Statistical Methods for Analysing Complex Survey Data: An Application to HIV/AIDS in Ethiopia

BY
Mohammed O. M. Mohammed

Thesis Submitted in Fulfillment of the Requirement for the Degree of PhD in Applied Statistics

School of Mathematics, Statistics and Computer Science

University of KwaZulu Natal

Pietermaritzburg

South Africa

July 2013

Declaration of Authorship

I, Mohammed.O.M Mohammed, declare that this thesis titled, 'Statistical methods for Analysing Complex Survey Data: An Application to HIV/AIDS in Ethiopia' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Mr. M. Mohammed Signed____________________ Date____________________
Prof.T. Zewotir Signed____________________ Date____________________
Dr.T. Achia Signed____________________ Date____________________

Abstract

The HIV/AIDS pandemic is currently the most challenging public health matter that faces third world countries, especially those in Sub-Saharan Africa. Ethiopia, in East Africa, with a generalised and highly heterogeneous epidemic, is no exception, with HIV/AIDS affecting most sectors of the economy. The first case of HIV in Ethiopia was reported in 1984. Since then, HIV/AIDS has become a major public health concern, leading the Government of Ethiopia to declare a public health emergency in 2002. In 2011, the adult HIV/AIDS prevalence in Ethiopia was estimated at 1.5%. Approximately 1.2 million Ethiopians were living with HIV/AIDS in 2010.

Surveys are an important and popular tool for collecting data. Analytical use of survey data especially health survey data has become very common, with a focus on the association of particular outcome variables with explanatory variables at the population level. In this study we used the data from the 2005 Ethiopian Demographic and Health Survey, (EDHS 2005), and identified key demographic, socioeconomic, sociocultural, behavioral and proximate determinants of HIV/AIDS risk factor. Usually most survey analysts ignore the complex survey design issues like clustering, stratification and unequal probability of selection (weights). This study deals with complex survey design and takes the design aspect into account, because failure to do so leads to bias parameters estimates and standard error, wide confidence intervals and statistical tests will be incorrect.

In this study, three statistical approaches were used to analyse the complex survey data. The first approach was a survey logistic regression used to model the binary outcome (HIV serostatus) and set of explanatory variables (the dependence of the HIV risk factors). The difference between survey logistic regression and the ordinary logistic regression is that survey logistic regression approach takes the study design into account during analysis. The second approach was a multilevel logistic regression model, that assumed that the data structure in the population was hierarchical, and that individual within household was selected from clusters that were randomly selected from a national sampling frame. We considered a three-level model for our analysis. This second approach considered the results from Frequentist and a Bayesian multilevel models. Bayesian methods can provide accurate estimates of the parameters and the uncertainty associated with them. The third approach used was a Spatial models approach where model parameters were estimated under the Integrated Nested Laplace Approximation (INLA) paradigm.

The results showed that, the age variables was highly significantly for contracting HIV and that young people were at higher risk. The place of residence was also found to be significant for acquiring HIV and people that who lived in urban area were found to be at higher risk. In terms of marital status, separated people were found to be at higher risk for contracting HIV. Unlike other studies, there was a linear correlation between the HIV risk and two predictors namely socio-economic status and education. However, the impact of socio-economic status was different for the male and female populations. The results showed that, the wealthier people were at higher risk compared to the poorer. The circumcised male were found to be at lower risk for contracting HIV compared to their counterpart. In terms of stigma related to HIV, the people with low stigma were found to be at higher probability for contracting HIV. The key finding was that both male and females who considered themselves to be at lower risk for contracting HIV were, in fact, the most likely to be HIV positive.

We mapped HIV prevalence in Ethiopia at regional levels. The analysis for all three methods are separately carried out for males and females. The study identifies the key factors associated with HIV risk in Ethiopia. The findings agree to large extent with others in the literature and could be used in the design of the policy and public health interventions to address trends in occurrence of the HIV epidemic.

Acknowledgements

First of all, I thank ALLAH for grace and mercy showered upon me by giving me the power to finish my PhD. This work could not have been accomplished without assistance of a number of people, and I would like to take this opportunity to thank them for their help.I would like to thank Prof. T Zewotir for his help. I would like to express my deep and sincere gratitude to my co supervisor, Dr Thomas Achia, whose guidance and stewardship anchored me through to the very end. He has taught me both consciously and unconsciously how the sound application of statistics can contribute to scientific knowledge in other disciplines. I appreciate all his contributions of time and ideas, which made my PhD experience productive and stimulating. The joy and enthusiasm he has for his research was contagious and motivational to me, even during tough times in the PhD pursuit.

I would also like to thank Professor Henry Mwambi for his support. Many thanks to my parents, brother, and sisters and my lovely uncle Musa for trusting me. To my friends and colleagues at UKZN and beyond, a very big thank you to all of you. I would like to thank Alneelain University for funding my PhD. Many thanks to my Sudanese friends in PMB who make me feel at home. My special thanks go to my best friend ever Mohammed Abdalaziz for everything. I extend my gratitude to Dr FGA Awad and his family for their hospitality and welcoming me and my wife at all times with open arms. Also I would like to thank Dr Ahmed Khidir for helping me a lot in Latex.

Finally, I would like to thank my lovely wife Saloma for moving to the other side of the world with me and for being there through both good and bad times. Thank you for all your love, friendship and continuous support and coming to the office in the night, especially during these last months, when my absentmindedness has reached new heights. I could not imagine doing this PhD without you, and I look forward to dive into our future adventures.

Contents

List of Figures

List of Tables

Dedication

To my mother

To my father

To my brothers and sisters

To my lovely wife Saloma
To my friends

Chapter 1

Introduction

1.1 Background

Ethiopia is a land locked country which is situated on the horn of Africa, with a total area of 1,127,127 square kilometers, bordered by six countries namely Eritrea, Djibouti, Somalia, Kenya, South Sudan and Sudan. Due to a high altitude differences within the country, the temperature of the country ranges from below 0^o C in the Simien mountains to a high of 48^0 C in the danakil depression, the hottest zone in the world. Although Ethiopia is categorised under least developed countries with a substantial proportion of people are living under the poverty level, there are so many different cultural heritages, and various ethnic groups who speak more than 80 different local languages. Ethiopia has its own unique alphabet, a diverse climate, and is known for maintaining of its independence for a long period of time, even during the African colonialism.

Ethiopia is divided into nine regions and two administrative cities. Each region except the regions of the two administrative cities is led by its own regional president. Each region has the right to decide and lead its people without the interference from the central government of the country, except in some sensitive issues like military and monetary policies which are managed at country level. A wide range of heterogeneities of different of cultures, traditions, and living standards are part of inter and intra regions of the country. The largest proportion of the country's population is in the Oromia region (36.7%), followed by the Amhara region (23.3%) and the Southern Nations, Nationalities and People regions (SNNP). The lowest proportion lives in Harari regional state which constitutes 0.2% of the country total population [29].

According to the 2007 population and housing census of Ethiopia, the Ethiopian population had soared to 76.9 million, with an annual growth rate of 2.6 million people per year. Nearly 83.8% of the population live in the rural areas and the rest 16.2% live in urban areas. The distribution of the population by sex of males and females is 50.5%, 49.5% , respectively. Also according to the Federal Democratic Republic of Ethiopia Population Census Commission (FDREPCC), the proportion of adults of working age (15-64), children (0-14) and elderly above 65 years is 51.9% , 45% and 3.2% respectively. The most dominant religion in Ethiopia is Orthodox Christianity, followed by Islam, Protestant and other traditional religions. There are variations among the proportions of the religious followers live in urban and rural areas. Most Muslim and traditional religious followers are live in rural areas, whereas the majority of Orthodox, Christians reside in urban areas.

Within the past two decades, like many other African countries, Ethiopia has experienced a growing HIV epidemic. The first laboratory diagnosed HIV infection in the country was confirmed in 1984 [75], and AIDS cases were first diagnosed in Addis Ababa in 1986 [97]. The pandemic started as a concentrated epidemic, and the initial cases were found among commercial sex workers [108]. A few years later, the infection spread to the general population, and the HIV positive cases were found among pregnant women visiting Antenatal Clinics (ANC) and among blood donors, specifically in the capital city Addis Ababa [53]. Today, the HIV/AIDS epidemic in Ethiopia is considered a generalised epidemic, which has affected all demographic, socio-economic and institutional populations of the society. According to the latest estimates of the joint United Nations Program on HIV/AIDS in 2006, an estimated 1.3 million Ethiopians were living with HIV, 220000 of whom were children [144]. A better understanding of the spread of the epidemic across various regions and population groups, and an understanding of how the knowledge of prevention methods,attitudes and behavioral factors affect the risk of HIV infection is critical in formulating effective prevention, treatment, care, supporting programmes and policies.

Previous studies on the prevalence and correlates of HIV infections in Sub-Saharan Africa has shown large differentials in the prevalence of HIV by age, sex, place of residence (urban, rural) and geographical region within and between countries [18, 111]. There are also other factors, such as, educational attainment, occupation and exposure to the media, that can influence risk taking behaviours [21, 41, 44, 78], and also lead to increased risk of HIV infection. Obviously, people with little education tend to have poor access to information on safe sex and are less likely to use condoms [94].

A number of studies have shown that having many sexual partners and having a casual sexual partners' increase the risk of getting infected with HIV [32, 132, 137, 151, 155]. Other studies have shown that having one regular partner can reduce the risk of getting HIV infected [110]. It has been argued that having concurrent sexual partners in a dense sexual network increases the risk of HIV infection by allowing the virus to spread rapidly to others [76, 81, 115]. Three recent clinical trial in Sub-Saharan Africa have shown that male circumcision can significantly reduce the risk of getting infected [8, 117, 154]. Knowledge of HIV prevention methods and positive attitudes of acceptance towards people living with HIV have been advocated to bring about a change in higher risk sexual behaviuors, and in turn reduce the spread of HIV infection [120]. Although, there are considerable efforts being made to promote knowledge of HIV prevention methods, reduce misconceptions, and promote accepting attitudes toward people living with HIV, stigma and discrimination against people living with HIV remain common in most countries [139].

1.2 Country profile

Ethiopia is one of the Sub-Saharan African countries which has been hard hit by the HIV pandemic and is thus home to a substantial number of infected people. Different surveys have presented varying results with regard to the estimated prevalence rate of HIV/AIDS in Ethiopia. For instance, Ethiopia's Federal Ministry of Health (FOMH) reported that the 2005 (ANC) survey showed an estimated prevalence rate of 3.5% nationally, 10.5% in urban areas, and 1.9% in rural areas. According to the country federal HIV/AIDS control office, the rate of prevalence in the year 2005/2006 was 2.1% nationally, 7.7% in urban areas, and 0.9% in rural areas. The national HIV prevalence figures do, however, overlook heterogeneities among different regions and places of residence such as urban versus rural areas in different parts of the country. Also when we look at individual regions of Ethiopia the HIV rates differ greatly for instance the prevalence rate ranges from 0.2% in the SNNP region to 6% in the Gambela region (EDHS 2005). The disparity between the regions with different prevalence rates is the result of the different socio-economic, cultural and demographic factors. Although the prevalence of the epidemic is low at national level when compared with many other Sub-Saharan countries, the number of people throughout the country living with HIV is still alarming and according to the estimate published in the 2005 Ministry of Health (MOH) report Ethiopia's HIV infected population had reached 1.32 million people in that year, which put Ethiopia among the top fifteen countries in the world in terms of the number of HIV infected individuals.

The government of Ethiopia, a long with many international and National Governmental Organizations (NGOS), has actively participated in the process of mitigating the of the epidemic within the country [10, 35, 39, 92]. The National HIV/AIDS Council Secretariat, headed by the president of Ethiopia, and the HIV/AIDS Prevention and Control Office (HAPCO) were the two main government organisations established in order to try and prevent and control the pandemic at national level. The Ethiopian Ministry of Health policy focuses on preventing the spread of the disease rather than finding a cure [109, 113]. In order to keep this agenda active the department's primary focus is therefore to identify the main factors or causes behind the spread of HIV/AIDS in the country and then take measures based on its findings to prevent further spread of the disease.

Most of the people in Ethiopia know very little about HIV/AIDS. However, they have traditionally been aware of a disease called Amenmin, and they are also at least aware of the presence of the epidemic of HIV/AIDS in Ethiopia, even if they do not know really anything more about it [112]. In 1985 and 1986 a total of 5265 Ethiopian recruit soldiers were tested for HIV; 0.1% of them were found to be infected and the presence of the disease in the country was officially recognised. By 1999, the HIV/AIDS prevalence of the Ethiopian population as a whole had risen to 2.8%.

There has been similarly an increase in the HIV prevalence rate among the country female commercial sex worker [107]. In 1988, for instance, 18.5% of the female commercial sex worker were infected, but by 1989 the prevalence had increased to 29.2%. The highest rates were found among female commercial sex worker working in the major transport routes in and around Addis Ababa. Similarly high rates were found among long distance truck drivers who work in the same routes [51].

Ethiopia is one of the poorest countries in world [145, 153]. It is currently facing three extremely serious problems, namely HIV/AIDS, recurrent drought and malaria [145]. Furthermore, the country has a history of civil war. These man-made and natural disasters negatively affect Ethiopia's agriculture dominated economy, thus helping to explain the low performance of the country's economy [4, 141]

1.3 Knowledge and misconceptions about HIV in Ethiopia

In Ethiopia, in response to the HIV/AIDS epidemic, different television and Movie actors have taken part in various programmes and used a variety of communication channels and approaches to promote widespread behavioural changes. However, these interventions have not yielded the intended outcomes or results [77]. Surveys on HIV/AIDS related to knowledge, attitudes and practices of Ethiopians show that, although there is a high level of awareness about HIV/AIDS, many people still lack adequate know how, when it comes to preventing the transmission of HIV, and many misconceptions surrounding the disease persist [150]. The studies also make it clear that the disparity between knowledge and practice is considerable. The preliminary findings of the round of the Behavioral Surveillance Survey (BSS) reveal that despite a high level of awareness about HIV/AIDS, only about 55% of the population know all three methods of HIV prevention, namely abstaining from sex, being faithful to one partner and using a condom (known as the ABC principle: Abstinence, Be Faithful to one partner and Condom). Furthermore, a study carried out in the town of Gambela revealed that only 0.9% of the 359 interviewed individuals knew about mother to child transmission of HIV/AIDS [118] and this is inspite of the fact that about 10% of new HIV infections in Ethiopia occur as a result of mother-to-child transmission according to Ministry of Health (MOH) report 2004.

Misconceptions about HIV transmission are also very common [12]. For instance, the preliminary findings of the second round of the BSS show that local misconceptions like eating uncooked eggs produced by a chicken that has swallowed a condom or eating raw meat prepared by an HIV infected person could result in your contracting the virus still remain high; such misconceptions were held at the time of the survey by more than 40% of those from all the studied groups, except the school youths, where the figure was 10%.

A thorough assessment of the HIV/AIDS prevention and transmission result of the second round of the BSS suggests a lack of comprehensive knowledge among the majority of the Ethiopian population. According to the findings of these surveys, less than 20% of the respondents knew all three preventive methods and had no misconceptions about how the disease is transmitted.

Small surveys carried out in different settings also reveal the presence of misconceptions among different sections of the Ethiopian population, for example, [16] interviewed 1214 students at Addis Ababa University and found that, in spite of a high level of knowledge about HIV/AIDS, 34% of the students still considered vaccination to be one of the modes of HIV transmission. In a similar study carried out at Jimma University, 58% of 500 interviewed students did not know that persistent use of condoms prevents HIV/AIDS, and 16% were not aware that a seemingly healthy looking person can transmit HIV [104]. All the above reviewed research results attest to the fact that, there is still a low level of HIV awareness and a low level of knowledge in Ethiopia with regard to HIV prevention and transmission.

1.4 Theoretical framework

The epidemic of HIV is spread unevenly in the world [6]. The prevalence of the pandemic is variable in different societies and regions. Some regions of the World are highly affected by this pandemic, and in other areas it is negligible. For instance, the society in Sub-Saharan Africa region is hard hit by HIV and the prevalence rate is very high as compared to the other region in the world. Many researchers investigate the factors that determine the prevalence of HIV in a given society by examining proximate factors that are associated with the prevalence of HIV driven by demographic, socio-economic and cultural variables. The risk of incidence and prevalence of HIV infection can be reduced or spread out widely depending on the culture, demographic and socioeconomic factors that determine prevention and the mitigating abilities of a given society [11, 92].

In this section the relationship between HIV infection and demographic, biological proximate, socio-economic and cultural factors that may affect the risk of transmission will be discussed based on the proximate determinant theoretical framework of Bonerma and Weir [19].

Many demographical and epidemiological studies are accompanied by theoretical frameworks to guide and visualise the relationship between different explanatory variables and the epidemics of some diseases. Furthermore, the collection analysis and interpretation of certain data also need to be guided by the theoretical frameworks that explain the association between the given variables [19]. The proximate determinants and the theoretical framework for factors affecting the risk of sexual transmission of

HIV/AIDS which is presented by Bonerma and Wier will be applied to this study to assess the association between the main socio-economical, demographical and cultural factors that fuel the prevalence of the of HIV epidemic in Ethiopia.

The proximate determinant framework was first introduced by [42], who constructed an analytical framework for the study of the sociology of fertility by setting intermediate variables through which social factors affect the fertility levels. This framework was further developed by different researchers; [20], who replaced the term intermediate variables used by [42] with proximate determinants. He established that the variation of the level of fertility can be decomposed into some proximate determinants which are marriage, postpartum infecundity, abortion and contraceptive methods. Next [20, 116] developed a child survival framework by using proximate determinants for fertility as abase. This framework has been further developed and widened by different scholars at different times along with the proximate determinants.

In 2005, Bonerma and Wier [19] developed and presented a theoretical framework that has guided the study and analysis of the distribution and determinant of HIV infection. As shown in Figure 1 the underlying determinants link to the biological determinants through the proximate determinants. These underlying determinants (socio-economic, socio-culture and demographic factors) together with intervention programmes, influence the proximate determinants, which affect the incidence and prevalence of the infection, that leads to disease and ultimately to death. Since the main focus of this study is to identify the socio-economic, socio-culture and demographical factors that which fuel the HIV/AIDS pandemic, the discussion and the focus will be on following the sequence of the schema. The theories and discussion traced below mainly focused on the predominant mode of transmission (heterosexual) of HIV/AIDS in most of the world especially in Sub-Saharan Africa and specifically in Ethiopia.

The underlying determinants that include the term as 'context' and intervention programmes are supposed to operate with the proximate determinants to determine the health outcome of an individual as shown in the above schema. Here, the first overview of proximate and biological determinants is separately assessed, and then their interaction with the underlying determinant is discussed in the context of Ethiopia.

The proximate determinants in the theoretical framework of HIV includes variables which are influenced by underlying determinants that determine the health outcomes of individuals. The proximate determinants can be categorised into three groups:

(1) biological mechanisms which determine the efficiency of transmission,

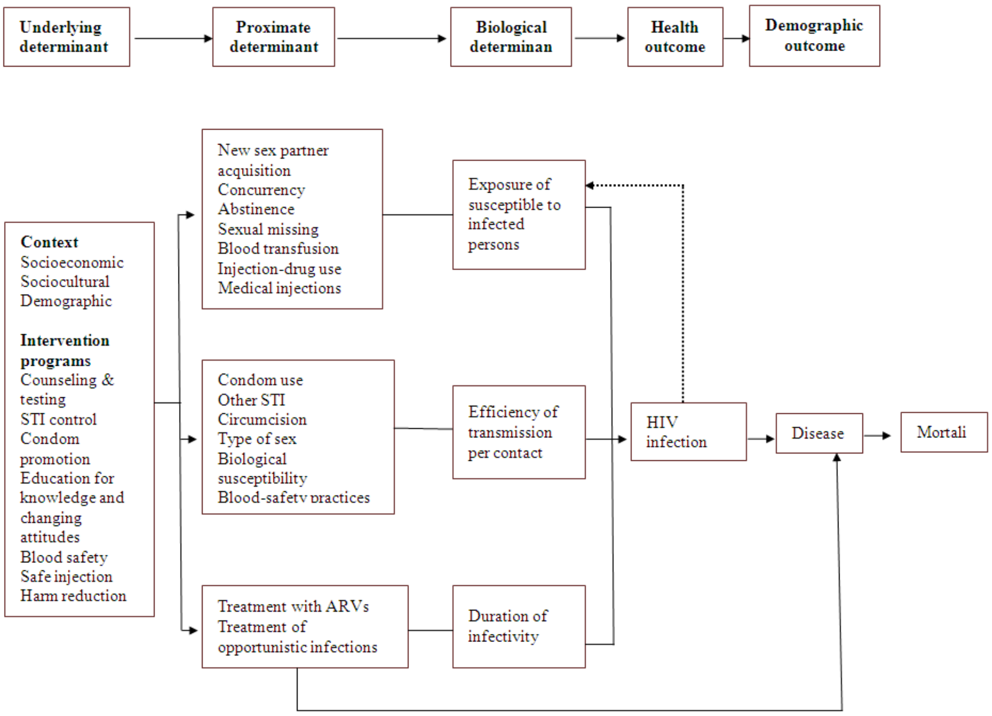


FIGURE 1: Proximate-determinants theoretical framework for factors that affect the transmission of HIV/AIDS. Source: Bonerma and Wier, (2005)

(2) physical barriers or practice that limit the dose of transmission of the virus that exposed to HIV and

(3) factors which influence both, the biological mechanisms of exposure means the virus load, which is the amount and concentration of the virus, and the biological susceptibility of the person exposed to infection and infectious virulence of pathogens.

Physical barriers or practice encompasses methods and practices are used to prevent the epidemic by reducing the exposure to the virus, these are condoms, gloves and bleached sharp equipments. Sexually Transmitted Diseases (STDs) are included in the factors that increase both biological exposure and the efficiency of transmission of the viruses.

Biological determinants are the basic components that determine the transmission and prevalence of an epidemic via three biological ways.

(1) the rate of exposure of a susceptible person to an infected person.

(2) the efficiency of transmission during exposures, which is determined by the proximate determinant of virus load, and

(3) the duration of the infectivity.

The efficiency of transmission of the virus may be protected or reduced by using a proximate determinant or physical barrier and practising, the biological component duration of infectivity of STDs has three stages as quoted by [19]. The duration of infectivity is dependent on the natural history of the infection and divided into an acute initial case during which the virus loads are high, a subsequent phase during which then virus loads decrease and the final phase which will increase again [19]. The duration of infectivity can be shortened by using some treatment.

HIV infection can be caused by the transfer of blood, semen, vaginal fluids, or breast milk form an infected to a healthier person. The transmission of these body fluids can be determined by two biological determinants: exposure susceptibility to infected persons and efficiency of transmission per contact. The proximate determinants, namely acquisition of new sex partner, number of concurrent sexual partners, abstinence, blood transfusion, and health care related injections serve to determine the exposure susceptibility to infected persons. Similarly, the efficiency of transmission is determined by the proximate determinant of: use physical barriers, blood safety practices, circumcision and STDs. The proximate determinant in turn is also affected by underlying determinants.

The underlying determinants determine the risk of exposure, and the transmission and infection of HIV/AIDS through the sequential links of proximate and biological determinants as shown in Figure 1. Demographic, socio-economic and cultural variables are included under the underlying determinants that operate with proximate determinants to determine the prevalence of HIV. Socio-economic factors are usually related to the building of attitudes of individuals and a healthy environment through education and level of income. Education influence the proximate determinants through raising of the awareness individuals on how to protect themselves from different kind of diseases.

Education can be used to build awareness about different issues, by accessing and perceiving different information through reading and experience. For instance, education may help an individual in the short and medium term by improving awareness of how people can protect themselves from being infected and how to practice safe by disease and methods which are mentioned in the proximate determinants that determine the prevalence of HIV. On the other side, in the long run education will help society to eradicate poverty and other related risky factors, through innovation and increasing the level of income of individuals. Also, the level of income affects the prevalence of HIV through practising risky behaviours that expose people to the epidemic. The impact of poverty on the prevalence of HIV ranges from direct causes of money constraints to buy protective barriers like condoms. Poverty is one of the socio-economic factors which is often associated with a low income levels. Low income levels (poverty) is usually mentioned as a driving force in risky sexual behaviours such as pursuing money to satisfy basic needs, that can reduce the risk of infection, to the lack of a balanced diet that can be related to weakening of the immunity systems of individuals.

Cultures and norms are other factors which affect the prevalence of HIV through the practice of harmful traditions like the cutting and tearing of human body parts in a traditional rituals . These practices influence the proximate determinants in two different ways. In the first place, the sharp equipments that are used for practicing these traditional practices may not be clean but infected, and in the second place this traditional process may be accompanied by blood transfusions, due to excessive bleeding during the process, which can lead to increasing of the efficiency of the transmission of HIV from infected to healthier persons. For instance, female circumcision is highly associated with HIV/AIDS infection because of the equipments that are not clean and may cause infections, large amounts of bleeding during the genital mutilation that may be accompanied by blood transfusion, or vaginal tearing during sexual intercourse. In some cultures of different societies, polygamy is widely practised. These traditions influence the proximate determinants through acquisition of many sexual partners, and therefore the susceptibility of an individual to HIV/AIDS.

Demographic factors are associated with the proximate determinants of the biological susceptibility of individuals and the acquisition of new sexual partners together with the frequency of sexual perform within each partner the susceptibility of individuals to HIV. The Women are biological more susceptible to HIV infection than men. Concerning, the demographic variables of age, and marital status of an individual, the proximate determinant of acquiring new sexual partners play a great role in determining the prevalence of HIV [157]. Different studies show that young people especially women are at high risk of HIV infection due to various cultural, biological and social factors. Physiologically and behaviourally youths are rushing to sex and practising unsafe sex. Therefore, many are at high risk of acquiring STD, which facilitate the transmission of HIV.

Among all the people in the world infected by STD, two thirds are youth aged twenty five and below. Similarly, due to the fact that the youth are more attracted to money, material goods and sex than older age group, they are more frequently exposed to unsafe sex specially if they are at low economic and strive to satisfy their material aspiration [119]. The marital status of an individual interacts with the proximate determinant, through the acquisition of new sexual partners.

1.5 Statement of the problem

Survey data is usually used to draw inferences about a population under study. However, complex survey is hard to handle due to three main reasons:

(1) the data is often clustered due to the use of multi-stage stratified cluster samples or longitudinal survey designs, implying that observations within the same cluster are correlated.

(2) sample units are often selected with unequal probabilities (weight). When these probabilities are related to the outcome variable, sampling becomes informative and the model holding for the sample is then different from the model holding in the population.

(3) survey data is almost inevitably subject to nonresponse, often of considerable magnitude, which again may affect the model holding for the responding units if the response probability is correlated with the outcome (not missing at random nonresponse).

Several approaches have been proposed for modelling and analysing complex survey data. These approaches differ in the conditions underlying their use, the data requirements, the inference objectives that they accommodate, statistical efficiency, computational demands, and the skills required for their implementation. This heterogeneity means that no single approach can be considered as best, or even operational, in all situations.

Different factors interact in a complex manner to contribute infection and spread HIV infection, which makes the control and prevention of the epidemic difficult [118]. The United Nations (UN) Declaration of commitment to HIV/AIDS identifies poverty and illiteracy as the major contributing factors to increasing the risk of HIV/AIDS in developing countries. In a similar way, [91] identifies poverty as a major deriving force behind HIV/AIDS in Ethiopia.

The socio-economic status of individuals in society presents a unique challenge for them to access basic social services including HIV/AIDS related messages. As a result, they are not exposed to all the known risk factors of HIV infection [72, 73]. The hearing and visually impaired also face significant disadvantages in most societies. Too often, they are not considered as a target group for HIV prevention education and HIV outreach efforts because there is misconception that they are not sexually active and therefore exposed to the risk of HIV infection [47, 72]

Population mobility, poverty and gender relationships are mentioned as factors at the macro level. Individual sexual behaviours, the number of sexual partners, and use of protection during sexual intercourse are taken as factors at the micro level in determining HIV infection [54]. According to MOH report 2004 and Addis Ababa City Administration Health Bureau (AACAHB),Unexpected sex with multiple sexual partners is the major means of HIV infection nationally and locally, in Addis Ababa. It is understood that there is a direct link between individual behaviour and individual sexual practices that influence HIV infection [28, 43]. Individual behaviour can also be shaped by available target specific HIV/AIDS education. However, studies indicate that HIV/AIDS related information is less accessible to persons with visual and hearing impairments because radio and television messages miss the deaf, while the television and billboard messages do not reach the blind [9, 72, 156]. This implies that it is not only necessary but also mandatory to have target specific HIV/AIDS intervention programs for sensory disabled populations.

1.6 The objectives of the study

The primary objectives of this study are

(1) to describe the patterns and distribution of HIV infection among adult women and men in various population groups in Ethiopia as well as to establish factors which most contribute to HIV infection.

(2) to assess the relationship between HIV status and selected demographic, socioeconomic and cultural factors.

(3) to develop statistical models for the prevalence of HIV in Ethiopia using demographic and health survey data from 2005 and strategies for addressing complex sampling design when computing advanced statistical procedures such as survey logistic regression, multilevel modeling and Spatial modeling.

(4) to develop HIV risk maps are for HIV distribution in Ethiopia for males and female.

1.7 Significance of the study

The significance of the study stems from two important considerations. Firstly, HIV is a serious pandemic that has affected and is affecting many sections of the World's populations, especially in developing countries. HIV has no cure and is still menacing. It is, therefore, critical to study and monitor the different dynamics and aspects of the disease and the global and local responses to control it. Secondly, there is a need to investigate and expose the extent of the different risk factors and levels of HIV through statistical models and techniques which play an important role in investigating and predicting factors affecting different aspect of people lives. The former helps in gaining insight and resolving real life problems. As the ongoing resolutions in computing relieves the burden of calculating and graphing, the emphasis on statistical concepts and insight from the data becomes both more important and practical.

1.8 Sample design

The 2005 EDHS sample was designed to provide estimates for health and demographic variables of interest for the following domains: Ethiopia as a whole, the urban and rural areas of Ethiopia (each as a separate domain), and 11 geographic areas (9 regions and 2 city administrations), namely: Tigray, Affar, Amhara, Oromiya, Somali, Benishangul-Gumuz, SNNP, Gambela, Harari, Addis Ababa and Dire Dawa. In general, DHS sample is stratified, clustered and selected in two stages. In the 2005 EDHS a representative sample of approximately 14,500 households from 540 clusters was selected. The sample was selected in two stages. In the first stage, 540 clusters (145 urban and 395 rural) were selected from the list of Enumeration Areas (EA) from the 1994 Population and Housing Census sample frame, then divided into zones and each zone into weredas. In addition to these administrative units, each wereda was subdivided into convenient areas called census EAs. Each EA was either totally urban or rural and the EAs were grouped by administrative wereda. Demarcated cartographic maps as well as census household and population data was also available for each census EA. The 1994 Census provided an adequate frame for drawing the sample for the (EDHS 2005). As in the (EDHS 2000), the EDHS (2005) sampled three of seven zones in the Somali region (namely Jijiga, Shinile and Liben). In the Affar region the incomplete frame used in 2000 was improved adding a list of villages not previously included, to improve the regions representativeness in the survey. However, despite efforts to cover the settled population, there may have been some bias in the representativeness of the regional estimates for both the Somali and Affar regions, primarily because the census frame excluded some areas in that had a predominantly nomadic population.

The 540 EAs selected for the EDHS are not distributed by region proportional to the census population. Thus, the sample for the (EDHS 2005) must be weighted to produce national estimates. As part of the second stage, a complete household listing was carried out in each selected cluster. The listing operation lasted three months from November 2004 to January 2005. Between 24 and 32 households from each cluster were then systematically selected for participation in the survey.

Because of the way the sample was designed, the number of cases in some regions appear small since they are weighted to make the regional distribution nationally representative.

1.9 Layout of the thesis

This thesis is organised as follows. In Chapter 2, we describe the survey logistic regression model and discuss approaches used to estimate parameters to carry out statistical inference.

Chapter 3 is a stand alone as a research article for the application of the survey logistic regression model to HIV prevalence data.

In Chapter 4, we describe the hierarchical Bayesian multilevel models and discuss the model estimation methods.

Chapter 5 is a stand alone as a research article for the application of multilevel models to HIV prevalence data.

In Chapter 6, we describe the spatial modeling and mapping for complex survey data, and how parameter estimation is carry out using INLA technique. Furthermore, we discuss the model selection criteria.

Chapter 7 is standing as a research article of spatial analysis and modeling of HIV prevalence data using integrated nested laplace approximation technique (INLA).

Chapter 8 summaries the thesis and gives directions for future research.

1.10 The research contributions

In this section the contributions of the research is outlined

(1) most of the researchers use logistic regression and χ^2 test to fit the HIV data models, which is means that they ignore the sample design. However, in this research we used survey logistic regression which is takes into account the sample design aspects(unequal selection), is used

(2) One of the most important contribution of this thesis is a comparison between ferquentist multilevel models, and Bayesian multilevel models, and the results are similar

(3) using Integrated Nested Laplace Approximation (INLA) to estimate and mapping HIV data in Ethiopia.

Chapter 2

Survey logistic regression models for analyzing complex survey data

2.1 Introduction

In Demographic and Health Surveys most data sets are collected using complex survey designs involving stratification, multi-stage clustering, and unequal sampling weights embedded in the plan. The population is usually divided into subpopulation, referred to as strata. Within each stratum, a sample is selected from the sampling units independently. For the statistical data analysts it will therefore be important to realize that the variance of estimates in models fitted will decrease if the sampling units within each stratum are homogeneous. Failure to account for the stratification in the analysis may result in overestimation of the standard error, and hence too wide confidence interval. We shall discuss this feature in the subsequent discussion.

The second common feature in complex survey data is clustering. In clustering, the total population is divided into groups (or clusters) and a sample of the groups is selected. In multistage clustering, the clusters selected at first stage are called Primary Sampling Units (PSUs). It is within these PSUs where desired samples are obtained using probabilistic, or non-probabilistic sampling, approach. In general, a failure, by a data analyst, to account for the clustering in the analysis may lead to underestimation of variabilities and incorrect conclusions.

Finally, the impact of the unequal selection of respondents in each PSU that occurs in demographic and health surveys needs to be addressed careful by statisticians. Sampling weight are used to model unequal units in the population that are represented by a single PSU sampled. The unequal selection probabilities inherent in the design of these surveys must be taken into account in the analysis to reduce the bias of the estimates and the underestimation of variabilities.

This chapter introduces and interrogates key statistical properties of the Survey Logistic Regression model and discusses approaches used to estimate parameters and to carry out statistical inference.

2.2 Model description

The logistic regression models is a member of the broader class of models referred to in the literature as Generalized Linear Models (GLM). These models are specifically designed to model the relationship between binary data and a set of predictor variables. The standard logistic regression model is, however, inappropriate for analyzing survey data with clustering and stratification that may be present in survey designs. Therefore, some adjustments to the classical methods that take account of the survey design elements are necessary in order to make valid inferences from the survey data [33]. Therefore, logistic regression models used to analyze data from the complex sampling designs have been referred to as survey logistic regression models in the literature [5, 96]. The statistical properties of the survey logistic regression models are similar to those of the ordinary logistic regression models. Survey logistic regression models simply account further for complexity in design. For survey logistic regression, the first stage in each stratum sampled is correctly modeled as the primary sampling units (PSUs).

Let $y_{hij}, i = 1, 2, \ldots, n_h; j = 1, \ldots, n_{hi}; h = 1, \ldots, H$ denote a dichotomous random variable taking value 1 if the j^{th} individual in the i^{th} household nested within h^{th} cluster is HIV seropositive, and 0 otherwise. Also let $\pi_{hij} = P(y_{hij} = 1)$ denote the probability that diagnosis is positive in the $i-th$ household within $j-th$ PSU nested within $h-th$ stratum.

Suppose $\boldsymbol{X}_{hij}$ is the row vector corresponding to the characteristics of the j^{th} individual in the i^{th} household nested within h^th cluster. Furthermore, we assume that the rows of the matrix $\boldsymbol{Y}$ represent $n = \sum_{h=1}^{H} \sum_{i=1}^{n_i} n_{hi}$ observation of the response variables Y and the rows of the $\boldsymbol{X}$ are n observations of the r explanatory variables $X_1, \ldots, X_r$.

We shall assume that Y_{hij} belongs to the exponential family of distributions with the sampling distribution defined as follows:

$$f\left(y_{hij}, \theta_{hij}, \phi\right) = \exp\left\{\frac{y_{hij}\theta_{hij} - b(\theta_{hij})}{a(\phi)} + c(y_{hij}, \phi)\right\}, \tag{2.1}$$

where $f(.)$ denotes the density function of y_{hij}, θ_{hij} is known as the natural parameter and ϕ is known as the dispersion parameters.

Furthermore, we will assume a model for the mean vector $\boldsymbol{\mu}_{hij} = E[y_{hij}]$ which can be re-write as

$$\boldsymbol{\mu}_{hij} = \boldsymbol{m}(\boldsymbol{X}_{hij}, \boldsymbol{\beta}), \tag{2.2}$$

where $m(.)$ denotes a vector-valued function of $\boldsymbol{X}_{hij}$ and $q \times 1$ vector $\boldsymbol{\beta}$ of unknown parameters. The model need to be transformed to a linear model by using an appropriate link function and the mean of the distribution which provides the relationship between the linear predictor. More specifically, the linear model of the (GLM) is given by

$$\eta_{hij} = g(\mu_{hij}) = \boldsymbol{X}'_{hij}\boldsymbol{\beta}, \tag{2.3}$$

where $\boldsymbol{X}'_{hij}$ denotes a vector of covariates and $g : \mathcal{R} \rightarrow \mathcal{R}$ denotes the link function for a binary outcome as in this study based on survey logistic regression the link function is $\eta_{hij} = logit(\mu_{hij})$ thus the generalized logit model can be written as

$$logit(\pi_{hij}) = \mathbf{X}'_{hij}\boldsymbol{\beta}. \tag{2.4}$$

2.3 Estimation Methods

In this section, we introduce the Maximum Likelihood Estimation (MLE) technique used to estimate parameters in the survey logistic regression model. Due to the complex design properties such as unequal probability of selection, clustering and stratification the ordinary MLE may not work properly. As a result, we resort to a Pseudo Maximum Likelihood Estimation (PLME) approach which takes sample design aspects into account as follows:

$$\ell = \log(\boldsymbol{\beta}|\boldsymbol{y}) = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} w_{hij}\varphi_{hij} \ln f\left(y_{hij}, \theta_{hij}, \phi\right), \tag{2.5}$$

where w_{ijh} , φ_{hij} and $f(y_{hij};\boldsymbol{\theta})$ denotes the weight, frequency and the probability density function for the j^{th} individual in the i^{th} household nested within h^{th} stratum, respectively.

In order to estimate the unknown parameters we need to differentiate the log-Likelihood function given in (2.5) with respect to $\boldsymbol{\beta}$. This results in the gradient function, which can be used to determine the desired result. That is,

$$\mathcal{D}(\boldsymbol{\beta}) = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} w_{hij}\varphi_{hij}\boldsymbol{D}'_{hij}\boldsymbol{\Sigma}^{-1}[\boldsymbol{y}_{hij} - \hat{\boldsymbol{\mu}_{hij}}] = 0, \tag{2.6}$$

where $\boldsymbol{D}_{hij} = \left[\frac{\partial \boldsymbol{\mu}_{hij}}{\partial \boldsymbol{\eta}'_{hij}}\right]\boldsymbol{A}_{hij}$, and $\boldsymbol{\Sigma}$ denotes the covariance matrix of y_{hij}.

In general, equation (2.6) do not have a close form of solution and an iterative scheme is required to obtained the maximum likelihood estimates of the unknown parameters, β.

2.3.1 Parameters estimation

In this subsection we derive expressions for the maximum likelihood estimators in a typical survey logistic regression.

Assuming that the outcomes variable y_{hij} follows Bernoulli distribution with density function

$$f(y_{hij}) = \pi_{hij}^{y_{hij}}(1-\pi_{hij})^{1-y_{hij}}, \tag{2.7}$$

the mean and variance of y_{hij} are respectively,

$$\mu_{hij} = \frac{exp\{\mathbf{x}'_{hij}\boldsymbol{\beta}\}}{1+exp\{\mathbf{x}'_{hij}\boldsymbol{\beta}\}} \text{ and } \sigma^2 = \mu_{hij}(1-\mu_{hij})$$

and where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_r)'$ denote the vector of parameters.

The log-likelihood function that forms the basis for maximum likelihood estimation is given by

$$\ell = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} w_{hij}\varphi_{hij}\left\{y_{hij}\log(\mu_{hij}) + (1-y_{hij})\log(1-\mu_{hij})\right\}, \tag{2.8}$$

Substituting the values of μ_{hij} into this expression we get the

$$\ell = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} w_{hij}\varphi_{ijh}\left[y_{hij}\log\left(\frac{e^{x'_{hij}\boldsymbol{\beta}}}{1+e^{x'_{hij}\boldsymbol{\beta}}}\right) + (1-y_{hij})\log\left(1-\frac{e^{x'_{hij}\boldsymbol{\beta}}}{1+e^{x'_{hij}\boldsymbol{\beta}}}\right)\right]. \tag{2.9}$$

In order to obtain the unknown parameters we have to differentiate the log-likelihood with respect to $\boldsymbol{\beta}$ to get the following equation

$$\begin{aligned}\frac{\partial\ell}{\partial\boldsymbol{\beta}} &= \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} w_{hij}\varphi_{hij}\frac{e^{x'_{hij}\boldsymbol{\beta}}}{\left(1+e^{x'_{hij}\boldsymbol{\beta}}\right)^2}\left[\frac{y_{hij}}{1-\left(1+e^{x'_{hij}\boldsymbol{\beta}}\right)^{-1}} - \frac{1-y_{hij}}{1+e^{x'_{hij}\boldsymbol{\beta}}}\right]\mathbf{X}'_{hij},\\ &= \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} \varphi_{hij}\boldsymbol{D}'_{hij}\left[\boldsymbol{\sigma}^2(y_{hij})\right]^{-1}[y_{hij}-\mu_{hij}],\end{aligned}$$

where $\boldsymbol{D}_{hij} = \mu_{hij}(1-\mu_{hij})\mathbf{X}'_{hij}$.

The Fisher information matrix for the parameters of the Bernoulli model follows as

$$\begin{aligned}\mathfrak{I} &= -E\left[\frac{\partial^2\ell}{\partial\boldsymbol{\beta}\boldsymbol{\beta}'}\right],\\ &= \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} Ax'^2_{hij}cd^{-2}\left[y_{hij}(1-2d^{-1}) - (1-y_{hij})((1+c)^{-1} - 3(1+c)^{-2}c)\right],\end{aligned}$$

where $A = \varphi_{hij}w_{hij}$, $c = exp(x'_{ijh}\boldsymbol{\beta})$ and $d = (1+exp(x'_{hij}\boldsymbol{\beta}))$.

After simplifying the previous equation we get the following equation which is referred to as the Fisher Information

$$\mathfrak{I} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} w_{hij}\varphi_{hij}\boldsymbol{D}'_{hij}\left[\boldsymbol{\sigma}^2(y_{hij})\right]^{-1}\boldsymbol{D}_{hij}, \tag{2.10}$$

2.3.2 Approximate covariance matrix

This section provides The iterative algorithm to obtain maximum likelihood estimates of the elements of β. The Fisher Scoring algorithm may be introduced as follows. If

$\hat{\boldsymbol{\beta}}^{(t)}$ denotes the t^{th} successive approximation to $\hat{\boldsymbol{\beta}}$, then the $(t+1)^{st}$ approximation is obtained from

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + \left\{ I_n(\hat{\boldsymbol{\beta}}^{(t)}) \right\}^{-1} g(\hat{\boldsymbol{\beta}}^{(t)}), \tag{2.11}$$

where the gradient vector $g(.)$ is

$$g(\boldsymbol{\beta}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} g_{hij}(\boldsymbol{\beta}), \tag{2.12}$$

where

$$g_{hij}(\boldsymbol{\beta}) = w_{hij}\varphi_{hij} D'_{hij} \{\Sigma(y_{hij})\}^{-1}[y_{hij} - \mu_{hij}], \tag{2.13}$$

and $I_n(\boldsymbol{\beta})$ is the Fisher information matrix (2.10). The results derived are based on [17] and use of first order-Taylor linearization. Denote the contribution to the gradient vector of each first-stage element for a given sampling stage by g_{hij}, where h denotes stratum, and i the i^{th} unit within this stratum. The index j denotes a typical final stage element contained within the PSU h, then

$$\hat{w}(\hat{\boldsymbol{\beta}}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} g_{hij}(\hat{\boldsymbol{\beta}}) = 0, \tag{2.14}$$

Using the first order-Taylor expansion of $\hat{w}(\hat{\boldsymbol{\beta}})$ at $\hat{\beta} = \beta$, it follows that

$$0 = \hat{w}(\hat{\boldsymbol{\beta}}) \approx \hat{w}(\hat{\boldsymbol{\beta}}) + \frac{\partial \hat{w}(\boldsymbol{\beta})}{\partial \beta'}(\hat{\beta} - \beta), \tag{2.15}$$

Taking variances on both sides of (2.15)

$$Cov\left(\hat{w}(\hat{\boldsymbol{\beta}})\right) \approx \frac{\partial \hat{w}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} Cov(\hat{\boldsymbol{\beta}}) \left(\frac{\partial \hat{w}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\right)', \tag{2.16}$$

It is obviously $\frac{\partial^2 lnL}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \frac{\partial}{\partial \boldsymbol{\beta}} \left[\frac{\partial g(\boldsymbol{\beta})}{\partial (\boldsymbol{\beta}')}\right]$ is a non-singular matrix,

$$Cov(\hat{\boldsymbol{\beta}}) \approx \left[\frac{\partial^2 lnL}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right]^{-1} Cov\left(\hat{w}(\hat{\boldsymbol{\beta}})\right) \left[\frac{\partial^2 lnL}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right]^{-1},$$

where $E\left[\frac{\partial^2 lnL}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right] = -I_n(\boldsymbol{\beta})$.

Therefore, an approximate expression for the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$Cov(\hat{\boldsymbol{\beta}}) \approx I_n^{-1}(\boldsymbol{\beta}) G I_n^{-1}(\boldsymbol{\beta}), \tag{2.17}$$

where $G = Cov\left(\hat{w}(\hat{\boldsymbol{\beta}})\right)$. Using the results derived by [56], it follows that, under single stage sampling with replacement (WR) or without replacement(WOR),

$$G = \sum_{h=1}^{H} \frac{n_h - f_h}{n_h - 1} \sum_{i=1}^{n_h} \left(t_{hi.} - t'_{h..}\right)\left(t_{hi..} - t'_{h..}\right)', \tag{2.18}$$

where $n_h = \sum_{j=1}^{n_{hi}} m_{hij}$, with m_{hij} being the number of cases with identical response patterns within stratum h, cluster i and ultimate sampling unite (USU) j. The term $\varphi_h = \frac{n_h}{N_h}$, which is called the sampling rate for stratum h, $t_{hij} = g_{hij}(\hat{\boldsymbol{\beta}})$ is given by substituting $\hat{\boldsymbol{\beta}}$ in equation (2.13) and as a consequence $t_{hi.} = \sum_{j=1}^{m_{hij}} t_{hij}$ and $t'_{h..} = \frac{1}{n_h} \sum_{i=1}^{n_h} t_{hi}$.

2.4 Conclusion

This chapter introduced the survey logistic models that take into account the sample design aspect.

When using a complex sampling design to draw a sample from a finite population, the sample design should be incorporated into the analysis of the survey data, in order to make valid statistical inferences for the finite population.

In this chapter, we defined a survey logistic regression model and showed how the parameter estimates are conducted.

The parameter estimates have been carried out using the pseud MLE method as the ordinary MLE method does not incorporate the design aspect into account.

The strength of a survey logistic model is that, it can handle the complex survey design aspect.

Chapter 3

Application of the survey logistic regression model to HIV prevalence data

This Chapter is stand alone as a research article of application of the survey logistic regression to HIV prevalence data in Ethiopia. The aims of this paper is to identify key demographic, socio-economic, socio-cultural, behavioural and proximate determinants as risk factors for HIV prevalence among men and women in Ethiopia. The analysis has done separately for men and women because the biological and social circumstances association with transmission of HIV differ by sex [85].

Prevalence and determinants of HIV infection in Ethiopia[1]

Mohammed O. M. Mohammed[2], T. N. O. Achia, and T.Zewotir
School of Mathematics, statistics and Computer Science,
University of KwaZulu-Natal, Private Bag X01 Scottsville 3209, Pietermaritzburg,
South Africa

Abstract. The HIV/AIDS pandemic is currently the most challenging public health issue to face third world countries, especially those in sub-Saharan Africa. Ethiopia, which is located in East Africa and has a generalised and highly heterogeneous epidemic, is no exception, with HIV affecting most sectors of its economy.

This study used data from the 2005 Ethiopia Demographic and Health Survey to identify key demographic, socio-economic, socio-cultural, behavioural and proximate determinants as risk factors for HIV prevalence among men and women in Ethiopia. In accordance with the sampling design, a survey logistic regression was carried out with a binary response variable, HIV serostatus, and with a separate analysis for both the male and female populations.

The results indicate that women are at higher risk of HIV infection, with a prevalence rate of 2%, as compared to men, who have a prevalence rate of 1%. Marital status and education were the main socio-demographic factors highly associated with HIV prevalence. Separated women (OR=9.76,95% CI=4.73-20.15) were at greater risk when compared with married women. Unlike other studies, there was a linear correlation between the HIV status and the two predictors namely socio-economic status and education. However, the impact of socio-economic status was different for both the male and female populations. Increased wealth was positively related to HIV infection: the wealthiest men were 3.63 times more likely than the poorest men to be HIV infected . The men and women from Addis Ababa city were at significantly higher risk for infection with HIV (OR=1.51 and 1.92, respectively) compared to Afar region. Uncircumcised men were nearly 2 times more likely to be HIV-positive than those who were circumcised. A key findings was that both men and women who considered themselves to be at low risk for acquiring HIV were, in fact, the most likely to be HIV-positive.

This study identifies key factors associated with HIV prevalence in Ethiopia. The findings agree to a large extent with the existing literature and can be used in the design of policy and public health interventions to address trends in the HIV pandemic.

Keywords: EDHS 2005; HIV; Survey Logistic Regression; Complex survey Design; HIV prevalence

[1] Submitted to the SAHARA - Journal of social aspects of HIV/AIDS research alliance

[2] Corresponding Author. 209509598@stu.ukzn.ac.za

3.1 Introduction

AIDS is a disease caused by the human immunodeficiency virus (HIV), which weakens the immune system of a person. AIDS was first recognized in the 1980s. Since then it has become an epidemic that has spread rapidly all over the world. The number of people living with HIV/AIDS grows substantially every year (HIV/AIDS and Youth, 2003). According to Global Health Policy Report of the United States (US), the number of the people living with HIV in the world is 33.4 million, and it is thought that more than 2 million people died due to this epidemic in the year 2008 (Epidemic, 2008). The number of people living with HIV in the world rose from 8 million in 1990 to 35.3 million by the end of 2009 [148].

Different studies have all revealed that the prevalence rate of the pandemic is extremely high in developing countries, especially when compared with first world countries [49]. The region of Sub-Saharan Africa is the most severely affected by this pandemic. According to the 2002 estimates from the World Health Organization (WHO), the Sub- Saharan Africa region accounts for more than 67.7% of the total number of people around the world living with HIV/AIDS .

Ethiopia is one of sub-Saharan Africa hardest hit nation in terms of the HIV pandemic and thus is home to a substantial number of infected people. The first case of HIV in Ethiopia was diagnosed in Addis Ababa hospital in 1986 [98]. Different surveys have produced varying results with regards to the estimated prevalence rate of HIV/AIDS in Ethiopia. For instance, Ethiopia's Federal Ministry of Health (FMOH) reported that the 2005 Antenatal Clinic (ANC) survey showed an estimated prevalence rate of 3.5% nationally, 10.5% in urban areas, and 1.9% in rural areas. According to the country's Federal HIV/AIDS Control Office (HAPCO), the prevalence in 2005/2006 was 2.1% nationally, 7.7% in urban areas, and 0.9% in rural areas. The national HIV/AIDS prevalence figures do, however, overlook the heterogeneities among different regions and places of residence. The disparity between the regions different prevalence rates is the result of the different socio-economic, cultural and demographic factors that fuel the HIV prevalence rates in those specific regions. While the prevalence rate of the epidemic is low at national level when compared with many other Sub-Saharan Africa countries, the number of people throughout the country living with HIV is still enormous, and according to the estimate published in the 2005 Ministry of Health (MOH) report, Ethiopia's HIV infected population reached 1.32 million people that year, which puts Ethiopia among the top fifteen countries in the world in terms of its number of HIV infected individuals.

Sampling design is commonly ignored in statistical analysis of the survey data, mainly because of the possibility to incorporate in the model specifications with certain characteristics of the sampling design. It has been shown that ignoring the sampling design leads to biased and misleading estimates of the standard error, Confidence intervals and statistical tests [22, 95].

The logistic regression model is classified under the Generalised Linear Models (GLM) and is used to model the binary data. Standard statistical methods are inappropriate for analyzing complex survey data due to the clustering and stratification used in the survey design. Some adjustments to the classical methods that do not take into account of the survey design are therefore necessary in order to make valid inference from the survey data [33]. Furthermore, logistic regression models can be used to analyze data from complex sampling designs, and these are known as survey logistic regression models. Such models are based on the same theory as are ordinary logistic regression models. The difference between them is that the survey logistic accounts for the complexity of survey design.

The aim of this study was therefore Survey to identify key demographic, socio-economic, socio-cultural, behavioural and proximate determinants as risk factors for HIV prevalence among men and women population of Ethiopia.

3.2 Methodology

3.2.1 Data

This study is based on a secondary analysis of the data from the (EDHS 2005). The survey was conducted from April 27 to August 30, 2005. The survey sample was designed to provide national, urban/rural, and regional estimates of key Health and demographic indictors. In the first stage, 540 clusters were selected from the list of Enumeration Areas (EA) in the 1994 population and housing census. Fieldwork was successfully completed in 535 of the 540 clusters. In the second stage, 24 to 32 households were selected systematically from each cluster for the survey sample. The survey administered the women's questionnaire to all eligible women in age 15-49 in the sampled households. The men's questionnaire was administered to all eligible men aged 15-59 in every household. In order to test for HIV the survey collected blood specimens from all eligible women and men in the household selected for the male interview. The response rates for HIV testing were 83% among women and 76% among men . The analysis used data from 14070 women and 6033 men age who had completed interview, who reported ever having had sexual intercourse, and who had a valid HIV test results. Because of the way the sample was designed, the number of cases in some regions appears small since they are weighted to make the regional distribution nationally representative.

In order to carry out the test of HIV, in a random sample of 50% of the households selected from the survey, all eligible women and men were asked to provide their consent for a blood draw and subsequent testing for HIV; in case of young in age 15-17 the consent was obtained from their parent. Everyone for whom consent was obtained provided three to four drops of blood from a finger prick collected on a filter paper with a special bar code label. The blood samples were transported to Addis Ababa to be tested for HIV at the Ethiopia Health and Nutrition Research Institute (EHNRI), a national laboratory. HIV testing was conducted using standard laboratory and quality control procedures. The blood collection and HIV testing protocol allowed anonymous linking of the HIV test result to an individual's socio-demographic and behavioural characteristics obtained from the individual questionnaires. Bar codes were used to make this link, after household and cluster identification codes were scrambled to ensure that all potential identifiers had been destroyed.

3.2.2 Covariates

The independent variables included sociodemographic characteristics (age, region, place of residence, education, religion, marital status), Socio-cultural factors (decision making ability, wealth index, stigma, circumcision), sexual behaviour characteristics (number of sexual partners in the last 12 months, history of STI in the last 5 years, age at first sex).

Principle Component Analysis (PCA) [48] was used to generate the stigma, media exposure and the ability of decision making. Stigma is defined as an attribute or label that sets a person apart from others and links the labeled person to undesirable characteristics [36]. Stigma related to AIDS has been defined as "the prejudice, discounting, discrediting, and discrimination that are directed at people perceived to have AIDS" [64]. A stigma index was created based on responses to the questions "willing to care for relative with AIDS", "person with AIDS allowed to continue" and "would buy vegetables from vendor with AIDS ". Based on factor scores, respondents were classified as having low, medium or high HIV-related stigma. A media exposure index was also computed using PCA based on responses to questions posed on the frequency of watching television, the frequency of listening to radio, and the frequency of reading newspapers. The respondents were then classified as having low, medium or high media exposure. The decision making index was computed based on the respondents answer to the questions: Final say on own health care, final say on making large household purchases, final say on making household purchases for daily need, final say on visits to family or relatives, final say on food to be cooked each day, and final say on deciding what to do with money husband earns. The decision making index was a trichotomous variable with levels defined as independent, consults and subservient.

3.2.3 Analysis plan

Separate analysis were carried out for the male and female data because the biological and social circumstances associated with transmission of HIV differ by sex. The HIV serostatus was the primary response variable in this analysis. Several other variables identified from the literature were cross-classified with this variable. SAS 9.2 (SAS Institute, Cary, NC) was used to carry out statistical analyses. Univariate analysis was done to assess the distribution of the sample and to compute overall prevalence of HIV. Both point estimates and robust 95% confidence intervals (based on robust standard errors after adjusting for strata and clustering at Primary Sampling Unit (PSU) level are presented. Bivariate and multivariate survey logistic regression models were used to assess the unadjusted and adjusted association, respectively, of different socioeconomic, demographic characteristics with HIV. Both bivariate and multivariate regression models were fitted after applying sampling weights and adjusting for multistage clustered sampling designs effects using PROC SURVEYLOGISTIC in SAS 9.2 (SAS Institute, Cary, NC)

3.3 Results

In this section we present the results obtained from the analysis, and we compute the Adjusted Odd Ratios (AOR), Odd Ratios (OR) and the 95% Confidence Interval (CI) which are based on the estimated coefficients reported in the tables that follow. As we mentioned earlier the analysis was done separately for females and males, and so for this reason we present their results separately.

3.3.1 Summary statistics

We now present the effect of the selected covariates on HIV prevalence separately for males and females . Our results suggest an association between age and HIV prevalence for both male and female populations. Prevalence of HIV was highest amongst men in 40-44 age groups (2.84%). The females in the 35-39 age group are at highest prevalence (4.44%). In terms of regions, the males of Gambela are at highest prevalence (6.18%), followed by Addis Ababa (3.3%), Affar (2.21) and Harari (2.05%) and the rest of the regions have less than (2%) prevalence. The Addis Ababa females have the highest prevalence (6.06%), followed by Harari (4.59%) and Dire Dawa (4.36%). All the other regions have prevalence less than (4%). HIV prevalence is higher in urban areas (2.62%, 7.73%) for both males and females respectively. In terms of number of children have dead, the HIV prevalence rate is highest amongst those with no children dead with (1.9%, 2.41%) for males and females respectively. The education level the highest prevalence found among those with highest education level followed by those with higher education for both males and females. In terms of wealth index for both males and females, the highest prevalence are found among the richest (2.21%, 6.09%) respectively. All the others categories of the wealth have prevalence 1% or below. The HIV prevalence is highest amongst the separated men (2.64%), followed by married (1.24%). The prevalence rate is very low among those never married. The separated females have highest prevalence compared the other categories. Religion for both males and females, the Orthodox have the highest prevalence rate (1.6%, 2.91%) respectively, and the rest have prevalence rates less than 1%. In terms of media exposure, both males and females the high risk are among those with high media exposure (1.68%, 4.58%) respectively. For both males and females HIV risk increase with lower HIV-related stigma. The uncircumcised male has the highest prevalence rate (1.11%). The highest prevalence among consults females (2.01%), followed by independent (1.86%) and very low for subservient females.

Table 1: Design adjusted HIV seroprevalence rates among the female respondents by selected covariates

Variable	Prevalence (N)	Variable	Prevalence (N)
Age, $p = 0.0011$		Age at first sex, $p = 0.9438$	
15-19	0.69(1405)	Under 14 yrs	2.51(802)
20-24	1.72(1083)	14-17 yrs	2.47(2341)
25-29	2.11(1103)	18+ yrs	2.24(1373)
30-34	1.48(727)	Religion, $p < 0.001$	
35-39	4.44(679)	Orthodox	2.91(2833)
40-44	3.06(516)	Protestant	0.96(1018)
45-49	0.85(429)	Other	0.79(2091)
Region, $p < 0.001$		Birth in the last 5 years, $p = 0.6180$	
Tigray	2.56(564)	no	1.99(2984)
Afar	3.25(295)	yes	1.74(2958)
Amhara	1.83(822)	STI, $p = 0.7155$	
Oromiya	2.23(965)	No	1.86(5905)
Somali	1.3(258)	Yes	2.72(22)
Ben-Gumz	0.9(389)	Other wives, $p = 0.1635$	
SNNP	0.1(997)	No other	2.47(2270)
Gambela	5.51(342)	1 other	1.52(3186)
Harari	4.59(345)	2 other	1.45(486)
Addis Abeba	6.06(673)	Media exposure, $p < 0.001$	
Dire Dawa	4.36(292)	low	0.67(2984)
Place of Residence, $p < 0.001$		medium	1.21(980)
Urban	7.73(1628)	high	4.58(1946)
Rural	0.65(4314)	HIV knowledge, $p = 0.0036$	
No child who have died, $p = 0.2769$		low	3.08(1960)
never had children	1.29(1942)	medium	2(1492)
no child dead	2.41(2310)	high	1.04(1727)
one child died	1.72(881)	Stigma, $p < 0.001$	
2+ child dead	1.84(809)	high	0.61(1476)
Education, $p < 0.001$		medium	3.02(1217)
No	1.03(3630)	low	5.66(1145)
Primary	2.45(1323)	Life time partners, $p < 0.001$	
Secondary	6.05(853)	Never had sex	0.12(1418)
Higher	1.05(136)	One	1.44(3288)
Wealth index, $p < 0.001$		More	4.95(1219)
Poorest	0.34(1224)	Decision making, $p = 0.0768$	
Poorer	1(942)	independent	1.86(1102)
Middle	0.43(927)	consults	2.01(1373)
Richer	0.21(879)	subservient	0.72(1229)
Richest	6.09(1970)	Total	1.86(3704)
Marital status, $p < 0.001$			
Never married	0.7(1557)		
Married, only wife	1.59(3230)		
Married, other wives	1.45(486)		
separated	6.43(669)		

3.3.2 Logistic regression results

Table 3 presents results of the survey logistic regression model fitted to the male and female data sets separately. The age variable was found to be significant in terms of HIV prevalence. Women aged 20 to 49 years are significantly more affected than are younger women aged 15 to 19 years. Those women living in Addis Ababa are more than twice as likely as those living in Affar region to be HIV-positive; whereas the risk for contracting HIV is much lower in SNNP region. The odds of acquiring HIV is significantly higher among women who have never married (OR=2.26, 95% CI=1.12, 4.57) when compared with married women. Women living in rural areas have a significantly decreased odds of HIV when compared with their urban counterpart. In terms of media exposure, those women who are exposed to lower media have higher odds of HIV (OR=8.38, 95% CI=4.75-14.79) compared with those women have higher media exposure. Women with primary education are nearly twice as likely to be HIV-positive compared to those with no education; although the odds for being HIV-positive is higher for the secondary education category. Also the findings indicate an

Table 2: Design adjusted HIV seroprevalence rates among male respondents by selected covariates

Variable	Prevalence (N)	Variable	Prevalence (N)
Age, $p < 0.001$		Marital status, $p < 0.001$	
15-19	0.12(1080)	Never married	0.29(2034)
20-24	0.36(890)	Married/Living together	1.24(2877)
25-29	0.7(702)	Separated	2.64(196)
30-34	1.94(636)	Age at first sex, $p < 0.001$	
35-39	1.82(543)	Not had	0.16(1560)
40-44	2.84(414)	Less	0.12(58)
45-49	0.01(365)	14-17	1.49(766)
50-54	0.88(287)	18+	1.21(2662)
55-59	0.34(190)	Religion, $p < 0.001$	
Region, $p = 0.0044$		Orthodox	1.6(2493)
Tigray	1.97(474)	Protestant	0.38(811)
Affar	2.21(233)	Other	0.17(1801)
Amhara	1.42(814)	Stigma, $p = 0.7166$	
Oromiya	0.41(959)	No	0.92(5069)
Somali	0(193)	Yes	0(24)
Benishangul-Gumuz	0(332)	No of other wives, $p = 0.9457$	
SNNP	0.37(822)	Not married	1.39(30)
Gambela	6.18(296)	1	1.25(2662)
Harari	2.05(280)	2	1.08(185)
Addis Ababa	3.3(518)	Circumcised, $p = 0.7376$	
Dire Dawa	1.73(186)	No	1.11(404)
Place of residence, $p < 0.001$		Yes	0.9(4693)
Urban	2.61(1158)	Ever paid for sex, $p = 0.4377$	
Rural	0.64(3949)	No	1.24(3066)
No of children dead, $p < 0.001$		Yes	2.27(67)
never had children	0.45(2286)	Media exposure, $p = 0.0105$	
no child dead	1.92(1495)	Low	0.6(1442)
one child died	0.44(646)	Medium	0.64(1881)
More than one	0.7(680)	High	1.68(1763)
Education, $p = 0.0036$		Stigma, $p < 0.001$	
No	0.77(2113)	High	0.24(1747)
Primary	0.52(1722)	Medium	1.23(1361)
Secondary	2.1(1099)	Low	1.71(1795)
Higher	1.65(173)		
Wealth index, $p < 0.001$			
Poorest	0.62(1034)		
Poorer	0.29(859)		
Middle	0.85(818)		
Richer	0.37(822)		
Richest	2.21(1574)		

inverted U-shaped relationship between education and HIV risk. Wealth is positively and monotonically related to being HIV prevalence, with those in wealthiest quintile being 5 times more likely to be infected than those in the poorer quintile. Middle-aged men have higher odds of HIV infection than younger men because of the former's greater sexual activity. Those men living in urban areas are 4 times as likely to be positive compared to those in urban areas. Religion is also a significant factor, with those men reporting that they are Orthodox being nearly 10 times as likely to be HIV-positive as others. The education factor has been found to have a significant effect on HIV prevalence, the more educated men are at higher risk (OR=2.51, 95% CI =1.14, 4.19) than their lesser educated counterparts. Men who are uncircumcised are being nearly 1.5 times as likely to be HIV-positive as circumcised men. Unlike the women, separated men are at higher risk (OR=2.16, 95% CI=1.47, 3.17) when compared with married men. Richer men have a higher risk of contracting the disease (OR=3.36, 95% CI=1.42-9.32) than do poorer men. Men who have little exposure to the media are at a higher odds of HIV (OR=0.27, 95% CI=0.13-0.55) compared to those who are highly exposed to the media.

Table 3: Distribution of female and male respondents by selected covariates and HIV serostatus

	Female		Male	
Variable	OR (95% CI)	AOR(95% CI)	OR (95% CI)	AOR(95% CI)
Age				
15-19	1	1	1	1
20-24	2.53(0.95,6.76)	1.12(0.62,2.01)	2.91(0.36,3.23)	1.84(0.23,14.49)
25-29	3.12(1.23,7.92)	0.46(0.23,0.91)	5.69(0.78,4.33)	2.76(0.39,9.19)
30-34	2.17(0.67,6.19)	0.78(0.44,1.38)	15.93(2.12,9.56)	6.47(0.89,7.23)
35-39	6.73(2.64,17.12)	1.15(0.64,2.07)	14.89(1.81,2.00)	6.94(0.86,5.91)
40-44	4.56(1.51,13.78)	0.83(0.39,1.80)	23.51(2.92,8.13)	11.21(1.35,9.12)
45-49	1.24(0.49,3.11)	0.001(0.01,0.03)	0.08(0.01,0.56)	0.04(0.05,0.27)
50-54			7.14(0.93,5.61)	4.08(0.50,3.25)
55-59			2.74(0.38,1.57)	1.36(0.18,10.41)
Region				
Afar	1		1	
Addis Abeba	1.92(0.91,4.07)		1.51(0.48,4.72)	
Amhara	0.56(0.23,1.37)		0.64(0.19,2.15)	
Ben-Gumz	0.27(0.09,0.80)		0.04(0.02,0.07)	
Dire Dawa	1.36(0.57,3.22)		0.78(0.23,2.64)	
Gambela	1.74(0.75,4.04)		2.92(0.94,9.06)	
Harari	1.43(0.68,3.01)		0.93(0.31,2.80)	
Oromiya	0.68(0.30,1.55)		0.18(0.04,0.80)	
SNNP	0.03(0.01,0.09)		0.17(0.03,0.81)	
Somali	0.39(0.19,0.81)		0.001(0.002,0.06)	
Tigray	0.78(0.19,0.81)		0.89(0.26,3.02)	
Residence				
Rural	1		1	
Urban	12.81(7.97,20.6)		4.14(2.28,7.49)	
Marital status-female				
Never married	1			
Married, only wife	2.29(1.13,4.63)			
Married, other wives	2.09(0.60,7.33)			
separated	9.76(4.73,20.15)			
Marital status-male				
Never married			1	1
Married or living together			4.29(2.22,8.28)	1.68(0.93,3.01)
Separated			9.24(5.22,16.37)	2.56(1.47,4.45)
STI				
No	1		1	1
Yes	1.48(1.15,1.91)		0.01(0.01,0.04)	0.01(0.01,0.003)
Birth in the past 5 years				
Yes	1			
No	1.15(0.75,1.75)			
Total number of children who had died				
No child dead	1	1	1	
Never had children	0.74(0.35,1.59)	0.43(0.27,0.67)	0.36(0.11,1.22)	
One child died	1.41(0.70,2.81)	1.55(0.71,3.39)	0.22(0.08,0.69)	
Two or more child died	1.07(0.42,2.68)	2.62(0.86,7.95)	0.23(0.11,1.22)	
Education level				
No education	1	1	1	1
Primary	2.42(1.43,4.11)	5.94(2.09,6.82)	0.67(0.28,1.61)	1.07(0.4,2.88)
Secondary	6.21(3.76,10.25)	6.99(1.84,6.55)	2.75(1.46,5.20)	3.33(1.57,7.09)
Higher	1.03(0.68,1.54)	1.27(0.33,4.81)	2.15(1.10,4.19)	1.03(0.46,2.31)
Wealth index				
Poorest	1	1	1	
Poorer	2.99(0.66,3.55)		0.46(0.09,2.48)	
Middle	1.29(0.25,6.63)	0.80(0.16,4.09)	1.38(0.44,4.35)	
Richer	0.64(0.16,2.53)	0.01(0.002,0.06)	0.59(0.17,2.06)	
Richest	19.27(4.78,7.32)	2.81(0.60,3.08)	3.63(1.42,9.32)	
Religion				
Other	1	1	1	1
Protestant	1.21(0.48,3.07)	1.36(0.45,4.12)	2.27(0.36,13.93)	2.20(0.34,4.37)
Orthodox	3.75(2.02,6.98)	2.32(1.17,4.60)	9.74(2.16,14.82)	7.79(1.75,4.72)
Number of the partners in the past years				
Never had sex	1	1		
One partner	12.57(5.15,30.70)	3.02(2.05,5.76)		
More than one partner	44.84(1.12,3.97)	6.21(3.51,11.0)		
Age at first sex				
Not has sex			1	1

Continue on Next Page...

Table3—Continued

	Female		Male	
Variable	OR (95% CI)	AOR(95% CI)	OR (95% CI)	AOR(95% CI)
Less than 14 years			0.75(0.22,2.51)	0.64(0.21,1.94)
14-17 years			9.54(2.41,7.81)	2.29(0.85,6.13)
More than 18 years			7.72(2.30,5.93)	2.11(0.84,5.31)
Media Exposure				
Low	1		1	
Medium	1.68(0.74,3.80)		1.07(0.47,2.43)	
High	6.58(3.83,11.29)		2.83(1.34,5.99)	
Stigma				
High	1	1	1	1
Medium	5.03(2.91,8.72)	2.76(1.51,5.06)	0.71(0.38,1.34)	0.81(0.41,1.58)
Low	9.71(5.91,15.96)	3.74(1.87,7.48)	0.14(0.05,0.36)	0.22(0.08,0.60)
Decision making ability				
Independent	1	1		
Consults	1.08(0.63,1.87)	1.98(1.05,3.73)		

3.4 Discussion and Conclusions

In an effort to improve the understanding of the HIV/AIDS epidemic in Sub-Saharan Africa, in general, and in Ethiopia, in particular, this study has reported on national HIV seroprevalence in Ethiopia and assessed key variables for their association with HIV serostatus at the individual level. A demographic and health survey is often conducted to obtained information about the prevalence of certain diseases and unhealthy behaviours, as well as people's related exposure to potential risk factors. Binary outcomes that measure the presence or absence of certain medical conditions are common in survey research. It can be a very important task for health researchers to examine the relationship between multiple health conditions and predictors, including, in complex surveys, behaviours as well as demographic and socio-economic measures.

The objective of this study was to identify key demographic, socio-economic, socio-cultural, behavioural and proximate determinants as risk factors for HIV prevalence among men and women in Ethiopia. We carried out an analysis for individuals within the age group (15-49) for female and (15-59) for male. Our variables were categorized as key demographic, socio-economic, socio-cultural, behavioural and proximate determinants, and were analysed by using survey logistic regression model contained in SAS 9.2.

The key demographic factor in this analysis was region: both men and women from Addis Ababa city had double the risk for infection with HIV as compared with respondents from Affar region [83]. The findings showed that women and men aged 35-39 years are at a significantly higher risk of being HIV-positive compared with the reference category aging from 15-19 years. Both men and women in rural are found at lower risk [86]. Urban residence was associated with much higher HIV prevalence rate, and there were large differentials by geographical region. Huge regional differences in HIV prevalence have also been observed in Sub-Saharan Africa [86, 111].

Wealth was positively related to risk for HIV for both women and men, yet education did not show the same relationship to the outcome variable. Because economic status and education level typically correlated for many outcomes, this finding is intriguing and would benefit from further exploration. Individuals who think they have only a small risk of acquiring HIV are, in fact, at highest risk of being HIV-positive, compared with those who think they have no risk.

The study found that the separated women and men had higher likelihood of HIV infection [111]. This was expected as some of the separated may be infected by their formal spouses as a result of HIV infection. Another strong correlate of HIV infection in men was male circumcision. Male circumcision had a strong protective effect on the likelihood of being HIV-positive. This is consistent with recent clinical trials in South Africa, Kenya, and Uganda that showed male circumcision can significantly reduce the risk of HIV infection [8, 154]. Our findings add to the large body of research indicating that circumcision has a positive effect against HIV infection among men. However, because circumcision is very closely correlated with ethnicity and culture, because not all studies find a protective effect of circumcision, and because there could be a confounding relationship between circumcision and ulcerative STIs, well controlled evidence from other national and disciplinary context should be considered before widespread planning and implementation of circumcision focused interventions.

The study shows that HIV is a multidimensional epidemic, with demographic, social, biological, and behavioural factors exerting influence on individual probability of becoming infected with HIV. Although all of these factors contribute to the risk profile for a given individual, ultimately, the findings suggest that differences in biological factors is more significant in assessing risk for HIV than differences in sexual behavior. The ways in which these interesting factors affect risk differ by sex, which implies that program interventions may require gender-specific approaches. Furthermore, the study provides a closer look at the spread of HIV infection in Ethiopia, specifically with regard to its different regions. The findings are useful in terms of identifying higher-prevalence and higher-risk populations, and for strengthening prevention, care and support, and treatment programmes.

Chapter 4

Reflect the Hierarchical aspect of the models under both frequentist and Bayesian approaches

4.1 Introduction

Multilevel models are statistical models of parameters that vary at more than one level. These models can be perceived as a sequence of linear or nonlinear models. In standard multiple linear regression techniques, the researchers try to explain the variations in outcome variables in terms of one or more independent variables. In multilevel modeling the effect is assumed to be spread over two or more interdependent levels. Emphasizing the similarity between regression and multilevel modelling, [135] states that multilevel modelling as form of regression that allows the analyst to use both unites (primary) and groups of units (secondary) data in the same model. [3] discussed multilevel as the name for a range of techniques including both fixed and random effects developed from hierarchical approaches to analysis in different fields [69, 123].

In geographical studies, we can often have a hierarchy of effects regions containing cities, and countries containing regions. Failure to incorporate these hierarchical level effects will lead to incorrect inferences . However, while standard approaches to multilevel analysis are well established, there is none the less much scope for refinment and development of this extremely useful methodology.

The distinguishing feature of cluster data is that observations within a cluster tend to be more a like than observations from different clusters or, stated otherwise, they are correlated. Heterogeneity between clusters therefore introduces an additional source of variation which complicates analysis. When this extra variation can not be explained by measured covariates, we require statistical analysis methods which explicitly acknowledge clustering in the data.

The aim of this chapter is to introduce and interrogates key statistical properties of the multilevel models and then discusses how parameter estimation and inference was carried out.

4.2 Multilevel model description

In this section, we briefly introduce multilevel model for normally distributed response variables. According to the (EDHS 2005) we consider a three hierarchical multilevel model. From the data we assign cluster to level 3, household to level 2 and individual to level 1.

Let $y_{hij}, i = 1, 2, \ldots, n_h; j = 1, \ldots, n_{hi}; h = 1, \ldots, H$ denotes the dichotomous random variable taking, value 1 if HIV diagnosis for j^{th} individual in the i^{th} household nested within h^{th} cluster is positive, and 0 otherwise. We assume $y_{hij} \sim Binomial(1, \pi_{hij})$, the model can be written model as follows

$$\text{logit}(y_{hij}) = \boldsymbol{x}'_{hij}\boldsymbol{\beta} + \boldsymbol{z}'_{3,hij}\boldsymbol{u}_h^3 + \boldsymbol{z}'_{2,hij}\boldsymbol{u}_{ih}^{(2)} + \boldsymbol{\epsilon}_{hij}, \tag{4.1}$$

where $\boldsymbol{x}_{hij}$ is the vector of the covariates having fixed effects $\boldsymbol{\beta}$, $\boldsymbol{z}'_{3,hij}$ is the vector of the covariates having random effect $\boldsymbol{u}_h^{(3)}$ at the cluster level, $\boldsymbol{z}'_{2,hij}$ is a vector of covariates having random effects $\boldsymbol{u}_{ih}^{(2)}$ at household level and $\boldsymbol{\epsilon}_{hij}$ is the vector of the error term. The random terms are assumed to be mutually independent and distributed as follows:

$$\boldsymbol{u}_h^{(3)} \sim N(\mathbf{0}, \boldsymbol{\Omega}^3)$$

$$\boldsymbol{u}_{ih}^{(2)} \sim N(\mathbf{0}, \boldsymbol{\Omega}^2)$$

$$\boldsymbol{\epsilon}_{hij} \sim N(\mathbf{0}, \sigma_e^2)$$

Model (4.1) can be written in matrix form as special case of the general linear mixed model as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}, \tag{4.2}$$

where $\boldsymbol{y}$ is the vector of the responses, $\boldsymbol{X}$ is the design matrix for the fixed effects, $\boldsymbol{u}$ is a vector of the random effects obtained by stacking $\boldsymbol{u}_{ih}^{(2)}$ on top of the cluster effects $\boldsymbol{u}_{h}^{(3)}$, $\boldsymbol{\beta}$ is a vector of the fixed effects, $\boldsymbol{Z}$ is the design matrix for the random effects, and $\boldsymbol{\epsilon}$ are error terms obtained by stacking $\boldsymbol{\epsilon}_{hij}$

4.3 Estimation Methods

There is no single agreed upon way to estimate the parameters in the multilevel model. Many methods of estimation can be applied, including Maximum Likelihood (ML), Restricted Maximum Likelihood (REML), and Bayesian estimation [93, 123]. These methods of estimation can be done by using many different algorithms. For instance, ML estimation can be carried out using the Expectation Maximisation (EM) algorithm, the Newton Raphson algorithm, Fisher scoring algorithm and Iterative Generalised Least Squares(IGLS), Bayesian estimation will be carried out using Gibbs sampling.

4.3.1 Maximum Likelihood Estimation(ML)

The idea behind ML estimation is to select parameter estimates that maximise the likelihood of the data. We consider how likely it is that we would have obtained the data for each of many different values for the fixed and variance parameters, and then pick the values for which the likelihood is the greatest. This involves an iterative algorithm that steps through possible values until the likelihood reaches its maximum. The objective of computational statisticians is to develop an algorithm that converges fairly fast across a wide range of applications. For illustration simplicity, we consider a two level model

$$g(\pi_{ij}) = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{u}_j, \tag{4.3}$$

with $\boldsymbol{\pi}_{ij} = P[\boldsymbol{y}_{ij} = 1|u_j](j = 1, \ldots; i = 1, \ldots, n_j)$ and $\boldsymbol{u}_j \sim N(0, \boldsymbol{\Omega}_u)$, $\boldsymbol{z}_{ij}'$ is a vector of covariates having random effect $\boldsymbol{u}_i$. Because of the local independence assumption, the conditional likelihood of the j takes the binomial form; that is, its contribution to the log marginal likelihood, obtained by integrating over the random effect, can be written as

$$\ell_i(\boldsymbol{\beta}, \boldsymbol{\Omega}_u) = log \int \prod_{i=1}^{n_j} (\boldsymbol{\pi}_{ij}^{y_{ij}} (1 - \boldsymbol{\pi}_{ij})^{1-y_{ij}} \phi(\boldsymbol{u}_j; \boldsymbol{\Omega}_u) d\boldsymbol{u}_j, \tag{4.4}$$

where $\phi(\boldsymbol{u}_j; \boldsymbol{\Omega}_u)$ is the normal density function $N(0, \Omega_u)$. The log marginal likelihood

$$\ell_i(\boldsymbol{\beta}, \boldsymbol{\Omega}_u) = \sum_{j=1}^{n} \ell_j(\boldsymbol{\beta}, \boldsymbol{\Omega}_u), \tag{4.5}$$

can be the maximized, using any standard optimization routine, to obtain estimates of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$. Unfortunately, equation (4.4) is intractable and necessitates the use of numerical integration techniques. As a result some algorithms are proposed to solve the problem.

ML estimation is available through a different algorithm and software packages. It can be carried out using the EM algorithm [45], or the Newton Raphson algorithm [100] implemented in proc NLMIXED (SAS institute 2000), and the Fisher scoring algorithm [103]. By combining Fisher scoring with the EM algorithm, relatively fast convergence can be obtained while avoiding inadmissible variance and covariance estimates. The estimates of the fixed effects and variance parameters also tend to be asymptotically efficient, which implies that when the sample size is large, the ML estimates will show minimum variance from sample to sample [123].

4.3.2 Restricted Maximum Likelihood Estimation(REML)

In REML estimation, maximum likelihood estimates are obtained for the variance parameters. These values are then used to obtain generalised least squares estimates of the fixed effects. REML estimates of variance parameters may be considered preferable to ML estimates because REML takes into account uncertainty in the fixed effects when the variance parameters are estimated. Since the uncertainty in the fixed effects is more pronounced with smaller sample sizes, we may suspect the difference in these methods would tend to be greater when sample sizes are smaller. Different empirical studies have found differences between ML and REML estimates under a variety of conditions [93], but these studies do not lead to uniform recommendations of one method over the other.

REML and ML estimates can be obtained from a variety of software packages such as HLM,SAS PROC MIXED and Mln and through different algorithms such as EM, Newoton Raphson, Fisher scoring, and restricted IGLS, and have been shown to have desirable properties under many conditions. Furthermore, under general conditions, the fixed effects are unbiased [88, 89], the estimates of the variance components are asymptotically unbiased [123], and as the sample size increases, the sampling distribution of the estimates become approximately normal [123]. Consequently, both ML and and REML are recommended for large sample sizes. When the sample size is small ML and REML estimates are questionable which may lead researchers to consider alternative methods.

4.3.3 Penalized Quasi Likelihood(PQL)

The PQL estimation procedure is described here for the two level regression models and we can extend to three level models which is suitable to our data. Consider a level-2 outcome Y_{ij} taking on a value of 1 if the individual is HIV positive and 0 otherwise with conditional probability p_{ij}. Then the logit model or the generalized liner model is

$$\ln\left[\frac{p_{ij}}{1-p_{ij}}\right] = \eta_{ij} = \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{u}_j,$$

for level-1 unit i nested within level-2 unit j. At level 1, we assume Y_{ij} conditionally distributed as Bernoulli, while the random effects vector $\boldsymbol{u}_j$ is distributed as $N(0, \sigma_u^2)$ across the level-2 units. Let us consider the variance σ_u^2 as T throughout the PQL estimation procedure.

The PQL approach can be derived as a nonlinear regression model. In binary outcomes with logit link, we start with level-1 model

$$Y_{ij} = p_{ij} + e_{ij}, \tag{4.6}$$

where $E(e_{ij}) = 0$ and $Var(e_{ij}) = p_{ij}(1-p_{ij})$. This is a nonlinear model which we can linearize by means of the first order Taylor series expansion. At this iteration s, we have

$$p_{ij} \approx p_{ij}^{(s)} + \frac{dp_{ij}}{d\eta_{ij}}(\eta_{ij} - \eta_{ij}^{(s)}), \tag{4.7}$$

and evaluate the derivative

$$\frac{dp_{ij}}{d\eta_{ij}} = p_{ij}(1-p_{ij}) = \omega_{ij}, \tag{4.8}$$

at $p_{ij}^{(s)}$ in equation (4.7) yields

$$Y_{ij} = p_{ij}^{(s)} + \omega_{ij}^{(s)}(\eta_{ij} - \eta_{ij}^{(s)}) + e_{ij}, \tag{4.9}$$

Algebraically rearranging this equation so that all known quantities are on left hand side of the equation produces

$$\frac{Y_{ij} - p_{ij}^{(s)}}{\omega_{ij}^{(s)}} + \eta_{ij}^{(s)} = \eta_{ij} + \frac{e_{ij}}{\omega_{ij}^{(s)}}, \tag{4.10}$$

This equation has the same form of the familiar two level hierarchical linear model

$$Y_{ij}^{*}(s) = \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{u}_j + \boldsymbol{\epsilon}_{ij}, \tag{4.11}$$

which gives a straightforward updating scheme. This is known as penalized quasi likelihood because it is obtained by optimizing a quasi likelihood with a penalty term on the random effects. Here

$$Y_{ij}^{*}(s) = \frac{(Y_{ij} - p_{ij}^{(s)})}{\omega_{ij}^{(s)} + \eta_{ij}^{(s)}}, \tag{4.12}$$

$$\epsilon_{ij} = \frac{e_{ij}}{\omega_{ij}^{(s)}} \sim N(0, \omega_{ij}^{(s)-1}), \tag{4.13}$$

and $\boldsymbol{u}_j \sim N(0,T)$.
The estimate of $\eta_{ij}^{(s)}$ can be written as below

$$\eta_{ij}^{(s)} = \boldsymbol{X}'_{ij}\hat{\boldsymbol{\beta}}^{(s)} + \boldsymbol{z}'_{ij}\boldsymbol{u}_j^{*}(s), \tag{4.14}$$

Where $\boldsymbol{u}_j^{*}(s)$ is the approximate posterior mode,

$$\boldsymbol{u}_j^{*(s)} = (\boldsymbol{z}'_j\boldsymbol{W}_j^{(s)}\boldsymbol{z}'_j + T^{(s-1)})^{-1}\boldsymbol{z}'_j\boldsymbol{W}_j^{(s)}(\boldsymbol{Y}_j^{*(s)} - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}^{(s)}), \tag{4.15}$$

where

$$\boldsymbol{W}_j^{(s)} = diag\{\omega_{ij}^{(s)}, ..., \omega_{nij}^{(s)}\}.$$

4.4 Bayesian Multilevel Model

4.4.1 Introduction

The Bayesian approach is another technique used to obtain parameter and precision estimates. In ferquentist approach the parameters are considered as unknown constants, but in the Bayesian approach all the parameters are considered as random variables and the data is used to update prior belief regarding these parameters. In the Bayesian approach the whole posterior distribution of the parameters of interest is obtained but it usually summarised by the posterior mean, median and the standard deviation. The Bayesian analysis, starts with an initial probability distribution for parameter $\boldsymbol{\theta}$ given by $\Pi(\boldsymbol{\theta})$, known as the prior. With observed data $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ we select a statistical model with density $p(\boldsymbol{y}|\boldsymbol{\theta})$ that describes the distribution of the data, given $\boldsymbol{\theta}$. Then we combine the prior beliefs about θ with the data y to create the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$. The posterior distribution reflects our updated beliefs, and may be summarised in the form of posterior statistics such as the posterior mean and standard deviation, and through the creation of credible intervals. Such intervals have a natural interpretation.

4.4.2 Bayesian Updating

Bayesian inference differs from classical inference in treating parameters as random variables and using the data to update prior knowledge about the parameters and functionals of those parameters. We also need model predications and these are provided as part of the updating process. Prior knowledge about parameters and updated (or posterior) knowledge about them, as well as implications for functionals and predication, are expressed in terms of densities. One of the benefits of modern Monte Carlo Markov Chain (MCMC) sampling methods [34, 58, 65, 133, 140] is the ease with which full marginal densities of parameters may be obtained.

The new Bayesian sampling based estimation techniques obtain samples from the posterior density, either of parameters themselves, or functionals of parameters. They improve considerably on multiple integration or analytical approximation methods that are not feasible with large numbers of parameters. Nevertheless many issues remain in the application of sampling based techniques, such as obtaining convergence, and the choice of efficient sampling methods. There are also general problems in Bayesian methods such as the choice of priors. The basis for Bayesian inference may be derived from simple probability theory. Thus the conditional probability theorem for events A and B is that

$$Pr(A|B) = \frac{Pr(A,B)}{Pr(B)} = Pr(B|A)\frac{Pr(A)}{Pr(B)},$$

By replacing B by observations $\boldsymbol{y}$, and A by a parameter set $\boldsymbol{\theta}$ and probabilities be densities the equation results in the relation

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta},\boldsymbol{y})}{p(\boldsymbol{y})} = p(\boldsymbol{y}|\boldsymbol{\theta})\frac{p(\boldsymbol{\theta})}{p(\boldsymbol{y})}, \tag{4.16}$$

where $p(\boldsymbol{y}|\boldsymbol{\theta})$ is the likelihood of $\boldsymbol{y}$ under a model and $p(\boldsymbol{\theta})$ is prior density, or the density of $\boldsymbol{\theta}$ before $\boldsymbol{y}$ is observed. This density expresses accumulated knowledge about $\boldsymbol{\theta}$, or, viewed another way, the degree of uncertainty about $\boldsymbol{\theta}$. It may also include working model assumptions, such as, one model might assume uncorrelated errors over time or space and another model might assume correlated errors.

Classical analysis via maximum likelihood focuses on the likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$ without introducing a prior, whereas a full Bayesian analysis updates the prior information about $\boldsymbol{\theta}$ with information contained in the data. The denominator $p(\boldsymbol{y}) = [\int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}]$ in (4.16) defines the marginal likelihood or prior predictive density of the data and may be set to be an unknown constant c. So posterior inferences about $\boldsymbol{\theta}$ under (4.16) can be written as

$$p(\boldsymbol{\theta},\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta})\frac{p(\boldsymbol{\theta})}{c},$$

or

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

4.5 Monte Carlo Markov Chain (MCMC) Techniques

The basis of modern Bayesian inference regarding $p(\boldsymbol{\theta}|\boldsymbol{y})$ is the use of iterative MCMC methods that involved repeated sampling from the posterior distribution, using long, possibly multiple, chains of parameters samples. One is then interested in posterior summaries of parameters or functionals from the MCMC out put in the form of expectations densities or probabilities. These summaries typically include posterior means and variances of the parameters themselves, or of functions $\Delta = \Delta(\boldsymbol{\theta})$ of the parameters, which analytically are

$$E(\boldsymbol{\theta}_k|\boldsymbol{y}) = \int \boldsymbol{\theta}_k p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}, \tag{4.17}$$

$$Var(\boldsymbol{\theta}_k|\boldsymbol{y}) = \int \boldsymbol{\theta}_k^2 p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} - [E(\boldsymbol{\theta}_k|\boldsymbol{y})]^2 = E(\boldsymbol{\theta}_k^2|\boldsymbol{y}) - [E(\boldsymbol{\theta}_k|\boldsymbol{y})]^2, \tag{4.18}$$

$$E[\Delta(\boldsymbol{\theta}/|\boldsymbol{y}] = \int \Delta(\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$$

$$\begin{aligned} Var[\Delta(\boldsymbol{\theta}|\boldsymbol{y}] = & \int \Delta^2 p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} - [E(\Delta|\boldsymbol{y})]^2 \\ & = E(\Delta^2|\boldsymbol{y}) - [E(\Delta|\boldsymbol{y})]^2, \end{aligned}$$

Often the major interest is in the marginal densities of the parameters themselves. Let the model dimension be d, so that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ then the marginal density of the jth parameter θ_j is obtained by integrating out all the other parameters

$$p(\theta_j|\boldsymbol{y}) = \int p(\boldsymbol{\theta}|\boldsymbol{y})d\theta_1 d\theta_2 \ldots d\theta_{j-1}d\theta_{j+1}\ldots d\theta_d.$$

Posterior probability estimates from an MCMC run might relate to the probability that $\boldsymbol{\theta}_k$ exceeds a threshold b, and provide an estimate of integral

$$Pr(\boldsymbol{\theta}_k > b|y) = \int \ldots \int_b^\infty \ldots \int p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}. \tag{4.19}$$

The predictive density for new or replicate data useful in model checking and comparison is

$$p(\boldsymbol{y}_{new}, \boldsymbol{y}) = \int p(\boldsymbol{y}_{new}, \boldsymbol{\theta}|y)d\boldsymbol{\theta} = \int p(\boldsymbol{y}_{new}|y, \boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta},$$

posterior probabilities might relate to the probability that $\boldsymbol{\theta}_j$ exceeds a threshold b, and involve integrals of the form

$$Pr(\boldsymbol{\theta}_j > b|\boldsymbol{y}) = \int (\boldsymbol{\theta}_j|\boldsymbol{y})d\boldsymbol{\theta}_j, \tag{4.20}$$

Such expectations, densities or probabilities may be obtained analytically for conjugate analysis, such as a binomial likelihood where the probability has a beta prior. Results can be obtained under asymptotic approximations [13], similar to those used in classical statistics, or by analytic approximations based on expanding the relevant integral [90]. Such approximations tend to be less good for posteriors that are not approximately normal or where there is multimodality. An alternative strategy facilitated by contemporary computer technology is to use sampling based approximations based on the Monte Carlo principle. One such sampling method is importance sampling [63, 106], and other precursors of modern Bayesian sampling include data augmentation for Bayes inference in missing data problems [138].

4.5.1 Monete Carlo Markov Chains (MCMC) Sampling Algorithm

The Metropolis-Hastings (M-H) algorithm is the baseline for MCMC sampling schemes and is based on a binary transition kernel. Following [79], the chain is updated from $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^*$ with probability

$$\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)}) = min\left(1, \frac{P(\boldsymbol{\theta}^*|\boldsymbol{y})f(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^*)}{P(\boldsymbol{\theta}^{(t)}|\boldsymbol{y})f(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})}\right), \tag{4.21}$$

with transition kernel

$$\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})f(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)}),$$

where f is known as a proposal or jumping density [34]. If the proposed update is rejected the next state is the same as the current state. The algorithm works most successfully when the proposal density matches, at least approximately, the shape of the target density $p(\boldsymbol{\theta}|\boldsymbol{y})$. The rate at which a proposal generated by f is accepted depends on how close $\boldsymbol{\theta}^*$ is to $\boldsymbol{\theta}^{(t)}$, and this depends on the variance σ^2 assumed in the proposal density. For a normal proposal density a higher acceptance rate follows from reducing σ^2, but with the risk that the posterior density will take longer to explore. Performance also tends to be improved if parameters are transformed to take the full range of positive and negative valued $(-\infty, \infty)$ so lessening the occurrence of skewed parameter densities.

If the proposal density is a symmetric density in with

$$f(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)}) = f(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^*),$$

then the Hastings algorithm reduces to an algorithm used by [60] for indirect simulation of energy distributions, whereby

$$\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)}) = min\left[1, \frac{p(\boldsymbol{\theta}^*|\boldsymbol{y})}{p(\boldsymbol{\theta}^{(t)}|\boldsymbol{y})}\right]. \tag{4.22}$$

A particular symmetric density in which

$$f(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)}) = f(|\ \boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*\ |),$$

leads to the random walk Metropolis [58]. While it is possible for the proposal density to relate to the entire parameter set, it is often computationally simpler to dived $\boldsymbol{\theta}$ into blocks or components, and use componentwise updating, where updating is used in a generic sense allowing for possible non-acceptance of proposed values. Thus let

$$\boldsymbol{\theta}_{[j]} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{(j-1)}, \boldsymbol{\theta}_{(j+1)}, \ldots, \boldsymbol{\theta}_d),$$

denote the parameter set omitting $\boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_j^{(t)}$ be the value of $\boldsymbol{\theta}_j$ after iteration t. At step j of iteration $t+1$ preceding $j-1$ parameters are already updated via the M-H Algorithm while $\boldsymbol{\theta}_{(j+1)}, \ldots, \boldsymbol{\theta}_d$ are still at their iteration t values [34]. Let the vector of partially updated parameters be denoted

$$\boldsymbol{\theta}_{[j]}^{(t,t+1)} = (\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \ldots, \boldsymbol{\theta}_{(j-1)}^{(t+1)}, \boldsymbol{\theta}_{(j+1)}^{(t)}, \ldots, \boldsymbol{\theta}_d^{(t)}).$$

The proposed value $\boldsymbol{\theta}_j^*$ for $\boldsymbol{\theta}_j^{(t+1)}$ is generated from the jth proposal density, denoted

$$f(\boldsymbol{\theta}_j^*|\boldsymbol{\theta}_j^{(t)}, \boldsymbol{\theta}_{[j]}^{(t,t+1)}).$$

Also governing the acceptance of a proposal are full conditional densities

$$p(\boldsymbol{\theta}_j^{(t)}|\boldsymbol{\theta}_{[j]}^{(t,t+1)}),$$

specifying the density of $\boldsymbol{\theta}_j$ conditional on other parameters $\boldsymbol{\theta}_{[j]}$. The candidate $\boldsymbol{\theta}_j^*$ is accepted with probability

$$\alpha(\boldsymbol{\theta}_j^{(t)}, \boldsymbol{\theta}_{[j]}^{(t,t+1)}, \boldsymbol{\theta}_j^*) = min\left[1, \frac{p(\boldsymbol{\theta}_j^*|\boldsymbol{\theta}_{[j]}^{(t,t+1)})f(\boldsymbol{\theta}_j^{(t)}|\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_{[j]}^{(t,t+1)})}{p(\boldsymbol{\theta}_j^{(t)}|\boldsymbol{\theta}_{[j]}^{(t,t+1)})f(\boldsymbol{\theta}_j^*|(\boldsymbol{\theta}_j^{(t)}, \boldsymbol{\theta}_{[j]}^{(t,t+1)})}\right].$$

The Gibbs sampler [27, 58, 66] is a special componentwise M-H algorithm whereby the proposal density for updating $\boldsymbol{\theta}_j$ is the full conditional $p(\boldsymbol{\theta}_j^*|\boldsymbol{\theta}_{[j]})$ so that proposals are accepted with probability 1. This sampler was originally developed by [61] for Bayesian image reconstruction, with the full potential for simulating marginal distribution by repeated draws recognized by [58]. The Gibbs sampler involves parameters-by-parameter updating which and when completed forms the transition from $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^{(t+1)}$:

1. $\boldsymbol{\theta}_1^{(t+1)} \sim f_1(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)}, ..., \boldsymbol{\theta}_d^{(t)};$
2. $\boldsymbol{\theta}_2^{(t+1)} \sim f_2(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_3^{(t)}, ..., \boldsymbol{\theta}_d^{(t)});$

.

.

.

.

d. $\boldsymbol{\theta}_d^{(t+1)} \sim f_d(\boldsymbol{\theta}_d|\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_3^{(t+1)}, ..., \boldsymbol{\theta}_{(d-1)}^{(t+1)}).$

Repeated sampling from Metropolis-Hasting H-M samplers such as the Gibbs sampler generates an autocorrelated sequence of numbers that, subject to regularity conditions, eventually the starting values

$$\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_1^{(0}, \boldsymbol{\theta}_2^{(0)}, ..., \boldsymbol{\theta}_d^{(0)}),$$

used to initialize the chain and converges to a stationary sampling distribution $p(\boldsymbol{\theta}|y)$.

The full conditional densities may be obtained from the joint density $p(\boldsymbol{\theta}, y) = p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})$ and in many cases reduce to standard densities from which sampling is straightforward. Full conditional densities can be obtained by abstracting out from the full model density (likelihood times prior) those elements including $\boldsymbol{\theta}_j$ are treating other components as constants [66]. Consider a conjugate model for Poisson count data $\boldsymbol{y}_i$ with mean $\boldsymbol{\mu}_i$ that are themselves gamma distributed; this a model appropriate for overdispersed count data with actual variability $Var(\boldsymbol{y})$ exceeding that under the Poisson model. Suppose $\mu_i \sim Gamma(\alpha, \beta)$, namely

$$f(\mu_i|\alpha, \beta) = \mu_i^{(\alpha-1)} \exp(-\beta\mu_i)\beta^\alpha/\Gamma(\alpha),$$

and further that $\alpha \sim E(a)$, and $\beta \sim Gamma(b, c)$, where a, b and c are preset constants; this prior structure is used by [62]. So the posterior density of

$$\boldsymbol{\theta} = (\mu_1, \mu_2, \ldots, \mu_n, \alpha, \beta),$$

given $\boldsymbol{y}$ is proportional to

$$e^{(-a\alpha)}\beta^{(b-1)}e^{(-c\beta)}\Pi_{i=1}^{n}\exp(-\mu)\mu_i^{(y_i)}\{\Pi_{i=1}^{n}\mu_i^{(\alpha-1)}\exp(-\beta\lambda_i)\}[\beta^{\alpha}|\Gamma(\alpha)]^{n},$$

where all constants are combined in the probability constant. It is apparent that the conditional densities of μ_i and $\boldsymbol{\beta}$ are $Gamma(y_i + \alpha, \beta + 1)$ and $Gamma(b + n\alpha, c + \sum \mu_i)$ respectively. The full conditional density of α is

$$f(\alpha|y, \beta, \mu) \propto \exp(-a\alpha)[\beta^{\alpha}|\Gamma(\alpha)]^{n}(\Pi_{i=1}^{n}\mu_i)^{(\alpha-1)}.$$

This density is non standard but log concave and can not be sampled directly. However, adaptive rejection sampling [67] may be used.

4.5.2 Monte Carlo Markov Chains(MCMC) Convergence

There are many unresolved questions around the assessment of convergence of MCMC sampling procedures [37]. It is generally accepted to be preferable to use two or more parallel chains with diverse starting values to ensure full convergence of the sample space of the parameters, and therefore diminish the chance that the sampling become trapped in a small part of the space [59, 60]. Single long runs adequate for straightforward problems, or as preliminaries to obtain inputs to multiple chains. Convergence for multiple chains may be assessed using Gelman-Rubin Scale Reduction Factors (SRF) that compare variation in the sampled parameter values within and between chains. Parameter samples from poorly identified models will show wide divergence in the sample paths between different chains and the variability of the sampled parameter values between chains will considerably exceeded the variability within any one chain. To measure the variability of samples $\theta_j^{(t)}$ within the jth chain $(j = 1, \ldots, J)$ define

$$V_j = \frac{\sum_{t=s+1}^{s+T}(\theta_j^{(t)} - \bar{\theta}_j)^2}{(T-1)},$$

over T iterations after an initial burn in of s iterations. Ideally the burn in period is a short initial set of samples where the effect of the initial parameter values tails off; during the burn in the parameter trace plots will show clear monotonic trends as they reach the region of posterior parameter space. Convergence is therefore assessed from iterations $s+1$ to $s+T$. Variability within chains V_w is then the average of the V_j. Between chain variance is measured by

$$V_B = \frac{T}{J-1}\sum_{j=1}^{J}(\bar{\theta}_j - \bar{\theta})^2,$$

where $\bar{\theta}$ is the average of the $\bar{\theta}_j$. The (SRF) compares a pooled estimator of $Var(\theta)$, given by $\frac{V_p=V_B}{T} + \frac{TV_w}{(T-1)}$, with the within sample estimate V_w. Specifically the SRF is $(\frac{V_p}{V_w})^{0.5}$ with values under 1.2 indicating convergence. Parameter samples obtained by MCMC methods are correlated, which means extra samples are needed to convey the same information. Additionally, as in any iterative estimation, there may be a delay in seeking the region of posterior density where the mode value is located. The extent of correlation, and the convergence towards the model region, will depend on a number of factors including the form of parametrisation, the sample size, the complexity of the model and the form of sampling.

More recently other proposed convergence statistics is that due to [23] and known as the Brooks-Gelman-Rubin(BGR) statistics are available. This is a ratio of parameter interval length, where for chain j the length of the $100(1-\alpha)\%$ interval for parameter $\boldsymbol{\theta}$ is obtained. That means the gap between 0.5α and $(1-0.5\alpha)$ points from T simulated values. This provides J within chain interval length, with mean I_u. For the pooled output of T_j samples , the same $100(1-\alpha)\%$ interval I_p is also obtained. Then the ratio $\frac{I_p}{I_u}$ should converge to one if there is convergent mixing over different chains.

Analysis of the sequence of samples from an MCMC chain amounts to an application of time series methods, with regards to problems such as assessing stationary in an autocorrelated sequence. Autocorrelation at lags 1, 2 and so on may be assessed from the full set of sampled values $\theta^{(t)}, \theta^{(t+1)}, \ldots$, or from sub samples k steps a part,

$$\theta^{(t)}, \theta^{(t+k)}, \theta^{(t+2k)}, \ldots.$$

If the chains are mixing satisfactory then the autocorrelation in the one step apart iterates of $\theta^{(t)}$ will fade to zero as the lag increases. Non vanishing autocorrelations at high lags means that less information about the posterior distribution is provided by each iterate and a higher sample size T is necessary to cover the parameter space. Slow convergence will show in trace plots that wander, and that exhibit short term trends rather than fluctuating rapidly around a stable mean.

Problems of convergence in MCMC sampling may reflect problems in model identifiability due to over fitting or redundant parameters. Running multiple chains often assists in diagnosing poor identifiability. This is illustrated most clearly when identifiability constraints are missing from a model, such as in discrete mixture models that are subject to label switching during MCMC updating [55]. One chain may have a different label to others so that obtaining a G-R statistic for some parameters is not sensible. Choice of diffuse priors tends to increase the chance of poorly identified models, especially in complex hierarchical models or small sample [57]. Elicitation of more informative priors or application of parameter constraints may assist identification and convergence. Correlation between parameters within the parameter set $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)$ increase the dependence between successive iterations. Reparameterisation to reduce the correlation such as centering predictor variables in regression usually improves convergence [57, 158].

An advantage of the Bayesian approach is that when the posterior distribution is simulated, the uncertainty of the parameter estimates is taken into account. So, the uncertainty in the parameter estimates for the fixed part is taken into account in the estimates for the random part. Moreover, simulating a large sample from the posterior distribution is useful because it provides not only point estimates of the unknown parameters, but also confidence intervals that do not rely on the assumption of normality for the posterior distribution. As a result, confidence intervals are also accurate for small samples [138]. However, the MCMC method does not deal very well with this extremely small data set. The number of MCMC iterations that are required are very large, and the autocorrelation is relatively high, especially in estimating the mean. This is indicate of data that contain very little information about the parameter estimates, which is not surprising given the small sample size.

As indicated earlier, all priors add some information to the data. As a result, Bayesian estimates are generally biased. When the sample size is reasonable, the amount of bias is small. This bias is acceptable, because Bayesian methods promise more precision, and better estimates of the sampling variance. Simulation research [24, 25]

4.6 Gibbs sampling

An MCMC is an iterative method and each iteration produces a set of parameter values, and after convergence the sequence of these, under general conditions, can be regarded as a serially correlated random draw from the joint posterior distribution of the parameters. MCMC algorithms enable us to successively sample from a convergent Marco Chain assuming the current values of the remaining components. The Gibbs sampling algorithms procedure can be outlined through a series of steps as follows:
Step 1

Sample a new set of fixed effects from the conditional posterior distribution

$$p(\boldsymbol{\beta}|\boldsymbol{y},\sigma_u^2,\sigma_e^2,u)\propto L(\boldsymbol{y};\boldsymbol{\beta},u,\sigma_e^2)p(\boldsymbol{\beta}),$$

a suitable diffuse prior is $p(\boldsymbol{\beta})\propto 1$

$$p(\boldsymbol{\beta}|\boldsymbol{y},\sigma_u^2,\sigma_e^2,u)\propto\left(\frac{1}{\sigma_e^2}\right)^{N/2}\prod_{i,j}\exp\left[-\frac{1}{2\sigma_e^2}(y_{ij}-u_j-(\boldsymbol{X}_{ij}\boldsymbol{\beta})^2\right],$$

so that we sample from

$$\hat{\boldsymbol{\beta}}\sim MVN(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{D}}),$$

$$\hat{\boldsymbol{\beta}}=\left[\sum_{i,j}\boldsymbol{X}'_{ij}\boldsymbol{X}_{ij}\right]^{-1}\left[\sum_{ij}\boldsymbol{X}'_{ij}(y_{ij}-u_j)\right],$$

where

$$\hat{\boldsymbol{D}}=\sigma_e^2\left[\sum_{i,j}\boldsymbol{X}'_{ij}\boldsymbol{X}_{ij}\right]^{-1}.$$

Note that this just the Ordinary Least Squares (OLS) applied to the adjusted response

$$\tilde{y}=\left[y_{ij}-(Zu)_{ij}\right],$$

and in the present case $(zu)_{ij}=u_j$. Other assumptions for the prior are possible. Suppose we assume multivariate normality, with $p(\boldsymbol{\beta})\sim MNV(\boldsymbol{0},\boldsymbol{V})$. This is a conjugate prior, since the posterior is also multivariate normally distributed. We sample from

$$\boldsymbol{\beta}\sim MNV(\hat{\boldsymbol{\beta}^*},\hat{\boldsymbol{D}^*}),$$

$$\hat{\boldsymbol{\beta}^*}=\left[\sum_{ij}\boldsymbol{X}'_{ij}\boldsymbol{X}_{ij}+\sigma_e^2\boldsymbol{V}^{-1}\right]^{-1}\left[\sum_{ij}\boldsymbol{X}'_{ij}\tilde{y_{ij}}\right],$$

$$\hat{\boldsymbol{D}}^{*} = \sigma_e^2 \left[\sum_{i,j} \boldsymbol{X}_{ij}^T \boldsymbol{X}_{ij} + \sigma_e^2 V^{-1} \right]^{-1} .$$

In the above and subsequent equations, the parameter terms on the right hand sides represent the current values and those on the left hand sides the new updated ones; for the practical issues we remove any indexing of the iterations where this leads to no ambiguity. The uniform prior distribution, extending over the whole real line, is improper in the sense that it is not a true probability distribution. Nevertheless, for independent normal priors the variance matrix, $\boldsymbol{V}$, for the fixed parameters is diagonal and if the elements are very large the posterior distribution for $\boldsymbol{\beta}$ is equivalent to assuming uniform prior distribution. The important requirement is that the posterior distribution exists, that is it.

We can sample each coefficient in turn but this would tend to produce more highly correlated chains compared to sampling them as a block where terms such as

$$\sum_{i,j} \boldsymbol{X}'_{ij} \boldsymbol{X}_{ij},$$

do not change and could be stored for use in each iteration.

Step 2

Sample a new set of residuals. Each residual $\boldsymbol{u}_j$ is assumed to have a prior distribution $u_j \sim N(0, \sigma_u^2)$ which leads to the following posterior distribution, where n_j is the number of level 1 units in the j^{th} level 2 unit

$$p(u_j|\boldsymbol{y}, \sigma_u^2, \sigma_e^2) \propto \left(\frac{1}{\sigma_e^2}\right)^{n_i/2} \prod_{i=1}^{n_j} \exp\left[-\frac{1}{\sigma_e^2}(y_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\beta} - u_j)^2\right] \times \left(\frac{1}{\sigma_u^2}\right)^{1/2} \exp\left[-\frac{1}{2\sigma_u^2} u_j^2\right],$$

so that we now sample from

$$u_j \sim N(\hat{u_j}, \hat{\boldsymbol{D}}),$$

$$\hat{u_j} = [n_j + \sigma_e^2 \sigma_u^{-2}]^{-1} \left[\sum_{i=1}^{n_j} (y_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\beta})\right],$$

$$\hat{\boldsymbol{D}} = \sigma_e^2 [n_j + \sigma_e^2 \sigma_u^{-2}]^{-1}.$$

When there are $p(>1)$ random coefficients with explanatory variable matrix $\boldsymbol{Z}$, this step is modified by sampling the residuals from a p variate normal distribution with

$$\hat{u_j} = \left[\sum_{i=1}^{n_j} \boldsymbol{Z}_{ij}^T \boldsymbol{Z}_{ij} + \sigma_e^2 \boldsymbol{\Omega}_u^{-1}\right]^{-1} \left[\sum_{i=1}^{n_j} \boldsymbol{Z}_{ij}^T (y_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\beta})\right],$$

$$\hat{\boldsymbol{D}} = \sigma_e^2 \left[\sum_{i=1}^{n_j} \boldsymbol{Z}_{ij}^T \boldsymbol{Z}_{ij} + \sigma_e^2 \boldsymbol{\Omega}_u^{-1}\right]^{-1}.$$

Step 3

Sample a new level 2 variance. An issue arises over the choice of appropriate diffuse prior distribution. We consider two probabilities, a uniform prior, as in the case of the fixed parameters, and an inverse gamma prior. The former choice often tend to produce positively biased estimates. The latter is commonly used. Rather than sampling the variance directly, it is simpler to sample its inverse σ_u^{-2}, refereed to as the precision which is assumed to have the Gamma prior distribution $p(\sigma_u^{-2}) \sim Gamma(\epsilon, \epsilon)$. We sample from

$$\sigma_u^{-2} \sim Gamma(a_u, b_u),$$

$$a_u = \frac{(j + 2\epsilon)}{2} \qquad b_u = (\epsilon + \sum_{j=1}^{j} \frac{u_j^2}{2}),$$

where j is the number of level 2 units.

For a uniform prior for σ_u^2 we sample from a Gamma distribution with

$$a_u = \frac{(j-2)}{2}, b_u = \sum_{j=1}^{J} \frac{u_j^2}{2}.$$

When there are $p(>1)$ random coefficients, we need to sample a new covariance matrix $\boldsymbol{\Omega}_u$; corresponding to the two choices for a single variance are a uniform prior over the space of covariance matrices or an inverse Wishart prior. We now sample from

$$\boldsymbol{\Omega}_u^{-1} \sim Wishart(\upsilon_u, S_u),$$

$$\upsilon_u = j + \upsilon_p, \quad S_u = \left(\sum_{j=1}^{j} \boldsymbol{u}_j^T \boldsymbol{u}_j + S_p\right),$$

where $\boldsymbol{u}_j$ is the row vector of residuals for the j^{th} level 2 unit and the prior

$$p(\boldsymbol{\Omega}_u^{-1}) \sim Wishart(\boldsymbol{\upsilon}_p, S_p).$$

A minimally informative or maximally diffuse choice for the prior would be to take υ_p equal to the order of $\boldsymbol{\Omega}_u$ and S_p equal to the value chosen to be close to the final estimate multiplied by υ_p; typically, we can use the maximum likelihood or approximate maximum likelihood estimate for S_p that is also used for the MCMC starting values.

Step 4

Sample a new level 1 variance. This is similar to the procedure for a single level 2 variance. We sample from

$$\sigma_e^2 \sim Gamma(a_e, b_e),$$

where

$$a_e = \frac{(N + 2\epsilon)}{2} \quad b_e = \frac{\sum_{i,j} e_{ij}^2}{2}.$$

Step 5

Compute the level 1 residuals.

4.7 Estimation method for Bayesian multilevel

Statistics is about uncertainty. We estimate unknown population parameters by statistics, calculated in a sample. In classical statistical inference, we express our uncertainty about how well an observed statistic estimates the unknown population parameter by examining its sampling distribution over an infinite number of possible samples. Since we generally have only one sample, the sampling distribution is based on a mathematical sampling model.

Lindley and Smith [99] introduced Bayesian estimation methods for multilevel models for the variance parameters when the fixed effects are estimated. Consequently, Bayesian estimation provides an appealing option for researchers working with small data set not like ML and REML estimates. This method can be carried out by using MCMC algorithms like the Gibbs sampler which is available in the R programme. All MCMC algorithms are iterative and at each iteration they are designed to yield a sample from the joint posterior distribution of the parameters of the model. These parameters will be regression coefficients, covariance matrices, residuals. After a specific number of iterations, we get a sample of values from the distribution of any parameter or set of parameters which can used to derive any desired distribution characteristics like the mode, mean and covariance matrix.

The Bayesian formulation for the multilevel models combine prior information about the fixed and random effects, with the likelihood based on the data. These parameters are considered as random variables described by the probability distributions, and the prior information for a parameter is incorporated into the model via a prior distribution.

Although Bayesian estimation is appealing in some cases, it also has some disadvantages. Prior distributions must be specified, but these specifications may be in conflict with some researchers desire not to let prior beliefs influence the results of their analysis [124]. The algorithm of Bayesian estimation is very computer intensive, making them impractical for large data sets.

4.8 Advantages of the Bayesian approach over a frequentist approach

In this section a brief list of advantages of Bayesian inference over frequentist inference approach is introduced.

(1) Bayesian inference allows informative priors so that prior knowledge or results of a previous model can be used to inform the current model.

(2) Bayesian inference can avoid troubles with model identification by manipulating prior distributions. Frequentist inference with any numerical approximation algorithm does not have prior distributions, and could be stuck in regions of flat density, which is considered a problem with model identification.

(3) the Bayesian approach treats the data as fixed, and the parameters as random because they are unknowns. The frequentist approach considers the unknown parameters as fixed, and the data as random, which is indicates that the estimation does not rely on the data it self, it also relies on hypothetical repeated

sampling in the future with similar data [128].

(4) the Bayesian approach estimates the full probability model. While the frequentist does not do so.

(5) the Bayesian approach estimates $p(hypothesis|data)$. In contrast, the frequentist approach estimates $p(data|hypothesis)$. Even though, the term(hypothesis testing) suggest it must be the hypothesis that is tested, given the data, not the opposite.

(6) the Bayesian approach has an axiomatic foundation [38] that is uncontested by the frequentist approach. Furthermore, the Bayesian approach is coherent to a frequentist, but a frequentist approach is incoherent to a Bayesian.

(7) the Bayesian approach has a strong decision theoretic foundation [14, 125]. The aims of the statistical inference are to facilitate decision making [125]. The most optimal decision is the Bayesian decision.

(8) the Bayesian approach includes uncertainty in the probability model, producing more realistic prediction. The frequentist approach does not includes the uncertainty of the parameters estimates, so it produces less realistic predications the Bayesian approach.

(9) the Bayesian approach is consistent with many of philosophies of science regarding to the epistemology, where knowledge can not be built entirely through experimentation, but requires prior knowledge [125].

(10) the Bayesian approach has the ability to compare different models with different methods using Deviance Informatin Criteria (DIC) including hierarchical models, but the the frequentist approach cannot.

(11) the Bayesian approach obeys the likelihood principle, Whereas the frequentist approach including MLE and the the General Method of Moments (GMM) or

the Generalised Estimating Equations (GEE), violates the likelihood principle [128].

(12) the Bayesian approach is protected against over fitting by integrating over model parameters whereas over fitting occurs in frequentist approach and is a serious problem in it.

(13) the Bayesian approach uses observed data only. But the frequentist approach uses both observed data and future data that are unobserved and hypothetical.

(14) the Bayesian approach uses the prior distribution which means that, more information is used and the 95% confidence interval of the posterior will be narrower than the 95% confidence intervals of the frequentist approach.

(15) the Bayesian approach uses probability interval to state the probability that the θ is between two points. The frequentist approach uses confidence intervals, which should be interpreted with a probability of zero or one that θ is in the region, so the frequentist never knows whether it is or is not, but can only say that if 100 repeated samples were drawn in the future, that it would be in the region for 95 samples.

(16) the Bayesian inference via using MCMC has a theoretical guarantee than the MCMC algorithm will converge if we run long enough but in frequentist inference we do not have guarantees for the convergence of the MLE.

(17) the Bayesian inference via MCMC or PMC is unbiased with respect to the sample size whether it is small or large. The frequentist becomes more biased when the sample size is small.

(18) the Bayesian inference via MCMC or PMC uses exact estimation with respect to sample size. When the frequentist uses approximate estimation it depends on asymptotic theory.

(19) the Bayesian inference with correlated predictors sometimes allows the hyperparameters to be distributed multivariate normal, therefore, including such correlation into the MCMC or PMC algorithm to enhance estimation. The frequentist inference does not use prior distribution, therefore, the confidence intervals will be wider and less certain with correlated predictors.

(20) The Bayesian inference with perfect priors is immune to singularities with matrix inversions, unlike frequentist inference.

4.8.1 Advantages of the frequentist approach over the Bayesian approach(disadvantages)

In this subsection the advantages of the frequentist approach over the Bayesian approach are listed

(1) frequentist models are good in handling large data sets, while Bayesian models (MCMC,PMC) have a problem handling large data set.

(2) frequentist models are always much easier to prepare, because many things do not need to be specified, such as prior distribution, initial values for numerical approximation, and the likelihood function. The Bayesian approach specify the prior distribution, initial values and so on. The frequentist approach is also well developed so it easy for anyone to apply it.

(3) frequentist models have a much short time to run compared with the Bayesian method. Simple frequentist models may be run in minutes, while the same model in Bayesian approach may take a week to run.

4.9 Conclusion

This chapter was introduced the multilevel level models and Bayesian approach for multilevel models. For both models we introduced the models description and how the parameters estimation was carried out. The multilevel models have strengths and weaknesses. One of the strength is that, Multilevel models have a very good ability to separately estimate the predictive effects of an individual predictor and its group-level mean, which is sometimes interpreted as the direct and contextual effects of the predictor. Multilevel models also take the sampling designs into account. Multilevel models provide a useful framework for thinking about problems with samples which have hierarchical structure. One advantage of the multilevel modeling approach is that it can deal with data in which the times of the measurements vary from subject to subject. The multilevel models are very weak to handle the missing data and also it has design issue problems. The biggest weakness of the Bayesian approach, is the time spent to run the model when using the MCMC methods of estimation, especially when the data sets are huge or the variables of the model are too many. It also takes a long time to draw a final decision from Bayesian models which means that models should be run with different priors on the model parameters. Furthermore, in order to make sure that the choice of the prior distribution does not affect the results, sensitivity analysis should have be performed

Chapter 5

Multilevel modeling of HIV prevalence data

This chapter is a stand alone research article of application of the Multilevel modeling to HIV prevalence data in Ethiopia. The aim of this paper was to model the HIV prevalence data via two paradigms of the multilevel modeling namely, the frequentist and the Bayesian approaches and compared the finding. The analysis was done separately for men and women because the biological and social circumstances association with transmission of HIV differ by sex.

Hierarchical multilevel models for analysing HIV/AIDS data in Ethiopia[1]

Mohammed O. M. Mohammed[2], T. N. O. Achia and T.Zewotir

School of Mathematics, statistics and Computer Science, University of KwaZulu-Natal, Private Bag X01 Scottsville 3209, Pietermaritzburg, South Africa

Abstract.

Ethiopia has an estimated 2 million people living with HIV and the third highest number of infections in Africa, according to UNAIDS. With a population of 83 million people and per capita income of less than US$ 100 annually, it is also one of the world's poorest countries.

The study used the Ethiopia Demographic Health Survey data (EDHS 2005) HIV data. Therefore, the objective of the study is to fit a model to data. We applied two approaches of hierarchical multilevel models, namely the frequentist multilevel and the Bayesian approach and compared the results. For the frequentist multilevel model we compute the estimates and the 95% confidence interval for each parameters and the deviance. We compute the estimates and the 2.5% quantile, the median and the 97.5% for each parameters and the deviance for the Bayesian approach. The analysis has been done separately for both females and the males. We compared the results of the two approaches, and the results are similar.

The findings indicate that, uncircumcised men were 4 times more likely to be HIV-positive than those who were not. Men aged 35-39 years had the highest risk of being HIV-positive, whereas the ages of highest risk for women were 25-29 years. Both wealthiest men and women have been found at higher risk for acquiring the disease.

This study reveals that HIV is a multidimensional social epidemic, with demographic, social, biological and behavioral factors all exerting influence on individual probability of becoming infected with HIV. Although all of these contribute to the risk profile for a given individual, the findings suggest that differences in biological factors such as circumcision and sexually transmitted infections may be more important in assessing risk for HIV than sexual behavior.

Keywords: HIV prevalence, Multilevel Models, Bayesian Hierarchical models, EDHS 2005, Markov Monte Carlo Chain(MCMC)

[1] Submitted to the Journal of Public Health

[2] Corresponding Author. 209509598@stu.ukzn.ac.za

5.1 Introduction

The emergence of the HIV epidemic is one of the biggest public health challenges the world has ever seen in recent history. In the last three decades HIV has spread rapidly and affected all sectors of society- young people and adults, men and women, and the rich and the poor. Sub-Saharan Africa is at the epicentre of the epidemic and continues to carry the full brunt of its health and socioeconomic impact. Ethiopia is among the countries most affected by the HIV epidemic. With an estimated adult prevalence of 1.5%, it has a large number of people living with HIV (approximately 800,000); and about 1 million AIDS orphans [82].

Findings from the most recent Antenatal Clinic (ANC) sentinel surveillance data in Ethiopia show a declining prevalence of infection rates among women aged 15-24 years attending ANC, from 5.6% in 2005, to 3.5% in 2007, to 2.6% in 2009. This trend was evident both in urban and rural areas. In urban centers the prevalence has halved, declining from 11.5 % in 2003 to 5.5% in 2009. The declining trend is even steeper in rural areas where prevalence declined from 4% in 2003, to 1.4% in 2009. Generally, 94% of the sentinel sites showed absolute decrease of which half of these were statistically significant [1].

Social research regularly involves problems that investigate the relationship between individuals and society. The general concept is that individuals interact with the social contexts to which they belong, that individual persons are influenced by the social groups or contexts to which they belong, and that these groups are in turn influenced by the individuals who make up the groups. Individuals and the social groups are conceptualised as a hierarchical system of individuals nested within groups, with individuals and groups defined at separate levels of this hierarchical system. Naturally, such systems can be observed at different hierarchical levels, and variables may be defined at each level. This leads to research into the relationships between variables characterising individuals and variables characterizing groups, a type of research generally referred to as multilevel research.

Multilevel models are becoming increasingly popular across a range of social sciences, as researchers come to appreciate that observed outcomes depend on variables organised in a nested hierarchy. There are many applications of multilevel modeling such as educational research where a number of well defined groups are organised within a hierarchical structure, for instance, the teacher-pupil relationships, leading to the analysis of effects on individual pupil behaviour coming from different hierarchical levels. In geographical studies, often envisage a hierarchy of effects from cities, regions containing cities, and countries containing regions. Failure to incorporate these effects emanating from different hierarchical levels will lead to incorrect inferences. However, while standard approaches to multilevel analysis are well established, there is none the less much scope for refinment and development of this extremely useful methodology.

The aim of this study is to fit the HIV prevalence of (EDHS 2005) via two hierarchical multilevel modeling approaches namely, the frequentist approach and the Bayesian approach. Furthermore, the results were compared and shown which approach performs better.

5.2 Methodology

5.2.1 Data

This study was based on a secondary analysis of the data from the (EDHS 2005). The EDHS survey was conducted from April 27 to August 30, 2005. The survey was designed to provide national, urban/rural, and regional estimates of key health and demographic indicators. In the first stage of the survey, 540 clusters were selected from the list of Enumeration Areas (EAs) utilised in Ethiopia's 1994 Population and Housing Census. Fieldwork was successfully completed in 535 of the 540 clusters. In the second stage, 24 to 32 households were selected systematically from each cluster within the survey sample. The survey involved administrating the women's questionnaire to all eligible women aged 15 to 49 within the households that were part of the sample. The men's questionnaire was administrated to all eligible men aged 15 to 59 within every sample household.

In order to test for HIV, those conducting the survey collected blood specimens from all the eligible women and men within each household selected for sampling. The response rates for HIV testing were 83% amongst women and 76% amongst the men. The analysis used data collected from the 14070 women and 6033 men who completed the interview, who reported ever having had sexual intercourse, and who had a valid HIV test results. Because of the way the survey was designed, the number of cases in some regions appears small since they were weighted to make the regional distribution nationally representative.

In order to carry out the HIV testing, a random sample of 50% of the households selected from the overall survey sample were selected, and all the eligible women and men in those homes were asked to provide their consent for a blood sample to be taken and for that blood to subsequently tested for HIV. When dealing with youths between the ages of 15 to 17 consent was obtained from a parent. Everyone form whom consent was obtained then provided three to four drops of blood. The blood was taken by way of a finger prick. Each sample was labeled with a special bar code label.

The blood samples were then transported to Addis Ababa to be tested for HIV at the Ethiopia Health and Nutrition Research Institute (EHNRI), which is a national laboratory. HIV testing was conducted using standard laboratory and quality control procedures. Blood collection and HIV testing protocol allowed for anonymous linking each HIV test result to that particular individual's socio-demographic and behavioural characteristics as obtained from his or her completed questionnaires. Barcodes were used to make this link, and this was done only after household and cluster identification codes were scrambled to ensure that all potential identifiers had been destroyed.

5.2.1.1 Covariates

The independent variables entertained includes socio-demographic characteristics (age, region, place of residence, education, religion, marital status), socio-cultural factors (decision making ability, wealth index, stigma, circumcision), sexual behaviour characteristics (number of sexual partners in the last 12 months, had any STI in the last 5 years, age at first sex).

Principle component analysis (PCA) [48] was used to generate the stigma, media exposure and the ability of decision making. Stigma is defined as an attribute or label that sets a person apart from others and links the labeled person to undesirable characteristics [36]. Stigma related to AIDS has been defined as "the prejudice, discounting, discrediting, and discrimination that are directed at people perceived to have AIDS" [64]. A stigma index was created based on responses to the questions "willing to care for relative with AIDS", "person with AIDS allowed to continue" and "would buy vegetables from vendor with AIDS ". Based on factor scores, respondents were classified as having low, medium or high HIV-related stigma. A media exposure index was also computed using PCA based on responses to questions posed on the frequency of watching television, the frequency of listening to radio, and the frequency of reading newspapers. The respondents were then classified as having low, medium or high media exposure. The decision making index was computed based on the respondents answer to the questions: Final say on own health care, final say on making large household purchases, final say on making household purchases for daily need, final say on visits to family or relatives, final say on food to be cooked each day, and final say on deciding what to do with money husband earns. The decision making index was a trichotomous variable with levels independent, consults and subservient.

5.3 Results

This section introduces HIV prevalence by categorical predictors for the data as well as the results obtained from an analysis for both frequentist and Bayesian approaches. In order to fit the frequentist multilevel model, we used lme4 Package in R software to analyse the data. For each parameter we compute the estimates and the 95% confidence intervals were computed and the deviance for the model calculated . We ran Bayesian multilevel models by fitting generalised linear mixed models using MCMC techniques in a MCMCglmm package in R software [74]. We ran 60000 MCMC iterations, with a burn-in period of 3000 iterations and the convergence of chain was tested by plotting both the fixed and the random effects of the model, and also tested by the means of the autocorrelation statistics. The priors were $(V = 1, \nu = 0.002)$, which required owing to the numerical problems of singularity in the mixed model equations. For sensitivity issues of the analysis the model was run many times with different priors, and the results were consistent. For each parameter we compute the estimates and 2.5%, median , 97.5% quantiles and the deviance were finally computed.

5.3.1 Summary statistics

This subsection presents the HIV prevalence rate by categorical predictors for the data

Table 4: Percent of females aged 15-49 years and males aged 15-54 years who are HIV-positive, with p values for χ^2 test, according to selected Demographic, Social, Biological, and Behavioral Characteristics, 2005 Ethiopia DHS

Variable	**Female**		**Male**	
	percentage	HIV positive	percentage	HIV positive
Age	$p = 0.0011$		$p < 0.001$	
15-19	0.69	14	0.12	1
20-24	1.72	26	0.36	8
25-29	2.11	38	0.70	14
30-34	1.48	17	1.94	19
35-39	4.44	27	1.82	12
40-44	3.06	13	2.84	11
45-49	0.85	7	0.01	1
50-54	-	-	0.88	3
55-59	-	-	0.34	1
Place of residence	$p < 0.001$		$p < 0.001$	
Urban	7.73	102	2.61	36
Rural	0.65	40	0.64	34
Region	$p < 0.001$		$p = 0.0044$	
Tigary	2.56	9	1.97	8
Afar	3.25	5	2.21	4
Amhara	1.83	13	1.42	11
Oromia	2.23	21	0.41	4
Somali	1.3	2	0	0
Ben-Gumz	0.9	4	0	50
SNNP	0.1	2	0.37	3
Gambela	5.51	16	6.18	15
Harari	4.59	15	2.05	6
Addis Abeba	6.06	42	2.05	6
Dirw Dawe	4.36	13	1.73	3
Religion	$p < 0.001$		$p < 0.001$	
Protestant	0.96	17	0.38	13
Others	0.79	17	0.17	7
Orthodox	2.91	108	1.60	50
Education level	$p < 0.001$		$p = 0.0036$	
No education	1.03	49	0.77	17
Primary	2.45	40	0.52	17
Secondary	6.05	51	2.10	31
Higher education	1.05	1.65	5	
Marital status	$p < 0.001$		$p < 0.001$	
Never married	0.7	23	0.29	14
Married, only wife	1.59	64	1.24	47
Married, other wives	1.45	7	-	-
Separated	6.43	48	2.64	9
Wealth index	$p < 0.001$		$p < 0.001$	
poorer	1	57	0.29	4
Poorest	0.34	8	0.62	10
Middle	0.43	9	0.85	9
Richer	0.21	8	0.37	8
Richest	6.09	110	2.21	39
STI	$p = 0.7155$		$p = 0.7166$	
NO	2.72	2	0.92	70
YES	1.86	5140	0	0
Media exposure	$p < 0.001$		$p = 0.0105$	
Low	0.67	30	0.60	8
Medium	1.21	12	0.64	19
High	4.58	98	1.68	43
Stigma	$p < 0.001$		$p < 0.001$	
Low	5.66	72	1.71	36
Medium	3.02	35	1.23	25
High	0.61	10	0.24	9
Birth in last 5 years	$p = 0.6180$			
Yes	1.74	61	-	-
No	1.99	81	-	-
Total number of children who had died	$p = 0.2769$		$p < 0.001$	
Never had children	1.29	40	0.45	21
No child died	2.41	71	1.92	35
One child died	1.72	18	0.44	8
Two or more child died	1.84	13	0.70	6
Number of the partners in the past years	$p < 0.001$			

Table 4 – *Continued from previous page*

Never had sex	0.12	8	-	-
One partner	1.44	59	-	-
More than one partner	4.95	74	-	-
Decision making ability	$p = 0.0768$			
Independent	1.86	26	-	-
Consults	2.01	28	-	-
Subservient	0.72	16	-	-
Circumcision			$p < 0.001$	
NO	-	-	1.11	13
YES	-	-	0.9	57
Age at first sex			$p < 0.001$	
Not had sex	-	-	0.16	4
Less than 14 years	-	-	00.12	1
14-17 years	-	-	1.49	18
more than 18 years	-	-	1.21	46

Table 4 presents the HIV prevalence rate with p value for χ^2 test for the men and women data. One of the key demographic characteristics was age . Male in the 40-44 age groups are at highest prevalence rate (2.84%), while females in 35-39 are at highest prevalence rate (4.44%). In terms of regions, males of Gambela are at highest prevalence (6.18%), followed by Addis Ababa (3.3%), Affar (2.21%) and Harari (2.05%) and the rest of the regions have less than (2%) prevalence . The Addis Ababa females have the highest prevalence rate (6.06%), followed by Harari (4.59%) and Dire Dawa (4.36%). All the other regions have prevalence rate less than (4%). HIV prevalence is higher in urban areas (2.62%, 7.73%) for males and females respectively. In terms of number of children that have died, the HIV prevalence is highest among those with no children dead with (1.92%, 2.41%) for males and females respectively. In terms of education level the highest prevalence was found among those with highest education level followed by those with higher education for both males and females. In terms of wealth index for both males and females, the highest prevalence are found among the richest (2.21%, 6.09%) respectively. All the others have prevalence 1% or below. The HIV prevalence rate is highest amongst the separated men (2.64%), followed by married (1.24%). The prevalence is very low among those never married. Also the separated females have highest prevalence . For the religion both males and females, the Orthodox have the highest prevalence (1.6%, 2.91%) respectively, and the rest have prevalence less than 1%. In terms of media exposure, both males and females the prevalence are among those with high media exposure (1.68%, 4.58%) respectively. For both males and females HIV prevalence increase with lower HIV-related stigma. The uncircumcised male has the highest prevalence rate (1.11%). The highest prevalence among the consultant females (2.01%), followed by independent (1.86%) and very low for subservient females.

5.3.2 The female data analysis results

This subsection presents the results obtained from the frequentist and the bayesian approaches for the female data.

Table 5 presents the results obtained from the female data for both approaches. The findings show women aged 25-29 years are at a significantly higher risk of being HIV-positive than women in the reference category aged 15-19. Those women lived in urban areas are double as likely to be HIV-positive as those in rural areas. Those with primary education are twice to as likely to be HIV-positive as those with no education; although the probability for acquiring HIV are higher for the higher and secondary education categories, there are not significantly different than those for women with no education. Wealth is positively related to being HIV-positive, the wealthiest women being 3 times more likely to be infected than those in poorest category. Regarding marital status, separated women are almost 2 times more likely to be HIV-positive, women who are married other wives are over 1.25 times more likely to be HIV-positive, and women who married only one wife are about 0.8 to be HIV-positive. Orthodox women are found to at higher risk for acquiring HIV. The women who have more than one partners are 3.25 times more likely to be infected than those in never had sex category. Women who report having had an STI in the year preceding are 2.25 more likely to be HIV-positive than women who did not. Regarding to stigma, Women who report having had low stigma are nearly 3 times likely to be HIV-positive than women with high stigma.

5.3.3 The male data analysis results

This subsection presents the results for the male data Table 6 presents the results obtained from the male data. For age group we note the inverted U-shaped relationship between age and infection with HIV, such that those in ages ranging from 20 to 44 years are the most to be infected, with risk peaking for those in the ages 35 to 39 years. All age groups are more likely to be infected than the reference group. Men living in urban areas are nearly 4 times more likely to be infected than those in lived in rural area. Regarding to education, Those with higher education are 23 times more likely to be infected than those with no education. With those in the wealthiest quintile being 10 times more likely to be infected than those in the poorest quintile.

Table 5: Comparison of Results for Frequentist and Bayesian Multilevel Model for the Female data

Parameters	Frequentist	Bayesian			
	Estimates of mean(95% C I)	Estimates of mean	2.5%	Median	97.5%
Fixed part					
Intercept	-6.98(-8.34,-3.81)	0.02	-0.004	-0.01	0.03
Age					
15-19	Ref	Ref	-	-	-
20-24	0.04(-0.79,0.87)	0.002	-0.001	0.003	0.02
25-29	0.53(-0.31,1.37)	0.01	-0.000002	0.01	0.03
30-34	0.01(-0.946,0.98)	0.73	-0.001	0.003	1.52
35-39	0.39(-0.545,1.32)	0.02	-0.0001	0.01	0.03
40-44	-0.26(-1.33,0.808)	-0.001	-0.001	-0.001	0.02
45-49	-0.66(-1.89,0.58)	-0.001	-0.003	-0.01	0.01
Religion					
Protestant	Ref	Ref	-	-	-
Others	-0.72(-1.39,-0.088)	-0.01	-0.002	-0.01	-0.002
Orthodox	0.02(-0.68,0.64)	-0.002	-0.001	-0.003	0.01
Residence					
Rural	Ref	Ref	-	-	-
Urban	0.77(0.04,1.57)	0.02	0.006	0.02	0.03
Education Level					
No Education	Ref	Ref	-	-	-
Higher Education	1.55(-0.13,3.22)	0.05	0.001	0.05	0.07
Primary	0.85(0.25,3.52)	0.06	0.002	0.06	0.09
Secondary	0.14(0.092,3.29)	0.05	0.002	0.05	0.08
Marital Status					
Never married	Ref	Ref	-	-	-
Married, only wife	-0.22(-0.57,0.90)	0.004	0.00008	0.03	0.05
Married, other wives	0.22(-0.54,1.40)	0.02	-0.001	0.004	0.02
Separated	0.68(0.42,1.37)	0.04	0.002	0.04	0.05
Wealth Index					
Poorest	Ref	Ref	-	-	-
Middle	0.29(-1.42,0.84)	-0.001	-0.001	–0.002	0.01
Poorer	0.04(-1.416,0.91)	-0.002	-0.001	-0.002	0.01
Richer	0.03(-1.38,0.86)	-0.003	-0.001	-0.003	0.01
Richest	1.22(-0.13, 1.98)	0.01	-0.000007	0.02	0.03
STI					
NO	Ref	Ref	-	-	-
Yes	0.81(-1.06,2.68)	0.05	-0.001	0.04	0.11
Media Exposure					
High	Ref	Ref	-	-	-
Medium	-0.74(-1.52,0.03)	-0.01	-0.001	-0.01	-0.001
Low	-0.19(-0.90,0.52)	-0.01	-0.002	-0.01	0.01
Birth in the past 5 years					
Yes	Ref	Ref	-	-	-
No	0.09(-0.95,1.12)	-0.03	-0.81	-0.03	0.73
Stigma					
High	Ref	Ref	-	-	-
Medium	0.67(-0.02,1.36)	0.004	-0.0004	0.005	0.01
Low	0.94(0.24,1.66)	0.02	0.0004	0.01	0.03
Number of the partners in the past years					
Never had sex	Ref	Ref	-	-	-
One partner	0.66(-0.68,0.93)	-0.03	-0.004	-0.04	-0.03
More than one partner	1.18(-0.68,0.93)	-0.09	-0.01	-0.09	-0.07
Total number of children who had died					
Never had children	Ref	Ref	-	-	-
No child died	-0.29(-0.86,0.69)	-0.01	-0.001	0.01	0.02
One child died	-0.37(-1.04,0.69)	-0.02	-0.001	0.003	0.01
Two or more child died	-0.20(-1.55,-0.68)	-0.02	-0.0001	-0.002	0.04
Decision making ability					
Independent	Ref	Ref	-	-	-
Consults	0.20(-0.67,0.26)	-0.004	-0.001	-0.004	0.01
Subservient	0.27(-0.54,0.67)	0.001	-0.0008	0.001	0.01
Random Part					
Household	2.09	0.0002	0.0001	0.0002	0.0003
Cluster	2.24	0.0005	0.0003	0.0005	0.0008
Individuals	2.24	0.02	0.02	0.02	0.02
Deviance	1011		-5816.077		

Table 6: Comparison of Results for Frequentist and Bayesian Multilevel Model for the male data

Parameters	Frequentist	Bayesian			
	Estimates of mean(95% C I)	Estimates of mean	2.5%	Median	97.5%
Fixed part					
Intercept	-14.40(-37.23,8.44)	0.03	-0.003	0.002	0.06
Age					
15-19	Ref	Ref	-	-	-
20-24	1.52(-7.26,10.29)	0.001	-0.01	0.0005	0.02
25-29	4.44(-4.95,13.84)	0.02	0.002	0.001	0.03
30-34	5.24(-5.04,5.52)	0.03	0.01	0.002	0.04
35-39	3.79(-6.81,14.39)	0.02	0.004	0.002	0.04
40-44	3.69(-8.45,15.83)	0.02	0.01	0.002	0.05
45-49	0.29(-12.51,13.09)	0.003	-0.02	0.0003	0.02
50-54	0.56(-14.76,15.89)	0.01	-0.01	0.001	0.03
55-59	-0.48(13.7,22.38)	0.01	-0.02	0.0005	0.03
Residence					
Rural	Ref	Ref	-	-	-
Urban	1.37(1.02,3.77)	0.02	0.003	0.001	0.03
Education Level					
No Education	Ref	Ref	-	-	-
Higher Education	3.17(4.81,6.47)	0.0002	-0.02	0.000002	0.02
Primary	-3.55(-13.85,6.74)	0.002	-0.02	0.0002	0.02
Secondary	-1.35(-11.49,8.79)	0.01	0.001	0.001	0.03
Marital Status					
Never married	Ref	Ref	-	-	-
Married or living together	-1.16(-10.00,7.69)	-0.01	-0.03.02	-0.001	0.01
Separated	2.71(-6.43,1.85)	0.02	0.01	0.002	0.04
Religion					
Protestant	Ref	Ref	-	-	-
Others	-4.82(-13.74,4.11)	-0.01	-0.02	-0.001	-0.01
Orthodox	-4.40(-12.63,3.82)	-0.01	-0.01	-0.0006	0.004
Wealth Index					
Poorest	Ref	Ref	-	-	-
Middle	0.15(-12.12,12.42)	0.002	-0.01	0.0001	0.01
Poorer	-3.25(-19.62,13.11)	-0.01	-0.02	-0.0006	0.004
Richer	-3.41(-17.22,10.39)	-0.003	-0.01	-0.0002	0.008
Richest	2.34(6.36,11.67)	-0.01	-0.02	-0.0005	0.007
Circumcision					
Yes	Ref	Ref	-	-	-
No	1.37(2.68,2.07)	-0.03	-0.04	-0.002	-0.01
Total number of children who had died					
No child dead	Ref	Ref	-	-	-
Never had children	-0.14(-13.81,13.52)	0.01	-0.006	0.001	0.03
One child died	0.64(1.89,3.18)	-0.0002	-0.01	-0.57	0.01
Two or more child died	-0.48(-13.81,13.52)	0.01	-0.005	-0.0005	0.02
Media Exposure					
Low	Ref	Ref	-	-	-
Medium	-1.12(-6.39,4.15)	-0.01	-0.4	-0.0003	0.01
High	-3.43(-13.85,6.99)	-0.02	-0.4	-0.0006	0.004
Age at first sex					
Not has sex	Ref	Ref	-	-	-
Less than 14 years	-2.05(-24.36,20.26)	-0.005	-0.04	-0.0004	0.03
14-17 years	-2.96(-10.36,4.44)	-0.01	-0.02	-0.0006	0.01
More than 18 years	1.08(-5.18,3.03)	-0.004	-0.01	-0.0003	0.01
Stigma					
Low	Ref	Ref	-	-	-
Medium	0.31(-6.85,7.46)	0.01	-0.002	-0.0006	0.02
High	-0.73(-8.99,7.53)	0.004	-0.005	0.0006	0.004
STI					
NO	Ref	Ref	-	-	-
Yes	0.02(-0.95,1.12)	-0.02	-0.06	-0.001	0.03
Random Part					
Household	2.09	0.0002	0.0001	0.0002	0.0004
Cluster	2.24	0.0003	0.0001	0.0002	0.0004
Individual	2.24	0.013	0.012	0.013	0.01
Deviance	448.7		-7555.704		

Regarding to marital status, separated men are almost 15 times more likely to be HIV-positive compared to those never married. However, the factor of circumcision status is among the strongest in the model: men who are uncircumcised are 4 times the risk of those who are circumcised to HIV-positive. The reporting STI in the past year is not significantly related to HIV related serostatus for men.

5.4 Discussion and Conclusions

If classical regression approaches are applied to multilevel data sets, the classical regression models fail to account for the dependency structure in the data. However, multilevel models do lead to valid inferences. The multilevel regression model is more complicated than the standard single-level multiple regression model. One difference is the number of parameters, which is much larger in the multilevel model. This poses problems when models are fitted that have many parameters, and also in model exploration. Another difference is that multilevel models often contain interaction effects in the form of cross-level interactions. Interaction effects are tricky, and analysts should deal with them carefully. Finally, the multilevel model contains several different residual variances, and no single number can be interpreted as the amount of explained variance.

In this study two types of the multilevel models were implemented, namely the frequentist and the Bayesian multilevel models and then the results compare. The results obtained from both approaches are identical [136]. The key demographic factor in this study is age. Women and men in earlier age were found at higher risk for acquiring the disease [121]. Urban residence was associated with much higher HIV infection rates, and there were large differentials by geographic region. Large regional variations in HIV prevalence have also observed in other Sub-Saharan Africa [86, 111]. The study found that the separated women and men had a higher likelihood of HIV infection. This was expected as they can get infected by their formal spouses. Higher education and higher household wealth were found to be positively associated with HIV infection. However, as expected, controlling for urban-rural residence, sexual behavior, and other factors that tend to be correlated with higher socioeconomic status diminished these associations considerably [110]. The most important biological factor associated with being HIV-positive for men was circumcision status: uncircumcised men were 4 times more likely to be HIV-positive than their counterpart. Other studies conducted in the region indicate that circumcised status may be over reported by as 10% [7].

One of the weaknesses of the multilevel model is the design issue. In multilevel models design where it is nested stricture of a population is to be modeled, the allocation of level 1 unites among level 2 units and the allocation of these among level 3 units and so on. It is clearly affects the precision of the resulting estimates of both fixed and random effect parameters. The problem of the design issue is more complicated when there are random cross associated with sampling more level 1 units within an existing level 2 units as opposed to selecting further level 1 units in a new level 2 unit. In the mean time there appears to be little empirical or theoretical work on issues of optimum design for multilevel models.

Multilevel models have a very good ability to separately estimate the predictive effects of an individual predictor and its group-level mean, which is sometimes interpreted as the direct and contextual" effects of the predictor. Multilevel models also take the sample stricture into account. Multilevel models provide a useful framework for thinking about problems with samples which have hierarchical structure. One advantage of the multilevel modeling approach is that it can deal with data in which the times of the measurements vary from subject to subject.

The other weakness of multilevel models is the missing values. The missing values are occur in surveys, because certain questions are not answered by particular groups of respondents. The main concerned with missing values is the outcome variables. In order to handle the missingness, a missing data mechanism should be used [101, 129]. Moreover, when the sample size is very small it leads to biased estimators and misleading statistical tests [52]. The other weakness of multilevel models is the power. Power considerations are more complex in multilevel designs because these studies typically have multiple purposes that may include detecting main and interaction effects within and across levels of analysis and testing variance components. Power analysis are further complicated by the numerous factors that impact power in a multilevel context. These include the number of units at each level, the magnitude of the intraclass correlation coefficient representing clustering effects, the presence of covariates at each level of the model and their relation to the outcome, effect size, and the alpha level used for statistical tests. In experimental studies designed to detect treatment effects, the level of randomisation has been found to impact power, such that randomization of the Level 2 units leads to less power than randomisation of the Level 1 units [46, 84]. The other critical issue in multilevel models is the model development and specification, and the selection of the predictors which are a critical part of the design of study [102]. In multilevel modeling variable selections can be complicated due to predictors selected for each level or cross levels. Furthermore, the process of variable selection can take many forms. The biggest challenge in multilevel modeling is how to specify the covariance structure. To the best of our knowledge there are weaknesses in the estimation procedure. There is no single agreed on the methods of estimation of multilevel parameters. Many methods of estimation can be applied, such as, ML, REML, and Bayesian estimation. There are also some issues in test of hypothesis and statistical inference.

The biggest limitation of the Bayesian approach, is the time spent to run the model when using the MCMC methods of estimation, especially when the data sets are huge or the variables of the model are too many. It also takes a long time to draw a final decision from Bayesian models which means that models should be run with different priors on the model parameters. Furthermore, in order to make sure that the choice of the prior distribution does not affect the results, sensitivity analysis should have be performed. This was not done in this thesis, because the parameter estimates obtained from the Bayesian approach were similar to that of the frequentist approach (refer Tables 5,6).

Chapter 6

Spatial modeling and mapping for complex survey data sets

6.1 Introduction

Spatial modeling deals with a specific form of disaggregation, in which an area is divided into a number of similar units: typically grid squares or polygons. The spatial model may be linked to a geographical information system (GIS) for data input and display. The transition from non-spatial to spatial modeling is often considered to be very significant, and there are a number of modeling packages that have their spatial modeling capabilities: indeed, many are labeled as landscape or landuse modeling tools [126].

Complex issues arise in spatial modeling, many of which are neither clearly defined nor completely resolved. The most fundamental of these is the problem of defining the spatial location of the entities being studied. Other issues in spatial modeling include the limitations of mathematical knowledge, the assumptions required by existing statistical techniques, and problems in computer based calculations. The classification techniques of spatial modeling is difficult because of the large number of different fields of research involved, the different fundamental approaches which can be chosen, and the many forms the data can take. These problems represent a challenge in spatial modeling because of the power of maps as media of presentation. When results are presented as maps, the presentation combines spatial data which are generally accurate with analytic results which may be in accurate, leading to an impression that analytic results are more accurate than the data would indicate [114]. The aim of the chapter is to give a general account of the spatial models.

6.2 The Spatial Models description

The spatial models allow the estimates to borrow strength from other regions through random effects terms. This section presents spatial models.

6.2.1 Conditional Auto-Regressive (CAR) Model

The CAR model, known in the literature as Auto-Normal model or Gauss-Markov model [31] is used, usually, to analyse phenomena that occur in a geographical area. More specifically, the CAR model is a continuous Markov field characterized by conditional probability density function and particularly suited to model spatial phenomena strongly tied to specific local context [15, 40]. Its utility is also largely attributed to the existence of a clear link between the conditional probability distribution and the joint distribution [15]

Let $S = \{1, 2, 3, \ldots, n\}$ be a spatial domain and the neighborhood $\aleph_i$ of area i, $i \in S$ implies that $\aleph_i = \{j \in S : j$ is a neighborhood of $i\}$. Furthermore, assign the random variable $X_i, i \in S$, and define the corresponding random field $\boldsymbol{X}$ as the following vector: $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)'$.

In the Gaussian CAR model, we assume that each observation of the out come variable X_i has conditional distribution:

$$X_i|\boldsymbol{X}_{(-i)} \sim N\left(\sum_{j\in\aleph_i} b_{ij}x_j, \sigma_i^2\right), \tag{6.1}$$

where b_{ij} is the weight of each observation on the mean of X_i and also denotes the spatial dependence parameter. The b_{ij} is not zero only if $j \in \aleph_i$. Traditionally, we set $b_{ij} = 0$ since we are not regressing any observation on itself. self. The $\boldsymbol{X}_{(-i)}$ denotes a vector of all observation except X_i. As we know X_i depends only on a set neighbor $\boldsymbol{X}_{(-i)}$ only if location j is in the neighborhood set $\aleph_i$ of X_i. The σ_i^2 is a potential unique variance for X_i.

The Gaussian processes are specified by their mean and covariance function [152]. The full conditional probability density function of the CAR model can be written as

$$f(x_i|x_{j\in\aleph_i}) = \sqrt{\frac{1}{2\pi\sigma_i^2}}\exp\left\{-\frac{[(x_i-\mu_i)-\rho\sum_{j\in N_i}\beta_{i,j}(x_j-\mu_j)]^2}{2\sigma_i^2}\right\}, \tag{6.2}$$

where $i, j \in S$, $\mu_i \in \Re, \sigma_i^2 \in \Re^+, |\rho| < 1, \beta_{(ij)} \in \Re, \beta_{(ij)} = \beta_{(ji)}, \beta_{(ii)} = 0$. Also the conditional joint probability density function of all the observation is

$$f(x) = \frac{1}{(2\pi)^{n/2}\det(\boldsymbol{B}^{-1}\Sigma_D)^{1/2}}exp\left[-\frac{(x-\boldsymbol{\mu})'\Sigma_D^{-1}\boldsymbol{B}(x-\boldsymbol{\mu})}{2}\right], \tag{6.3}$$

Where $\boldsymbol{\mu}$ is a n dimensional vector as $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)'$, while $\boldsymbol{B}$ is $n \times n$ invertible matrix of the following type

$$\mathbf{B} = (\mathbf{I} - \rho\beta) \text{ with } b_{ij} = \begin{cases} 1 \text{ if } i = j \\ -\rho\beta_{(ij)} \text{ if } j \in \aleph_i \\ 0 \text{ otherwise} \end{cases} \tag{6.4}$$

The matrix $\boldsymbol{\Sigma}_D$ is $n \times n$ diagonal matrix; where $\boldsymbol{\Sigma}_D = \text{diag}(\sigma_i^2, \ldots, \sigma_n^2) = \Sigma_{D_{ii}} = \sigma^2$ such that Σ_D is symmetric. The CAR model for $\boldsymbol{X}$ in equation (6.2) can be written as follows:

$$X_i|X_j \sim N\left[\mu_i + \rho\sum_{i\in N_i}\beta_{(ij)}(x_j-\mu_j), \sigma_i^2\right], i \in S \tag{6.5}$$

In matrix and vector notation equation (6.5) can written as

$$\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{B}^{-1}\Sigma_D), \tag{6.6}$$

The necessary and sufficient condition for (6.6) to be a valid joint probability density function is that the covariance matrix must not only be symmetric, but also positive definite. For X_i to be a Gaussian random variables, we defined a symmetric weighted adjacency matrix $\boldsymbol{W}$, where

$$\mathbf{W} = (\omega_{(ij)}) \text{ with } \omega_{(ij)} = \begin{cases} 1 \text{ if } i = j \\ \varphi(i,j) \text{ with } j \in N_i \\ 0 \text{ otherwise} \end{cases} \tag{6.7}$$

Where $\forall i, j \in S, \omega_{(ij)} = \omega_{(ji)}$; $\varphi(i, j)$ is a measure that quantifies the proximity between region i and region j; if $\varphi(i, j) = 1$, then i and j are neighbors. The indicator $\varphi(i, j) = 1$ represent the distance between the centroid of region i and j. It is not necessary that $\boldsymbol{W}$ is symmetric . Let $\boldsymbol{W}$ be the diagonal of the adjacency matrix $\boldsymbol{W}$. The adjacency matrix of normalization $\boldsymbol{W}_D$ can be defined as follows:

$$\boldsymbol{W}_D = diag(\omega_{(1+)}, \omega_{(2+)}, \ldots, \omega_{(n+)}). \tag{6.8}$$

Where

$$\omega_{(i+)} = \sum_{j \in N_i} \omega_{(ij)}, i, j \in S,$$

Then we can define a matrix of interaction $\boldsymbol{\beta}$ as a normalized adjacency matrix given as follows:

$$\boldsymbol{\beta} = \boldsymbol{W}_D^{-1}\boldsymbol{W}, \tag{6.9}$$

Where $\beta_{(ij)} = \frac{\omega_{(ij)}}{\omega_{(i+)}}$; $\beta_{(ij)}\sigma_j^2 = \beta_{(ji)}\sigma_j^2, i, j \in S$. Suppose once again that the matrix $\boldsymbol{W}_D$ corresponding to a constant diagonal matrix need to be normalized as follows:

$$\boldsymbol{W}_D = \sigma^2 \boldsymbol{W}_D^{-1}, \tag{6.10}$$

Where $\sigma_i^2 = \frac{\sigma^2}{\omega_{(i+)}}, i \in S, \sigma^2 \in \Re_+$. The conditional joint probability density function can be written as follows:

$$f(x_1, \ldots, x_n) \propto exp\{-\frac{1}{2\sigma^2}\boldsymbol{q}'(\Sigma_{D_w} - \boldsymbol{W})^{-1}\Sigma_D \boldsymbol{q}\} \tag{6.11}$$

This is a multivariate Gaussian distribution where $\boldsymbol{q} = (x - \mu)$ and $\boldsymbol{B} = (\Sigma_{D_w} - \boldsymbol{W})$. The CAR model for the random field X can be written as follows:

$$x_i | x_j \sim N\left(\sum_j \frac{\omega_{ij}}{\omega_{i+}} x_j, \frac{\sigma_i^2}{\omega_{i+}}\right)$$

Note that

$$(\boldsymbol{I} - \beta) = \frac{\Sigma_D - \boldsymbol{W}}{\sigma^2}.$$

From the conditional joint probability density function of the CAR model, we have

$$f(x_1, x_2, \ldots, x_n) \propto exp\left[-\frac{1}{2\sigma^2}\left(\sum_j (x_i - x_j)\right)^2\right] \tag{6.12}$$

Then the joint probability density function of (6.12) can be written a more compact form as follows:

$$\boldsymbol{X} \sim N\left(\boldsymbol{\mu}, \left[\frac{1}{\sigma^2}(\boldsymbol{W}_D - \rho\boldsymbol{W})\right]^{-1}\right)$$

The row stochastically of $\hat{\boldsymbol{W}} = diag\frac{1}{\omega_{i+}}\boldsymbol{W}$ indicates that the distribution is not suitable. This property can be fixed by the parameter ρ. We can redefine $\Sigma_{D_x}^{-1} = (\Sigma_{D_w} - \rho\boldsymbol{W})^{-1}$ and select ρ so that $\Sigma_{D_x}^{-1}$ is non singular, preferably with $\rho \in \left(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}\right)$, where $\lambda_1 < \lambda_2, \ldots, < \lambda_n$ are the ordered eigenvalues of $\Sigma_{D_w}^{-1}\boldsymbol{W}\Sigma_{D_w}^{1/2}$. By simplifying the bounds, and replacing $\boldsymbol{W}$ by $\hat{\boldsymbol{W}}$ we get the following

$$\Sigma_{D_x}^{-1} = \Sigma_{D_w}(\boldsymbol{I} - \Theta\hat{\boldsymbol{W}}), \tag{6.13}$$

If $|\Theta| < 1$, then $\Sigma_{D_W}(\boldsymbol{I} - \Theta\hat{\boldsymbol{W}})$ is non singular, if $\rho \in \left(\frac{1}{\lambda_1}, 1\right)$. Here ρ is the additional parameter which makes X_i independent when it is equal to 0. We can therefore express the CAR model as follows

$$\boldsymbol{X} = \boldsymbol{B}X_i + \epsilon, \tag{6.14}$$

Where $\boldsymbol{X} \sim N\left(0, (\boldsymbol{I} - \beta)^{-1}\Sigma_D\right)$ and $\epsilon \sim N\left(0, \Sigma_D(1 - \boldsymbol{I})'\right)$.

6.2.2 Convolution spatial models

Recently, convolution based models for spatial data have been gained popularity as a result of their flexibility in handling spatial dependence and their ability to accommodate very large data sets. The flexibility of the convolution models are due to the frameworks based on moving average(MA) construction which guarantees a valid spatial covariance function. This modeled approach to spatial modeling has been used:

(1) to provide an alternative to the standard classes of parametric variogram/covariance functions commonly used in geostatistics

(2) to specify Gaussian process models with nonstationary and anisotropic covariance functions

(3) to create non-Gaussian classes of models for spatial data [26].

Let y_i be a dichotomous random variable taking value 1 if the individual is HIV positive and 0 otherwise. Assume $y_i \sim \text{Bin}(n_i, \pi_i)$ then a convolution model can be defined as follows:

$$\text{logit}(\pi_i) = \boldsymbol{X}_i'\boldsymbol{\beta} + u_i + v_i, \tag{6.15}$$

where u_i is the spatially structured random effect with $u_i|\boldsymbol{u}_{-i} \sim N(\frac{\phi}{n_i}\sum_{j\in\aleph_i} u_j, \frac{1}{\tau^2 n_j})$ and v_i is the spatially unstructured random effect distributed as $v_i \sim N(0, \sigma^2)$. The posterior distribution can be defined as follows:

$$f(\boldsymbol{u}, \boldsymbol{v}, \kappa, \lambda, \beta|\boldsymbol{y}) \propto \prod_{i=1}^{n} f(y_i|u, v, \kappa, \lambda, \beta)\prod_{i=1}^{p} f(\beta_i)\prod_{i=1}^{n} f(u_i|\kappa)f(\kappa)\prod_{i=1}^{n} f(v_i|\lambda)f(\lambda), \tag{6.16}$$

This simplifies to

$$f(u, v, \kappa, \lambda|\boldsymbol{y}) \propto \prod_{i=1}^{n}\binom{n_i}{y_i}p_i^{y_i}(1-p_i)^{n_i-y_i} \times \frac{1}{\kappa^{n/2}} \times \exp -\frac{1}{2\kappa}\sum_{i\neq j}(u_i-u_j)^2 \times$$
$$\frac{1}{\lambda^{n/2}}\exp -\frac{1}{2\lambda}\sum_{i=1}^{n} v_i^2 \times \prod_{j=1}^{p} e^{\frac{-1}{2c}\beta_j^2} \times e^{-\kappa/0.005}\kappa^{0.05} \times e^{-\lambda/0.005}\lambda^{0.05},$$

where $\prod_{i=1}^{n}\binom{n_i}{y_i}p_i^{y_i}(1-p_i)^{n_i-y_i}$ is the likelihood, $\frac{1}{\kappa^{n/2}} \times \exp -\frac{1}{2\kappa}\sum_{i\neq j}(u_i-u_j)^2$ the structured CAR prior, $\frac{1}{\lambda^{n/2}}\exp -\frac{1}{2\lambda}\sum_{i=1}^{n} v_i^2$ the unstructured exchangeable prior, $\prod_{j=1}^{p} e^{\frac{-1}{2c}\beta_j^2}$ the normal prior and $e^{-\kappa/0.005}\kappa^{0.05} \times e^{-\lambda/0.005}\lambda^0.05$ the hyperprior.

6.3 The Integrated Nested Laplace Approximation

In this section we discuss how we performed approximate Bayesian inference under a subclass of structured additive regression models, named *latent Gaussian models*. Structure additive models are a flexible and extensively used class of models.

Suppose that the observation (or response) variable y_i is assumed to belong to an exponential family, where the mean X_i is linked to a structured additive predictor η_i through a link-function $g()$, so that $g(\mu_i) = \eta_i$. The structured additive predictor η_i accounts for effects of various covariates in an additive way:

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta}\beta_k z_{ki} + \epsilon_i \tag{6.17}$$

Here, the $\{f^{(j)}()\}$s are unknown functions of the covariates $\boldsymbol{u}$, the $\{\beta_k\}$s represent the linear effect of covariates $\boldsymbol{z}$ and the ϵ_is are unstructured terms. This class of model has a wealth of applications, thanks to the very different forms that the unknown functions $\{f^{(j)}\}$ can take. Latent Gaussian models are a subset of all Bayesian additive models with a structured additive predictor (6.17); namely those which assign a Gaussian prior to α, $\{f^{(j)}\}, \{\beta_k\}$ and $\{\epsilon_i\}$. Let $\boldsymbol{x}$ denote the vector of all the latent Gaussian variables, and $\boldsymbol{\theta}$ the vector of hyperparameters, which are not necessarily Gaussian.

To simplify the following discussion, denote generically $\pi(|)$ as the conditional density of its arguments, and let $\boldsymbol{x}$ be all the n Gaussian variables α, $\{f^{(j)}\}, \{\beta_k\}$ and $\{\epsilon_i\}$. The density $\pi(\boldsymbol{x}|\boldsymbol{\theta}_1)$ is Gaussian with (assumed) zero mean, precision matrix $\boldsymbol{Q}(\boldsymbol{\theta}_1)$ with hyperparameters θ_1. Denote by $N(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$, the N($\boldsymbol{\mu},\boldsymbol{\Sigma}$) Gaussian density with mean $\boldsymbol{\mu}$ and covariance (inverse precision) $\boldsymbol{\Sigma}$ at configuration $\boldsymbol{x}$. Note that we have included $\{\eta_i\}$ instead of $\{\epsilon_i\}$ into $\boldsymbol{x}$, as it simplifies the notation later on.

The distribution for the n_d observational variables $\boldsymbol{y} = \{y_i : i \in \mathcal{I}\}$ is denoted by $\pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}_2)$ and we assume that $\{y_i : i \in \mathcal{I}\}$ are conditionally independent given $\boldsymbol{x}$ and $\boldsymbol{\theta}_2$. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$ with dim($\boldsymbol{\theta}$) $= m$. The posterior then reads (for a non-singular $\boldsymbol{Q}(\boldsymbol{\theta})$)

$$\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) \approx \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\prod_{i\in\mathcal{I}}\pi(y_i|x_i,\boldsymbol{\theta}) \tag{6.18}$$

The imposed linear constraints (if any) are denoted by $\boldsymbol{Ax} = \boldsymbol{e}$ for a $k \times n$ matrix $\boldsymbol{A}$ of rank k. The main aim is to approximate the posterior marginals $\pi(x_i|\boldsymbol{y})$, $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and $\pi(\theta_j|\boldsymbol{y})$.

Many, but not all, latent Gaussian models in the literature satisfy two basic properties which we shall assume throughout the study. The first is that the latent field $\boldsymbol{x}$, which is often of large dimension admit conditional independence properties. Hence, the latent field is a Gaussian Markov Random Field (GMRF) with a sparse precision matrix $\boldsymbol{Q}(\theta)$. This means that we can use numerical methods for sparse matrices, which are much quicker than general dense matrix calculations. The second property is that the number of hyperparameters m, is small, say $m \leq 6$. Both properties are usually required to produce fast inference.

6.3.1 MCMC approaches

The common approach to inference for latent Gaussian models is Markov chain Monte Carlo (MCMC). It is well known however that MCMC tends to exhibit poor performance when applied to such models. Various factors explain this. First, the components of the latent field $\boldsymbol{x}$ are strongly dependent on each other. Second, $\boldsymbol{\theta}$ and $\boldsymbol{x}$ are also strongly dependent, especially when n is large. A common approach to (try to) overcome this first problem is to construct a joint proposal based on a Gaussian approximation to the full conditional of $\boldsymbol{x}$. The second problem requires, at least partially, a joint update of both and x. One suggestion is to use the one-block approach: make a proposal for $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$, update $\boldsymbol{x}$ from the Gaussian approximation conditional on $\boldsymbol{\theta}'$ then accept/reject jointly. Some models can alternatively be reparameterised to overcome the second problem. Independence samplers can also sometimes be constructed. For some (observational) models, auxiliary variables can be introduced to simplify the construction of Gaussian approximations. Despite all these developments, MCMC remains painfully slow from the end users point of view.

6.3.2 INLA

The posterior marginals of interests can be written as

$$\pi(x_i|\boldsymbol{y}) = \int \pi(x_i|\boldsymbol{\theta y})\pi(\boldsymbol{\theta}|\boldsymbol{y}) \text{ and } \pi(\theta_j|\boldsymbol{y}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j}$$

This form can be used to construct nested approximations

$$\tilde{\pi}(x_i|\boldsymbol{y}) = \int \tilde{\pi}(x_i|\boldsymbol{\theta y})\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \text{ and } \tilde{\pi}(\theta_j|\boldsymbol{y}) = \int \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j} \tag{6.19}$$

Here, $\tilde{\pi}(.|.)$ is an approximated (conditional) density of its arguments. Approximations to $\pi(x_i|\boldsymbol{y})$ are computed by approximating $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and $\pi(xi|\boldsymbol{\theta},\boldsymbol{y})$, and using numerical integration to integrate out $\boldsymbol{\theta}$. The integration is possible as the dimension of $\boldsymbol{\theta}$ is small. The nested approach makes Laplace approximations very accurate when applied to latent Gaussian models. The approximation of $\pi(\theta_j|\boldsymbol{y})$ is computed by integrating out $\boldsymbol{\theta}_j$ from $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$.

This approach is based on the following approximation $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ of the marginal posterior of $\boldsymbol{\theta}$

$$\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \approx \left. \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\tilde{\pi}_G(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})} \tag{6.20}$$

where $\tilde{\pi}_G(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})$ is the Gaussian approximation to the full conditional of $\boldsymbol{x}$, and $\boldsymbol{x}^*(\boldsymbol{\theta})$ is the mode of the full conditional for $\boldsymbol{x}$, for a given θ. The proportionality sign (6.20) comes from the fact that the normalizing constant for $\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})$ is unknown.

6.4 The model selection Criteria

This section, introduce the method used to select the best fitting model from a number of proposed models. D.J Spiegelhalter et al [134] introduced the Deviance Information Criterion (DIC) which is used to select the best fitting model for a number of proposed models under the Bayesian approach. Based on the likelihood function $f(\boldsymbol{X}|\theta)$, the deviance is usually defined as follows:

$$D(\theta) = -2\ln f(\boldsymbol{X}|\theta), \tag{6.21}$$

Also the posterior mean of deviance can be defined as follows:

$$D' = -\frac{2}{M}\sum_{m=1}^{M} f(\boldsymbol{X}|\theta^m), \tag{6.22}$$

Where θ^m is the sample parameter value. The deviance of the posterior expected parameter estimate can be defined as follows:

$$\hat{D}(\hat{\theta}) = -2\ln(\boldsymbol{X}|\hat{\theta}), \tag{6.23}$$

That is, given any sample parameter value θ^m, furthermore, the deviance of the posterior expected parameter estimate can be defined as follows:

$$\hat{D}(\theta^m) = -2\ln f(\boldsymbol{X}|\theta^g), \tag{6.24}$$

The effective number of parameter identify by the model pD can be expressed as follows:

$$pD = D' - D(\hat{\theta}), \tag{6.25}$$

Where $D(\hat{\theta})$ is the deviance calculated at the posterior mean of the parameters. Therefore, the DIC can be defined as follows:

$$DIC = pD + \hat{D}, \tag{6.26}$$

Where D is the deviance calculated at the posterior mean of the parameters and p is the number of parameters. Smaller DIC values correspond to the best models. However, DIC does not work properly in models with many random effects [122].

6.5 Conclusion

This chapter was introduced the statistical spatial models. We defined CAR models, and Convolution model. In order to estimate their parameters we used INLA technique instated of MCMC. INLA provides fast and accurate Bayesian approximation to posterior marginals in latent Gaussian model compared to MCMC. Furthermore we introduce the model selection criteria (DIC). The strength of spatial models is that, it can allow the estimates to borrow strength from other regions through random effects terms.

Chapter 7

Spatial analysis and modeling of HIV prevalence data using Integrated Nested Laplace Approximation technique

This Chapter is stand alone as a research article of application of the spatial analysis and modeling of HIV prevalence data using Integrated Nested Laplace Approximation technique. The study had two primary objectives, namely:

(1) to conduct spatial analysis and modeling HIV prevalence data using the Integrated Laplace Nested Approximation (INLA)

(2) to map HIV prevalence in Ethiopia.

The analysis was done separately for men and women because the biological and social circumstances association with transmission of HIV differ by sex.

Spatial analysis and modeling of HIV prevalence data using Integrated Nested Laplace Approximation technique[1]

Mohammed O. M. Mohammed[2], T. N. O. Achia

School of Mathematics, statistics and Computer Science, University of
KwaZulu-Natal, Private Bag X01 Scottsville 3209, Pietermaritzburg,
South Africa

Abstract. Ethiopia with an estimated 1.1 million people living with HIV, has one of the largest populations of HIV infected people in the world.

In this study we made use of data from the Ethiopia Health Demographic Survey data (EDHS 2005). Our approach was to analyse males and females data separately . The study had two primary objectives, namely to conduct spatial analysis and modeling HIV prevalence data using the Integrated Laplace Nested Approximation (INLA), and to map HIV prevalence in Ethiopia. In order to achieve the objectives of the study we fitted the following four models:the Generalised Linear Model (GLM), Generalised Linear Mixed Model(GLMM), Conditional Auto-Regressive model (CAR), and convolution model. We used Deviance Information Criteria (DIC) to select the best model.

The findings revealed that GLMM has the smallest DIC for both the female and male data sets (DIC=1064.77,653.50) respectively, which indicated that GLMM was the best suited model to our data. National HIV prevalence in Ethiopia was found to be 1.4% in the analysis of the study sample, uncircumcised men were 6 times more likely to be HIV-positive than those who were not. Men aged 40 to 44 years had the highest prevalence of being HIV-positive, whereas the ages of highest risk for women were 25 to 29 years. The wealthiest women were 3.56 times more likely than the poorest women to be HIV-positive.

In this study we demonstrate that by using INLA, one can directly compute very accurate approximation to the posterior marginals. Furthermore, INLA provides more precise estimates in a matter of only minutes or even seconds . Another advantage of our approach is the heightened ability of INLA to compute model comparison criteria and various predictive measures so that models can be compared.

Keywords: Mapping HIV prevalence; INLA; DIC; Bayesian inference; Spatial modeling; Generalized Linear Mixed Models

[1] Submitted to the Journal of Statistics Sinica

[2] Corresponding Author. 209509598@stu.ukzn.ac.za

7.1 Introduction

The HIV/AIDS epidemic has already claimed more than 25 million lives and another 39.5 million people worldwide are currently estimated to be living with the disease [50, 146]. Most of the people now living with HIV/AIDS (95%) reside in low and middle-income countries, which is also where the most new HIV infections and AIDS related deaths occurs [50, 143]. The highest rate of HIV/AIDS in the world is been found in Sub-Saharan Africa, followed by the Caribbean; there is also concern about the epidemic in parts of eastern Europe and Asia [147]. HIV currently causes more deaths worldwide than does any other disease, HIV/AIDS is considered a threat to the well-being of the economic sector as well as to the social and political stability of many nations [87]

Sub-Saharan Africa is however the hardest hit by HIV and contains almost two-thirds (62% or 24.7 million people) of the people living with HIV/AIDS, even though it only contains about 11% of the world's total population [80]. The region is also home to 91% of the 2.3 million children living with HIV/AIDS globally [143]. In several countries, more than 10% of the adults are already estimated to be HIV positive [149]. There is evidence that the epidemic may be slowing or stabilising in certain eastern and western African countries, but there are also signs that the epidemic is still on the increase in a few countries [2].

Ethiopia HIV/AIDS epidemic pattern continues to be generalised and heterogeneous, but with marked regional variations. At the national level, the epidemiologic trend in 2004 has been stable. However, HIV prevalence appears to be declining in urban areas, according to the analysis of data from antenatal care (ANC) sites, which provide data that has been collected consistently for more than ten years. For example, HIV prevalence amongst pregnant women attending ANC in Addis Ababa has declined from 23% in 1996 to 10% in 2007.

INLA is a new tool for Bayesian inference on latent Gaussian models when the focus is on posterior marginal distribution [130]. INLA substitute Markov Chain Monte Carlo(MCMC) simulations with accurate, deterministic approximations for posterior marginal distribution. The quality of such approximations is extremely high, such that even very long MCMC runs are unable to detect any error in them. A detailed description of the INLA method and a thorough comparison with the MCMC result can be found in [130].

No much work has been conducted on spatial modelling and mapping of HIV prevalence in Ethiopia. In general, this study focused on using data from the Ethiopia Demographic Health Survey data (EDHS 2005). The main objectives were to map HIV prevalence in Ethiopia and make use of the spatial technique via INLA approach.

7.2 Methodology

7.2.1 Data

This study was based on a secondary analysis of the data from the (EDHS 2005). The EDHS survey was conducted from April 27 to August 30, 2005. The survey sample was designed to provide national, urban/rural, and regional estimates of key health and demographic indicators. In the first stage of the survey, 540 clusters were selected from the list of Enumeration Areas (EAs) utilised in Ethiopia's 1994 Population and Housing Census. Fieldwork was successfully completed in 535 of the 540 clusters. In the second stage, 24 to 32 households were selected systematically from each cluster within the survey sample. The survey involved administrating the women's questionnaire to all eligible women aged 15 to 49 within the households that were part of the sample. The men's questionnaire was administrated to all eligible men aged 15 to 59 within every sample household.

In order to test for HIV, those conducting the survey collected blood specimens from all the eligible women and men within each household selected for sampling. The response rates for HIV testing were 83% amongst women and 76% amongst the men. The analysis used data collected from the 14070 women and 6033 men who completed the interview, who reported ever having had sexual intercourse, and who had a valid HIV test results. Because of the way the sample was designed, the number of cases in some regions appears small since they were weighted to make the regional distribution nationally representative.

In order to carry out the HIV testing, a random sample of 50% of the households selected from the overall survey sample was selected, and all the eligible women and men in those homes were asked to provide their consent for a blood sample to be taken and for that blood to subsequently tested for HIV. For the youths between the ages of 15 to 17, after consent was obtained from their parents, three to four drops of blood sample was taken. The blood was taken by way of a finger prick. The blood were placed filter paper. Each sample was labeled with a special bar code label.

The blood samples were then transported to Addis Ababa to be tested for HIV at the Ethiopia Health and Nutrition Research Institute (EHNRI), which is a national laboratory. HIV testing was conducted using standard laboratory and quality control procedures. Blood collection and HIV testing protocol allowed for anonymous linking each HIV test result to that particular individual's socio-demographic and behavioural characteristics as obtained from his or her completed questionnaires. Barcodes were used to make this link, and this was done only after household and cluster identification codes were scrambled to ensure that all potential identifiers had been destroyed.

7.2.1.1 Covariates

The independent variables used includes socio-demographic characteristics (age, region, place of residence, education, religion, marital status), socio-cultural factors (decision making ability, wealth index, stigma, circumcision), sexual behaviour characteristics (number of sexual partners in the last 12 months, Had any STI in the last 5 years, age at first sex).

Principle Component Analysis (PCA) [48] was used to generate the stigma, media exposure and the ability of decision making. Stigma is defined as an attribute or label that sets a person apart from others and links the labeled person to undesirable characteristics [36]. Stigma related to AIDS has been defined as "the prejudice, discounting, discrediting, and discrimination that are directed at people perceived to have AIDS" [64]. A stigma index was created based on responses to the questions "willing to care for relative with AIDS", "person with AIDS allowed to continue" and "would buy vegetables from vendor with AIDS ". Based on factor scores, respondents were classified as having low, medium or high HIV-related stigma. A media exposure index was also computed using PCA based on responses to questions posed on the frequency of watching television, the frequency of listening to radio, and the frequency of reading newspapers. The respondents were then classified as having low, medium or high media exposure. The decision making index was computed based on the respondents answer to the questions: Final say on own health care, final say on making large household purchases, final say on making household purchases for daily need, final say on visits to family or relatives, final say on food to be cooked each day, and final say on deciding what to do with money husband earns. The decision making index was a trichotomous variable with levels independent, consults and subservient.

7.2.2 Summary statistics

This subsection presents the HIV prevalence rate by categorical predictors for the data

Table 7: Percent of females aged 15-49 years and males aged 15-54 years who are HIV-positive, with p values for χ^2 test, according to selected Demographic, Social, Biological, and Behavioral Characteristics, 2005 Ethiopia DHS

Variable	**Female**		**Male**	
	percentage	HIV positive	percentage	HIV positive
Age	$p = 0.0011$		$p < 0.001$	
15-19	0.69	14	0.12	1
20-24	1.72	26	0.36	8
25-29	2.11	38	0.70	14
30-34	1.48	17	1.94	19
35-39	4.44	27	1.82	12
40-44	3.06	13	2.84	11
45-49	0.85	7	0.01	1
50-54	-	-	0.88	3
55-59	-	-	0.34	1
Place of residence	$p < 0.001$		$p < 0.001$	
Urban	7.73	102	2.61	36
Rural	0.65	40	0.64	34
Region	$p < 0.001$		$p = 0.0044$	
Tigary	2.56	9	1.97	8
Afar	3.25	5	2.21	4
Amhara	1.83	13	1.42	11
Oromia	2.23	21	0.41	4
Somali	1.3	2	0	0
Ben-Gumz	0.9	4	0	50
SNNP	0.1	2	0.37	3
Gambela	5.51	16	6.18	15
Harari	4.59	15	2.05	6
Addis Abeba	6.06	42	2.05	6
Dirw Dawe	4.36	13	1.73	3
Religion	$p < 0.001$		$p < 0.001$	
Protestant	0.96	17	0.38	13
Others	0.79	17	0.17	7
Orthodox	2.91	108	1.60	50
Education level	$p < 0.001$		$p = 0.0036$	
No education	1.03	49	0.77	17
Primary	2.45	40	0.52	17
Secondary	6.05	51	2.10	31
Higher education	1.05	1.65	5	
Marital status	$p < 0.001$		$p < 0.001$	
Never married	0.7	23	0.29	14
Married, only wife	1.59	64	1.24	47
Married, other wives	1.45	7	-	-
Separated	6.43	48	2.64	9
Wealth index	$p < 0.001$		$p < 0.001$	
poorer	1	57	0.29	4
Poorest	0.34	8	0.62	10
Middle	0.43	9	0.85	9
Richer	0.21	8	0.37	8
Richest	6.09	110	2.21	39
STI	$p = 0.7155$		$p = 0.7166$	
NO	2.72	2	0.92	70
YES	1.86	5140	0	0
Media exposure	$p < 0.001$		$p = 0.0105$	
Low	0.67	30	0.60	8
Medium	1.21	12	0.64	19

Table 7 – *Continued from previous page*

High	4.58	98	1.68	43
Stigma	$p < 0.001$		$p < 0.001$	
Low	5.66	72	1.71	36
Medium	3.02	35	1.23	25
High	0.61	10	0.24	9
Birth in last 5 years	$p = 0.6180$			
Yes	1.74	61	-	-
No	1.99	81	-	-
Total number of children who had died	$p = 0.2769$		$p < 0.001$	
Never had children	1.29	40	0.45	21
No child died	2.41	71	1.92	35
One child died	1.72	18	0.44	8
Two or more child died	1.84	13	0.70	6
Number of the partners in the past years	$p < 0.001$			
Never had sex	0.12	8	-	-
One partner	1.44	59	-	-
More than one partner	4.95	74	-	-
Decision making ability	$p = 0.0768$			
Independent	1.86	26	-	-
Consults	2.01	28	-	-
Subservient	0.72	16	-	-
Circumcision			$p < 0.001$	
NO	-	-	1.11	13
YES	-	-	0.9	57
Age at first sex			$p < 0.001$	
Not had sex	-	-	0.16	4
Less than 14 years	-	-	00.12	1
14-17 years	-	-	1.49	18
more than 18 years	-	-	1.21	46

Table 7 presents the HIV prevalence rate with p value for χ^2 test for the men and women data. One of the key demographic characteristics was age .The male in the 40-44 age groups are at highest prevalence rate (2.84%), the females in 35-39 are at highest prevalence rate (4.44%). In terms of regions, the males of Gambela are at highest prevalence rate (6.18%), followed by Addis Ababa (3.3%), Affar (2.21%) and Harari (2.05%) and the rest of the regions have less than (2%) prevalence rate. The Addis Ababa females have the highest prevalence rate (6.06%), followed by Harari (4.59%) and Dire Dawa (4.36%). All the other regions have prevalence rate less than (4%). HIV prevalence rate is higher in urban areas 2.62%, 7.73% for both males and females respectively. In terms of number of children have died, the HIV prevalence is highest amongst those with no children dead with (1.92%, 2.41%) for males and females respectively. In terms of education level the highest prevalence was found among those with highest education level followed by those with higher education for both males and females. In terms of wealth index both males and females, the highest prevalence are found among the richest (2.21%, 6.09%) respectively. All the others have prevalence rates 1% or below. The HIV prevalence rate is highest amongst the separated men (2.64%), followed by married (1.24%). The prevalence rate is very low among those never married. Also the separated females have highest prevalence rate. In terms of religion both males and females, the Orthodox have the highest prevalence

(1.6%, 2.91%) respectively, and the rest have prevalence less than 1%. In terms of media exposure, for both males and females the high prevalence are among those with high media exposure (1.68%, 4.58%) respectively. For both males and females HIV prevalence increase with lower HIV-related stigma. The uncircumcised males have the highest prevalence (1.11%). The highest prevalence among consults females (2.01%), followed by independent (1.86%) and very low for subservient females.

7.3 Results

In this section we present the results we obtained from the analysis which were in keeping with four different models, namely GLM, GLMM, CAR, and convolution model. For each model we calculated DIC and the effective number of parameters (PD) in order to select the best model. After that we computed the mean, standard error, the quantiles, and the 2.5% and 975% quantile confidence intervals for the best model. Besides all that, we also present in this section the maps for HIV prevalence in Ethiopia, the mean posterior, and the bound of 0.025, 0.975 quantile confidence limits for the best model. As was mentioned earlier the analysis was conducted separately for females and males, and for this reason we now present them separately.

7.3.1 Results of the female data

In this sub-section we present the results obtained from the female data

Table 8: Deviance Information Criteria and effective number of parameters for the Female data

	GLM	GLMM	CAR	Convolution
DIC	1080.91	1064.77	1080.91	1065.50
PD	30.25	35.98	30.25	36.23

Table 8 presents the DIC and PD for the different models we fitted and it demonstrates that the GLMM has the lowest DIC (1064.77) which means it is the best model.

Table 9 presents the results of the GLMM for the female data. The women aged 25 to 29 are at significantly higher prevalence of being HIV positive than women in the reference category aging from 15 to 19 years. Those women living in rural areas are half as likely to be HIV positive compared to those in urban areas. The women with primary education are nearly 1.28 times as likely to HIV positive as those with no

Table 9: Posterior mean, standard deviation and quantiles for the parameters in GLMM for female data

Parameters	mean	sd	2.5%quant	97.5%quant
Intercept	-6.72	1.09	-8.99	-4.71
Age				
15-19	Ref	Ref	Ref	Ref
20-24	0.04	0.38	-0.69	0.79
25-29	0.51	0.39	-0.23	1.28
30-34	0.04	0.44	-0.82	0.91
35-39	0.39	0.43	-0.43	1.24
40-44	-0.34	0.49	-1.29	0.62
45-49	-0.69	0.57	-1.84	0.39
Religion				
Protestant	Ref	Ref	Ref	Ref
Orthodox	0.0008	0.31	-0.59	0.63
Others	-0.85	0.38	-1.59	-0.11
Residence				
Rural	Ref	Ref	Ref	Ref
Urban	0.73	0.34	0.07	1.42
Education Level				
No Education	Ref	Ref	Ref	Ref
Higher Education	-1.52	0.77	-3.19	-0.17
Primary	0.25	0.26	-0.27	0.76
Secondary	0.10	0.29	-0.48	0.69
Martial Status				
Never married	Ref	Ref	Ref	Ref
Married, only wife	-0.15	0.37	-0.86	0.58
Married, other wives	0.31	0.55	-0.81	1.37
Separated	0.73	0.38	0.002	1.48
Wealth Index				
Poorest	Ref	Ref	Ref	Ref
Middle	0.43	0.51	-0.56	1.43
poorer	0.10	0.53	-0.95	1.14
Richer	0.18	0.53	-0.85	1.21
Richest	1.27	0.51	0.29	2.28
Media Exposure				
High	Ref	Ref	Ref	Ref
Low	-0.26	0.33	-0.90	0.38
Medium	-0.79	0.35	-1.51	-0.12
Number of partners in the past years				
Never had sex	Ref	Ref	Ref	Ref
One partner	1.99	0.48	1.07	2.97
More than one partner	3.12	0.49	2.17	4.12
Total number of children who had died				
Never had children	Ref	Ref	Ref	Ref
No child died	-0.31	0.28	-0.85	0.24
One child died	-0.47	0.37	-1.21	0.25
Two or more child died	-0.23	0.43	-1.08	0.59
Stigma				
High	Ref	Ref	Ref	Ref
Low	0.93	0.32	0.33	1.57
Medium	0.65	0.31	0.06	1.26
STI				
No	Ref	Ref	Ref	Ref
Yes	-0.93	0.83	-2.42	0.87
Decision Making ability				
Independent	Ref	Ref	Ref	Ref
Consults	0.28	0.21	-0.12	0.69
Subservient	0.22	0.27	-0.32	0.75

education. Wealth is positively and monotonically related to being HIV positive, with those in richest quintile being 3.5 times more likely to be infected than those in the poorest quintile. Regarding marital status, separated women are twice more likely to be HIV positive as those never married. Those women with more than one partner are 22 times as likely to be HIV positive as those never had sex. There are higher prevalence for women who are at low stigma compared with women who report that they are at high stigma. The risk of being HIV positive does not vary according to reported decision making.

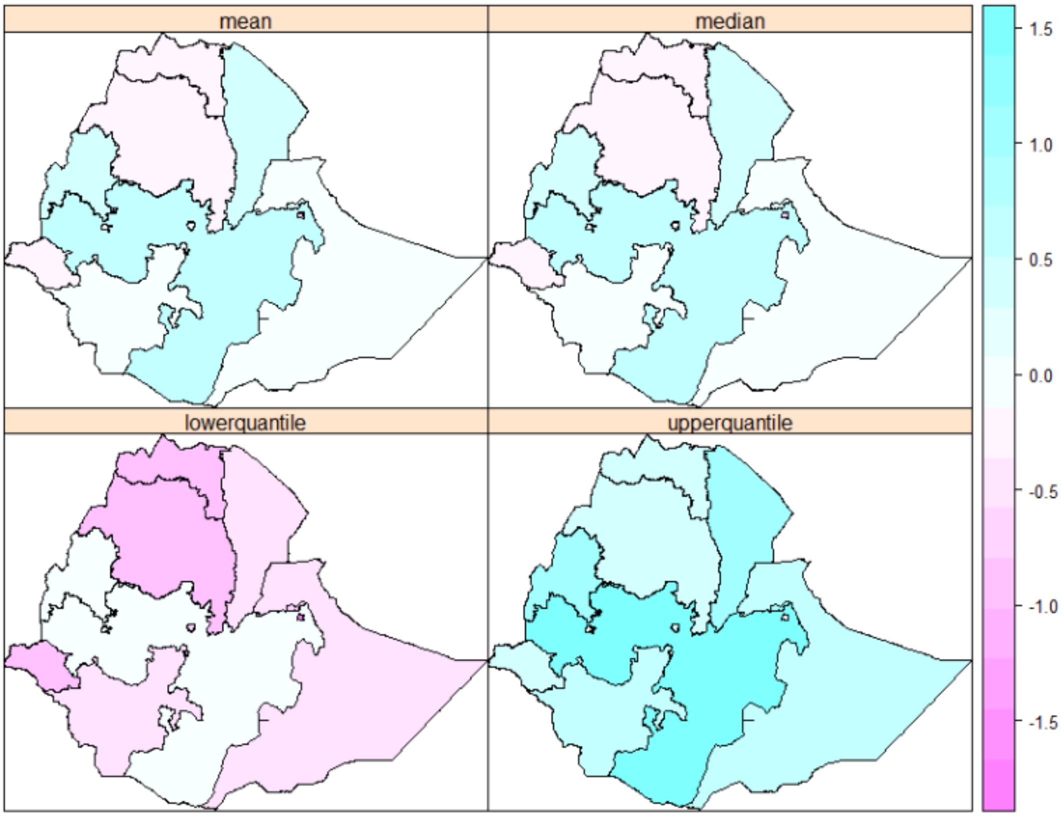


FIGURE 2: Maps of posterior mean, median, lower quantile, and upper quantile and for the female data

Figure 2 represent the maps of posterior mean, median, lower quantile, and upper quantile and for the Female data .

7.3.2 Results of the male data

In this subsection we present the results obtained from the male data

TABLE 10: Deviance Information Criteria and effective number of parameters for the male data

	GLM	GLMM	CAR	Convolution
DIC	653.52	653.50	653.52	653.50
PD	28.84	28.84	28.84	28.84

Table 10 presents the DIC and PD for the different models which we fit and it demonstrates that the GLMM has the lowest one (653.50) which mean it is the best model.

Table 11 present the results of the GLMM model for the male data, we note the inverted U-shaped relationship between age and infection with HIV, such those in ages ranging from 25 to 44 years are the most likely to be infected, with risk peaking for

Table 11: Posterior mean, standard deviation and quantiles for the parameters in GLMM for male data

Parameters	mean	sd	2.5%quant	97.5%quant
Intercept	-8.55	1.22	-11.22	-6.39
Age				
15-19	Ref	Ref	Ref	Ref
20-24	1.89	1.08	0.04	4.27
25-29	2.53	1.09	0.65	4.93
30-34	2.95	1.10	1.05	5.38
35-39	2.76	1.12	0.82	5.23
40-44	3.06	1.14	1.08	5.56
45-49	0.49	1.49	-2.45	3.39
50-54	2.19	1.25	-0.05	4.89
55-59	1.26	1.50	-1.69	4.19
Residence				
Rural	Ref	Ref	Ref	Ref
Urban	0.86	0.49	-0.07	1.87
Education Level				
No Education	Ref	Ref	Ref	Ref
Higher Education	-0.40	0.66	-1.74	0.84
Primary	-0.03	0.39	-0.82	0.75
Secondary	0.26	0.47	-0.67	1.18
Martial Status				
Never married	Ref	Ref	Ref	Ref
Married or living together	0.49	0.52	-0.53	1.50
separated	1.25	0.55	0.16	2.31
Wealth Index				
Poorest	Ref	Ref	Ref	Ref
Middle	-0.16	0.49	-1.13	0.80
poorer	-1.15	0.61	-2.43	-0.01
Richer	0.50	0.53	-1.55	0.52
Richest	0.85	0.61	-1.73	0.66
Circumcision				
Yes	Ref	Ref	Ref	Ref
No	1.81	0.41	0.98	2.59
Media Exposure				
High	Ref	Ref	Ref	Ref
Low	-0.98	0.51	-2.02	-0.01
Medium	-0.37	0.36	-1.08	0.33
Age at first sex				
Not has sex	Ref	Ref	Ref	Ref
Less than 14 years	1.02	1.17	-1.53	3.07
14-17 years	1.14	0.63	-0.03	2.45
more than 18 years	1.00	0.61	-0.11	2.27
Total number of children who had died				
Never had children	Ref	Ref	Ref	Ref
No child died	-0.19	0.43	-0.99	0.69
One child died	-0.49	0.55	-1.58	0.59
Tow or more child died	-0.50	0.63	-1.75	0.71
Stigma				
High	Ref	Ref	Ref	Ref
Low	0.75	0.42	-0.04	1.59
Medium	0.85	0.41	0.07	1.69

those in ages 45 to 49. All ages groups are more likely to be infected than the reference group(ages 15-19 years). Men living in urban areas are 2.4 times as likely to be HIV positive as those living in rural areas. The probability of being HIV positive is positively, but no significantly, related to the number of one's children that has died. Those with secondary education are 1.3 times as likely to HIV positive as those with no education. Wealthiest men are 2.33 times as likely to HIV positive as those in poorest quintile. Circumcision factor has been found highly significant for acquiring HIV, men who are not circumcised are 6 times as likely to HIV positive as those who are circumcised.

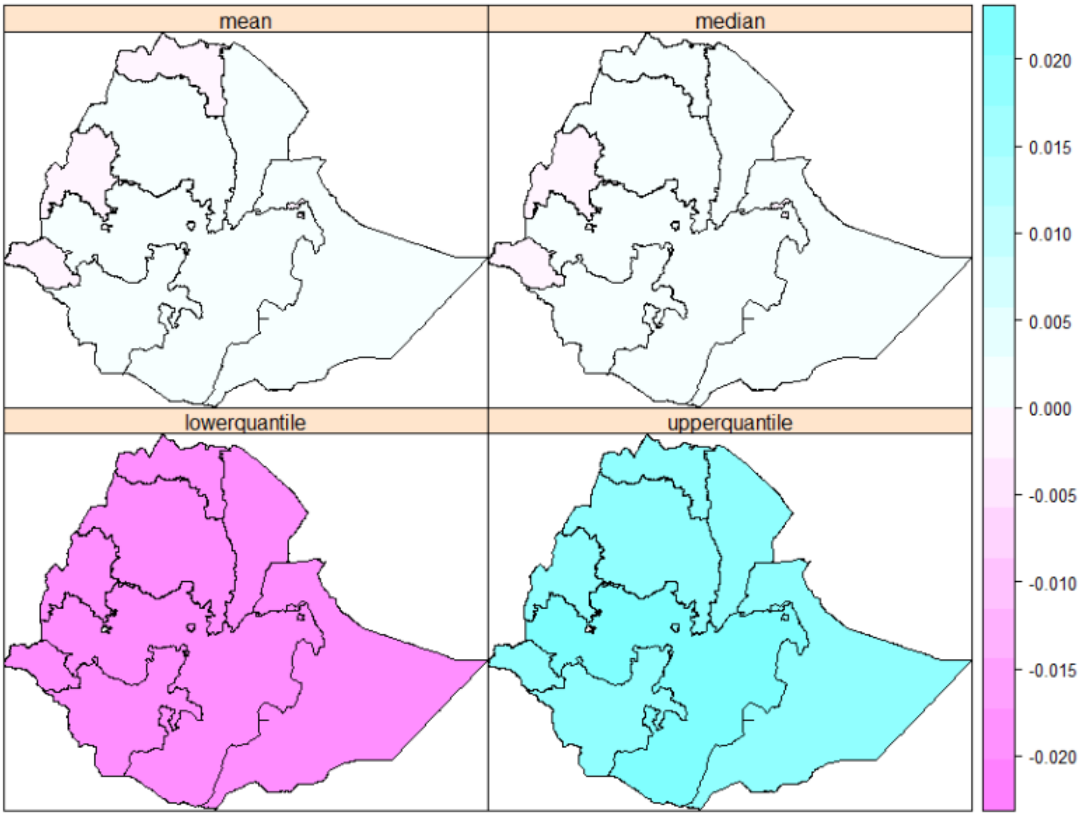


Figure 3: Maps of posterior mean, median, lower quantile, and upper quantile and for the male data

Figure (3) represent the maps of posterior mean, median, lower quantile, and upper quantile and for the male data

7.4 Discussion and Conclusions

The study made use of the (EDHS 2005) data. Our analysis was carried out separately for the male and female populations. The main objectives of this study were to map HIV prevalence in Ethiopia and use spatial modeling by using INLA approach. In order to achieve the objectives, we fitted four different models, namely GLM, GLMM, CAR, and convolution model. For each model we computed DIC, in order to select the best model, the GLMM, has the lower DIC for both males and females. We drew the maps of Ethiopian's HIV prevalence and poster mean, with 0.025 quantiles and 0.975 for the GLMM. Our findings are in the line with those contained in the literature [127]

The findings for both men and women demonstrate the urban residence were found at higher prevalence . Wealth was positively related to risk for HIV for both men and women, the wealthiest are most likely to have high prevalence HIV [68]. Separated men and women were be at higher risk to be HIV positive. The outstanding biological factor associated with being HIV positive for men was circumcision, uncircumcised men were 6 times more likely to HIV positive compared to circumcised men. Other study conducted in Sub Saharan Africa indicates the uncircumcised men are 80% to contract the disease [154]

As shown by our study, INLA is a powerful inference approach for Bayesian hierarchical models. The advantages of the INLA approach are not only computational; it also allows for greater automation and parallel implementation. In practice, INLA can be used as a black box to analyze latent Gaussian models. The main drawback of the INLA approach is that its computational cost is exponential with respect to the number of hyperparameters m. In most application m is small, but applications where m goes up to 10 do exist. However, this situation is not always as severe as it appears at first glance. We show in our study that, by using INLA, one can directly compute very accurate approximation to the posterior marginals.

The findings of this study demonstrate that the GLMM is the best model fitting the data. This finding in turn indicates that the GLMM enjoying increased popularity because of its ability to model the correlated observations. INLA allows the computation of many Bayesian GLMMs in a reasonable amount of time, enabling an extensive comparison of different models and prior distributions.

Chapter 8

Discussion and Recommendations

8.1 Summary

The study has focused on statistical methods for analysing complex survey design. More specifically, we have been concerned about statistical methods for binary outcome data, which is gaining more and more attention in the applied statistics field. In the analysis of complex survey data, such as stratified multi-stage cluster samples,unequal probability selection, and ignoring design effects such as (clustering and stratification), usually leads to biased estimates of parameters, standard error, wide confidence intervals and misleading statistical tests. The challenge in analysing the complex data is how to incorporate the design aspect in the estimation procedure.

In this study, we try to give more insight into statistical methods for analysing data in a complex survey design, when the outcome variable is binary. These methods have been presented with in-depth analysis of a practical data set with a binary outcome. The study used (EDHS 2005) data and the outcome variable is the HIV serostatus (positive or negative). The explanatory variables are classified as demographic, socio-economic, socio-cultural, behavioural and proximate determinants of HIV risk. The analysis was done separately for both male and female because the biological and social circumstances associated with the transmission of HIV differs by sex.

In order to analyse complex survey design we have applied three models; survey logistic regression, multilevel models and spatial models. Survey logistic regression was used to model the binary outcome variable which is HIV serostatus, and as mentioned survey logistic takes into account the sample design. For each parameters we calculate the OR, AOR and 95% confidence interval. The findings agreed with the literature. A multilevel model and a Bayesian multilevel model were used to model the data. We used a frequentist multilevel model first because it takes into account multistage sampling, such as what we have in our data: for example individuals within households nested within regions. We then used the Bayesian approach, and finally spatial models. The four different fitted models were Generalized Linear Models (GLM), Generalized Linear Mixed Models (GLMM), Conditional Auto-Regressive (CAR) model, and Convolution model. We used INLA to estimate the parameters of the models. In order to select the best model we used Deviance Information Criteria (DIC), and established that GLMM is the best model because it has the smallest DIC. We mapped HIV prevalence in Ethiopia.

HIV testing was successfully conducted for 83% of eligible women and 76% of eligible men. For both sexes combined, the coverage was 80%. Refusals were the most important reason for non-response on the HIV testing component of the survey for both women (13%) and men (17%). The findings indicated that the national HIV prevalence in Ethiopia was 1.4%. HIV among women is around 2%, while for men it is just 1%. HIV prevalence levels rise with age, peaking among women in their late 30s and among men in their early 40s, indicating that young women are more vulnerable to HIV infection compared with young men. Urban residents have a significantly higher risk of HIV infection than rural residents (6% versus 1%). The risk of HIV infection among rural women and men is almost identical, while urban women are more than three times as likely as urban men to be infected. HIV prevalence levels are highest in Gambela and Addis Ababa. Other regions in which HIV prevalence exceeds the national average include Harari, Dire Dawa, Affar, Tigray and Amhara. HIV infection levels increase proportionately with education for both women and men, and are markedly higher among those with a secondary or higher education compared to those with no education. Wealthy women and men were found to have a higher risk for contracting HIV compared to the poor. The outstanding biological factor associated with being HIV-positive for men was circumcision status. The findings indicated that uncircumcised men were at higher risk compared to those who were circumcised. Our findings add to the large body of research indicating that circumcision has a protective effect against HIV infection among men. The study provides a closer look at the spread of HIV infection in Ethiopia, specifically with regard to the different regions. The findings are useful in terms of identifying population of higher-prevalence and higher risk, and for strengthening prevention, care and support, and treatment programmes

For the Ethiopian DHS, the sample size was designed to represent all administrative units in Ethiopia. The sample was selected from each of Ethiopia's eleven geographic/administrative regions: nine regional states (Tigray, Affar, Amhara, Oromia, Somali, Benishangul-Gumuz, SNNP, Gambela and Harari) and two city administrations (Addis Ababa and Dire Dawa). The sampling frame used was from the Population and Housing Census results. In general, a DHS sample is stratified, clustered and selected in two stages. In the 2005 EDHS a representative sample of approximately 17,817 households from 540 clusters was selected. The sampling was done in two stages. In the first stage, 540 clusters from which 145 urban and 395 rural clusters were selected. In the second stage, the representative households were selected. Furthermore, all women age 15-49 and all men age 15-59 who were either permanent residents of the selected households or visitors who stayed in the household the night before the survey were eligible to be interviewed. Because the sample is not self weighting at the national level and to ensure statistical reliability, for this study the sampling weight was used for each analysis used.

8.2 Limitations of the research and future work direction

This section summarises the limitations concerning the data sets which were used in this study. The methodological limitations of this research is discussed and recommended areas for future research, based on these limitations, are made.

8.2.1 The Data Limitation

The study mainly uses secondary data from (EDHS 2005). A major limitation is that our selection of EDHS clusters within a 15 kilometer radius around the Antenatal Clinic (ANC) surveillance sites was based on the assumption that 15 kilometers is a reasonable maximum distance which women would travel for ANC care, yet it may not reflect a true catchment area for an ANC site. A previous analysis of ANC attendees at sentinel surveillance sites in Uganda showed that these distances correspond reasonably well with the actual administrative areas where clients were living . Moreover, the distance women are prepared to travel for ANC may vary from one region to another and may be different for urban and rural areas. For a more meaningful comparison, the catchment areas should be defined by examining the ANC client records for each surveillance site.

Another source of bias may be due to displacement of (Global Positioning System) GPS coordinates of EDHS clusters (five kilometers in rural areas and two kilometers in urban areas) to protect the confidentiality of the survey participants. However, because the displacement was random and the results from individual ANC catchment areas were aggregated up to the national level, any effect of such bias is expected to be small. The comparison between the ANC surveillance survey and the EDHS may also be affected by differences in HIV testing protocols and differences in the definitions of urban and rural areas.

In the analysis of nonresponse bias in the EDHS, one limitation is that the estimates are only adjusted to the extent that the socio-demographic and behavioural characteristics included in the analysis are correlated with the risk of HIV infection. Another limitation is that the adjustments for respondents not interviewed and not tested are based on limited information available from the household questionnaires. Moreover, the adjustments for nonresponse do not account for any bias due to exclusion of population members not living in households, such as those living on the street or in institutions (e.g., prisons, boarding schools, military barracks, refugee camps, and brothels).

In summary, the HIV prevalence estimate derived from the ANC surveillance survey appears to have overestimated HIV prevalence among women in the general population. The EDHS estimate compared well with the ANC surveillance estimate when the comparison was restricted to women residing within the catchment areas of the ANC surveillance sites. Patterns by age and urban/rural residence point to possible sources of bias in the ANC estimates. On the other hand, the analysis of nonresponse suggests bias in EDHS estimates due to significantly higher predicted HIV rates among the non-responders. However, this bias did not have any significant effect on the national estimates of HIV prevalence.

Overall, the EDHS provided high-quality, reliable, representative national prevalence estimates of HIV, and its associated characteristics and risk factors. This data is useful for: identifying geographic regions with elevated rates of HIV, and higher-risk and vulnerable populations; providing a better understanding of risky and protective sexual behaviours; assessing availability and access to health services; and planning for prevention, care, and treatment programmes. Data from the EDHS is also useful for calibrating prevalence estimates from surveillance systems and improving the accuracy of national estimates. However, to provide nationally representative trend data on HIV prevalence and associated risk factors, there is a need to carry out similar surveys at regular intervals.

8.2.2 Methodological limitations and future direction

Three methods were applied when analyzing the data in the study (survey logistic regression, multilevel, spatial models). And each method has its strengths and weaknesses. The weaknesses therefore, can be considered limitations when using these methods.

Avenues for further work in this research are discussed in this section. In the future one can extends survey logistic models so that they can accommodates the Bayesian approach by incorporating the sample weights. Another possibility is a comparison of the methods used to analyse the complex survey data by computer simulation to show which model is the effective. One can also fit the models and estimates to the parameters using INLA and MCMC and compare both methods. Furthermore, one can apply a multilevel approach to meta analysis, because meta analysis can be viewed as a special case of multilevel analysis. It is possible our future research may include the development of statistical techniques to handle complex survey design, missing data, and measurement error in data from longitudinal surveys and household surveys. Another future application in multilevel models is applying the bootstrap method in the framework of hierarchical linear modeling[70, 71]. Small area estimation from unit-level for complex survey data with generalised linear or non-linear mixed models, are also possible direction for future work. The theory of generalised linear mixed models relies largely on the normality of random effects [105]. In real life, however, the normality of area effects is not always justified. There are some research activities in the Empirical Best Linear Unbiased Predictor (EBLUP) estimation under non-normal area effects, but to the best of our knowledge, no paper has yet been published. Thus, there is obviously a lot of work to be done concerning the robustness of EBLUP estimators with respect to the non-normality and the possibility of applying distributions other than normal. The biggest limitation to using Bayesian methods, is the time required to run models when using MCMC methods of estimation, especially when data sets are huge or there are many variables to be included in the model. Another limitation is that models should be run with different priors on the parameters, which also can increase the time taken to draw final conclusions from Bayesian models.

Fitting Bayesian multilevel models using MCMC techniques is computer intensive especially for the huge data sets encountered in health care research. INLA can be used as an alternative technique. INLA provides marginal posterior samples for the parameters of interest in a fraction of the time that MCMC techniques would require. Some of the analysis in this study performed used INLA, resulting in models which took days to fit being realised in a matter of minutes with posterior samples from INLA are as accurate as those from MCMC techniques. Another advantage of INLA is that there is less programming effort compared to MCMC techniques. All further research will target INLA in implementing Bayesian multilevel models.

A very important step in modeling data is to check various features of the fitted model. This usually involves checking goodness of fit to the model, checking the model assumption, and detecting possible influential observations. To the best of our knowledge, little work has been done on model checking and model diagnostics for multilevel models, so this can also be considered as a future direction of this thesis. Another future direction could be the use of small area estimation via M-quantile regression [30, 131, 142].

Appendix A

The posterior distribution of the Variance component and the fixed effects for Bayesian multilevel model

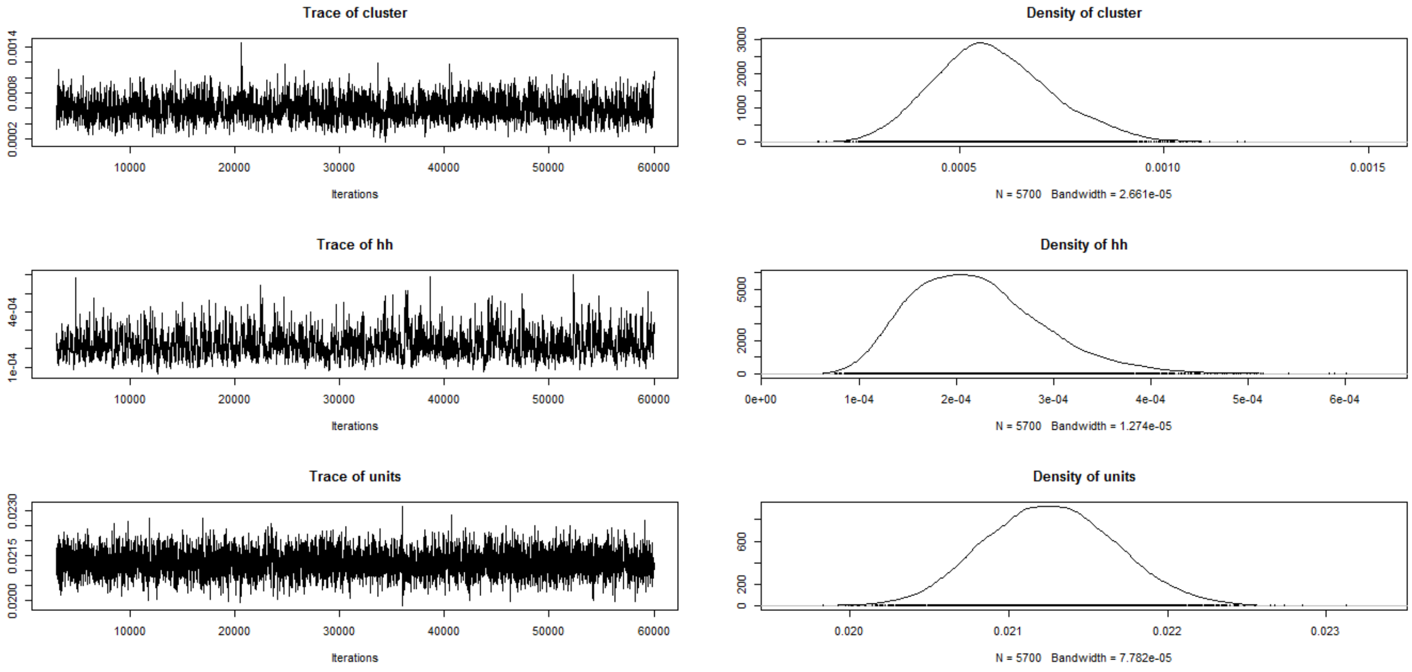


FIGURE 4: The posterior distribution of the variance component for the female data

Figure (4) showed the posterior distribution of the random effect part of the model, based on the analysis of 60000 iterations, 2000 burn in period, and with thin=10 in MCMCglmm.

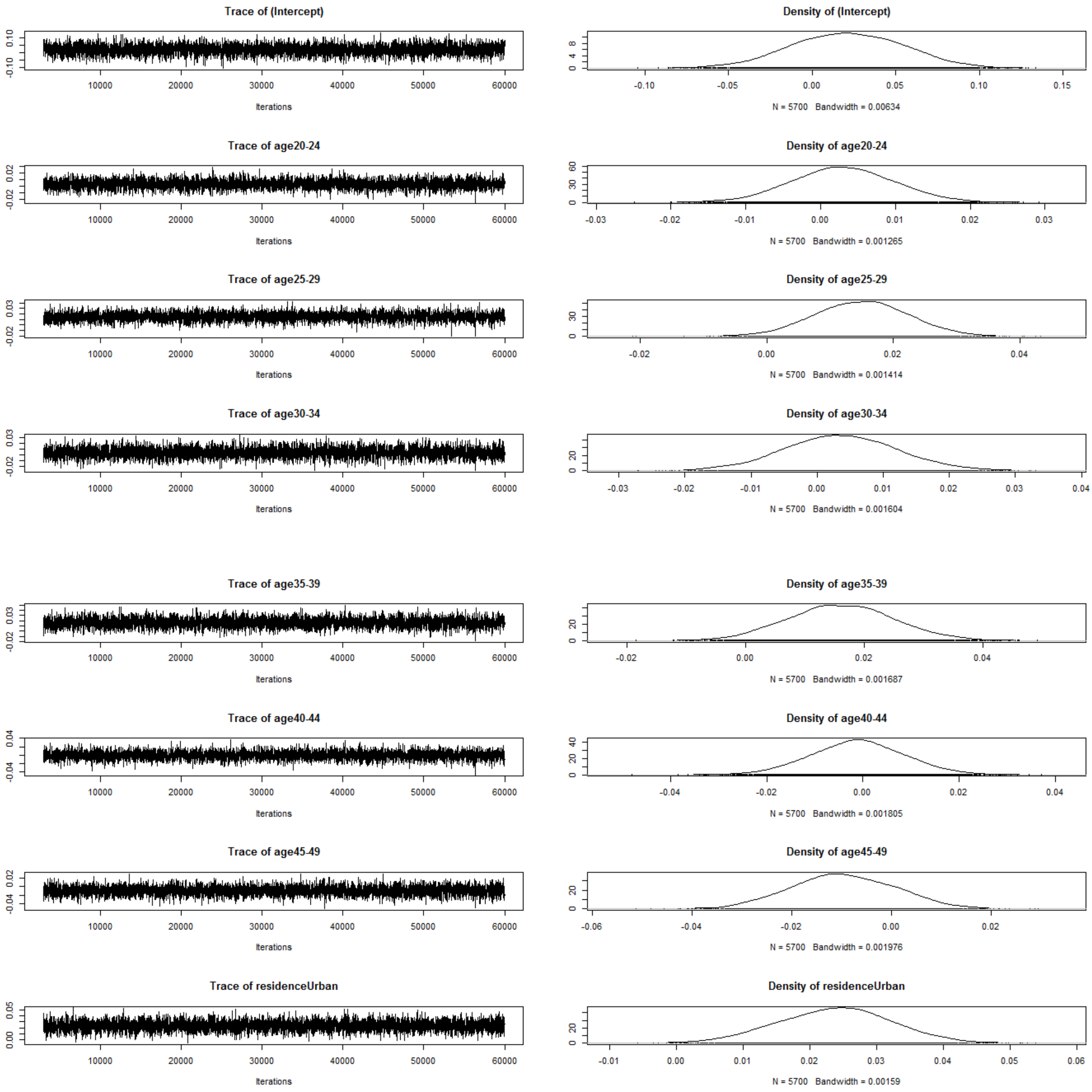
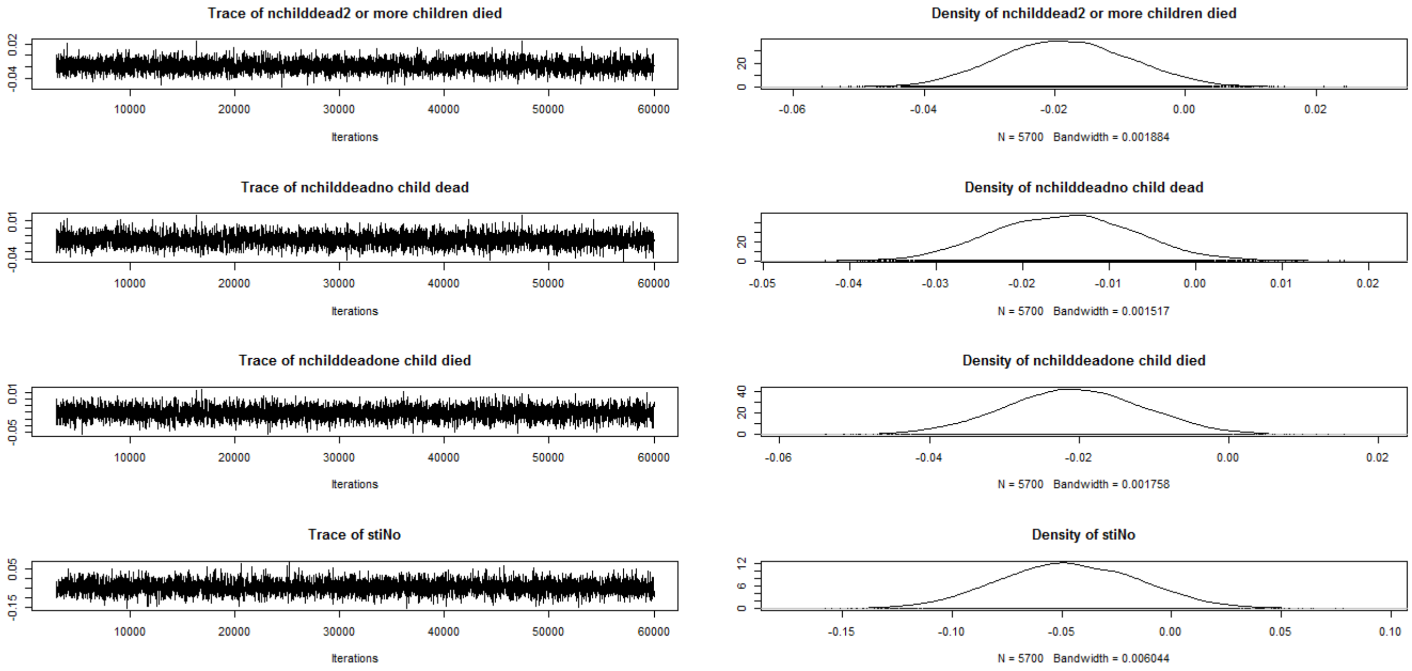
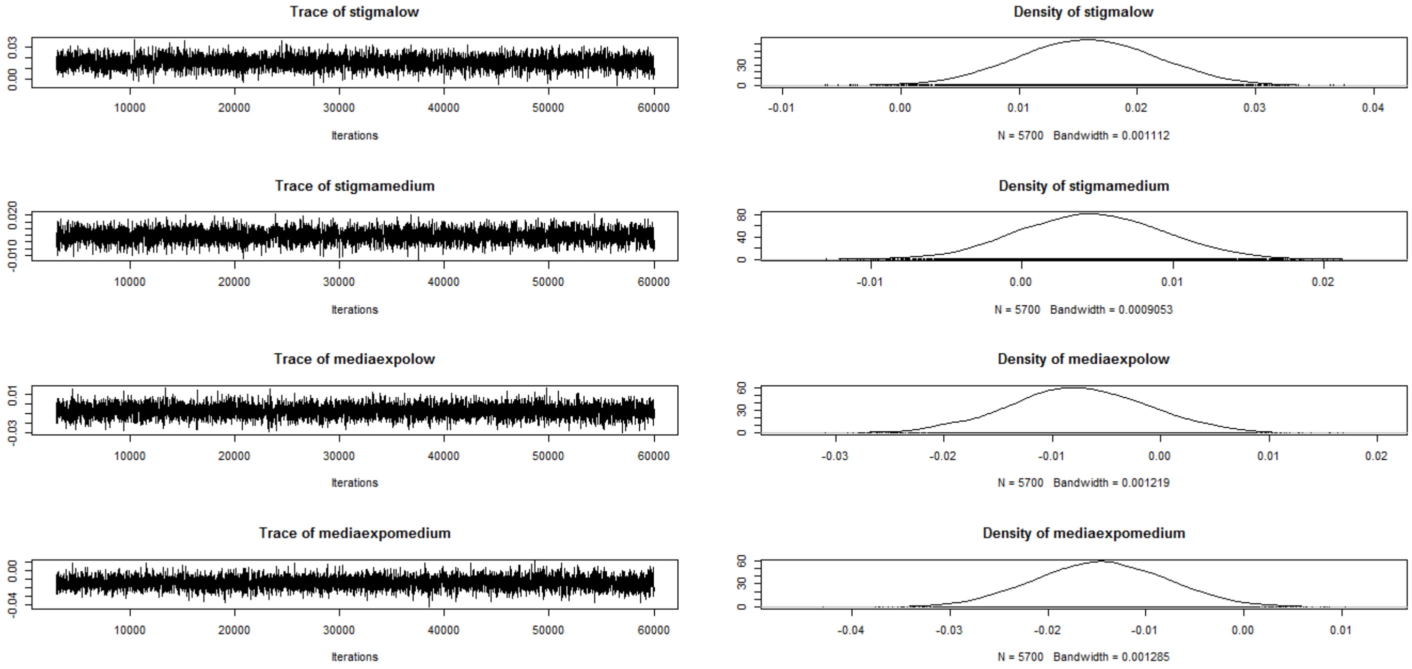
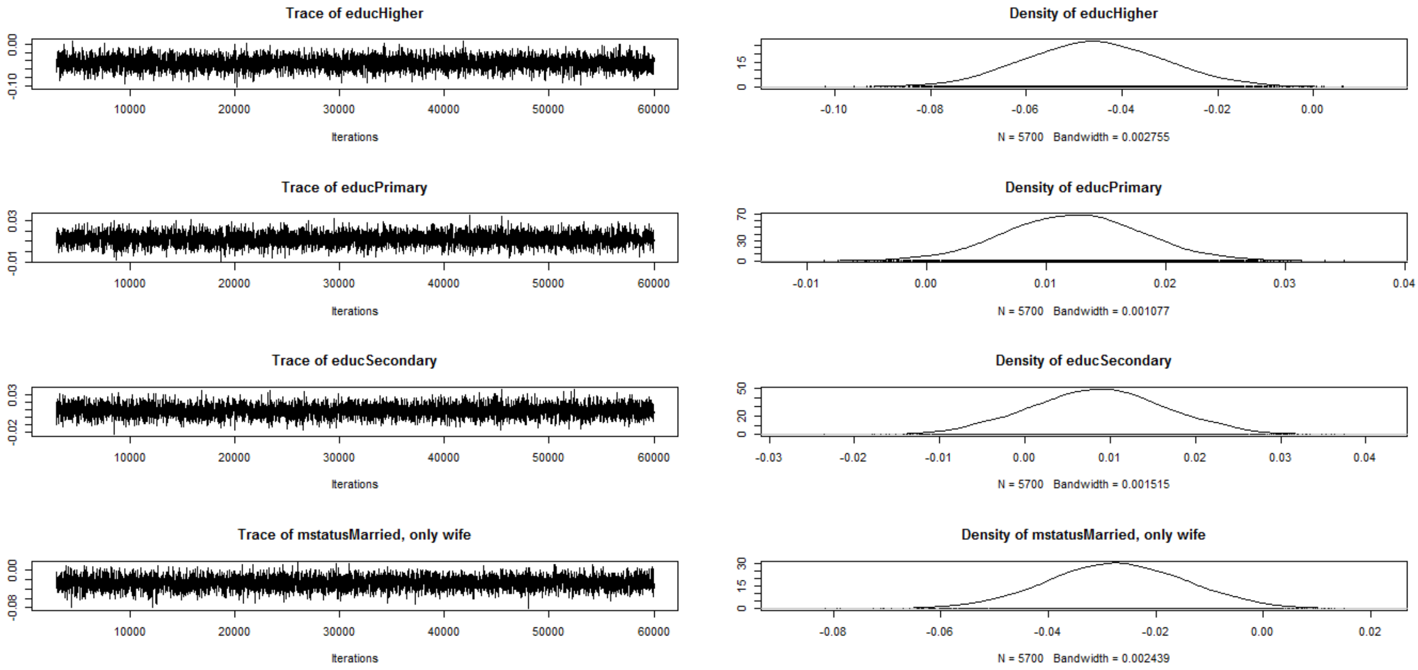
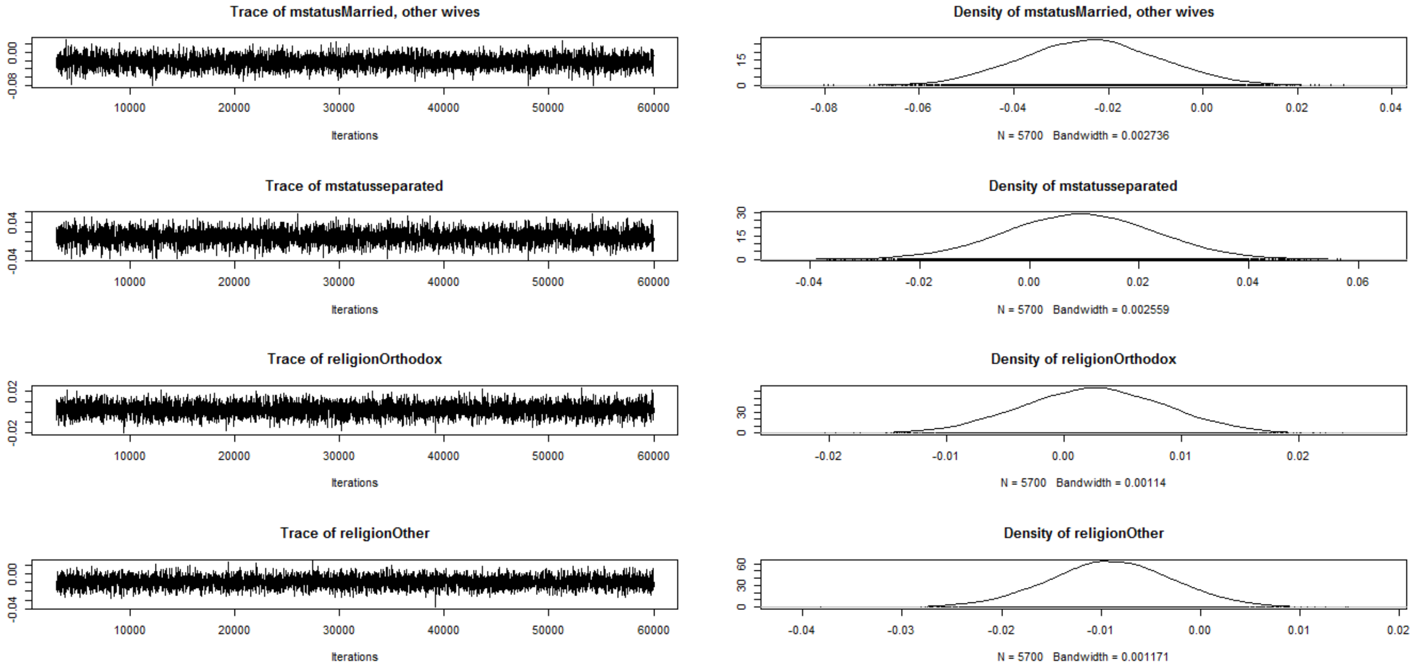


Figure 5 showed the posterior distribution of the fixed effect part of the model, based on the analysis of 60000 iterations, 2000 burn in period, and with thin=10 in MCMCglmm.


Trace of nchilddead2 or more children died
-0.04
0.02
10000
20000
30000
40000
50000
60000
Iterations
Density of nchilddead2 or more children died
0
20
-0.06
-0.04
-0.02
0.00
0.02
N = 5700 Bandwidth = 0.001884
Trace of nchilddeadno child dead
-0.04
0.01
10000
20000
30000
40000
50000
60000
Iterations
Density of nchilddeadno child dead
0
20
-0.05
-0.04
-0.03
-0.02
-0.01
0.00
0.01
0.02
N = 5700 Bandwidth = 0.001517
Trace of nchilddeadone child died
-0.05
0.01
10000
20000
30000
40000
50000
60000
Iterations
Density of nchilddeadone child died
0
20
40
-0.06
-0.04
-0.02
0.00
0.02
N = 5700 Bandwidth = 0.001758
Trace of stiNo
-0.15
0.05
10000
20000
30000
40000
50000
60000
Iterations
Density of stiNo
0
6
12
-0.15
-0.10
-0.05
0.00
0.05
0.10
N = 5700 Bandwidth = 0.006044


Trace of stigmalow
0.00
0.03
10000
20000
30000
40000
50000
60000
Iterations
Density of stigmalow
0
30
-0.01
0.00
0.01
0.02
0.03
0.04
N = 5700 Bandwidth = 0.001112
Trace of stigmamedium
-0.010
0.020
10000
20000
30000
40000
50000
60000
Iterations
Density of stigmamedium
0
40
80
-0.01
0.00
0.01
0.02
N = 5700 Bandwidth = 0.0009053
Trace of mediaexpolow
-0.03
0.01
10000
20000
30000
40000
50000
60000
Iterations
Density of mediaexpolow
0
30
60
-0.03
-0.02
-0.01
0.00
0.01
0.02
N = 5700 Bandwidth = 0.001219
Trace of mediaexpomedium
-0.04
0.00
10000
20000
30000
40000
50000
60000
Iterations
Density of mediaexpomedium
0
30
60
-0.04
-0.03
-0.02
-0.01
0.00
0.01
N = 5700 Bandwidth = 0.001285


Trace of educHigher
0.00
-0.10
10000 20000 30000 40000 50000 60000
Iterations
Density of educHigher
15
0
-0.10 -0.08 -0.06 -0.04 -0.02 0.00
N = 5700 Bandwidth = 0.002755
Trace of educPrimary
0.03
-0.01
10000 20000 30000 40000 50000 60000
Iterations
Density of educPrimary
70
30
0
-0.01 0.00 0.01 0.02 0.03 0.04
N = 5700 Bandwidth = 0.001077
Trace of educSecondary
0.03
-0.02
10000 20000 30000 40000 50000 60000
Iterations
Density of educSecondary
50
20
0
-0.03 -0.02 -0.01 0.00 0.01 0.02 0.03 0.04
N = 5700 Bandwidth = 0.001515
Trace of mstatusMarried, only wife
0.00
-0.08
10000 20000 30000 40000 50000 60000
Iterations
Density of mstatusMarried, only wife
30
15
0
-0.08 -0.06 -0.04 -0.02 0.00 0.02
N = 5700 Bandwidth = 0.002439


Trace of mstatusMarried, other wives
0.00
-0.08
10000 20000 30000 40000 50000 60000
Iterations
Density of mstatusMarried, other wives
15
0
-0.08 -0.06 -0.04 -0.02 0.00 0.02 0.04
N = 5700 Bandwidth = 0.002736
Trace of mstatusseparated
0.04
-0.04
10000 20000 30000 40000 50000 60000
Iterations
Density of mstatusseparated
30
15
0
-0.04 -0.02 0.00 0.02 0.04 0.06
N = 5700 Bandwidth = 0.002559
Trace of religionOrthodox
0.02
-0.02
10000 20000 30000 40000 50000 60000
Iterations
Density of religionOrthodox
30
0
-0.02 -0.01 0.00 0.01 0.02
N = 5700 Bandwidth = 0.00114
Trace of religionOther
0.00
-0.04
10000 20000 30000 40000 50000 60000
Iterations
Density of religionOther
60
30
0
-0.04 -0.03 -0.02 -0.01 0.00 0.01 0.02
N = 5700 Bandwidth = 0.001171

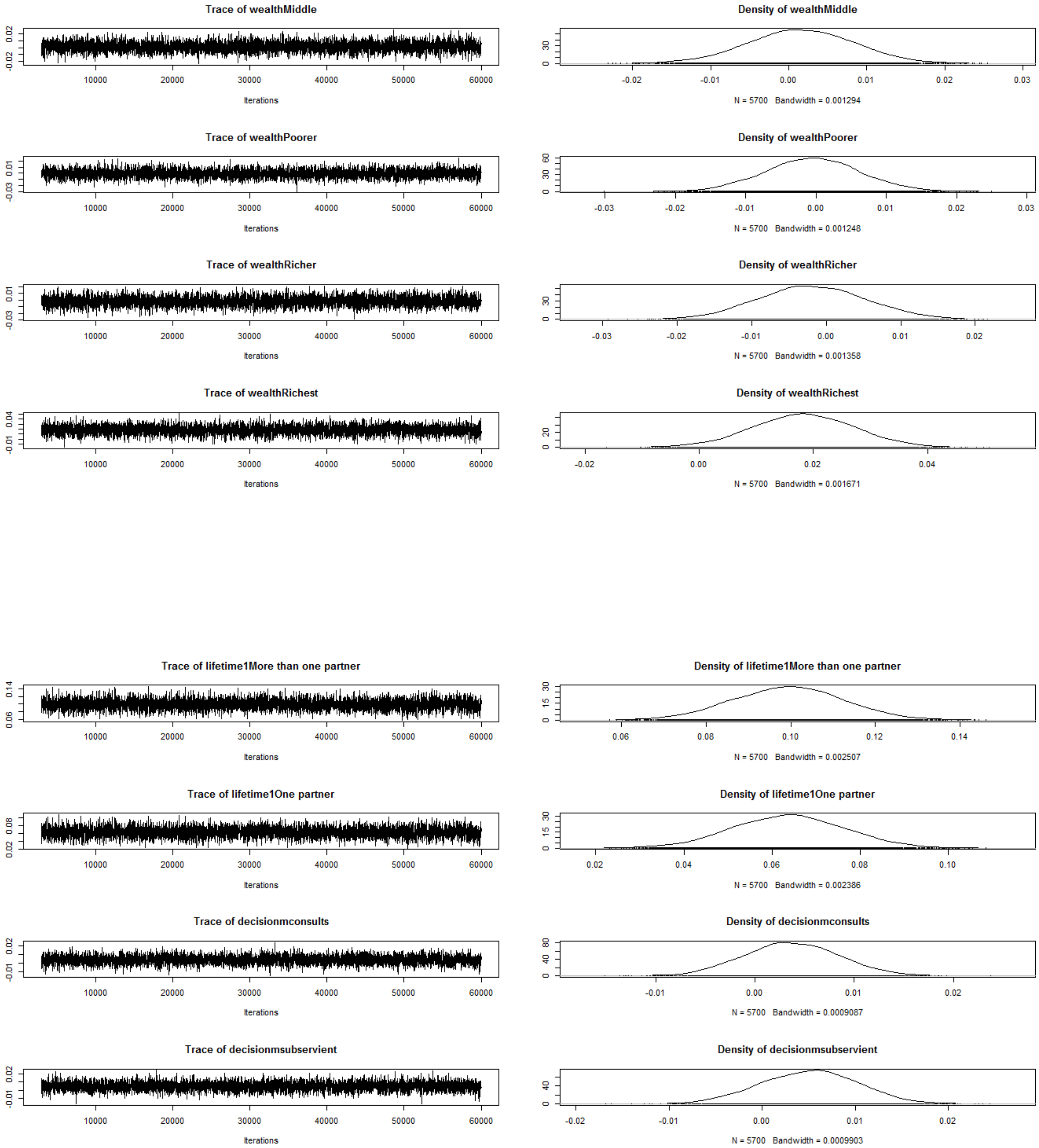


Figure 5: The posterior distribution for the female data .

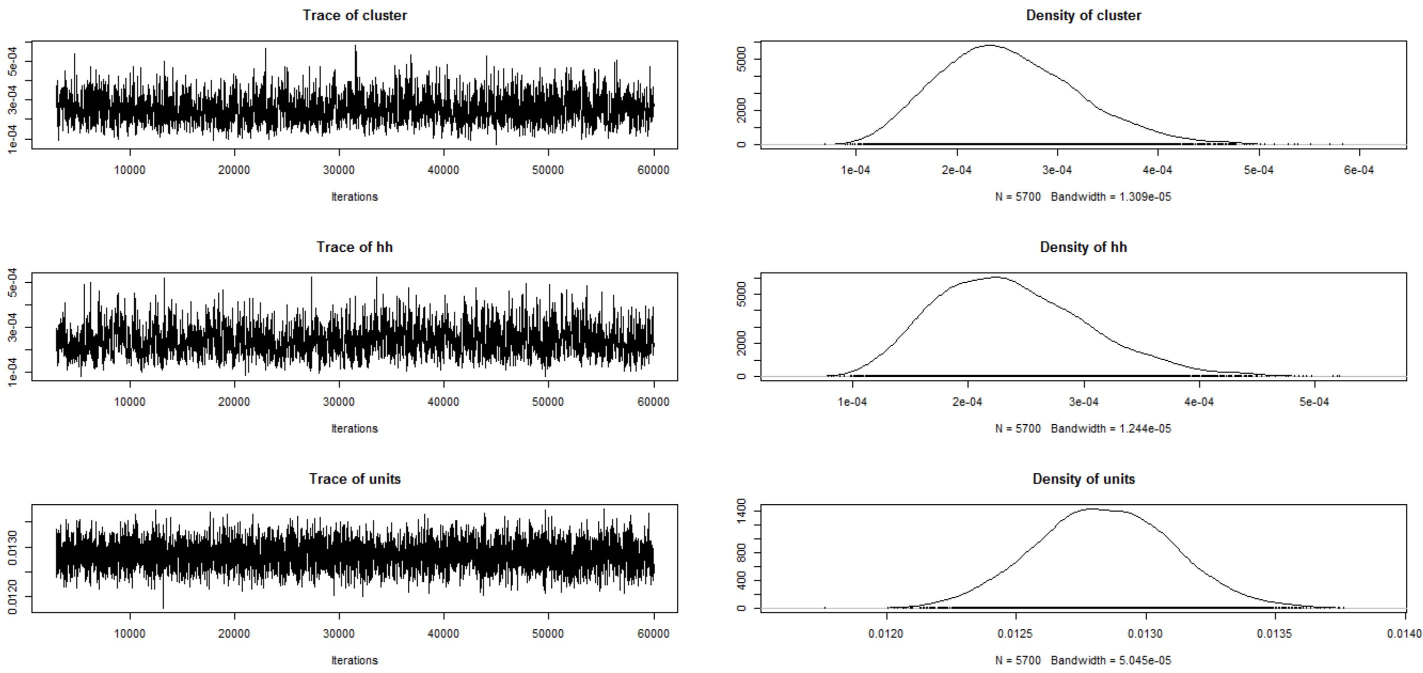
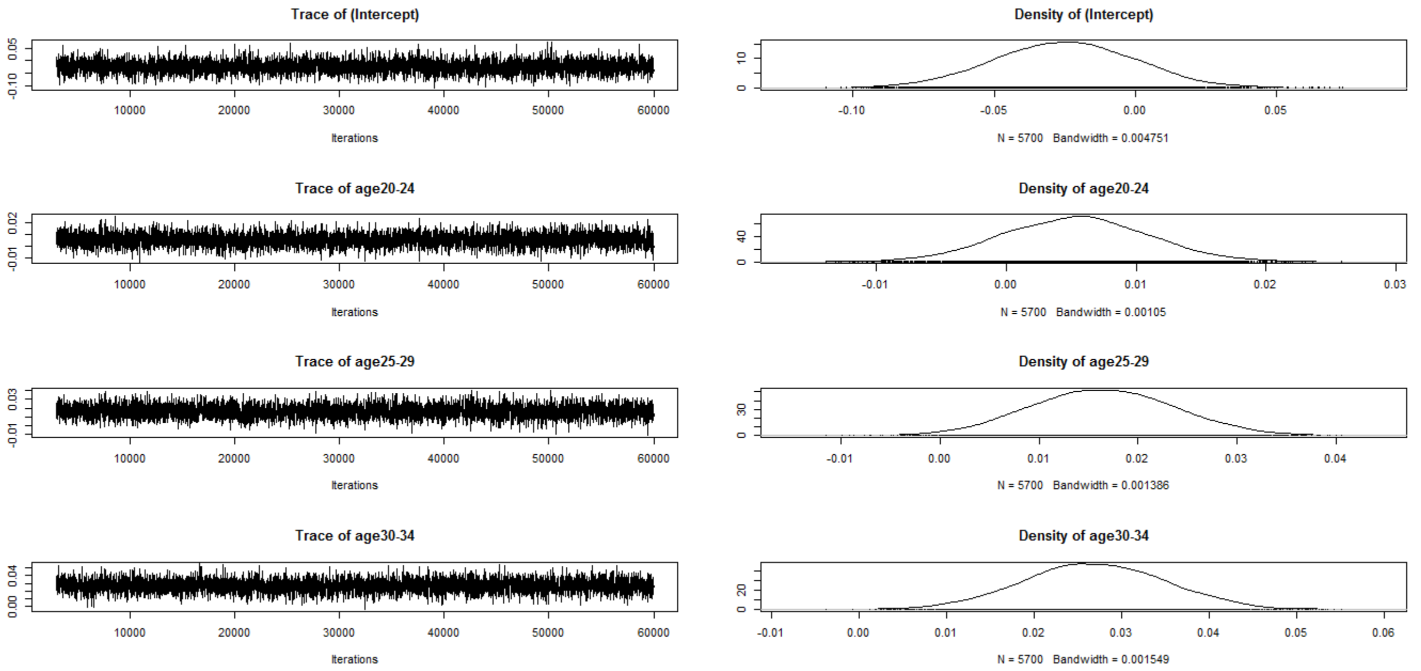
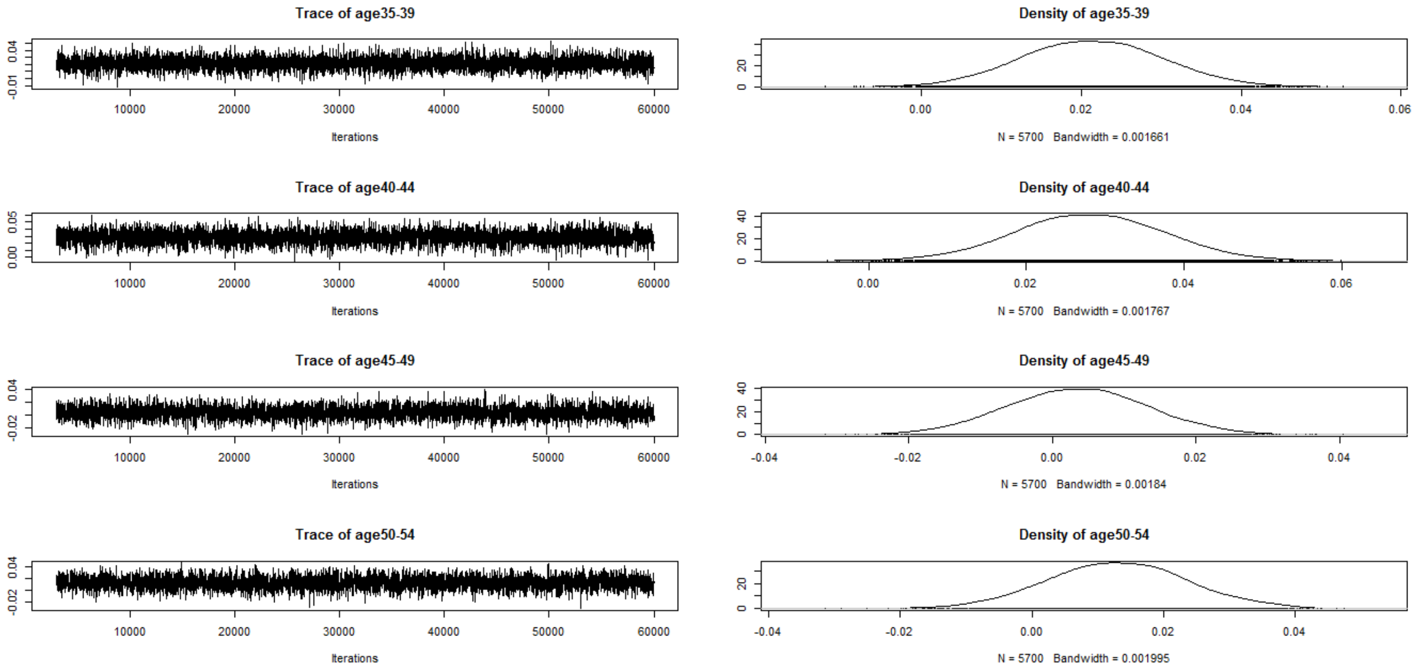
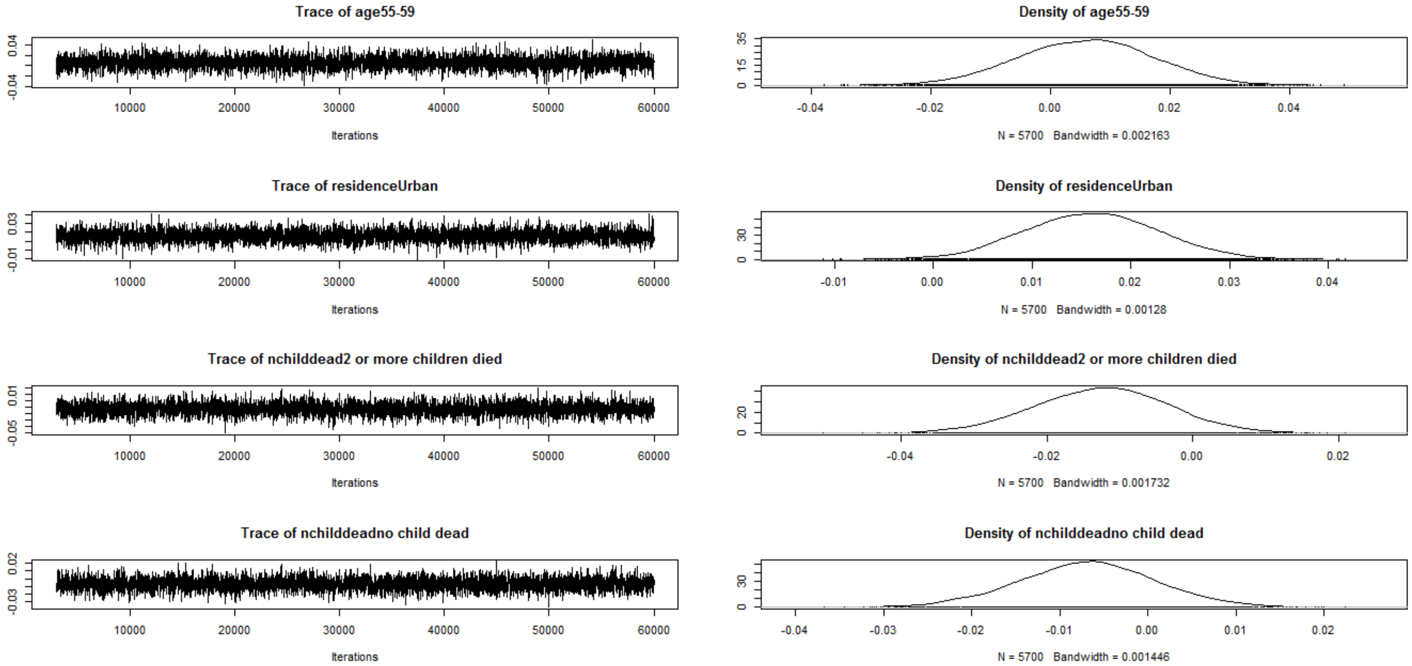


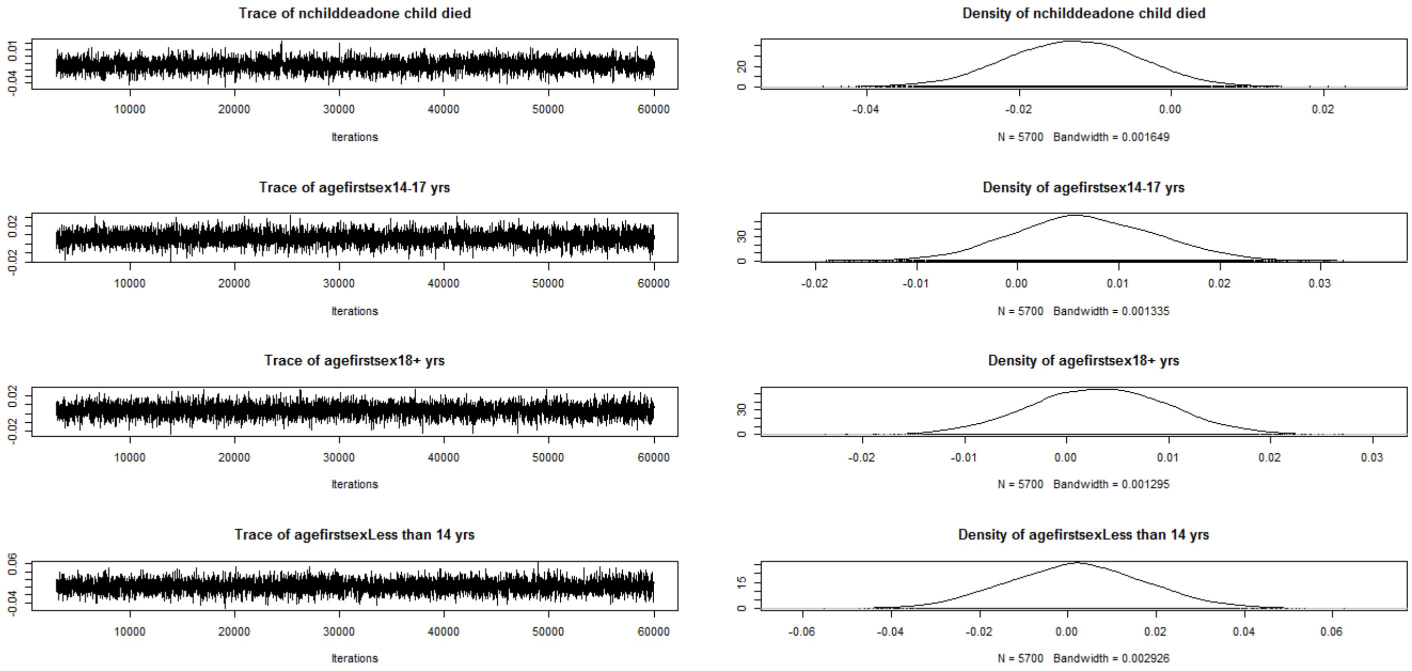
FIGURE 6: The posterior distribution of the variance component for the male data

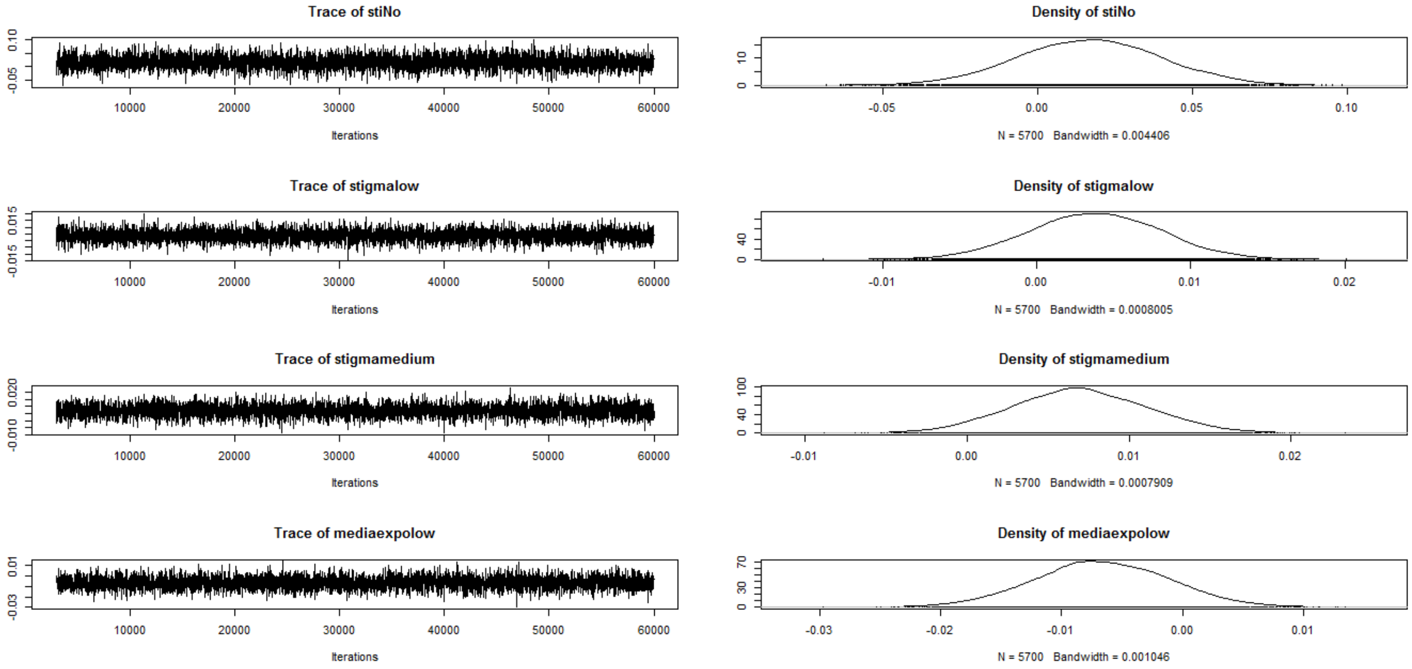
Figure (6) showed the posterior distribution of the random effect part of the model for the male, based on the analysis of 60000 iterations, 2000 burn in period, and with thin=10 in MCMCglmm.

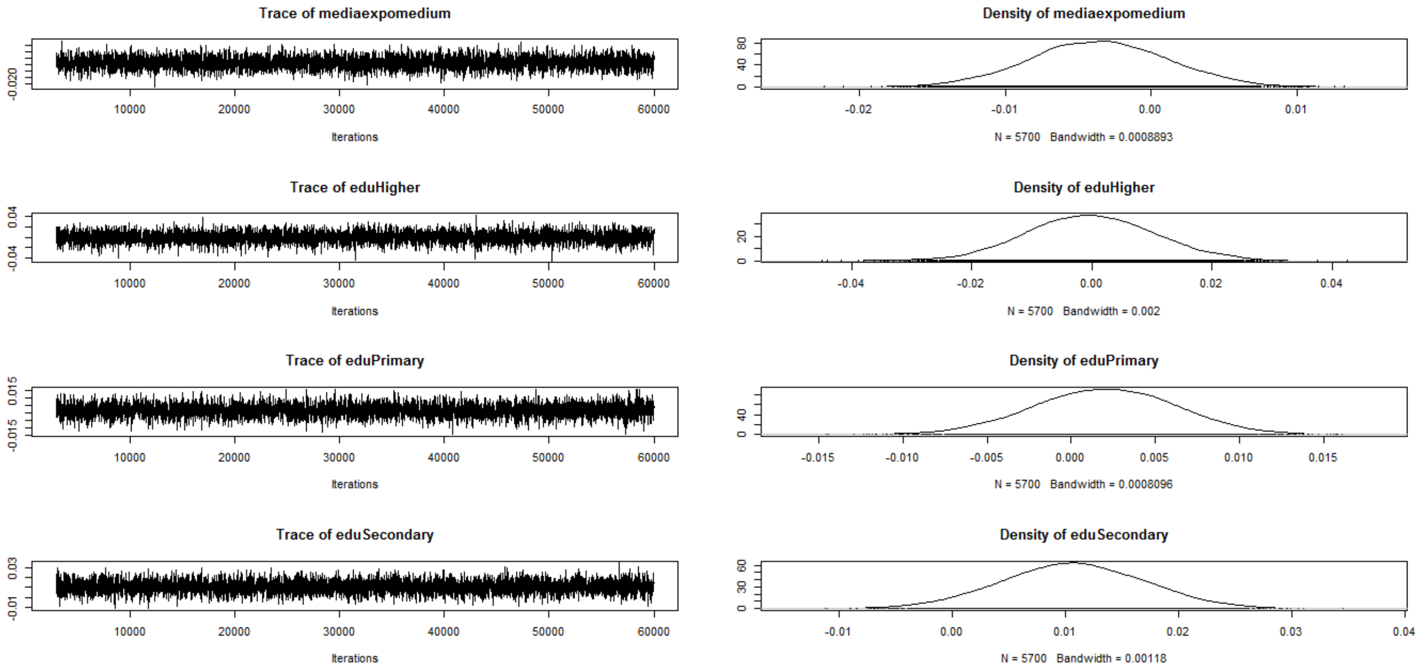



Trace of age35-39
-0.01
0.04
10000
20000
30000
40000
50000
60000
Iterations
Density of age35-39
0
20
0.00
0.02
0.04
0.06
N = 5700 Bandwidth = 0.001661
Trace of age40-44
0.00
0.05
10000
20000
30000
40000
50000
60000
Iterations
Density of age40-44
0
20
40
0.00
0.02
0.04
0.06
N = 5700 Bandwidth = 0.001767
Trace of age45-49
-0.02
0.04
10000
20000
30000
40000
50000
60000
Iterations
Density of age45-49
0
20
40
-0.04
-0.02
0.00
0.02
0.04
N = 5700 Bandwidth = 0.00184
Trace of age50-54
-0.02
0.02
0.04
10000
20000
30000
40000
50000
60000
Iterations
Density of age50-54
0
20
-0.04
-0.02
0.00
0.02
0.04
N = 5700 Bandwidth = 0.001995


Trace of age55-59
-0.04
0.04
10000
20000
30000
40000
50000
60000
Iterations
Density of age55-59
0
15
35
-0.04
-0.02
0.00
0.02
0.04
N = 5700 Bandwidth = 0.002163
Trace of residenceUrban
-0.01
0.03
10000
20000
30000
40000
50000
60000
Iterations
Density of residenceUrban
0
30
-0.01
0.00
0.01
0.02
0.03
0.04
N = 5700 Bandwidth = 0.00128
Trace of nchilddead2 or more children died
-0.05
0.01
10000
20000
30000
40000
50000
60000
Iterations
Density of nchilddead2 or more children died
0
20
-0.04
-0.02
0.00
0.02
N = 5700 Bandwidth = 0.001732
Trace of nchilddeadno child dead
-0.03
0.02
10000
20000
30000
40000
50000
60000
Iterations
Density of nchilddeadno child dead
0
30
-0.04
-0.03
-0.02
-0.01
0.00
0.01
0.02
N = 5700 Bandwidth = 0.001446


Trace of nchilddeadone child died
-0.04 0.01
10000 20000 30000 40000 50000 60000
Iterations
Density of nchilddeadone child died
0 20
-0.04 -0.02 0.00 0.02
N = 5700 Bandwidth = 0.001649
Trace of agefirstsex14-17 yrs
-0.02 0.02
10000 20000 30000 40000 50000 60000
Iterations
Density of agefirstsex14-17 yrs
0 30
-0.02 -0.01 0.00 0.01 0.02 0.03
N = 5700 Bandwidth = 0.001335
Trace of agefirstsex18+ yrs
-0.02 0.02
10000 20000 30000 40000 50000 60000
Iterations
Density of agefirstsex18+ yrs
0 30
-0.02 -0.01 0.00 0.01 0.02 0.03
N = 5700 Bandwidth = 0.001295
Trace of agefirstsexLess than 14 yrs
-0.04 0.06
10000 20000 30000 40000 50000 60000
Iterations
Density of agefirstsexLess than 14 yrs
0 15
-0.06 -0.04 -0.02 0.00 0.02 0.04 0.06
N = 5700 Bandwidth = 0.002926


Trace of stiNo
-0.05 0.10
10000 20000 30000 40000 50000 60000
Iterations
Density of stiNo
0 10
-0.05 0.00 0.05 0.10
N = 5700 Bandwidth = 0.004406
Trace of stigmalow
-0.015 0.015
10000 20000 30000 40000 50000 60000
Iterations
Density of stigmalow
0 40
-0.01 0.00 0.01 0.02
N = 5700 Bandwidth = 0.0008005
Trace of stigmamedium
-0.010 0.020
10000 20000 30000 40000 50000 60000
Iterations
Density of stigmamedium
0 40 100
-0.01 0.00 0.01 0.02
N = 5700 Bandwidth = 0.0007909
Trace of mediaexpolow
-0.03 0.01
10000 20000 30000 40000 50000 60000
Iterations
Density of mediaexpolow
0 30 70
-0.03 -0.02 -0.01 0.00 0.01
N = 5700 Bandwidth = 0.001046



Trace of mediaexpomedium

Density of mediaexpomedium

Trace of eduHigher

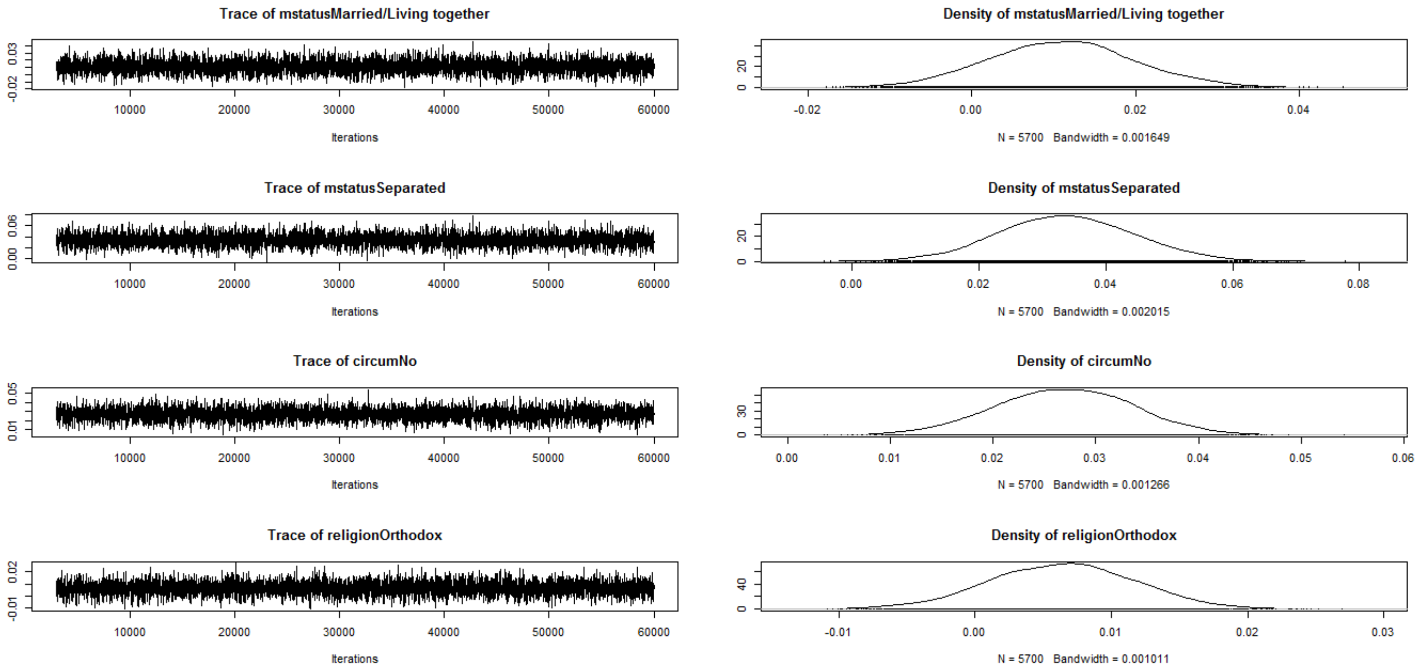
Density of eduHigher

Trace of eduPrimary

Density of eduPrimary

Trace of eduSecondary

Density of eduSecondary



Trace of mstatusMarried/Living together

Density of mstatusMarried/Living together

Trace of mstatusSeparated

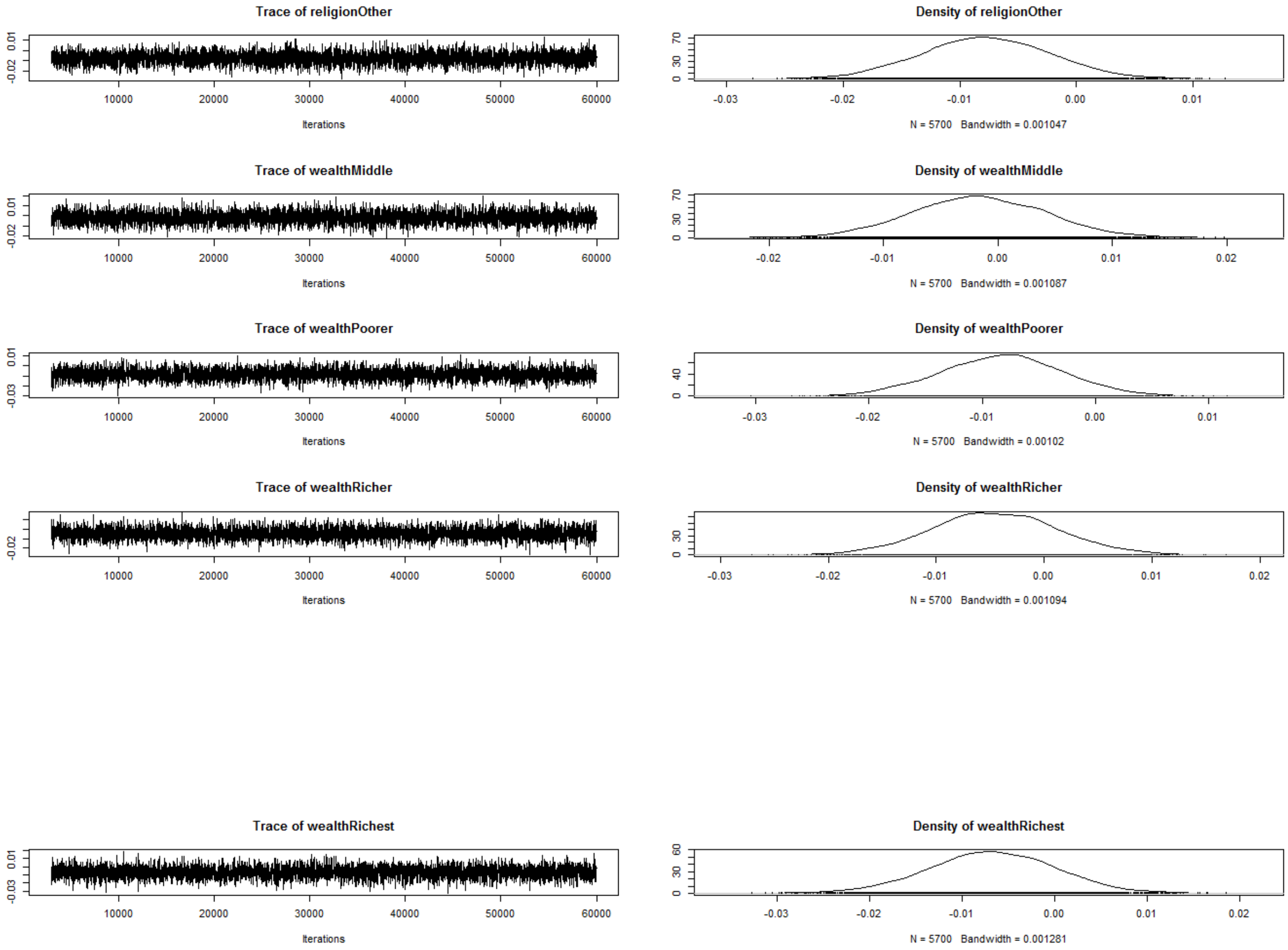
Density of mstatusSeparated

Trace of circumNo

Density of circumNo

Trace of religionOrthodox

Density of religionOrthodox



Trace of religionOther

Density of religionOther

Trace of wealthMiddle

Density of wealthMiddle

Trace of wealthPoorer

Density of wealthPoorer

Trace of wealthRicher

Density of wealthRicher

Trace of wealthRichest

Density of wealthRichest

FIGURE 7: The posterior distribution for the male data .

Figure 7 showed the posterior distribution of the fixed effect part of the model for the male, based on the analysis of 60000 iterations, 2000 burn in period, and with thin=10 in MCMCglmm.

Appendix B

The Posterior mean together with 0.025 quantile, median and 0.975 of the GLMM

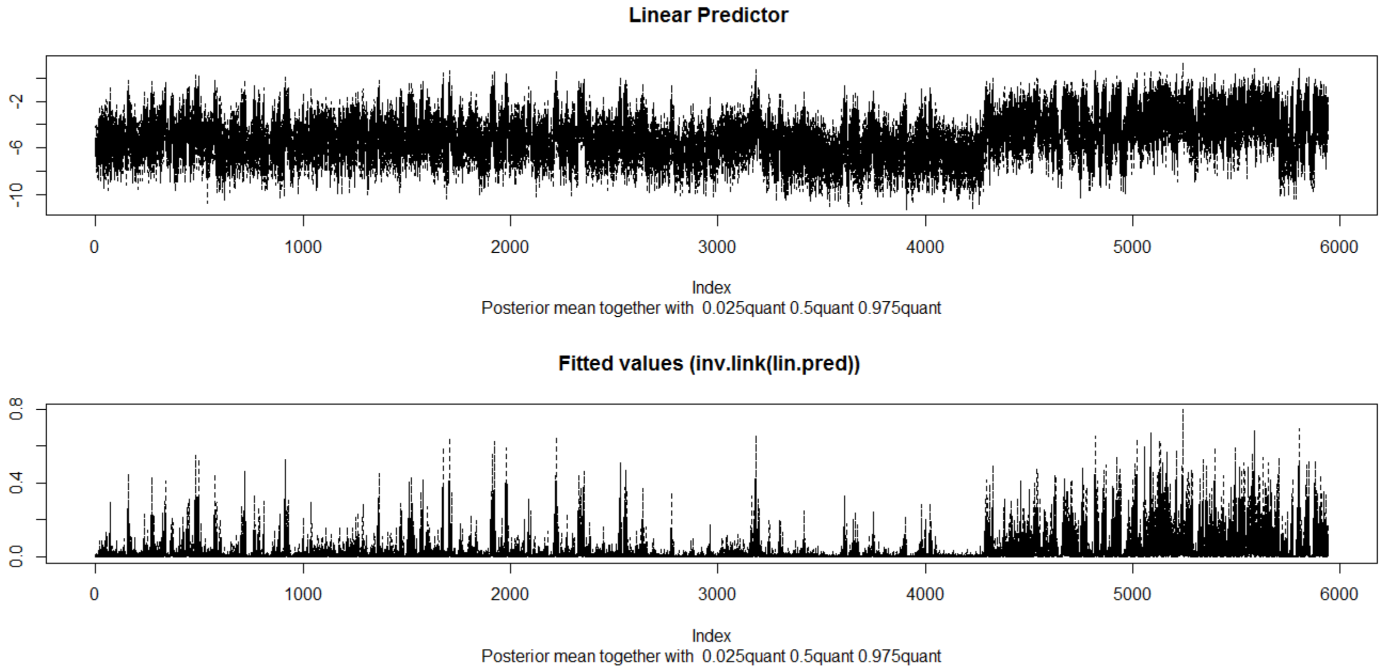
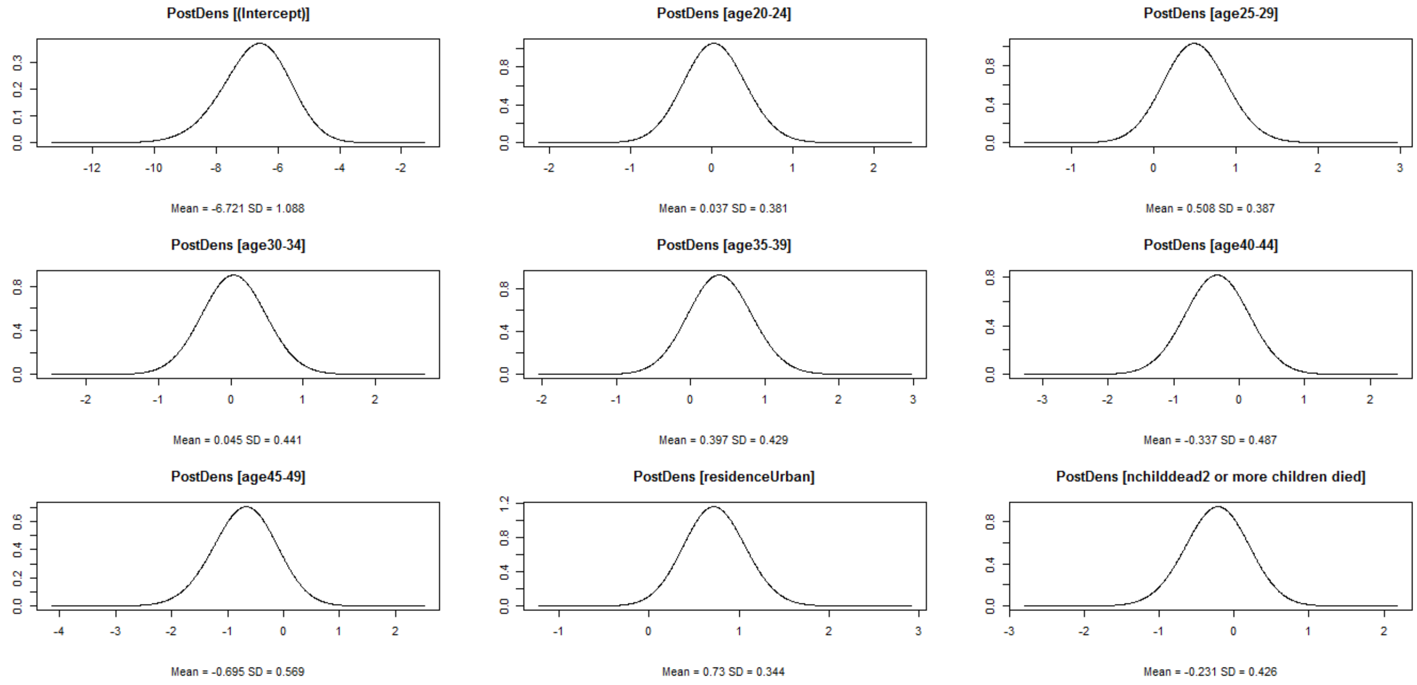
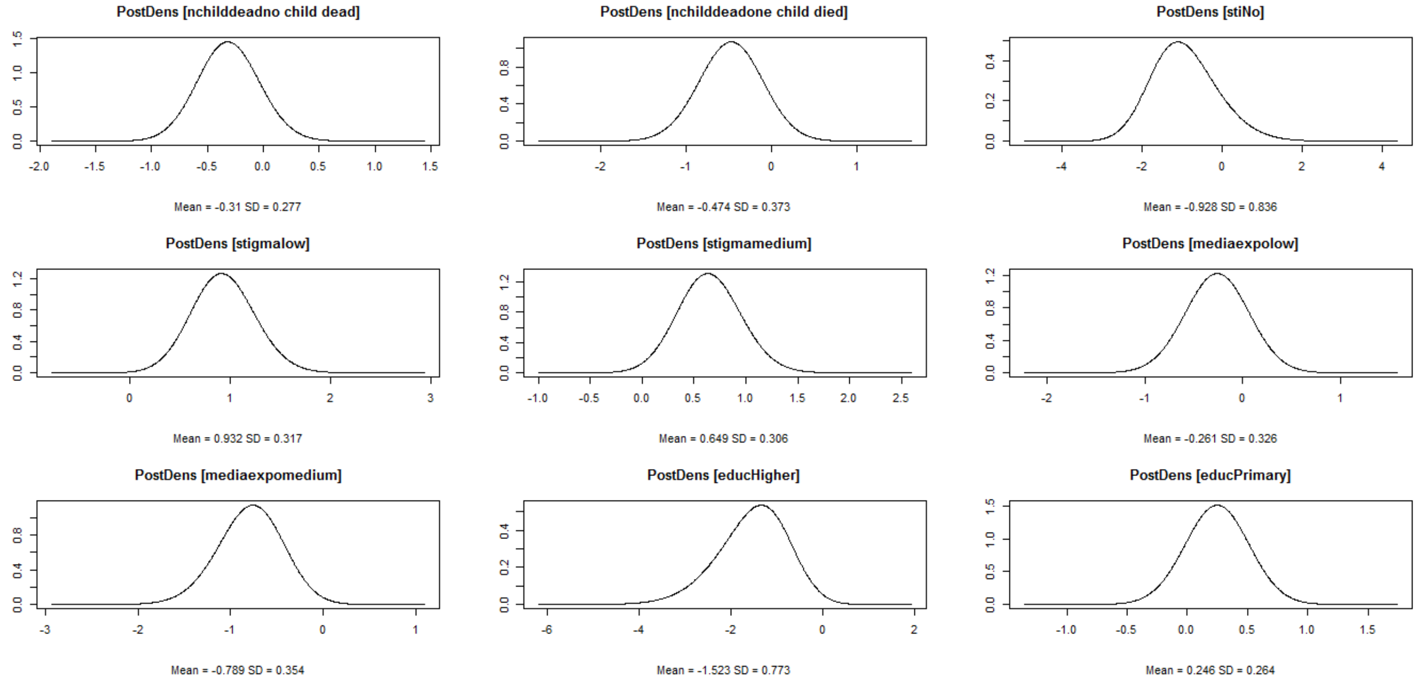
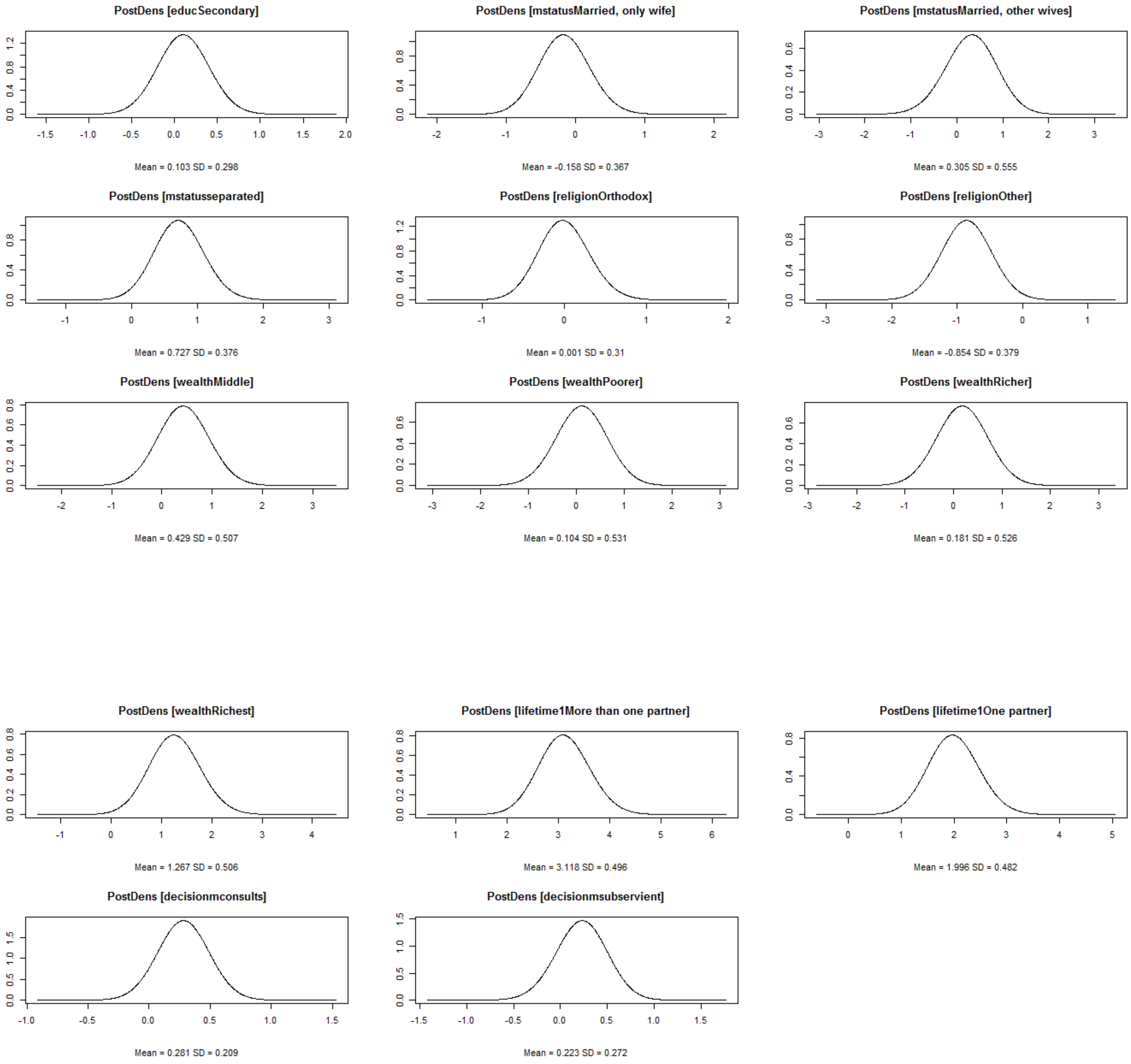


FIGURE 8: The posterior mean together with 0.025 quantile, median and 0.975 quantile for the female data

Figure (12) showed The posterior mean together with 0.025 quantile, median and 0.975 quantile for the female data.


PostDens [(Intercept)]
Mean = -6.721 SD = 1.088
PostDens [age20-24]
Mean = 0.037 SD = 0.381
PostDens [age25-29]
Mean = 0.508 SD = 0.387
PostDens [age30-34]
Mean = 0.045 SD = 0.441
PostDens [age35-39]
Mean = 0.397 SD = 0.429
PostDens [age40-44]
Mean = -0.337 SD = 0.487
PostDens [age45-49]
Mean = -0.695 SD = 0.569
PostDens [residenceUrban]
Mean = 0.73 SD = 0.344
PostDens [nchilddead2 or more children died]
Mean = -0.231 SD = 0.426


PostDens [nchilddeadno child dead]
Mean = -0.31 SD = 0.277
PostDens [nchilddeadone child died]
Mean = -0.474 SD = 0.373
PostDens [stiNo]
Mean = -0.928 SD = 0.836
PostDens [stigmalow]
Mean = 0.932 SD = 0.317
PostDens [stigmamedium]
Mean = 0.649 SD = 0.306
PostDens [mediaexpolow]
Mean = -0.261 SD = 0.326
PostDens [mediaexpomedium]
Mean = -0.789 SD = 0.354
PostDens [educHigher]
Mean = -1.523 SD = 0.773
PostDens [educPrimary]
Mean = 0.246 SD = 0.264



PostDens [educSecondary]

Mean = 0.103 SD = 0.298

PostDens [mstatusMarried, only wife]

Mean = -0.158 SD = 0.367

PostDens [mstatusMarried, other wives]

Mean = 0.305 SD = 0.555

PostDens [mstatusseparated]

Mean = 0.727 SD = 0.376

PostDens [religionOrthodox]

Mean = 0.001 SD = 0.31

PostDens [religionOther]

Mean = -0.854 SD = 0.379

PostDens [wealthMiddle]

Mean = 0.429 SD = 0.507

PostDens [wealthPoorer]

Mean = 0.104 SD = 0.531

PostDens [wealthRicher]

Mean = 0.181 SD = 0.526

PostDens [wealthRichest]

Mean = 1.267 SD = 0.506

PostDens [lifetime1More than one partner]

Mean = 3.118 SD = 0.496

PostDens [lifetime1One partner]

Mean = 1.996 SD = 0.482

PostDens [decisionmconsults]

Mean = 0.281 SD = 0.209

PostDens [decisionmsubservient]

Mean = 0.223 SD = 0.272

FIGURE 9: The posterior densities for the fixed effects part of the GLMM for the female data

Figure 13 showed The posterior densities for the fixed effects part of the GLMM for the female data. Figure 14 The posterior mean, 0.025%, median and 0.975% for the region for the female data. Figure 15 The posterior densities for the region for the female data.

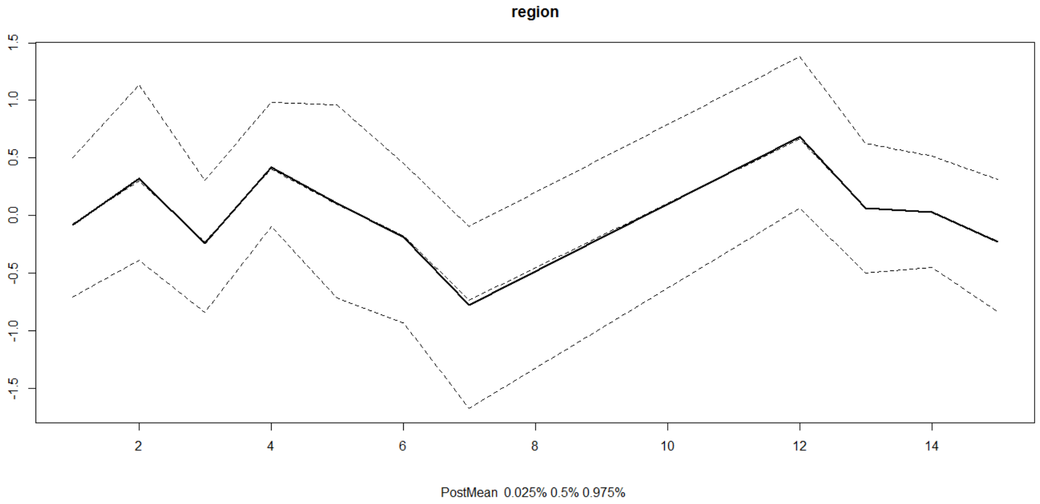


Figure 10: The posterior mean, 0.025%, median and 0.975% for the region for the female data .

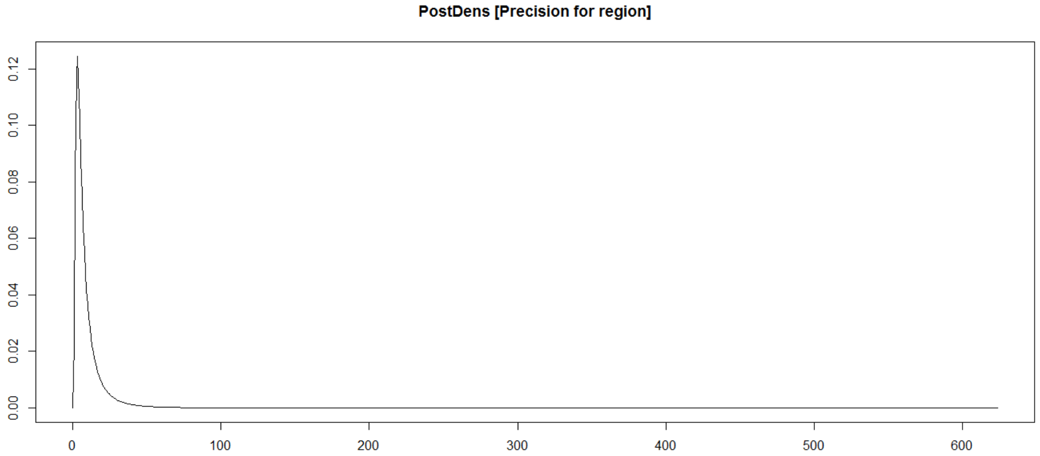


Figure 11: The posterior densities for the region for the female data .

Figure (12) showed The posterior mean together with 0.025 quantile, median and 0.975 quantile for the male data.

Figure 13 showed The posterior densities for the fixed effects part of the GLMM for the male data. Figure 14 The posterior mean, 0.025%, median and 0.975% for the region for the male data. Figure 15 The posterior densities for the region for the male

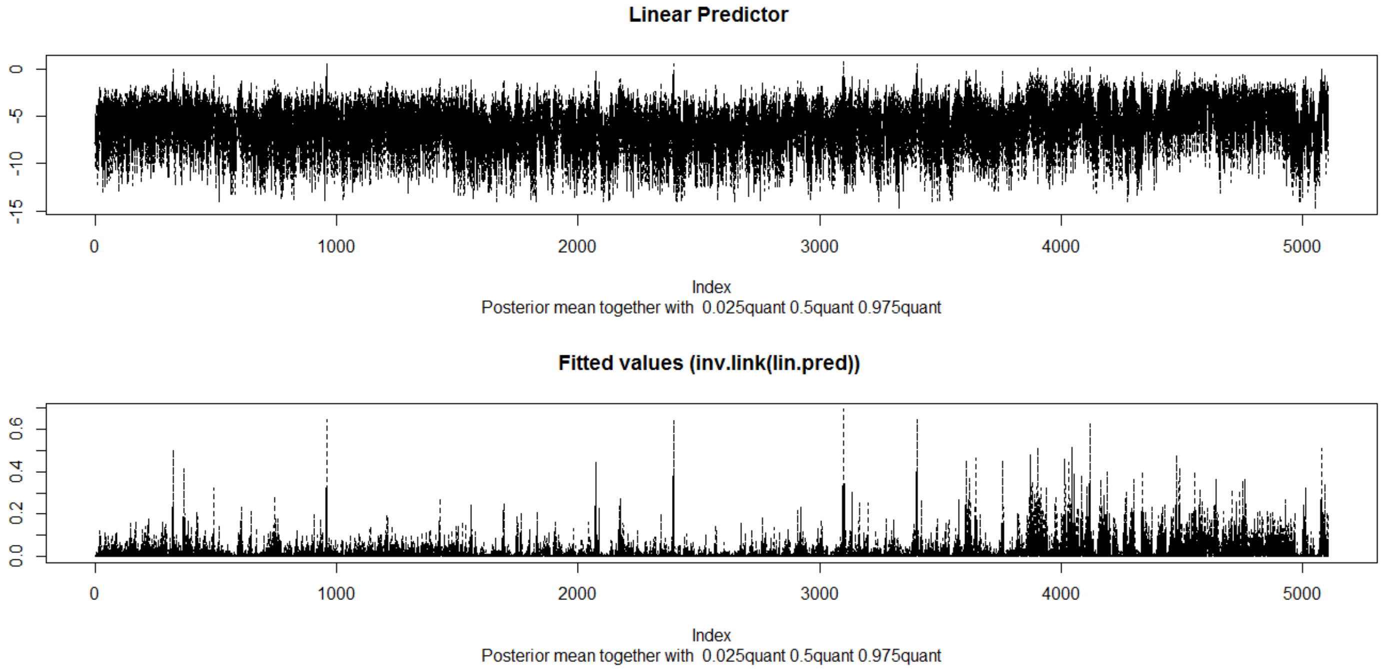
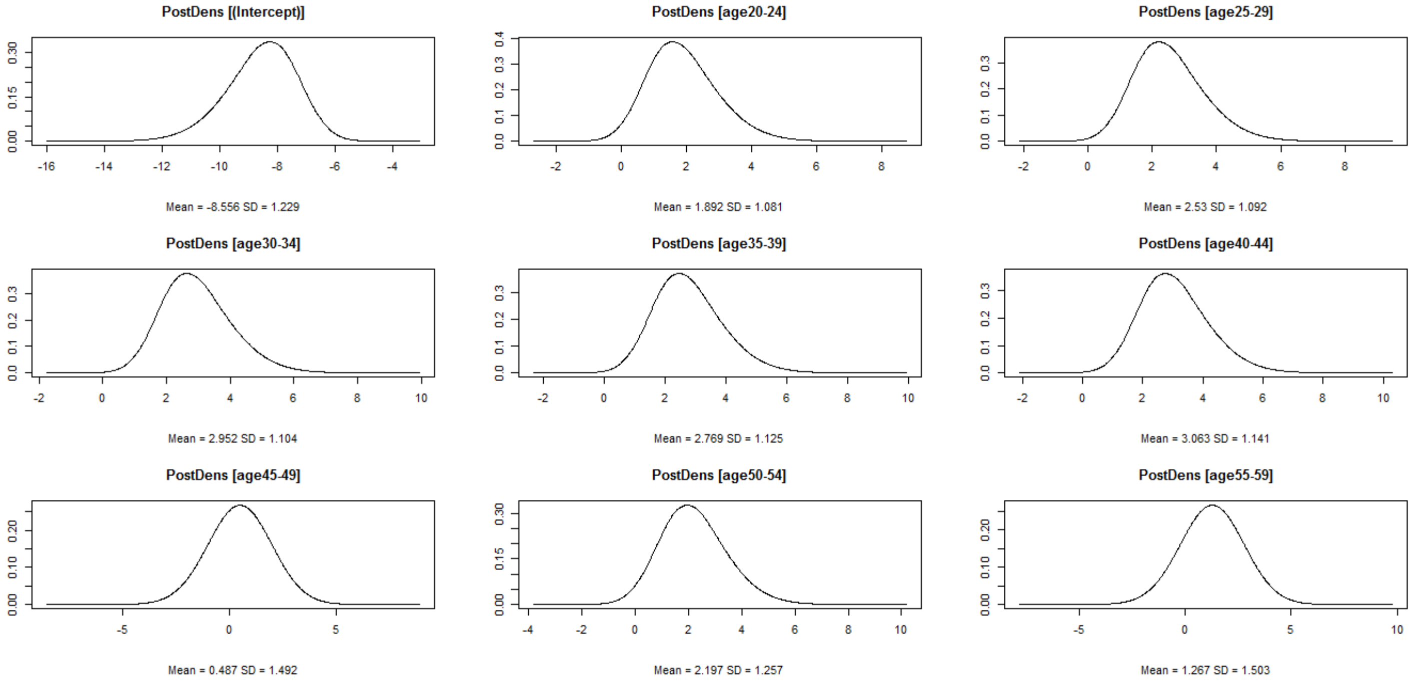
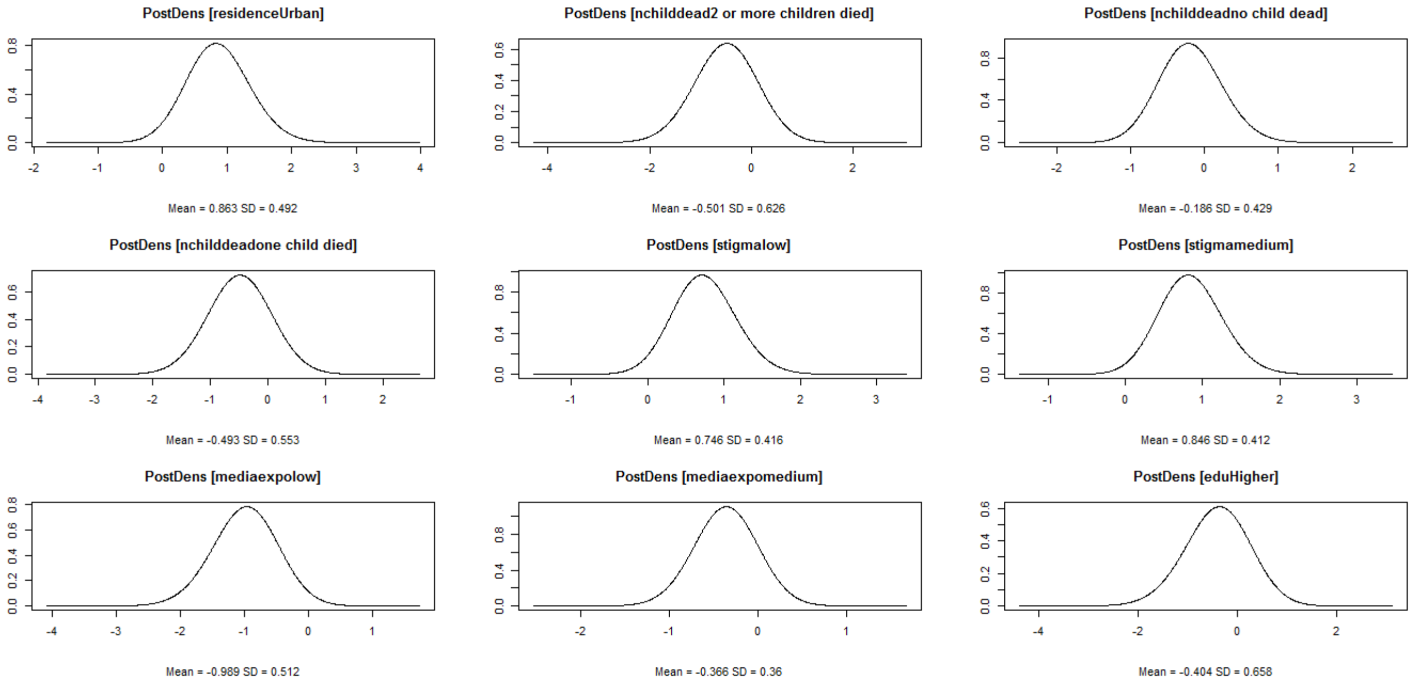
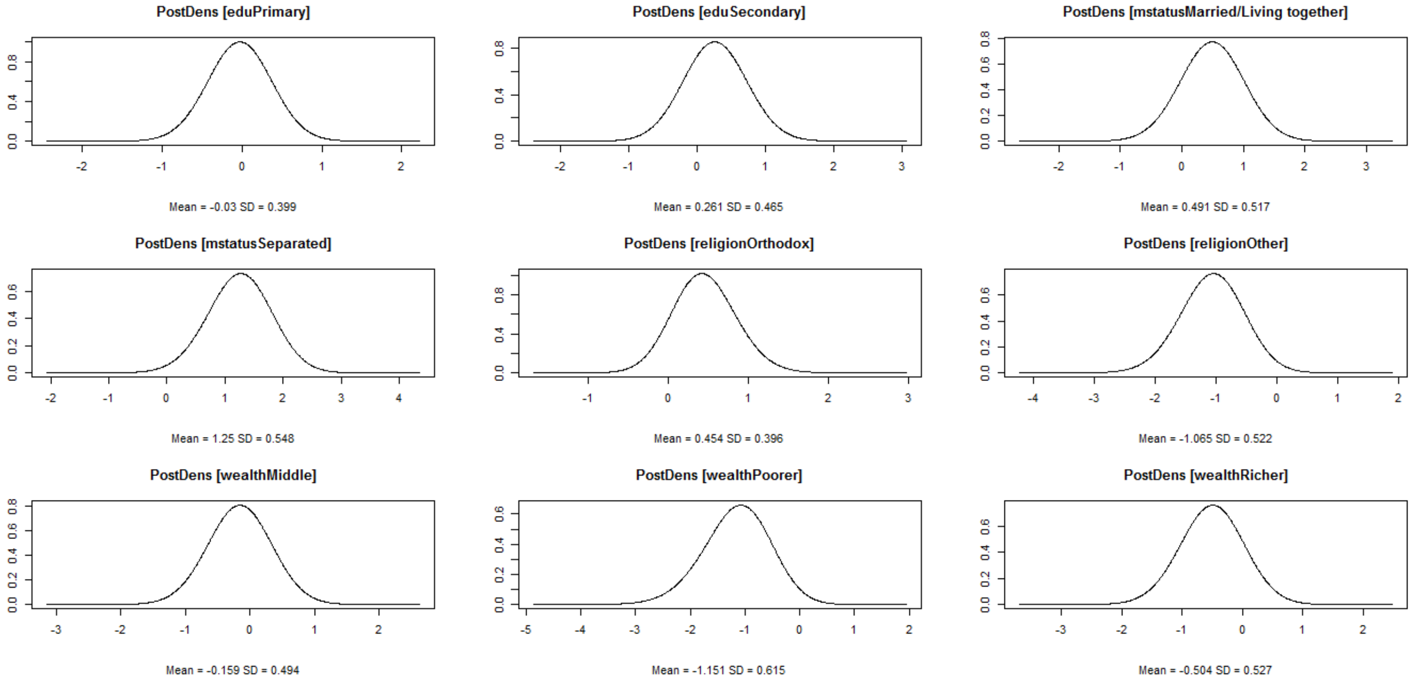


Figure 12: The posterior mean together with 0.025 quantile, median and 0.975 quantile for the male data




PostDens [residenceUrban]
0.0 0.4 0.8
-2 -1 0 1 2 3 4
Mean = 0.863 SD = 0.492
PostDens [nchilddead2 or more children died]
0.0 0.2 0.4 0.6
-4 -2 0 2
Mean = -0.501 SD = 0.626
PostDens [nchilddeadno child dead]
0.0 0.4 0.8
-2 -1 0 1 2
Mean = -0.186 SD = 0.429
PostDens [nchilddeadone child died]
0.0 0.2 0.4 0.6
-4 -3 -2 -1 0 1 2
Mean = -0.493 SD = 0.553
PostDens [stigmalow]
0.0 0.4 0.8
-1 0 1 2 3
Mean = 0.746 SD = 0.416
PostDens [stigmamedium]
0.0 0.4 0.8
-1 0 1 2 3
Mean = 0.846 SD = 0.412
PostDens [mediaexpolow]
0.0 0.2 0.4 0.6 0.8
-4 -3 -2 -1 0 1
Mean = -0.989 SD = 0.512
PostDens [mediaexpomedium]
0.0 0.4 0.8
-2 -1 0 1
Mean = -0.366 SD = 0.36
PostDens [eduHigher]
0.0 0.2 0.4 0.6
-4 -2 0 2
Mean = -0.404 SD = 0.658


PostDens [eduPrimary]
0.0 0.4 0.8
-2 -1 0 1 2
Mean = -0.03 SD = 0.399
PostDens [eduSecondary]
0.0 0.4 0.8
-2 -1 0 1 2 3
Mean = 0.261 SD = 0.465
PostDens [mstatusMarried/Living together]
0.0 0.2 0.4 0.6 0.8
-2 -1 0 1 2 3
Mean = 0.491 SD = 0.517
PostDens [mstatusSeparated]
0.0 0.2 0.4 0.6
-2 -1 0 1 2 3 4
Mean = 1.25 SD = 0.548
PostDens [religionOrthodox]
0.0 0.4 0.8
-1 0 1 2 3
Mean = 0.454 SD = 0.396
PostDens [religionOther]
0.0 0.2 0.4 0.6
-4 -3 -2 -1 0 1 2
Mean = -1.065 SD = 0.522
PostDens [wealthMiddle]
0.0 0.2 0.4 0.6 0.8
-3 -2 -1 0 1 2
Mean = -0.159 SD = 0.494
PostDens [wealthPoorer]
0.0 0.2 0.4 0.6
-5 -4 -3 -2 -1 0 1 2
Mean = -1.151 SD = 0.615
PostDens [wealthRicher]
0.0 0.2 0.4 0.6
-3 -2 -1 0 1 2
Mean = -0.504 SD = 0.527

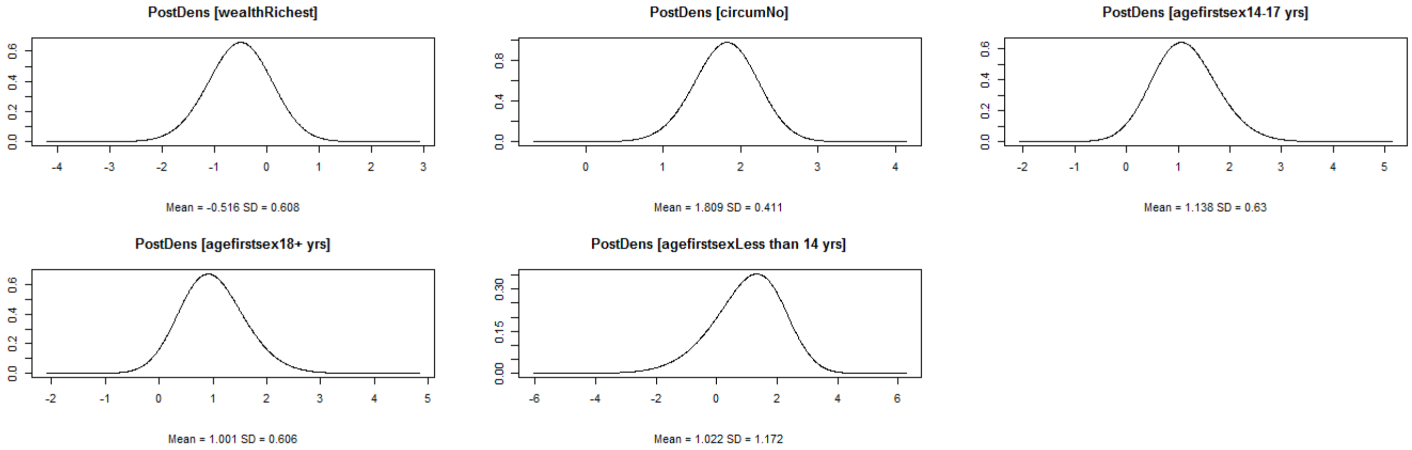


FIGURE 13: The posterior densities for the fixed effects part of the GLMM for the male data

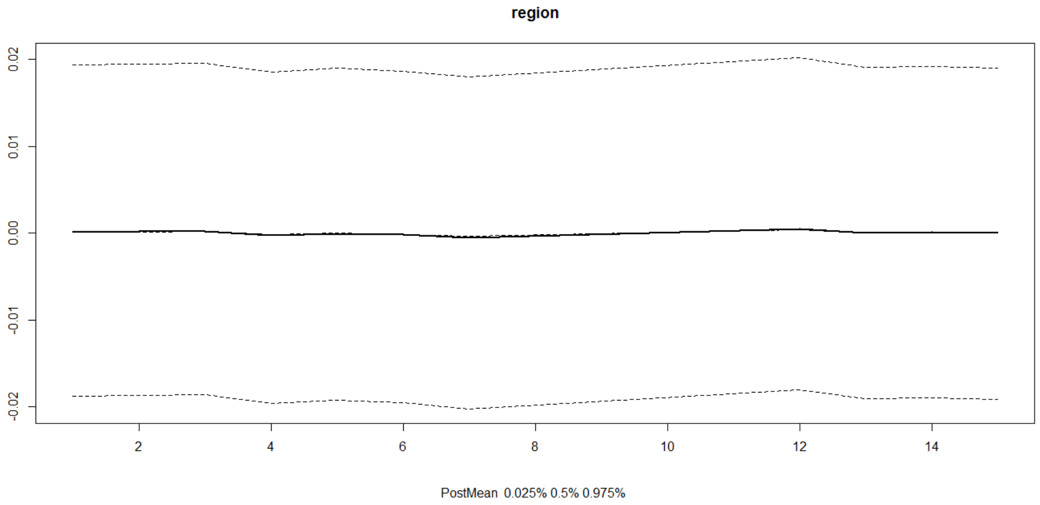


FIGURE 14: The posterior mean, 0.025%, median and 0.975% for the region for the male data .

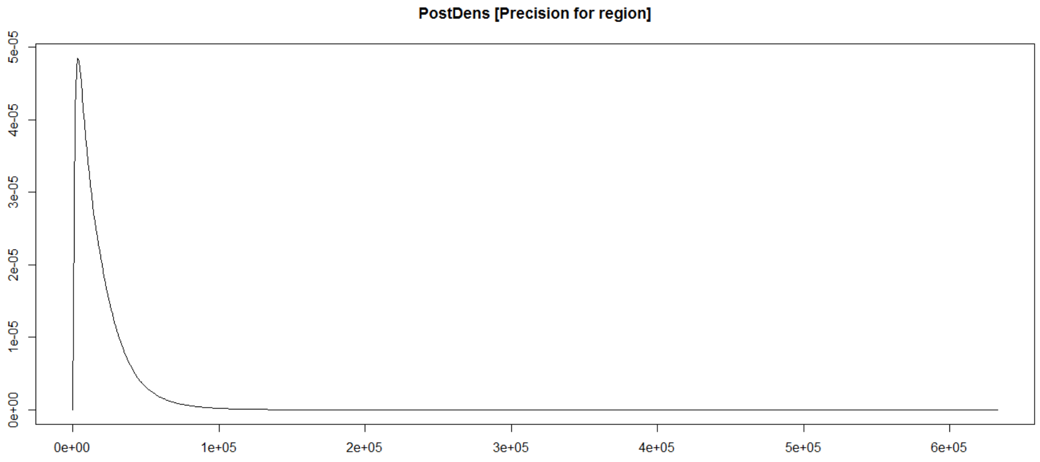


Figure 15: The posterior densities for the region for the female data .

data.

Bibliography

[1] Abuye, C., Urgaa, A. K., and Kebede, A. A. Ethiopian health and nutrition research institute.

[2] Africa, S.-S. The hiv/aids epidemic in south africa august 2006.

[3] Aitkin, M., and Longford, N. Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society. Series A (General)* (1986), 1–43.

[4] Amsalu, A., and Gebremichael, D. An overview of climate change impacts and responses in ethiopia in 2009. *ETHIOPIAN ENVIRONMENT REVIEW NO* (2010).

[5] Archer, K., and Lemeshow, S. Goodness-of-fit test for a logistic regression model fitted using survey sample data. *Stata Journal 6*, 1 (2006), 97.

[6] Ateka, G. K. Factors in hiv/aids transmission in sub-saharan africa. *Bulletin of the World Health Organization 79*, 12 (2001), 1168–1168.

[7] Auvert, B., Buve, A., Lagarde, E., Kahindo, M., Chege, J., Rutenberg, N., Musonda, R., Laourou, M., Akam, E., Weiss, H., et al. Male circumcision and hiv infection in four cities in sub-saharan africa. *Aids 15* (2001), S31–S40.

[8] Auvert, B., Taljaard, D., Lagarde, E., Sobngwi-Tambekou, J., Sitta, R., and Puren, A. Randomized, controlled intervention trial of male circumcision for reduction of hiv infection risk: the anrs 1265 trial. *PLoS medicine 2*, 11 (2005), e298.

[9] Bachou, H., Tylleskär, T., Kaddu-Mulindwa, D., and Tumwine, J. Bacteraemia among severely malnourished children infected and uninfected

with the human immunodeficiency virus-1 in kampala, uganda. *BMC infectious diseases 6*, 1 (2006), 160.

[10] Bärnighausen, T., Hosegood, V., Timaeus, I., and Newell, M. The socioeconomic determinants of hiv incidence: evidence from a longitudinal, population-based study in rural south africa. *AIDS (London, England) 21*, Suppl 7 (2007), S29.

[11] Bärnighausen, T., Tanser, F., and Newell, M. Lack of a decline in hiv incidence in a rural community with high hiv prevalence in south africa, 2003–2007. *AIDS research and Human Retroviruses 25*, 4 (2009), 405–409.

[12] Bekele, A., and Ali, A. Effectiveness of iec interventions in reducing hiv/aids related stigma among high school adolescents in hawassa, southern ethiopia. *Ethiopian Journal of Health Development 22*, 3 (2008), 232–242.

[13] Bernardo, J., and Smith, A. M (1994). bayesian theory. *C hichester: Wiley*.

[14] Bernardo, J., and Smith, A. Bayesian theory. *Measurement Science and Technology 12*, 2 (2001), 221.

[15] Besag, J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)* (1974), 192–236.

[16] Beyene Petros., S. B., and Mekonnen, Y. Aids and college students in addis ababa: A study on knowledge, attitude and behavior. *Ethiopian Journal of Health Development 11(2)* (1997), 115–123.

[17] Binder, D. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review/Revue Internationale de Statistique* (1983), 279–292.

[18] Bingenheimer, J. Wealth, wealth indices and hiv risk in east africa. *International Family Planning Perspectives 33*, 2 (2007), 83.

[19] Boerma, J. T., and Weir, S. S. Integrating demographic and epidemiological approaches to research on hiv/aids: the proximate-determinants framework. *Journal of Infectious Diseases 191*, Supplement 1 (2005), S61–S67.

[20] Bongaarts, J., and Potter, R. G. *Fertility, biology, and behavior: An analysis of the proximate determinants*. Academic Press New York, 1983.

[21] Bradley, H., Bedada, A., Brahmbhatt, H., Kidanu, A., Gillespie, D., and Tsui, A. Educational attainment and hiv status among ethiopian voluntary counseling and testing clients. *AIDS and Behavior 11*, 5 (2007), 736–742.

[22] Brogan, D. J. Pitfalls of using standard statistical software packages for sample survey data. *Encyclopedia of biostatistics 5* (1998), 4167–4174.

[23] Brooks, S., and Gelman, A. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics 7*, 4 (1998), 434–455.

[24] Browne, W. *Applying MCMC methods to multi-level models*. University of Bath, 1998.

[25] Brownel, W., and Draper, D. Implementation and performance issues in the bayesian and likelihood fitting of multilevel models. *Computational statistics 15* (2000), 391–420.

[26] Calder, C., and Cressie, N. Some topics in convolution-based spatial modeling. *Proceedings of the 56th Session of the International Statistics Institute* (2007), 22–29.

[27] Casella, G., and George, E. Explaining the gibbs sampler. *The American Statistician 46*, 3 (1992), 167–174.

[28] Catania, J., Kegeles, S., and Coates, T. Towards an understanding of risk behavior: an aids risk reduction model (arrm). *Health Education & Behavior 17*, 1 (1990), 53–72.

[29] CENTRAL STATISTICAL AGENCY. Compilation of economic statistics in ethiopia. Tech. rep., 2007.

[30] Chambers, R., and Tzavidis, N. M-quantile models for small area estimation. *Biometrika 93*, 2 (2006), 255–268.

[31] Chellappa, R. Two-dimensional discrete gaussian markov random field models for image processing. *Progress in pattern recognition 2* (1985), 79–112.

[32] Chen, L., Jha, P., Stirling, B., Sgaier, S., Daid, T., Kaul, R., and Nagelkerke, N. Sexual risk factors for hiv infection in early and advanced hiv epidemics in sub-saharan africa: systematic overview of 68 epidemiological studies. *PLoS One 2*, 10 (2007), e1001.

[33] Chen, Z., and Mantel, H. Analysis of binary data from a complex survey with misclassification in an ordinal covariate.

[34] Chib, S., and Greenberg, E. Understanding the metropolis-hastings algorithm. *The American Statistician 49*, 4 (1995), 327–335.

[35] Cohen, D., and UNDP, H. Poverty and hiv/aids in sub-saharan africa, 1998.

[36] Corrigan, P. W., Penn, D. L., et al. Lessons from social psychology on discrediting psychiatric stigma. *American Psychologist 54* (1999), 765–776.

[37] Cowles, M., and Carlin, B. Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association 91*, 434 (1996), 883–904.

[38] Cox, R. Probability, frequency and reasonable expectation. *American journal of physics 14*, 1 (1946), 1–13.

[39] CRDA. Annual report crda. Tech. rep., 2006.

[40] Cressie, N. Statistics for spatial data. *Terra Nova 4*, 5 (1992), 613–617.

[41] Currier, J., Kendall, M., Zackin, R., Henry, W., Alston-Smith, B., Torriani, F., Schouten, J., Mickelberg, K., Li, Y., Hodis, H., et al. Carotid artery intima–media thickness and hiv infection: traditional risk factors overshadow impact of protease inhibitor exposure. *AIDS (London, England) 19*, 9 (2005), 927.

[42] Davis, K., and Blake, J. Social structure and fertility: An analytic framework. *Economic development and cultural change* (1956), 211–235.

[43] DAW, W. UNAIDS, 2000. In *The HIV/AIDS pandemic and its gender implications, Report of the expert group meeting, DAW & Department of Economic and Social Affairs, 17th of November* (2000).

[44] De Walque, D. How does the impact of an hiv/aids information campaign vary with educational attainment? evidence from rural uganda. *Journal of Development Economics 84*, 2 (2007), 686–714.

[45] Dempster, A., Laird, N., and Rubin, D. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* (1977), 1–38.

[46] Donner, A., Birkett, N., and Buck, C. Randomization by cluster sample size requirements and analysis. *American Journal of Epidemiology 114*, 6 (1981), 906–914.

[47] Driedger, D., and D'Aubin, A. Discarding the shroud of silence: An international perspective on violence, women and disability. *Canadian Woman Studies 12*, 1 (1991).

[48] Dunteman, G. H. *Principal components analysis*, vol. 69. SAGE Publications, Incorporated, 1989.

[49] Dye, C., Scheele, S., Dolin, P., Pathania, V., Raviglione, M. C., et al. Global burden of tuberculosis. *JAMA: the journal of the American Medical Association 282*, 7 (1999), 677–686.

[50] Epidemic., G. A. Report on the global aids epidemic. Tech. rep., 2008.

[51] Essex, M., Mboup, S., Kanki, P., and Kalengayi, M. *AIDS in Africa*. Raven press, 1994.

[52] Fears, T., Benichou, J., and Gail, M. A reminder of the fallibility of the wald statistic. *The American Statistician 50*, 3 (1996), 226–227.

[53] Fontanet, A., Messele, T., Dejene, A., Enquselassie, F., Abebe, A., Cutts, F., de Wit, T., Sahlu, T., Bindels, P., Yeneneh, H., et al. Age-and sex-specific hiv-1 prevalence in the urban community setting of addis ababa, ethiopia. *Aids 12*, 3 (1998), 315.

[54] Fontanet, A., and Piot, P. State of our knowledge: the epidemiology of hiv/aids. *Health Transition Review* (1994), 11–22.

[55] Frühwirth-Schnatter, S. Markov chain monte carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association 96*, 453 (2001), 194–209.

[56] Fuller, W. Regression analysis for sample survey. *Sankhya Series C 37* (1975), 117–132.

[57] Gelfand, A., Sahu, S., and Carlin, B. Efficient parametrisations for normal linear mixed models. *Biometrika 82*, 3 (1995), 479–488.

[58] Gelfand, A., and Smith, A. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association 85*, 410 (1990), 398–409.

[59] Gelman, A., and Rubin, D. Inference from iterative simulation using multiple sequences. *Statistical science 7*, 4 (1992), 457–472.

[60] Gelman, A., and Shirley, K. Inference from simulations and monitoring convergence. *Handbook of Markov Chain Monte Carlo: Methods and Applications.* (2010), 131–143.

[61] Geman, S., and Geman, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 6 (1984), 721–741.

[62] George, E., Makov, U., and Smith, A. Conjugate likelihood distributions. *Scandinavian Journal of Statistics* (1993), 147–156.

[63] Geweke, J. Bayesian inference in econometric models using monte carlo integration. *Econometrica: Journal of the Econometric Society* (1989), 1317–1339.

[64] Gilbert, P., and Andrews, B. *Shame: Interpersonal behavior, psychopathology, and culture*. Oxford University Press on Demand, 1998.

[65] Gilks, W., Richardson, S., and Spiegelhalter, D. *Markov Chain Monte Carlo in practice: interdisciplinary statistics*, vol. 2. Chapman & Hall/CRC, 1995.

[66] Gilks, W., Richardson, S., and Spiegelhalter, D. *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC, 1996.

[67] Gilks, W., and Wild, P. Adaptive rejection sampling for gibbs sampling. *Applied Statistics* (1992), 337–348.

[68] Gillespie, S., Kadiyala, S., and Greener, R. Is poverty or wealth driving hiv transmission? *Aids 21* (2007), S5–S16.

[69] Goldstein, H. *Multilevel models in education and social research.* London, England: Charles Griffin & Co; New York, NY, US: Oxford University Press, 1987.

[70] Goldstein, H. Bootstrapping in multilevel models. *Handbook of advanced multilevel analysis* (2011), 163–171.

[71] Goldstein, H. *Multilevel statistical models*, vol. 922. Wiley, 2011.

[72] Groce, N. Hiv/aids and disability: capturing hidden voices: the world bank/yale university global survey on hiv/aids and disability.

[73] Groce, N., and Trasi, R. Rape of individuals with disability in the age of aids: The folk belief of virgin cleansing. *Lancet 363*, 9422 (2004), 1663–1664.

[74] Hadfield, J. Mcmc methods for multi-response generalized linear mixed models: the mcmcglmm r package. *Journal of Statistical Software 33*, 2 (2010), 1–22.

[75] Hailù, K., Bekura, D., Buttò, S., Verani, P., Titti, F., Sernicola, L., Rapicetta, M., Rossi, G., and Pasquini, P. Serological survey of human immunodeficiency virus (hiv) in ethiopia. *Journal of medical virology 28*, 1 (1989), 21–24.

[76] Halperin, D., and Epstein, H. Concurrent sexual partnerships help to explain africa's high hiv prevalence: implications for prevention. *The Lancet 364*, 9428 (2004), 4–6.

[77] HAPCO. Behavioral survey surveillance (bss) study in ethiopia. Tech. rep., 2002.

[78] Hargreaves, J., and Glynn, J. Educational attainment and hiv-1 infection in developing countries: a systematic review. *Tropical Medicine & International Health 7*, 6 (2002), 489–498.

[79] Hastings, W. Monte carlo sampling methods using markov chains and their applications. *Biometrika 57*, 1 (1970), 97–109.

[80] Haub, C., and Cornelius, D. *World Population Data Sheet of the Population Reference Bureau, Inc.* Population Reference Bureau, 2001.

[81] Helleringer, S., and Kohler, H. The structure of sexual networks and the spread of hiv in sub-saharan africa: evidence from likoma island (malawi).

[82] Hladik, W., Shabbir, I., Jelaludin, A., Woldu, A., Tsehaynesh, M., and Tadesse, W. Hiv/aids in ethiopia: where is the epidemic heading? *Sexually transmitted infections 82*, suppl 1 (2006), i32–i35.

[83] Hong, R., Mishra, V., and Govindasamy, P. *Factors associated with prevalent HIV infections among Ethiopian adults: Further analysis of the 2005 Ethiopia Demographic and Health Survey*. Macro International, 2008.

[84] Hsieh, F. Sample size formulae for intervention studies with the cluster as unit of randomization. *Statistics in medicine 7*, 11 (2006), 1195–1201.

[85] Joe, G. W., and Simpson, D. D. Hiv risks, gender, and cocaine use among opiate users. *Drug and Alcohol Dependence 37*, 1 (1995), 23–28.

[86] Johnson, K., and Way, A. Risk factors for hiv infection in a national adult population: evidence from the 2003 kenya demographic and health survey. *JAIDS Journal of Acquired Immune Deficiency Syndromes 42*, 5 (2006), 627–636.

[87] Joint United Nations Programme on HIV/AIDS. *AIDS Scorecards: Overview: UNAIDS Report on the Global AIDS Epidemic 2010*. UNAIDS, 2010.

[88] Kackar, R., and Harville, D. Unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communications in statistics-theory and methods 10*, 13 (1981), 1249–1261.

[89] Kackar, R., and Harville, D. Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association 79*, 388 (1984), 853–862.

[90] Kass, R., Tierney, L., and Kadane, J. Asymptotics in bayesian computation. *Bayesian statistics 3* (1988), 261–278.

[91] Kidanu, A., and Banteyerga, H. Aids and poverty. in dessalegn rahmato (ed.), some aspects of poverty in ethiopia. *Studies on Poverty 1* (2003), 32–44.

[92] Kloos, H., Mariam, D., and Lindtjorn, B. The aids epidemic in a low-income country: Ethiopia. *Human Ecology Review 14*, 1 (2007), 39.

[93] Kreft, I., De Leeuw, J., and de Leeuw, J. *Introducing multilevel modeling*. Sage London, 1998.

[94] Lagarde, E., Auvert, B., Chege, J., Sukwa, T., Glynn, J., Weiss, H., Akam, E., Laourou, M., Carael, M., Buve, A., et al. Condom use and its association with hiv/sexually transmitted diseases in four urban communities of sub-saharan africa. *Aids 15* (2001), S71.

[95] Landis, J. R., Lepkowski, J. M., Eklund, S. A., and Stehouwer, S. A. A statistical methodology for analyzing data from a complex survey: the first national health and nutrition examination survey. *Vital and health statistics. Series 2, Data evaluation and methods research*, 92 (1982), 1.

[96] Lee, E., and Forthofer, R. *Analyzing complex survey data*. No. 71. Sage Publications, Inc, 2006.

[97] Lester, F., Ayehunie, S., and Zewdie, D. Acquired immunodeficiency syndrome: seven cases in an addis ababa hospital. *Ethiopian medical journal 26*, 3 (1988), 139.

[98] Lester, G., Merritt, A., Neuwirth, L., Vetro-Widenhouse, T., Steible, C., and Rice, B. Effect of alpha 2-adrenergic, cholinergic, and nonsteroidal anti-inflammatory drugs on myoelectric activity of ileum, cecum, and right ventral colon and on cecal emptying of radiolabeled markers in clinically normal ponies. *American journal of veterinary research 59*, 3 (1998), 320.

[99] Lindley, D., and Smith, A. Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)* (1972), 1–41.

[100] Lindstrom, M., and Bates, D. Newtonraphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association 83*, 404 (1988), 1014–1022.

[101] Little, R. Regression with missing x's: a review. *Journal of the American Statistical Association 87*, 420 (1992), 1227–1237.

[102] Little, T., Lindenberger, U., and Nesselroade, J. On selecting indicators for multivariate measurement and modeling with latent variables: When" good" indicators are bad and" bad" indicators are good. *Psychological Methods 4*, 2 (1999), 192.

[103] Longford, N. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika 74*, 4 (1987), 817–827.

[104] Mamo, Y., Belachew, T., Abebe, W., Gebre-Selassie, S., and Jira, C. Pattern of widal agglutination reaction in apparently healthy population of jimma town, southwest ethiopia. *Ethiopian medical journal 45*, 1 (2007), 69.

[105] McCulloch, C., Searle, S., and Neuhaus, J. Generalized, linear, and mixed models, 2001.

[106] McFadden, D. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica: Journal of the Econometric Society* (1989), 995–1026.

[107] Measure, D. Macro international inc. *Uzbekistan 1996 Final Report* (2008).

[108] Mehret, M., Khodakevich, L., Zewdie, D., Gizaw, G., Ayehunie, S., Shanko, B., Gebrehewot, B., Gemeda, A., Adal, G., Abebe, T., et al. Hiv-1 infection among employees of the ethiopian freight transport corporation. *Ethiop J Health Dev 4*, 2 (1990), 177–182.

[109] Merson, M., O'Malley, J., Serwadda, D., and Apisuk, C. The history and challenge of hiv prevention. *The Lancet 372*, 9637 (2008), 475–488.

[110] Mishra, V., Assche, S., Greener, R., Vaessen, M., Hong, R., Ghys, P., Boerma, J., Van Assche, A., Khan, S., and Rutstein, S. Hiv infection does not disproportionately affect the poorer in sub-saharan africa. *Aids 21* (2007), S17.

[111] Mishra, V., Vaessen, M., Boerma, J., Arnold, F., Way, A., Barrere, B., Cross, A., Hong, R., and Sangha, J. Hiv testing in national population-based surveys: experience from the demographic and health surveys. *Bulletin of the World Health Organization 84*, 7 (2006), 537–545.

[112] MOH. Strategic plan for multi sectoral response against hiv/aids for 2004-2008. addis ababa (ethiopia):ministry of health. Tech. rep., Ethiopian Ministry of Health, 2004.

[113] Molla, M., Berhane, Y., and Lindtjørn, B. Traditional values of virginity and sexual behaviour in rural ethiopian youth: results from a cross-sectional study. *BMC Public Health 8*, 1 (2008), 9.

[114] Monmonier, M. Ho w to lie with maps.

[115] Morris, M., and Kretzschmar, M. Concurrent partnerships and the spread of hiv. *Aids 11*, 5 (1997), 641.

[116] Mosley, W. H., and Chen, L. C. An analytical framework for the study of child survival in developing countries. *Population and development review 10* (1984), 25–45.

[117] National Institute of Health. Adult male circumcision significantly reduced risk of acquiring hiv. trials kenya and uganda stopped early. Tech. rep., 2006.

[118] Negash, Y., Gebre, B., Benti, D., and Bejiga, M. A community based study on knowledge, attitude and practice (kap) on hiv/aids in gambella town, western ethiopia. *Ethiopian Journal of Health Development 17*, 3 (2004), 205–213.

[119] Nyblade, L., Pande, R., Mathur, S., MacQuarrie, K., Kidd, R., and Banteyerga, H. Disentangling hiv and aids stigma in ethiopia, tanzania and zambia. *Washington DC: International Center for Research on Women* (2008).

[120] Pérez V, R., Barrales C, I., Jara P, J., Palma R, V., and Ceballos M, A. Knowledge of hiv/aids among adolescents in chillán, chile. *Midwifery 24*, 4 (2008), 503–508.

[121] Pettifor, A. E., van der Straten, A., Dunbar, M. S., Shiboski, S. C., and Padian, N. S. Early age of first sex: a risk factor for hiv infection among women in zimbabwe. *Aids 18*, 10 (2004), 1435–1442.

[122] Plummer, M. Penalized loss functions for bayesian model comparison. *Biostatistics 9*, 3 (2008), 523–539.

[123] Raudenbush, S., and Bryk, A. *Hierarchical linear models: Applications and data analysis methods*, vol. 1. Sage Publications, Inc, 2002.

[124] Raudenbush, S., Bryk, A., Cheong, Y., and Congdon, R. Hlm 5, hierarchical linear and nonlinear modeling. chicago: Scientific software international, 2000.

[125] Robert, C. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer, 2007.

[126] Rogerson, P. Spatial models of search. *Geographical Analysis 14*, 3 (1982), 217–228.

[127] Roos, M., and Held, L. Sensitivity analysis in bayesian generalized linear mixed models for binary data. *Bayesian Analysis 6*, 2 (2011), 259–278.

[128] Rossi, P., and Allenby, G. Bayesian statistics and marketing. *Marketing Science 22*, 3 (2003), 304–328.

[129] Rubin, D. Multiple imputation for survey nonresponse, 1987.

[130] Rue, H., Martino, S., and Chopin, N. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology) 71*, 2 (2009), 319–392.

[131] Salvati, N., Tzavidis, N., Pratesi, M., and Chambers, R. Small area estimation via m-quantile geographically weighted regression. *Test* (2012), 1–28.

[132] Shelton, J., Halperin, D., Nantulya, V., Potts, M., Gayle, H., and Holmes, K. Partner reduction is crucial for balanced abc approach to hiv prevention. *Bmj 328*, 7444 (2004), 891–893.

[133] Smith, A., and Roberts, G. Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)* (1993), 3–23.

[134] Spiegelhalter, D., Best, N., Carlin, B., and Van Der Linde, A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*, 4 (2002), 583–639.

[135] Stephen, G. What is multi-level modeling for? *British journal of Educational studies 51(1)* (2003), 46–63.

[136] Stewart, C. *Multilevel modelling of event history data: comparing methods appropriate for large datasets.* PhD thesis, University of Glasgow, 2010.

[137] Stoneburner, R. Sexual partner reductions explain human immunodeficiency virus declines in uganda: comparative analyses of hiv and behavioural data in uganda, kenya, malawi, and zambia. *International journal of epidemiology* (2004), 1–10.

[138] Tanner, M., and Wong, W. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association 82*, 398 (1987), 528–540.

[139] Tarwireyi, F. Stigma and discrimination: coping behaviours of people living with hiv and aids in an urban community of mabvuku and tafara, harare, zimbabwe. *The Central African journal of medicine 51*, 7-8 (2005), 71.

[140] Tierney, L. Markov chains for exploring posterior distributions. *the Annals of Statistics* (1994), 1701–1728.

[141] Tronvoll, K. Human rights violations in federal ethiopia: When ethnic identity is a political stigma. *International Journal on Minority and Group Rights 15*, 1 (2008), 49–79.

[142] Tzavidis, N., Salvati, N., Pratesi, M., and Chambers, R. M-quantile models with application to poverty mapping. *Statistical Methods & Applications 17*, 3 (2008), 393–411.

[143] UNAIDS. Unaids report. Tech. rep., 1999.

[144] UNAIDS. Overview of the global aids epidemic: Report on the global aids epidemic. washington: Usa. Tech. rep., 2006.

[145] UNAIDS. Global report unaids report on the global aids epidemic. Tech. rep., 2010.

[146] UNAIDS. Unaids on the global aids epidemic. Tech. rep., 2010.

[147] UNAIDS. Unaids report. Tech. rep., 2010.

[148] UNAIDS, W. Report on the global aids epidemic. *Geneva, Switzerland* (2008).

[149] UNAIDS, W. Report on the global aids epidemic. *Geneva, Switzerland* (2008).

[150] UNFPA. Annual report. Tech. rep., 2008.

[151] Vermund, S. Casual sex and hiv transmission. *American Journal of Public Health 85*, 11 (1995), 1488–1489.

[152] Waller, L. A., Carlin, B. P., Xia, H., and Gelfand, A. E. Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association 92*, 438 (1997), 607–617.

[153] WB. Working for a world free of poverty. Tech. rep., The World Bank, 2011.

[154] Williams, B. G., Lloyd-Smith, J. O., Gouws, E., Hankins, C., Getz, W. M., Hargrove, J., De Zoysa, I., Dye, C., and Auvert, B. The potential impact of male circumcision on hiv in sub-saharan africa. *PLoS Medicine 3*, 7 (2006), e262.

[155] Wilson, D. Partner reduction and the prevention of hiv/aids. *Bmj 328*, 7444 (2004), 848–849.

[156] Yousafzai, A., Edwards, K., D'Allesandro, C., and Lindstrom, L. Hiv/aids information and services: The situation experienced by adolescents with disabilities in rwanda and uganda. *Disability and Rehabilitation 27*, 22 (2005), 1357–1363.

[157] Zuma, K., Gouws, E., Williams, B., and Lurie, M. Risk factors for hiv infection among women in carletonville, south africa: migration, demography and sexually transmitted diseases. *International journal of STD & AIDS 14*, 12 (2003), 814–817.

[158] Zuur, G., Garthwaite, P., and Fryer, R. Practical use of mcmc methods: lessons from a case study. *Biometrical journal 44*, 4 (2002), 433–455.