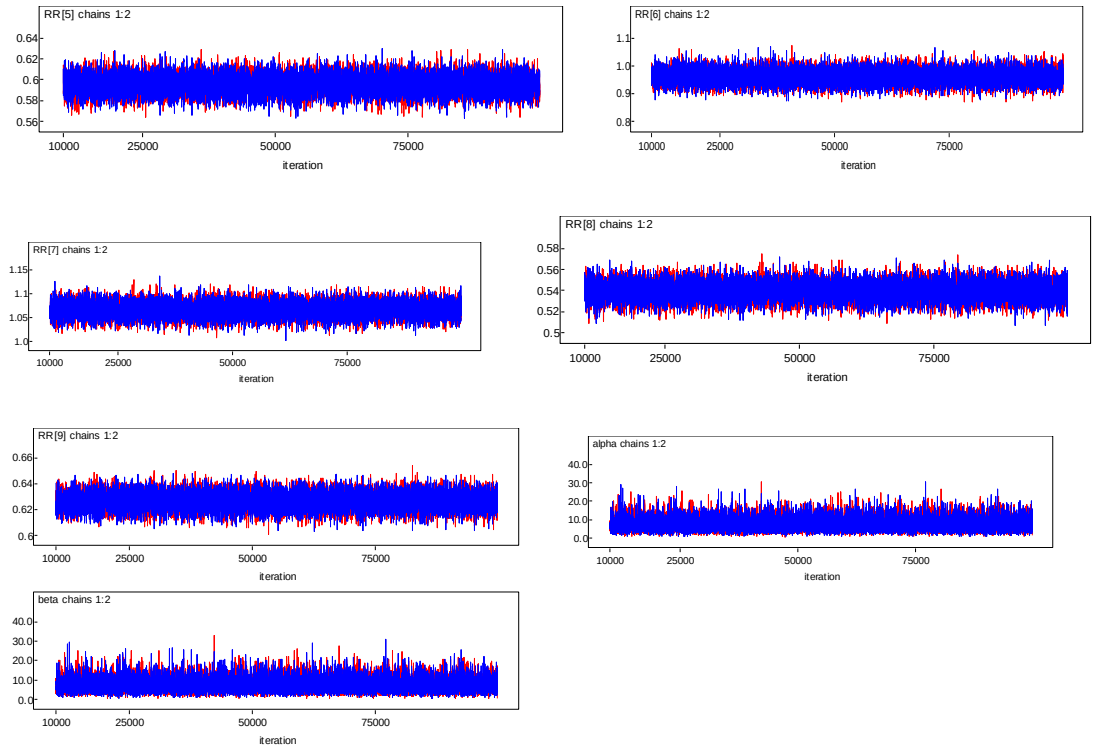


Figure 5.2: TB History plot by fitting Poisson-Gamma model

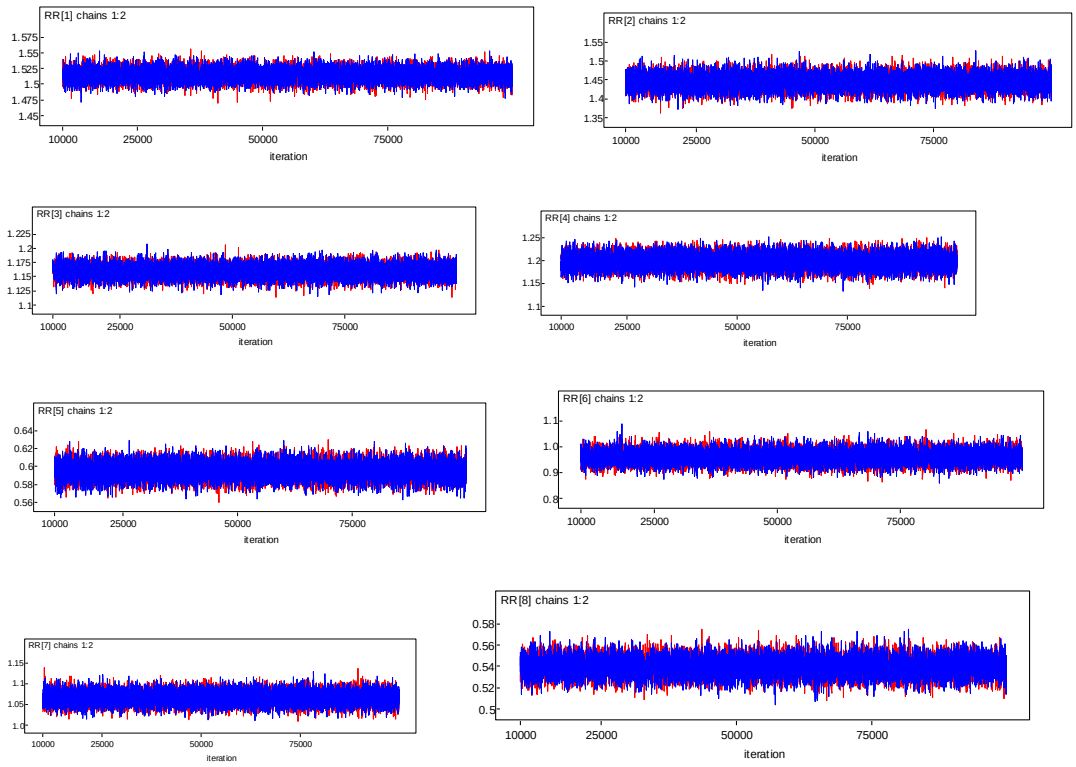
C.2 Poisson GLMMs/Test Code

```
# The model
model
{
# The likelihood
        for (i in 1:N) {
                Tbobserved[i]   ~ dpois(mu[i])
                terms[i]<-beta+b[i]
                log(mu[i]) <- log(Tbexpected[i])+terms[i]
                RR[i] <- exp(terms[i])                  # Area-specific relative risk (for maps)
                SMR[i]<-Tbobserved[i]/Tbexpected[i]

        }
# Non-spatial for random effects:

for (i in 1:9)
{
        b[i] ~dnorm(0.0,tau.b)
        }

        # Other priors:
        beta ~ dflat()
        tau.b  ~ dgamma(0.5, 0.0005)                    # prior on precision
        sigma <- sqrt(1 / tau.b)                        # standard deviation
```

}

# Initialization
list(beta=0,b=c(0,0,0,0,0,0,0,0,0),tau.b=1)
list(beta=0.5,b=c(0.5,0.5,0.5,0,2,0,0,0,0),tau.b=1)
Data
 # The data
list(N=9)
Tbobserved[] Tbexpected[]
20832 13748.26412
5373 3716.141225
10146 8747.751034
5850 4882.521073
4175 7019.760688
1361 1417.892798
4674 4384.696052
3813 7073.764867
8773 14006.20815
END

Figure 5.3: History: Complete trace plots for TB fixed effects parameter fitting Poisson-GLMM model



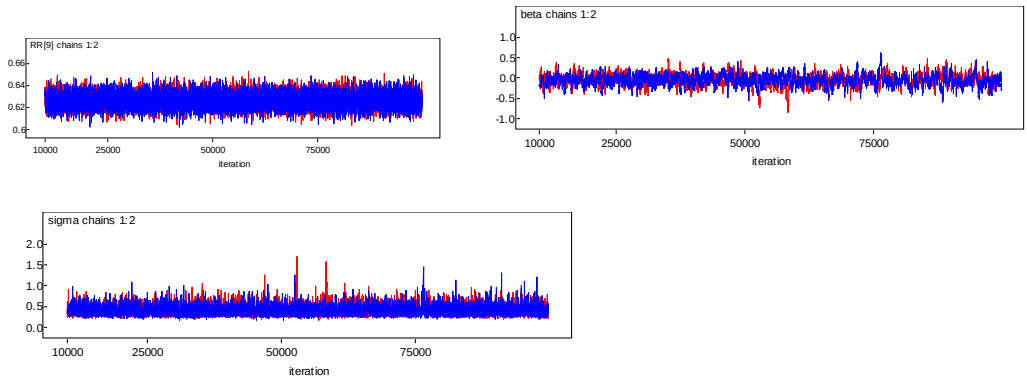


Figure 5.3: TB History plot by fitting Poisson-GLMMs

C.3 Spatial CAR Model /Test Code

```
# The model
model
{
# The likelihood
        for (i in 1 : N) {
                Tbobserved[i]   ~ dpois(mu[i])
                terms[i]<-beta+ b[i]
                log(mu[i]) <- log(Tbexpected[i])+terms[i]
                SMR[i]<-Tbobserved[i]/Tbexpected[i]
                RR[i] <- exp(terms[i])                          # Area-specific relative risk (for
maps)
        }

# CAR prior distribution for random effects:
        b[1:N] ~ car.normal(adj[], weights[], num[], tau)
        for(k in 1:sumNumNeigh) {
                weights[k] <- 1
        }

        # Other priors:
        beta~dnorm(0.0,0.001)
        tau ~ dgamma(0.5, 0.0005)                       # prior on precision
        sigma <- sqrt(1 / tau)                  # standard deviation
}
```

Data

```
# The data
list(N=9)

Tbobserved[]  Tbexpected[]
20832 13748.26412
5373 3716.141225
10146 8747.751034
5850 4882.521073
4175 7019.760688
1361 1417.892798
4674 4384.696052
3813 7073.764867
```

8773 14006.20815
END

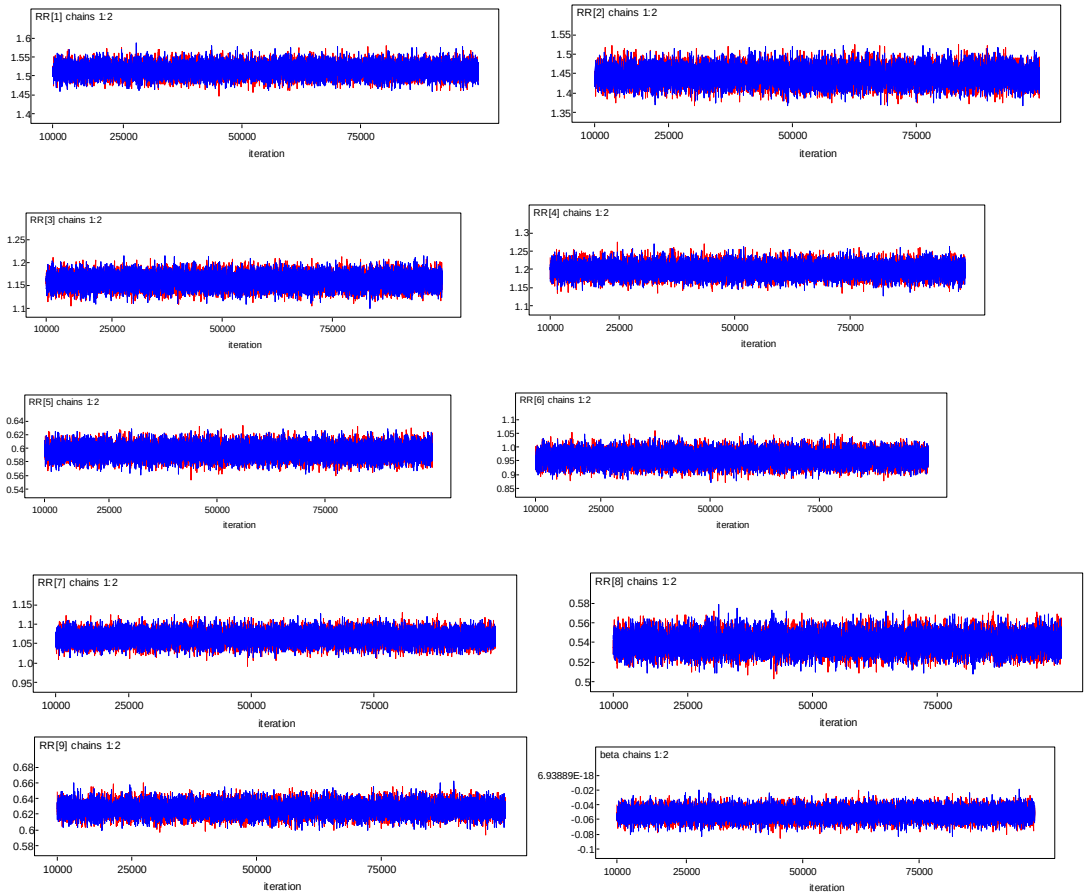
Adjecency information

list(num = c(3, 6, 4, 4, 3, 4, 4, 2, 4
),
adj = c(
4, 3, 2,
9, 7, 6, 4, 3, 1,
8, 6, 2, 1,
9, 5, 2, 1,
9, 7, 4,
8, 7, 3, 2,
9, 6, 5, 2,
6, 3,
7, 5, 4, 2
),
sumNumNeigh = 34)

Initialization

list(beta=0,b=c(0,0,0,0,0,0,0,0,0),tau=1)
list(beta=0.5,b=c(1,0,1,0,0,0,0,0,0),tau=1.4)

Figure 5.4: History: Complete trace plots for TB fixed effects parameter fitting Spatial CAR model



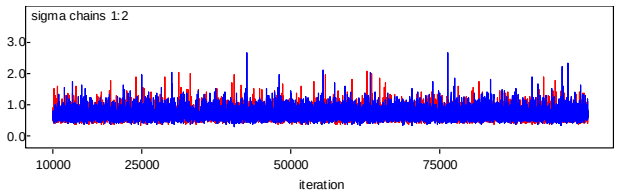


Figure 5.4: TB History plot by fitting Spatial CAR model

C.4 Spatial Model with convolution priors /Test Code

 # The model

```
model
{
# The likelihood
        for (i in 1 : N) {
                Tbobserved[i]   ~ dpois(mu[i])
                terms[i]<-beta+ b[i]+v[i]
                log(mu[i]) <- log(Tbexpected[i])+terms[i]
                SMR[i]<-Tbobserved[i]/Tbexpected[i]
                RR[i] <- exp(terms[i])                          # Area-specific relative risk (for
maps)
                v[i]~dnorm(0.0,1.0E-3)

                v[i]~dnorm(0.0,tau.v)
}


# CAR prior distribution for random effects:
        b[1:N] ~ car.normal(adj[], weights[], num[], tau)
        for(k in 1: sumNumNeigh) {
                weights[k] <- 1
        }

        # Other priors:
        beta ~ dnorm(0.0,1.0E-4)

        tau  ~ dgamma(0.5, 0.0005)                      # prior on precision
        tau.v  ~ dgamma(0.5, 0.0005)                    # prior on precision
        sigma <- sqrt(1 / tau)                  # standard deviation
        sigma.v<-sqrt(1/tau.v)
}
```

Data

 # The data

```
list(N=9)
Tbobserved[]   Tbexpected[]
20832  13748.26412
5373  3716.141225
10146  8747.751034
5850  4882.521073
4175  7019.760688
1361  1417.892798
4674  4384.696052
```

3813 7073.764867
8773 14006.20815
END

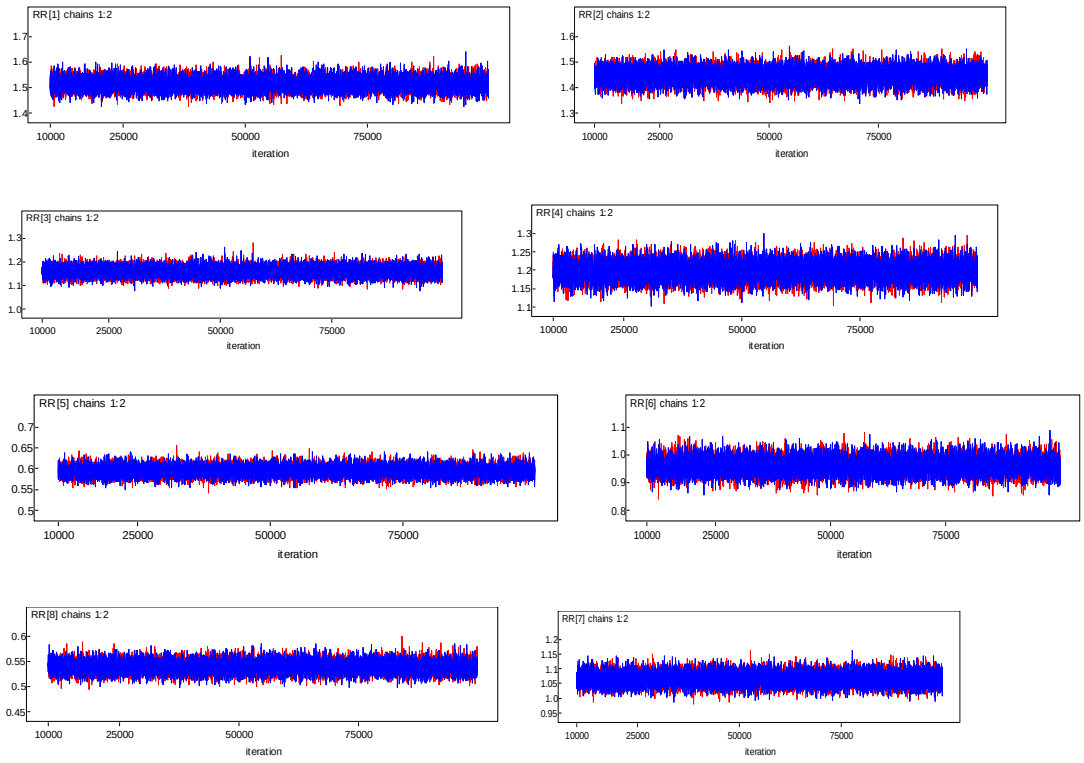
Adjecency information

```
list( num = c(3, 6, 4, 4, 3, 4, 4, 2, 4
),
adj = c(
4, 3, 2,
9, 7, 6, 4, 3, 1,
8, 6, 2, 1,
9, 5, 2, 1,
9, 7, 4,
8, 7, 3, 2,
9, 6, 5, 2,
6, 3,
7, 5, 4, 2
),
sumNumNeigh = 34)
```

Initialization

```
list(beta=0,b=c(0,0,0,0,0,0,0,0,0),v=c(0,0,0,0,0,0,0,0,0),tau=1,tau.v=1)
list(beta=1,b=c(1,0,1,0,0,1,0,0,0),v=c(1,0,1,0,1,0,0,0,0),tau=1,tau.v=1.5)
```

Figure 5.5: History: Complete trace plots for TB fixed effects parameter fitting Spatial Convolution model



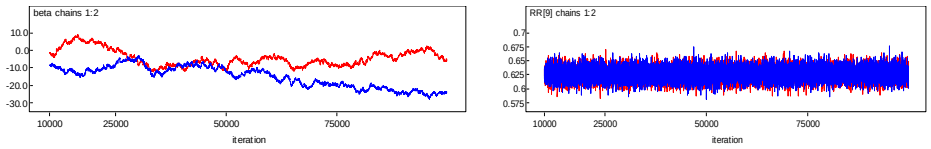


Figure 5.5: TB History plot by fitting spatial Convolution model.

APPENDIX D

HIV DEATH MODELING CODES AND HISTORY PLOTS

D.1 Poisson-gamma /Test CODE

```
# The model
model
{
# The likelihood
        for (i in 1:N) {
                HIVobserved[i]    ~ dpois(mu[i])
                mu[i]<- HIVexpected[i]*theta[i]
                RR[i] <- theta[i]                                   # Area-specific relative risk (for
maps)
                SMR[i]<-HIVobserved[i]/HIVexpected[i]
                theta[i]~dgamma(alpha,beta)
        }
        # Other priors:
        alpha~dgamma(0.1,0.0001)
        beta~dgamma(0.1,0.0001)

}

list(N=9)

HIVobserved[]  HIVexpected[]
4528 2900.172
998 783.9133
1622 1845.322
958 1029.959
278 1480.806
359 299.1019
762 924.9437
1563 1492.198
2643 2954.584

END
# Initialization
list(theta=c(1,1,1,1,1,1,1,1,1),alpha=1,beta=1)
list(theta=c(1,1,1,1,1,1,1,1,1),alpha=1.5,beta=1)
```

Figure 5.6: History: Complete trace plots for HIV fixed effects parameter fitting

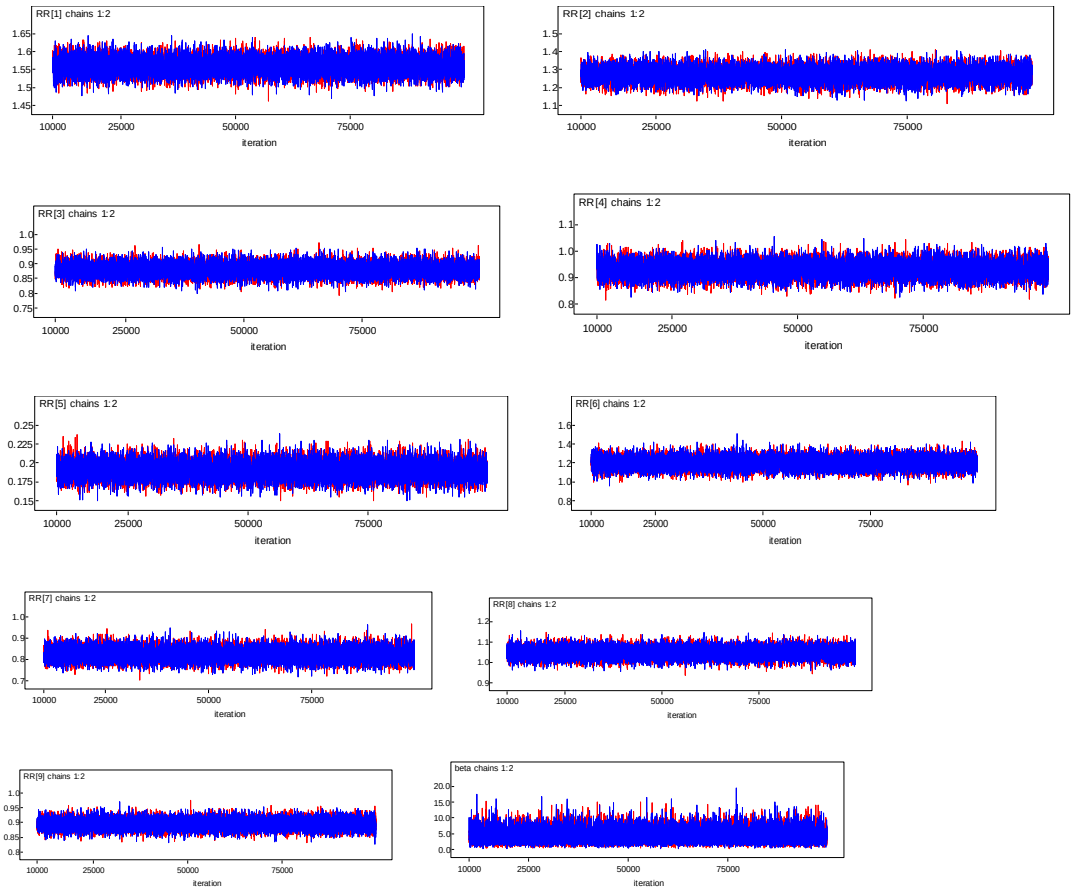


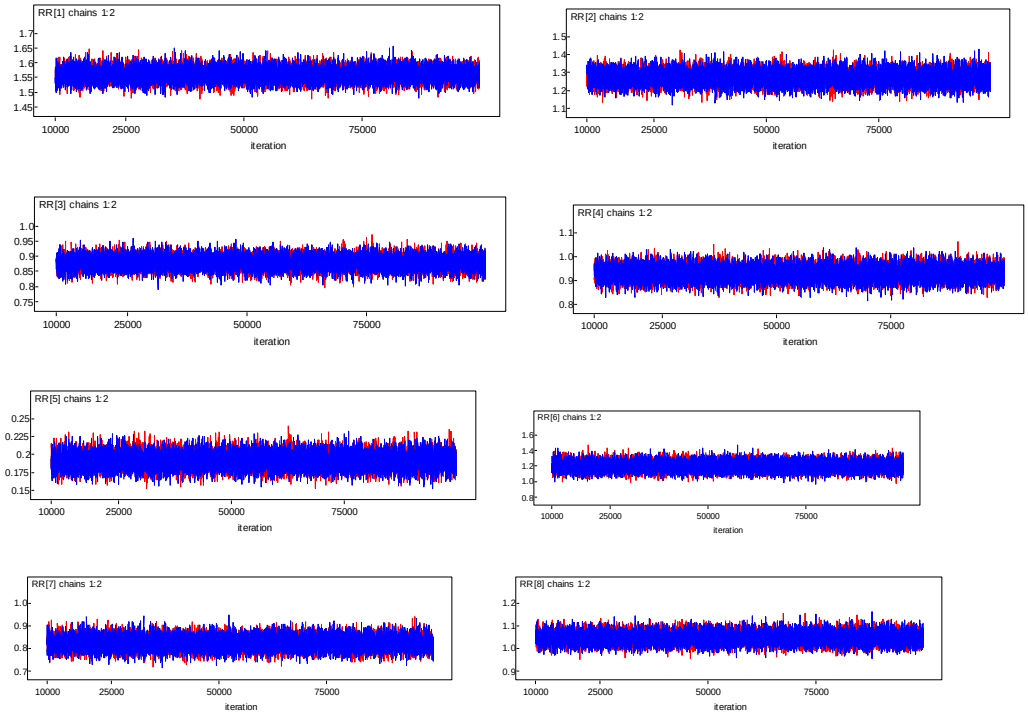
Figure 5.6: HIV history plot by fitting Poisson-Gamma

D.2 Poisson-GLMMs /Test CODE

```
# The model
model
{
# The likelihood
        for (i in 1:N) {
                HIVobserved[i]   ~ dpois(mu[i])
                terms[i]<-beta+b[i]
                log(mu[i]) <- log(HIVexpected[i])+terms[i]
                RR[i] <- exp(terms[i])                  # Area-specific relative risk (for maps)
                SMR[i]<-HIVobserved[i]/HIVexpected[i]

        }
# Non-spatial for random effects:

for (i in 1:9)
{
        b[i] ~dnorm(0.0,tau.b)
        }
```

```
        # Other priors:
        beta ~ dflat()
        tau.b  ~ dgamma(0.5,  0.0005)                    # prior on precision
        sigma <- sqrt(1 / tau.b)                         # standard deviation
}

# Initialization
list(beta=0,b=c(0,0,0,0,0,0,0,0,0),tau.b=1)
list(beta=0.5,b=c(0.5,0.5,0.5,0,2,0,0,0,0),tau.b=1)

Data
list(N=9)
HIVobserved[]  HIVexpected[]
4528 2900.172
998 783.9133
1622 1845.322
958 1029.959
278 1480.806
359 299.1019
762 924.9437
1563 1492.198
2643 2954.584
END
```

Figure 5.7: History: Complete trace plots for HIV fixed effects parameter fitting



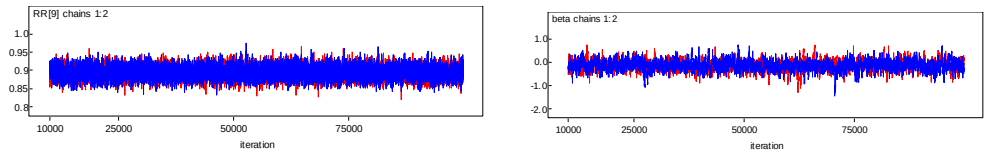


Figure 5.7: HIV history plot by fitting Poisson-GLMMs

D.3 Spatial CAR Model /Test Code

 # The model

```
model
{
# The likelihood
        for (i in 1 : N) {
                HIVobserved[i]   ~ dpois(mu[i])
                terms[i]<-beta+ b[i]
                log(mu[i]) <- log(HIVexpected[i])+terms[i]
                SMR[i]<-HIVobserved[i]/HIVexpected[i]
                RR[i] <- exp(terms[i])                          # Area-specific relative risk (for
maps)
        }

# CAR prior distribution for random effects:
        b[1:N] ~ car.normal(adj[], weights[], num[], tau)
        for(k in 1:sumNumNeigh) {
                weights[k] <- 1
        }

        # Other priors:
        beta~dnorm(0.0,0.001)
        tau ~ dgamma(0.5, 0.0005)                       # prior on precision
        sigma <- sqrt(1 / tau)                  # standard deviation
}
```

Data

 # The data

```
list(N=9)
HIVobserved[]  HIVexpected[]
4528 2900.172
998 783.9133
1622 1845.322
958 1029.959
278 1480.806
359 299.1019
762 924.9437
1563 1492.198
2643 2954.584

END


Adjecency information
list( num = c(3, 6, 4, 4, 3, 4, 4, 2, 4
```

),
adj = c(
4, 3, 2,
9, 7, 6, 4, 3, 1,
8, 6, 2, 1,
9, 5, 2, 1,
9, 7, 4,
8, 7, 3, 2,
9, 6, 5, 2,
6, 3,
7, 5, 4, 2
),
sumNumNeigh = 34)
Initialization
list(beta=0,b=c(0,0,0,0,0,0,0,0,0),tau=1)
list(beta=0.5,b=c(1,0,1,0,0,0,0,0,0),tau=1.4)

Figure 5.8: History: Complete trace plots for HIV fixed effects parameter fitting

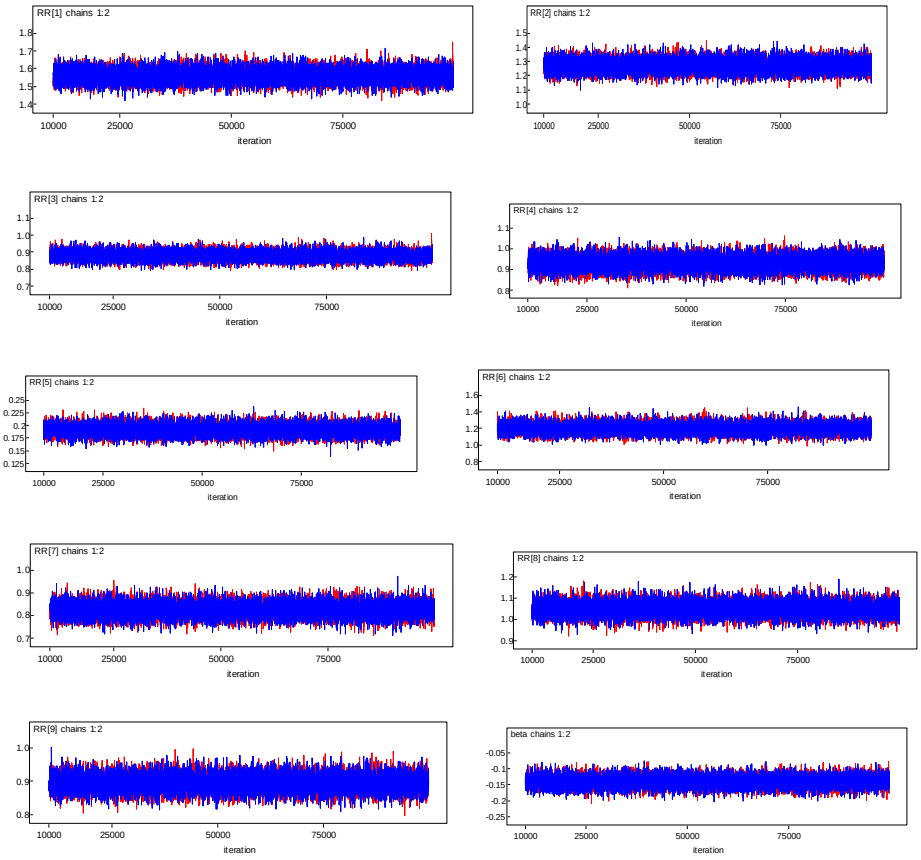


Figure 5.8: HIV history plot by fitting Spatial CAR model

D.4 Spatial Model with convolution priors /Test CODE

 # The model

model
{
# The likelihood

```
        for (i in 1 : N) {
                HIVobserved[i]    ~ dpois(mu[i])
                terms[i]<-beta+ b[i]+v[i]
                log(mu[i]) <- log(HIVexpected[i])+terms[i]
                SMR[i]<-HIVobserved[i]/HIVexpected[i]
                RR[i] <- exp(terms[i])                          # Area-specific relative risk (for
maps)
                v[i]~dnorm(0.0,1.0E-3)
               v[i]~dnorm(0.0,tau.v)
        }


# CAR prior distribution for random effects:
        b[1:N] ~ car.normal(adj[], weights[], num[], tau)
        for(k in 1: sumNumNeigh) {
                weights[k] <- 1
        }

        # Other priors:
        beta ~ dnorm(0.0,1.0E-4)

        tau  ~ dgamma(0.5, 0.0005)                      # prior on precision
        tau.v  ~ dgamma(0.5, 0.0005)                    # prior on precision
        sigma <- sqrt(1 / tau)                  # standard deviation
        sigma.v<-sqrt(1/tau.v)
}
```

Data

 # The data

```
list(N=9)

HIVobserved[]  HIVexpected[]
4528 2900.172
998 783.9133
1622 1845.322
958 1029.959
278 1480.806
359 299.1019
762 924.9437
1563 1492.198
2643 2954.584
END


Adjecency information
list( num = c(3, 6, 4, 4, 3, 4, 4, 2, 4
),
adj = c(
4, 3, 2,
9, 7, 6, 4, 3, 1,
8, 6, 2, 1,
9, 5, 2, 1,
9, 7, 4,
8, 7, 3, 2,
9, 6, 5, 2,
6, 3,
7, 5, 4, 2
```

),
sumNumNeigh = 34)

Initialization
list(beta=0,b=c(0,0,0,0,0,0,0,0,0,0),v=c(0,0,0,0,0,0,0,0,0,0),tau=1,tau.v=1)
list(beta=1,b=c(1,0,1,0,0,1,0,0,0),v=c(1,0,1,0,1,0,0,0,0),tau=1,tau.v=1.5)

END

Figure 5.9: History: Complete trace plots for HIV fixed effects parameter fitting

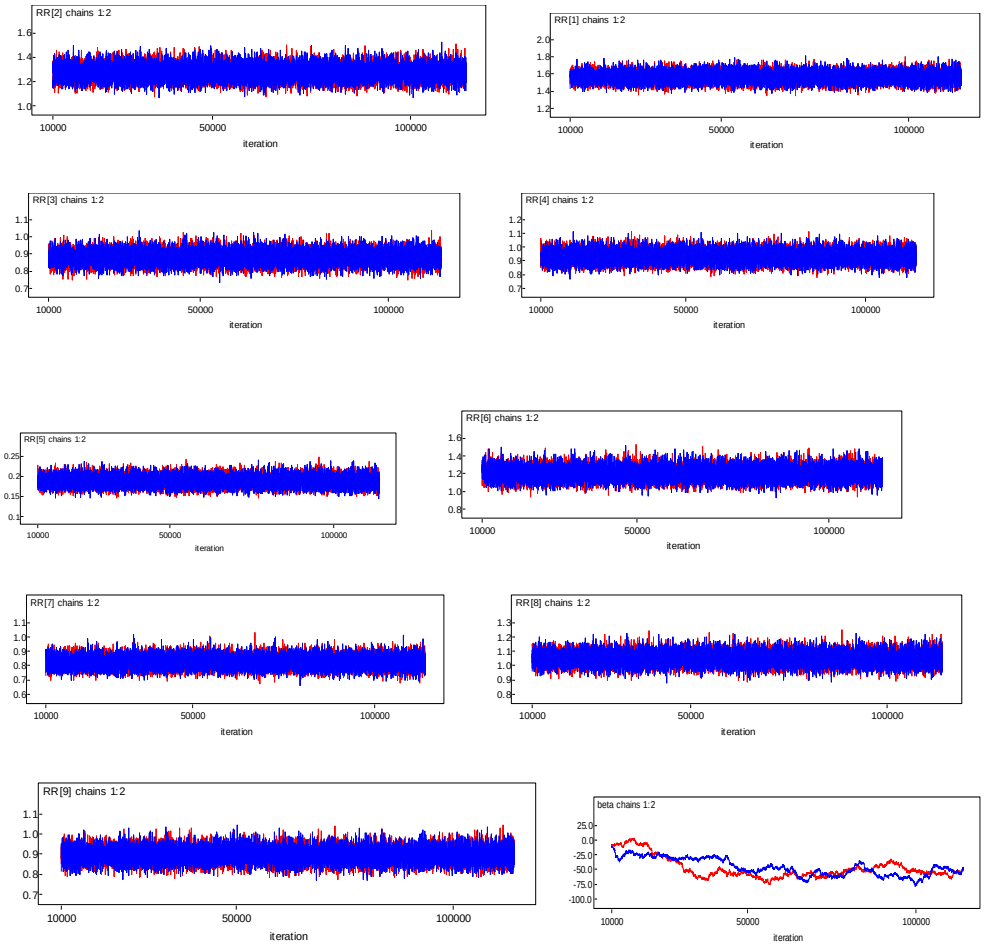


Figure 5.9: HIV history plot by fitting Spatial Convolution model