Modelling depression in South Africa

Tahzeeb Ghoor

August, 2020

Modelling depression in South Africa

by

Tahzeeb Ghoor

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

Thesis Supervisor: Ms. Danielle Roberts

Thesis Co-supervisor: Dr. Siaka Lougue

UNIVERSITY OF KWAZULU-NATAL
SCHOOL OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE
WESTVILLE CAMPUS, DURBAN, SOUTH AFRICA

Declaration - Plagiarism

I, Tahzeeb Ghoor, declare that

1. The research reported in this thesis, except where otherwise indicated, is my original research.
2. This thesis has not been submitted for any degree or examination at any other university.
3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowlegded as being sourced from other persons.
4. This thesis 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 thesis does not contain text, graphics or tables copied and pasted from the internet, unless specifically acknowledged, and the source being detailed in the thesis and in the reference sections.

Tahzeeb Ghoor (Student)	Date
Ms. Danielle Roberts (Supervisor)	Date
Dr. Siaka Lougue (Co-supervisor)	Date

Disclaimer

This document describes work undertaken as a Masters 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

Depression is considered to be the leading cause of disability worldwide, with approximately 350 million individuals, of all ages, affected. The mental disorder is predominant in females and poverty is associated with an increased prevalence. The 12-month prevalence in South Africa is approximately 16.5%, with a lifetime prevalence of common mental disorders among adults of 38% (World Health Organization (WHO), 2017). In order to assist individuals in dealing with depression, it is important for such individuals to be identified at an early stage in order to provide them with the necessary support before their depression becomes unmanageable, thus putting them at risk for self-inflicted harm.

The objective of this study was to investigate the prevalence and risk determinants of depression among South African individuals between the ages 15 to 49 years old and to determine which factors contribute the most to this mental illness. This study made use of data from the 2016 South African General Household Survey which was carried out using a multistage cluster sampling technique. The sample was not spread geographically in proportion to the population, but rather equally across the enumeration areas. The response variable of interest was binary, indicating whether an individual considered himself/herself depressed or not. Three statistical approaches were applied. The first was the survey logistic regression model which is a design-based approach. In this approach, parameter estimates and inferences were based on the sampling weights, and only inferences concerning the effects of certain covariates on the response variable were of interest. The second was a generalized linear mixed model which is a model-based approach. In this approach, interest was also on estimating and accounting for the proportion of variation in the response variable that was attributable to each of the multiple levels of sampling. This approach also accounted for possible correlations in the data where individuals in the same household or cluster tend to be more alike than those from other households or clusters. Lastly, a Bayesian network was applied to model the conditional dependence among the variables. This approach is a type of probabilistic graphical model that uses Bayesian inference for calculations of the probabilities.

The results indicated that substance abuse, the person's perceived health status and gender were significantly associated with depression. Each of the three techniques were then used to classify the depression status of the individuals, and their performances in this classification were compared. The purpose of being able to classify an individual's depression status, based on their individual and household factors, is to be able to identify a depressed individual in order to target them for intervention. The generalized linear mixed model proved to be the better performing technique in terms of classification. Thus, we recommend that when using data based on a complex survey design, this technique is considered in classifying the occurrence of an event of interest.

Acknowledgements

I would first like to thank my supervisor Ms. Danielle Roberts, for her continuous support, guidance, advice and encouragement throughout this thesis. I would also like to extend a thanks to my co-supervisor Mr. Siaka Lougue for introducing me to Bayesian Networks. Finally, I would like to thank the staff members of the Department of Statistics at the University of KwaZulu-Natal for their moral and academic support throughout the years.

Many thanks to Statistics South Africa for allowing me access to the General Household Survey (GHS) data.

I am truly grateful to family and friends who have supported me throughout the years of me doing my MSc.

Contents

List of Figures

List of Tables

Abbreviations

AIC	Alkaike's Information Criterion
BAN	Bayesian Network-augmented Naive
BRR	Balanced Repeated Replication
BN	Bayesian Network
CIA	Central Intelligence Agency
CS	Compound Symmetry
DAG	Directed Acyclic Graph
EA	Enumeration Area
GHS	General Household Survey
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
HIV	Human Immunodeficiency Virus
IRWLS	Iteratively Reweighted Least Squares
JRR	Jackknife Repeated Replication
LM	General Linear Model
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimation
MGL	Marginal Quasi-Likelihood
OLR	Ordinary Logistic Regression
OLS	Ordinary Least Squares
OR	Odds Ratios
NIH	National Institute of Mental Health
PDF	Probability Distribution Function
PML	Pseudo-Maximum Likelihood
PQL	Penalized Quasi-Likelihood
PSU	Primary Sampling Unit
SC	Schwarz Criterion
SRL	Survey Logistic Regression
TAN	Tree-augmented Naive
VC	Variance Components
QIC	Quasi-Likelihood Under the Independence Model Criterion
QL	Quasi-Likelihood
UN	Unstructured
WHO	World Health Organization

Chapter 1

Introduction

Depression is a common mental disorder, which is characterized by sadness, loss of interest, feelings of low self-worth or guilt and feelings of tiredness and poor concentration. These problems can become chronic or recurrent and lead to substantial impairments in an individual's ability to take care of his or her everyday responsibilities (World Health Organization (WHO), 2017). Depression is considered to be the leading cause of disability worldwide, with approximately 350 million individuals, of all ages, affected. The mental disorder is predominant in females (WHO, 2012).

The burden of depression is carried among the poorest individuals of South Africa. Mental health care is under-funded compared to other health priorities in the country (Pandey et al., 2017). The 12-month prevalence in South Africa is approximately 16.5%, with a lifetime prevalence of common mental disorders among adults of 38% (Bateman, 2014). Furthermore, South Africa has the eighth highest rate of suicide in the world, with approximately 8 000 people committing suicide each year. Close to two thirds of all suicide victims are between the ages of 20 and 39, with approximately 4.6 male suicides for every female suicide (Malan, 2014).

Counselling is extremely useful in assisting individuals with depression. Traditional African medicine has a great effect on the counselling process. Although there are substantial differences between African medicine and Westernized conceptualizations of depression, they both value the emotional psychosocial contexts (Starkowitz, 2013). However, many studies have shown that there are barriers to seeking help and treatments for individuals with mental disorders. Such barriers include the stigma around mental disorders, low levels of knowledge of the disease and its treatment, and financial difficulties (Nglazi et al., 2016).

Mental disorders such as depression cause a significant economic burden on a coun-

try at both the individual and societal level. This is due to an individual who suffers from mental disorders having a reduced quality of life, disrupted work and life roles, and increased morbidity and mortality (Nglazi et al., 2016).

1.1 Literature Review

Genetic factors play an important role in the development of depression. There have been studies that indicate that if one identical twin is depressed, there is approximately a 76% chance that the other will develop clinical depression. However, if the identical twins were raised apart, there is a 67% chance of the second twin developing depression. This is a higher rate of depression when compared to the the general public prevalence (All-About-Depression.com, 2017). Depression is a familial disorder, however, an individual's environment also plays a significant part in their depression status. Major depression is a complex disorder that does not result from either genetic or environmental influences alone, but rather from a combination of the two (Sullivan et al., 2000).

Depression can seriously affect individuals with chronic illnesses such as cancer, thyroid deficiencies, and the Human Immunodeficieny Virus (HIV), among others. Some of these diagnoses can be life changing and can cause severe psychological and emotional stress (Smith, 2015). There has been evidence of a strong association between HIV and depression (Galambos et al., 2004).

Looking at the social risk factors for depression, first consider abuse. There is a strong relationship between abuse and depression. Children who are abused or neglected have a high risk of suffering from depression and/or other mental illnesses as they mature. Sexual abuse does not appear to affect individuals as much as physical or any other type of abuse, however, adults with a history of childhood sexual abuse have a higher prevalence than those who have not experienced such trauma (National Institue of Mental Health (NIH), 2007). The lack of social support can also cause depression. If an individual has prolonged social isolation and has only a few, if any, supportive relationships, this can be a source of depression. Feelings of exclusion can bring an episode in people who are prone to mood disorders (Healthline, 2017).

Major life events affect an individual's depression status. Events such as being unemployed, getting a divorce, a loss of a loved one or other stressful situations can increase an individual's risk of depression. A deep sadness can occur, where some individuals will feel better after a few months and others experience more serious,

long term periods of depression. An individual is likely to suffer from depression if he/she has grieving symptoms that occur for more than two months (Healthline, 2017). Depression is a mental illness that frequently co-occurs with substance use. Individuals who abuse substances are more likely to suffer from depression and vice versa, i.e. individuals may consume alcohol or drugs to relieve feelings of sadness. However, people fail to realize that substances such as alcohol can increase feelings of sadness and fatigue (Smith, 2017).

Age and gender play important roles in depression. Depression is twice as common in adolescent and adult females than in males. Depression disorders equally affect males and females in pre-pubertal children. Epidemiological data suggests that there is a higher depression prevalence among females than males, however some believe that both genders are affected equally, although women are more likely to be diagnosed with the disorder (Piccinelli & Wilkinson, 2000).

The triggers for depression among different genders are different. Females often show internalizing symptoms whereas as males show externalizing symptoms. For example, women display more sensitivity to interpersonal relationships and men display more sensitivity to external career and goal orientated factors. Women also experience illnesses such as premenstrual dysphoric disorder, postpartum depression and postmenopausal depression and anxiety, that are associated with changes in ovarian hormones and could contribute to the increased prevalence of depression in women (Albert, 2015).

All ethnic and cultural groups are exposed to depression. The depression rate across different groups are consistent, however, ethnic and cultural differences often impact the way in which members express their feelings and their willingness to seek treatment (Haggerty, 2016). Poverty is associated with increased prevalence of mental health disorders. Brown & Moran (1997) showed how financial hardship was related to the risk of having a chronic episode of depression, lasting at least a year. A study on depression in the United States, which made use of Household Population survey data from 2009-2012, showed that more than 15% of people living below the federal poverty level had depression compared with 6.2% of persons living at or above the poverty level (Pratt, 2014).

Very few studies on the prevalence and risk factors of depression in the South African population have been carried out recently. Data for these studies were obtained from the Composite International Diagnostic Interview with logistic regression being the primary method of analysis (Herman et al., 2009; Nglazi et al., 2016; Tomlinson et al.,

2009). Since there are only a few studies available, this may be attributed to the the higher burden of other communicable and non-communicable diseases than receive more attention, such as HIV, hypertension and diabetes. However, depression is often a co-morbidity of these diseases, and thus should not be taken lightly (Nglazi et al., 2016).

1.2 Aims and Objectives of the Study

This study aimed at using various statistical methods to analyze data on depression that was collected based on a complex survey design. The specific objectives of the study are:

- To investigate the risk determinants of depression among South African individuals aged 15 to 49 years old and to determine which factors contribute the most to this mental illness.
- To explore various statistical methods of classifying a person's depression status.

1.3 Thesis Structure

This chapter provided an introduction to the topic, as well as a review of literature on depression and associated risk factors. In addition, the chapter provded the objectives of the study. Chapter 2 introduces the data, the variables of interest and some descriptive analysis of the data. Furthermore, Chapter 2 discusses the statistical methods appropriate for the analysis of the data, therefore presenting a brief overview of the methods considered in this thesis. Chapter 3 introduces the generalized linear model and the survey logistic regression model which presents the first statistical approach applied, the results of which are presented at the end of the chapter. Chapter 4 covers a brief overview of the generalized linear mixed model which is the second statistical approach applied. The results of this second approach are presented at the end of Chapter 4. The last statistical approach applied, Bayesian Networks, is discussed in Chapter 5, where the results are presented at the end of the chapter. The final chapter, Chapter 6, provides a comparison of the performance of the three approaches in classifying an individual's diabetes status. In addition, this chapter discusses the limitations of the thesis, the conclusion and the future direction of this study.

Chapter 2

Data Description and Exploration

This chapter discusses the area of study, the data set and the variables of the study. Exploratory analyses is also done to enable the individual to get a general understanding of the data.

2.1 Study Area

The Republic of South Africa is at the southern tip of Africa. It shares a boarder with Botswana, Lesotho, Mozambique, Namibia, Swaziland and Zimbabwe and has a coastline of 2,798 km. The country has an area of 1,219,090 square kilometres with approximately 4,620 square kilometres being taken up by water and 1,214,470 is land (CIA World Factbook, 2017). South Africa is divided up into 9 provinces, namely, Limpopo, North West, Gauteng, Mpumalanga, Northern Cape, Free State, KwaZulu Natal, Western Cape and Eastern Cape, as shown in Figure 2.1 on the next page. There is approximately 55 520 396 people living in South Africa with slightly more females than males. The population growth rate was estimated to be 0.99% in 2017. The South African birth rate is close to 20.5 births per 1000 individuals and the death rate is approximately 9.6 deaths per 1000 individuals (Statistics South Africa, 2011).

2.2 The Data Set

This study made use of data from the 2016 South African General Household Survey (SAGHS). The SAGHS is an annual survey which is aimed at determining the progress of development in South Africa. The performance of programmes and the quality of services are consistently measured across various sectors in the country.

Figure 2.1: Map of South Africa

2.2.1 Sampling Procedure and Data Collection

The SAGHS was nationally represented and made use of a stratified multi-stage cluster sampling design to carry out the survey. South Africa was stratified into its 9 provinces which were then subdivided into 62 districts. These districts were further divided into 233 enumeration areas (EAs). The sample was not spread geographically in proportion to the population, but rather equally across the EAs.

The survey consisted of a two-stage sample design where the first stage involved selecting clusters from a list of EAs covered in the 2011 Population Census. These areas made up the primary sampling units (PSUs) or clusters. A total of 3,324 clusters with probability proportional to size were collected. A list of all households in the 3,324 clusters was drawn up and provided the sampling frame from which the dwelling units were selected for the survey. The second stage of the selection process involved systematic sampling of the dwelling units from the list of households in each cluster. The final sample then consisted of 21,228 households.

Data was collected from all the individuals in the selected households using the GHS

questionnaire. This questionnaire was designed to collect demographic, biographical and health information, among others, from each household member, as well as information on the household's dwelling unit, such as household assets, access to electricity, and type of water sources.

2.3 Study Variables

The response variable of interest is the self-reported depression status of the sampled individuals between the ages of 15 to 49 years. This is a binary response, indicating whether or not the individual believed they were depressed.

The independent variables considered comprise of individual and household related factors as shown by Figure 2.2. These variables are based on the literature as well as the availability of information in the SAGHS data.

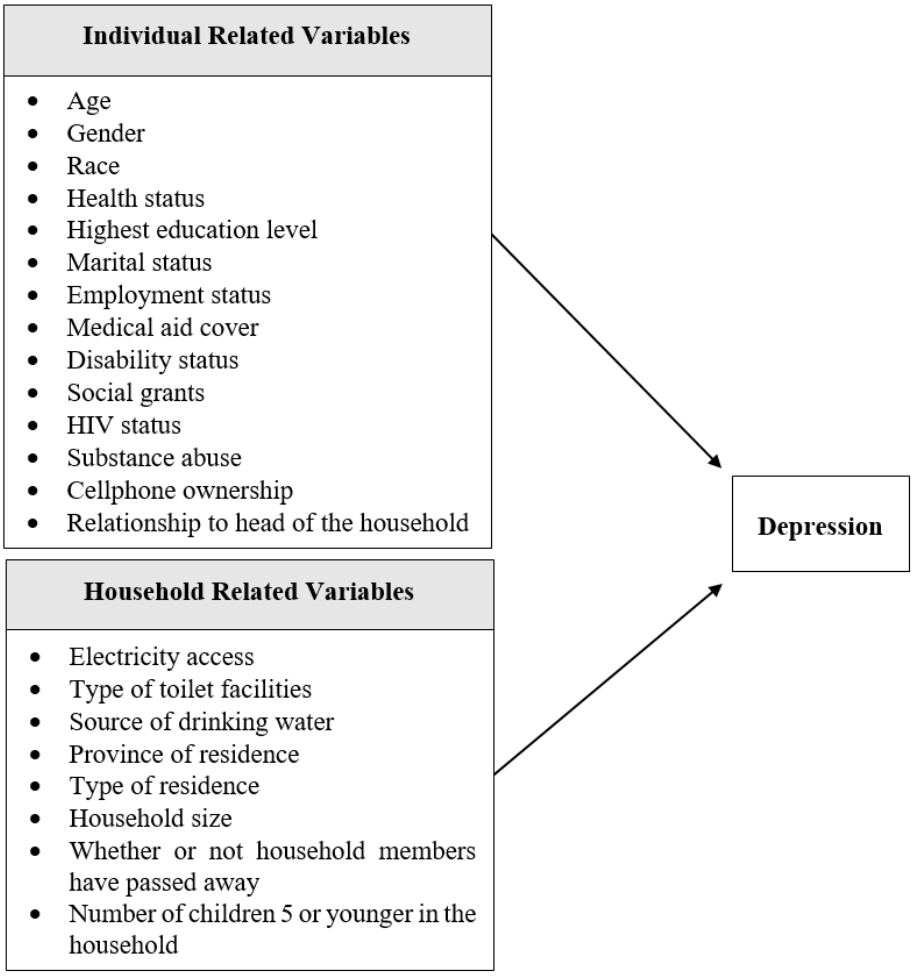


Figure 2.2: Potential factors associated with depression

2.4 Data Exploration

Before any statistical modelling of the data is done, it is ideal to first carry out some exploratory analyses, which is presented in this section. This enables an individual to get a general understanding of the data. In addition, this section provides more detail regarding the explanatory variables.

The final data set used in this study comprised of 29840 individuals from 15073 households. The total number of individuals who reported to suffer from diabetes was 100, resulting in an observed prevalence of 0.3%. It is acknowledged that this observed prevalence is significantly low, particularly compared to the national 12-month prevalence of 16.5%. This is most likely due to it being based on self-reporting, where individuals are concerned about the stigma concerning depression or any mental health issues, thus leading to under-reporting of depression in the survey.

Tables 2.1 and 2.2 present the distribution of the data set according to the individual and household explanatory variables of interest, respectively. The majority of the sample was made up of individuals from the Black race group. Most of the individuals perceived their health to be good, however this was based on self-reporting. The majority of the sampled individuals had a secondary school level of education and most of the individuals were employed at the time of the survey. However, a fair amount of the sampled individuals were not economically active during the time of the survey (40.7% of the sample). This refers to those who are retired, disabled or working as a 'housewife'. Only 14.7% of the sampled individuals were on a private medical aid scheme. Only 3.7% of the sample reported to be HIV positive, which is also based on self-reporting. Once again, this is significantly lower than the current prevalence of HIV in South Africans between the ages of 15 to 49 years old, at 19% (Avert, 2019). In addition, only 0.2% of the sampled reported to suffer from substance abuse, which is based on alcohol or drug use. The majority of the sampled individuals were single and owned a cellphone.

Table 2.1: Distribution of the sample according to the individual level variables

Variable		%
Age group	15-19 years	17.7
	20-29 years	33.1
	30-39 years	28.2
	40-49 years	21.0
Gender	Male	47.9
	Female	52.1
Race	African/Black	82.9
	Coloured	10
	Indian/Asian	1.9
	White	5.2
Health status	Excellent	35.0
	Very Good	21.1
	Good	37.8
	Fair	4.8
	Poor	1.4
Highest education level	No schooling	1.4
	Primary School	11.2
	Secondary School	75.2
	Higher Education	12.2
Employment Status	Employed	43.7
	Unemployed	15.7
	Not Economically Active	40.7
Medical aid scheme	Yes	14.7
	No	85.3
Disability	Not disabled	94.0
	Disabled	6.0
Social grants	Yes	8.5
	No	91.5
Marital status	Married	21.3
	Partnered	11.1
	Divorced/widowed/separated	2.9
	Single	64.7
HIV/AIDS	Yes	3.7
	No	96.3
Substance abuse	Yes	0.2
	No	99.8
Cellphone ownership	Yes	88.2
	No	11.8
Relationship to household head	Head	32.3
	Spouse/partner	13.8
	Other	53.9

It can be seen from Table 2.2 below that the majority of the households had flush toilet facilities, access to electricity and protected water sources for drinking water, which comprised of protected wells (private and public), boreholes and protected springs. Unprotected water sources included open wells (private and public), unprotected springs, rainwater and surface water (rivers/streams, ponds/lakes and dams). Individuals from households with unimproved toilet facilities, which included pit latrines without ventilations, comprised of 17.1% of the sample. Most of the sampled households were from urban areas of residence. A total of 3.6% of the sampled individuals had lost a household member.

Table 2.2: Distribution of the sample according to the household level variables

Variable		%
Type of toilet facility	Flush Toilet	59.6
	VIP Latrine	20.0
	Unimproved Facility	17.1
	Other	3.3
Access to electricity	Yes	93.8
	No	6.2
Source of drinking water	Tap into household	43.9
	Protected water	50.7
	Unprotected water	5.4
Province	Western Cape	10.9
	Eastern Cape	13.3
	Northern Cape	4.5
	Free State	5.4
	KwaZulu-Natal	17.8
	North West	6.6
	Gauteng	22.3
	Mpumalanga	8.2
	Limpopo	11.0
Type of residence	Urban	64.7
	Traditional	31.5
	Farms	3.8
Household members passed away	Yes	3.4
	No	96.6

Figure 2.3 displays the observed prevalence of depression according to the province of residence. The North West Province had the highest prevalence of depression at 0.61% and KwaZulu-Natal had the lowest at 0.06%.

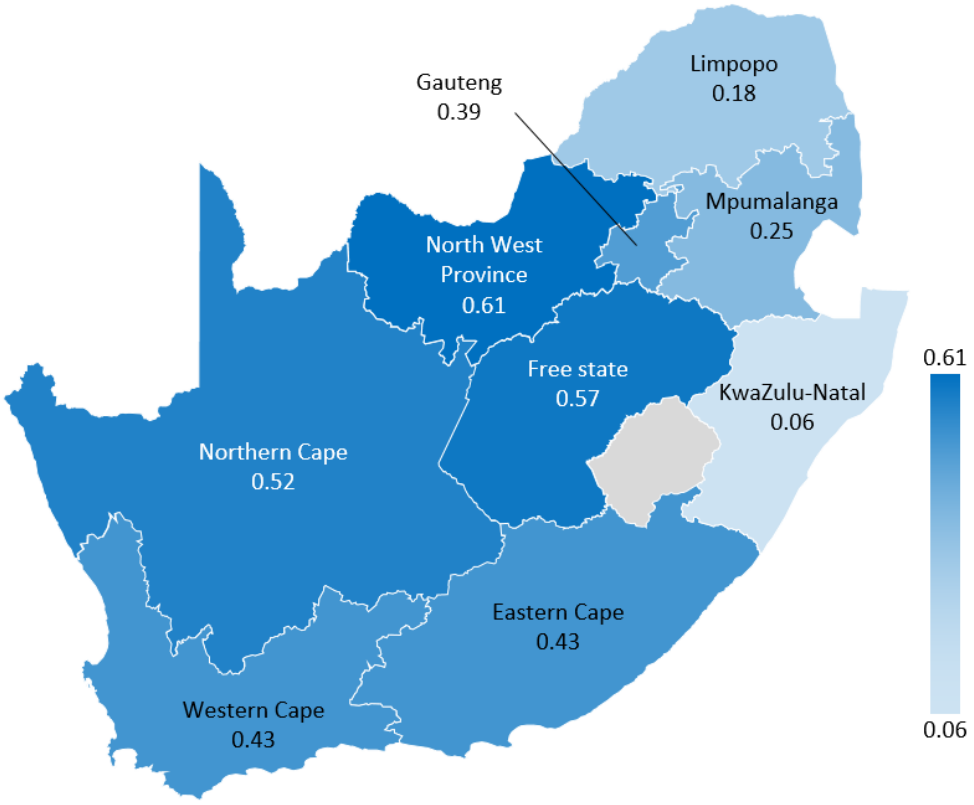


Figure 2.3: Observed prevalence of depression according to province in South Africa

Figure 2.3 displays the observed prevalence of depression according disability status, substance abuse, HIV status and health status. In general, those who suffered from health related issues, such as HIV and disability, had a higher observed prevalence of depression at 1.00% and 1.73%, respectively, compared to those who reported not to be suffering from these issues. The prevalence of depression decreased with an increase in an individual's perception of their health. Individuals abusing alcohol or drugs had a substantially higher prevalence of depression (17.86%) compared to those who reported not to be abusing these substances (0.3%). The treemap in Figure 2.5 on the next page presents the observed prevalence of depression according to the demographic variables race and gender. The White race group had a prevalence of 0.77%, which is more than the that of the other race groups combined. It should be noted that the Indian/Asian race group is not represented in this figure, this is due to none of the individuals in this race group having reported to suffer from depression. The prevalence of depression for females (0.48%) was almost three

times that of males (0.18%).

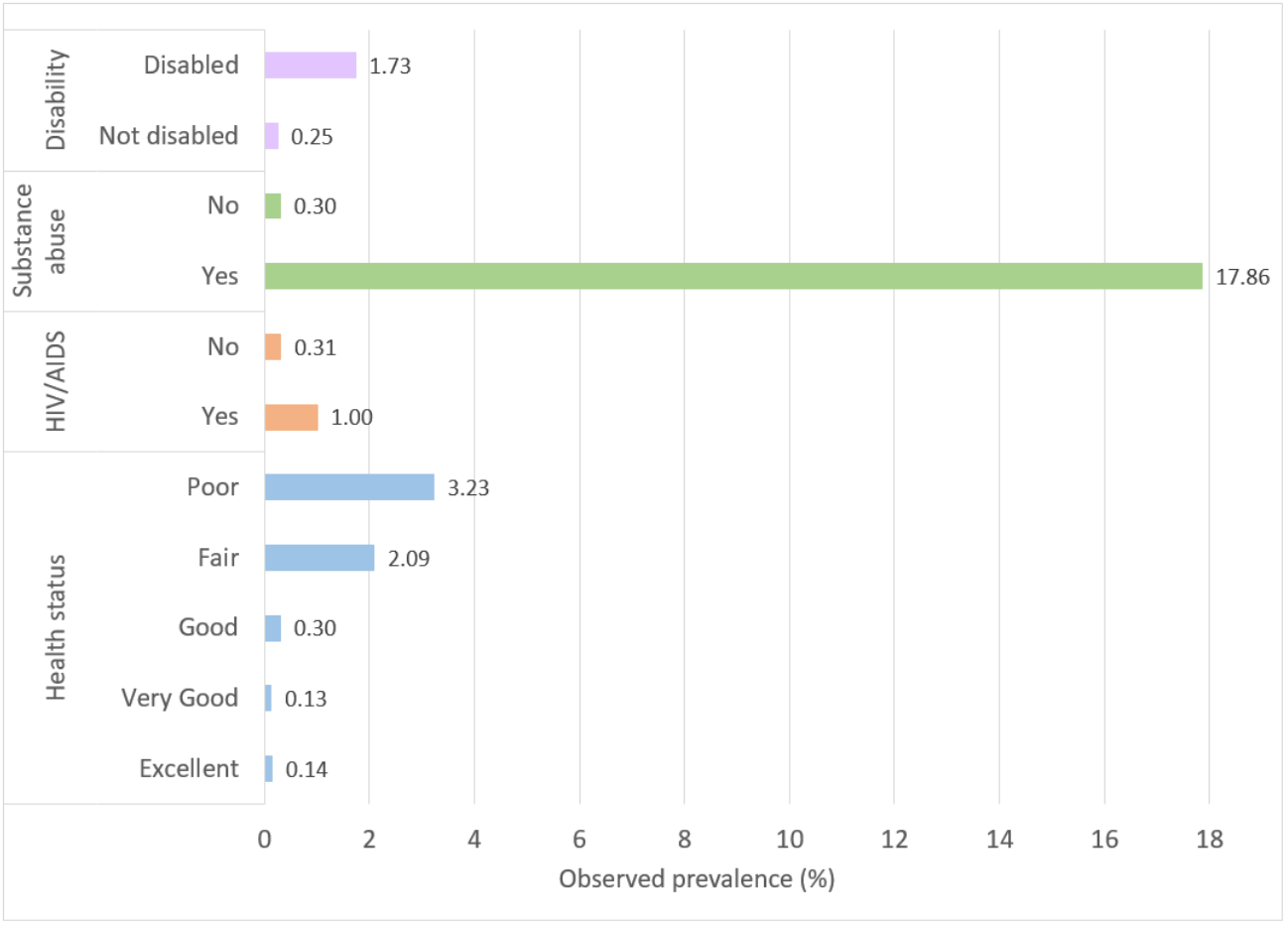


Figure 2.4: Observed prevalence of depression according disability status, substance abuse, HIV status and health status

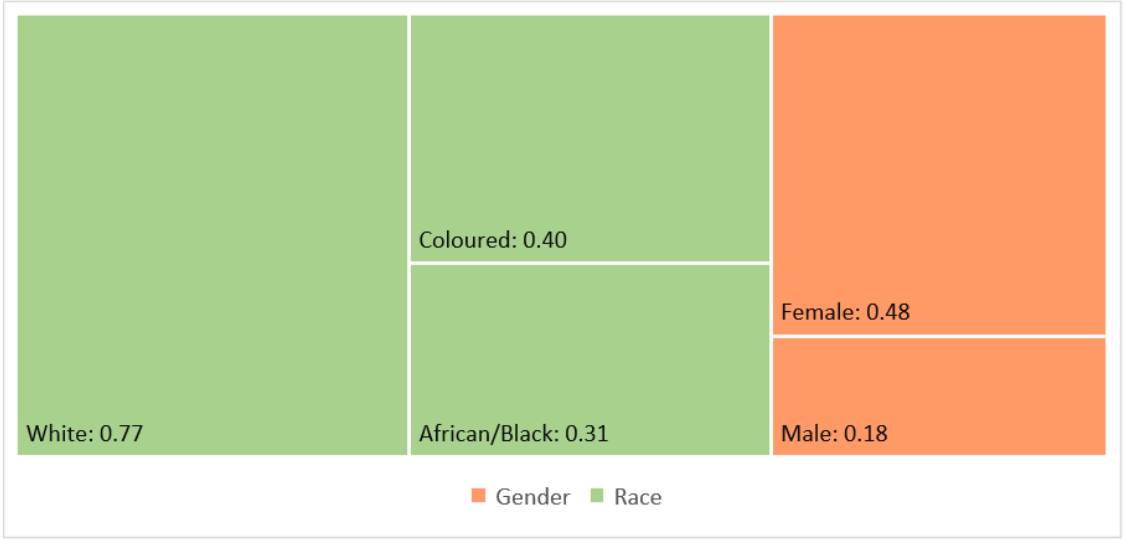


Figure 2.5: Observed prevalence of depression according race and gender

The parallel plot in Figure 2.6 below displays some descriptive statistics for the continuous covariates, age, the number of members in a household and the number of children five or younger in the household, according to depression status. Both depressed and not depressed individuals had the same minimum age of 15, minimum number of household members at 1 and minimum number of younger children in the household at 0. However, the average and maximum of these covariates differed for each group of individuals. The average age was higher for those who were depressed, however, the average household size and average number of young children in the household was slightly lower for those who were depressed compared to those who were not depressed.

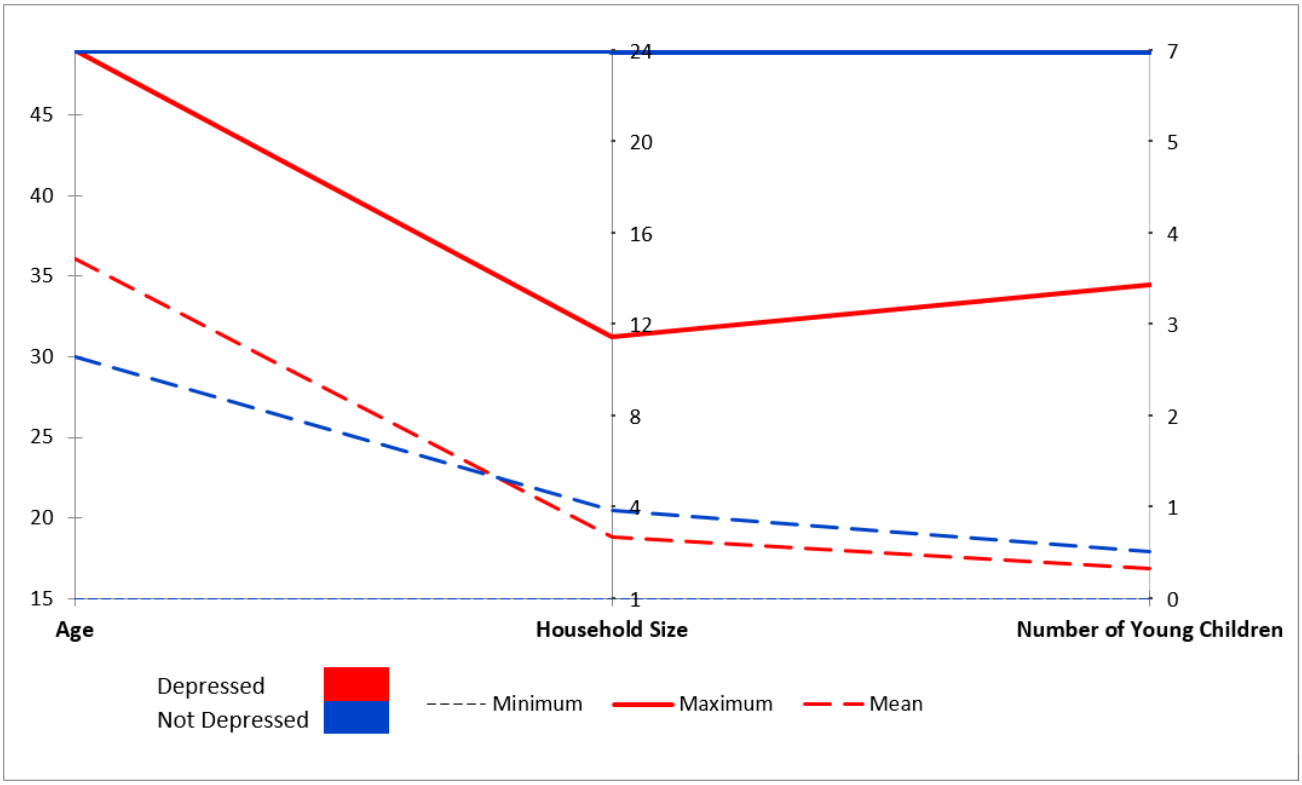


Figure 2.6: Descriptive statistics for the continuous covariates according to depression status

The observed prevalence of depression according the cellphone ownership, whether or not the individual was on a social grant, whether or not the individual was on a medical aid scheme and employment status is presented in Figure 2.7. A higher prevalence of depression was observed for individuals not on social grants (0.34%) and for individuals on a medical aid scheme (0.39%), as well as for those that own cellphones (0.35%). However, individuals not employed had the highest prevalence of depression at 0.41% compared to those employed at 0.33%.

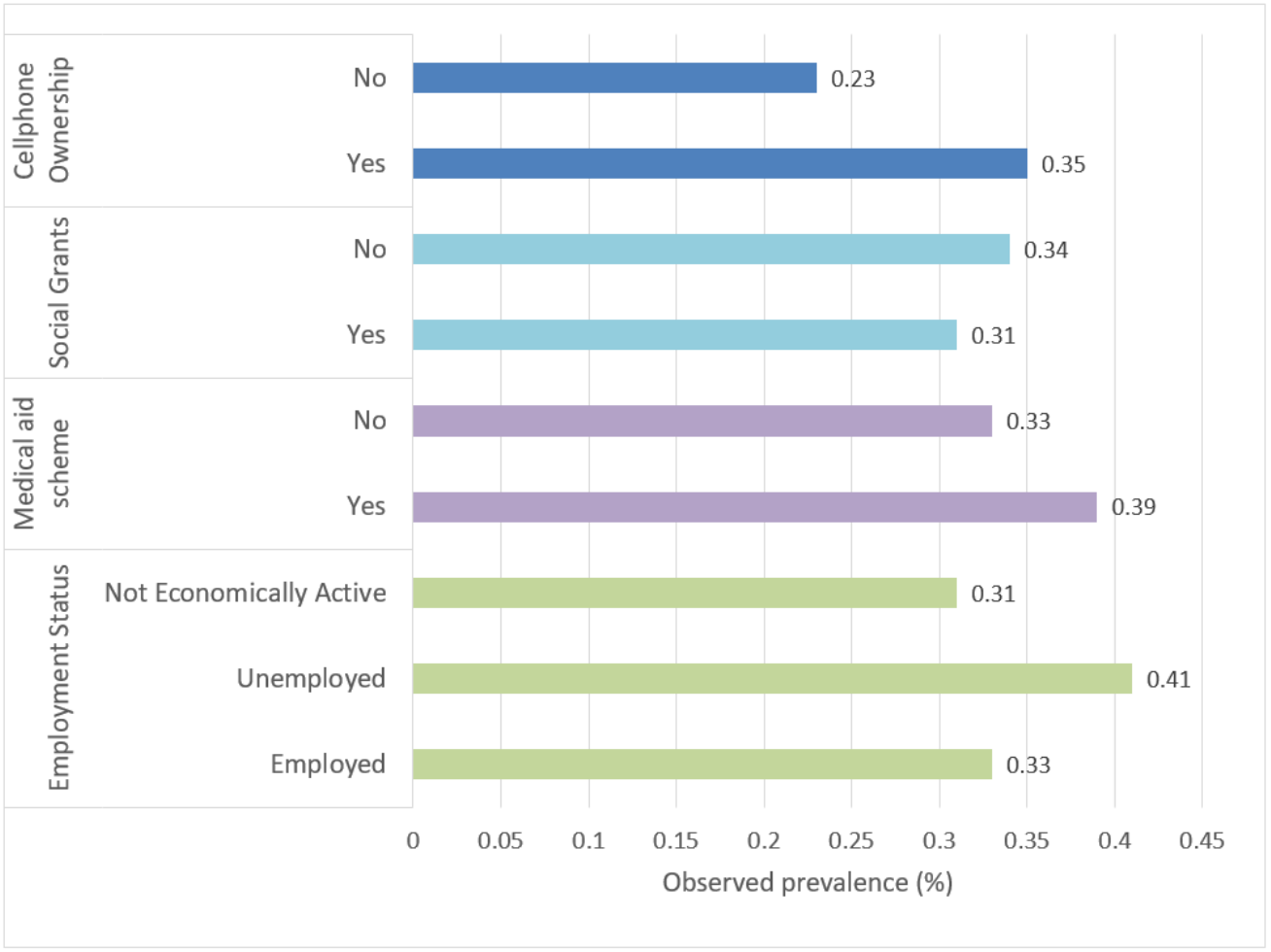


Figure 2.7: Observed prevalence of depression according to cellphone ownership, social grants, medical aid scheme and employment status

Figure 2.8 on the next page displays the observed prevalence according to highest education level. Those with no schooling had the highest observed prevalence (0.72%) followed by those with higher education such as a degree or diploma (0.63%). Figure 2.9 presents the observed prevalence of depression according to marital status, relationship to the head of household and whether or not a household member had passed away. A similar prevalence was observed between those or had and had not lost a household family member. A substantially higher prevalence of depression was seen in individuals who were divorced, separated or widowed (1.53%) compared to those who were married (0.41%), partnered (0.33%) or single (0.26%). The highest prevalence of depression was observed in individuals who were the head of their household (0.59%) rather than a spouse or partner (0.51%).

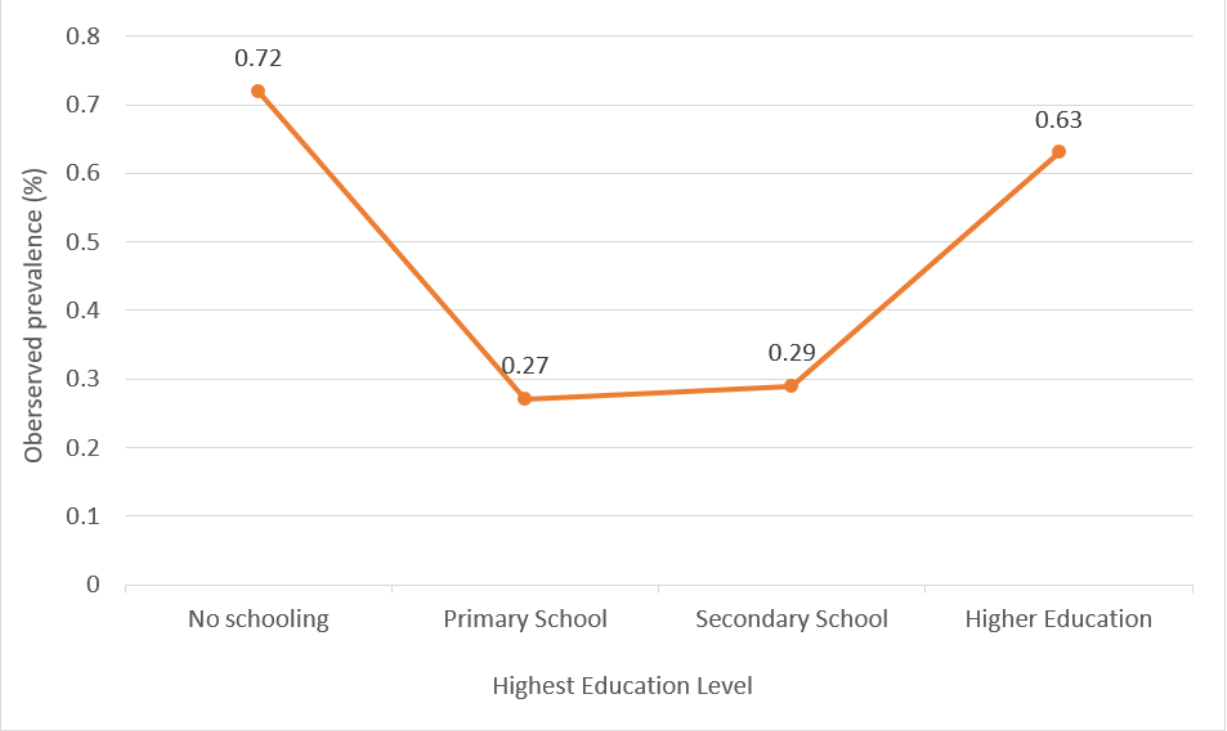


Figure 2.8: Observed prevalence of depression according to the highest level of education

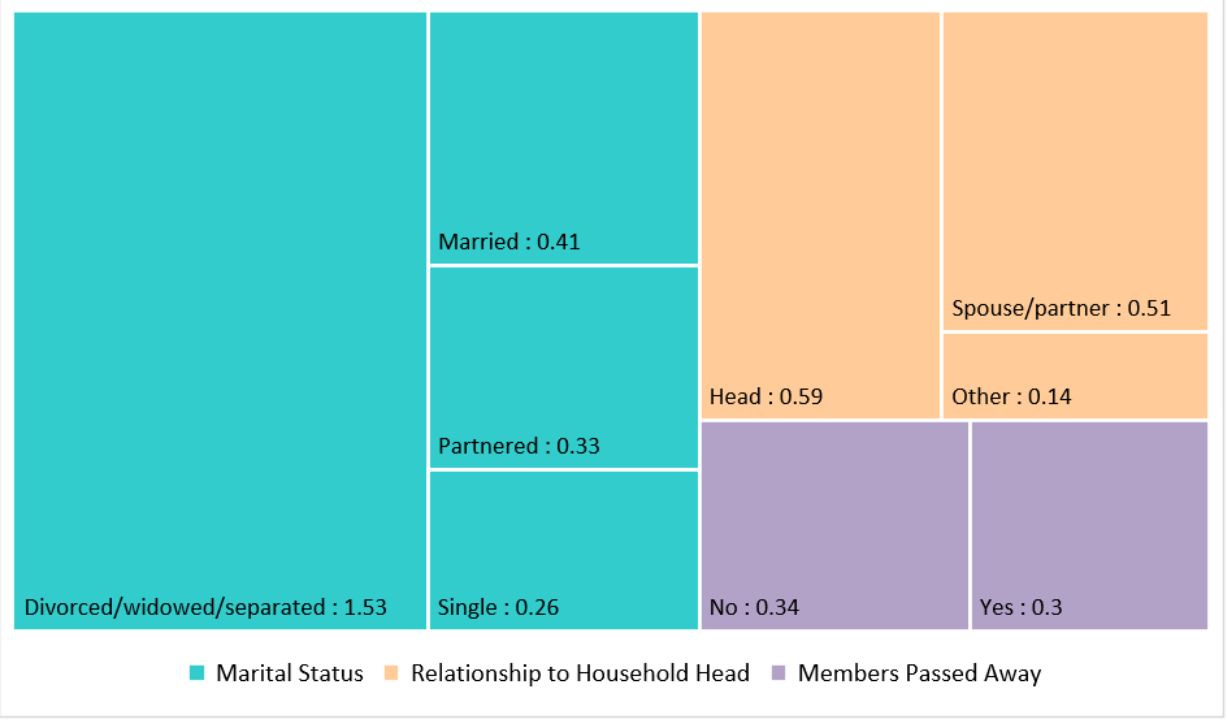


Figure 2.9: Observed prevalence of depression according to marital status, relationship to the head of household and whether or not a household member had passed away

Figure 2.10 below presents the observed prevalence of depression according to the household characteristics: access to electricity, type of toilet facility, source of water for drinking and type of place of residence. These household characteristics are associated with an individual's socio-economic status, where individuals residing in households with flush toilets, tap water and access to electricity are perceived to be at a higher socio-economic status compared to those living in households without access to these facilities. In addition, individuals residing in households with access to these facilities had a higher prevalence of depression, as well as those residing in urban areas compared to those living in other places of residence.

Based on these exploratory results, it is observed that the majority of the individuals who reported to suffer from depression were at a higher socio-economic status, where they could afford a medical aid scheme and were not on a social grant.

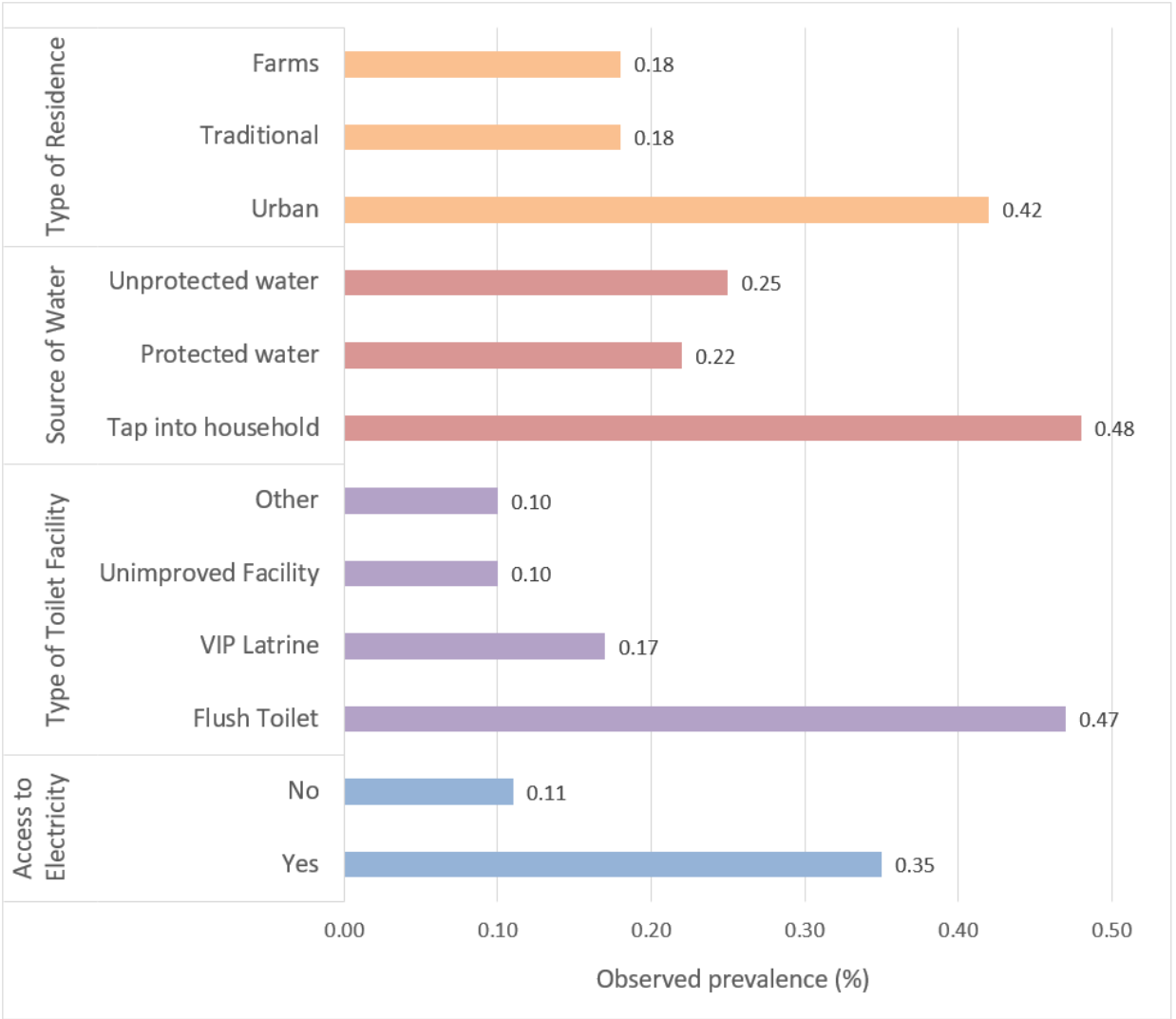


Figure 2.10: Observed prevalence of depression according to household characteristics

2.5 Statistical Methods Considered for this Study

The response variable of interest is binary, namely, the self-reported status of depression of the individual. Thus, logistic regression, which is a class of generalized linear models (GLMs), would usually be selected for the analysis. However, in this study, three different statistical approaches are considered. The first two take into account the complex survey design used to collect the data. These include a design-based approach and a model-based approach. For the design-based approach, a survey logistic regression model is used where parameter estimates and inferences are based on the sampling weights, and only inferences concerning the effects of certain covariates on the response variable are of interest (Heeringa et al., 2010). For the model-based approach, a generalized linear mixed model is applied. In this approach, interest is also on estimating the proportion of variation in the response variable that was attributable to each of the multiple levels of sampling (Heeringa et al., 2010). This approach also accounts for possible correlations in the data where individuals in the same household or cluster tend to be more alike than those from other households or clusters. The last statistical approach considered in this thesis is a Bayesian network, which is applied to model the conditional dependence among the variables. This approach is a type of probabilistic graphical model that uses Bayesian inference for calculations of the probabilities. Each of the three techniques will be used to classify the depression status of the individuals, and their performances in this classification will be compared.

The following three chapters give an overview of the three approaches as well as presents the results of each applied to the SAGHS data.

Chapter 3

Generalized Linear Models

Linear statistical models are commonly used for various analysis procedures (Ravishanker & Dey, 2001). The general linear model (LM) forms the basis of the statistical analysis that is used in applied and social research. However, the LM assumes that the response variable is continuous and is thus not appropriate in the case of a discrete, binary response. In this chapter, appropriate methods of dealing with discrete responses are discussed.

3.1 Generalized Linear Models

In cases where the response variable is not normal, the general linear model cannot be applied. Instead, the generalized linear model (GLM) is used to model the data. The GLM was developed by Nelder and Wedderburn in 1972. It can be used to fit a binary response variable that follows a general distribution of the exponential family (Agresti, 2002).

If Y_i for $i = 1, \ldots, n$ is a response variable from a distribution that is a member of the exponential family, then the probability density function for Y_i is given by

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

where θ_i is the natural or canonical parameter, $b(\theta_i)$ is a function of the natural parameter, $a(\phi) = \frac{\phi}{\omega_i}$ is a function of the scale parameter, where ω_i is the weight for the i^{th} observation and $c(y_i, \phi)$ is a constant.

The mean and variance of the response variable are:

$$E(Y_i) = b'(\theta_i) = \mu_i \qquad i = 1, 2, ..., n, \tag{3.2}$$

$$var(Y_i) = a_i(\phi)b''(\theta_i) = a_i(\phi)v(\mu_i) \qquad i = 1, 2, ..., n, \tag{3.3}$$

where $v(\mu_i) = b''(\theta_i)$

The GLM consists of three components:

1. **Random Component**: This consists of the independent response variables Y_i which belong to the exponential family with a probability distribution given by Equation 3.1.

2. **Systematic component**: This component relates a vector $\boldsymbol{\eta} = (\eta_1, \eta_2, ..., \eta_n)'$ to a set of explanatory variables through a link function. Let $\boldsymbol{x}_i = (1, x_{1i}, ..., x_{pi})'$ be a $(p+1)$-dimensional vector of covariates and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ be a vector of the unknown regression coefficients. The distribution of Y_i depends on $\boldsymbol{x}_i$ through the linear predictor, η_i, such that

$$\begin{aligned} \eta_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} \\ &= \boldsymbol{x}_i'\boldsymbol{\beta}. \end{aligned}$$

3. **A link function**: This component, denoted by g, is a monotonic and differentiable function that links the mean response $\mu_i = E(y_i)$ to the linear predictor $\eta_i = \boldsymbol{x_i'\beta}$ as follows

$$\eta_i = g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}. \tag{3.4}$$

Distributions such as the Poisson, Binomial, Chi-Square and Gamma distribution belong to the exponential family. Each member of the exponential family of distributions has a unique canonical link function. The logit is the link function for the Binomial distribution. The GLM with a logit link is referred to as a logistic regression model which is discussed in Section 3.2

3.1.1 Parameter Estimation

The maximum likelihood (ML) method can be used for the parameter estimation of GLMs. Advances in statistical theory and computer software have allowed this method of estimation to become the most popular technique in applied statistics (Wu, 2005). The log-likelihood function for a single observation is given by

$$\ell_i = \ln\ f(y_i, \theta_i, \phi) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi). \tag{3.5}$$

Since $Y_i, i = 1, ..., n$, are independent, the joint log-likelihood function is

$$\ell(\boldsymbol{\beta}, \boldsymbol{y}) = \sum_{i=1}^{n} \ell_i. \tag{3.6}$$

The ML estimation of $\beta_j, j = 0, ..., p$ is the solution to the equation

$$\frac{\partial \ell_i}{\partial \beta_j} = 0. \tag{3.7}$$

To obtain the solution, we use the chain rule

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}. \tag{3.8}$$

Using Equation 3.5, we obtain

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a_i(\phi)} = \frac{y_i - \mu_i}{a_i(\phi)}. \tag{3.9}$$

Since $\mu_i = b'(\theta_i),\ Var(Y_i) = a_i(\phi)v(\mu_i),$ and $\eta_i = \sum_j \beta_j x_{ij}$,

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = v(\mu_i),$$

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}.$$

Therefore,

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{y})}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i(\phi)} \frac{1}{v(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \\ &= \sum_{i=1}^{n} (y_i - \mu_i) W_i \frac{\partial \eta_i}{\partial \mu_i} x_{ij}, \end{aligned}$$

where W_i is referred to as the iterative weights, which is given by

$$\begin{aligned} W_i &= \frac{1}{a_i(\phi)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 v_i^{-1} \\ &= \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2, \end{aligned} \tag{3.10}$$

where $v_i = v(\mu_i)$ is the variance function. Since $\eta_i = g(\mu_i)$, $\frac{\partial \mu_i}{\partial \eta_i}$ depends on the link function of the model.

Therefore, solving for the equation below will give the ML estimate for $\boldsymbol{\beta}$

$$\sum_{i=1}^{n}(y_i - \mu_i)W_i\frac{\partial \eta_i}{\partial \mu_i}x_{ij} = 0. \tag{3.11}$$

Equation 3.11 is a non-linear function of $\boldsymbol{\beta}$. Iterative procedures such as *Newton Raphson* and *Fisher Score* are therefore required to solve this equation.

Newton Raphson

The iterative equation for Newton Raphson is given by

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \hat{\boldsymbol{\theta}}^{(t)} - (\boldsymbol{H}^{(t)})^{-1}\boldsymbol{U}^{(t)}, \tag{3.12}$$

where $\hat{\boldsymbol{\theta}}^{(t)}$ is the approximation of $\boldsymbol{\theta}$ at the t^{th} iteration. $\boldsymbol{U}^{(t)} = \frac{\partial \ell_p}{\partial \boldsymbol{\theta}}$ evaluated at $\hat{\boldsymbol{\theta}}^{(t)}$, where $\boldsymbol{U}$ is called the score. $\boldsymbol{H}^{(t)}$ is the Hessian matrix, $\boldsymbol{H}$, with the following elements evaluated at $\hat{\boldsymbol{\theta}}^{(t)}$:

$$\boldsymbol{H}_{jk} = \frac{\partial^2 \ell_p}{\partial \theta_j \partial \theta_k}. \tag{3.13}$$

Fisher Score

The Fisher score equation is given by

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \hat{\boldsymbol{\theta}}^{(t)} + (\boldsymbol{\mathcal{I}}^{(t)})^{-1}\boldsymbol{U}^{(t)}, \tag{3.14}$$

where $\boldsymbol{\mathcal{I}} = -E(\boldsymbol{H})$ is known as the information matrix.

Both Newton Raphson and Fisher Score require an appropriate starting value $\hat{\boldsymbol{\theta}}^{(0)}$, whereafter the process will continue until the difference between the successive approximations is negligible i.e. the algorithm converges.

Applying the Newton Raphson to Equation 3.11, the iterative equation will be

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} - (\boldsymbol{H}^{(t)})^{-1}\boldsymbol{U}^{(t)}, \tag{3.15}$$

and the iterative Fisher Score equation given by

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + (\boldsymbol{\mathcal{I}}^{(t)})^{-1}\boldsymbol{U}^{(t)}, \tag{3.16}$$

with information matrix

$$\begin{aligned} \boldsymbol{\mathcal{I}} &= -E(\boldsymbol{H}) \\ &= -E\left(\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right) \\ &= \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}, \end{aligned} \tag{3.17}$$

where $\boldsymbol{W}$ is known as the weight matrix with diagonal elements given in Equation 3.10. Equation 3.16 can be written as

$$\begin{aligned} \boldsymbol{\mathcal{I}}^{(t)}\hat{\boldsymbol{\beta}}^{(t+1)} &= \boldsymbol{\mathcal{I}}^{(t)}\hat{\boldsymbol{\beta}}^{(t)} + \boldsymbol{U}^{(t)} \\ &= \boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{z}^{(t)}, \end{aligned} \tag{3.18}$$

where $\boldsymbol{W}^{(t)}$ is the weight matrix evaluated at $\hat{\boldsymbol{\beta}}^{(t)}$, and $\boldsymbol{z}^{(t)}$ has the following elements evaluated at $\hat{\boldsymbol{\beta}}^{(t)}$

$$z_i = \eta_i + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right). \tag{3.19}$$

z_i is known as the adjusted dependent variable or the working variable. We, therefore, obtain

$$\hat{\boldsymbol{\beta}}^{(t+1)} = (\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{z}^{(t)}. \tag{3.20}$$

Each iteration step is as a result of a weighted least squares regression of z_i on the independent variables x_i, with weight W_i. Therefore, Fisher scoring can be regarded as iteratively reweighted least squares (IRWLS) carried out on a transformed version of the response variable (Knypstra, 2008).

It follows that the asymptotic variance of the estimate of $\boldsymbol{\beta}$ is the inverse of the information matrix given in Equation 3.17 and can be estimated by

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}, \tag{3.21}$$

where $\widehat{\boldsymbol{W}}$ is $\boldsymbol{W}$ evaluated at $\hat{\boldsymbol{\beta}}$ and depends on the link function of the model. The dispersion parameter, ϕ, in function $a_i(\phi)$ that is used in the calculation of W_i resolves to 0 in the IRWLS procedure. The estimate of $\boldsymbol{\beta}$ is therefore the same under

any value of ϕ. However, the value of ϕ is necessary for the calculation of the variance of $\hat{\boldsymbol{\beta}}$, so when ϕ is unknown, it can be estimated using a moment estimator (McCulloch & Searle, 2001), given by

$$\hat{\phi} = \frac{1}{n-p-1} \sum_{i=1}^{n} \frac{\omega_i (y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}, \tag{3.22}$$

where ω_i is the weight defined in Equation 3.1.

3.1.2 Measure of Fit

Assessing the goodness-of-fit of the model of interest is an important step in statistical analysis. One way in which this is done is by using the deviance. The deviance is a measure of discrepancy between the predicted values from the fitted model and the actual values of the data set. Assume for a fitted model, there are $p+1$ parameters and $\ell(\hat{\boldsymbol{\mu}}, \phi, \boldsymbol{y})$ is the log-likelihood function maximized over $\hat{\boldsymbol{\beta}}$ for a fixed value of the dispersion parameter ϕ, and $\ell(\boldsymbol{y}, \phi, \boldsymbol{y})$ is the maximum log-likelihood achievable under the saturated model where the number of parameters equals the number of observations, the scaled deviance is

$$D^s = \frac{-2\left[\ell(\hat{\boldsymbol{\mu}}, \phi, \boldsymbol{y}) - \ell(\boldsymbol{y}, \phi, \boldsymbol{y})\right]}{\phi}. \tag{3.23}$$

If $\phi = 1$, the deviance is defined as

$$D^s = -2\left[\ell(\hat{\boldsymbol{\mu}}, \phi, \boldsymbol{y}) - \ell(\boldsymbol{y}, \phi, \boldsymbol{y})\right]. \tag{3.24}$$

The scaled deviance converges asymptotically to a χ^2 distribution with $n-p-1$ degrees of freedom. Therefore, when testing at α level of significance, the fitted model is rejected if the calculated deviance is greater that or equal to $\chi^2_{n-p-1;\alpha}$

The *Generalized Pearson's Chi-Square* is another commonly used measure of goodness of fit. This is given by

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}, \tag{3.25}$$

which asymptotically follows a χ^2 distribution with $n-p-1$ degrees of freedom, where $v(\hat{\mu}_i)$ is the estimated variance function for the distribution. Similar to the deviance, the smaller the value of the χ^2 statistic, the better the fit of the model. The scaled Pearson's χ^2 statistic is $\dfrac{\chi^2}{\phi}$ (Wu, 2005). The value of the Pearson's χ^2 test for linear models is the residual sum of squares, since $v(\hat{\mu}_i)$ is generally taken as one,

which leads to both the deviance and Pearson's χ^2 statistic having exact χ^2 distributions. These measures of goodness-of-fit for other distributions have asymptotic χ^2 distributions. When samples are small, no distribution is superior to another. The deviance, however, has an advantage over Pearson's χ^2 statistic as it is additive for nested models (Nelder & Wedderburn, 1972).

3.1.3 Likelihood Ratio Test

In order to test whether an independent variable has no effect on the response variable (in other words their parameter is equal to zero), given the other variables in the model, the deviances of the full model and the reduced model can be compared. The test statistic is compared using the following

$$D_{reduced} - D_{full}. \tag{3.26}$$

Both the deviances above involve the log-likelihood for the saturated model, therefore, resulting in the following test statistic

$$\chi^2 = -2\left[\text{log-likelihood}(reduced\,model) - \text{log-likelihood}(full\,model)\right]. \tag{3.27}$$

The test statistic has an asymptotic χ^2 distribution. The degrees of freedom is the difference in the number of parameters between the full model and the reduced model. This test is known as the *Likelihood Ratio Test*.

As seen in the previous section, if $\phi \neq 1$, the scaled deviance can be used. Applying the scaled deviance definition, Equation 3.27 becomes

$$T = \frac{-2\left[\text{log-likelihood}(reduced\ model) - \text{log-likelihood}(full\ model)\right]}{\phi}. \tag{3.28}$$

When $\phi \neq 1$ and unknown, the value of ϕ can be estimated by using Equation 3.22.

3.1.4 Wald Test

The Wald test is a hypothesis test usually performed on parameters that have been estimated by the maximum likelihood. Suppose a hypothesis test on a single parameter β_j needs to be performed. The test statistic for this test is

$$z_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}. \tag{3.29}$$

The standard error of β_j is the square root of the diagonal elements in the inverse of

the information matrix given in Equation 3.17. The null hypothesis for the Wald test is $H_0 : \beta_j = 0$, i.e. the variable has no effect on the response. Therefore, for large values of the test statistic, the null hypothesis is rejected and it can be concluded that the corresponding variable is significant to the model and thus has a significant effect on the response.

3.1.5 Quasi-Likelihood Function

The maximum likelihood requires a known distribution in order to evaluate the function, however, it is not always practical to have a known distribution. Wedderburn (1974) proposed the quasi-likelihood (QL) method as a way of fitting regression models without having to state a specific distribution. He considered both linear and non-linear models and showed that only the relationship between the mean and the variance of observations need to be specified in order to find the quasi-likelihood function (Elder, 1996). The issue of over-dispersion is not an issue when the QL method is used (Agresti, 2007).

Wedderburn (1974) used the following relation to determine the quasi-likelihood (specifically quasi-log-likelihood) function $Q(y_i; \mu_i)$ for each observation

$$\frac{\partial Q(y_i; \mu_i)}{\partial \mu_i} = \frac{\omega_i (y_i - \mu_i)}{\phi \, v(\mu_i)}, \tag{3.30}$$

where ω_i is the known weight associated with observation y_i.

Therefore, from Equation 3.30, we can obtain

$$Q(y_i; \mu_i) = \int_{y_i}^{\mu_i} \frac{\omega_i (y_i - t)}{\phi \, v(\mu_i)} dt + \text{some function of } y_i. \tag{3.31}$$

Thus, the maximum quasi-likelihood estimates of $\boldsymbol{\beta}$ can be obtained from Equation 3.31 using Fisher Scoring. Equation 3.22 can be used to estimate ϕ 3.22.

The QL above explains the case for which observations are independent. This can, however, be extended to cases for which the observations are correlated. It can be shown that properties of the quasi-likelihood function are similar to properties of the ordinary log-likelihood function. It is therefore possible to carry out goodness-of-fit and hypothesis tests that were previously discussed. In addition, when the distribution of the response comes from the exponential family, the log-likelihood function is identical to the quasi-log-likelihood function (Wedderburn, 1974).

3.2 Logistic Regression

The logistic regression model is a special case of a generalized linear model and is used to a model binary response. The binary response is usually coded as 0 or 1 (Hilbe, 2009). Let the response variable be defined as

$$Y_i = \begin{cases} 1 & \text{if the outcome is a success e.g. an individual is suffering from depression,} \\ 0 & \text{if the outcome is a failure e.g. an individual is not suffering from depression.} \end{cases} \tag{3.32}$$

Thus, Y_i follows a Bernoulli distribution, so we let $P(Y_i = 1) = \pi_i$ be the probability of success and $P(Y_i = 0) = 1 - \pi_i$ be the probability of failure. Therefore,

$$E(Y_i) = \pi_i, \tag{3.33}$$

and

$$Var(Y_i) = \pi_i(1 - \pi_i). \tag{3.34}$$

Suppose we use the General Linear Model to model Y_i:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i \qquad i = 1, ..., n. \tag{3.35}$$

Thus, by Equation 3.33

$$\begin{aligned} E(Y_i) &= \pi_i \\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} \qquad i = 1, ..., n \\ &= \boldsymbol{x}_i' \boldsymbol{\beta}. \end{aligned} \tag{3.36}$$

Since π_i is a probability, it is limited by $0 \leq \pi_i \leq 1$. However, using Equation 3.35 to model a binary outcome would result in the value of $E(Y_i)$ outside of its range. Therefore, a model for $E(Y_i)$ that restricts its value between 0 and 1 would be more suitable (Rencher & Schaaljie, 2008). Such a model is given by the logistic regression model, given by

$$logit(\pi_i) = \ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \boldsymbol{x}_i' \boldsymbol{\beta} \qquad i = 1, 2, ..., n, \tag{3.37}$$

where $\boldsymbol{x}_i'$ is a vector of explanatory variables corresponding to the i^{th} observation

and $\boldsymbol{\beta}$ is a vector of unknown parameters (Lugga, 2012). The left hand side of Equation 3.37 is referred to as the logit link, denoted by η_i in the GLM. Thus, Equation 3.36 will become

$$\pi_i = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}. \tag{3.38}$$

This is known as the logistic regression model which is a class of the GLM with a logit link. The value of the link, η_i, is allowed to range freely while restricting that of $E(Y_i) = \pi_i = \mu_i$ between 0 and 1. There are three commonly used links to model a binary response, namely, the logit link, the probit link and the complementary log-log link, however the logit link is the most convenient to interpret (Lugga, 2012). The maximum likelihood estimates of $\boldsymbol{\beta}$ can be found using the iterative equations discussed previously in sub-section 3.1.1. The score $\boldsymbol{U}$ can be found as follows.

A binary variable has the following probability distribution

$$f(y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}. \tag{3.39}$$

Equation 3.39 can be expressed as

$$f(y_i) = exp\left[y_i \ \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \ln(1-\pi_i)\right]. \tag{3.40}$$

The equation above is in the same form of Equation 3.1 where $a_i(\phi) = 1$, therefore, the dispersion parameter $\phi = 1, c(y_i, \phi) = 0$ and the canonical parameter $\theta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$ which results in $\pi_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}$. Therefore, $b(\theta_i) = \ln(1 + e^{\theta_i})$

Since, for the logistic regression model, $E(Y_i) = \pi_i = \mu_i$, $Var(Y_i) = \mu_i(1-\mu_i) = v_i$ and the link function

$$\begin{aligned}\eta_i &= \ln\left(\frac{\mu_i}{1-\mu_i}\right) \\ &= \ln(\mu_i) - \ln(1-\mu_i).\end{aligned}$$

It follows

$$\begin{aligned}\frac{\partial \eta_i}{\partial \mu_i} &= \frac{\partial}{\partial \mu_i}[\ln(\mu_i) - \ln(1-\mu_i)] \\ &= \frac{1}{\mu_i} + \frac{1}{1-\mu_i} \\ &= \frac{1}{\mu_i(1-\mu_i)},\end{aligned}$$

and

$$v_i^{-1} = \frac{1}{\mu_i(1-\mu_i)}.$$

Therefore,

$$\begin{aligned}W_i &= \frac{1}{a_i(\phi)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 v_i^{-1} \\ &= [\mu_i(1-\mu_i)]^2 \frac{1}{\mu_i(1-\mu_i)} \\ &= \mu_i(1-\mu_i).\end{aligned} \tag{3.41}$$

Thus, the score $\boldsymbol{U}$ given in Equation 3.11 can reduce to

$$\sum_{i=1}^{n}(y_i - \mu_i)x_{ij}, \tag{3.42}$$

where $\mu_i = \pi_i$ is given by Equation 3.38. Using this value of $\boldsymbol{U}$ in the iterative equations of Newton Raphson and Fisher Scoring given by Equation 3.15 and Equation 3.16, respectively, the ML estimate of $\boldsymbol{\beta}$ can be obtained. The variance of $\hat{\boldsymbol{\beta}}$ can be calculated by using Equation 3.21 where the diagonal elements in the weight matrix $\boldsymbol{W}$ are given by Equation 3.41.

The odds of an event occurring is given by

$$\frac{\pi_i}{1-\pi_i}. \tag{3.43}$$

Taking e^{β_j} gives the odds ratio corresponding to a one unit increase in the corresponding independent variable x_{ij}, while all other independent variables remain constant. In general, for a k unit change in the independent variable, the odds ratio is defined as $e^{k\beta_j}$. This explains how much more or less likely an event is to occur if there is a change in one of the independent variables (Kutner et al., 2005).

3.3 Survey Logistic Regression

The logistic regression model is used to model binary response. However, the results are only valid if the data was obtained using simply random sampling. Thus, in the case of survey data that consists of clustering and stratification, the logistic regression model is no longer appropriate as failure to account for the survey design can result in overestimation of standard errors which can lead to incorrect results (Heeringa et al., 2010). It is therefore necessary to make adjustments to the classical logistic regression model in order to make valid inferences. Logistic regression that is used in the analysis of data from complex survey designs, where sampling weights are accounted for, is referred to as survey logistic regression and is a design-based statistical approach (Roberts et al., 1987)

Survey logistic regression (SLR) models have the same theory as ordinary logistic regression (OLR) models. The only difference is in the estimation of the variance of the parameter estimates. If the data is selected using simple random sampling, then the SLR and OLR models give identical estimates. However, if the data is selected from a complex survey design, the regression coefficient estimates and the standard errors will differ due to incorporating the sampling weight (Moeti, 2007).

3.3.1 The Model

Let's consider the survey logistic regression model for a binary response variable $Y_{hij}(j = 1, ..., n_{hi}; i = 1, ..., n_h; h = 1, ..., H)$ which equals 1 if the j^{th} individual in the i^{th} household, within the h^{th} cluster, is suffering from depression, and 0 otherwise. Also, let $\pi_{hij} = P(Y_{hij}) = 1$ be the probability that this individual is suffering from depression. Then the survey logistic regression model is given by

$$logit(\pi_{hij}) = \boldsymbol{x}'_{hij}\boldsymbol{\beta}, \tag{3.44}$$

with

$$\pi_{hij} = \frac{\exp(\boldsymbol{x}'_{hij}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}'_{hij}\boldsymbol{\beta})}, \tag{3.45}$$

where x_{hij} is the row of the design matrix corresponding to the response of the j^{th} individual in the i^{th} household within the h^{th} cluster, and $\boldsymbol{\beta}$ is the vector of unknown regression coefficients that need to be estimated.

The maximum likelihood estimation is used to estimate parameters of an ordinary logistic regression model. However, calculation of standard errors of parameter es-

timates can be complicated for data obtained from complex survey designs (Vittinghoff et al., 2011). The survey logistic regression model has the same form as the ordinary logistic regression model. Therefore, the probability distribution of the response variable is given by

$$f(y_{hij}) = \pi_{hij}^{y_{hij}}(1-\pi_{hij})^{1-y_{hij}}, \tag{3.46}$$

which yields

$$E(Y_{hij}) = \pi_{hij} = \frac{e^{\boldsymbol{x}'_{hij}\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}'_{hij}\boldsymbol{\beta}}}, \tag{3.47}$$

and

$$Var(Y_{hij}) = \pi_{hij}(1-\pi_{hij}) = \frac{e^{\boldsymbol{x}'_{hij}\boldsymbol{\beta}}}{(1+e^{\boldsymbol{x}'_{hij}\boldsymbol{\beta}})^2}. \tag{3.48}$$

The log-likelihood function is then given by

$$\ell = \ln L(y) = \sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}\sum_{h=1}^{H} \ln f(y_{hij}). \tag{3.49}$$

Sampling weights are not incorporated in the above log-likelihood function. Therefore, the maximum likelihood estimates of the model's parameters which are obtained using this function are only valid for simple random sampling where observations are unweighted. As a result, we resort to a likelihood function that incorporates the sampling weights, namely, a pseudo-likelihood function. The pseudo-maximum likelihood (PML) is a method of estimation that uses the pseudo-likelihood function.

3.3.2 Pseudo-Likelihood Function

The PML method is similar to that of the maximum-likelihood method. However, the sampling weights are accounted for as follows:

$$P\ell = \sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}\sum_{h=1}^{H} w_{hij} \ln f(y_{hij}), \tag{3.50}$$

where $P\ell$ represents the pseudo-log-likelihood function with w_{hij} being the weight associated with the Y_{hij}^{th} observation. Substituting the probability distribution of y_{hij} given by Equation 3.46 into Equation 3.50, the pseudo-likelihood function for the survey logistic regression model obtained is

$$P\ell = \sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}\sum_{h=1}^{H} w_{hij}[y_{hij}\ln(\pi_{hij}) + (1+y_{hij})\ln(1-\pi_{hij})]. \tag{3.51}$$

The parameter estimates can be obtained if Equation 3.51 is maximized with respect to $\boldsymbol{\beta}$.

$$\frac{\partial P\ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}\sum_{h=1}^{H} w_{hij}\boldsymbol{D'}\left[\frac{y_{hij}-\pi_{hij}}{\pi_{hij}(1-\pi_{hij})}\right] = \mathbf{0}, \tag{3.52}$$

where $\boldsymbol{D'} = \dfrac{\partial(\pi(\boldsymbol{\beta}))}{\partial \boldsymbol{\beta}}$ is a vector of partial derivatives.

It can be shown that the above equation can be simplified into the following estimating equations

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}\sum_{h=1}^{H} w_{hij}(y_{hij}-\pi_{hij})\boldsymbol{x'}_{hij} = \mathbf{0}. \tag{3.53}$$

These estimating equations are nonlinear functions of $\boldsymbol{\beta}$, and therefore, requires iterative procedures such as Newton-Raphson and Fisher Scoring to be solved.

Incorporating Equation 3.15 into Equation 3.53, the Newton-Raphson iterative procedure is given by

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} - (\boldsymbol{H}^{(t)})^{-1}\boldsymbol{S}(\hat{\boldsymbol{\beta}})^{(t)}, \tag{3.54}$$

where $\boldsymbol{S}(\hat{\boldsymbol{\beta}})$ is Equation 3.53 evaluated at $\hat{\boldsymbol{\beta}}^{(t)}$ and $\boldsymbol{H}^{(t)}$ is the Hessian matric, $\boldsymbol{H}$, evaluated at $\hat{\boldsymbol{\beta}}^{(t)}$ where

$$\boldsymbol{H} = \frac{\partial^2 P\ell}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta'}}. \tag{3.55}$$

Incorporating Equation 3.16 into Equation 3.53, the Fisher Scoring iterative procedure is given by

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} - (\boldsymbol{I}^{(t)})^{-1}\boldsymbol{S}(\hat{\boldsymbol{\beta}})^{(t)}. \tag{3.56}$$

The most commonly used methods of variance estimation for the survey logistic regression model are Taylor series approximation, which is based on a linearization technique, Jackknife repeated replication (JRR) and balanced repeated replication (BRR), which are based on a resampling technique (Heeringa et al., 2010). Only the Taylor series approximation method will be considered for this thesis.

3.3.3 Taylor Series Approximation

The Taylor series approximation method is based on the simplicity associated with estimating the variance of a linear statistic, even with a complex sample design.

The parameter estimates, $\widehat{\boldsymbol{\beta}}$, are defined by

$$\boldsymbol{S}(\widehat{\boldsymbol{\beta}}) = 0. \tag{3.57}$$

Using the linear terms of the Taylor series expansion, the following approximate linearized expression is obtained

$$\boldsymbol{S}(\widehat{\boldsymbol{\beta}}) \simeq \boldsymbol{S}(\boldsymbol{\beta}) + \frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{3.58}$$

Therefore,

$$\boldsymbol{S}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{S}(\boldsymbol{\beta}) \simeq \frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{3.59}$$

Taking the variances of both sides of Equation 3.59, the variance approximation of $\boldsymbol{S}(\widehat{\boldsymbol{\beta}})$ can be expressed by

$$Var\left[\boldsymbol{S}(\widehat{\boldsymbol{\beta}})\right] = \left[\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right] Var(\widehat{\boldsymbol{\beta}}) \left[\frac{\partial \boldsymbol{S}(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}\right]'. \tag{3.60}$$

Equation 3.60 can be written as

$$Var(\widehat{\boldsymbol{\beta}}) = \left[\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]^{-1} Var\left[\boldsymbol{S}(\widehat{\boldsymbol{\beta}})\right] \left[\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]^{-1}. \tag{3.61}$$

This leads to the following sandwich type variance estimator

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}) = \left[\mathcal{I}(\widehat{\boldsymbol{\beta}})\right]^{-1} Var\left[\boldsymbol{S}(\widehat{\boldsymbol{\beta}})\right] \left[\mathcal{I}(\widehat{\boldsymbol{\beta}})\right]^{-1}, \tag{3.62}$$

where $\mathcal{I}(\widehat{\boldsymbol{\beta}}) = \dfrac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \dfrac{\partial^2 P\ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}$ is the information matrix evaluated at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ and $Var\left[\boldsymbol{S}(\widehat{\boldsymbol{\beta}})\right]$ denotes the variance-covariance matrix for the $p+1$ estimating equations. Since each estimating equation is a sample total of the individual scores for the n survey respondents, standard formulae can be used to estimate the variances and covariances (Heeringa et al., 2010).

3.3.4 Model Checking

Goodness-of-Fit

Generally after fitting a model, a test of goodness-of-fit is needed to assess whether the model used is the best for fitting the data. This test measures the discrepancy between observed values and expected values for the model used in the analysis. The main tools used for assessing the goodness-of-fit of the fitted generalized linear model is the log-likelihood ratio (deviance) and the Pearson Chi-square statistics. The goodness-of-fit tests are based on independent observations that are identically distributed. In many cases, however, the observations are not i.i.d.. In GLMs, for example, the observations are independent but not identically distributed (Jiang, 2001). In a complex survey design, observations that are in the same cluster tend to be more homogenous than observations that are from different clusters. The appropriate test of goodness-of-fit in this situation is an extension to the Hosmer-Lemeshow goodness-of-fit test as it accounts for the design of a study making it appropriate in measuring the fit of the survey logistic regression model (Hosmer & Lemeshow, 1980).

Hosmer & Lemeshow (1982) developed a set of goodness-of-fit tests to avoid problems associated with the asymptotic distribution of a Chi-square test. The Hosmer-Lemeshow goodness-of-fit-test suggests that observations have to be partitioned into g (where g is preferably 10) equal sized groups base on their ordered estimated probabilities, $\hat{\pi}_i$. The Hosmer-Lemeshow test statistic, which follows a Chi-square distribution with 8 degrees of freedom, is

$$\chi^2_{HL} = \sum_{j=1}^{10} \frac{(O_j - E_j)^2}{E_j(1 - \frac{E_j}{n_j})}, \tag{3.63}$$

where
n_j = number of observations in the j^{th} group,
$O_j = \sum_i y_i$ = observed number of cases in the j^{th} group,
$E_j = \sum_i = \hat{\pi}_i$ = expected number of cases in the j^{th} group.

Currently, there is no formal goodness-of-fit testing procedure for fitting a survey logistic regression model to sample survey data (Archer & Lemeshow, 2006). However, the proposed goodness-of-fit test is the F-adjusted mean residual test as follows:

Once the survey logistic regression model is fitted, the residuals for the j^{th} observa-

tion in the the i^{th} cluster are obtained. The residuals are calculated as follows

$$\hat{r}_{ij} = y_{ij} - \hat{\pi}(x_{ij}). \tag{3.64}$$

Then using a grouping strategy proposed by Graubard et al. (1997), the observations are grouped into deciles of risk based on their estimated probabilities and sampling weights (Archer & Lemeshow, 2006). The mean residuals by decile of risk $\widehat{\boldsymbol{M}}' = (\widehat{M}_1, \widehat{M}_2, ..., \widehat{M}_{10})$ with

$$\widehat{M}_g = \frac{\sum_i \sum_j w_{ij}\hat{r}_{ij}}{\sum_i \sum_j w_{ij}} \qquad g = 1, 2, ..., 10, \tag{3.65}$$

where w_{ij} denotes the sampling weight associated with observations y_{ij}.

The Wald test statistic for testing g categories is given by

$$\widehat{W} = \widehat{\boldsymbol{M}}' \left[\widehat{Var}(\widehat{\boldsymbol{M}})\right]^{-1} \widehat{\boldsymbol{M}}, \tag{3.66}$$

where $\widehat{Var}(\widehat{\boldsymbol{M}})$ denotes the variance-covariance matrix of $\widehat{\boldsymbol{M}}$ which can be obtained through linearization (Archer et al., 2007). This test statistic is Chi-square with $g-1$ degrees of freedom. However, the Chi-square distribution was found not to be an appropriate reference distribution. The F-corrected Wald statistic is therefore used instead. The test statistic is given by

$$F = \frac{(f-g+2)}{fg} W. \tag{3.67}$$

This test statistic is approximately F-distributed with $g-1$ numerator degrees of freedom and $f-g+2$ denominator degrees of freedom, where f is the difference between the number of sampled clusters and the number of strata, with g denoting the number of categories (Archer & Lemeshow, 2006). From this, the F-adjusted mean residual test statistic is obtained as follows

$$\widehat{Q}_m = \frac{(f-8)}{10f} \widehat{\boldsymbol{M}}' \left[\widehat{Var}(\widehat{\boldsymbol{M}})\right]^{-1} \widehat{\boldsymbol{M}}, \tag{3.68}$$

since $g = 10$ deciles of risk.

In addition to the above, information criteria such as Akaike's Information Criteria (AIC) and Schwarz Criterion (SC) can be used to compare the goodness-of-fit of two nested models.

Testing Model Parameters

Inferences about the parameters of the survey logistic regression model cannot rely on the likelihood ratio test since the SLR model is based on the pseudo-likelihood which is an approximation to the true likelihood function (Hosmer Jr et al., 2013). It is, therefore, appropriate to use the Wald test instead. The null hypothesis for the Wald test is

$$H_0 = \boldsymbol{C}\boldsymbol{\beta} = \mathbf{0}, \tag{3.69}$$

with test statistic

$$W = (\boldsymbol{C}\widehat{\boldsymbol{\beta}})' \left[\boldsymbol{C}\, \widehat{Var}(\widehat{\boldsymbol{\beta}})\, \boldsymbol{C}' \right]^{-1} (\boldsymbol{C}\widehat{\boldsymbol{\beta}}), \tag{3.70}$$

where

- $\boldsymbol{C}$ is a matrix of constants that defines the hypothesis to be tested,
- $\widehat{Var}(\widehat{\boldsymbol{\beta}})$ is the estimated variance-covariance matrix for $\hat{\boldsymbol{\beta}}$ given by Equation 3.60.

The decision on whether to reject, or not to reject, the null hypothesis is based on the Chi-square test statistic with q degrees of freedom, where q is the number of independent rows of the matrix $\boldsymbol{C}$, and the p-value. Thus, the null hypothesis is rejected at a p-value associated with this test statistic, of less that 0.05, else it is not rejected. The Wald test statistic can be approximated to an F-statistic using Equation 3.67 with $g - q$ degrees of freedom.

3.3.5 Survey Logistic Regression for Classification

The survey logistic regression model can be used as a classification technique by obtaining the estimated probability of the event of interest occurring, given by

$$\widehat{\pi}_{hij} = \frac{\exp(\boldsymbol{x}'_{hij}\widehat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{x}'_{hij}\widehat{\boldsymbol{\beta}})}, \tag{3.71}$$

Then, based on a comparing the estimated probability to a specified threshold, the event of interest can be classified as either occurring or not occurring. In the case of a binary response, such as depression status, the default threshold is 0.5. For example, if $\widehat{\pi}_{hij} > 0.5$, we classify the individual as having depression. However, this threshold can be changed depending on the domain under analysis. In this study, as the event is rare with only 0.3% of the individuals stating they have depression, we will

consider a threshold of 1% as well as the default of 50% when using the statistical models to classify the individual's depression status.

3.4 Classification Performance Measures

The performance of a model with regards to classification can be calculated by making use of a confusion matrix. A confusion matrix shows where the model gets 'confused' in making predictions. Information on actual values in the rows and predicted values in the columns are represented. The following terms are defined with respect to classifying whether an individual suffers from depression or not:

1. True Positive (TP): the number of cases correctly identified for those that suffer from depression.
2. False Positive (FP): the number of cases incorrectly identified for those that suffer from depression.
3. True Negatives (TN): the number of cases correctly identified as not suffering from depression
4. False Negative (FN): the number of cases incorrecly identified as not suffering from depression.

A general form of the confusion matrix is depicted in Table 3.1 below (Beleites et al., 2013).

Table 3.1: General Confusion Matrix

Observed	Predicted	
	Yes	No
Yes	TP	FN
No	FP	TN

From the confusion matrix, we can determine the model performance by assessing the following fractions:

- Accuracy measures the proportion of actual positives and negatives that are correctly identified, given by

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3.72}$$

- Sensitivity measures the proportion of actual positives that are correctly identified, given by

$$Sensitivity = \frac{TP}{TP + FN}. \tag{3.73}$$

- Specificity measures the proportion of actual negatives that are correctly identified, given by

$$Specificity = \frac{TN}{TN + FP}. \tag{3.74}$$

- Precision measures the proportion of actual positives out of all those that are predicted to be positive, given by

$$Precision = \frac{TP}{TP + FP}. \tag{3.75}$$

3.5 Survey Logistic Regression Model Applied to SAGHS data

The analysis in this section was done using SAS version 9.4. The procedure PROC SURVEYLOGISTIC was used to fit the survey logistic regression model to the data. PROC SURVEYLOGISTIC incorporates the complex survey design in the analysis, including designs with stratification and unequal weighting. For this study, the household weights are equal to the inverse of the probability of a household being selected to take part in the survey. Each individual received their household's sampling weight, and thus their observation was weighted according to this household weight. The weights were adjusted for non-response and missing observations. The Taylor series approximation method was used for variance estimation of the SLR model

Before the final SLR model was obtained, a univariate SLR model was fitted for each explanatory variable to assess its association with depression status. Only those variables that were significant using a relaxed p-value of 20% were selected for the final SLR model. Based on the univariate models, the following variables were not included in the final model:

- Employment status
- Medical aid cover
- Social grants
- Cellphone ownership

- Electricity access

- Whether or not a household member had passed away

Furthermore, to avoid possible confounding effects between the explanatory variables, all meaningful two-way interactions effects were explored. Only significant interactions where the full model achieved convergence as well as a substantially improved AIC value were selected for the final model.

In SAS, the model's predictive accuracy can be assessed using statistics such as the concordance index (*c*), Somers' D (SD), Goodman-Kruskal Gamma (GKG), and Kendall's Tau-a (KT). These statistics are calculated as follows:

$$\begin{aligned} c &= [n_t - 0.5(t - n_c - n_d)]t^{-1} \\ SD &= (n_c - n_d)t^{-1} \\ GKG &= (n_c - n_d)(n_c + n_d)^{-1} \\ KT &= (n_c - n_d)[0.5N(N-1)]^{-1} \end{aligned}$$

where

- n_c is the number of concordant pairs (a pair of observations with different observed responses is concordant if the observation with the lower ordered response value, $y = 0$, has a lower predicted mean score than the observation with the higher ordered response value, $y = 1$),

- n_d is the number of discordant pairs (the opposite to concordant pairs),

- N is the sum of observation frequencies in the data and t is the total number of pairs,

- $t - n_c - n_d$ is the number of tied pairs, which are paired observations with different responses that are neither concordant nor discordant.

The concordance index *c* is equal to the area under the receiver operating characteristic (ROC) curve and ranges from 0 to 1. A value of 0 implies that there is no association. The predictive accuracy is poor if *c* is between 0.5 and 0.6, moderate between 0.6 and 0.7, acceptable between 0.7 and 0.8 and excellent if *c* is greater than 0.8. Somers' D statistic, which ranges from -1 (all pairs disagree) to 1 (all pairs agree), is used to determine the strength and direction of relation of the pairs of observations (UCLA: Statistical Consulting Group, 2019). The Goodman-Kruskal Gamma statistic, which also ranges from -1 (no association) to 1 (perfect association), is calculated similarly to Somers' D, however it ignores the tied pairs. The last statistic, Kendall's

Tau-a, is a modification to Somers' D where it takes into account the difference between the total number of paired observations and the number of paired observations with different responses (SAS Institute Inc., 2013). In addition to choosing the significant two-way interaction effects based on those that substantially reduced the AIC value, the interaction effects were also selected based on those that maximized the four statistics given above.

The final SLR model is presented in Table 3.2 below. The main effects for variables relationship to household head, race, health status and substance abuse had a significant effect on the likelihood of depression at a 5% level of significance. In addition, disability status and highest education level were significant at a 10% level of significance. The interaction effects between substance abuse and disability status, as well as highest education level and gender were both had a significant effect on the likelihood of depression. This final model resulted in a concordance index of 0.896, thus, the predictive accuracy of the final SLR model is in an acceptable range.

Table 3.2: Analysis of effects for the final SLR model

Effect	**F-Value**	***p*-value**
Age	0.10	0.758
Number of young children	1.44	0.231
Household size	0.11	0.745
Marital status	0.48	0.696
Province	1.42	0.183
Relationship to household head	6.79	0.001
HIV/AIDS	0.11	0.737
Type of residence	1.10	0.332
Race	680.6	<0.001
Health status	15.20	<0.001
Type of toilet facility	1.92	0.124
Source of water	2.23	0.107
Substance abuse	57.93	<0.001
Disability status	3.58	0.059
Highest education level	2.51	0.057
Gender	0.19	0.665
Substance abuse * Disability status	7.06	0.008
Highest education level * Gender	3.44	0.016

Table 3.3 on the next page displays the estimated adjusted odds ratios (OR) and their 95% confidence intervals (CI) for the variables that were not included in the interactions.

Table 3.3: Odds ratio estimates (OR) and corresponding 95% confidence intervals (CI) for the variables not included in interactions for the SLR model

Variables	Odds Ratio (95% CI)
Age	0.995 (0.964; 1.027)
Number of young children	0.774 (0.509; 1.177)
Household size	0.977 (0.849; 1.124)
Marital status (ref = single)	
Divorced/widowed/separated	1.282 (0.598; 2.746)
Married	0.677 (0.255; 1.801)
Partnered	0.844 (0.342; 2.083)
Province (ref = Western Cape)	
Eastern Cape	1.234 (0.548; 2.776)
Free State	1.078 (0.391; 2.972)
Gauteng	0.808 (0.360; 1.817)
KwaZulu-Natal	0.157 (0.037; 0.664)*
Limpopo	0.697 (0.227; 2.139)
Mpumalanga	0.776 (0.254; 2.367)
North West	1.129 (0.439; 2.902)
Northern Cape	1.009 (0.341; 2.987)
Relationship to household head (ref = head of household)	
Spouse/partner	0.594 (0.220; 1.603)
Other	0.194 (0.081; 0.464)*
HIV/AIDS (ref = yes)	
No	1.151 (0.507; 2.613)
Type of residence (ref = urban)	
Farms	0.491 (0.100; 2.420)
Traditional	1.802 (0.624; 5.203)
Race (ref = white)	
African/Black	0.521 (0.243; 1.119)
Coloured	0.552 (0.209; 1.461)
Indian/Asian	<.001 (<.001; <.001)*
Health status (ref = Poor)	
Excellent	0.050 (0.018; 0.135)*
Very Good	0.037 (0.012; 0.117)*
Good	0.102 (0.043; 0.245)*
Fair	0.548 (0.241; 1.249)

Continued on next page

Table 3.3 – *Continued from the previous page*

Variables	Odds Ratio (95% CI)
Source of water (ref = tap)	
Protected water	0.624 (0.337; 1.155)
Unprotected water	1.482 (0.446; 4.918)
Type of toilet facility (ref = flush toilet)	
Unimproved Facility	0.225 (0.064; 0.791)*
VIP Latrine	0.354 (0.110; 1.141)
Other	0.213 (0.021; 2.148)

*significant at 5% level of significance

Significance of the factors were assessed based on the inclusion of 1 in the 95% confidence interval for the odds ratio. No significant differences in the odds of depression were seen with an increase in age, number of household members and the number of children five and under in the household (Table 3.3. In addition, there were no significant differences in the odds of depression based on marital status, whether or not the individual has HIV/AIDS, type of residence and source of drinking water. The inability of the SLR model to detect significance differences for these variables may be attributed to the small number of cases in the data. However, the results suggest that individuals residing in KwaZulu-Natal had a significantly lower odds of depression compared to those residing in the Western Cape (OR = 0.157, 95% CI: 0.037–0.664). Those who had a relationship to the household head other than being a spouse or partner had a significantly lower odds compared to those who were the head of household (OD = 0.194, 95% CI: 0.081–0.464). Individuals from the Indian/Asian race group had a substantially lower odds of depression compared to those from the White race group. This result is due to no Indian/Asian individuals reported to have suffered from depression in the survey. The reported health status of an individual was significantly associated with their depression status, where the likelihood of depression was much lower for those who had a positive perception of their health. Compared to individuals residing in households that have flush toilet facilities, those in households with unimproved toilet facilities had a significantly lower likelihood of depression (OR = 0.225, 95% CI: 0.064–0.791).

Figures 3.1 and 3.2 present the estimated log-odd of depression according to the interaction between substance abuse and disability, and the interaction between highest education level and gender, respectively. A positive log-odds is associated with a higher likelihood of depression, and a negative log-odds is associated with a decreased likelihood of depression.

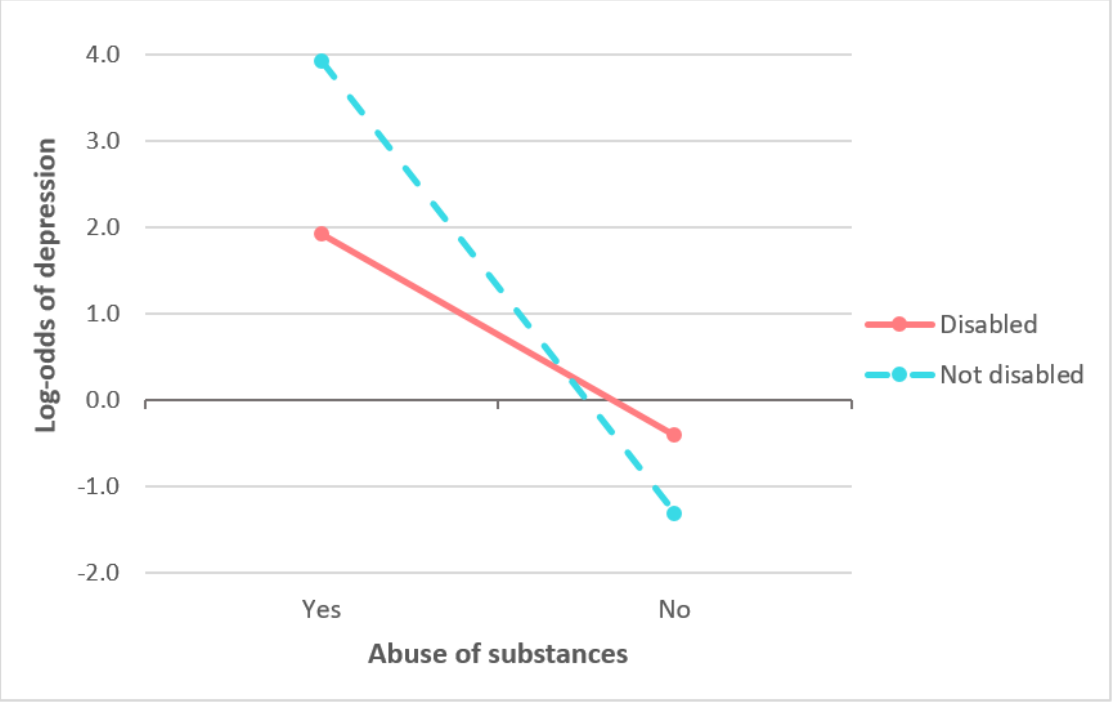


Figure 3.1: The estimated log-odds of depression associated with substance abuse and disability for the SLR model

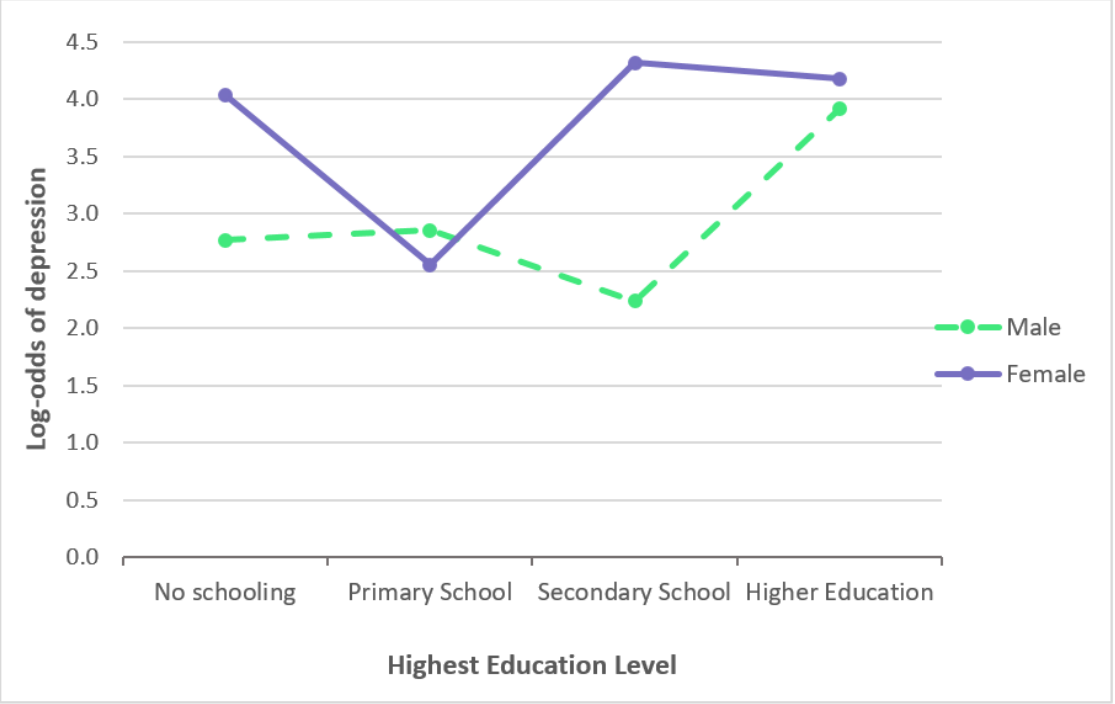


Figure 3.2: The estimated log-odds of depression associated with highest education level and gender for the SLR model

Interestingly, the log-odds of depression was higher for those who were not disabled but abusing substances, compared to those who were disabled and abusing substances (Figure 3.1). However, there was a decrease in the log-odds for those who were not abusing substances, regardless of whether they were disabled or not. Thus, in general, individuals abusing substances had a substantially higher likelihood of being depressed. Based on Figure 3.2, females in general had a higher likelihood of depression, except for those with only a primary school level of education as males with this same level of education had the higher likelihood. A significant difference in the log-odds of depression was seen between males and females with a secondary school level of education.

The predicted probabilities of an individual's depression status for each observation was extracted from the SLR model. The classification can be seen in Tables 3.4 and 3.5, which were based on a 50% and 1% threshold, respectively. An accuracy of 99.6% was achieved based on the 50% threshold, however it resulted in a sensitivity of only 1% and a precision of 25%. These calculations can be seen below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{1 + 29737}{1 + 99 + 3 + 29737} = \frac{29738}{29840} = 0.996$$

$$Sensitivity = \frac{TP}{TP + FN} = \frac{1}{99 + 1} = \frac{1}{100} = 0.01$$

$$Precision = \frac{TP}{TP + FP} = \frac{1}{1 + 3} = \frac{1}{4} = 0.25$$

Table 3.4: Confusion Matrix for the Survey Logistic Regression Model based on a 50% threshold

Observed	Predicted		Total
	Yes	No	
Yes	1	99	100
No	3	29737	29740
Total	4	29836	29840

Using a threshold of 1%, an accuracy of 94.1% was achieved, with a sensitivity of 67% and a precision of 3.7%. These calculations can be seen below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{67 + 28019}{67 + 33 + 1721 + 28019} = 0.941$$

$$Sensitivity = \frac{TP}{TP + FN} = \frac{67}{67 + 33} = \frac{67}{100} = 0.67$$

$$Precision = \frac{TP}{TP + FP} = \frac{67}{67 + 1721} = \frac{67}{1788} = 0.037$$

Table 3.5: Confusion Matrix for the Survey Logistic Regression Model based on a 1% threshold

Observed	Predicted		Total
	Yes	No	
Yes	67	33	100
No	1721	28019	29740
Total	1788	28052	29840

3.6 Summary

This chapter considered an overview of generalized linear models, which is a class of models used to model non-normal responses. An extension of this class of models was presented, namely the survey logistic regression model, which is used to model a binary response based on data obtained from a complex survey design. The survey logistic regression model is a design-based approach as it incorporates the sampling weights in the estimation of the parameters and their standard errors.

This chapter concluded with the presentation of the results from the survey logistic regression model applied to the SAGHS data. Two interaction effects were included in the final model. The model accounted for the design of the survey and assumed that the observations were independent. However, it did not account for the effects of clustering where observations may be correlated. Thus, the next chapter discusses an approach that accounts for such clustering in the survey design.

Chapter 4

Generalized Linear Mixed Models

Generalized linear mixed models (GLMMs) are an extension of the GLM in which random effects are added to the model. The use of GLMMs can allow random effects to be properly specified and computed, where errors can also be correlated. Generalized Linear Mixed Models focus on non-normal distributions, but the model can include normally distributed data as a special case. This model can overcome the over-dispersion in the data and at the same time, accommodate the population heterogeneity. The main difference in the structure of GLMMs as compared with GLMs is the incorporation of the random effects term into the linear predictor. However, the nature of the data may also dictate the use of GLMMs rather than the GLM. These have models become more applicable in many practical situations, however, parameter estimation is not as straight forward due of the inclusion of the random effects (McCulloch & Neuhaus, 2005).

A variable is considered a fixed effect when interest is on the effect of only those factor levels that are included in the model (McCulloch & Neuhaus, 2005). However, the variable is considered a random effect if the factor levels of the variable that are included in the model only represents a sample from a larger population of potential factor levels, and inferences are to be made on the whole population of factor levels. A random effect is used to model the random variation in the response variable based on the different levels of the factor (Jiang, 2007). Hence, mixed models are often applied in the modelling hierarchical or multilevel data, where observations can be placed in levels of hierarchy in the data. Such data includes clustered, repeated measures and longitudinal data where observations within the same cluster or from the same individual tend to be more homogeneous compared to those from another cluster or individual. This means that these observations can no longer be treated as independent. However, the inclusion of a random effect in the model allows for the correlation structure of the observations to be modelled and accounted for.

As the data used in this study was based on a multistage cluster sampling design, the clusters included in the study only represent a random sample of clusters. In addition, the clusters were generally based on some kind of community, such as a village or residential area. Therefore, individuals in the same cluster may have more in common and thus be more alike compared to those from different clusters. It is for this reason that we consider the GLMM to model the depression status of an individual, as it can account for possible correlations in the data.

4.1 The GLMM

Let Y_{ij} indicate the j^{th} response, $j = 1, 2..., n_i$, for the i^{th} cluster, $i = 1, 2, ..., m$, and $\boldsymbol{y}_i$ indicates the $n_i \times 1$ vector of responses for the i^{th} cluster. In the GLMM, responses Y_{ij} of $\boldsymbol{y}_i$ are assumed to be conditionally independent given a vector of random effects, γ_i which are normally distributed. Observations in GLMMs arise from a distribution in the exponential family with the following form

$$f(y_{ij}|\theta_{ij}, \phi) = \exp\left\{\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right\}, \tag{4.1}$$

which follows the same form as Equation 3.1 in Chapter 3, and thus the parameters in the above equation are similarly defined as those in Equation 3.1.

The mean μ_{ij} is the conditional mean of Y_{ij} that is modelled through a linear predictor η_{ij}, containing fixed regression parameters $\boldsymbol{\beta}$, as well as subject-specific parameters γ_i. Thus,

$$\begin{aligned} \eta_{ij} &= g(\mu_{ij}) = g\big[E(y_{ij}|\gamma_i)\big] \\ &= \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{\gamma}_i, \end{aligned} \tag{4.2}$$

or in matrix form

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma}, \tag{4.3}$$

where $g(.)$ is the known link function that links the conditional mean of $\boldsymbol{y}$ and the linear form of predictors. $\boldsymbol{X}$ is the $n \times (p+1)$ design matrix for fixed effects. $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of fixed effects regression coefficients. $\boldsymbol{Z}$ is the $n \times q$ design matrix for the random effects and $\boldsymbol{\gamma}$ is a $q \times 1$ vector of random effect coefficients. Thus, it is assumed $\boldsymbol{\gamma} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{G})$ where $\boldsymbol{G}$ depends on unknown variance components.

There are two approaches used to estimate the parameters in a GLMM: the Bayesian

approach and the maximum likelihood approach. The method of maximum likelihood is the most commonly used method of estimation and has a variety of optimality properties (Searle et al., 2006). This study focusses on the maximum likelihood approach.

4.2 Maximum Likelihood Estimation

In linear mixed models, estimation of parameters is based on the marginal likelihood of the data and can be evaluated analytically (Jiang, 2007). With GLMMs, to obtain maximum likelihood estimates, the marginal likelihood is maximized. This is done by integrating over the distribution of q-dimensional random effects. The contribution of the i^{th} cluster to the likelihood is given by

$$f_i(y_{ij}|\boldsymbol{\beta}, \boldsymbol{G}, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\gamma}_i, \boldsymbol{\beta}, \phi) f(\boldsymbol{\gamma}_i|\boldsymbol{G}) d\boldsymbol{\gamma}_i, \tag{4.4}$$

where $f(\boldsymbol{\gamma}_i|\boldsymbol{G})$ is the distribution of random effects.

Therefore, the complete likelihood function for $\boldsymbol{\beta}$, $\boldsymbol{G}$ and ϕ is given by

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{G}, \phi) &= \prod_{i=1}^{m} f_i(y_{ij}|\boldsymbol{\beta}, \boldsymbol{G}, \phi) \\ &= \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\gamma}_i, \boldsymbol{\beta}, \phi) f(\boldsymbol{\gamma_i}|\boldsymbol{G}) d\boldsymbol{\gamma}_i. \end{aligned} \tag{4.5}$$

In the case on modelling a non-normal response, parameter estimation for the GLMM becomes difficult. Generally, unlike the linear mixed models, the likelihood function under the GLMM does not have a closed-form expression (Jiang, 2007). This is due to the likelihood involving high-dimensional integrals that cannot be evaluated analytically. Thus, approximations are required to evaluate the likelihood function given in Equation 4.5. There have been a number of proposed methods of approximation (Hedeker, 2005), however, there are three basic approaches:

- approximation of the integrand.
- approximation of the integral itself.
- approximation of the data.

4.3 Laplace Approximation

The exact likelihood function can sometimes be difficult to evaluate. A common method used for an approximation of the likelihood function is the Laplace approximation, which is based on an approximation of the integrand (Jiang, 2007). Suppose an integral has the form

$$\int e^{Q(\boldsymbol{x})} dx, \tag{4.6}$$

where $Q(\boldsymbol{x})$ is a known and unimodal function, and $\boldsymbol{x}$ is a $q \times 1$ vector of variables. If $Q(\boldsymbol{x})$ is minimized when $\boldsymbol{x} = \widehat{\boldsymbol{x}}$, then the second-order Taylor series expansion of $Q(\boldsymbol{x})$ around $\widehat{\boldsymbol{x}}$ is given by

$$Q(\boldsymbol{x}) \approx Q(\widehat{\boldsymbol{x}}) + \frac{1}{2}(\boldsymbol{x} - \widehat{\boldsymbol{x}})' Q''(\widehat{\boldsymbol{x}})(\boldsymbol{x} - \widehat{\boldsymbol{x}}), \tag{4.7}$$

where $Q''(\widehat{\boldsymbol{x}})$ is the Hessian of Q evaluated at $\widehat{\boldsymbol{x}}$. The approximation to this integral uses as many different estimates of $\widehat{\boldsymbol{x}}$ as necessary according to the different modes of function Q.

Since $\boldsymbol{\gamma} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{G})$, it can be shown that the integral in the likelihood Equation 4.5 is proportional to the integral in Equation 4.6, where the function Q is given by

$$Q(\gamma) = \phi^{-1} \sum_{j=1}^{n_i} \left[y_{ij}(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{\gamma}) - b(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{\gamma}) \right] - \frac{1}{2}\boldsymbol{\gamma}'\boldsymbol{G}\boldsymbol{\gamma}. \tag{4.8}$$

Thus, Laplace's method can be applied. This approximation method tends to be better for large cluster sizes and can be improved by adding higher-order terms to the Taylor series expansion.

4.4 Gaussian Quadrature

The Laplace approximation is based on a linearization method of the integrand. The Gauss-Hermite Quadrature and the Adaptive Gauss-Hermite Quadrature are alternatives that provide an approximation of the integral or numerical integration. Due to their relation with Gaussian densities, they give approximations to an integral in the following form (Liu & Pierce, 1994)

$$\int h(x) e^{-x^2} dx. \tag{4.9}$$

In order to apply the Gauss-Hermite Quadrature and the Adaptive Gauss-Hermite Quadrature methods, the likelihood contribution for the i^{th} cluster in Equation 4.4

must be represented in the form of the integral in Equation 4.9, which is done by standardizing the random effects such that they have an identity variance-covariance matrix $\boldsymbol{I}$.

Let $\boldsymbol{\delta}_i = \boldsymbol{G}^{-\frac{1}{2}}\boldsymbol{\gamma}_i$. Thus, $\boldsymbol{\delta}_i$ has a normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{I}$. The linear predictor therefore becomes $\boldsymbol{\theta}_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{G}^{\frac{1}{2}}\boldsymbol{\delta}_i$, which now contains the variance components in $\boldsymbol{G}$. This leads to the likelihood contribution for the i^{th} cluster given by

$$\begin{aligned} f(y_{ij}|\boldsymbol{\beta}, \boldsymbol{G}, \phi) &= \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\gamma}_i, \boldsymbol{\beta}, \phi) f(\boldsymbol{\gamma_i}|\boldsymbol{G}) d\boldsymbol{\gamma_i} \\ &= \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\delta}_i, \boldsymbol{\beta}, \boldsymbol{G}, \phi) f(\boldsymbol{\delta}_i) d\boldsymbol{\delta}_i. \end{aligned} \tag{4.10}$$

Thus, this equation is now in the form of Equation 4.9 and therefore can be approximated using the Gauss-Hermite quadrature or adaptive Gauss-Hermite quadrature.

In Gauss-Hermite quadrature, the integral in Equation 4.9 is approximated by

$$\int h(x)e^{-x^2}dx \approx \sum_{i=1}^{L} w_i h(x_i), \tag{4.11}$$

where the values of x_i are solutions of the L^{th} order to the Hermite polynomial with weights w_i. The values of x_i and w_i for $i = 1, 2, \ldots, 20$ are found in tables given by Abramowitz & Stegum (1972). If L increases, the approximation improves , however when the sum is taken from 1 to L, the Gauss-Hermite quadrature gives exact solutions for all polynomials of degree $2L-1$ (McCulloch & Searle, 2001). The quadrature points x_i are chosen independently of the function $h(x)$ and thus may result in x_i not lying in the region of interest (Pinheiro & Bates, 1995). This method can sum over a large number of points which can prove useful in the event of a large number of random effects in a model (Hedeker, 2005).

The adaptive Gauss-Hermite quadrature is a method that is used to overcome the issue in which x_i does not lie in the region of interest. In the adaptive Gauss-Hermite quadrature method, the quadrature points are rescaled and shifted such that the integrand in Equation 4.9 is sampled in a suitable range (Liu & Pierce, 1994). In order to obtain the same level of accuracy as the Gauss-Hermite quadrature, this method requires significantly less quadrature points. However, this adaptive Gauss-Hermite quadrature is much more time consuming to compute as the mode and curvature is

calculated for each cluster in the data set (Hartzel et al., 2001). The adaptive Gauss-Hermite quadrature reduces to the Laplace Approximation when $L = 1$.

Newton Raphson and Fisher Scoring iterative procedures can be used to maximize the likelihood after applying these numerical approximations. These methods work relatively well in the case of a single random effect or even when there are two or three nested random effects in the model. However, for more complicated structures, these methods fail (McCulloch & Searle, 2001).

4.5 Penalized Quasi-Likelihood

The Penalized Quasi-Likelihood is based on the decomposition of the data into the mean and an appropriate error term using the Taylor series expansion of the mean. The mean is a non-linear function of the linear predictor since it is the inverse of the link function (Molenberghs & Verbeke, 2005). Consider the decomposition

$$\begin{aligned} Y_{ij} &= \mu_{ij} + \epsilon_{ij} \\ &= h(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{\gamma}_i) + \epsilon_{ij}, \end{aligned} \tag{4.12}$$

where $h(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{\gamma}_i) = g^{-1}(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{\gamma}_i)$ is the inverse of the link function. The error terms are assumed to follow a distribution with a mean of zero and variance equal to $Var(Y_{ij}) = \phi v(\mu_{ij})$. Assuming the natural or canonical link function, $v(\mu_{ij}) = h'(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{\gamma}_i)$ where h' is the derivative with respect to μ_{ij} . In order to obtain an approximation of the mean and the parameters, the Taylor series expansion of Equation 4.12 is carried out. When this is done about current estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{i}}$, the method is referred to as Penalized Quasi-Likelihood (PQL) (Goldstein & Rasbash, 1996). This gives

$$\begin{aligned} Y_{ij} &\approx h(\boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}'_{ij}\widehat{\boldsymbol{\gamma}}_i) \\ &\quad + h'(\boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}'_{ij}\widehat{\boldsymbol{\gamma}}_i)\boldsymbol{x}'_{ij}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \\ &\quad + h'(\boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}'_{ij}\widehat{\boldsymbol{\gamma}}_i)\boldsymbol{z}'_{ij}(\boldsymbol{\gamma}_i - \widehat{\boldsymbol{\gamma}}) + \epsilon_{ij} \\ &= \widehat{\mu}_{ij} + v(\widehat{\mu}_{ij})\boldsymbol{x}'_{ij}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + v(\widehat{\mu}_{ij})\boldsymbol{z}'_{ij}(\boldsymbol{\gamma}_i - \widehat{\boldsymbol{\gamma}}) + \epsilon_{ij}, \end{aligned} \tag{4.13}$$

where $\widehat{\mu}_{ij}$ is equal to its current predictor $h(\boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}'_{ij}\widehat{\boldsymbol{\gamma}}_i)$ for the conditional mean $E(Y_{ij}\boldsymbol{\gamma}_i)$. This becomes

$$\boldsymbol{y}_i \approx \widehat{\boldsymbol{\mu}}_i + \widehat{\boldsymbol{V}}_i\boldsymbol{X}_i(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + \widehat{\boldsymbol{V}}_i\boldsymbol{Z}_i(\boldsymbol{\gamma}_i - \widehat{\boldsymbol{\gamma}}) + \boldsymbol{\epsilon}_i, \tag{4.14}$$

where $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are appropriate design matrices and $\widehat{\boldsymbol{V}}_i$ is the diagonal matrix with elements $v(\widehat{\mu}_{ij}) = h'(\boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}'_{ij}\widehat{\boldsymbol{\gamma}}_i)$. Rearranging the terms in the above equation and multiplying by $\widehat{\boldsymbol{V}}_i^{-1}$ gives

$$\boldsymbol{y}_i^* \equiv \widehat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i) + \boldsymbol{X}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\widehat{\boldsymbol{\gamma}} \approx \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\gamma} + \boldsymbol{\epsilon}_i^*, \tag{4.15}$$

where $E(\boldsymbol{\epsilon}_i^*) = E(\widehat{\boldsymbol{V}}_i^{-1}\boldsymbol{\epsilon}_i) = \boldsymbol{0}$. Equation 4.15 can be viewed as can be viewed as a linear mixed model for the pseudo response $\boldsymbol{y}_i^*$. Thus, methods of fitting linear mixed models become available in order to obtain updated estimates for $\boldsymbol{\beta}, \boldsymbol{G}$ and ϕ. Specifically, estimates can be obtained by optimizing the quasi-likelihood function that includes a penalty term on the random effects of the form

$$\frac{1}{2}\boldsymbol{\gamma}'\boldsymbol{G}\boldsymbol{\gamma}.$$

This results in optimizing a penalized quasi-likelihood function below

$$L_{PQL} = \sum Q_i - \frac{1}{2}\boldsymbol{\gamma}'\boldsymbol{G}\boldsymbol{\gamma}, \tag{4.16}$$

where Q_i is the quasi-likelihood function by McCulloch & Nelder (1989). More information on this procedure can be found by Breslow & Clayton (1993).

4.6 Marginal Quasi-Likelihood

This method is similar to the PQL method, however, the Taylor series expansion of the mean in Equation 4.12 is carried out about the current estimate of $\widehat{\boldsymbol{\beta}}$ for the fixed effects but about $\widehat{\boldsymbol{\gamma}}_i = \boldsymbol{0}$ for the random effects. The current predictor of $\hat{\mu}_{ij}$ will, therefore, be of the form $h(\boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}})$. Therefore, the pseudo data in Equation 4.15 can be represented as

$$\boldsymbol{y}_i^* \equiv \widehat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i) + \boldsymbol{X}_i\widehat{\boldsymbol{\beta}} \approx \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\gamma} + \boldsymbol{\epsilon}_i^*, \tag{4.17}$$

which satisfies the approximate linear mixed model. The same procedure of obtaining the updated estimates of $\boldsymbol{\beta}$, $\boldsymbol{G}$ and ϕ for the PQL method can be followed for the MQL method, however, the resulting estimates will be referred to as marginal quasi-likelihood estimates (Breslow & Clayton, 1993).

The PQL and MGL methods may result in biased estimates. These estimates may be biased towards zero (Hedeker, 2005). Various methods of dealing with these bias estimates have been proposed. Beslow & Lin (1995) and Lin & Breslow (1996) proposed the inclusion of bias correction terms and Kuk (1995) proposed the use of iterative bootstrap. Goldstein & Rasbash (1996) showed that including a second order term in the Taylor series expansion improves the accuracy of the approximations.

4.7 Model Selection

Inferences about parameters can be done by using a number of various tests. In cases where the fixed effect parameter estimates are obtained using the numerical approximations, inferences can be done using the the likelihood ratio test and Wald test for GLMs, or the F-test for the survey logistic regression model. The likelihood ratio test can be used to compare two nested models with the same covariance structure but with different mean structures. However, as the PQL and MQL methods are not likelihood-based, the likelihood ratio test cannot be used for model selection in the case of estimates being obtained using these methods (Hedeker, 2005).

The likelihood ratio test can also be used for comparing nested models with *different* covariance structures but with the same mean. Inferences on the variance components will also be valid for F-tests. However, if the variance parameter to be tested takes values on the boundary of the parameter space, the normal approximation fails and thus the test statistics for these tests will not have the traditional Chi-square distribution under the null hypothesis (Zhang & Lin, 2008). Self & Liang (1987), Stram & Lee (1994) and Zhang & Lin (2008) have shown that testing the null hypothesis of no random effects can be carried out using a mixture of Chi-squared distributions rather that the classical single Chi-square distribution.

After fitting the GLMM, it can be used to classify the depression status of an individual based on a similar procedure to that discussed for the SLR model in Sub-section 3.3.5.

4.8 GLMM Applied to SAGHS Data

The analysis in this section was also done using SAS version 9.4. In SAS, the procedure PROC GLIMMIX allows a GLMM to be fitted to the data. The RANDOM statement specifies the random effects to be included in the model. In order to account for any heterogeneity between clusters in the SAGHS data, an intercept term that varied at cluster level was included in the model, thus resulting in a random intercept model. There were 3131 clusters in the final SAGHS data. The logit link function was used with a binary distribution specified. Along with being computationally less demanding, this model was fitted using the Laplace approximation as this method is likelihood based and therefore allows for the comparison of models using model selection criteria such as AIC and BIC.

The need for a random intercept was assessed by testing if the corresponding covari-

ance parameter should be equal to zero. This was done using the COVTEST statement in SAS, which produces likelihood ratio tests for covariance parameters. The parameter under the null hypothesis fell on the boundary of the parameter space, therefore, the p-value for the test was determined using a linear combination of central Chi-Square probabilities. The table below shows the result of this test when fitting the full GLMM to the data. This result indicates that the null hypothesis of the covariance parameter equal to zero was rejected, thus suggesting the random cluster effect was significant in the model. In addition, The variance component for the random cluster effect was estimated as 52.0 with a standard error of 12.2. This estimate is far from zero, thus confirming again the need for this random effect in the model, which also suggests that there was high variation between the clusters.

Table 4.1: Test of covariance parameters based on the likelihood.

Label	**DF**	**-2Log Likelihood**	$\boldsymbol{\chi^2}$	***P*-value**
No G - side effects	1	999.62	120.23	$<$0.0001

The label SAS gives the covariance parameter of the random effect is 'G-side effects'. SAS can distinguish between two types of random effects. The covariance parameters for the random components in the model that are contained in the variance-covariance matrices $\boldsymbol{G}$ and $\boldsymbol{R}$ are referred to as G-side and R-side effects, respectively. Including a random intercept in the model according to the different clusters will result in the inclusion of G-side effects in the model. R-side effects are also called residual effects, and such an effect allows for an overdispersion effect to be added to the model which acts as a multiplier on the variance function, thus lifting the restriction of the dispersion parameter $\phi = 1$.

In order to determine which covariance structure of $\boldsymbol{G}$ is best suited for the data, each of the covariance structures (VC (Variance Components), AR(1) , UN (Unstructured) and CS (Compound Symmetry)) were tested (SAS Institute Inc., 2013). VC yielded the lowest AIC value, thus it was selected. The final GLMM was fitted using the same variables as those in the final SLR model. It should be noted that the SAS GLIMMIX procedure can include a weight statement so that the parameter estimates are weighted, just as in the case of the SLR model. However, upon attempting to incorporate the sampling weights in the GLMM, the model was highly overdispersed. Thus, the final results below are based on the unweighted GLMM, which did not suffer from residual overdispersion.

The significance of the variables in the final GLMM is presented in Table 4.2. The GLMM found fewer significant factors compared to the SLR model. Once again,

this could be attributed to the small number of depressed cases in the data. However, both the GLMM and SLR model found the following variables significant: relationship to head, health status, substance abuse, substance abuse as well as the interaction between the highest education level and gender.Highest education level was a significant factor for depression at a 10% level of significance.

Table 4.2: Analysis of effects for the final GLMM

Effect	**F-Value**	***p*-value**
Age	0.08	0.783
Number of young children	0.82	0.365
Household size	2.39	0.122
Marital status	1.49	0.216
Province	0.22	0.988
Relationship to household head	7.04	0.001
HIV status	0.25	0.619
Type of residence	0.01	0.988
Race	0.42	0.737
Health status	16.07	$<$0.001
Type of toilet facility	0.24	0.867
Source of water	1.33	0.264
Substance abuse	16.40	$<$0.001
Disability status	0.75	0.387
Highest education level	2.25	0.081
Gender	5.60	0.018
Substance abuse * Disability status	7.13	0.008
Highest education level * Gender	4.47	0.004

Table 4.3 on the next page gives the estimated adjusted odds ratios (OR) and their 95% confidence intervals (CI) for the variables that were not included in the interactions. The odds of depression for those who were not the household head, or a spouse/partner to the household head, were 0.162 times less likely to have depression (95% CI: 0.062–0.423). No significant differences in the odds of depression were seen for the different race groups or provinces. Again, the reported health status of an individual proved to be a significant factor for depression, where the odds of depression was much lower for those who perceived their health to be better than poor.

Table 4.3: Odds ratio estimates (OR) and corresponding 95% confidence intervals (CI) for the variables not included in interactions for the GLMM

Variables	Odds Ratio (95% CI)
Age	1.006 (0.965; 1.048)
Number of young children	0.786 (0.467; 1.322)
Household size	0.851 (0.693; 1.044)
Marital status (ref = single)	
Divorced/widowed/separated	1.423 (0.469; 4.320)
Married	0.459 (0.169; 1.244)
Partnered	0.434 (0.138; 1.359)
Province (ref = Western Cape)	
Eastern Cape	0.974 (0.122; 7.754)
Free State	0.924 (0.086; 9.889)
Gauteng	0.943 (0.153; 5.800)
KwaZulu-Natal	0.218 (0.009; 5.124)
Limpopo	0.788 (0.048; 12.88)
Mpumalanga	0.644 (0.043; 9.590)
North West	1.614 (0.170; 15.33)
Northern Cape	0.774 (0.064; 9.424)
Relationship to household head (ref = head of household)	
Spouse/partner	0.649 (0.243; 1.736)
Other	0.162 (0.062; 0.423)*
HIV/AIDS (ref = yes)	
No	1.310 (0.451; 3.803)
Type of residence (ref = urban)	
Farms	1.147 (0.040; 33.08)
Traditional	0.887 (0.124; 6.329)
Race (ref = white)	
African/Black	0.570 (0.114; 2.861)
Coloured	1.135 (0.172; 7.475)
Indian/Asian	0.002 (<.001; >999)
Health status (ref = Poor)	
Excellent	0.008 (0.002; 0.034)*
Very Good	0.044 (0.013; 0.142)*
Good	0.008 (0.002; 0.040)*
Fair	0.303 (0.098; 0.938)*

Continued on next page

Table 4.3 – *Continued from the previous page*

Variables	Odds Ratio (95% CI)
Source of water (ref = tap)	
Protected water	0.540 (0.233; 1.248)
Unprotected water	1.319 (0.153; 11.39)
Type of toilet facility (ref = flush toilet)	
Unimproved Facility	0.528 (0.085; 3.270)
VIP Latrine	0.792 (0.136; 4.619)
Other	1.578 (0.107; 23.32)

*significant at 5% level of significance

Figures 4.1 and 4.2 display the estimated log-odd of depression according to the interaction between substance abuse and disability, and the interaction between highest education level and gender, respectively, for the GLMM. These interaction effects display similar patterns to that of the SLR model. Individuals abusing substances had a higher likelihood of depression compared to those not abusing substances, regardless of their disability status.

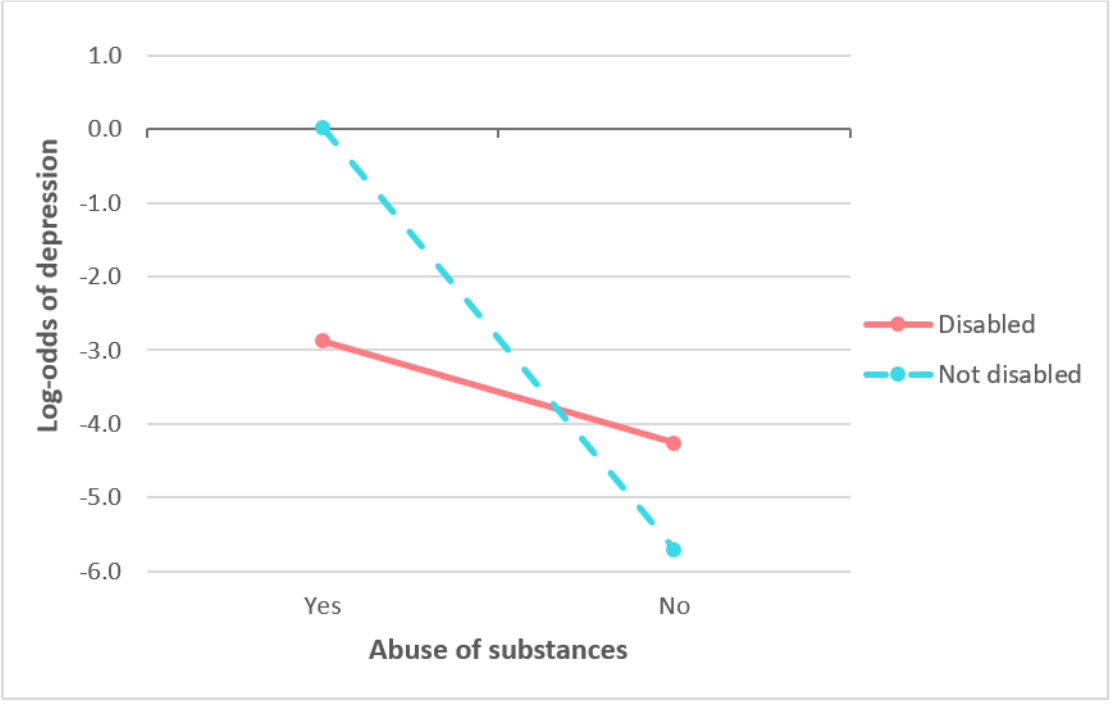


Figure 4.1: The estimated log-odds of depression associated with substance abuse and disability for the GLMM

Based on Figure 4.2 below, we can see again that females generally had a higher likelihood of depression, where females with a secondary education level had the highest chance of being depressed. However, their male counterparts (those with a secondary education level), had the lowest chance of being depressed. This Figure 4.2 clearly displays the need to account for the interaction between education level and gender. If the interaction between the two variables was not accounted for, the effect of education level on the likelihood of depression would have been the same for both males and females, which would not have been appropriate.

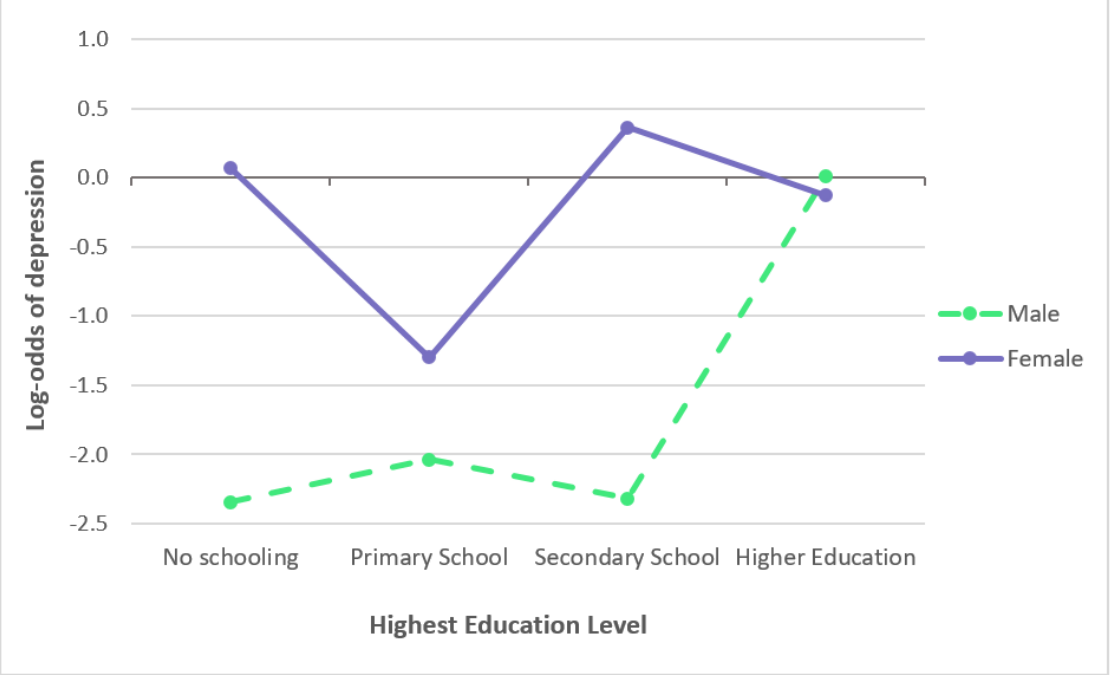


Figure 4.2: The estimated log-odds of depression associated with highest education level and gender for the GLMM

In a similar manner to that of the SLR model, the estimated probabilities based on the GLMM were extracted for each of the observations in the SAGHS data set. Based on a threshold of 50% and 1%, the depression status of the individuals was classified, the confusion matrices for which are given in Tables 4.4 and 4.5, respectively. The threshold of 50% yielded an accuracy of 99.8% with a sensitivity of 36% and precision of 80%. The calculations can be seen by the equations that follow.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{36 + 29731}{36 + 29731 + 9 + 64} = \frac{29767}{29840} = 0.998$$

$$Sensitivity = \frac{TP}{TP + FN} = \frac{36}{36 + 64} = \frac{36}{100} = 0.36$$

$$Precision = \frac{TP}{TP + FP} = \frac{36}{36 + 9} = \frac{36}{45} = 0.80$$

Table 4.4: Confusion Matrix for the Generalized Linear Mixed Model based on a 50% threshold

Observed	Predicted		Total
	Yes	No	
Yes	36	64	100
No	9	29731	29740
Total	45	29795	29840

Based on a 1% threshold, the accuracy decreased to 98.3%, with an increase in the sensitivity to 96% but a substantially lower precision of 15.9%. This can be seen by the following equations

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{96 + 29233}{96 + 29233 + 507 + 4} = \frac{29329}{29840} = 0.983$$

$$Sensitivity = \frac{TP}{TP + FN} = \frac{96}{96 + 4} = \frac{96}{100} = 0.96$$

$$Precision = \frac{TP}{TP + FP} = \frac{96}{96 + 507} = \frac{96}{45} = 0.189$$

Table 4.5: Confusion Matrix for the Generalized Linear Mixed Model based on a 1% threshold

Observed	Predicted		Total
	Yes	No	
Yes	96	4	100
No	507	29233	29740
Total	603	29237	29840

4.9 Summary

This chapter extended on the theory of Chapter 3 by considering the inclusion of a random effect in the generalized linear model to produce a generalized linear mixed model. A random effect accounts for possible correlations in the data that may be present due to multiple stages of sampling. In this study, the random effect was incorporated into the model to account for the effect of clustering. The results showed that this clustering effect was significant, indicating a high amount of variation in the

to converge. This chapter concluded with the presentation of the results from the GLMM applied to the SAGHS data, which indicated similar results to that of the SLR model.

So far, the two approaches that have been applied to model and classify an individual's depression status are based on classical statistical techniques. Next, we consider a machine learning technique that focuses on determining the important variables in classifying an individual's depression status. Such a technique is referred to as a Bayesian network.

Chapter 5

Bayesian Networks

5.1 Introduction to Bayesian Networks

Bayesian networks are probabilistic graphical models, which allow the representation of dependencies among variables and aid in specifying the joint probability distributions among the variables (Han et al., 2012). Bayesian networks are a common classification technique. Bayesian networks, also known as Bayesian belief networks, were introduced by Judea Pearl in the early 1980s as a representation device, allowing an individual to systematically and locally assemble probabilistic beliefs into a coherent whole. Some of these beliefs could be read off directly from the Bayesian network, however, many were implied by this representation and required computational work to be made explicit. Computing and explicating these beliefs has been the subject of much research and become known as the problem of inference in Bayesian networks. The computed beliefs form the basis of decision making (Darwiche, 2010). There has been a lot of progress on inferences in Bayesian networks since it was first introduced and a lot more can be done. People always exceed the ability of existing algorithms by building more complex networks (Darwiche, 2010).

A Bayesian network is defined by two components:

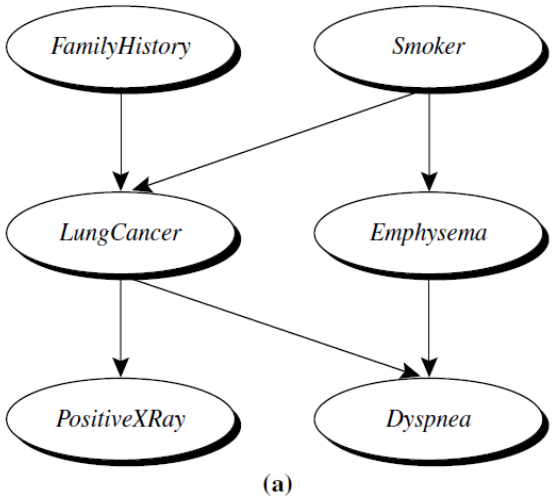
i) a directed acyclic graph (DAG).

ii) a set of conditional probability tables for each variable in the DAG (Darwiche, 2010).

The DAG consists of nodes, representing variables, and arcs, representing the dependency relations between the variables. If there is an arc from node X to node Y, then X is the parent of Y and Y is a descendent or child of X. Furthermore, a property of a DAG is that it must not consist of any loops or cycles. Each variable in

the Bayesian network is conditionally independent of its non-descendants, given its parents. The variables in a BN may be discrete or continuous. In addition, the variables may correspond to actual variables given in the data or to "hidden variables" that are believed to form a relationship.

Figure 5.1 is an example of a simple Bayesian network, taken from Han et al. (2012). The arcs, given by the arrows, allow a representation of causal knowledge. For example, having lung cancer is dependent on a person?s family history of lung cancer, as well as whether or not the person is a smoker. However, the variable PositiveXRay is independent of whether the patient has a family history of lung cancer or is a smoker, *given* information regarding whether or not the patient has lung cancer. This means that once we know the outcome of the variable LungCancer, then the variables FamilyHistory and Smoker do not provide any additional information regarding PositiveXRay. The DAG also indicates that the variable LungCancer is *conditionally independent* of Emphysema, given its parents, FamilyHistory and Smoker.



	FH, S	*FH, ~S*	*~FH, S*	*~FH, ~S*
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

(b)

Figure 5.1: An example of a Bayesian network. (a) A DAG (b) The conditional probability table for the values of the variable LungCancer (LC) showing each possible combination of the values of its parent nodes, FamilyHistory (FH) and Smoker (S) (Han et al., 2012).

The Bayesian network has one conditional probability table (CPT) for each node in the network. The CPT for a variable Y gives the conditional distribution $P(Y|Parents(Y))$. Part (b) in Figure 5.1 above gives a CPT for the variable LungCancer. The conditional probability for each known value of LungCancer is given for each possible combi-

nation of the values of its parents. For example, we see that

$$P(LungCancer = yes|FamilyHistory = yes, Smoker = yes) = 0.8.$$

These conditional probability tables can be used in calculations of probabilities of certain joint or conditional events of interest occurring in the network.

Consider n random variables $X_1, X_2, ..., X_n$, a directed acyclic graph with n numbered nodes, and suppose node i corresponds to variable i. As each variable is conditionally independent of its non-descendants in the network, given its parents, the joint probability distribution of the variables in the network is given by

$$P(X_1, X_2, ...X_n) = \prod_{i=1}^{n} P(X_i|Parents(X_i)) \quad \text{for } i = 1, 2, ..., n. \tag{5.1}$$

The value of $P(X_i|Parents(X_i)$ is taken from the CPT for variable X_i.

Bayesian networks are used for modelling knowledge in many domains with uncertain knowledge. For example, medicine, engineering, text analysis and data fusion (Gulyás, 2006).

5.2 Training a Bayesian Network

The training of a BN refers to constructing the DAG and estimating the CPT for each node of the DAG for a given data set. Thus, the training of a BN involves two steps: (1) creating the structure of the network, and (2) estimating the probability values in the tables associated with each node. The network topology, which refers to the layout of the nodes and arcs, can constructed by human experts or inferred from the data. Several algorithms exist for training the network topology from the data. However, human experts usually have good experience and understanding of the direct conditional dependencies that hold in the domain under analysis, which helps in network design. In this case, experts must specify conditional probabilities for the nodes that participate in direct dependencies. These probabilities can then be used to compute the remaining probability values (Tan et al., 2014).

If the network topology is known and all of the variables are observable, then training the network is simple where only the CPT entries need to be computed. In this case, computing the probabilities is done similarly to that of a naïve Bayesian classification technique. However, if the network topology is given but some of the

variables are hidden, the gradient descent method can be used to train the network. Such a training method results in what is called an adaptive probabilistic network. If both the topology of a BN and the values in the CPTs need to be determined, this presents a very difficult computational task (Han et al., 2012).

The trained BN can be used for inference where it can provide estimated probabilities of the event of interest occurring, which can then be used to classify the event as occurring or not.

5.3 Types of Bayesian Network Structures

Several types of Bayesian network structures can be trained (Dean, 2018).

- Naïve Bayesian network, which connects the response/target variable to each explanatory/input variable. However, there are no other connections between the variables because the input variables are assumed to be conditionally independent on each other. This is the simplest of the Bayesian networks.
- Tree-Augmented Naive (TAN), which connects the target variable to each input variable and connects the input variables in a tree structure. The tree that connects the input variables is based on the maximum spanning tree algorithm. In a TAN, each node can have at most two parents, one of which must be the target variable.
- Bayesian Network Augmented Naive (BAN), which connects the target variable to each input variable and creates a Bayesian network structure between the input variables. In a BAN, each node can have multiple parents, one of which must be the target variable.
- Parent Child (PC), which connects the target variable to each input variable. However, input variables are allowed to be either a parent of the target variable or a child of the target variable. The Bayesian Information Criterion is used to determine whether an input variable is a parent of the target variable or a child of the target variable.
- Markov Blanket, which creates a set of connections between the target variable and the input variables, but also permits connections between certain input variables. Only the children of the target variable are allowed to have an additional parent node. All other variables are conditionally independent of the target variable and thus do not affect the classification model. The Bayesian Information Criterion is used to determine whether an input variable is a parent of the target variable or a child of the target variable

5.4 Bayesian Network Applied to SAGHS Data

In SAS Enterprise Miner, the procedure PROC HPBNET allows a Bayesian Network model to be fitted to the data. The HPBNET procedure can learn different types of Bayesian network structures, including those discussed in Section 5.3. PROC HPBNET performs efficient variable selection through independence tests, and it selects the best model automatically from the specified parameters (Dean, 2018).

Table 5.1 below shows the variable selection results. PROC HPBNET removes variables where the p-value is greater than 0.05. The procedure also automatically bins the values of continuous variables into groups. However, for the purpose of this analysis, the age of the individual was incorporated into the Bayesian network as qualitative.

Table 5.1: Variable Selection

Variable	Selected	χ^2 Statistic	p-value	Degrees of Freedom	Conditional Variables
Age Group	Yes	43.95	< 0.001	6	
Number of children ≤ 5 years	No	10.03	0.187	7	
Disability	Yes	111.31	< 0.001	1	
Highest education level	No	32.03	0.058	21	Age group
Employment status	No	0.91	0.635	2	
Gender	Yes	21.01	< 0.001	1	
Type of residence	No	14.61	0.405	14	Age group
Household size	No	25.14	0.196	20	
Marital status	No	27.90	0.143	21	Age group
Province of residence	No	31.46	0.494	32	Race
Medical aid cover	No	0.43	0.511	1	
Health status	Yes	252.79	< 0.001	4	
Substance Abuse	Yes	515.74	< 0.001	1	
HIV status	No	5.76	0.330	5	Health status
Social grants	No	0.03	0.852	1	
Cellphone ownership	No	1.37	0.241	1	
Electricity Access	No	4.07	0.771	7	Age group
Death in household	No	0.04	0.827	1	
Race	No	13.18	0.356	12	Highest education level
Relationship to head	Yes	41.64	< 0.001	2	
Type of toilet facility	No	26.61	0.874	36	Province of residence
Source of drinking water	No	23.96	0.156	18	Province of residence

Table 5.1 indicates the highest education level, type of residence, marital status and electricity access are conditionally independent of depression given the age group, therefore, PROC HPBNET removes it from the network. Similarly, province of residence is conditionally independent of depression given the race of the individual, HIV status is conditionally independent of depression given health status and race is conditionally independent of depression given the highest education level.

Table 5.2 below presents the structure of the Bayesian network, where the parent nodes and child nodes are indicated for the network. In addition, the DAG for this structure is given by Figure 5.2. The health status of an individual is the parent of depression. Depression is the parent of disability, age group, gender, substance abuse, as well as the individual's relationship to the head of the household.

Table 5.2: Structure of the Network

Parent Node	**Child Node**
Health Status	Depression
Depression	Age group (in years)
Relationship to head of the household	Age group (in years)
Depression	Disability
Depression	Relationship to head of the household
Substance Abuse	Gender
Depression	Substance Abuse
Depression	Relationship to head of the household
Disability	Relationship to head of the household
Gender	Relationship to head of the household

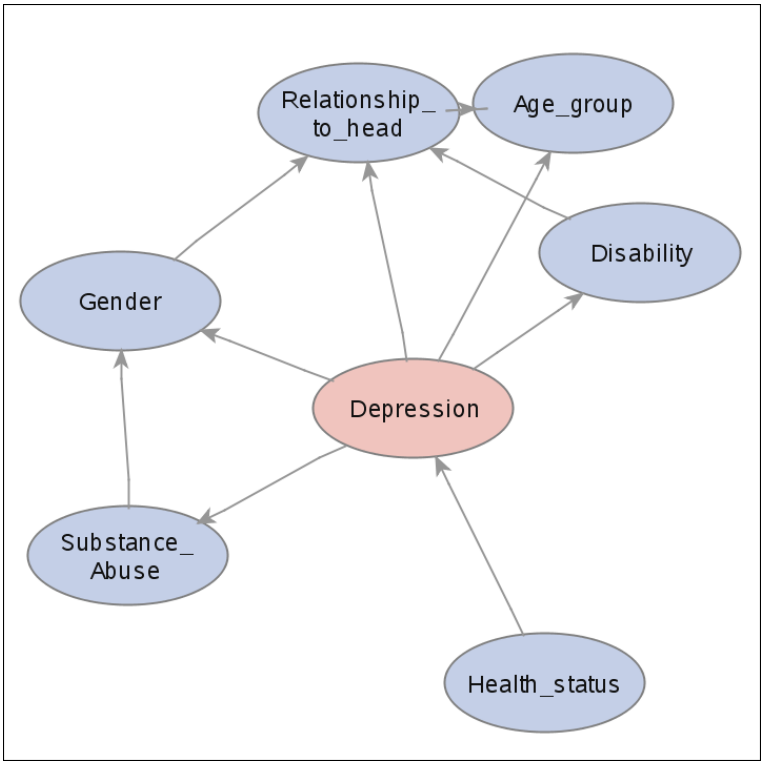


Figure 5.2: Resulting DAG for the Bayesian Network applied to the SAGHS Data.

In summary, from the structure, one can infer that depression is dependent on:

- Health status
- Age group of the individual
- Disability
- Gender
- Substance Abuse
- Relationship to head of the household

However, depression is conditionally independent of

- Number of children 5 years or younger in the household
- Highest education level
- Employment status
- Type of residence

- Household size
- Marital status
- Province of residence
- Medical aid cover
- HIV status
- Social grants
- Cellphone ownership
- Electricity access
- Whether or not household members have passed away
- Race

The conditional probability tables for the Bayesian Network is given by Table A.1 in Appendix A. This conditional probability table together with the network structure indicated by Table 5.2 determine the Bayesian Network.

The Bayesian Network indicates that, given an individual who has a poor health status, the probability of depression is 3.46%. It can also be inferred that, given the individual suffers from depression, the probability that the individual is disabled is 31.373%, the probability that the individual is female is 77.174% and the probability that the individual suffers from substance abuse is 10.784%. If the individual is depressed, the probability that he/she is the head of the household is 50% and the probability the he/she is the spouse is 32.143%. The probability that the individual is between the ages of 40 and 44 years is 25%, given the individual is depressed. If the individual is depressed, the probability that he/she is between the ages of 30 and 34 years old is 21.875% and if he/she is between the ages of 35-39 the probability is 15.625%. These are ages where the individual is at the prime of his/her working life and the responsibilities are greatest. Probabilities of certain variables of the Bayesian Network are shown in Table 5.3.

Table 5.3: Probability table of Bayesian Network

Child Node	Condition	Parent Node	Condition	Probability
Disability	Yes	Depression	Yes	0.31373
Gender	Male	Depression	Yes	0.22826
Substance Abuse	Yes	Depression	Yes	0.10784
Relationship to head	Head	Depression	Yes	0.50000
Relationship to head	Spouse	Depression	Yes	0.32143
Relationship to head	Other	Depression	Yes	0.17857
Age group	25 to 29	Depression	Yes	0.10938
Relationship to head	Head	Age group	45 to 49	0.19159
Depressed	Yes	Health status	Poor	0.03457

The Bayesian Network also indicates that if your age is between the ages of 45 and 49 years, the probability of being the head of the household is 19.159%.

The resulting Bayesian network was used to estimate the probability of depression for each individual in the sample. Using this estimated probability, the individual was classified as having depression or not based on two thresholds, 50% and 1%. Table 5.4 below presents the confusion matrix based on a 50% threshold. Based on this threshold, the table indicates that the model provides a 99.6% accuracy, however, the sensitivity is only 3% with a precision of 18.8%. This calculation can be seen below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{3 + 29727}{3 + 29727 + 13 + 97} = \frac{29730}{29840} = 0.996$$

$$Sensitivity = \frac{TP}{TP + FN} = \frac{3}{3 + 97} = \frac{3}{100} = 0.003$$

$$Precision = \frac{TP}{TP + FP} = \frac{3}{3 + 13} = \frac{96}{45} = 0.188$$

Table 5.4: Confusion Matrix for the Bayesian Network based on a 50% threshold

Observed	Predicted		Total
	Yes	No	
Yes	3	97	100
No	13	29727	29740
Total	16	29824	29840

Table 5.5 below presents the confusion matrix based on a 1% threshold. Based on this threshold, the model provides a 93.7% accuracy, however, the sensitivity is much higher at 59% compared to that using a 50% threshold. Although, the precision is lower 3.1%. The equations below indicate how these statistics are calculated.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{59 + 27888}{59 + 27888 + 1852 + 41} = \frac{27947}{29840} = 0.996$$

$$Sensitivity = \frac{TP}{TP + FN} = \frac{59}{59 + 41} = \frac{59}{100} = 0.59$$

$$Precision = \frac{TP}{TP + FP} = \frac{59}{59 + 1852} = \frac{96}{1911} = 0.031$$

Table 5.5: Confusion Matrix for the Bayesian Network based on a 1% threshold

Observed	Predicted		Total
	Yes	No	
Yes	59	41	100
No	1852	27888	29740
Total	1911	27929	29840

5.5 Summary

This chapter considered a Bayesian network to explore the important variables in determining an individual's depression status. This technique is a common classification machine learning technique which allows the representation of dependencies among the variables. A brief overview of Bayesian networks was discussed in this chapter. The chapter then concluded by presenting the results of the Bayesian network applied to the SAGHS data, where the variables that depression status was dependent on and conditionally independent on were indicated. The results of the Bayesian network were then used to classify the depression status of the individuals in the data.

The final chapter discusses and compares the results of the three approaches applied in this study. In addition, the limitations of the study are discussed, as well as possible future directions of the study.

Chapter 6

Discussion and Conclusion

6.1 Comparison of Model Accuracies

Table 6.1 below presents a comparison of the performance of the models considered in this thesis with regards to classifying an individual's depression status. Classification was done based on two thresholds, 50% and 1%. This means an individual was classified as suffering from depression if their estimated probability, which was extracted for each fitted model, was more than 0.5 for the 50% threshold and more than 0.01 for the 1% threshold. All of the models produced a very high accuracy for both thresholds, as well as a very high specificity. However, these two measures should not be focused on as these models might have achieved these accuracies by chance due to the excessive number of non-depressed cases in the data compared to the number of depressed cases. While none of the models produced sufficient performance statistics in terms of all four measures simultaneously, the GLMM based on a 50% threshold seems to be performing best with having obtained an 80% precision. The Bayesian network produced the lowest precision for both thresholds, which is unsurprising as this technique is not able to accommodate the design of the survey used to collect the data.

Table 6.1: Classification statistics (%) for the different models based on thresholds of 50% and 1%

Model	Accuracy	Sensitivity	Specificity	Precision
SLR model (50%)	99.6	1	99.99	25.9
SLR model (1%)	94.1	67	94.20	3.7
GLMM (50%)	99.8	36	99.97	80.0
GLMM (1%)	98.3	96	98.30	15.9
Bayesian Network (50%)	99.6	3	99.96	18.8
Bayesian Network (1%)	93.7	59	93.77	3.1

6.2 Limitations

The limitations of this study are numerous. The biggest limitation lies in the number of depressed cases, with only 0.3% of the sampled individuals having reported that they suffer from depression. This has restricted the ability of the statistical models to detect any significant association between the explanatory variables and depression status. Therefore, the results should be considered with caution. In addition, the data was based on a cross sectional design, therefore no causal effect can be established between the explanatory variables and the response. Many of the variables were also based on self-reporting, such as health status, depression status and disability status, which may cause bias in the data. This may be particularly true of depression status, as based on other studies, the 12-month prevalence of depression in SA is 16.5%, which is substantially higher than the 0.3% seen in the nationally representative SAGHS data (Bateman, 2014). This can be attributed to numerous reasons, such as the stigma around depression, where individuals with depression may be considered as weak, therefore restricting their willingness to be open about their depression status (Barney et al., 2006). Furthermore, many individuals may not be aware of their depression status as their knowledge of the signs and symptoms of depression may be low.

6.3 Conclusion

This study aimed to investigate the risk determinants of depression among South African individuals aged 15 to 49 years old and to determine which factors contribute the most to this mental illness. This was done by applying three approaches to the SAGHS data, which was obtained using a complex survey design. The survey logistic regression model, a generalized linear mixed model and a Bayesian network were the three approaches considered. The SLR model and GLMM account for the survey design in various ways. The SLR model incorporates the sampling weights in the estimation of the parameters, and the GLMM accounts for the effect of clustering by way of a random intercept at cluster level. The Bayesian network, which is a probabilistic graphical model, does not account for the survey design, which may explain this approach having the worst performance with regards to classification. The Bayesian network is also unable to determine the significance of a variable in modelling depression, only the importance of the variable. However, it is able to accommodate dependencies among all of the variables, unlike the SLR model and GLMM. Thus, each approach has their own advantages.

The SLR model and GLMM indicated that a person's perceived health status, abuse

of substance, and gender were significantly associated with depression. An individual with a poor perception of their health had a significantly higher odds of depression compared to those with a more positive perception. The two models found a significant interaction between disability status and substance abuse, as well as between highest level of education and gender. Individuals abusing substances had a higher likelihood of depression. Although, we are unable to state whether the individual is depressed because they are abusing substances, or if the they are abusing substances because they are depressed, however either relation has been shown to be true (Nglazi et al., 2016). In general, women had a substantially higher chance of being depressed compared to men, particularly those with only a secondary level of education. This is a common finding (Pratt, 2014). Thus, possible risk determinants of depression consist of poor health, abuse of substance and being female, particularly being female with a secondary level of education.

While the Bayesian network showed that depression was dependent on the age group of the individual, the SLR model and GLMM did not indicate that age was significantly associated with depression. Although, the latter models incorporated age as a continuous effect rather than categorical. The exploratory data analysis revealed that the average age of the depressed individuals was higher than those who were not depressed, which is in agreement to other findings (Pratt, 2014). Based on the explanatory data analysis, it appeared that individuals in a higher socioeconomic bracket were more likely to be depressed, which was confirmed by the results of the SLR model where individuals residing in households with flush toilet facilities had a significantly higher likelihood of depression compared to those residing in households with unimproved toilet facilities. This funding may be due to these individuals experiencing the pressure of sustaining such a lifestyle, or the pressure of the job that affords them such a lifestyle. This result contradicts that of Brown & Moran (1997) and Pratt (2014).

The aim of this study was to use the applied methods to classify an individual's depression status and asses which method performed the best in this classification. The purpose of being able to classify an individual's depression status, based on their individual and household factors, is to be able to identify a depressed individual in order to apply the appropriate intervention before they inflict self-harm. It is also important to be able to identify individuals at a higher risk of depression in order equip them with the knowledge of how to handle such an illness. Even though the performance of the three methods was restricted due to the limitations of the data, the GLMM proved to be the better performing technique. Thus, we recommend that when using data based on a complex survey design, a GLMM is considered in classifying the occurrence of an event of interest.

6.4 Future Direction

There is still much to do in order to assist individuals suffering from mental illnesses such as depression. South Africa especially has a long way to go in helping such individuals, particularly with the stigma around depression. As the data utilized in this study was from 2016, a future direction includes obtaining an updated SAGHS data set and exploring methods that are not sensitive to the low number of cases, in addition to being able to accommodate the survey design.

References

Abramowitz, M., & Stegum, I. (1972). Handbook of mathematical functions with formulas, graphs and mathematical tables. *New York:Dover*.

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons, Inc., 2 ed.

Agresti, A. (2007). An introduction to categorical data analysis. *John Wiley & Son, Inc.*, *61*, 439–447.

Albert, P. R. (2015). Why is depression more prevalent in women? *Journal of psychiatry & neuroscience: JPN*, *40*(4), 219.

All-About-Depression.com (2017). Genetic causes of depression. `http://www.allaboutdepression.com/cau_03.html`. [Online; accessed September-2019].

Archer, K., & Lemeshow, S. (2006). Goodness-of-fit test for a logistic regression model fitted using survey sample data. *Strata Journal*, *6*(1), 97–105.

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

Avert (2019). HIV and AIDS in South Africa. `https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/south-africa`. [Online; accessed June-2019].

Barney, L. J., Griffiths, K. M., Jorm, A. F., & Christensen, H. (2006). Stigma about depression and its impact on help-seeking intentions. *Australian & New Zealand Journal of Psychiatry*, *40*(1), 51–54.

Bateman, C. (2014). The Price of Depression. `http://www.sadag.org/images/pdf/price_of_depression.pdf`. [Online; accessed February-2019].

Beleites, C., Salzer, R., & Sergo, V. (2013). Validation of soft classification models using partial class memberships: An extended concept of sensitivity and co. applied

to grading of astrocytoma tissues. *Chemoetrics amd Intelligent Laboratry Systems*, *122*, 12–22.

Beslow, N., & Lin, X. (1995). Bias correction in generalized linear mixed models with single component of dispersion. *Biometrka*, *82*, 81–91.

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

Brown, G. W., & Moran, P. M. (1997). Single mothers, poverty and depression. *Psychological Medicine*, *27*(1), 21–33.

CIA World Factbook (2017). Africa : South africa. `https://www.cia.gov/library/publications/the-world-factbook/geos/sf.html`. [Online; accessed March-2019].

Darwiche, A. (2010). Inference in bayesian networks: A historical perspective. *Heuristics, Probability and Causality. A Tribute to Judea Pearl. College Publications*.

Dean, J. (2018). *SAS/STAT 13.1 User's Guide The SURVEYLOGISTIC Procedure*. Cary, NC: SAS Institute Inc.

Elder, J. A. (1996). *Development of Quasi-Likelihood Techniques for the Analysis of Pseudo-Proportional Data*. Ph.D. thesis, Virginia Commonwealth University.

Galambos, N. L., Leadbeater, B. J., & Barker, E. T. (2004). Gender differences in and risk factors for depression in adolescence: A 4-year longitudinal study. *International Journal of Behavioral Development*, *28*(1), 16–25.

Goldstein, H., & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society*, *A159(3)*, 505 – 513.

Graubard, B., Kom, E., & Midthune, D. (1997). Time-to-event analysis of longitudinal follow-up of a survey: choice of time-scale. *American Journal of Epidemiology*, *145*(1), 72–80.

Gulyás, C. (2006). *Creation of a Bayesian network-based meta spam filter, using the analysis of different spam filters*. Ph.D. thesis, Master Thesis, Budapest, 16th May.

Haggerty, J. (2016). Risk factors for depression. `http://psychcentral.com/lib/risk-factors-for-depression/`. [Online; accessed August-2019].

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Elsevier Inc, 3rd ed.

Hartzel, J., Agresti, A., & Caffo, B. (2001). Multinomial logit random effects models. *Statistical Modelling*, *1*, 81–102.

Healthline (2017). Risk factors for depression. `http://www.healthline.com/health/depression/risk-factors#1`. [Online; accessed September-2019].

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

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied Survey Data Analysis*. Statistics in the Social and Behavioral Sciences Series. Chapman & Hall/CRC. Taylor & Francis Group, LLC.

Herman, A. A., Stein, D. J., Seedat, S., Heeringa, S. G., Moomal, H., & Williams, D. R. (2009). The South African Stress and Health (SASH) study: 12-month and lifetime prevalence of common mental disorders. *South African Medical Journal*, *99*, 339–344.

Hilbe, J. (2009). *Logistic Regression Models*. Texts in statistical science. Chapman & Hall, 1st ed.

Hosmer, D., & Lemeshow, S. (1980). Goodness of fit tests for multiple logistic regression model. *Communications in Statistics*, *9*(10), 1043–1069.

Hosmer, D., & Lemeshow, S. (1982). A review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology*, *115*(1), 92–106.

Hosmer Jr, D., Lemeshow, S., & Sturdivant, R. (2013). *Applied Logistic Regression*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 3rd ed.

Jiang, J. (2001). A nonstandard χ^2-test with application to generalized linear model diagnostics. *Statistics and probability letters*, *53*(1), 101–109.

Jiang, J. (2007). Linear and generalized linear mixed models and their applications. *Springer Science and Business Media, LLC*.

Knypstra, S. (2008). Generalised linear models. `http://pyqrs.eu/sk/pyqrs/miscellaneous-texts/GLM.pdf`.

Kuk, A. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society*, *B57(2)*, 395 – 407.

Kutner, M. H., Neter, J., Nachtsheim, C. J., & Li, W. (2005). *Applied Linear Statistical Models*. Irwin, 5 ed.

Lin, X., & Breslow, N. (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. (1994). A note on gauss-hermite quadrature. *Biometrika*, *81*, 624–629.

Lugga, J. (2012). *Analysis of a Binary Response: An Application to Entrepreneurship Success in South Sudan*. Master's thesis, School of Mathematics, Statistics and Computer Science: University of Kwa Zulu Natal.

Malan, M. (2014). Concern over rate of teen depression and suicide.

McCulloch, C., & Searle, S. (2001). Generalized,linear and mixed models. *John Wiley & Sons,Inc*.

McCulloch, P., & Nelder, J. (1989). Generalized linear models. *Chapman & Hall,London*.

McCulloch, P., & Neuhaus, J. (2005). Generalized linear mixed models. *John Wiley & Sons,Inc*.

Moeti, A. (2007). *Factors Affecting the Health Status of the People of Lesotho*. Master's thesis, School of Statistics and Actuarial Science: University of KwaZulu Natal Pietermaritzburg.

Molenberghs, G., & Verbeke, G. (2005). Models for discrete longitudinal data. *Springer-Verlag: New York.*.

National Institue of Mental Health (NIH) (2007). History of childhood abuse or neglect increases risk of major depression. `http://www.nimh.nih.gov/news/science-news/2007/history-of-child-abuse-or-neglect-increases-risk-of-major-depression.shtml`. [Online; accessed September-2019].

Nelder, J., & Wedderburn, R. (1972). Generalized linear models. *Journal of the royal statistical society*, *A135*, 370–384.

Nglazi, M., Joubert, J., Stein, D., et al. (2016). Epidemiology of major depressive disorder in South Africa (1997-2015):a systematic review protocol. *BMJ Open*, *6*, 117 – 149.

Pandey, M., Kulkarni, V., & Gaiha, R. (2017). South Africa: Aging, Depression and Disease in South Africa. `http://www.allafrica.com/stories/201702210374.html`. [Online; accessed February-2019].

Piccinelli, M., & Wilkinson, G. (2000). Gender differences in depression. *The British Journal of Psychiatry*, *177*(6), 486–492.

Pinheiro, J., & Bates, D. (1995). Approximation of the log-likelihood function in the non-linear mixed effects model. *Journal of Computational and Graphical Statistics*, *4*, 12–35.

Pratt, L. A. (2014). Depression in the U.S. Household Population, 2009-2012. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

Ravishanker, N., & Dey, D. K. (2001). *A first course in linear model theory*. CRC Press.

Rencher, A., & Schaaljie, G. (2008). *Linear Models in Statistics*. John Wiley & Sons, Inc.

Roberts, G., Rao, N., & Kumar, D. (1987). Logistic regression analysis of sample survey data. *Biometrika*, *1*, 1–12.

SAS Institute Inc. (2013). *Exploring SAS Enterprise Miner: Special Collection*. Cary, NC: SAS Institute Inc.

Searle, S., Casella, G., & McCullocg, C. (2006). Variance components. *John Wiley & Sons,Inc*.

Self, S. G., & Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82(398)*, 605 – 610.

Smith, H. R. (2015). Depression in cancer patients: Pathogenesis, implications and treatment. *Oncology letters*, *9*(4), 1509–1514.

Smith, K. (2017). Substance abuse and depression. `http://www.psycom.net/depression-substance-abuse`. [Online; accessed September-2019].

Starkowitz, M. (2013). *African traditional healers understanding of depression as a mental illness: implications for social work practice*. Ph.D. thesis, University of Pretoria.

Statistics South Africa (2011). 2011 census.

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

Sullivan, P. F., Neale, M. C., & Kendler, K. S. (2000). Genetic epidemiology of major depression: review and meta-analysis. *American Journal of Psychiatry*, *157*(10), 1552–1562.

Tan, P.-N., Steinbach, M., & Kumar, V. (2014). *Introduction to Data Mining*. Pearson Education Ltd, 1 ed.

Tomlinson, M., Grimsrud, A. T., Stein, D. J., Williams, D. R., & Myer, L. (2009). The epidemiology of major depression in South Africa: Results from the South African Stress and Health study. *South African Medical Journal*, *99*, 367–373.

UCLA: Statistical Consulting Group (2019). PROC LOGISTIC—SAS ANNOTATED OUTPUT. `https://stats.idre.ucla.edu/sas/output/proc-logistic/`. [Online; accessed November-2019].

Vittinghoff, E., Glidden, D., Shiboski, S., & McCulloch, C. (2011). *Regression methods in biostatistics: linear, logistic, survival and repeated measures models*. Springer Science & Business Media.

Wedderburn, R. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, *61*, 439–447.

WHO (2012). Depression. `http://www.who.int/mental_health/management/depression/who_paper_depression_wfmh_2012.pdf`. [Online; accessed January-2019].

World Health Organization (WHO) (2017). Depression. `http://www.who.int/topics/depression/en/`. [Online; accessed January-2019].

Wu, Z. (2005). Generalized linear models in family studies. *Journal of marriage & family*, *67*(4), 22 – 31.

Zhang, D., & Lin, X. (2008). Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. *Springer New York*, *192*, 19 – 361.

Appendix A

Probability Table for the Bayesian Network

Table A.1: Probability Table

Parent Node	Parent Condition	Child Node	Child Condition	Value
Health status	Excellent	Depression	Yes	0.00153
Health status	Excellent	Depression	No	0.99847
Health status	Very good	Depression	Yes	0.00143
Health status	Very good	Depression	No	0.99857
Health status	Good	Depression	Yes	0.00317
Health status	Good	Depression	No	0.99690
Health status	Fair	Depression	Yes	0.02153
Health status	Fair	Depression	No	0.97847
Health status	Poor	Depression	Yes	0.03457
Health status	Poor	Depression	No	0.96543
Depression	Yes	Age group	15-19 years	0.01563
Depression	Yes	Age group	20-24 years	0.07813
Depression	Yes	Age group	25-29 years	0.10938
Depression	Yes	Age group	30-34 years	0.21875
Depression	Yes	Age group	35-39 years	0.15625
Depression	Yes	Age group	40-44 years	0.25000
Depression	Yes	Age group	45-49 years	0.17188
Depression	No	Age group	15-19 years	0.31580
Depression	No	Age group	20-24 years	0.25193
Depression	No	Age group	25-29 years	0.18096
Depression	No	Age group	30-34 years	0.11535
Depression	No	Age group	35-39 years	0.06661
Depression	No	Age group	40-44 years	0.04345
Depression	No	Age group	45-49 years	0.02590
Relationship to head	Head	Age group	15-19 years	0.01398
Relationship to head	Head	Age group	20-24 years	0.06605
Relationship to head	Head	Age group	25-29 years	0.13920
Relationship to head	Head	Age group	30-34 years	0.19931
Relationship to head	Head	Age group	35-39 years	0.19159
Relationship to head	Head	Age group	40-44 years	0.19827

Continued on next page

Table A.1 – *Continued from previous page*

Parent Node	Parent Condition	Child Node	Child Condition	Value
Relationship to head	Head	Age group	45-49 years	0.19159
Relationship to head	Spouse/Partner	Age group	15-19 years	0.01313
Relationship to head	Spouse/Partner	Age group	20-24 years	0.08192
Relationship to head	Spouse/Partner	Age group	25-29 years	0.14827
Relationship to head	Spouse/Partner	Age group	30-34 years	0.21439
Relationship to head	Spouse/Partner	Age group	35-39 years	0.20321
Relationship to head	Spouse/Partner	Age group	40-44 years	0.17647
Relationship to head	Spouse/Partner	Age group	45-49 years	0.16262
Relationship to head	Other	Age group	15-19 years	0.31580
Relationship to head	Other	Age group	20-24 years	0.25193
Relationship to head	Other	Age group	25-29 years	0.18096
Relationship to head	Other	Age group	30-34 years	0.11535
Relationship to head	Other	Age group	35-39 years	0.06661
Relationship to head	Other	Age group	40-44 years	0.04345
Relationship to head	Other	Age group	45-49 years	0.02590
Depression	Yes	Disability	No	0.68630
Depression	Yes	Disability	Yes	0.31370
Depression	No	Disability	No	0.94090
Depression	No	Disability	Yes	0.05910
Depression	Yes	Gender	Male	0.22826
Depression	Yes	Gender	Female	0.77174
Depression	No	Gender	Male	0.47885
Depression	No	Gender	Female	0.52115
Substance Abuse	Yes	Gender	Male	0.79167
Substance Abuse	Yes	Gender	Female	0.20833
Substance Abuse	No	Gender	Male	0.47885
Substance Abuse	No	Gender	Female	0.52115
Depression	Yes	Substance Abuse	Yes	0.10784
Depression	Yes	Substance Abuse	No	0.89216
Depression	No	Substance Abuse	Yes	0.00158
Depression	No	Substance Abuse	No	0.99842
Depression	Yes	Relationship to head	Head	0.56604
Depression	Yes	Relationship to head	Spouse/Partner	0.24528
Depression	Yes	Relationship to head	Other	0.18868
Depression	No	Relationship to head	Head	0.22440
Depression	No	Relationship to head	Spouse/Partner	0.24470
Depression	No	Relationship to head	Other	0.53090
Disability	No	Relationship to head	Head	0.54545
Disability	No	Relationship to head	Spouse/Partner	0.04545
Disability	No	Relationship to head	Other	0.40909
Disability	Yes	Relationship to head	Head	0.50000
Disability	Yes	Relationship to head	Spouse/Partner	0.32143
Disability	Yes	Relationship to head	Other	0.17857
Gender	Male	Relationship to head	Head	0.55556
Gender	Male	Relationship to head	Spouse/Partner	0.22222

Continued on next page

Table A.1 – *Continued from previous page*

Parent Node	Parent Condition	Child Node	Child Condition	Value
Gender	Male	Relationship to head	Other	0.22222
Gender	Female	Relationship to head	Head	0.22442
Gender	Female	Relationship to head	Spouse/Partner	0.24466
Gender	Female	Relationship to head	Other	0.53091

Appendix B

SAS Codes

```
*———————————— FINAL SLR MODEl————————————*
proc surveylogistic data =SAGHS_2016;
class
    MaritalStatus SubstanceAbuse Disability EduLevel (ref='Higher Education')
    Gender RelationHH (ref='Head') TypeRes HIV Race
    HealthStatus (ref='Poor') WaterSource (ref='Tap into household')
    Province ToiletFacility(ref='Flush Toilet') / param=reference ;
model Depression (descending) = Age Child5yr Hholdsz MaritalStatus EduLevel|Gender
                    Province RelationHH HIV TypeRes HealthStatus
                    Race Disability|SubstanceAbuse WaterSource ToiletFacility ;
weight person_wgt;
OUTPUT OUT=SLR_predicted pred=p ;
run;

*————————————Final GLMM ————————————*

proc glimmix data=SAGHS_2016 method=laplace;
class
    MaritalStatus SubstanceAbuse Disability EduLevel (ref='Higher Education')
    Gender RelationHH (ref='Head') TypeRes HIV Race
    HealthStatus (ref='Poor') WaterSource (ref='Tap into household')
    Province ToiletFacility(ref='Flush Toilet') ;
model Depression (descending) = Age Child5yr Hholdsz MaritalStatus EduLevel|Gender
                    Province RelationHH HIV TypeRes HealthStatus
                    Race Disability|SubstanceAbuse WaterSource
                    ToiletFacility / link=logit dist=binary oddsratio solution ;
random int / subject = PSU type=VC ;
covtest zerog;
OUTPUT OUT=glmm_predicted pred=p predicted(ilink) ;
run;
```

```
*————————————————BAYESIAN NETWORK MODEL————————————————*

proc hpbnet data=GHS_2016;
    target Depression;
    input Age_group
          Gender
          Race
          Health_Status
          Highest_education_level
          Marital_Status
          Medical_Aid_Cover
          employ_status
          disability
          social_grants
          HIV
          Substance_Abuse
          Cellphone_ownership
          relationship_to_head
          Electricity
          waterAccess
          ToiletFacility
          Province
          TypeRes
          HouseholdSize
          Death_of_Household_Member
          Child_Under_5 /level=NOM

output network=network varselect=varselect;
run;

proc print data=network noobs label;
var _parentnode_ _childnode_;
where _type_="STRUCTURE";
run;

proc print data=network noobs label;
var _parentnode_ _parentcond_ _childnode_ _childcond_ _value_;
where _type_="PROBABILITY";
run;

proc print data=varselect noobs label;
run;
```