Modelling Obesity Risk Factors among Adult Females in South Africa via a GLMM: Classical and Bayesian Approaches

Telissa Pillay

December, 2016

Modelling Obesity Risk Factors among Adult Females in South Africa via a GLMM: Classical and Bayesian Approaches

by

Telissa Pillay

A dissertation submitted to the
University of KwaZulu-Natal
in fulfilment of the requirements
for the degree of
MASTER OF SCIENCE (STATISTICS)

SCHOOL OF MATHEMATICS, STATISTICS & COMPUTER SCIENCE

UNIVERSITY OF
KWAZULU-NATAL

WESTVILLE CAMPUS, DURBAN, SOUTH AFRICA

Declaration - Plagiarism

I, Telissa Pillay, declare that

1. The research reported in this dissertation, except where otherwise indicated, is my original research.
2. This dissertation has not been submitted for any degree or examination at any other university.
3. This dissertation does not contain other persons' data, pictures, graphs or other information, unless specifically acknowlegded as being sourced from other persons.
4. This dissertation does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then
   (a) their words have been re-written but the general information attributed to them has been referenced, or
   (b) where their exact words have been used, then their writing has been placed in italics and referenced.
5. This dissertation does not contain text, graphics or tables copied and pasted from the internet, unless specifically acknowledged, and the source being detailed in the dissertation and in the reference sections.

_____ _____

Telissa Pillay (Student) Date

_____ _____

Mr. M.J. Hammujuddy (Supervisor) Date

Disclaimer

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

Abstract

Obesity has reached epidemic proportions and has emerged as a serious public health concern in South Africa, especially among females. Obesity is a major contributor to the burden of non-communicable diseases, and thus, imposes substantial costs to the health care system as well as the economy. Therefore, understanding the risk factors associated with obesity is imperative in informing policy and developing effective prevention strategies. The etiology of obesity arises from a multilevel framework. In this study, a generalized linear mixed model (GLMM), which is a model-based statistical approach suitable for handling hierarchically structured discrete data, was employed to identify risk factors associated with obesity among adult females in South Africa. Obesity is classified as a body mass index (BMI) ≥ 30 $\mathrm{kg/m^2}$. Therefore, the response variable of interest was binary, indicating whether the female was obese or not. The GLMM was applied to a subset of the data set from the National Income Dynamics Study (NIDS) which is the first national panel study of individuals of all ages in South Africa. In fitting the GLMM, different classical and Bayesian estimation methods were used and different link functions for binary data were explored. Results were obtained using the Laplace approximation, adaptive Gauss-Hermite quadrature, penalized quasi-likelihood (PQL), Markov chain Monte Carlo (MCMC) and integrated nested Laplace approximation (INLA) methods. This study confirmed that these methods differ in terms of computational speed. Moreover, in identifying the key determinants of obesity among adult females in South Africa, this study found that age, ethnicity, marital status, education level, employment status, household income, household expenditure on food and geographical type of residence are highly significant contributing risk factors.

Acknowledgements

It is with immense gratitude that I acknowledge the guidance and encouragement of my supervisor, Mr. M. J. Hammujuddy. This work could not have been accomplished without his continued support. His sage advice and insightful criticisms throughout my undergraduate and Honours degrees have been invaluable to me.

To the staff members in the Discipline of Statistics at the University of KwaZulu-Natal (Westville Campus), thank you for the support and encouragement. A special thanks goes to Ms. Danielle Roberts for all her advice and to Mr. K. Chinhamu for his constant motivation.

I hereby acknowledge the financial assistance of the National Research Foundation (NRF) towards this research. Opinions expressed, and conclusions arrived at, are those of mine and are not necessarily to be attributed to the NRF.

My deep and sincere gratitude to my family for always believing in me. I am grateful to my brother for his encouragement and always being there for me. I am deeply indebted to my parents for their selfless and unparalleled love and support, and for the countless sacrifices they have made so that I could be where I am today. This journey would not have been possible if not for them, and I am forever grateful. This milestone is dedicated to them.

Finally, to my one and only, Ashlyn, I am truly grateful for all the love, help and support you have given me. Thank you for always being so patient and understanding. I could not imagine doing this degree without you, and I look forward to dive into our future adventures.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Obesity has become a global epidemic with more than 500 million obese adults worldwide. According to the World Health Organization (2014), the prevalence of obesity has more than doubled since 1980 and the number of obese adults is expected to rise to more than a billion by 2030. The escalating prevalence of obesity is associated with increased morbidity and mortality from comorbidities such as cardiovascular disease, diabetes mellitus (type 2) and various types of cancer, and thus, imposes a significant economic burden on already strained healthcare systems (Malik et al., 2012; McCormick et al., 2012). Furthermore, obesity imposes substantial costs to the economy, such as economic disenfranchisement, loss of productivity, reduction in tax revenue, and increased government expenditure on incapacity and unemployment benefits (McCormick et al., 2012; Some et al., 2016).

Globally, the prevalence of obesity is higher among females. In regions such as Africa and South East Asia, the prevalence among females is more than double that among males (World Health Organization, 2014). In the past, the African continent has been grappled with undernutrition and the burden of infectious diseases such as HIV and tuberculosis. However, in recent years, the rapid rise in the prevalence of obesity and associated comorbidities poses a major concern for the continent (Micklesfield et al., 2013). In order to inform policy and develop effective prevention strategies to reduce the prevalence of obesity, it is imperative to identify and understand the risk factors associated with the epidemic (Affenito et al., 2012; Sartorius et al., 2015).

1.1 Background

South Africa (SA) is the southernmost country on the African continent, and has an area of 1,22 million square kilometres (Stats SA, 2012). On the south of SA lies 2 798 kilometres of coastline that stretches along the Atlantic Ocean and the Indian Ocean (WWF-SA, 2016), to the north lies the neighbouring countries of Namibia, Botswana and Zimbabwe, to the north east are Mozambique and Swaziland, and enclaved is the Kingdom of Lesotho (Mofuoa, 2015). SA is made up of nine provinces which are administratively divided into 52 district councils. The district councils consist of 8 metropolitan and 44 district municipalities, as seen in Figure 1.1[1].

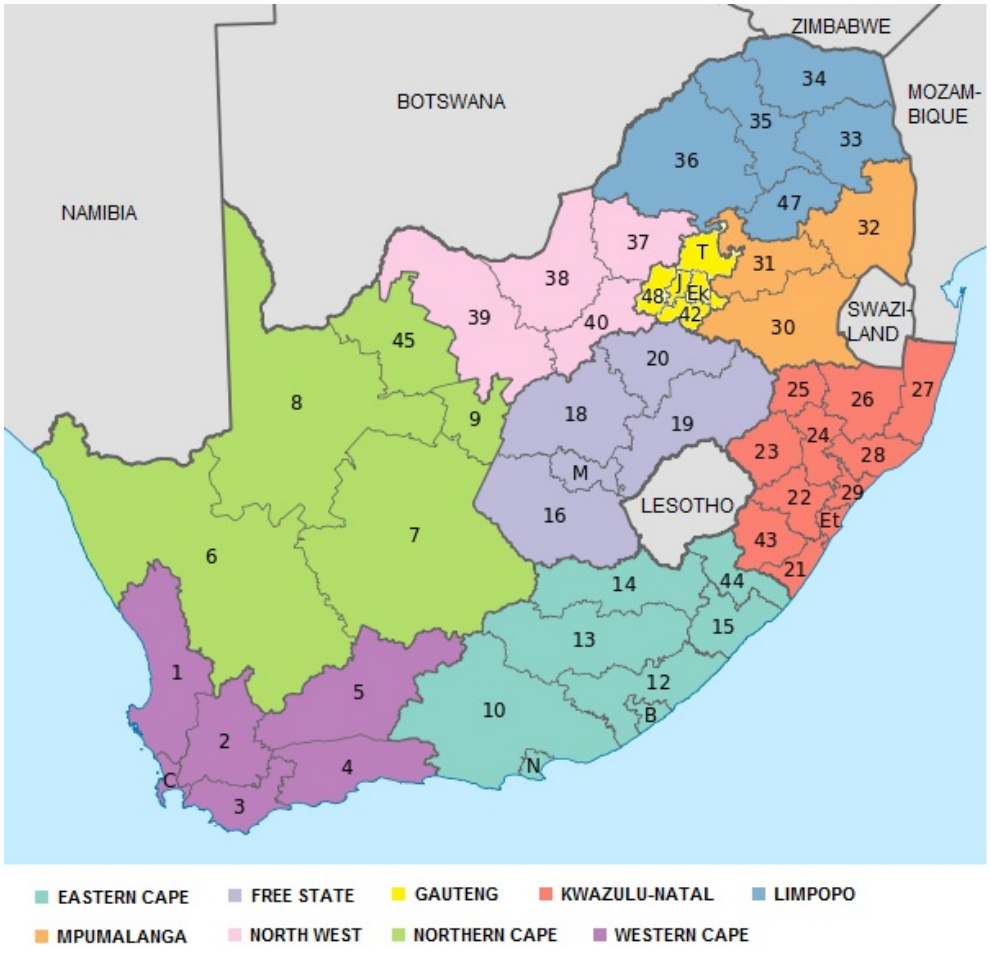


Figure 1.1: Map of South Africa

[1]District municipalities are numbered according to their district codes and metropolitan municipalities are abbreviated according to their names.

The population of SA is currently estimated at 54,96 million, and with a population of 51,77 million in 2011, the annual population growth rate is estimated at 1.65% (Stats SA, 2012, 2015). Approximately 51% of the population are females. There are four population groups in SA: Africans, Whites, Asians/Indians and Coloureds. Approximately 80% of the population are Africans (Stats SA, 2015).

SA is categorized as a middle-income country and has one of the largest economies in Africa (World Bank Group, 2016). Moreover, the country is undergoing a rapid epidemiological transition and has the highest prevalence of obesity, predominantly among females, in the Sub-Saharan region (Goedecke et al., 2006; Micklesfield et al., 2013). In 2008, the prevalence of obesity among South African adult females was at 33% (Ardington & Gasealahwe, 2012). A study undertaken in 2013 revealed that this figure had risen to 42%, and was approximately three times the obesity prevalence among males (The GBD 2013 Obesity Collaboration, 2014).

The SA government has recently taken heed to the obesity epidemic and has introduced a national strategy for the prevention and control of obesity while encouraging additional research in the field (Department of Health: RSA, 2015). The strategy aims to reduce the prevalence of obesity by 10% between 2015-2020. Studies have shown that the high obesity prevalence among females in SA is attributed to factors such as African ethnicity, urban residence (Puoane et al., 2002), being married, lack of exercise (Alaba & Chola, 2014), high income quintiles (Case & Menendez, 2009), high household expenditure on food and crime (Sartorius et al., 2015). Other studies (Butzlaff & Minos, 2016; Cois & Day, 2015; Malhotra et al., 2008) have found significant associations between adult obesity prevalence and factors such as age, education, employment status, smoking and alcohol consumption. Although these studies have identified this wide variety of risk factors associated with obesity, there is still a need to further investigate these factors, especially among South African females, in order to evaluate and improve existing governmental interventions as well as informing and developing new strategies.

1.2 Objectives

Obesity is a problem arising from a complex system in which individual behaviour is influenced not only by individual factors, but also by multiple levels of socioenvironmental factors, such as institutions, families and neighbourhoods, social networks, culture and the physical environment, that are heterogeneous and interdependent. Modelling data from such a complex system requires methods that address multiple factors and levels as well as accounting for the apparent heterogeneity in the data (Huang et al., 2009; Lakerveld et al., 2012). Such data analysis methods include both design-based and model-based approaches. For design-based methods, such as survey logistic regression, sampling weights are incorporated in the model to account for the complex sampling design. This method, however, is useful when only inferences on certain explanatory variables are of interest. This method also assumes the observations are independent. When heterogeneity that is attributable to multiple levels of sampling in the data is also of interest, a model-based approach is more appealing (Heeringa et al., 2010). Thus, this method accounts for possible correlations that may exist among the observations. Therefore, in this study, a generalized linear mixed model (GLMM), which is a model-based statistical approach suitable for handling hierarchically structured discrete data, was employed. The objectives of this study were to:

- Account for heterogeneity in the distribution of obesity among females in SA by fitting a GLMM using data from the National Income Dynamic Study (NIDS).
- Identify significant risk factors associated with obesity among females in SA.
- Examine and compare the different classical and Bayesian estimation methods used in fitting a GLMM.

1.3 Overview

In this chapter, we introduced some background information. In Chapter 2, the National Income Dynamic Study (NIDS) is introduced and exploratory data analyses are performed on the NIDS data set. Chapter 3 introduces GLMMs and some classical estimation methods are discussed, and fitted to the data set. Chapter 4 gives an overview of the Bayesian approaches where the Markov chain Monte Carlo (MCMC) and the integrated nested Laplace approximation (INLA) methods are discussed, and applications thereof are illustrated. The last chapter discusses results obtained from the different estimation methods, highlights conclusions and presents possible areas for further study.

Chapter 2

National Income Dynamics Study

This study analyzes data from the National Income Dynamics Study (NIDS) which is the first national panel study of individuals of all ages in South Africa. NIDS was established by the South African Presidency with the aim of tracking changes in the well-being of the South African population. The NIDS surveys were conducted by the South African Labour and Development Research Unit (SALDRU) based in the School of Economics at the University of Cape Town where ethical approval was granted by the Faculty of Commerce Ethics Committee (Leibbrandt et al., 2009).

2.1 Sampling procedure

The NIDS surveys were carried out in 2008 (Wave 1), 2010/2011 (Wave 2) and 2012 (Wave 3). A stratified, two-stage random cluster sample design was employed to select households to be included in the sample at baseline. The sample was stratified according to the 52 district councils in SA from which clusters of dwelling units were systematically drawn. The target population was private households, residents in workers' hostels, convents and monasteries. Other collective living quarters were excluded from the sampling frame (Leibbrandt et al., 2009).

At baseline, 28 226 individuals from 7 296 households were successively interviewed, resulting in a household level response rate of 69% and an individual response rate within households of 93%. The second and third waves of the NIDS survey provides data on 28 551 individuals from 6 787 households and 32 633 individuals from 8 040 households, respectively (de Villiers et al., 2013). In this study, data from the third

wave were used. However, the analysis was restricted to females aged $\geq$ 15 years old. Females who were pregnant at the time of interview were excluded from the sample as pregnancy influences weight. Furthermore, those with missing data entries were also excluded, resulting in a complete case analysis. Thus, the final sample in this study was made up of 10 411 females from 6 459 households.

2.2 Data collection

The selected households were visited and interviewed by trained fieldworkers. Three questionnaires were administered: household, adult and child questionnaires. A proxy questionnaire was also used for those individuals who were unavailable or unable to answer their own adult questionnaire. These questionnaires were designed to collect data on a wide range of information that includes basic demographics, education, employment, health (including anthropometric data), household income and expenditure.

Quality controllers were employed to verify and check the completeness of the data obtained during fieldwork. An in-field and telephonic call-back strategy were used to validate the professionalism of fieldworkers ensuring that the correct households were being interviewed, gaining insight on refusals to participate as well as obtaining key missing data in cases that did not warrant the questionnaires being sent back to field (Leibbrandt et al., 2009).

2.3 Variables of interest

The body mass index (BMI) is defined as an individual's body mass (in kilograms) divided by the square of her height (in metres). In the NIDS survey, weight and height measurements for all individuals were taken and their individual BMIs were computed. The response variable of interest was obesity which is classified as BMI $\geq 30\ \mathrm{kg/m^2}$. The independent variables were based on the paper by Sartorius et al. (2015). The demographic variables were age, population group and marital status. The lifestyle variables included exercise frequency, smoking, alcohol consumption, depression and total household expenditure on food. The variables education level,

employment status and total household income were categorized as socio-economic variables. The environmental variables were geographical type and crime.

The independent variables that were not categorical were recoded. These include age, total household expenditure on food and total household income. Physical exercise was coded 0 if the female exercised less than once a week, 1 if exercised one to two times a week, and 2 if exercised more than twice a week. Out of the five smoking-related questions in the adult questionnaire, current smoking status was used. Alcohol consumption was dichotomous and coded 0 if the female never or no longer drinks alcohol, and 1 if not. The 10-item Center for Epidemiologic Studies Depression Scale (CES-D) was used in screening for depressive symptoms. A total score of 10 or higher suggests the presence of significant depressive symptoms (Zhang et al., 2012). Geographical type was categorized into urban, traditional and farm areas. Urban areas are defined as built-up areas established through cities, towns and suburbs. Traditional areas are communally-owned land under the jurisdiction of traditional leaders, and farms are land used for commercial farming. The variable crime was dichotomous and based on whether or not the individual perceived crime (burglaries, muggings or thefts) to be common in their neighborhood or not. In the following section, we explore the NIDS data set descriptively.

2.4 Exploratory data analysis

Of the 10 411 females in the sample, 3 601 had a BMI of 30 $\mathrm{kg/m^2}$ or greater, resulting in an observed obesity prevalence of 34.6%. Table 2.1 displays the distribution of females in the sample according to the different independent variables. The percentage of females in the sample ranged from 9.9% between the ages of 65 years and older to 28.0% between the ages of 15 and 24 years. Figure 2.1 shows that the observed prevalence of obesity peaked at 50.3% in the age group 55 to 64 years. Furthermore, there was an increase in the observed prevalence of obesity as age increased, followed by a decline after the age of 65 years.

Table 2.1: Percentage of females according to the different explanatory variables

Variable	Percentage
Age	
15-24	28.0
25-34	20.2
35-44	16.2
45-54	14.7
55-64	11.0
65+	9.9
Population group	
African	82.9
Coloured	13.5
Asian/Indian	1.0
White	2.6
Marital status	
Married	23.8
Living with partner	6.3
Widow	11.5
Divorced or separated	2.5
Never married	55.9
Exercise frequency	
Less than once a week	87.3
1 to 2 times a week	6.4
More than twice a week	6.2
Current smoker	
No	92.6
Yes	7.4
Alcohol consumption	
No	86.8
Yes	13.2
Depression	
No	73.6
Yes	26.4

Continued on next page

Table 2.1 – *Continued from previous page*

Variable	Percentage
Household expenditure on food	
Quartile I	25.0
Quartile II	26.5
Quartile III	24.1
Quartile IV	24.4
Education	
No schooling	12.1
Primary	21.5
Secondary	64.7
Tertiary	1.7
Employment status	
Unemployed	70.7
Employed	29.3
Household income quintile	
I	21.2
II	20.9
III	19.8
IV	19.6
V	18.5
Geographical type	
Urban	45.5
Traditional	46.6
Farms	7.9
Crime	
No	54.3
Yes	45.7

The majority of the females were African, making up more than eighty percent of the sample. Only 1.0% of the females were Asian/Indian. Figure 2.2 shows the observed prevalence of obesity corresponding to the different population groups in the study. The observed prevalence according to population group ranged from 26.9% to 41.7%, with the observed prevalence among African and Coloured females not differing much. Although the observed prevalence of obesity was highest among white females, this group only made up 2.6% of the sample.

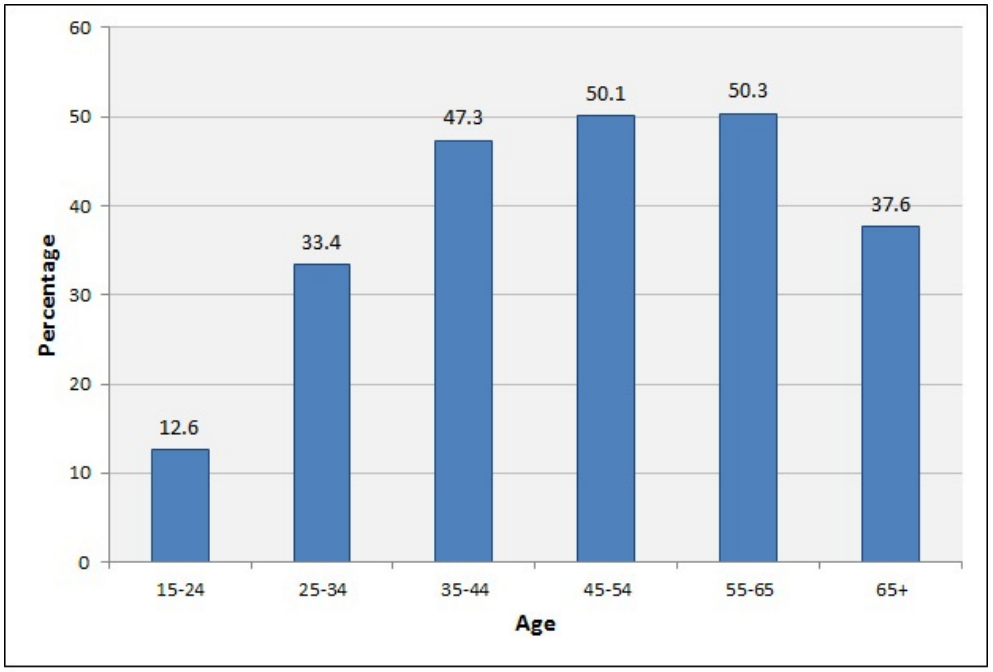


Figure 2.1: Observed prevalence of obesity according to the different age groups

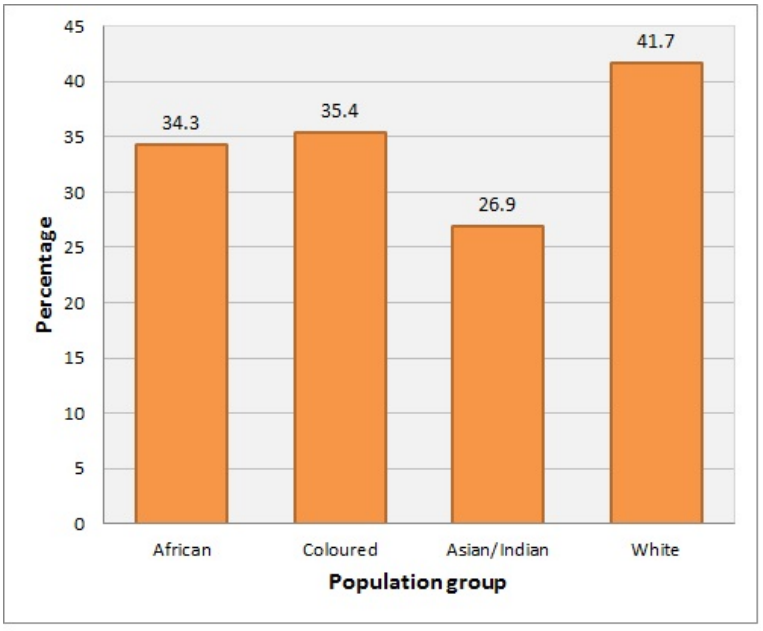


Figure 2.2: Observed prevalence of obesity according to the different population groups

More than half of the females in the sample (55.9%) reported never being married, followed by married (23.8%) and widowed (11.5%). Figure 2.3 shows that those married had the highest observed obesity prevalence at 52.7%, followed by those

divorced or separated with an observed prevalence of 46.3%. Those reported never being married had the lowest observed prevalence at 25.6%. Only a small proportion of the females in the sample (12.6%) reported having exercised at least once a week, with the majority (87.3%) having exercised less than once a week. Of those who exercised less than once a week, 35.2% had a BMI $\geq$ 30 $\mathrm{kg/m^2}$, as seen in Figure 2.4. The observed prevalence among those females who exercised one to two times a week (28.9%) and those who exercised more than twice a week (31.4%) was not much different.

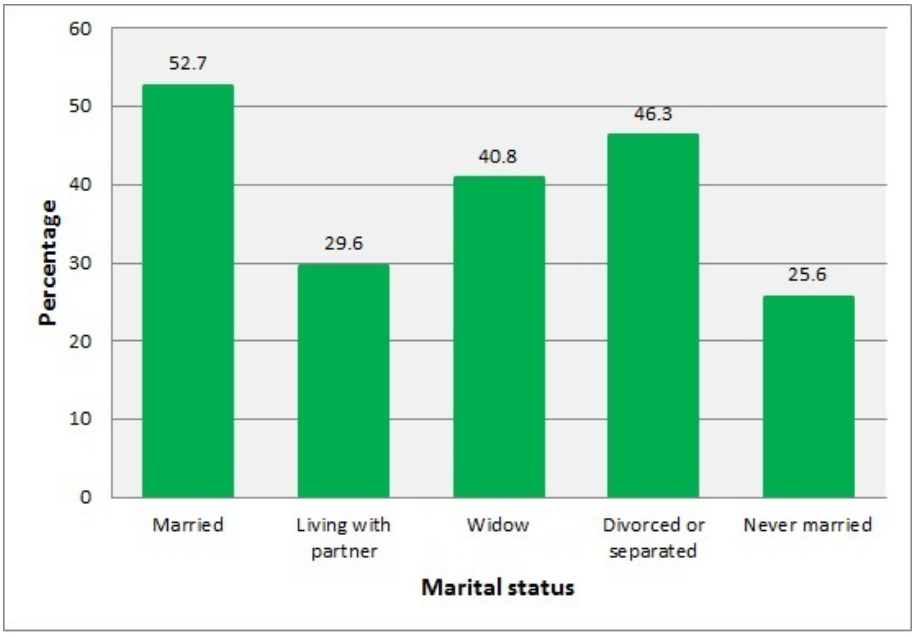


Figure 2.3: Observed prevalence of obesity according to marital status

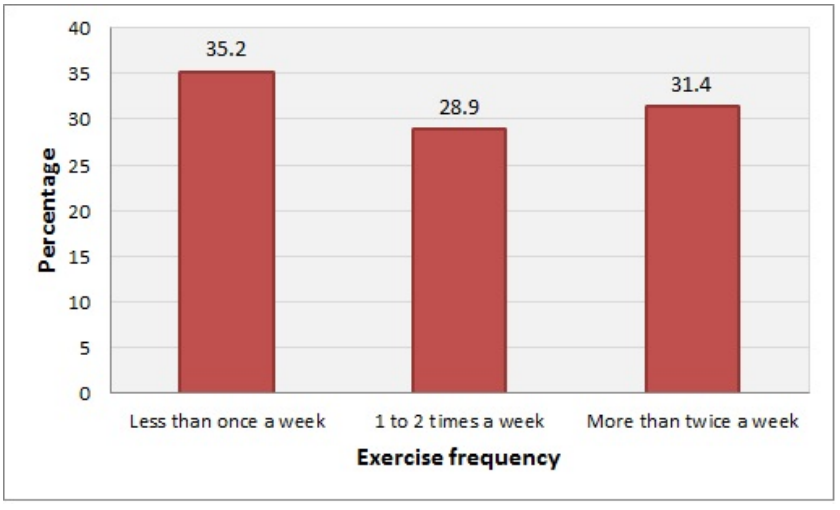


Figure 2.4: Observed prevalence of obesity according to exercise frequency

Table 2.1 shows that 7.4% of the female study sample were smokers and 13.2% consumed alcohol. Figure 2.5 shows the observed prevalence of obesity according to these two lifestyle choices. The observed prevalence among non-smokers (35.1%) was slightly higher compared to smokers (28.2%). Similarly, the observed prevalence among those who did not consume alcohol (35.3%) was slightly higher than those who did (29.9%). More than a quarter of the females in the sample (26.4%) were classified as suffering from depression according to the CES-D. Figure 2.6 reveals that the observed prevalence of obesity among those classified as suffering from depression (35.6%) and those not classified as suffering from depression (34.2%) was not much different.

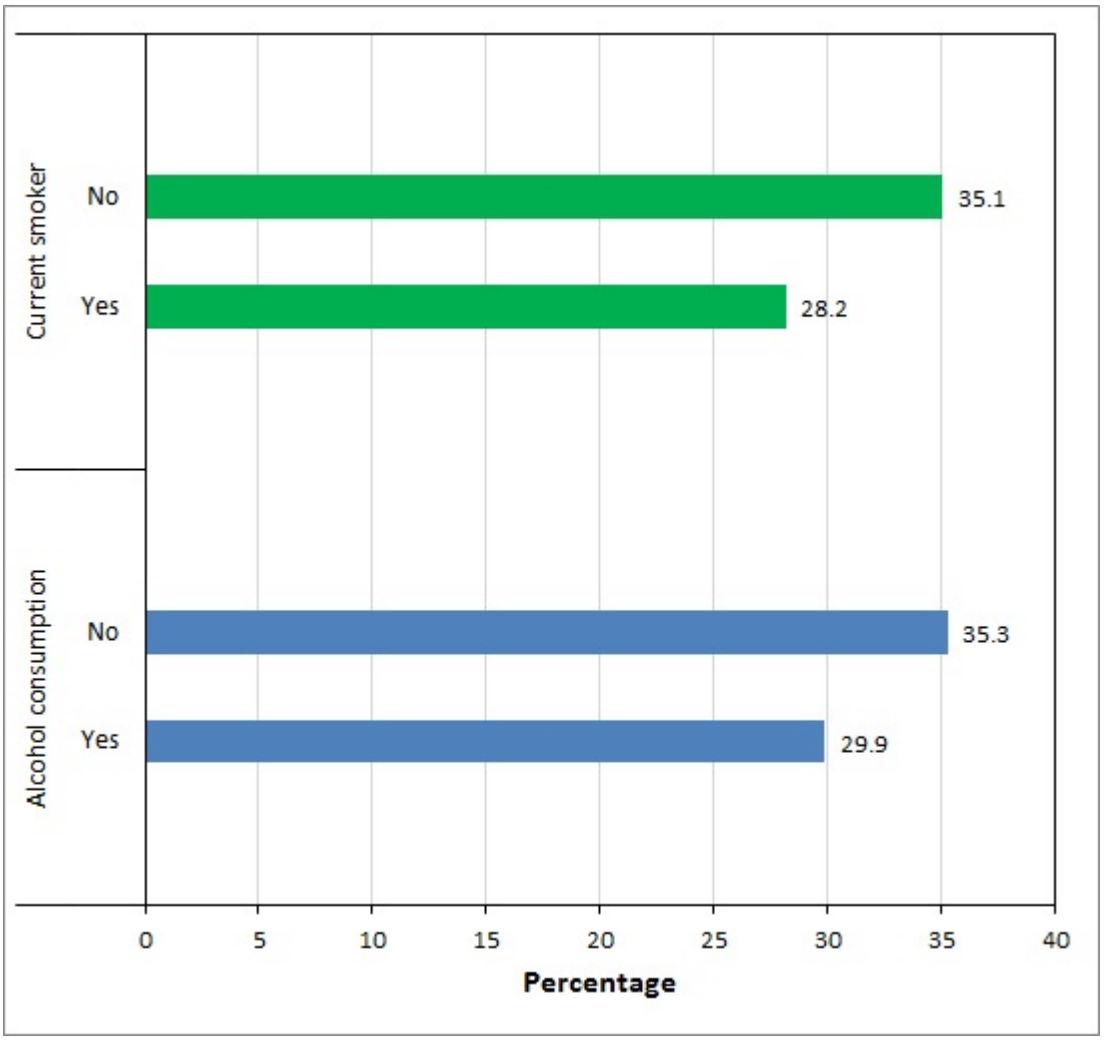


Figure 2.5: Observed prevalence of obesity according to smoking status and alcohol consumption

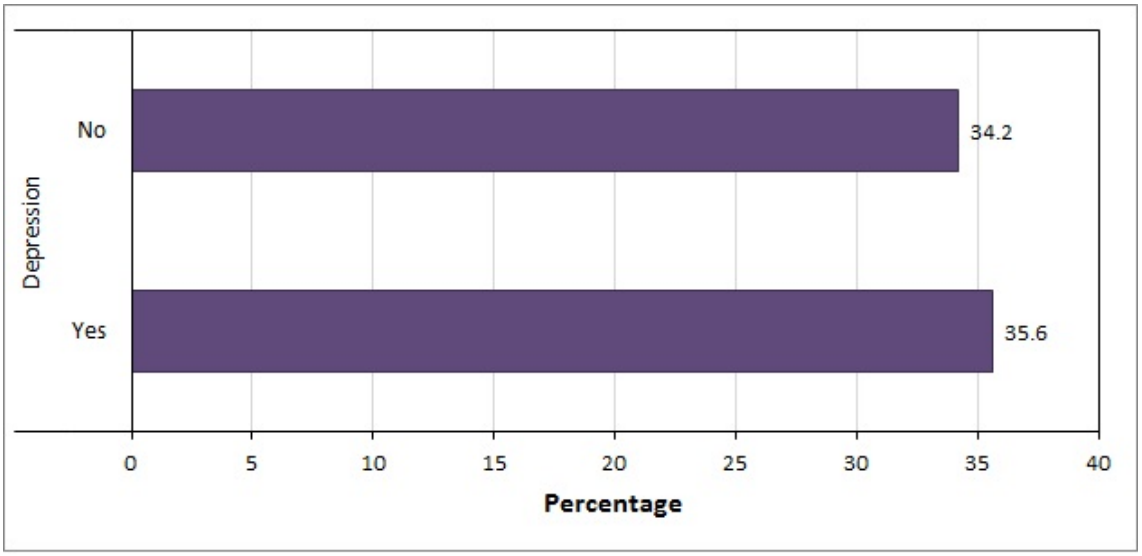


Figure 2.6: Observed prevalence of obesity according to depression

The percentage of females in the sample with a total household expenditure on food within the different quartiles ranged from 24.1% to 26.5%. The observed prevalence of obesity according to the different quartiles ranged from 30.0% to 39.6%, as seen in Figure 2.7. Those females with a total household expenditure on food within the highest quartile had the highest observed prevalence, with the observed prevalence among those with a total expenditure on food within the second and third quartiles not differing by much.

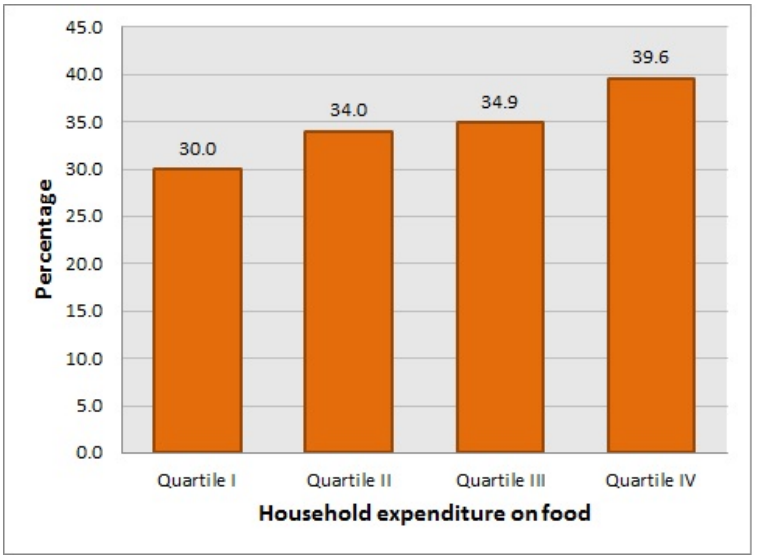


Figure 2.7: Observed prevalence of obesity according to household expenditure on food

Most of the females in the sample (64.7%) reported having received up to secondary education followed by primary education (21.5%). As seen in Figure 2.8, those females with secondary education had the lowest observed obesity prevalence at 32.4%. In contrast, those having received tertiary education had the highest observed prevalence at 47.8%. However, this group only made up 1.7% of the sample. A total of 70.7% of the females in the sample reported being unemployed. Figure 2.9 shows the observed prevalence of obesity according to employment status. Out of the females who were unemployed, 31.1% were classified as obese. The observed prevalence among those females who were employed was slightly higher at 42.9%.

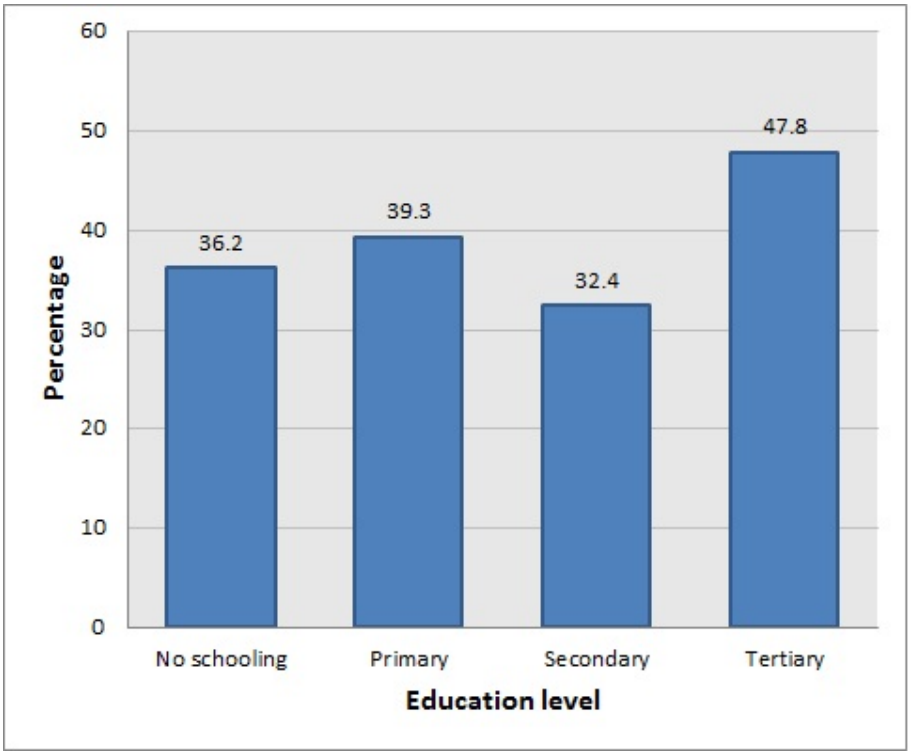


Figure 2.8: Observed prevalence of obesity according to education level

The percentage of females in the sample according to the different household income quintiles ranged from 18.5% to 21.2%. Females belonging to the highest household income quintile had the highest observed obesity prevalence at 43.3% as seen in Figure 2.10. The observed prevalence among those females belonging to the first and second household income quintile, and those belonging to the third and fourth household income quintile were not much different.

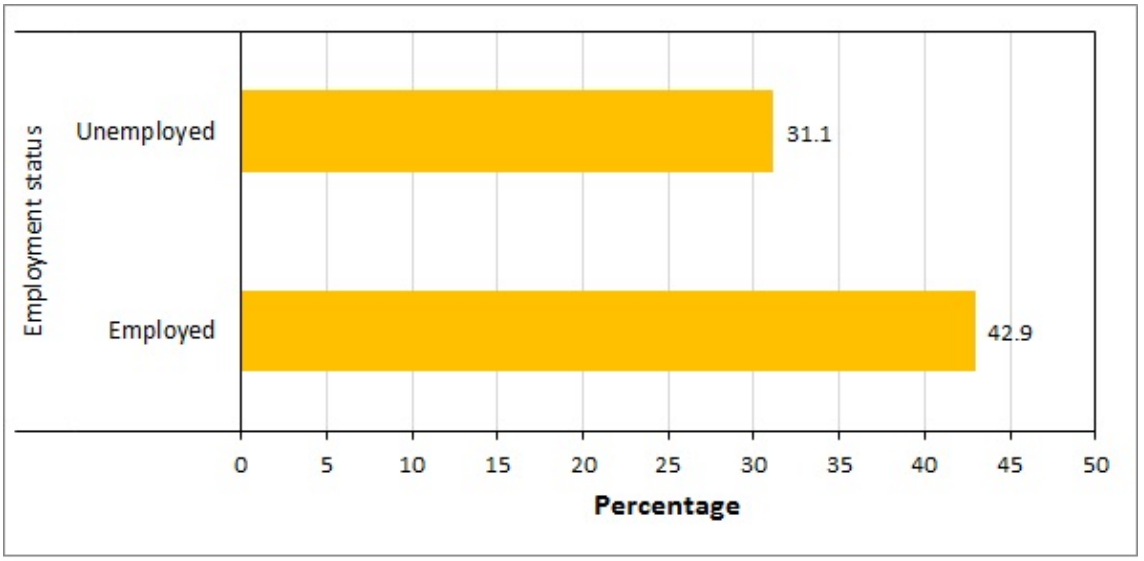


Figure 2.9: Observed prevalence of obesity according to employment status

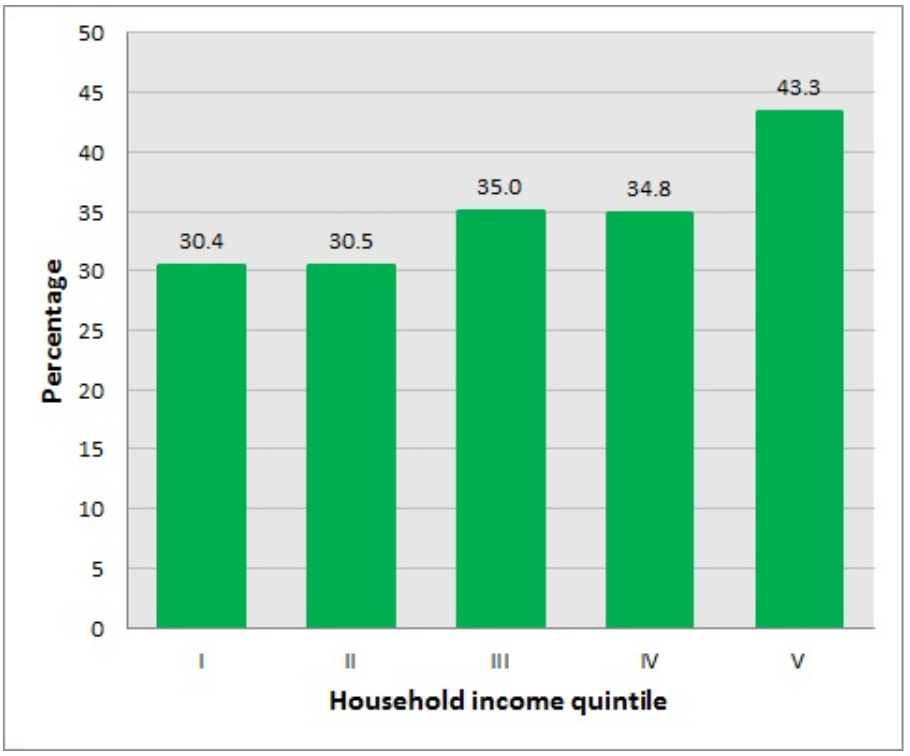


Figure 2.10: Observed prevalence of obesity according to household income

Females who lived in urban and traditional areas made up 45.5% and 46.6% of the sample, respectively. Only 7.9% of the female study sample lived on farms. Figure 2.11 shows that the observed prevalence of obesity was highest among those liv-

ing in traditional areas (38.7%) followed by farms (31.8%). However, the observed prevalence among those living in farms and those in urban areas was not much different. Out of the total number of females in the sample, 54.3% reported living in areas where crime was common. Figure 2.12 shows that the observed prevalence among this group was 34.4%, and the observed prevalence among those reported living in areas where crime was not common was 34.8%. This suggests that crime is not a significant determinant of obesity in females in SA.

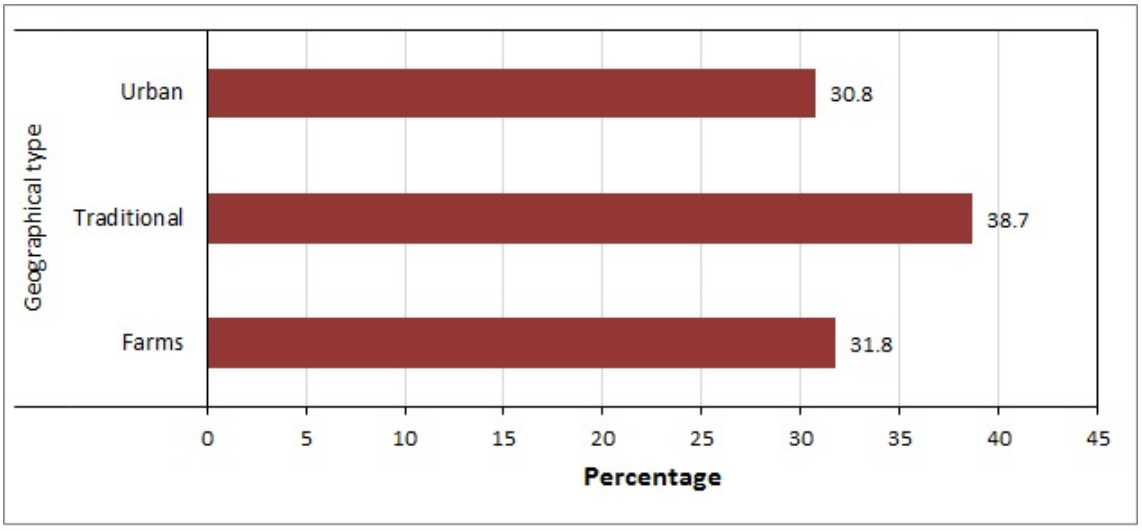


Figure 2.11: Observed prevalence of obesity according to geographical type

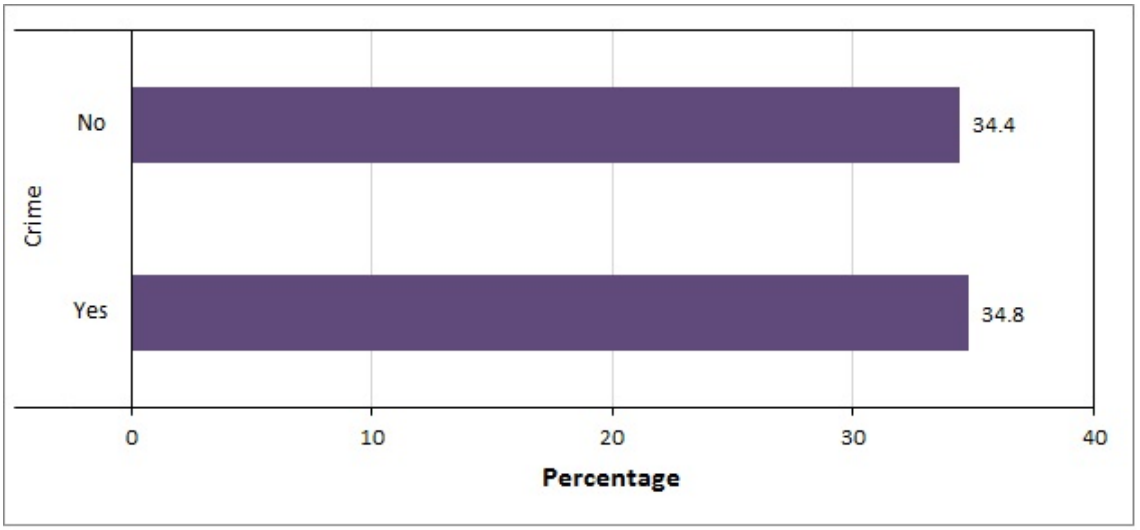


Figure 2.12: Observed prevalence of obesity according to crime

2.4.1 Test of Association

A chi-square test was used to test for associations between the explanatory variables and obesity among females in the NIDS data. The results of this test, summarized in Table 2.2, show that there is a positive association between obesity among females and age, population group, marital status, smoking status, alcohol consumption, household expenditure on food, education level, employment status, household income and geographical type of residence (all p-values<0.05). No association was found between obesity among females and exercise frequency (p-value=0.1053), depression (p-value=0.1780) and crime (p-value=0.6787).

Table 2.2: Cross tabulation of obesity status and explanatory variables

Effect	Numerator DF	F-Value	P-Value
Age	5	1040.96	<0.0001
Population group	3	9.40	0.0245
Marital status	4	611.73	<0.0001
Exercise frequency	2	14.21	0.1053
Current smoker	1	15.08	<0.0001
Alcohol consumption	1	15.44	<0.0001
Depression	1	1.81	0.1780
Household expenditure on food	3	52.79	<0.0001
Education	3	52.19	<0.0001
Employment status	1	132.40	<0.0001
Household income quintile	4	98.94	<0.0001
Geographical type	2	69.33	<0.0001
Crime	1	0.17	0.6787

To further explore the relationship between obesity among females and the demographic, lifestyle, socioeconomic and environmental variables, a survey logistic model was fitted to the NIDS data. The procedure and results of this analysis are presented in Appendix D. This method, however, is design-based, and thus does not account for possible correlations in the observations. In the case of obesity it is necessary to account for the effects of clustering since individuals from the same cluster tend to be more homogeneous compared to those from different clusters (Huang et al., 2009). Therefore, the next two chapters focuses on GLMMs. Both classical and Bayesian estimation methods are discussed, and applications thereof are illustrated.

Chapter 3

Generalized Linear Mixed Models: The Classical Approach

3.1 Introduction

The simplest statistical model is the classical linear model (LM) where responses are assumed to be independently Gaussian distributed with constant variance and where the mean of the response variable is a linear combination of explanatory variables. Although LMs are versatile and robust, they are not suitable for modelling discrete data (Rencher & Schaalje, 2008). The analysis of discrete data can be performed within the framework of generalized linear models (GLMs). This class of models, introduced by Nelder & Wedderburn (1972), are an extension of LMs. In GLMs, the assumption of independence among responses is maintained but their distribution belong to the exponential family. Furthermore, a suitable transformation of the mean results in a linear combination of explanatory variables and the variance is a function of the mean (McCullagh & Nelder, 1989). The linear combination forms the *linear predictor* and is referred to as the *systematic component*, while the response variable is known as the *random component* (Agresti, 2015).

Another extension of the LM is the linear mixed model (LMM). In LMs, explanatory variables may be continuous or categorical. The categorical variables, commonly known as factors, usually comprise of several fixed levels, and may be crossed or nested (Searle et al., 2006). In a statistical analysis, the focus is essentially on the fixed effects of these levels on the response variable. By contrast, an LMM contains

random effects as well as fixed effects. Random effects are due to an infinite set of levels of a factor from which only a random sample of those levels are considered to be present in the data (McCulloch et al., 2008). LMMs are often used in the modelling of hierarchical or multilevel data (Ker, 2014) where observations are obtained within clusters. Moreover, observations within the same cluster tend to be correlated. Thus, in LMMs, the assumption of independence among observations is relaxed. The correlation structure of the observations is accounted for by the inclusion of random effects in the model (Hedeker, 2005).

The generalized linear mixed model (GLMM) is a combination of the aforementioned extensions of the LM: The GLM and LMM (Breslow, 2003). In GLMMs, an unobserved vector of random effects is introduced into the linear predictor of a GLM. Moreover, the observations are assumed to be conditionally independent given the random effects (Breslow & Clayton, 1993). In this chapter, we present an overview of GLMMs from a classical perspective.

3.2 GLMM

Let $\boldsymbol{y}_k = (y_{k1}, \ldots, y_{kn_k})'$ denote a vector of responses where y_{ki} represents the i^{th} response from the k^{th} cluster; $i = 1, \ldots, n_k$ and $k = 1, \ldots, K$. Let $\boldsymbol{\gamma}_k = (\gamma_{k1}, \ldots, \gamma_{kc})'$ denote a q-vector of random effects, where γ_{kj} $(j = 1, \ldots, c)$ is the j^{th} random effect associated with cluster k, having q_j levels such that $q = \sum_{j=1}^{c} q_j$. The random effects are assumed to be independently Gaussian distributed with $E(\boldsymbol{\gamma}_k) = \mathbf{0}$ and $Var(\boldsymbol{\gamma}_k) = \boldsymbol{G}(\boldsymbol{\varphi})$; that is, $\gamma_k \sim N(\mathbf{0}, \boldsymbol{G})$, where $\boldsymbol{G} \equiv \boldsymbol{G}(\boldsymbol{\varphi})$ with $\boldsymbol{\varphi}$ being a $c \times 1$ vector of variance components (Lin & Breslow, 1996; Zhang & Lin, 2008).

Given the vector $\boldsymbol{\gamma}_k$, the responses y_{ki} are assumed to be conditionally independent with density belonging to the exponential family of distributions which, in its canonical form (McCulloch et al., 2008), is given by

$$f(y_{ki}|\theta_{ki}, \phi) = \exp\left\{\frac{y_{ki}\,\theta_{ki} - b(\theta_{ki})}{a_{ki}(\phi)} + c(y_{ki}, \phi)\right\} \tag{3.1}$$

where θ_{ki} is called the *natural parameter*, ϕ is referred to as the *dispersion* or *scale pa-*

rameter and $a_{ki}(\phi)$, $b(\theta_{ki})$ and $c(y_{ki}, \phi)$ are known functions. The function $a_{ki}(\phi)$ has the form $a_{ki}(\phi) = \phi/\tau_{ki}$, where τ_{ki} is a known weight associated with y_{ki} (Agresti, 2015). It can be shown that the conditional mean and conditional variance of y_{ki} are, respectively, given by

$$E(\, y_{ki}|\boldsymbol{\gamma}_k) = \mu_{ki}^{\boldsymbol{\gamma}_k} \tag{3.2}$$

and

$$Var(\, y_{ki}|\boldsymbol{\gamma}_k) = \phi\, \tau_{ki}^{-1}\, V(\, \mu_{ki}^{\boldsymbol{\gamma}_k}) \tag{3.3}$$

where $V(\, \mu_{ki}^{\boldsymbol{\gamma}_k})$ is known as the *(conditional) variance function*. For binary data, the response y_{ki} takes on the value 1 if the outcome is a success, and 0 otherwise. Thus, y_{ki} follows a Bernoulli distribution with conditional mean given by Equation 3.2 and conditional variance

$$Var(\, y_{ki}|\boldsymbol{\gamma}_k) = \phi\, \tau_{ki}^{-1}\, \mu_{ki}^{\boldsymbol{\gamma}_k}(1 - \mu_{ki}^{\boldsymbol{\gamma}_k})$$

In GLMs, a monotone and twice differentiable function g (Wedderburn, 1976) is used to transform the mean of the response in order to achieve a linear relationship between the mean and the systematic component (McCullagh & Nelder, 1989). Similarly, in GLMMs, the conditional mean of y_{ki} is transformed such that

$$g(\, \mu_{ki}^{\boldsymbol{\gamma}_k}) = \eta_{ki}^{\boldsymbol{\gamma}_k} = \boldsymbol{x}'_{ki}\, \boldsymbol{\beta} + \boldsymbol{z}'_{ki}\boldsymbol{\gamma}_k \tag{3.4}$$

Or, more compactly, as

$$g(\boldsymbol{\mu}_k^{\boldsymbol{\gamma}_k}) = \boldsymbol{\eta}_k^{\boldsymbol{\gamma}_k} = \mathbf{X}'_k\, \boldsymbol{\beta} + \mathbf{Z}'_k\boldsymbol{\gamma}_k \tag{3.5}$$

where

- $\boldsymbol{\mu}_k^{\boldsymbol{\gamma}_k}$ is an $n_k \times 1$ conditional mean vector, with $g(\boldsymbol{\mu}_k^{\boldsymbol{\gamma}_k}) = (g(\mu_{k1}^{\boldsymbol{\gamma}_k}), \ldots, g(\mu_{kn_k}^{\boldsymbol{\gamma}_k}))'$.
- $\mathbf{X}_k$ is an $n_k \times p$ design matrix associated with fixed effects vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$.
- $\mathbf{Z}_k = [\mathbf{Z}_{k1}, \ldots, \mathbf{Z}_{kc}]$, is an $n_k \times q$ design matrix associated with the q-vector $\boldsymbol{\gamma}_k$, where $\mathbf{Z}_{kj}$ is an $n_k \times q_j$ design matrix for the j^{th} random effect.

The function g is called the *link function* and $\boldsymbol{\eta}_k^{\boldsymbol{\gamma}_k}$ is known as the *linear predictor*.

Equation 3.5 can also be written as

$$g(\boldsymbol{\mu}_k^{\boldsymbol{\gamma}_k}) = \boldsymbol{\eta}_k^{\boldsymbol{\gamma}_k} = \mathbf{X}_k' \boldsymbol{\beta} + \sum_{j=1}^{c} \mathbf{Z}_{kj}' \boldsymbol{\gamma}_{kj} \tag{3.6}$$

The model described herein is a GLMM suited to hierarchical data[1]. A more general form of GLMMs uses general design matrices for both the fixed and random components (Zhao et al., 2006).

When there is a direct relationship between the natural parameter and the linear predictor, then $g(\cdot)$ is referred to as the *canonical* link function (Agresti, 2002). In the case of binary data, the canonical link function is given by

$$g(\mu_{ki}^{\boldsymbol{\gamma}_k}) = \text{logit}(\mu_{ki}^{\boldsymbol{\gamma}_k}) = \ln\left(\frac{\mu_{ki}^{\boldsymbol{\gamma}_k}}{1-\mu_{ki}^{\boldsymbol{\gamma}_k}}\right) = \boldsymbol{x}_{ki}' \boldsymbol{\beta} + \boldsymbol{z}_{ki}' \boldsymbol{\gamma}_k \tag{3.7}$$

which is referred to as the (conditional) logit link. This logit transformation ensures that $E(\, y_{ki}|\boldsymbol{\gamma}_k)$ is bounded between 0 and 1 (Rencher & Schaalje, 2008). From Equation 3.7, we obtain

$$\mu_{ki}^{\boldsymbol{\gamma}_k} = \frac{\exp(\boldsymbol{x}_{ki}' \boldsymbol{\beta} + \boldsymbol{z}_{ki}' \boldsymbol{\gamma}_k)}{1 + \exp(\boldsymbol{x}_{ki}' \boldsymbol{\beta} + \boldsymbol{z}_{ki}' \boldsymbol{\gamma}_k)} \tag{3.8}$$

In matrix notational form, we have

$$\text{logit}(\mu_{ki}^{\boldsymbol{\gamma}_k}) = \mathbf{X}_k' \boldsymbol{\beta} + \mathbf{Z}_k' \boldsymbol{\gamma}_k \tag{3.9}$$

This is commonly referred to as the random effects logistic regression model which is a class of the GLMM with a logit link (Kuss, 2002).

Other non-canonical link functions for binary data, which also bounds $E(\, y_{ki}|\boldsymbol{\gamma}_k)$ between 0 and 1, are the probit link and and the complementary log-log link. The GLMM with a probit link is given by

$$\text{probit}(\mu_{ki}^{\boldsymbol{\gamma}_k}) = \Phi^{-1}(\mu_{ki}^{\boldsymbol{\gamma}_k}) = \boldsymbol{x}_{ki}' \boldsymbol{\beta} + \boldsymbol{z}_{ki}' \boldsymbol{\gamma}_k \tag{3.10}$$

or equivalently

$$\mu_{ki}^{\boldsymbol{\gamma}_k} = \Phi(\boldsymbol{x}_{ki}' \boldsymbol{\beta} + \boldsymbol{z}_{ki}' \boldsymbol{\gamma}_k) \tag{3.11}$$

[1]The notation for higher-order GLMMs is straightforward.

where the probit link Φ is the standard Normal cumulative distribution function (Finney, 1971). The GLMM with a complementary log-log link has the form

$$\log[-\log(1-\mu_{ki}^{\gamma_k})] = \boldsymbol{x}'_{ki}\boldsymbol{\beta} + \boldsymbol{z}'_{ki}\boldsymbol{\gamma}_k \tag{3.12}$$

or equivalently

$$\mu_{ki}^{\gamma_k} = 1 - \exp[-\exp(\boldsymbol{x}'_{ki}\boldsymbol{\beta} + \boldsymbol{z}'_{ki}\boldsymbol{\gamma}_k)] \tag{3.13}$$

This model is based on the extreme value distribution, known as the Gumbel distribution, which is asymmetric (Agresti, 2002). In contrast, the logit and probit links both approach 0 and 1 symmetrically and asymptotically. Therefore, these two links often produce similar results (Cox & Snell, 1989; Finney, 1971).

3.3 Maximum likelihood estimation

The method of maximum likelihood is a standard method of estimation in parametric models where parameter estimates maximize the likelihood function of the observed data (Searle et al., 2006). In GLMMs, the marginal likelihood function, which is obtained by integrating over the distribution of the random effects, is maximized (Molenberghs & Verbeke, 2005). Let the contribution of the k^{th} cluster to the marginal likelihood be

$$f_k(y_{ki}\,|\boldsymbol{\beta}, \boldsymbol{G}, \phi) = \int_{\mathbb{R}^q} \prod_{i=1}^{n_k} f_{ki}(y_{ki}\,|\boldsymbol{\gamma}_k, \boldsymbol{\beta}, \phi) f(\boldsymbol{\gamma}_k \,|\, \boldsymbol{G})\, d\boldsymbol{\gamma}_k \tag{3.14}$$

where $f(\boldsymbol{\gamma}_k \,|\, \boldsymbol{G})$ is the distribution of the random effects, then the likelihood function of $\boldsymbol{\beta}, \boldsymbol{G}$ and ϕ is given jointly by

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{G}, \phi) &= \prod_{k=1}^{K} f_k(y_{ki}\,|\boldsymbol{\beta}, \boldsymbol{G}, \phi) \\ &= \prod_{k=1}^{K} \int_{\mathbb{R}^q} \prod_{i=1}^{n_k} f_{ki}(y_{ki}\,|\boldsymbol{\gamma}_k, \boldsymbol{\beta}, \phi) f(\boldsymbol{\gamma}_k \,|\, \boldsymbol{G})\, d\boldsymbol{\gamma}_k \end{aligned} \tag{3.15}$$

(Molenberghs & Verbeke, 2005). Maximization of Equation 3.15 requires the evaluation of q-dimensional integrals which, except in the case of normality assumptions,

are analytically intractable. Thus, closed form expressions for the likelihood function of a GLMM are typically unavailable (Jiang, 2007). Several numerical approximation methods have been proposed to circumvent the computational difficulties associated with the likelihood function in Equation 3.15. These methods involve either approximations of the integrand, approximations of the integral or approximations of the data (Molenberghs & Verbeke, 2005).

3.3.1 Laplace approximation

The Laplace approximation method is used to approximate integrals of the form

$$\int e^{Q(\boldsymbol{t})} dt \tag{3.16}$$

where $Q(\boldsymbol{t})$ is a known unimodal function and $\boldsymbol{t}$ is a q-dimensional vector of variables (Tuerlinckx et al., 2006). Let $\widehat{\boldsymbol{t}}$ be the value of $\boldsymbol{t}$ for which the function Q is maximized. Then, $Q(\boldsymbol{t})$ can be approximated by the second-order Taylor series expansion about $\widehat{\boldsymbol{t}}$; that is,

$$Q(\boldsymbol{t}) \approx Q(\widehat{\boldsymbol{t}}) + \frac{1}{2}(\boldsymbol{t} - \widehat{\boldsymbol{t}})' \ddot{Q}(\widehat{\boldsymbol{t}})(\boldsymbol{t} - \widehat{\boldsymbol{t}}) \tag{3.17}$$

where $\ddot{Q}(\widehat{\boldsymbol{t}})$ is the Hessian matrix of Q with entries $\dfrac{\partial^2 Q(\boldsymbol{t})}{\partial \boldsymbol{t}\, \partial \boldsymbol{t}'}$ evaluated at $\widehat{\boldsymbol{t}}$ (Molenberghs & Verbeke, 2005). When $Q(\boldsymbol{t})$ in Equation 3.16 is replaced by its approximation given in Equation 3.17, the resultant integrand resembles a multivariate Gaussian distribution with mean vector $\widehat{\boldsymbol{t}}$ and variance-covariance matrix $[-E(\ddot{Q}(\widehat{\boldsymbol{t}}))]^{-1}$. Therefore, the Laplace approximation to the integral in Equation 3.16 is given by

$$\int e^{Q(\boldsymbol{t})} dt \approx (2\pi)^{\frac{q}{2}} |-\ddot{Q}(\widehat{\boldsymbol{t}})|^{-\frac{1}{2}} e^{Q(\widehat{\boldsymbol{t}})} \tag{3.18}$$

Since $\boldsymbol{\gamma}_k \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{G})$, the integrals in the likelihood function in Equation 3.15 can be expressed in the form of the integral given in Equation 3.16, where the function Q

becomes

$$Q(\boldsymbol{\gamma}_k) = \phi^{-1}\,\tau_{ki} \sum_{i=1}^{n_k} \left[y_{ki}(\boldsymbol{x}'_{ki}\boldsymbol{\beta} + \boldsymbol{z}'_{ki}\boldsymbol{\gamma}_k) - b(\boldsymbol{x}'_{ki}\boldsymbol{\beta} + \boldsymbol{z}'_{ki}\boldsymbol{\gamma}_k) \right] - \frac{1}{2}\,\boldsymbol{\gamma}'_k \boldsymbol{G}\,\boldsymbol{\gamma}_k \qquad (3.19)$$

so that the Laplace approximation method can be applied. The accuracy of the Laplace approximation can be improved by including higher-order terms in the Taylor series expansion (Raudenbush et al., 2000). However, this method works well provided that the sample size of the clusters n_k are sufficiently large (Tuerlinckx et al., 2006).

3.3.2 Gaussian quadrature

The Gauss-Hermite quadrature and the adaptive Gauss-Hermite quadrature, because of their relation to Gaussian densities, are used to approximate integrals of the form

$$\int h(t)e^{-t^2}dt \qquad (3.20)$$

where h is a known and smooth function (Liu & Pierce, 1994). Suppose that a standardization of the random effects $\boldsymbol{\gamma}_k$ is given by

$$\delta_k = \boldsymbol{G}^{-\frac{1}{2}}\,\boldsymbol{\gamma}_k$$

such that δ_k follows a Gaussian distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{I}$, where $\boldsymbol{I}$ represents an identity matrix. The linear predictor then has the form $\theta_{ki} = \boldsymbol{x}'_{ki}\boldsymbol{\beta} + \boldsymbol{z}'_{ki}\,\boldsymbol{G}^{\frac{1}{2}}\,\boldsymbol{\delta}_k$, where the variance components in $\boldsymbol{G}$ are included. The likelihood contribution of the k^{th} cluster in Equation 3.14 is then

$$f_k(y_{ki}\,|\boldsymbol{\beta}, \boldsymbol{G}, \phi) = \int \prod_{i=1}^{n_k} f_{ki}(y_{ki}\,|\boldsymbol{\gamma}_k, \boldsymbol{\beta}, \phi) f(\boldsymbol{\gamma}_k\,|\,\boldsymbol{G})\,d\boldsymbol{\gamma}_k \qquad (3.21)$$

$$= \int \prod_{i=1}^{n_k} f_{ki}(y_{ki}\,|\boldsymbol{\delta}_k, \boldsymbol{\beta}, \boldsymbol{G}, \phi) f(\boldsymbol{\delta}_k) d\boldsymbol{\delta}_k, \qquad (3.22)$$

which is proportional to the form of the integral given in Equation 3.20. Therefore, approximations to this integral can be obtained using the Gauss-Hermite quadrature

or the adaptive Gauss-Hermite quadrature.

In the Gauss-Hermite quadrature, the integral in Equation 3.20 is approximated by

$$\int h(t)e^{-t^2}dt \approx \sum_{j=1}^{J} w_j\, h(t_j) \tag{3.23}$$

where the nodes or quadrature points t_j are solutions to the J^{th} order Hermite polynomial and w_j are the corresponding quadrature weights. The values of t_j and w_j, for $j = 1, \ldots, 20$, can be obtained from tables reported by Abramowitz & Stegun (1974). Alternatively, these values may be computed via an algorithm for any value of J (McCulloch et al., 2008). If $h(t)$ is a polynomial of degree $(2J - 1)$, then with J quadrature points, the Gauss-Hermite quadrature yields exact solutions. A major disadvantage with this method is that, due to the quadrature points t_j being selected independently of the function $h(t)$, t_j may not lie within the region of interest (Molenberghs & Verbeke, 2005). Furthermore, factors such as large sample sizes within clusters and large variances associated with random effects have a negative influence on the accuracy of the approximations. Increasing the number of quadrature points can improve the accuracy of the approximations. However, this also increases the computational complexity (Capanu et al., 2013).

An improved version of the Gauss-Hermite quadrature, known as the adaptive Gauss-Hermite quadrature, addresses the problems mentioned above by centering the quadrature points with respect to the mode of the integrand for each cluster and scaling them according to the estimated curvature at that mode (Tutz, 2012). As a result, more quadrature points lie within the region of interest. This approximation method uses a significantly lower number of quadrature points to achieve the same level of accuracy as the Gauss-Hermite quadrature. However, both these methods become computationally infeasible when the number of random effects is large. The adaptive Gauss-Hermite quadrature is also much more time consuming as it requires the mode and curvature for each cluster to be computed (Capanu et al., 2013; Tuerlinckx et al., 2006). When $J = 1$, the adaptive Gauss-Hermite quadrature is equivalent to approximating the integrand using the Laplace approximation method (Molenberghs & Verbeke, 2005).

3.3.3 Penalized quasi-likelihood

The concept of quasi-likelihood (QL) was introduced by Wedderburn (1974) for parameter estimation when the distributional form of the observations is not known. The definition of a QL depends only on the specification of a mean-variance relationship for the observations. This concept is exploited in the penalized quasi-likelihood (PQL) approach. Consider a decomposition of the data into the mean, which is a non-linear function of the linear predictor conditional on the random effects, and an error term, such that

$$\begin{aligned} Y_{ki} &= \mu_{ki}^{\boldsymbol{\gamma}_k} + \epsilon_{ki} \\ &= h(\boldsymbol{x}'_{ki}\boldsymbol{\beta} + \boldsymbol{z}'_{ki}\,\boldsymbol{\gamma}_k) + \epsilon_{ki} \end{aligned} \tag{3.24}$$

where $h(\cdot) = g^{-1}(\cdot)$ is the inverse of the link function and ϵ_{ki} are error terms assumed to follow a distribution with mean zero and variance $Var(Y_{ki}|\boldsymbol{\gamma}_k) = \phi\,\tau_{ki}^{-1}\,V(\mu_{ki}^{\boldsymbol{\gamma}_k})$. For the canonical link function, the variance function has the form

$$V(\mu_{ki}^{\boldsymbol{\gamma}_k}) = \dot{h}(\boldsymbol{x}'_{ki}\boldsymbol{\beta} + \boldsymbol{z}'_{ki}\,\boldsymbol{\gamma}_k)$$

where $\dot{h}$ denotes the derivative of h with respect to $\mu_{ki}^{\boldsymbol{\gamma}_k}$ (Molenberghs & Verbeke, 2005). Let $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}_k$ denote the current estimates of the fixed and random effects, respectively. Then, the PQL method approximates the mean in Equation 3.24, and hence the parameters, by a linear Taylor series expansion about $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}_k$. This yields

$$\begin{aligned} Y_{ki} &\approx h(\boldsymbol{x}'_{ki}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}'_{ki}\,\widehat{\boldsymbol{\gamma}}_k) \\ &\quad + \dot{h}(\boldsymbol{x}'_{ki}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}'_{ki}\,\widehat{\boldsymbol{\gamma}}_k)\,\boldsymbol{x}'_{ki}\,(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \\ &\quad + \dot{h}(\boldsymbol{x}'_{ki}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}'_{ki}\,\widehat{\boldsymbol{\gamma}}_k)\,\boldsymbol{z}'_{ki}\,(\boldsymbol{\gamma}_k - \widehat{\boldsymbol{\gamma}}_k) + \epsilon_{ki} \\ &= \widehat{\mu}_{ki}^{\boldsymbol{\gamma}_k} + V(\widehat{\mu}_{ki}^{\boldsymbol{\gamma}_k})\,\boldsymbol{x}'_{ki}\,(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + V(\widehat{\mu}_{ki}^{\boldsymbol{\gamma}_k})\,\boldsymbol{z}'_{ki}\,(\boldsymbol{\gamma}_k - \widehat{\boldsymbol{\gamma}}_k) + \epsilon_{ki} \end{aligned} \tag{3.25}$$

where $\widehat{\mu}_{ki}^{\boldsymbol{\gamma}_k} = h(\boldsymbol{x}'_{ki}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}'_{ki}\,\widehat{\boldsymbol{\gamma}}_k)$ is the current predictor for the conditional mean $E(Y_{ki}|\boldsymbol{\gamma}_k)$ and conditional variance $V(\widehat{\mu}_{ki}^{\boldsymbol{\gamma}_k}) = \dot{h}(\boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}'_{ki}\,\widehat{\boldsymbol{\gamma}}_k)$. Equation 3.25 can be rewritten more compactly as

$$\boldsymbol{y}_k \approx \widehat{\boldsymbol{\mu}}_k^{\boldsymbol{\gamma}_k} + \widehat{\boldsymbol{V}}_k\,\boldsymbol{X}_k\,(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + \widehat{\boldsymbol{V}}_k\,\boldsymbol{Z}_k\,(\boldsymbol{\gamma}_k - \widehat{\boldsymbol{\gamma}}_k) + \boldsymbol{\epsilon}_k \tag{3.26}$$

where $\boldsymbol{X}_k$ and $\boldsymbol{Z}_k$ are appropriate design matrices and $\widehat{\boldsymbol{V}}_k$ is a diagonal matrix with elements $V(\widehat{\mu}_{ki}^{\boldsymbol{\gamma}_k})$ on the main diagonal. Multiplying Equation 3.26 by $\widehat{\boldsymbol{V}}_k^{-1}$ and rearranging the terms yields

$$\boldsymbol{y}_k^* \equiv \widehat{\boldsymbol{V}}_k^{-1}(\boldsymbol{y}_k - \widehat{\boldsymbol{\mu}}_k^{\boldsymbol{\gamma}_k}) + \boldsymbol{X}_k\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_k\widehat{\boldsymbol{\gamma}}_k \approx \boldsymbol{X}_k\boldsymbol{\beta} + \boldsymbol{Z}_k\boldsymbol{\gamma}_k + \boldsymbol{\epsilon}_k^* \tag{3.27}$$

where $\boldsymbol{\epsilon}_k^* = \widehat{\boldsymbol{V}}_k^{-1}\boldsymbol{\epsilon}_k$ has mean zero and variance $\boldsymbol{\Gamma}_k \equiv Var(\boldsymbol{\epsilon}_k^*) \approx \phi\,\tau_k^{-1}\,V(\mu_k^{\boldsymbol{\gamma}_k})$ (Tuerlinckx et al., 2006). Thus, the mean and variance of $\boldsymbol{y}_k^*$ are given by

$$\begin{aligned} E(\boldsymbol{y}_k^*) &\approx E(\boldsymbol{X}_k\boldsymbol{\beta} + \boldsymbol{Z}_k\boldsymbol{\gamma}_k + \boldsymbol{\epsilon}_k^*) \\ &= \boldsymbol{X}_k\boldsymbol{\beta} \end{aligned} \tag{3.28}$$

and

$$\begin{aligned} Var(\boldsymbol{y}_k^*) &\approx Var(\boldsymbol{X}_k\boldsymbol{\beta} + \boldsymbol{Z}_k\boldsymbol{\gamma}_k + \boldsymbol{\epsilon}_k^*) \\ &\approx \boldsymbol{Z}_k\,Var(\boldsymbol{\gamma}_k)\,\boldsymbol{Z}_k' + \boldsymbol{\Gamma}_k \\ &= \boldsymbol{Z}_k\boldsymbol{G}\boldsymbol{Z}_k' + \boldsymbol{\Gamma}_k \\ &= \boldsymbol{\Psi} \end{aligned} \tag{3.29}$$

respectively. In the context GLMs, the vector $\boldsymbol{y}_k^*$ is known as the *adjusted* or *working* dependent variable. Breslow & Clayton (1993) have shown that the same results are obtained when the working dependent variable is defined directly as a linearized form of the link function $g(\cdot)$ applied to the data; that is,

$$\boldsymbol{y}_k^* \equiv g(\widehat{\boldsymbol{\mu}}_k^{\boldsymbol{\gamma}_k}) + \dot{g}(\widehat{\boldsymbol{\mu}}_k^{\boldsymbol{\gamma}_k})(\boldsymbol{y}_k^* - \widehat{\boldsymbol{\mu}}_k^{\boldsymbol{\gamma}_k})$$

Equation 3.27 can be viewed as a linear mixed model (LMM) with $\boldsymbol{y}_k^*$ as the response vector. Therefore, estimation methods developed for LMMs can be employed to obtain updated estimates for the fixed and random effects (Molenberghs & Verbeke, 2005). According to Harville (1977), the estimates of the fixed effects parameter $\boldsymbol{\beta}$ and the random effects $\boldsymbol{\gamma}_k$ are given by

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}_k'\widehat{\boldsymbol{\Psi}}^{-1}\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k'\widehat{\boldsymbol{\Psi}}^{-1}\boldsymbol{y}_k^* \tag{3.30}$$

and

$$\begin{aligned}\widehat{\gamma}_k &= \boldsymbol{G}\boldsymbol{Z}_k'\widehat{\boldsymbol{\Psi}}^{-1}(\boldsymbol{y}_k^* - \boldsymbol{X}_k\widehat{\boldsymbol{\beta}}) \\ &= \boldsymbol{G}\boldsymbol{Z}_k'\widehat{\boldsymbol{P}}\boldsymbol{y}_k^* \end{aligned} \tag{3.31}$$

where $\widehat{\boldsymbol{\Psi}}$ is the estimated variance of $\boldsymbol{y}_k^*$ and $\widehat{\boldsymbol{P}} = \widehat{\boldsymbol{\Psi}}^{-1} - \widehat{\boldsymbol{\Psi}}^{-1}\boldsymbol{X}_k(\boldsymbol{X}_k'\widehat{\boldsymbol{\Psi}}^{-1}\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k'\widehat{\boldsymbol{\Psi}.}^{-1}$ These parameter estimates are then used to update $\boldsymbol{y}_k^*$ which in turn is used to update the parameter estimates. This iteration process continues until convergence is achieved (Tuerlinckx et al., 2006). The resulting estimates are called penalized quasi-likelihood (PQL) estimates. The standard implementation of PQL holds ϕ fixed at unity but Wolfinger & O'Connell (1993) showed that, if desired, ϕ can be estimated from the data.

3.3.4 Marginal quasi-likelihood

Similar to the PQL approach, the marginal quasi-likelihood (MQL) method approximates the mean in Equation 3.24 by a linear Taylor series expansion. However, this is done about the current estimates $\widehat{\boldsymbol{\beta}}$ for the fixed effects and about $\widehat{\boldsymbol{\gamma}}_k = \boldsymbol{0}$ for the random effects (Molenberghs & Verbeke, 2005). Thus, the current predictor for the conditional mean has the form $\widehat{\mu}_{ki}^{\boldsymbol{\gamma}_k} = h(\boldsymbol{x}_{ki}'\widehat{\boldsymbol{\beta}})$. The working dependent variable $\boldsymbol{y}_k^*$ is then given by

$$\boldsymbol{y}_k^* \equiv \widehat{\boldsymbol{V}}_k^{-1}(\boldsymbol{y}_k - \widehat{\boldsymbol{\mu}}_k^{\boldsymbol{\gamma}_k}) + \boldsymbol{X}_k\widehat{\boldsymbol{\beta}} \approx \boldsymbol{X}_k\boldsymbol{\beta} + \boldsymbol{Z}_k\boldsymbol{\gamma}_k + \boldsymbol{\epsilon}_k^* \tag{3.32}$$

which again approximates an LMM. Therefore, the parameter estimates can be obtained using the same procedure as for the PQL approach. However, the resulting estimates are referred to as marginal quasi-likelihood estimates (Breslow & Clayton, 1993). The PQL and MQL methods are more flexible and computationally faster than the Laplace approximation and the Gaussian quadratures. However, both methods tend to produce estimates that are biased towards zero, especially in the case of binary data when the sample sizes within clusters are relatively small and when the variance components are relatively large (Tuerlinckx et al., 2006).

3.4 Model assessment

Once the parameter estimates are obtained, inferences about model parameters can be made. When testing the significance of the fixed effects parameters in a GLMM, the null hypothesis to be tested is of the form $H_0 : \boldsymbol{C\beta} = \mathbf{0}$, where $\boldsymbol{C}$ is a matrix of constants of full row rank d (Tuerlinckx et al., 2006). Such a hypothesis can be tested using the likelihood ratio test, the Wald test or the score test. These tests are asymptotically equivalent under the null hypothesis and follow a χ^2-distribution with d degrees of freedom. These statistics can also be used to test the significance of the random effects in a GLMM. Testing whether certain random effects should be included in the model is equivalent to testing whether the corresponding variance components in $\boldsymbol{G}$ are statistically zero. Thus, the null hypothesis involves testing whether the variance parameters lie on the boundary of the model parameter space. In such case, the test statistics do not have the traditional χ^2-distribution, but rather follow a mixture of χ^2-distributions (Self & Liang, 1987; Stram & Lee, 1994; Zhang & Lin, 2008). The likelihood ratio and score tests are based on likelihood theory. Both the PQL and MQL methods are not likelihood-based (Hedeker, 2005). Therefore, these tests cannot be used for model selection when the estimates are obtained using these methods. The Wald test, however, can still be used to test the significance of the model parameters (Bolker et al., 2009).

Another commonly used model selection tool is the Akaike Information Criterion (AIC). AIC is defined as

$$AIC = -2\ln(L) + 2p \tag{3.33}$$

where L is the likelihood of the fitted model maximized over p parameters (Akaike, 1987). AIC is evaluated for each model in a competing set, and the model with the smallest AIC value is selected as the best-fitting model.

3.5 Application

The analysis in this section was performed using SAS (SAS Intitute Inc., 2013), version 9.4, where the procedure PROC GLIMMIX was used to fit a GLMM to the NIDS data. All demographic, lifestyle, socio-economic and environmental variables, discussed in Chapter 2, were selected for inclusion in the model. An intercept, varying across clusters, was also included. This random intercept was used to account for heterogeneity among clusters, and thereby, account for possible correlation among observations within the same cluster. The model was first fitted using the Laplace approximation method. All three link functions for binary data were fitted. However, the logit link produced the lowest AIC value and was therefore selected. Model diagnostics are provided in Appendix C. To confirm the need for the random intercept in the model, the COVTEST statement in SAS, which produces likelihood ratio tests for covariance parameters, was used. The result of this test, given in Table 3.1, indicated that the covariance parameter was highly significant (p-value < 0.0001), and thereby confirmed the necessity of including the random cluster effect in the model.

Table 3.1: Test of covariance parameters based on the likelihood

Label	DF	-2Log Likelihood	χ^2	P-value
No G - side effects	1	11912	31.91	<0.0001

For valid inference, an appropriate covariance structure for the data needs to be selected, thus the model was fitted using various covariance structures. Table 3.2 gives the different covariance structures fitted and their corresponding AIC values.

Table 3.2: AIC Goodness-of-Fit Statistic for GLMM

Covariance Structure	AIC
Variance components (VC)	11 946.57
Autoregressive (AR(1))	11 948.57
Compound symmetry (CS)	11 948.57
Heterogeneous compound symmetry (CSH)	11 948.57
Unstructured (UN)	11 947.57

The best suited covariance structure was VC, as this produced the lowest AIC value. Model selection of the fixed effects was achieved using a backward selection procedure. Based on the p-values obtained in the type III analysis of fixed effects, insignificant fixed effects were removed from the model one at a time until only significant fixed effects were left. Furthermore, all two-way and higher-order interaction terms were explored, however, none were found to be significant. The resulting model with one variance component, denoted by GLMM(1), is presented in Table 3.3.

Table 3.3: Type III analysis of fixed effects for GLMM(1)

Effect	**Numerator DF**	**F-Value**	**P-Value**
Age	5	112.26	<0.0001
Population group	3	12.60	<0.0001
Marital status	4	23.03	<0.0001
Current smoker	1	32.65	<0.0001
Alcohol consumption	1	9.77	0.0018
Household expenditure on food	3	3.26	0.0206
Education	3	10.90	<0.0001
Employment status	1	3.73	0.0494
Household income quintile	4	5.29	0.0003
Geographical type	2	18.11	<0.0001

Age, ethnicity, marital status, smoking status, alcohol consumption, household expenditure on food, education level, employment status, household income and geographical type of residence were all found to be significantly associated with obesity among females. This was consistent with the results obtained with the chi-square test of association in Section 2.4.1. The Pearson chi-square statistic over its degrees of freedom, which is a measure of variability in the marginal distribution of the data, was 1. This indicates that the variability in the data was properly modelled and, hence there was no residual overdispersion. The estimated variance component for cluster effect was 0.0367 with a standard error of 0.0125. The parameter estimates, odds ratios (OR) with their 95% confidence intervals and p-values for GLMM(1) are given in Table 3.4. The results were obtained after approximately 25 seconds.

Table 3.4: Estimates and OR with 95% confidence intervals for GLMM(1)

Parameter	Estimate	Std. Error	OR (95% C.I.)	p-value
Intercept	-2.1868	0.1314		<0.001
Age				
15-24	ref	...	1	...
25-34	1.167	0.076	3.213 (2.766, 3.732)	<0.001
35-44	1.709	0.083	5.521 (4.689, 6.500)	<0.001
45-54	1.881	0.091	6.561 (5.494, 7.836)	<0.001
55-64	1.981	0.101	7.246 (5.947, 8.830)	<0.001
65+	1.600	0.112	4.954 (3.977, 6.170)	<0.001
Population group				
African	ref	...	1	...
Coloured	-0.205	0.091	0.814 (0.681, 0.974)	0.024
Asian/Indian	-1.201	0.248	0.301 (0.185, 0.589)	<0.001
White	-0.619	0.152	0.538 (0.400, 0.724)	<0.001
Marital status				
Married	ref	...	1	...
Living with partner	-0.599	0.103	0.549 (0.449, 0.672)	<0.001
Widow	-0.397	0.081	0.672 (0.574, 0.787)	<0.001
Divorced or separated	-0.364	0.137	0.695 (0.531, 0.910)	0.008
Never married	-0.548	0.061	0.578 (0.513, 0.651)	<0.001
Current smoker				
No	ref	...	1	...
Yes	-0.579	0.101	0.561 (0.460, 0.684)	<0.001
Alcohol consumption				
No	ref	...	1	...
Yes	-0.228	0.073	0.796 (0.690, 0.918)	0.002
Household expenditure on food				
Quartile I	ref	...	1	...
Quartile II	0.163	0.064	1.177 (1.037, 1.335)	0.011
Quartile III	0.151	0.069	1.163 (1.016, 1.331)	0.029
Quartile IV	0.219	0.077	1.245 (1.071, 1.447)	0.004
Education				
No schooling	ref	...	1	...
Primary	0.273	0.080	1.313 (1.123, 1.536)	0.001
Secondary	0.460	0.084	1.584 (1.344, 1.868)	<0.001
Tertiary	0.176	0.184	1.193 (0.832, 1.711)	0.337

Continued on next page

Table 3.4 – *Continued from previous page*

Parameter	Estimate	Std. Error	OR (95% C.I.)	p-value
Employment status				
Unemployed	ref	...	1	...
Employed	0.102	0.053	1.108 (1.001, 1.229)	0.048
Household income quintile				
I	ref	...	1	...
II	-0.061	0.072	0.941 (0.818, 1.083)	0.396
III	0.082	0.074	1.086 (0.939, 1.255)	0.267
IV	0.073	0.078	1.075 (0.924, 1.252)	0.349
V	0.313	0.088	1.367 (1.151, 1.623)	<0.001
Geographical type				
Urban	ref	...	1	...
Traditional	0.352	0.063	1.422 (1.257, 1.607)	<0.001
Farms	0.029	0.097	1.030 (0.851, 1.245)	0.763

The results revealed an increase in odds as age increased followed by a decline after the age of 65 years. Compared to females aged 15-24 years, females aged 55-64 years had the highest odds (OR = 7.246; 95% CI: 5.947-8.830), followed by those aged 45-54 years (OR = 6.561; 95% CI: 5.494-7.836). Females aged 65+ years were approximately five times more likely to be obese compared to those aged 15-24 years (OR = 4.954; 95% CI: 3.977-6.170). White (OR = 0.538; 95% CI: 0.400-0.724), Coloured (OR = 0.814; 95% CI: 0.681-0.974) and Asian/Indian (OR = 0.301; 95% CI: 0.185-0.589) females were less likely to be obese compared to African females. Females living with their partners, widowed, divorced or separated, or never married, were associated with a lower risk of obesity compared to those who are married, with the odds ranging from 0.549 to 0.695.

Female smokers were associated with a lower risk of obesity compared to non-smokers (OR = 0.561; 95% CI: 0.460-0.684). Similarly, females who consumed alcohol were significantly less likely to be obese (OR = 0.796; 95% CI: 0.690-0.918). Although the odds of obesity for females with a total household expenditure on food within the highest quartile was only 1.245 (95% CI: 1.071-1.447), they were most at risk for obesity compared to those with a total household expenditure on food within the first quartile. Those with a total household expenditure on food within the second and third quartiles were also more likely to be obese compared to those with a total household expenditure on food within the first quartile (OR = 1.177; 95% CI: 1.037-1.335 and OR = 1.163; 95% CI: 1.016-1.331).

Females with a primary or secondary education were associated with a higher risk of obesity compared to those with no schooling (OR = 1.313; 95% CI: 1.123-1.536 and OR = 1.584; 95% CI: 1.344-1.868, respectively). Although the odds of obesity for those with a tertiary education was higher, they were not significantly different compared to those with no schooling (p-value = 0.337). Females who were employed were at a higher risk of obesity compared to their unemployed counterparts (OR = 1.108; 95% CI: 1.001-1.229). Those belonging to the highest household income quintile were more likely to be obese compared to those belonging to the lowest household income quintile (OR = 1.367; 95% CI: 1.151-1.623). All other household income quintiles were not significantly different to the lowest household income quintile. Females in traditional areas were more likely to be obese compared to those living in urban areas (OR = 1.422; 95% CI: 1.257-1.607). There was no significant difference between those living in farms and those living in urban areas (p-value = 0.763).

Individuals living in the same household often display similar lifestyle and socio-economic patterns, and thus, may be more homogeneous than individuals from different households, even those within the same cluster. In the NIDS data set, there are up to 14 females living in the same household, with a mean of 1.61 females per household. Therefore, in order to account for possible correlation within households, households nested within clusters were included as an additional random effect in the model. The test of covariance parameters for this model, given in Table 3.5, produced a significant result with a p-value < 0.0001. Thus suggesting that both the heterogeneity among clusters and the heterogeneity among households nested within clusters have a significant effect on obesity among females. Furthermore, the model with both variance components, denoted by GLMM(2), produced a lower AIC value than that of GLMM(1).

Table 3.5: Test of covariance parameters based on the likelihood

Label	**DF**	**-2Log Likelihood**	χ^2	**P-value**
No G - side effects	2	11916	85.39	<0.0001

Model selection of the fixed effects was carried out for this model using the same procedure mentioned above. The resulting model, given in Table 3.6, was the same as GLMM(1).

Table 3.6: Type III analysis of fixed effects for GLMM(2)

Effect	Numerator DF	F-Value	P-Value
Age	5	102.98	<0.0001
Population group	3	12.09	<0.0001
Marital status	4	22.57	<0.0001
Current smoker	1	31.81	<0.0001
Alcohol consumption	1	9.55	0.0020
Household expenditure on food	3	3.40	0.0171
Education	3	10.35	<0.0001
Employment status	1	4.26	0.0390
Household income quintile	4	4.94	0.0006
Geographical type	2	18.19	<0.0001

The Pearson chi-square statistic over its degrees of freedom was 0.79, which once again indicated that there was no residual overdispersion. The estimated variance components for cluster effect and household by cluster effect are given in Table 3.7.

Table 3.7: Covariance parameter estimates for GLMM(2)

Covariance Parameter	Subject	Estimate	Standard Error
Intercept	Cluster	0.0375	0.0140
Intercept	Household(Cluster)	0.5164	0.0892

Table 3.8 presents the results of the fixed effects for GLMM(2). These results were obtained within 10 minutes. Compared to GLMM(1), GLMM(2) produced slightly higher standard errors, and hence, wider confidence intervals. This was expected due to the additional source of variation in the model. However, both models produced similar parameter estimates and therefore, similar conclusions can be drawn.

Table 3.8: Estimates and OR with 95% confidence intervals for GLMM(2)

Parameter	Estimate	Std. Error	OR (95% C.I.)	p-value
Intercept	-2.392	0.149		<0.001
Age				
15-24	ref	...	1	...
25-34	1.258	0.084	3.518 (2.985, 4.145)	<0.001
35-44	1.870	0.094	6.489 (5.396, 7.804)	<0.001
45-54	2.065	0.103	7.886 (6.441, 9.656)	<0.001
55-64	2.170	0.115	8.762 (6.993, 10.978)	<0.001
65+	1.748	0.125	5.740 (4.489, 7.341)	<0.001
Population group				
African	ref	...	1	...
Coloured	-0.224	0.101	0.800 (0.656, 0.975)	0.027
Asian/Indian	-1.328	0.282	0.265 (0.153, 0.461)	<0.001
White	-0.695	0.170	0.499 (0.357, 0.697)	<0.001
Marital status				
Married	ref	...	1	...
Living with partner	-0.679	0.114	0.507 (0.405, 0.634)	<0.001
Widow	-0.454	0.090	0.635 (0.532, 0.758)	<0.001
Divorced or separated	-0.397	0.153	0.673 (0.498, 0.908)	0.010
Never married	-0.610	0.069	0.543 (0.475, 0.622)	<0.001
Current smoker				
No	ref	...	1	...
Yes	-0.631	0.112	0.532 (0.427, 0.663)	<0.001
Alcohol consumption				
No	ref	...	1	...
Yes	-0.250	0.081	0.779 (0.665, 0.913)	0.002
Household expenditure on food				
Quartile I	ref	...	1	...
Quartile II	0.190	0.073	1.209 (1.048, 1.396)	0.010
Quartile III	0.169	0.079	1.184 (1.015, 1.382)	0.032
Quartile IV	0.256	0.088	1.292 (1.087, 1.535)	0.004
Education				
No schooling	ref	...	1	...
Primary	0.299	0.089	1.349 (1.134, 1.605)	0.001
Secondary	0.497	0.094	1.643 (1.368, 1.974)	<0.001
Tertiary	0.157	0.205	1.170 (0.783, 1.747)	0.444

Continued on next page

Table 3.8 – *Continued from previous page*

Parameter	Estimate	Std. Error	OR (95% C.I.)	p-value
Employment status				
Unemployed	ref	...	1	...
Employed	0.121	0.059	1.129 (1.006, 1.267)	0.039
Household income quintile				
I	ref	...	1	...
II	-0.070	0.081	0.932 (0.795, 1.092)	0.385
III	0.088	0.084	1.092 (0.926, 1.288)	0.296
IV	0.081	0.089	1.085 (0.912, 1.291)	0.359
V	0.347	0.101	1.415 (1.162, 1.723)	0.001
Geographical type				
Urban	ref	...	1	...
Traditional	0.398	0.071	1.488 (1.296, 1.709)	<0.001
Farms	0.033	0.110	1.034 (0.833, 1.283)	0.762

The analysis was also carried out using the adaptive Gauss-Hermite quadrature and PQL estimation methods. The procedure `PROC GLIMMIX` was used once again. To explore the impact of different numbers of quadrature points on parameter estimation using the adaptive Gauss-Hermite quadrature, different numbers of quadrature points (J = 3, 5, 10, 20) were used. However, this led to negligible differences in parameter estimation with no differences between parameter estimates for quadrature points 10 and 20. Both the adaptive Gauss-Hermite quadrature and PQL methods produced results very similar to that obtained using the Laplace approximation method. However, these methods differed in terms of computational speed, with the adaptive Gauss-Hermite quadrature with 20 quadrature points taking up to 25 minutes to run GLMM(1) and up to 60 minutes to run GLMM(2).

Chapter 4

Bayesian Analysis of the NIDS Data Set

The Bayesian approach is more appealing than the classical approach for inference in GLMMs as it takes into account the ease with which uncertainty in parameters is estimated (Zhao et al., 2006). A Bayesian analysis is often performed using Markov chain Monte Carlo (MCMC) methods. However, the integrated nested Laplace approximation (INLA) is becoming a computationally convenient alternative to MCMC (Fong et al., 2010). INLA may be regarded as a novel numerical inferential procedure which renders MCMC sampling redundant as it approximates posterior distributions in a fully automated way (Roos & Held, 2011). This chapter outlines the Bayesian approach to GLMMs and presents results obtained using both the MCMC and INLA methods in the analysis of the NIDS data set.

4.1 Bayesian inference

Bayesian inference is an approach to statistical inference that exploits Bayes theorem where all unknown parameters are treated as random variables and all forms of uncertainty are expressed in terms of probability statements (Gelman et al., 2014). Let $\boldsymbol{\vartheta}$ be a vector of unknown parameters. The joint probability density of the observed data $\boldsymbol{y}$ and the unknown parameters has the form

$$f(\boldsymbol{\vartheta}, \boldsymbol{y}) = f(\boldsymbol{\vartheta})\, f(\boldsymbol{y}|\boldsymbol{\vartheta}) \tag{4.1}$$

where $f(\boldsymbol{\vartheta})$ is the *prior* distribution of $\boldsymbol{\vartheta}$ and $f(\boldsymbol{y}|\boldsymbol{\vartheta})$ is the *sampling* distribution of

$\boldsymbol{y}|\boldsymbol{\vartheta}$. The prior distribution represents an assumption about the nature of the parameters before observing the data while the sampling distribution reflects information about the parameters contained in the data (Wade, 2000). The parameters of the prior distribution are usually referred to as *hyperparameters* and can be chosen based on previous studies with similar data and/or expert opinion. They can also be non-informative when no prior knowledge about the unknown parameters exists (Glickman & van Dyk, 2007).

In the Bayesian paradigm, all inference is based on the *posterior* distribution $f(\boldsymbol{\vartheta}|\boldsymbol{y})$. According to Bayes theorem, the posterior distribution is

$$f(\boldsymbol{\vartheta}|\boldsymbol{y}) = \frac{f(\boldsymbol{\vartheta}, \boldsymbol{y})}{f(\boldsymbol{y})} = \frac{f(\boldsymbol{\vartheta})\, f(\boldsymbol{y}|\boldsymbol{\vartheta})}{f(\boldsymbol{y})} \tag{4.2}$$

where $f(\boldsymbol{y})$ is the marginal probability density of $\boldsymbol{y}$ such that $f(\boldsymbol{y}) = \int_{\boldsymbol{y}} f(\boldsymbol{\vartheta})\, f(\boldsymbol{y}|\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}$. Moreover, $f(\boldsymbol{y})$ does not depend on $\boldsymbol{\vartheta}$ and, is thus, considered to be a normalizing constant; that is, a constant which ensures that the posterior distribution integrates to one. Therefore, an equivalent form of Bayes theorem is given by

$$f(\boldsymbol{\vartheta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\vartheta})\, f(\boldsymbol{\vartheta}) \tag{4.3}$$

When the sampling distribution is regarded as a function of $\boldsymbol{\vartheta}$, for given $\boldsymbol{y}$, it is called the likelihood function and is denoted by $l(\boldsymbol{\vartheta}|\boldsymbol{y})$ (Box & Tiao, 1992). Thus, Equation 4.3 becomes

$$f(\boldsymbol{\vartheta}|\boldsymbol{y}) \propto l(\boldsymbol{\vartheta}|\boldsymbol{y})\, f(\boldsymbol{\vartheta}) \tag{4.4}$$

Thus, the posterior distribution is proportional to the product of the likelihood function and the prior distribution of the parameters.

All statistical inference can be deduced from appropriate summaries of the posterior distribution. These summaries are typically expressed in terms of posterior expectations of functions of $\boldsymbol{\vartheta}$ and are of the form

$$I = \int g(\boldsymbol{\vartheta})\, f(\boldsymbol{\vartheta}|\boldsymbol{y}) d\boldsymbol{\vartheta} \tag{4.5}$$

where $g(\boldsymbol{\vartheta})$ is some function of $\boldsymbol{\vartheta}$ (Gilks et al., 1996). I in Equation 4.5 is typically high-dimensional and is analytically intractable. Evaluation of such integrals requires computational methods such as the Markov chain Monte Carlo methods.

4.2 Markov chain Monte Carlo methods

The Markov chain Monte Carlo (MCMC) method is a general computational method based on sampling directly from the posterior distribution and then using those samples to estimate the quantities of interest (Brooks, 1998). The samples are drawn sequentially from a target distribution with each sample depending only on the previous sample drawn. Hence, the samples form a Markov chain. A Markov chain is a sequence of random variables $\boldsymbol{\vartheta}^{(0)}$, $\boldsymbol{\vartheta}^{(1)}$, ..., where the random variable $\boldsymbol{\vartheta}^{(t)}$ depends only on the previous state of the chain $\boldsymbol{\vartheta}^{(t-1)}$ (Gilks et al., 1996). Monte Carlo integration then uses the Markov chain samples to approximate the posterior expectation in Equation 4.5. This gives

$$\widehat{I} = \frac{1}{T-B} \sum_{t=B+1}^{T} g(\boldsymbol{\vartheta}^{(t)}) \tag{4.6}$$

where $\widehat{I}$ is called an ergodic average, T is the sample size generated from the target distribution and B indicates the amount of burn-in; that is, the number of initial samples that are discarded in order to minimize the effect of the initial values on the posterior inference (Craiu & Rosenthal, 2014). If the Markov chain has the target distribution as a stationary (or invariant) distribution, then under certain conditions, $\widehat{I}$ will converge to the target distribution (Craiu & Rosenthal, 2014; Roberts, 1996).

Several MCMC methods have been proposed in the literature. However, the most commonly used are the Metropolis-Hastings algorithm and the Gibbs sampler. These two methods are outlined in the following sections.

4.2.1 The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm, proposed by Hastings (1970), is a generalization of the Metropolis algorithm. Suppose $\boldsymbol{\vartheta}^{(t)}$ is the t^{th} sample, $t = 1, \ldots, T$, from the target distribution $f(\boldsymbol{\vartheta}|\boldsymbol{y})$. With the Metropolis-Hastings algorithm, $\boldsymbol{\vartheta}^{(t)}$ is chosen by first sampling a candidate value $\boldsymbol{\vartheta}^{*}$ from a proposal distribution $q(\boldsymbol{\vartheta}^{*}|\boldsymbol{\vartheta}^{(t-1)})$.

The candidate value $\boldsymbol{\vartheta}^*$ is then accepted with probability α where

$$\alpha = \min\left(1\,,\ \frac{f(\boldsymbol{\vartheta}^*|\boldsymbol{y})\, q(\boldsymbol{\vartheta}^{(t-1)}|\boldsymbol{\vartheta}^*)}{f(\boldsymbol{\vartheta}^{(t-1)}|\boldsymbol{y})\, q(\boldsymbol{\vartheta}^*|\boldsymbol{\vartheta}^{(t-1)})}\right) \tag{4.7}$$

If $\boldsymbol{\vartheta}^*$ is accepted, the algorithm is repeated with $\boldsymbol{\vartheta}^*$ as the starting value. If $\boldsymbol{\vartheta}^*$ is rejected, the starting value remains unchanged. The Metropolis-Hastings algorithm can be summarized as follows:

Step 1: Choose an initial value $\boldsymbol{\vartheta}^{(0)}$.

Step 2: Set $t = 1$.

Step 3: Generate $\boldsymbol{\vartheta}^*$ from $q(\boldsymbol{\vartheta}^*|\boldsymbol{\vartheta}^{(t-1)})$.

Step 4: Compute α.

Step 5: Set

$$\boldsymbol{\vartheta}^{(t)} = \begin{cases} \boldsymbol{\vartheta}^* & \text{with probability } \alpha \\ \boldsymbol{\vartheta}^{(t-1)} & \text{otherwise.} \end{cases}$$

Step 6: If $t < T$, set $t = t + 1$ and return to step 3.
Else, end iteration.

The Metropolis-Hastings algorithm will eventually converge to the target distribution. However, the rate of convergence depends on the form of the proposal distribution (Roberts, 1996; Tierney, 1996).

4.2.2 The Gibbs sampler

A special case of the Metropolis-Hastings algorithm is the Gibbs sampler which uses the full conditional posterior distribution $f(\vartheta_j|\boldsymbol{\vartheta}_{-j}, \boldsymbol{y})$ as the proposal distribution where $\boldsymbol{\vartheta}_{-j} = (\vartheta_1, \ldots, \vartheta_{j-1}, \vartheta_{j+1}, \ldots, \vartheta_k)'$. This distribution leads to an acceptance probability $\alpha = 1$. Thus, the proposed value is accepted at every iteration (Gelman et al., 2014; Gilks et al., 1996). The Gibbs sampler can be summarized as follows:

Step 1: Choose an initial value $\vartheta^{(0)}$.

Step 2: Set $t = 1$.

Step 3: Generate each component of $\vartheta^{(t)}$ as follows:

- $\vartheta_1^{(t)}$ from $f(\vartheta_1|\vartheta_2^{(t-1)}, \vartheta_3^{(t-1)}, \ldots, \vartheta_k^{(t-1)}, \boldsymbol{y})$
- $\vartheta_2^{(t)}$ from $f(\vartheta_2|\vartheta_1^{(t)}, \vartheta_3^{(t-1)}, \ldots, \vartheta_k^{(t-1)}, \boldsymbol{y})$
- $\vartheta_3^{(t)}$ from $f(\vartheta_3|\vartheta_1^{(t)}, \vartheta_2^{(t)}, \vartheta_4^{(t-1)}, \ldots, \vartheta_k^{(t-1)}, \boldsymbol{y})$

  $\vdots$
- $\vartheta_j^{(t)}$ from $f(\vartheta_j|\vartheta_1^{(t)}, \vartheta_2^{(t)}, \ldots, \vartheta_{j-1}^{(t)}, \vartheta_{j+1}^{(t-1)}, \ldots, \boldsymbol{y})$

  $\vdots$
- $\vartheta_k^{(t)}$ from $f(\vartheta_k|\vartheta_1^{(t)}, \vartheta_2^{(t)}, \ldots, \vartheta_{k-1}^{(t)}, \boldsymbol{y})$

Step 4: If $t < T$, set $t = t + 1$ and return to step 3.

Else, end iteration.

Gibbs sampling is often appealing and works well when the full conditional posterior distributions are easy to sample from (SAS Intitute Inc., 2008).

4.2.3 Assessing convergence

Convergence diagnostics are tools used to determine whether the MCMC algorithm has reached its stationary or target distribution. Various convergence diagnostics have been proposed. However, these diagnostics test for conditions that are only necessary, but not sufficient, for convergence. Therefore, Cowles & Carlin (1996) recommend using a variety of diagnostics rather than relying on a single statistic or plot. A commonly used statistical diagnostic is the Gelman-Rubin criterion, first proposed by Gelman & Rubin (1992). This diagnostic involves running multiple MCMC chains and then comparing the variances within each chain and between chains. For M parallel MCMC chains, let $\boldsymbol{\vartheta}^t$, $t = 1, \ldots, n$, be the set of a single Markov chain output and let $\boldsymbol{\vartheta}_m^t$ denote the simulations for each $\boldsymbol{\vartheta}^t$, $m = 1, \ldots, M$. Then, the between-chain variance B is given by

$$B = \frac{n}{M-1} \sum_{m=1}^{M} (\bar{\boldsymbol{\vartheta}}_m^{\cdot} - \bar{\boldsymbol{\vartheta}}_{\cdot}^{\cdot})^2 \tag{4.8}$$

where $\bar{\boldsymbol{\vartheta}}_m^{\cdot} = \frac{1}{n}\sum_{t=1}^{n}\boldsymbol{\vartheta}_m^t$ and $\bar{\boldsymbol{\vartheta}}_{\cdot}^{\cdot} = \frac{1}{M}\sum_{m=1}^{M}\bar{\boldsymbol{\vartheta}}_m^{\cdot}$, and the within-chain variance W is computed from

$$W = \frac{1}{M}\sum_{m=1}^{M} s_m^2 \tag{4.9}$$

where $s_m^2 = \frac{1}{n-1}\sum_{t=1}^{n}(\boldsymbol{\vartheta}_m^t - \bar{\boldsymbol{\vartheta}}_m^{\cdot})^2$ (SAS Intitute Inc., 2008). The estimated posterior marginal variance of ϑ^t is a weighted average of B and W, and is given by

$$\widehat{Var}(\boldsymbol{\vartheta}^t|\boldsymbol{y}) = \frac{n-1}{n}W + \frac{1}{n}B \tag{4.10}$$

Assuming that the starting points in each chain are appropriately overdispersed, this quantity will overestimate the true variance $Var(\boldsymbol{\vartheta}^t|\boldsymbol{y})$. In contrast, W will underestimate $Var(\boldsymbol{\vartheta}^t|\boldsymbol{y})$ early in the sampling run as the individual chains would not have had the time to range over all of the stationary distribution (Gelman et al., 2014). However, when $n \to \infty$, both $\widehat{Var}(\boldsymbol{\vartheta}^t|\boldsymbol{y})$ and W will converge to the true variance. Therefore, the Gelman-Rubin diagnostic monitors convergence from

$$\widehat{R} = \sqrt{\frac{\widehat{Var}(\boldsymbol{\vartheta}^t|\boldsymbol{y})}{W}} \tag{4.11}$$

This is known as the potential scale reduction. This quantity declines to 1 as $n \to \infty$. Therefore, values of $\widehat{R}$ close to 1, usually less than 1.1, suggest that convergence has occurred (Gelman et al., 2014). Graphical diagnostic methods are also useful in monitoring convergence. Trace plots are commonly used. These are plots of the iterations against the simulated values. If all of the values lie within a region without any strong periodicities and tendencies, then convergence can be assumed. Other graphical methods involve plots of autocorrelations and ergodic means (Ntzoufras, 2009; SAS Intitute Inc., 2008).

4.3 Integrated nested Laplace approximation

The integrated nested Laplace approximation (INLA) is a computationally convenient alternative to MCMC methods. It approximates the posterior marginal distributions more accurately in a fully automated way. The INLA approach was introduced by Rue et al. (2009) for Bayesian inference on the broad class of latent Gaussian models. These models assume that the response variable y_i is conditionally independent given some underlying latent field $\boldsymbol{\xi}$ and a vector of hyperparameters $\boldsymbol{\vartheta}$. INLA approximates the posterior marginal distributions of the latent variables as well as those of the hyperparameters of the latent Gaussian model. The posterior marginal distributions of interest are given by

$$f(\xi_i|\boldsymbol{y}) = \int f(\xi_i|\boldsymbol{\vartheta}, \boldsymbol{y})\, f(\boldsymbol{\vartheta}|\boldsymbol{y})\, d\boldsymbol{\vartheta} \tag{4.12}$$

and

$$f(\vartheta_j|\boldsymbol{y}) = \int f(\boldsymbol{\vartheta}|\boldsymbol{y})\, d\boldsymbol{\vartheta_{-j}} \tag{4.13}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, k$. The INLA approach consists of three steps. The first step approximates the full posterior marginal distributions of the hyperparameters $f(\boldsymbol{\vartheta}|\boldsymbol{y})$. This is done by first approximating $f(\boldsymbol{\xi}|\boldsymbol{\vartheta}, \boldsymbol{y})$, the full conditional of $\boldsymbol{\xi}$, by a multivariate Gaussian density $\tilde{f}_G(\boldsymbol{\xi}|\boldsymbol{\vartheta}, \boldsymbol{y})$ evaluated at its mode, and then approximating $f(\boldsymbol{\vartheta}|\boldsymbol{y})$ using the Laplace approximation

$$\tilde{f}(\boldsymbol{\vartheta}|\boldsymbol{y}) \propto \left. \frac{f(\boldsymbol{\xi}, \boldsymbol{\vartheta}, \boldsymbol{y})}{\tilde{f}_G(\boldsymbol{\xi}|\boldsymbol{\vartheta}, \boldsymbol{y})} \right|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*(\boldsymbol{\vartheta})} \tag{4.14}$$

where $\boldsymbol{\xi}^*(\boldsymbol{\vartheta})$ is the mode of the full conditional for $\boldsymbol{\xi}$, for a given $\boldsymbol{\vartheta}$ (Martins et al., 2013). According to Rue & Martino (2007), this approximation is particularly accurate; even long MCMC runs are unable to detect any error in it.

The second step approximates the full conditional posterior marginal distribution for the latent variables $f(\xi_i|\boldsymbol{\vartheta}, \boldsymbol{y})$. For this approximation, three options are available. These options vary in terms of computational speed and accuracy. The fastest option uses the marginals of the Gaussian approximation $\tilde{f}_G(\boldsymbol{\xi}|\boldsymbol{\vartheta}, \boldsymbol{y})$ computed in the

previous step. This often leads to reasonable results. However, there can be errors in the location of the posterior mean and/or errors that are due to lack of skewness (Rue & Martino, 2007). One way to improve on the Gaussian approximation is to use the Laplace approximation which is given by

$$\tilde{f}_{LA}(\xi_i|\boldsymbol{\vartheta}, \boldsymbol{y}) \propto \left. \frac{f(\boldsymbol{\xi}, \boldsymbol{\vartheta}, \boldsymbol{y})}{\tilde{f}_G(\xi_{-i}|\xi_i, \boldsymbol{\vartheta}, \boldsymbol{y})} \right|_{\xi_{-i}=\xi^*_{-i}(\xi_i, \boldsymbol{\vartheta})} \tag{4.15}$$

where ξ_{-i} denotes the vector $\boldsymbol{\xi}$ with the i^{th} component excluded, $\tilde{f}_G(\xi_{-i}|\xi_i, \boldsymbol{\vartheta}, \boldsymbol{y})$ is the Gaussian approximation of $f(\xi_{-i}|\xi_i, \boldsymbol{\vartheta}, \boldsymbol{y})$ and $\xi^*_{-i}(\xi_i, \boldsymbol{\vartheta})$ is the modal configuration (Martins et al., 2013). The Laplace approximation is the most accurate of the three options. However, its computation can be very time consuming. The third option is derived from a Taylor series expansion of the Laplace approximation $\tilde{f}_{LA}(\xi_i|\boldsymbol{\vartheta}, \boldsymbol{y})$, up to third order. This is known as the simplified Laplace approximation. This option corrects the Gaussian approximation for location and skewness, but with a lower computational cost than the Laplace approximation.

In the third step, the full posterior marginal distributions computed in the previous steps are combined, and the posterior marginal distributions of interest are obtained by integrating out the relevant terms. The approximation for the posterior marginal distribution of the latent variables are obtained using the expression

$$\tilde{f}(\xi_i|\boldsymbol{y}) = \int \tilde{f}(\xi_i|\boldsymbol{\vartheta}, \boldsymbol{y})\, \tilde{f}(\boldsymbol{\vartheta}|\boldsymbol{y})\, d\boldsymbol{\vartheta} \approx \sum_{b=1}^{B} \tilde{f}(\xi_i|\boldsymbol{\vartheta}_b, \boldsymbol{y})\, \tilde{f}(\boldsymbol{\vartheta}_b|\boldsymbol{y})\, \Delta_b \tag{4.16}$$

which is evaluated using numerical integration on selected values of $\boldsymbol{\vartheta}$ with area weights Δ_b, $b = 1, \ldots, B$. In a similar manner, the approximation for the posterior marginal distribution of the hyperparameters $\tilde{f}(\vartheta_j|\boldsymbol{y})$ can be obtained. Rue et al. (2009) discuss two strategies for the selection of the integration points $\boldsymbol{\vartheta}_b$, namely the GRID strategy and the central composite design (CCD) strategy. The latter strategy is computationally less demanding and accurate enough for the computation of $\tilde{f}(\xi_i|\boldsymbol{y})$. However, when interest is on obtaining a more accurate estimate of $\tilde{f}(\vartheta_j|\boldsymbol{y})$, the GRID strategy may be necessary (Martino & Rue, 2010).

4.4 Model selection

The deviance information criterion (DIC) is a popular Bayesian model selection criterion designed to compare complex hierarchical models (Spiegelhalter et al., 2002). The DIC is a measure of model fit and complexity, and is defined as

$$DIC = \bar{D} + p_D \tag{4.17}$$

where $\bar{D}$ is the posterior mean of the deviance of the model and p_D is the effective number of parameters. Smaller values of the DIC indicate a better trade-off between model fit and model complexity. Thus, the model with the smallest value of DIC will be the optimal model (Adrion & Mansmann, 2012). Another useful quantity for comparing models from a Bayesian approach is the marginal likelihood (Rue et al., 2009). The marginal likelihood is the normalizing constant of the posterior distribution and for a certain model M is given by

$$f(\boldsymbol{y}|M) = \int L(\boldsymbol{y}|\boldsymbol{\vartheta}, M)\, f(\boldsymbol{\vartheta}|M)\, d\boldsymbol{\vartheta} \tag{4.18}$$

which is the average of the likelihood over the prior distribution. Hence, the value of the marginal likelihood will be larger when both the prior distribution and the likelihood are concentrated over the same parameter space, and the value will be smaller when the prior distribution emphasizes regions of the parameter space where the likelihood is low (Xie et al., 2011). Therefore, the larger the value of the marginal likelihood, the better the model fit (Roos & Held, 2011).

For the GLMM described in Section 3.2, and using Equation 4.4, the posterior distribution is then given by

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{\gamma}_k, \boldsymbol{\varphi}|\boldsymbol{y}) &\propto f(\boldsymbol{\beta}, \boldsymbol{\gamma}_k|\boldsymbol{\varphi})\, f(\boldsymbol{\varphi}) \prod_{k=1}^{K} f(\boldsymbol{y}_k|\boldsymbol{\beta}, \boldsymbol{\gamma}_k, \boldsymbol{\varphi}) \\ &\propto f(\boldsymbol{\varphi})\, f(\boldsymbol{\beta})|\boldsymbol{G}(\boldsymbol{\varphi})|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}\boldsymbol{\gamma}_k' \boldsymbol{G}(\boldsymbol{\varphi})^{-1} \boldsymbol{\gamma}_k + \sum_{k=1}^{K} \ln f(\boldsymbol{y}_k|\boldsymbol{\beta}, \boldsymbol{\gamma}_k) \right\} \end{aligned} \tag{4.19}$$

The parameters in Equation 4.19 can be estimated using the methods described above.

4.5 Application

4.5.1 MCMC

Bayesian analysis of the NIDS data was performed using the `R` software for Windows (R Core Team, 2016). In `R`, the package `MCMCglmm` allows for a GLMM to be fitted using the Gibbs sampler (Hadfield, 2010). This package, however, only allows for the default logit link function to be fitted for binary data. In Bayesian analysis, prior distributions for the parameters need to be specified. Fong et al. (2010) proposed a prior specification based on the Gamma distribution $\Gamma(0.5, 0.0164)$ for the variance components in a GLMM. This specification was derived using a classical Lemma which states that if $\boldsymbol{\gamma}_k$ follows a Gaussian distribution with zero mean and Gamma precision, that is $\boldsymbol{G}^{-1} \sim \Gamma(\alpha_1, \alpha_2)$, then the marginal posterior distribution of $\boldsymbol{\gamma}_k$ is a Student-t distribution with $2\alpha_1$ degrees of freedom, location $\mathbf{0}$ and scale $\sqrt{\alpha_2/\alpha_1}$ (Grilli et al., 2014). Thus, by selecting a marginal Student-t distribution with one degree of freedom for $\boldsymbol{\gamma}_k$ and imposing $\exp^{\boldsymbol{\gamma}_k} \in [0.1, 10]$ with probability 0.95, the hyperparameters $\alpha_1 = 0.5$ and $\alpha_2 = 0.0164$ are obtained. Grilli et al. (2014) compared this prior specification with two other commonly used specifications for variance components in GLMMs with binary outcomes and concluded that the $\Gamma(0.5, 0.0164)$ specification was the best of the three considered. Therefore, for estimation of the variance components in the analysis of the NIDS data set, we use the aforementioned prior specification. For the regression coefficients, the default multivariate Normal distribution with zero mean vector, and variance-covariance matrix with variances $1e + 10$, was used. The Gelman-Rubin diagnostic, together with trace plots, were used to monitor model convergence. For GLMM(1), 60000 iterations were used. This led to a potential scale reduction $\widehat{R} = 1$ for all parameters. Results were obtained within 30 minutes. For GLMM(2), 90000 iterations were used in order to obtain $\widehat{R} < 1.1$ for all parameters. Results for this model were obtained in approximately 80 minutes. The DIC values for GLMM(1) and GLMM(2) were 11 884.8 and 11 393.7, respectively. Thus, as with the classical estimation methods, GLMM(2) provides a better fit to the data. Therefore, only results obtained for this model are presented in this section. The results for GLMM(1), however, were very similar to those obtained using the classical methods. The results for GLMM(2) are given in Table 4.1.

Table 4.1: MCMC estimates and OR with 95% confidence intervals for GLMM(2)

Parameter	Estimate	Std. Error	OR (95% C.I.)	p-value
Intercept	-3.064	0.200		<0.001
Age				
15-24	ref	...	1	...
25-34	1.598	0.108	4.943 (4.003, 6.086)	<0.001
35-44	2.398	0.126	11.001 (8.525, 14.069)	<0.001
45-54	2.655	0.139	14.225 (10.859, 18.671)	<0.001
55-64	2.785	0.151	16.200 (12.025, 21.737)	<0.001
65+	2.242	0.166	9.412 (6.848, 13.040)	<0.001
Population group				
African	ref	...	1	...
Coloured	-0.299	0.134	0.742 (0.575, 0.968)	0.024
Asian/Indian	-1.716	0.373	0.180 (0.086, 0.368)	<0.001
White	-0.921	0.225	0.398 (0.257, 0.621)	<0.001
Marital status				
Married	ref	...	1	...
Living with partner	-0.899	0.149	0.407 (0.308, 0.552)	<0.001
Widow	-0.598	0.118	0.550 (0.440, 0.699)	<0.001
Divorced or separated	-0.504	0.199	0.604 (0.409, 0.883)	0.009
Never married	-0.793	0.091	0.452 (0.377, 0.535)	<0.001
Current smoker				
No	ref	...	1	...
Yes	-0.799	0.144	0.450 (0.340, 0.599)	<0.001
Alcohol consumption				
No	ref	...	1	...
Yes	-0.306	0.103	0.736 (0.605, 0.908)	0.003
Household expenditure on food				
Quartile I	ref	...	1	...
Quartile II	0.251	0.099	1.285 (1.060, 1.554)	0.010
Quartile III	0.220	0.104	1.246 (1.016, 1.523)	0.035
Quartile IV	0.341	0.116	1.406 (1.119, 1.758)	0.003
Education				
No schooling	ref	...	1	...
Primary	0.377	0.113	1.458 (1.169, 1.822)	<0.001
Secondary	0.628	0.123	1.874 (1.456, 2.361)	<0.001
Tertiary	0.162	0.267	1.306 (0.698, 1.984)	0.544

Continued on next page

Table 4.1 – *Continued from previous page*

Parameter	Estimate	Std. Error	OR (95% C.I.)	p-value
Employment status				
Unemployed	ref	...	1	...
Employed	0.164	0.077	1.178 (1.016, 1.372)	0.031
Household income quintile				
I	ref	...	1	...
II	-0.096	0.104	0.908 (0.742, 1.114)	0.364
III	0.109	0.111	1.115 (0.890, 1.379)	0.329
IV	0.099	0.117	1.104 (0.870, 1.380)	0.400
V	0.447	0.134	1.564 (1.217, 2.054)	<0.001
Geographical type				
Urban	ref	...	1	...
Traditional	0.519	0.095	1.680 (1.395, 2.026)	<0.001
Farms	0.040	0.147	1.041 (0.779, 1.385)	0.780

For GLMM(2), the MCMC method produced slightly different parameter estimates to those obtained using the classical estimation methods. In terms of standard errors and confidence intervals, MCMC produced slightly larger standard errors and, therefore, wider confidence intervals. However, the MCMC method led to the same inferences as the classical methods and hence, similar conclusions can be drawn. The estimated variance components for GLMM(2) using MCMC are given in Table 4.2. These estimates were inflated compared to those obtained using the classical methods.

Table 4.2: Variance component estimates for GLMM(2) using MCMC

Variance component	Estimate	Std. Error	95% C.I.
Cluster	0.0713	0.0280	0.0227-0.1268
Household within cluster	1.7283	0.2693	1.273-2.2220

4.5.2 INLA

In implementing the INLA method, the `R` package `INLA` was used. The prior distributions specified in the previous section were also employed in this analysis of the NIDS data set. GLMM(1) and GLMM(2) were fitted using all three link functions described in Section 3.2. The DIC and log-marginal likelihood (LML) values for these models are given in Table 4.3.

Table 4.3: DIC and LML for GLMM(1) and GLMM(2) with different link functions

Education Level	**GLMM(1)**			**GLMM(2)**		
	logit	probit	cloglog	logit	probit	cloglog
DIC	11 923.6	11 927.5	11 924.1	11 762.9	11 772.3	1.124e+278
LML	-6 108.7	-6.115.4	-6 113.9	-6 087.6	-6 090.2	-6 089.0

For both GLMM(1) and GLMM(2), the logit link function provides the best fit. In comparing these two models, GLMM(2) has a lower DIC and higher LML, and is, therefore, the better model. Results for this model, given in Table 4.4, were obtained within 15 minutes.

Table 4.4: INLA estimates and OR with 95% confidence intervals for GLMM(2)

Parameter	**Estimate**	**Std. Error**	**OR (95% C.I.)**
Intercept	-2.362	0.142	
Age			
15-24	ref	...	1
25-34	1.247	0.080	3.480 (2.974, 4.080)
35-44	1.847	0.089	6.341 (5.323, 7.561)
45-54	2.038	0.098	7.675 (6.341, 9.309)
55-64	2.143	0.109	8.525 (6.890, 10.559)
65+	1.728	0.119	5.629 (4.459, 6.828)
Population group			
African	ref	...	1
Coloured	-0.223	0.097	0.800 (0.660, 0.968)
Asian/Indian	-1.310	0.267	0.270 (0.158, 0.452)
White	-0.684	0.161	0.505 (0.368, 0.691)

Continued on next page

Table 4.4 – *Continued from previous page*

Parameter	Estimate	Std. Error	OR (95% C.I.)
Marital status			
Married	ref	…	1
Living with partner	-0.664	0.108	0.515 (0.416, 0.636)
Widow	-0.444	0.085	0.641 (0.543, 0.758)
Divorced or separated	-0.390	0.145	0.677 (0.509, 0.899)
Never married	-0.599	0.065	0.549 (0.484, 0.624)
Current smoker			
No	ref	…	1
Yes	-0.623	0.106	0.536 (0.435, 0.660)
Alcohol consumption			
No	ref	…	1
Yes	-0.246	0.077	0.782 (0.672, 0.908)
Household expenditure on food			
Quartile I	ref	…	1
Quartile II	0.185	0.069	1.203 (1.050, 1.380)
Quartile III	0.165	0.075	1.179 (1.018, 1.365)
Quartile IV	0.249	0.083	1.283 (1.090, 1.511)
Education			
No schooling	ref	…	1
Primary	0.294	0.084	1.342 (1.139, 1.582)
Secondary	0.491	0.089	1.634 (1.373, 1.944)
Tertiary	0.160	0.194	1.174 (0.802, 1.716)
Employment status			
Unemployed	ref	…	1
Employed	0.118	0.056	1.125 (1.009, 1.255)
Household income quintile			
I	ref	…	1
II	-0.069	0.077	0.933 (0.803, 1.084)
III	0.086	0.080	1.090 (0.931, 1.274)
IV	0.079	0.084	1.082 (0.918, 1.276)
V	0.342	0.095	1.408 (1.168, 1.697)
Geographical type			
Urban	ref	…	1
Traditional	0.389	0.068	1.476 (1.290, 1.685)
Farms	0.030	0.105	1.030 (0.838, 1.265)

For GLMM(2), the results obtained using the INLA method were very similar to those obtained using the classical methods and, therefore, lead to similar conclusions. The estimated variance components for GLMM(2) using INLA are given in Table 4.5. Furthermore, these estimates are consistent with those obtained using the classical methods.

Table 4.5: Variance component estimates for GLMM(2) using INLA

Variance component	**Estimate**	**Std. Error**	**95% C.I.**
Cluster	0.2037	0.036	0.1397-0.2769
Household within cluster	0.6487	0.059	0.5346-0.7618

Chapter 5

Discussions and Conclusions

In this study, a GLMM was used to investigate the relationship between obesity among females in SA and selected demographic, lifestyle, socioeconomic and environmental variables, and to identify significant risk factors associated with obesity. A survey logistic model (Appendix A) was used as an alternative method to achieve this purpose. However, this method is design-based, which assumes the observations are independent. Therefore, the GLMM was fitted to the NIDS data set, first to account for possible heterogeneity among clusters only (GLMM(1)), and then extended to account for possible heterogeneity among households nested within clusters (GLMM(2)). One of the objectives of this study was to examine and compare the different classical and Bayesian methods used in fitting a GLMM. For the classical approach, results were obtained using the Laplace approximation, adaptive Gauss-Hermite quadrature and PQL methods. The Bayesian approach was demonstrated via MCMC and INLA. For all estimation methods, the best fitting model was GLMM(2). Even though only the logit link could be explored under MCMC, this link function provided the best fit under all other estimation methods. These methods produced very similar results, except for MCMC, which produced slightly inflated parameter estimates for both the fixed effects and variance components. All methods differed significantly in terms of computational speed. The classical methods had shorter run-times compared to the Bayesian methods, which is one of the advantages of the classical approach over the Bayesian approach (Hall, 2012). However, with the adaptive Gauss-Hermite quadrature, the computational time increased considerably as the number of quadrature points were increased. For the Bayesian methods, the run-times for INLA was considerably shorter compared to MCMC.

With the GLMM, the variables age, population group, marital status, smoking status, alcohol consumption, household expenditure on food, education level, employment status, household income and geographical type were all found to be significantly associated with obesity among females. In contrast, the survey logistic model found employment status to be insignificant. Furthermore, this model found significant two-way interactions between population group and education level, population group and geographical type as well as education level and alcohol consumption. This may be an effect of taking sampling weights into consideration. Despite these differences and the slight differences in parameter estimates, the same conclusions can be drawn. Older females were associated with a higher risk of obesity, with those between the ages of 55 and 64 years old being most at risk. African females and those married were at significantly elevated risk of obesity which is consistent with the findings of Malhotra et al. (2008), Puoane et al. (2002) and Sartorius et al. (2015). Conversely, female smokers and those who consumed alcohol were associated with a lower risk of obesity. Total household expenditure on food was shown to be significantly associated with risk of obesity, with females with a total household expenditure on food within the highest quartile having the highest risk. Similar to the study by Puoane et al. (2002), females with a tertiary education, or those with no schooling, were associated with a lower risk of obesity. Compared to females who were unemployed, those who were employed were at a higher risk. Furthermore, females belonging to the highest household income quintile were also at a higher risk of obesity. This suggests that higher socioeconomic status is associated with increased risk of obesity. These results are in agreement with those found by Case & Menendez (2009), Kruger et al. (2012) and Sartorius et al. (2015). Compared to urban areas, a female's risk of obesity was greater in traditional areas and farms, which is in contrast to the findings of Puoane et al. (2002) and Sartorius et al. (2015). However, urbanization is associated with the adoption of a Westernized lifestyle, in particular changes in diet, and a study by Bourne et al. (2002) has shown that these changes are occurring in non-urban areas as well. A study by Yu & Lippert (2016) indicates that neighbourhood crime was associated with decreased physical activity and increased obesity. However, in this study, crime was found to be insignificant. Furthermore, exercise frequency was found to be insignificant which is consistent with the findings by Malhotra et al. (2008), but in contradiction to those by Alaba &

Chola (2014), Butzlaff & Minos (2016) and Cois & Day (2015). Depression was also found to be insignificant, with the distribution of obesity prevalence being almost the same for females classified as suffering from depression, and those who are not.

The etiology of obesity is one that is multifaceted. It can change over time and differ across regions. A study that takes these two factors into account through spatiotemporal modelling would be highly recommended. Furthermore, the prior distributions used in this study were based on those proposed by Fong et al. (2010). Even though Grilli et al. (2014) have shown that these prior specifications work well, further studies should investigate the sensitivity of these prior specifications. Finally, this study used a complete case analysis. Further studies should take missing data into consideration.

References

Abramowitz, M. & Stegun, I. A. (1974). *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. New York: Dover.

Adrion, C. & Mansmann, U. (2012). Bayesian model selection techniques as decision support for shaping a statistical analysis plan of a clinical trial: An example from a vertigo phase III study with longitudinal count data as primary endpoint. *BMC Medical Research Methodology*, 12(137).

Affenito, S. G., Franko, D. L., Striegel-Moore, R. H., & Thompson, D. (2012). Behavioral determinants of obesity: Research findings and policy implications. *Journal of Obesity*, (pp. doi:10.1155/2012/150732).

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley and Sons, Inc. Hoboken, New Jersey, 2nd edition.

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. John Wiley and Sons, Inc. Hoboken, New Jersey.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 317–332.

Alaba, O. & Chola, L. (2014). Socioeconomic inequalities in adult obesity prevalence in South Africa: A decomposition analysis. *International Journal of Environmental Research and Public Health*, 11, 3387–3406.

Ardington, C. & Gasealahwe, B. (2012). Health: Analysis of the NIDS wave 1 and 2 datasets. *SALDRU, University of Cape Town.*, SALDRU Working Paper Number 80 / NIDS Discussion Paper 2012/3.

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

Bourne, L. T., Lambert, E. V., & Steyn, K. (2002). Where does the black population of South Africa stand in the nutrition transition. *Public Health Nutrition*, 5(1A), 157–162.

Box, G. E. P. & Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis*. John Wiley and Sons, Inc. New York.

Breslow, N. (2003). Whither PQL? *UW Biostatistics Working Paper Series. Working Paper 192. http://biostats.bepress.com/uwbiostat/paper192*.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421), 9–25.

Brooks, S. P. (1998). Markov chain Monte Carlo method and its application. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 47(1), 69–100.

Butzlaff, I. & Minos, D. (2016). Understanding the drivers of overweight and obesity in developing countries: The case of South Africa. *GlobalFood Discussion Papers*, 78.

Capanu, M., Gnen, M., & Begg, C. B. (2013). An assessment of estimation methods for generalized linear mixed models with binary outcomes. *Stat Med*, 32(26).

Case, A. & Menendez, A. (2009). Sex differences in obesity rates in poor countries: Evidence from South Africa. *Economics and Human Biology*, 7, 271–282.

Cois, A. & Day, C. (2015). Obesity trends and risk factors in the South African adult population. *BMC Obesity*, 2(42), DOI10.1186/s40608–015–0072–2.

Cowles, M. K. & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434), 883–904.

Cox, D. R. & Snell, E. J. (1989). *Analysis of Binary Data*. Chapman and Hall, New York, 2nd edition.

Craiu, R. V. & Rosenthal, J. S. (2014). Bayesian computation via Markov chain Monte Carlo. *Annual review of statistics and its application*, 1, 179–201.

de Villiers, L., Brown, M., Woolard, I., Daniels, R. C., & Leibbrandt, M. (2013). National Income Dynamics Study wave 3 user manual. *Cape Town: Southern Africa Labour and Development Research Unit*.

Department of Health:RSA. (2015). Strategy for the prevention and control of obesity in South Africa: 2015-2020.

Finney, D. J. (1971). *Probit Analysis*. Cambridge University Press, Cambridge, UK, 3rd edition.

Fong, Y., Rue, H., & Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics*, 11(3), 397–412.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis*. Taylor and Francis Group, LLC. Boca Raton, 3rd edition.

Gelman, A. & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457–472.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice* (pp. 1–20).: Chapman and Hall. London.

Glickman, M. E. & van Dyk, D. A. (2007). Basic Bayesian methods. In *Methods in Molecular Biology, vol. 404: Topics in Biostatistics* (pp. 59–74).: Humana Press Inc. Totowa, NJ.

Goedecke, J. H., Jennings, C. L., & Lambert, E. V. (2006). Obesity in South Africa. In *K. Steyn, J. Fourie, N. Temple (eds). Chronic Diseases of Lifestyle in South Africa: 1995 - 2005* (pp. 65–79).: South African Medical Research Council, Technical Report. Cape Town.

Grilli, L., Metelli, S., & Rampichini, C. (2014). Bayesian estimation with integrated nested Laplace approximation for binary logit mixed models. *Journal of Statistical Computation and Simulation*, 2718–2726.

Hadfield, J. D. (2010). MCMC methods for multi-response generalised linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, 33(2), 1–22.

Hall, B. (2012). Bayesian inference. *Statisticat, LLC*.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358), 320–338.

Hastings, W. K. (1970). Monte Carlo sampling methods using markov chains and their applications. *Biometrika*, 57, 97–109.

Hedeker, D. (2005). Generalized linear mixed models. In *B. Everitt, and D. Howell (Eds.) Encyclopedia of Statistics in Behavioral Science*: John Wiley and Sons, Ltd.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied Survey Data Analysis*. Taylor and Francis Group, LLC.

Hosmer, D. & Lemeshow, S. (2000). *Applied Logistic Regression*. John Wiley and Sons, 2nd edition.

Huang, T. T., Drewnowski, A., Kumanyika, S. K., & Glass, T. A. (2009). A systems-oriented multilevel framework for addressing obesity in the 21st century. *Preventing Chronic Disease*, 6(3). $http://www.cdc.gov/pcd/issues/2009/jul/09\_0013.htm$. Accessed: [July 2016].

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and their Applications*. Springer Science + Business Media, Inc. New York.

Ker, H. W. (2014). Application of hierarchical linear models/linear mixed-effects models in school effectiveness research. *Universal Journal of Educational Research*, 2(2), 173–180.

Kruger, A., Wissing, M. P., Towers, G. W., & Doak, C. M. (2012). Sex differences independent of other psychosociodemographic factors as a predictor of body mass index in black South African adults. *Journal of Health, Population and Nutrition*, 30(56), PMID:22524120.

Kuss, O. (2002). How to use SAS for logistic regression with correlated data. *Statistics and Data Analysis*.

Lakerveld, J., Brug, J., Bot, S., Teixeira, P. J., Rutter, H., Woodward, E., Samdal, O., Stockley, L., Bourdeaudhuij, I. D., van. Assema, P., Robertson, A., Lobstein, T., Oppert, J. M., dny, R., & Nijpels, G. (2012). Sustainable prevention of obesity through integrated strategies: The SPOTLIGHT projects conceptual framework and design. *BMC Public Health*, 12(793). $http://www.biomedcentral.com/1471-2458/12/793$.

Leibbrandt, M., Woolard, I., & de Villiers, L. (2009). Methodology: Report on NIDS wave 1. *Technical Paper 1*.

Lin, X. & Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91(435), 1007–1016.

Liu, Q. & Pierce, D. A. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, 81(3), 624–629.

Malhotra, R., Hoyo, C., Østbye, T., Hughes, G., Schwartz, D., Tsolekile, L., Zulu, J., & Puoane, T. (2008). Determinants of obesity in an urban township of South Africa. *South African Journal of Clinical Nutrition*, 21(4), 315–320.

Malik, V. S., Willett, W. C., & Hu, F. B. (2012). Global obesity: trends, risk factors and policy implications. *Nature Reviews: Endocrinology*.

Martino, S. & Rue, H. (2010). Case studies in Bayesian computation using INLA. In *P. Mantovan and P. Secchi (Eds.) Complex Data Modeling and Computationally Intensive Statistical Methods* (pp. 99–114).: Springer-Verlag Italia.

Martins, T. G., Simpson, D., Lindgren, F., & Rue, H. (2013). Bayesian computing with INLA: New features. *Computational Statistics and Data Analysis*, 67, 68–83.

McCormick, B., Stone, I., & Team, C. A. (2012). Economic costs of obesity and the case for government intervention. *Obesity reviews*, 8(Suppl. 1), 161–164.

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall. London, 2nd edition.

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, linear, and mixed models*. John Wiley and Sons, Inc. Hoboken, New Jersey.

Micklesfield, L. K., Lambert, E. V., Hume, D. J., Chantler, S., Pienaar, P. R., Dickie, K., Puoane, T., & Goedecke, J. H. (2013). Socio-cultural, environmental and behavioural determinants of obesity in black South African women. *Cardiovascular journal of Africa*, 24, 369–375.

Mofuoa, K. (2015). Social embeddedness of agriculture for human progress in the nineteenth century Southern Africa: Evidence and lessons from Lesotho. *International Journal of Development Research*, 5(12), 6369–6379.

Molenberghs, G. & Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer Science + Business Media, Inc. New York.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, 135(3), 370–384.

Ntzoufras, I. (2009). *Bayesian Modelling Using WinBUGS*. John Wiley and Sons, Inc. Hoboken, New Jersey.

Puoane, T., Steyn, K., Bradshaw, D., Laubscher, R., Fourie, J., Lambert, V., & Mbananga, N. (2002). Obesity in South Africa: The South African demographic and health survey. *Obesity Research*, 10, 1038–1048.

R Core Team. (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. $https://www.R-project.org/$.

Raudenbush, S. W., Yang, M.-L., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, 9, 141–157.

Rencher, A. C. & Schaalje, G. B. (2008). *Linear models in statistics*. John Wiley and Sons, Inc. New Jersey, 2nd edition.

Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice* (pp. 45–58).: Chapman and Hall. London.

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

Rue, H. & Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *Journal of Statistical Planning and Inference*, 137(10), 3177–3192.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B*, 71(2), 319–392.

Sartorius, B., Veerman, L. J., Manyema, M., Chola, L., & Hofman, K. (2015). Determinants of obesity and associated population attributability, South Africa: Empirical evidence from a national panel survey, 2008-2012. *PLoS ONE*, 10(6), e0130218.doi:10.1371/journal.pone.0130218.

SAS Intitute Inc. (2008). *$SAS/STAT^{®}$ 9.2 Users Guide*. Cary, NC: SAS Institute Inc.

SAS Intitute Inc. (2013). *$SAS/STAT^{®}$ 13.1 Users Guide*. Cary, NC: SAS Institute Inc.

Searle, S. R., Casella, G., & McCulloch, C. E. (2006). *Variance Components*. John Wiley and Sons, Inc. Hoboken, New Jersey.

Self, S. G. & Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82, 605–610.

Some, M., Rashied, N., & Ohonba, A. (2016). The impact of obesity on employment in South Africa. *Studies in Economics and Econometrics*, 40(2), 87–104.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583–639.

Stats SA. (2012). *Census 2011, Statistical release*. Statistics South Africa, Pretoria.

Stats SA. (2015). *Mid-year population estimates 2015, Statistical release*. Statistics South Africa, Pretoria.

Stram, D. O. & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50(4), 1171–1177.

The GBD 2013 Obesity Collaboration (2014). Global, regional and national prevalence of overweight and obesity in children and adults 1980-2013: A systematic analysis. *Lancet*, 384(9945), 766–781.

Tierney, L. (1996). Introduction to general state-space Markov chain theory. In *Markov Chain Monte Carlo in Practice* (pp. 59–74).: Chapman and Hall. London.

Tuerlinckx, F., Rijmen, F., Verbeke, G., & Boeck, P. D. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, 59, 225–255.

Tutz, G. (2012). *Regression for Categorical Data*. Cambridge University Press. New York.

Wade, P. R. (2000). Bayesian methods in conservation biology. *Conservation Biology*, 14(5), 1308–1316.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61(3), 439–447.

Wedderburn, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, 63(1), 27–32.

Wolfinger, R. & O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48, 233–243.

World Bank Group. (2016). *Migration and Remittances Factbook*. World Bank Publications, 3rd edition.

World Health Organization. (2014). Global status report on noncommunicable diseases.

WWF-SA (2016). Oceans facts and futures: Valuing South Africa's ocean economy. *WWF-SA, Cape Town, South Africa*.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., & Chen, M.-H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology*, 60(2), 150–160.

Yu, E. & Lippert, A. M. (2016). Neighborhood crime rate, weight-related behaviors, and obesity: A systematic review of the literature. *Sociology Compass*, 10(3), 187–207.

Zhang, D. & Lin, X. (2008). Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. In *D. Dunson (Ed.) Random Effect and Latent Variable Model Selection*, volume 192 of Lecture notes in Statistics (pp. 19–36).: Springer New York. 101.

Zhang, W., OBrien, N., Forrest, J. I., Salters, K. A., Patterson, T. L., Montaner, J. S. G., Hogg, R. S., & Lima, V. D. (2012). Validating a shortened depression scale (10 item ces-d) among HIV-positive people in British Columbia, Canada. *PLoS ONE*, 7(7), e40793. doi:10.1371/journal.pone.0040793.

Zhao, Y., Staudenmayer, J., Coull, B. A., & Wand, M. P. (2006). General design Bayesian generalized linear mixed models. *Statistical Science*, 21(1), 35–51.

Appendix A

Survey Logistic Model Analysis

With the aim of accounting for the sampling design of the NIDS survey, a survey logistic model was fitted to the NIDS data. This was done using the SAS procedure `PROC SURVEYLOGISTIC`, which allows sampling weights to be specified in the analysis. The usual model selection procedures (stepwise, forward and backward) have not yet been included in SAS version 9.4 for `PROC SURVEYLOGISTIC`. Therefore, model selection was performed using similar steps to those suggested by (Hosmer & Lemeshow, 2000). In the first step, bivariate analyses of the relationship of obesity and all demographic, lifestyle, socio-economic and environmental variables were performed one at a time. The variables age, population group, marital status, current smoker, alcohol consumption, household expenditure on food, education level, employment status, household income quintile, and geographical type had a bivariate association with obesity at p-values less than 0.1. These variables were then selected for inclusion into a multivariate survey logistic model. To determine the final model, a backward selection procedure was performed and insignificant variables, based on the type III analysis of effects, were removed from the model one at a time until only significant variables were left. The remaining variables were age, population group, marital status, current smoker, alcohol consumption, household expenditure on food, education, household income quintile, and geographical type. Only employment status was found to be insignificant when included in the multivariate model. All interaction terms of the remaining variables were explored. The interaction terms that led to a large decrease in the deviance were selected. The final model, given in the following table, included three two-way interaction terms.

Type III analysis of effects for the final SLR model

Effect	**DF**	**Chi-Square**	**P-Value**
Age	5	97.51	<0.0001
Population group	3	35.67	<0.0001
Marital status	4	4.78	0.0030
Current smoker	1	7.87	0.0075
Alcohol consumption	1	13.41	0.0007
Household expenditure on food	3	3.42	0.0259
Education	3	19.25	<0.0001
Household income quintile	4	1.11	0.0363
Geographical type	2	17.78	<0.0001
Population group $*$ Education	8	14.87	<0.0001
Population group $*$ geographical type	5	16.97	<0.0001
Alcohol consumption $*$ Education	3	3.56	0.0224

For variance estimation of the model, the Taylor series approximation method, which is the default in SAS `PROC SURVEYLOGISTIC` was used. The predictive accuracy of the model was found to be in an acceptable range, with a concordance index (c) of 0.721, indicating that, in predicting the probability of a positive obesity result, 72.1% of the cases were predicted correctly. The parameter estimates, adjusted odds ratios (aOR) with their 95% confidence intervals, and the p-values are given in the table that follows. Compared to the GLMM fitted in this study, the survey logistic model produced slightly different results. With the survey logistic model, employment status was found to be insignificant. Furthermore, the two-way interactions between population group and education level, population group and geographical type, as well as education level and alcohol consumption were found to be significant. The relationship between population group and education level is presented in the figure on page 70. This figure reveals that African females who reported receiving up to a tertiary education were most at risk of being obese. For all levels of education, Asian/Indian females had the lowest risk of obesity. For White females, the risk of obesity increased as the level of education increased, but decreased at tertiary level.

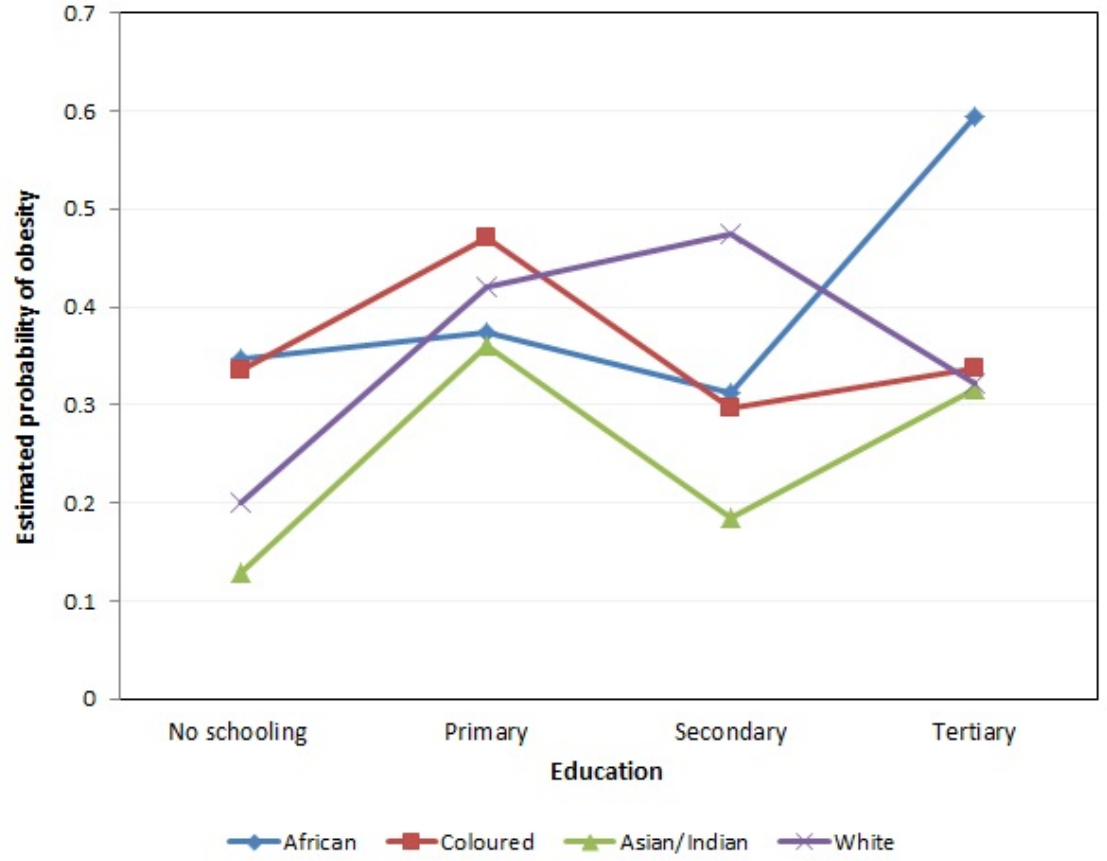
Estimates and aOR with 95% confidence intervals for the survey logistic model

Parameter	Estimate	Std. Error	aOR (95% C.I.)	p-value
Intercept	-2.4884	0.1592		<0.001
Age				
15-24	ref	…	1	…
25-34	1.245	0.097	3.472 (2.853, 4.225)	<0.001
35-44	1.735	0.105	5.669 (4.587, 7.007)	<0.001
45-54	1.977	0.118	7.223 (5.693, 9.165)	<0.001
55-64	2.116	0.109	8.299 (6.668, 10.328)	<0.001
65+	1.585	0.175	4.878 (3.428, 6.943)	<0.001
Population group				
African	ref	…	1	…
Coloured	-0.195	0.109	0.823 (0.507, 0.904)	0.024
Asian/Indian	-1.352	0.329	0.259 (0.225, 0.712)	<0.001
White	-0.689	0.234	0.502 (0.178, 0.668)	<0.001
Marital status				
Married	ref	…	1	…
Living with partner	-0.489	0.188	0.613 (0.420, 0.895)	0.012
Widow	-0.270	0.152	0.763 (0.561, 0.963)	0.048
Divorced or separated	-0.184	0.188	0.831 (0.569, 0.910)	0.033
Never married	-0.385	0.101	0.680 (0.555, 0.834)	<0.001
Current smoker				
No	ref	…	1	…
Yes	-0.386	0.138	0.680 (0.515, 0.897)	0.008
Alcohol consumption				
No	ref	…	1	…
Yes	-0.607	0.126	0.545 (0.283, 0.950)	0.006
Household expenditure on food				
Quartile I	ref	…	1	…
Quartile II	0.155	0.092	1.168 (1.030, 1.405)	0.009
Quartile III	0.184	0.101	1.202 (1.020, 1.474)	0.008
Quartile IV	0.439	0.141	1.551 (1.167, 2.061)	0.003
Education				
No schooling	ref	…	1	…
Primary	0.323	0.075	1.381 (1.189, 1.606)	<0.001
Secondary	0.535	0.111	1.707 (1.365, 2.136)	<0.001
Tertiary	0.884	0.251	2.421 (0.685, 3.629)	0.100

Continued on next page

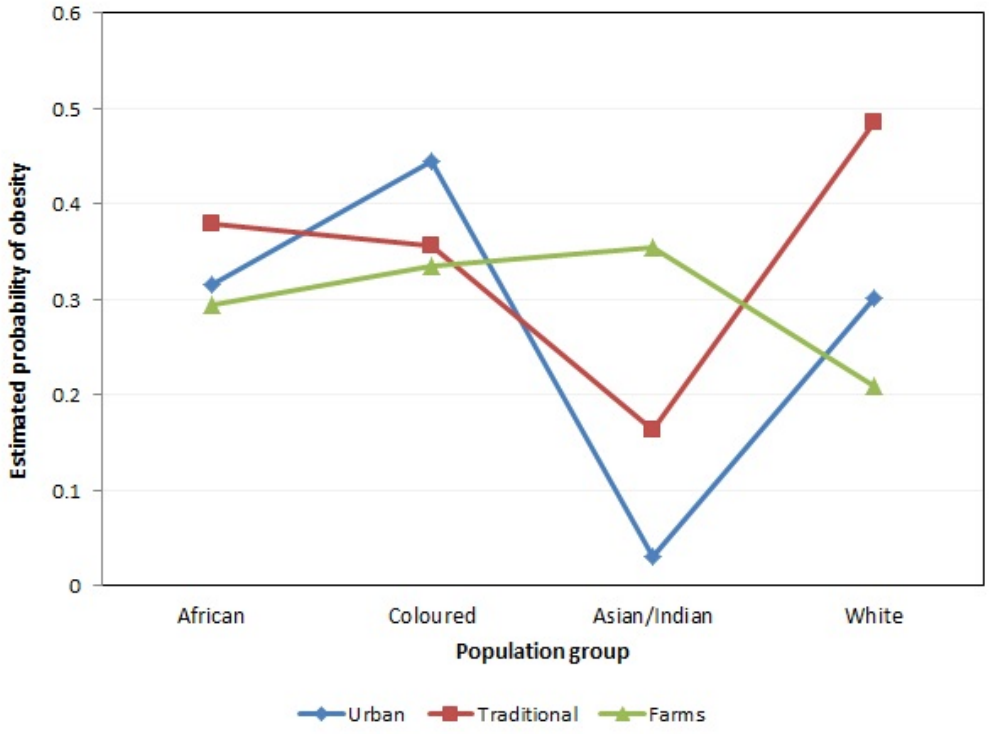
Continued from previous page

Parameter	Estimate	Std. Error	OR (95% C.I.)	p-value
Household income quintile				
I	ref	…	1	…
II	-0.001	0.113	0.999 (0.795, 1.256)	0.995
III	0.147	0.080	1.159 (0.986, 1.362)	0.072
IV	0.102	0.107	1.107 (0.892, 1.374)	0.346
V	0.206	0.176	1.229 (1.161, 1.752)	<0.001
Geographical type				
Urban	ref	…	1	…
Traditional	0.314	0.097	1.369 (1.126, 1.662)	<0.001
Farms	0.076	0.178	1.079 (0.647, 1.328)	0.674
Population group ∗ Education				
African and No schooling	ref	…	1	…
Coloured and Primary	0.481	0.250	1.618 (0.976, 2.678)	0.061
Coloured and Secondary	-0.380	0.291	0.634 (0.381, 1.230)	0.198
Coloured and Tertiary	-0.354	0.090	0.702 (0.115, 4.293)	0.696
Asian/Indian and Primary	1.532	0.486	4.627 (1.738, 12.33)	0.003
Asian/Indian and Secondary	0.412	0.190	1.510 (1.029, 2.217)	0.036
Asian/Indian and Tertiary	1.548	0.087	4.702 (0.811, 3.696)	0.083
White and Secondary	-0.415	0.104	0.660 (0.533, 0.824)	<0.001
White and Tertiary	-0.521	0.113	0.594 (0.433, 0.820)	<0.001
Population group ∗ geographical type				
African and Urban	ref	…	1	…
Coloured and Traditional	-1.064	0.078	0.345 (0.071, 1.669)	0.180
Coloured and Farms	-0.776	0.082	0.460 (0.087, 2.425)	0.351
Asian/Indian and Traditional	1.782	0.145	5.941 (4.531, 8.406)	<0.001
Asian/Indian and Farms	2.105	0.103	8.207 (3.397, 10.580)	<0.001
White and Traditional	1.105	0.318	3.019 (0.845, 10.794)	0.087
Alcohol consumption ∗ Education				
No and No schooling	ref	…	1	…
Yes and Primary	-0.147	0.145	0.863 (0.352, 2.115)	0.742
Yes and Secondary	0.588	0.103	1.800 (0.294, 4.059)	0.152
Yes and Tertiary	-1.063	0.168	0.345 (0.049, 0.533)	0.008

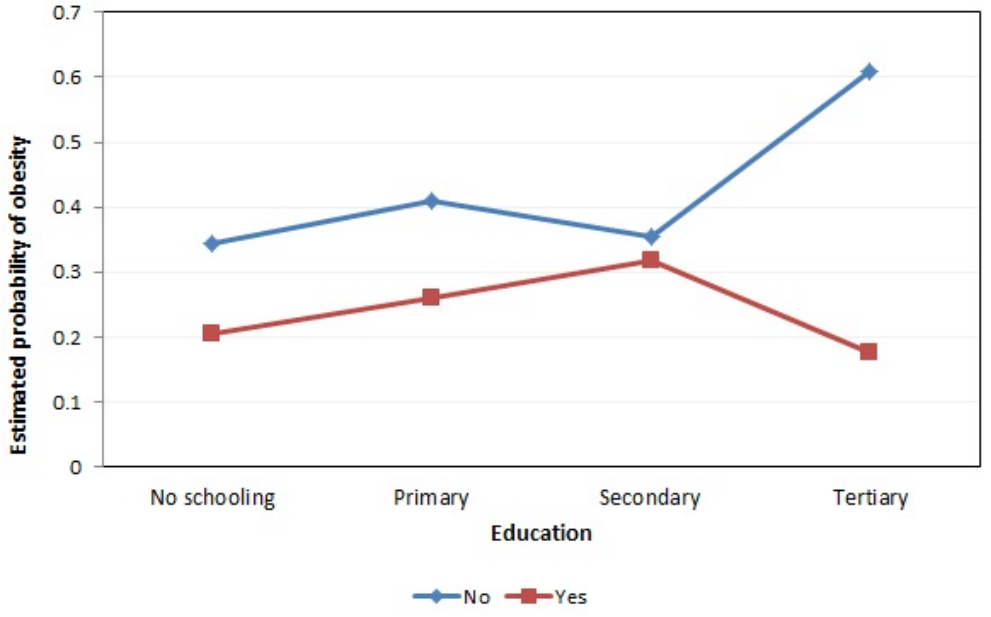


Estimated probability of obesity associated with the interaction of population group and education level

The first figure on the next page shows the estimated probability of obesity associated with population group and geographical type. Females living in urban or traditional areas had a higher risk of obesity compared to those living on farms, however excluding Asian/ Indian females who instead, had a higher risk living on farms than in urban or traditional areas. The estimated probability of obesity associated with education level and alcohol consumption, given in the second figure on page 71, indicates that females who received up to a tertiary education and did not consume alcohol were most at risk of being obese. The risk of obesity was low for females who consumed alcohol across all levels of education. For all other variables, the results obtained with the survey logistic model were similar to those obtained with the GLMM and therefore similar conclusions can be drawn.



Estimated probability of obesity associated with the interaction of population group and geographical type



Estimated probability of obesity associated with the interaction of education level and alcohol consumption

SAS codes

The following SAS codes were used to fit the final survey logistic model to the NIDS data:

```
proc surveylogistic data = dataF;

stratum DC / list;

cluster cluster;

class X1 X3 X4 X8 X9 X11 X12 X14 X15 / param=glm;

model BMI (descending) = X1 X3 X4 X8 X9 X11 X12 X14 X15 X3*X12
X3*X15 X9*X12/ clparm;

weight weights; run;

############################################################
```

where X1 = age of female; X3 = population group; X4 = marital status; X8 = current smoking status; X9 = alcohol consumption; X11 = total household expenditure on food; X12 = education; X14 = total household income; X15 = geographical type

Appendix B

Codes

The SAS codes used in the analysis of the NIDS data are given below:

GLMM(1)

```
proc glimmix data=dataF method=laplace
plots=studentpanel(conditional);

class X1 X3 X4 X8 X9 X11 X12 X13 X14 X15

cluster; model BMI(descending) = X1 X3 X4 X8 X9 X11 X12 X13 X14 X15/
link=logit dist=binary oddsratio solution;

random intercept/subject=cluster type=VC;
covtest zerog;
run;
```

GLMM(2)

```
proc glimmix data=dataF method=laplace
plots=studentpanel(conditional);

class X1 X3 X4 X8 X9 X11 X12 X13 X14 X15 cluster hhid;

model BMI (descending) = X1 X3 X4 X8 X9 X11 X12 X13 X14 X15/
link=logit dist=binary oddsratio solution;

random intercept/ subject=cluster;

random intercept/ subject=hhid(cluster);
covtest zerog;
run;
```

The R codes in the Bayesian analysis of the NIDS data is given below:

MCMC method

```
#####################################################

library(MCMCglmm)


####GLMM(2)


mc3 <- MCMCglmm(Yf ~ X1 + X3 + X4 + X8 + X9 + X11 + X12
                        + X13 + X14 + X15,
                    random=~cluster + cluster:hhid, data=dataF,
                    family="categorical", prior=list(R=list(V=1, fix=1),
                    G=list(G1=list(V=0.0164, nu=0.5),
                    G2=list(V=0.0164, nu=0.5))),
                    nitt=90000, slice=F)

summary(mc3$VCV) summary(mc3$Sol) plot(mc3) summary(mc3)


mc4 <- MCMCglmm(Yf ~ X1 + X3 + X4 + X8 + X9 + X11 + X12
                        + X13 + X14 + X15,
                    random=~cluster + cluster:hhid, data=dataF,
                    family="categorical", prior=list(R=list(V=1, fix=1),
                    G=list(G1=list(V=0.0164, nu=0.5),
                    G2=list(V=0.0164, nu=0.5))),
                    nitt=90000, slice=F)

mcmc2 <- mcmc.list(mc3, mc4) gelman.diag(mcmc2)


#############################################
```

INLA method

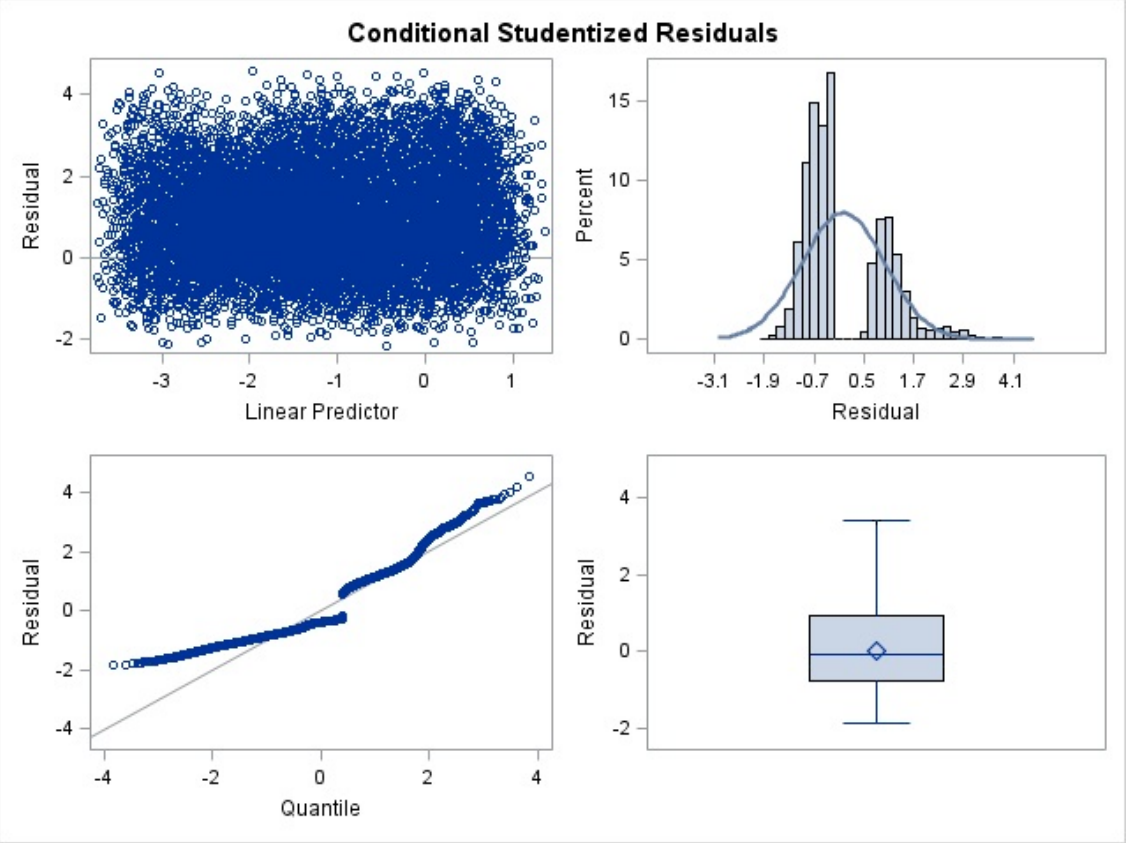
```
#####################################################

library(INLA)

####GLMM(2)

dataF.inla.fit.2 = inla(Yf ~ X1 + X3 + X4 + X8 + X9 + X11
                            + X12 + X13 + X14 + X15
                            + f(cluster, model="iid", param=c(0.5,.0164))
                            + f(cluster:hhid, model="iid", param=c(0.5,.0164)),
                            data=dataF, family="binomial",
                            control.compute=list(dic=TRUE))
dataF.hyperpar.2 = inla.hyperpar(dataF.inla.fit.2)
summary(dataF.inla.fit.2) F2 = inla.contrib.sd(dataF.inla.fit.2)
F2$hyper plot(dataF.inla.fit.2)

#############################################
```

where X1 = age of female; X3 = population group; X4 = marital status; X8 = current smoking status; X9 = alcohol consumption; X11 = total household expenditure on food; X12 = education; X13 = employment; X14 = total household income; X15 = geographical type

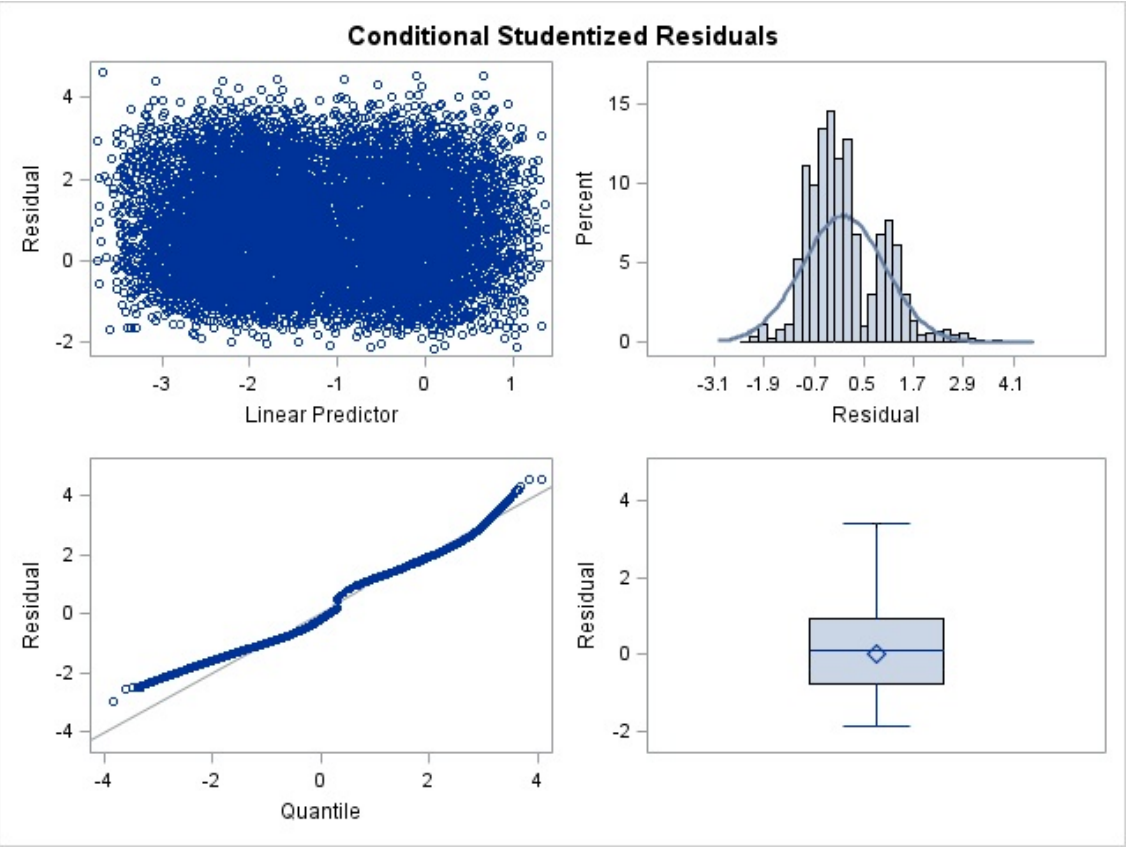
Appendix C

The model diagnostic plots obtained using the Laplace approximation method are given below:

GLMM(1)



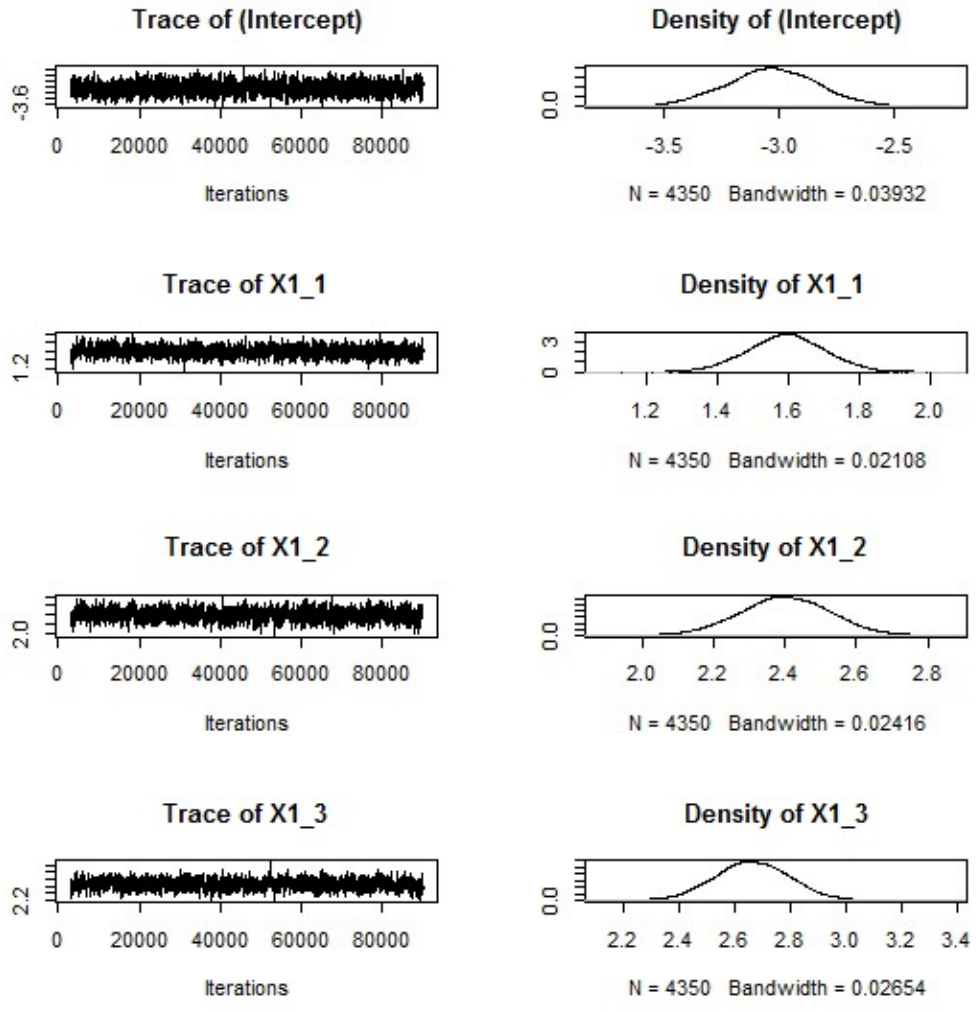
GLMM(2)

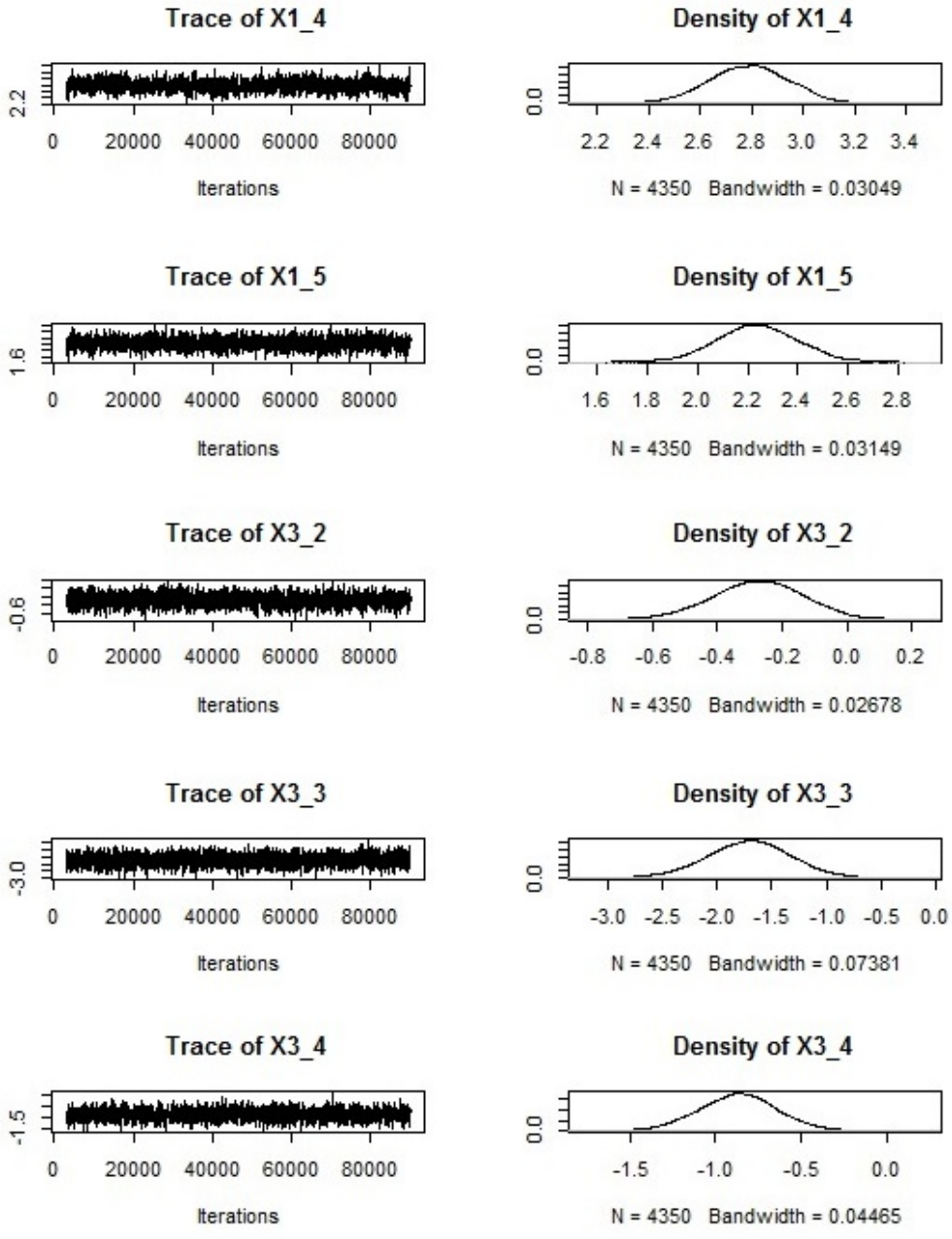

Conditional Studentized Residuals
Residual
Linear Predictor
Percent
Residual
Residual
Quantile
Residual

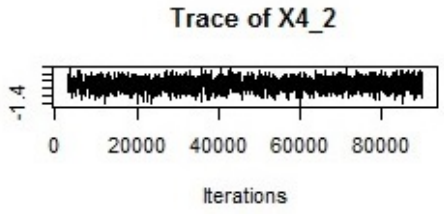
Appendix D

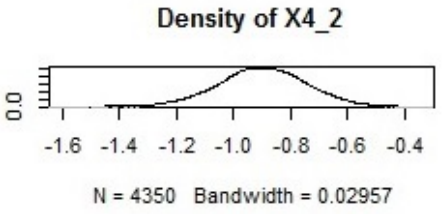
The plots obtained using the Bayesian estimation methods for GLMM(2) are given below:

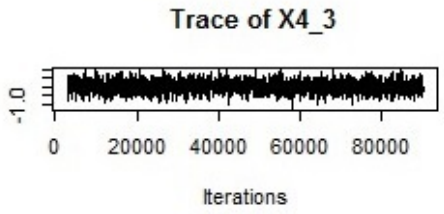
B.1 MCMC method

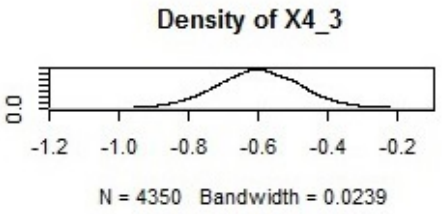


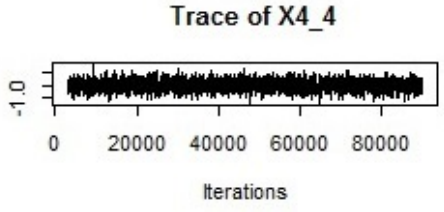

Trace of X1_4
2.2
0
20000
40000
60000
80000
Iterations
Density of X1_4
0.0
2.2
2.4
2.6
2.8
3.0
3.2
3.4
N = 4350 Bandwidth = 0.03049
Trace of X1_5
1.6
0
20000
40000
60000
80000
Iterations
Density of X1_5
0.0
1.6
1.8
2.0
2.2
2.4
2.6
2.8
N = 4350 Bandwidth = 0.03149
Trace of X3_2
-0.6
0
20000
40000
60000
80000
Iterations
Density of X3_2
0.0
-0.8
-0.6
-0.4
-0.2
0.0
0.2
N = 4350 Bandwidth = 0.02678
Trace of X3_3
-3.0
0
20000
40000
60000
80000
Iterations
Density of X3_3
0.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
N = 4350 Bandwidth = 0.07381
Trace of X3_4
-1.5
0
20000
40000
60000
80000
Iterations
Density of X3_4
0.0
-1.5
-1.0
-0.5
0.0
N = 4350 Bandwidth = 0.04465

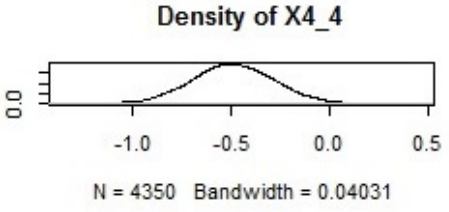

Trace of X4_2
-1.4
0
20000
40000
60000
80000
Iterations

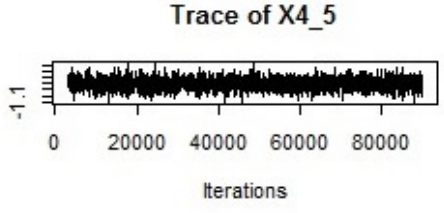

Density of X4_2
0.0
-1.6
-1.4
-1.2
-1.0
-0.8
-0.6
-0.4
N = 4350 Bandwidth = 0.02957

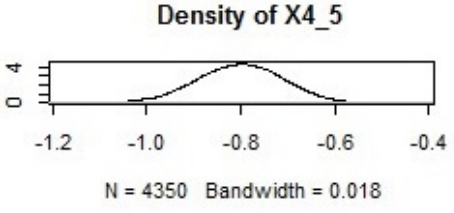

Trace of X4_3
-1.0
0
20000
40000
60000
80000
Iterations

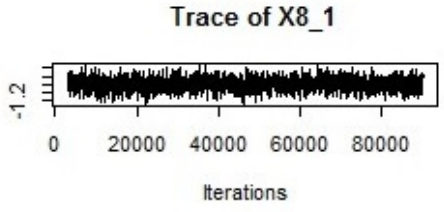

Density of X4_3
0.0
-1.2
-1.0
-0.8
-0.6
-0.4
-0.2
N = 4350 Bandwidth = 0.0239

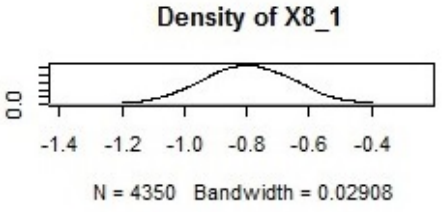

Trace of X4_4
-1.0
0
20000
40000
60000
80000
Iterations

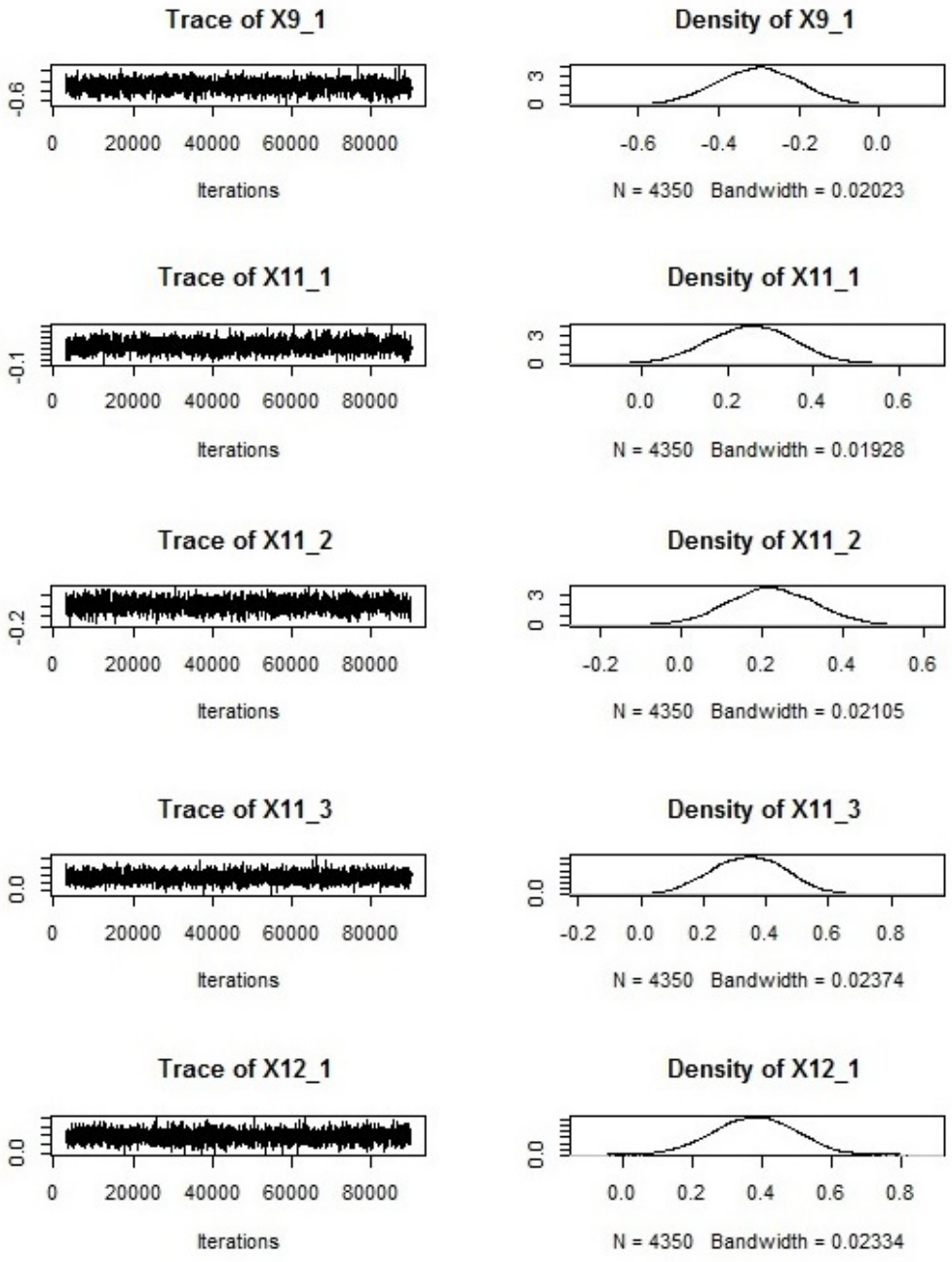

Density of X4_4
0.0
-1.0
-0.5
0.0
0.5
N = 4350 Bandwidth = 0.04031

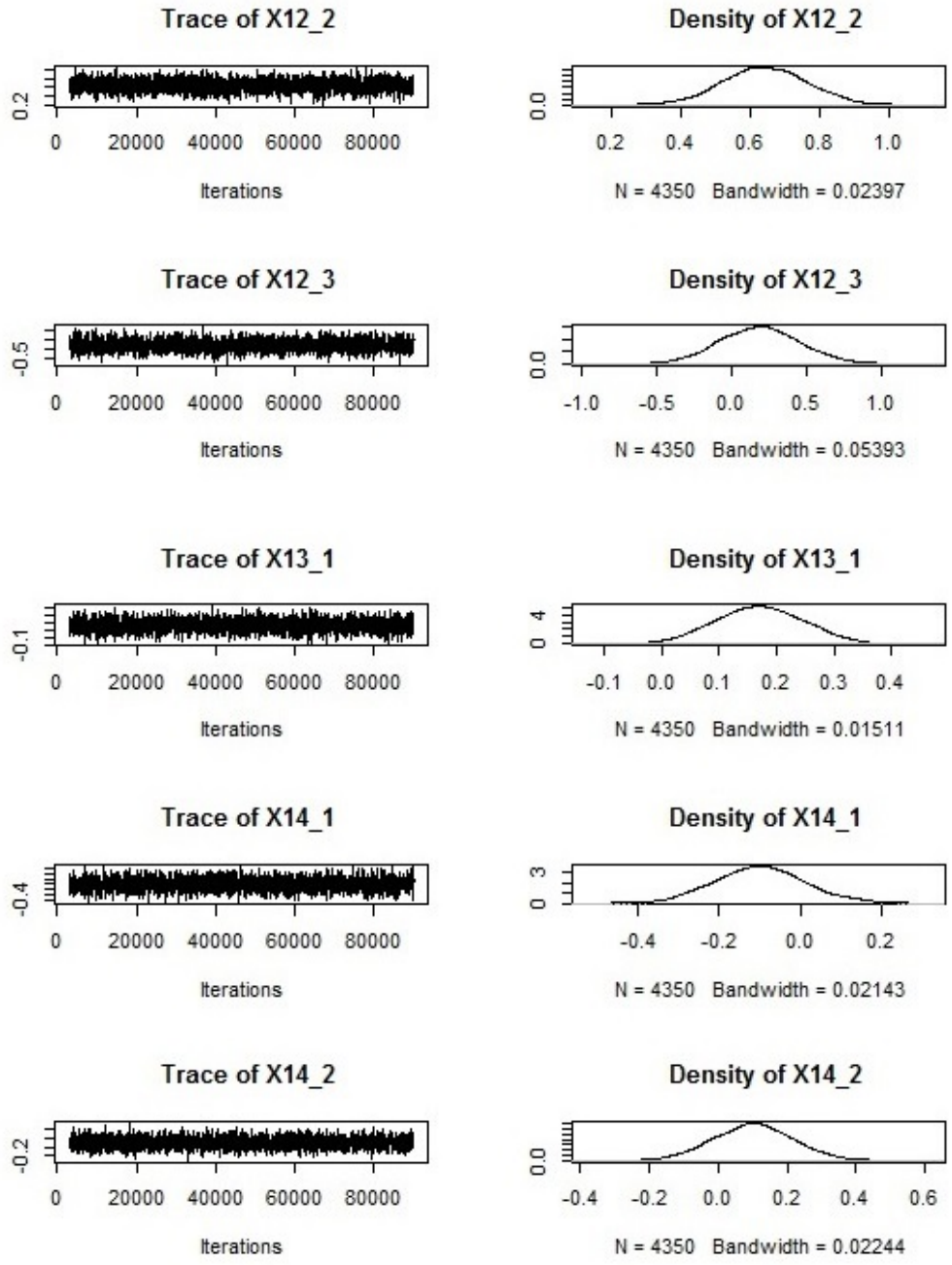

Trace of X4_5
-1.1
0
20000
40000
60000
80000
Iterations

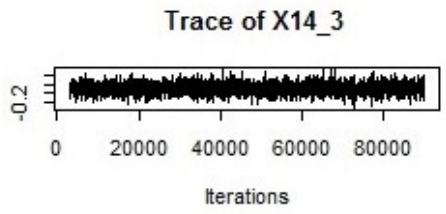

Density of X4_5
0 4
-1.2
-1.0
-0.8
-0.6
-0.4
N = 4350 Bandwidth = 0.018

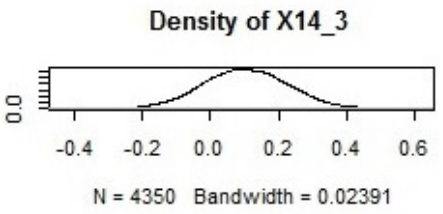

Trace of X8_1
-1.2
0
20000
40000
60000
80000
Iterations

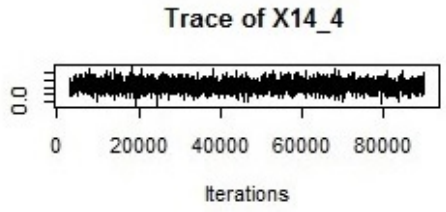

Density of X8_1
0.0
-1.4
-1.2
-1.0
-0.8
-0.6
-0.4
N = 4350 Bandwidth = 0.02908

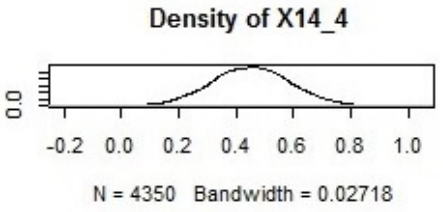

Trace of X9_1
-0.6
0
20000
40000
60000
80000
Iterations
Density of X9_1
0 3
-0.6
-0.4
-0.2
0.0
N = 4350 Bandwidth = 0.02023
Trace of X11_1
-0.1
0
20000
40000
60000
80000
Iterations
Density of X11_1
0 3
0.0
0.2
0.4
0.6
N = 4350 Bandwidth = 0.01928
Trace of X11_2
-0.2
0
20000
40000
60000
80000
Iterations
Density of X11_2
0 3
-0.2
0.0
0.2
0.4
0.6
N = 4350 Bandwidth = 0.02105
Trace of X11_3
0.0
0
20000
40000
60000
80000
Iterations
Density of X11_3
0.0
-0.2
0.0
0.2
0.4
0.6
0.8
N = 4350 Bandwidth = 0.02374
Trace of X12_1
0.0
0
20000
40000
60000
80000
Iterations
Density of X12_1
0.0
0.0
0.2
0.4
0.6
0.8
N = 4350 Bandwidth = 0.02334

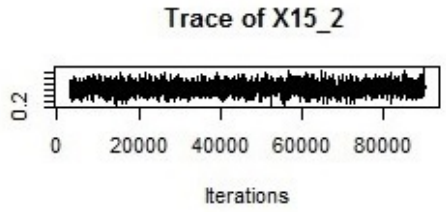

Trace of X12_2
Density of X12_2
Iterations
N = 4350 Bandwidth = 0.02397
Trace of X12_3
Density of X12_3
Iterations
N = 4350 Bandwidth = 0.05393
Trace of X13_1
Density of X13_1
Iterations
N = 4350 Bandwidth = 0.01511
Trace of X14_1
Density of X14_1
Iterations
N = 4350 Bandwidth = 0.02143
Trace of X14_2
Density of X14_2
Iterations
N = 4350 Bandwidth = 0.02244

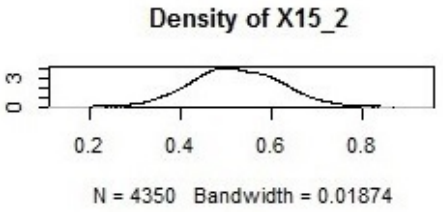

Trace of X14_3
-0.2
0
20000
40000
60000
80000
Iterations

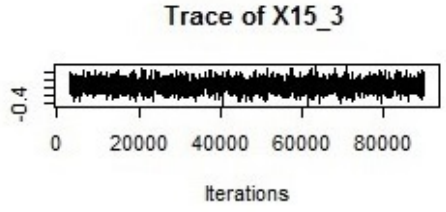

Density of X14_3
0.0
-0.4
-0.2
0.0
0.2
0.4
0.6
N = 4350 Bandwidth = 0.02391

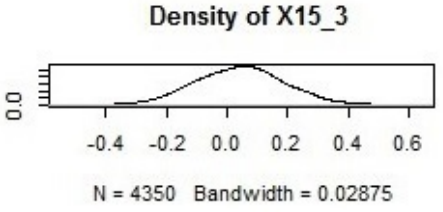

Trace of X14_4
0.0
0
20000
40000
60000
80000
Iterations

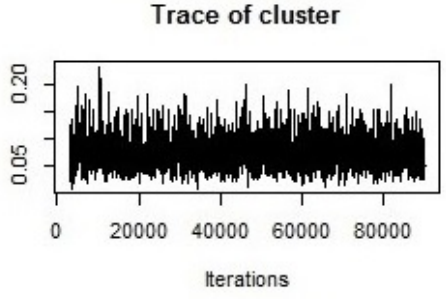

Density of X14_4
0.0
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
N = 4350 Bandwidth = 0.02718

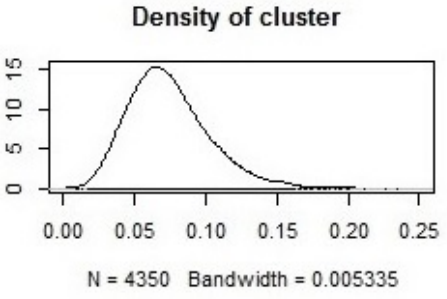

Trace of X15_2
0.2
0
20000
40000
60000
80000
Iterations

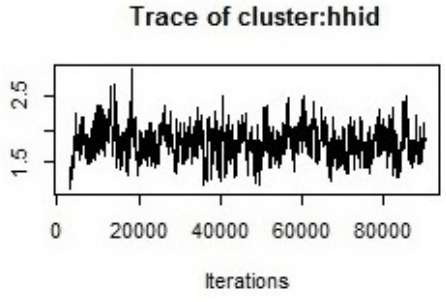

Density of X15_2
0 3
0.2
0.4
0.6
0.8
N = 4350 Bandwidth = 0.01874

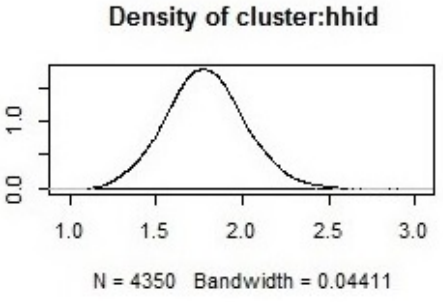

Trace of X15_3
-0.4
0
20000
40000
60000
80000
Iterations


Density of X15_3
0.0
-0.4
-0.2
0.0
0.2
0.4
0.6
N = 4350 Bandwidth = 0.02875

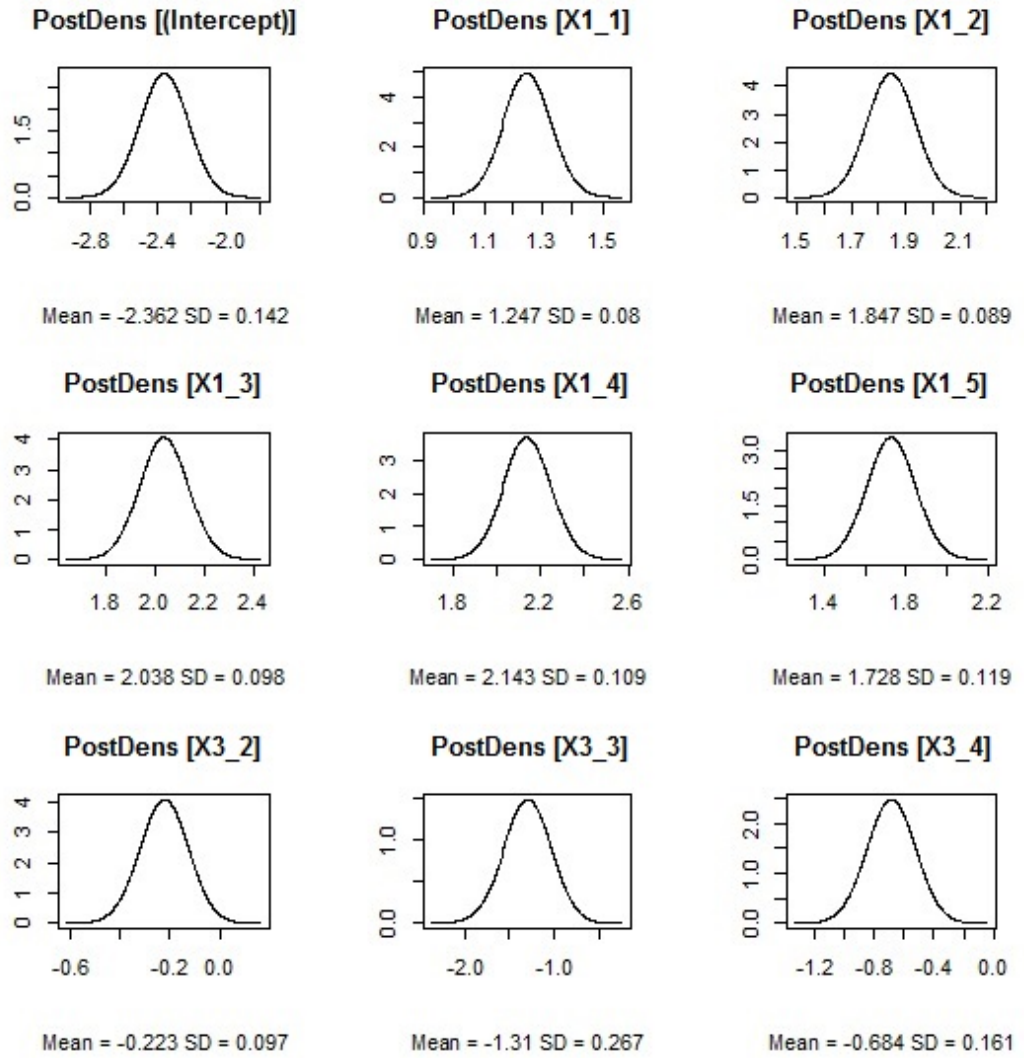

Trace of cluster
0.05
0.20
0
20000
40000
60000
80000
Iterations

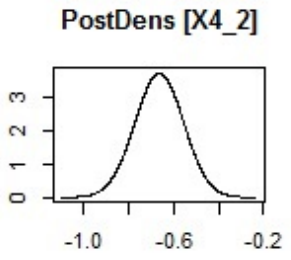

Density of cluster
0
5
10
15
0.00
0.05
0.10
0.15
0.20
0.25
N = 4350 Bandwidth = 0.005335


Trace of cluster:hhid
1.5
2.5
0
20000
40000
60000
80000
Iterations

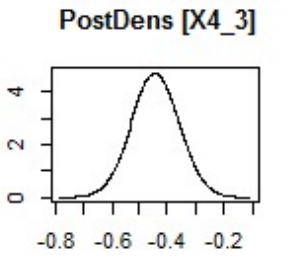

Density of cluster:hhid
0.0
1.0
1.0
1.5
2.0
2.5
3.0
N = 4350 Bandwidth = 0.04411

B.2 INLA method

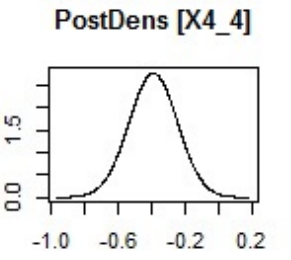




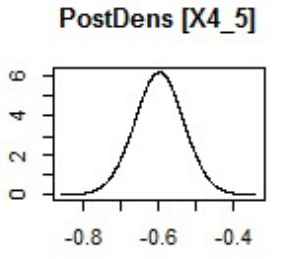
Mean = -0.664 SD = 0.108



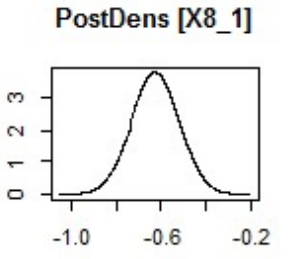
Mean = -0.444 SD = 0.085



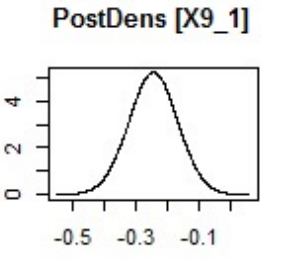
Mean = -0.39 SD = 0.145



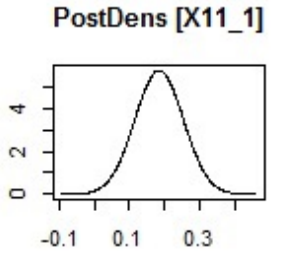
Mean = -0.599 SD = 0.065



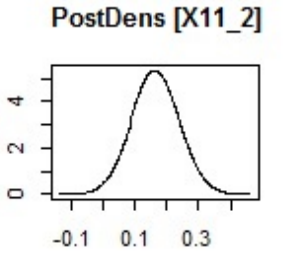
Mean = -0.622 SD = 0.106



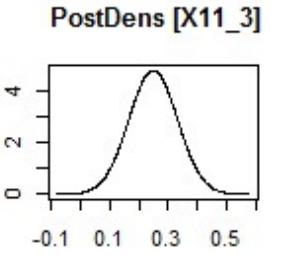
Mean = -0.246 SD = 0.077



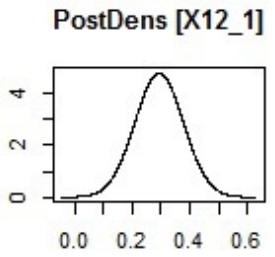
Mean = 0.185 SD = 0.069



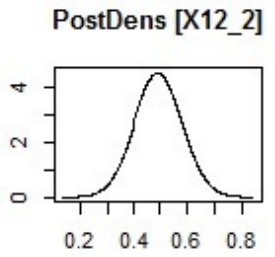
Mean = 0.165 SD = 0.075



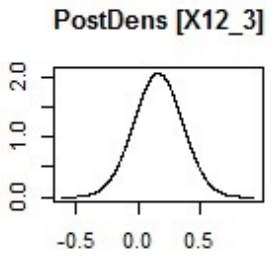
Mean = 0.249 SD = 0.083



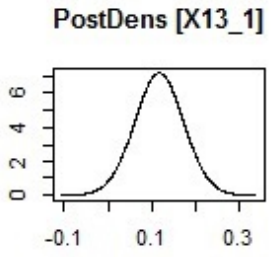
Mean = 0.294 SD = 0.084



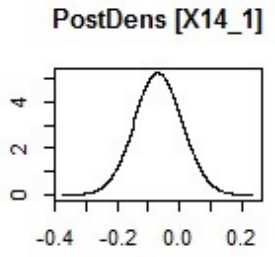
Mean = 0.49 SD = 0.089



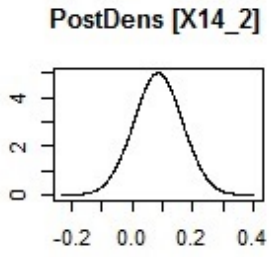
Mean = 0.16 SD = 0.194



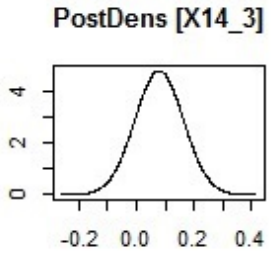
Mean = 0.118 SD = 0.056



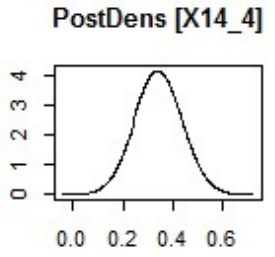
Mean = -0.069 SD = 0.077



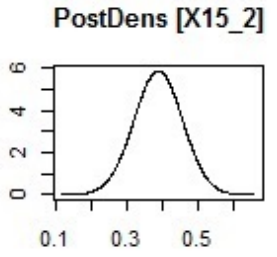
Mean = 0.086 SD = 0.08



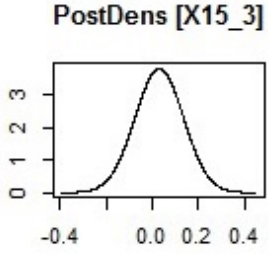
Mean = 0.079 SD = 0.084



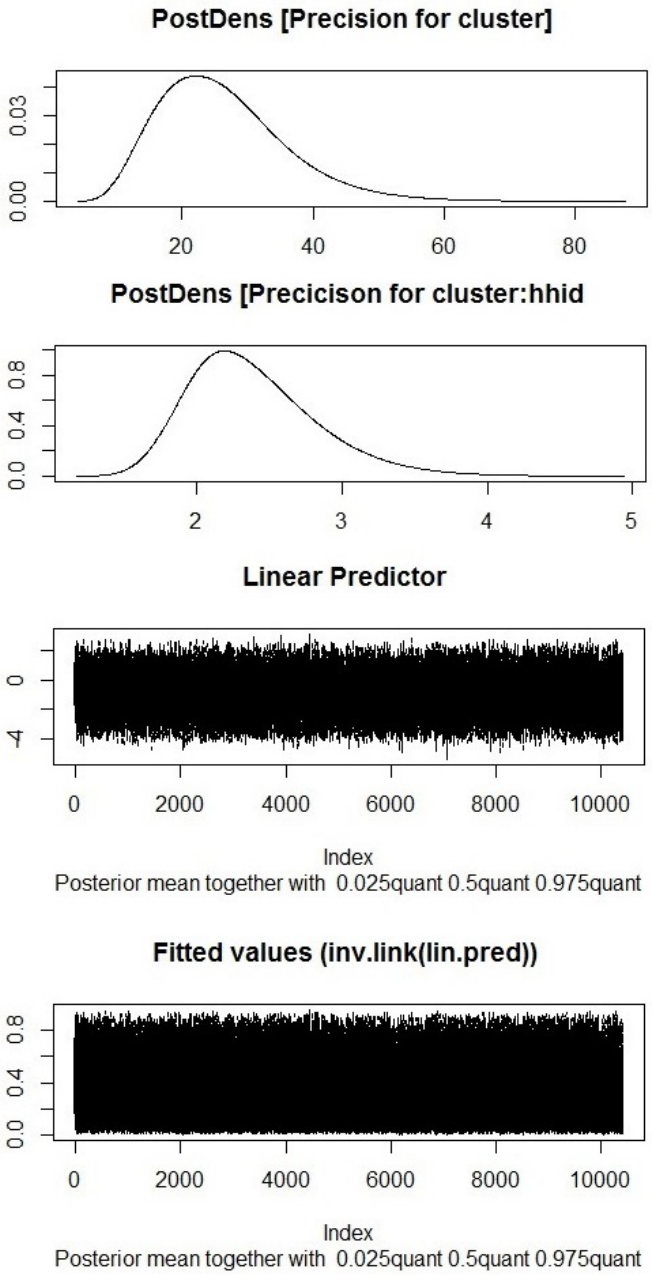
Mean = 0.342 SD = 0.095



Mean = 0.389 SD = 0.068



Mean = 0.03 SD = 0.105


PostDens [Precision for cluster]
0.00
0.03
20
40
60
80
PostDens [Precicison for cluster:hhid
0.0
0.4
0.8
2
3
4
5
Linear Predictor
-4
0
0
2000
4000
6000
8000
10000
Index
Posterior mean together with 0.025quant 0.5quant 0.975quant
Fitted values (inv.link(lin.pred))
0.0
0.4
0.8
0
2000
4000
6000
8000
10000
Index
Posterior mean together with 0.025quant 0.5quant 0.975quant