Modelling Tuberculosis Risk Factors Among Adult Men in South Africa

UNIVERSITY OF
KWAZULU - NATAL

INYUVESI
YAKWAZULU-NATALI

Muziwandile Nhlakanipho Mlondo

December, 2021

Modelling Tuberculosis Risk Factors Among Adult Men in South Africa

By

Muziwandile Nhlakanipho Mlondo

Submitted to the University of KwaZulu-Natal in

fulfillment of the academic requirements for the degree

of

MASTER OF SCIENCE

in

STATISTICS

under the supervision of

Prof. Sileshi Fanta Melesse

And

Co-supervisor Prof. Henry G Mwambi

UNIVERSITY OF KWAZULU-NATAL

SCHOOL OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE

PIETERMARITZBURG CAMPUS, SOUTH AFRICA

Declaration

I, Muziwandile Nhlakanipho Mlondo, declare that this dissertation titled 'Modelling Tuberculosis Risk Factors Among Adult Men in South Africa' and the work presented in it are my own original work. I confirm that:

- This thesis has not been submitted for any degree or examination at any other university.
- This thesis does not contain others persons' data, graphs or other information, unless specifically acknowledged as being sourced from other persons.
- This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources haven 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.
- 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.

[redacted]	14/12/2021
Muziwandile Nhlakanipho Mlondo	Date
[redacted]	10/02/2022
Prof Sileshi Fanta Melesse	Date
[redacted]	10/02/2022
Prof Henry Mwambi	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.

Dedication

To my late grandmother, father & mother, Thabisile Alice Mlondo, Mduduzi Comfort Mlondo, and Thembani Purity Nkosi.

Acknowledgments

First and foremost, I would like to thank God for the gift of life and the strength to complete this thesis.

I am so fortunate and grateful to all those I had the privilege of working with during Honours and my master's degree. I would especially like to thank Prof. Sileshi F. Melesse, the main supervisor of this work, for his constructive suggestion and comment on my work. I would like to appreciate my co-supervisor, Prof H. Mwambi. I would also like to appreciate Mr. Ashenafi Yirga for his contribution. I appreciate all their contribution of time, and insight, making this thesis possible.

I dedicate the work of this thesis to the loving memory of my late grandmother, Thabisile Alice Mlondo, who passed away in June 2020, and my father, Mduduzi Comfort Mlondo, who passed away in September 2019. Thank you for always believing in me and encouraging me to be the best person I can be. My desire is to make you proud, you motivated me to complete this research.

My deepest gratitude to my family for their supporting and believing in me. I greatly appreciate my mother, Thembani Nkosi, and my cousin Lungi Sithole for your motivation.

Last but not least, I would like to thank all my friends Talani Mhelembe, Thandi Mbuyazi, Thandokazi Jantjies, Thembelani Ntshele, and Mthunzi Sithole for all your advice, support, and encouragement.

Published Papers

The following papers are under review for publication from this thesis.

1. Risk factors associated with tuberculosis among men; a study of South Africa. *Under review*

2. Semi-parametric model to study the risk factors of tuberculosis among adult men in South Africa. *Under review*

Abstract

Tuberculosis is among the major public health problems not only in South Africa but worldwide. Tuberculosis is an underlying cause of more than 1.5 million deaths each year worldwide, making it the world's top infectious killer. There are more cases for men than women. Such a heavy burden requires an understanding of the tuberculosis status of the people, especially among men, and associated risk factors. Therefore, this study uses some statistical methods that are suitable to estimate the effect of the risk factors associated with tuberculosis among adult men. The study used the 2016 South African Demographic and Health Survey data.

The Generalized Linear Models, such as the binary logistic regression model that assumes a simple random sampling as a sampling method followed by survey logistics that incorporate the complex design by means of robust standard errors of estimates, were applied to the data. The findings revealed that models that account for complex design are more suitable than those that do not account for complexity. To account for variability between the primary sampling units generalized linear mixed model was then used. GLMMs accounts for correlation within clusters by means of random effects which also account for cluster to cluster heterogeneity. Further, a generalized additive mixed-effect model was used to fit nonlinear and non-normal data; the categorical variables were modeled parametrically and continuously by non-parametric models. The thesis also discussed limitations for each of these models.

The findings from this study revealed that the risk factors of tuberculosis are: any chronic disease, current age, region, race, number of times away from home, marital status, weight, and interaction effect of chronic disease and age, the interaction effect of smoking status and number of household members.

Contents

Table of figure

List of Tables

Acronyms/Abbreviations

BCG – Bacille Calmette-Guerin

BMI – Body Mass Index

GHS - General Household Survey

HIV – Human Immunodeficiency Virus

MUAC – Mid-Upper Arm Circumference

SADHS - South African Demographic and Health Survey

TB – Tuberculosis

Chapter One

INTRODUCTION

1.1 Background

One of the public health threats that remain in all countries is tuberculosis. An essential factor for human well-being is healthy organs. At the same time, tuberculosis is a transmissible disease that frequently strikes the lungs. However, it can also diffuse to other organs, such as the brain and spine (Hussain, 2020). A type of bacteria called Mycobacterium causes tuberculosis, which spreads through the air.

Mycobacterium tuberculosis is spread through the air in 1–5 micron-sized droplet nuclei. Coughing, sneezing, shouting, or singing produce infectious droplet nuclei in people with pulmonary or laryngeal tuberculosis. These microscopic particles can stay in the air for several hours, depending on the surroundings. Mycobacterium tuberculosis is spread through the air, not through contact with a surface. When a person inhales M. tuberculosis-containing droplet nuclei, the droplet nuclei travel through the mouth or nasal passages, upper respiratory tract, and bronchi to reach the lungs' alveoli.

However, being infected by the tuberculosis bacteria does not always mean you will get sick. The disease has two different forms, which are: Latent Tuberculosis and Active Tuberculosis. Latent TB is when someone has the bacteria, but their immune system prevents the bacteria from spreading. The infection is still alive but not active, thus not affecting the person or person who is not sick and can be active someday. Active TB is when someone has bacteria that can multiply (spreads) and attack their organs. TB can affect anyone anywhere, but the literature shows that it

is more noticeable among adult men than women (WHO, 2018). In this study, we will be looking at which factors affect the spread of tuberculosis and how to mitigate the spread in order to reduce the number of persons developing TB, thus the number of deaths.

WHO (2020) reported that ten million people contract tuberculosis (TB), even though the disease is preventable and curable. The death rate of tuberculosis is 1.5 million per year, making it the most virulent disease on earth (WHO, 2020). TB is the main opportunistic infection for individuals with Human immunodeficiency virus. Also, WHO examined that about a quarter of the population globally is believed to be infected with latent TB (WHO, 2018). Only 5-15% of people with latent tuberculosis will get sick with active tuberculosis disease. Without the legitimate treatment of tuberculosis, there is a high mortality rate of most HIV-positive individuals with tuberculosis. About two-third of new TB cases in 2019 resided in these eight countries: Pakistan, Bangladesh, India, Nigeria, Philippines, China, and South Africa, which are low-and middle-income countries (WHO, 2020).

"The clock is Ticking - let find, treat and end TB now"- 2021 Campaign theme.

Insights (StatsSA, 2013) has released a report dealing with mortality and the cause of death in South Africa. This is based on data obtained from deaths in 2010, which were recorded at Home Affairs department. The total number of deaths went down by 6.3 % in 2010 compared to 2009. Data shows that more males than females died due to tuberculosis, and the highest number of deaths recorded were among the age group 30-39 years. TB was the leading cause of death in South Africa; about 12% of death occurred in 2010 (StatsSA, 2013).

General household survey (GHS 2011) results showed that 2.9% of tuberculosis sufferers said they were sick a month before the survey was conducted (Lehohla, 2013). The age group, which is mostly affected by TB in South Africa were 25-64 years. The data showed that there were more black African compared to the other

race group who were sick with TB. Most sick or injured people before the survey and who suffered from TB reside in KwaZulu-Natal. GHS 2019 report showed that South Africa accounts 3% of cases globally.

In recent studies and statistics, they show that the disease cases are slowing down but not fast enough. WHO (2020) reported that the TB rate is falling at about 2% per year, and the health targets of the United Nations Sustainable Development Goals (SDGs) are to end the TB epidemic by 2030. On the other hand, the research shows that the number of tuberculosis deaths in 2019 reached 1.4 million (208,000 with HIV). Ten million people fell sick with tuberculosis worldwide in 2019, 5.6 million were men, 3.2 million were women, and 1.2 million were children (WHO, 2020).

The critical problems with tuberculosis are that it is not easy to diagnose the patient fast enough and treat them before spreading the germs to the communities. Another problem is to control the spread in the public areas and public transport. This study aims to help determine the relationship of the factors affecting the spread and determine the high-risk factors for tuberculosis.

1.2 Objectives of the Study

1.2.1 General Objectives

The overall purpose of this study is to investigate the factors affecting the risk of developing tuberculosis in South Africa.

1.2.2 The Main Purpose of this study

To fit some statistical models to data that considers the sampling procedure and determine factors associated with the risk of developing tuberculosis among South

African men using the 2016 South African Demographic and Health Survey (SADHS) data.

1.3 Importance of the Study

The introduction section of this study reveals that tuberculosis remains a worldwide public health threat causing more than a million deaths, and some people still suffer from tuberculosis each year. Tuberculosis has been a leading cause of death in South Africa since 2010 (Lehohla, 2013). This suggests that there should be interventions by health to help decrease the risk of developing tuberculosis. Therefore, studying factors affecting the risk of developing tuberculosis in South Africa is very important. Furthermore, it will assist in understanding areas needed to focus on to reduce the risk of developing tuberculosis. Moreover, this research will focus on men since the literature suggest that the gender that is more likely to have tuberculosis is male.

1.4 Literature Review

Tuberculosis development is a two-stage process in which a vulnerable person who is exposed to an infectious tuberculosis case becomes infected and may later acquire the disease based on a variety of factors (Lienhardt, 2001). Tuberculosis has long been considered a poverty illness, and several features of low socioeconomic status (SES), such as overcrowding, hunger, and a household level, are risk factors for the disease (Dubos & Dubos, 1987).

(Khwarwadkar, et al., 2022) conducted a study to investigate epidemiological and prediction model studies that explore how climate change may affect the risk factors for TB. They found a positive association between climate change and TB risk

factors which were HIV, diabetes, undernutrition, overcrowding, poverty, and indoor air pollution.

Coker et al. (2006) conducted a case-control study to investigate risk factors for pulmonary TB. Their primary outcomes measured the determinates related to the development of TB in a large city in Russia. They took all adults with pulmonary tuberculosis as their cases. Then they randomly sampled the controls; they had no history of TB. They found that poverty, living with a relative with tuberculosis, unemployment, drinking unpasteurized milk, living in overcrowded conditions, diabetes, and prison were independently associated with an increased risk of developing tuberculosis. As a result, these variables raise the chance of infection. Similar research was investigated by Narasimhan et al. (2013). They also look at the risk factors for tuberculosis, from exposure to the tuberculosis bacilli to the development of the active disease. The relative risk, prevalence (cohort study) was used in looking at the risk factors associated with the development of tuberculosis. They found that the relative risk of TB disease is 8.3 (6.1-10.8) higher for HIV-infected patients than persons without HIV with a 1.1% prevalence; thus, HIV infection remains the risk factor of TB. Also, the relative risk of TB disease is 4.0 (2.0-6.0) higher for children with malnutrition, 3.0 (1.5-7.8) higher for persons with diabetes compared without diabetes, 2.9 (1.9-4.3) higher for people who drink alcohol > 40g/day, 2.6 (1.6-4.3) higher for smokers than non-smokers, and 1.5 (1.2-2.3) higher for indoor pollution (ref). Thus, all these risk factors accentuate the progression of tuberculosis infection. HIV coinfection is the most critical risk factor for tuberculosis infection, and the disease becomes the deadliest. There is a high rate of smoking and diabetes among men in South Africa, so the risk of tuberculosis progression increases.

The case-control investigation for the risk factors of TB in adults in King County, Washington, was conducted by Buskin et al. (1994). The variables they used for their study include age, ethnicity, gender, socioeconomic factors (type of residence

and years of education), history of various medical conditions (HIV infection, silicosis, cancer, pancreatitis, diabetes, partial gastrectomy, or antrectomy, etc.), reasons for prior hospitalization, height, weight, place of birth, smoking status, and drinking history (Buskin et al., 1994).

They found that tuberculosis risk increases with age; people in their 70s and older are more prone to acquire the disease than people in their twenties (Buskin et al., 1994). Males risk was greater than females; also, persons of color were more likely than whites to acquire tuberculosis. They also found that people born outside of the United States had higher chances of being infect with TB than those born in the United States (Buskin et al., 1994). Relative to persons with higher income, more stable living situations, and postsecondary education, person's lowest SES were more likely to develop TB (Buskin et al., 1994).

Heavy drinkers are twice as likely as those with no alcohol consumption to have tuberculosis, whereas moderate drinking has no association (Buskin et al., 1994). Current and former smokers had a higher chance to acquire TB than nonsmokers (Buskin et al., 1994). They also found that the history of any underlying medical condition has an association with an increase in the risk of tuberculosis. They believed that targeting the groups with a higher risk of development can be an effective way to reduce the rate of tuberculosis. This has an effect of reducing the reproduction number of the disease.

Narayanan et al. (2007) conducted similar research using a cross-sectional survey method in South India to measure the independent association of risk factors age, sex, smoking, and alcohol with tuberculosis in the form of prevalence odds ratio (POR). They found that risk factors age, sex, smoking, and alcohol are all independently associated with pulmonary tuberculosis. These factors play a role in the progression of tuberculosis infection to disease. Age and sex have a strong association compared to smoking and alcoholism. They also believe that the

Directly Observed Therapy (DOT) strategy for treating tuberculosis controls the spread of the infection but does not affect the current rate of incidence.

A similar case-control study was done by Lienhardt et al. (2005). They also investigated the risk factors for tuberculosis in three West Africa counties (Guinee, Guinea Bissau, and Gambia). They examined the host-related factors (includes: sex, marital status, former TB, family history of TB, smoking, alcohol, drug, BCG scar, HIV status, history of worms, treatment of worms, history of asthma, treatment of asthma, diabetes and hemoglobin, environmental factors (includes: demographic and household-related factors, and socioeconomic factors) they used univariate analysis to analyze host-related factors and environmental factors.

Based on the host-related factors, they found that males, widowed/divorced, and single persons had a high risk of acquiring tuberculosis (Lienhardt et al., 2005). A previous case of TB in a family member enormously increases the risk. Risk of TB increase with smoking. Alcohol and drug abuse had a significant association with the development of tuberculosis; someone drinking alcohol had higher chances for risk of TB compared to a non-alcoholic. Drug use has a higher risk of tuberculosis. Drinking tea among adults has no association with the risk of tuberculosis. While on the other hand presence of BCG scar shows to reduce the risk of developing tuberculosis. HIV and, history of worm infection have been associated with a higher risk of tuberculosis. History of asthma or treatment of asthma had no association with the risk of tuberculosis. Diabetes and anemia had an association with tuberculosis. Diabetes and anemia were found to increase the risk of tuberculosis.

They recommended improving tuberculosis control to identify specific targets, such as enhanced higher education on TB and its risk factors.

Shetty et al. (2006) conducted a case-control study to evaluate the risk factors of TB in South India. They developed a logistic regression model that focused on the potential socio-demographic risk factors for tuberculosis. The variables that they included in their study were categorical variables which are marital status, religion,

education, employer, occupation, household size, household income, house possessions, single-roomed business, people per room, separate kitchen, smoking, alcohol, chronic disease, TB contact, BMI, and MUAC.

Most of these factors had no association with tuberculosis in univariate analysis. The only few variables included in the multivariable analysis were education, household income, persons per room, separate kitchen, alcohol, smoking, and chronic disease. Unlike most other researchers in the literature, they found from a multivariate logistic regression model that alcohol consumption and smoking have no association with TB, contradicting the other researchers.

The only significant factors were education, chronic disease, and a separate kitchen. Compared to those with no education, people with higher education are shown to have a reduced the risk of developing tuberculosis. Persons with at least one of the chronic diseases have higher chances of developing TB. At the same time, those without a separate kitchen have higher chances of developing tuberculosis compared to persons with a separate kitchen (Tekkel, et al., 2002).

Tekkel et al. (2002) conducted a case-control study to determine the risk factors for pulmonary tuberculosis in Estonia. Data were analyzed using a logistic regression model to obtain odds ratio and confidence interval. They included various demographic, socio-economic, and household characteristics such as age, place of residence, place of birth, ethnicity, marital status, education, income, previous years economic situation, occupation, unemployed, previous imprisonment, persons per room, heating in the last place of residence, central supply water, and central sewerage system in their study. They also looked at the habits and exposure, including regular smoking, passive smoking, alcohol consumption, drug abuse, nutrition, weight loss, contact with factors hazardous to health, residing/ working with tuberculosis patients, and contact with TB patients elsewhere.

They revealed that the main determinants for tuberculosis were marital status, education level, low income, being in prison, not having own place of residence,

unemployment, current smoking, alcohol consumption, insufficient food, and contact with tuberculosis patients (Tekkel, et al., 2002). Place of birth was not a risk factor. The risk for tuberculosis decreased for overweight or obese persons. Unlike the other reports in the literature, the risk factors for pulmonary tuberculosis in Estonia (Tekkel, et al., 2002) were different. The majority of people at risk were young and middle-aged men given the complexity of TB, this might be due to working environmental conditions which put these men at higher risk, and drug abuse did not have an impact.

1.5 Outline of the Thesis

This particular investigation consists of six chapters. The introduction section gave some background about tuberculosis, the purpose of the study, the significance of studying tuberculosis, and some existing information on risk factors of tuberculosis by reviewing previous research that has been done. Chapter two covers the source, study variables, descriptive analysis for SADHS 2016. Chapter three gives an overview of Generalized Linear Models, including logistic regression models and survey logistic regression and their application to the data. Chapter four covers the generalized linear mixed model and its application to data. Chapter five covers the generalized additive mixed model and its application to the data. Chapter six gives a conclusion and discussion of the results from different statistical approaches used in this thesis.

Chapter Two

Exploratory data analysis

2.1 Introduction

For any analysis, it is important to summarize and understand the characteristics of the outcome variable and all possible exploratory variables. The exploratory analysis will be conducted in this chapter to explore the distributional properties of the variables. The 2016 South African Demographic and Health Survey was utilized for this investigation. The 2016 South Africa Demographic and Health Survey was implemented by Statistics South Africa (StatsSA) with the South African Medical Research Council (SAMRC), with the demand of the National Department of Health (NDoH). Survey data collection was conducted from 27 June 2016 to 4 November 2016. The 2016 SADHS primary objective was to deliver the latest estimates of fundamental demographic and health indicators to help policymakers and program managers to evaluate and design programs and plan strategies to improve the health of the citizens. The sampling frame used for the 2016 SADHS is the Statistics South Africa Master Sample Frame (MSF), which was conducted using Census 2011 enumeration areas (EAs). EAs of the manageable size were treated as primary sampling units (PSUs), whereas large EAs were split into conceptual PSUs (National Department of Health, 2019). Since SA comprises nine provinces, PSUs were used to ensure survey precision across regions (National Department of Health, 2019). Each region was stratified into the urban, farm, and traditional areas, resulting in 26 sampling strata (National Department of Health, 2019). A total of 750 PSUs were selected from 26 sampling strata, 468 selected PSUs in urban areas, 224 PSUs

in traditional areas, and 58 PSUs in farm areas (National Department of Health, 2019). South African Demographic and Health 2016 shows that adult men with tuberculosis are approximately 5.5% (National Department of Health, 2019).

2.2 Study Variables

2.2.1 Dependent variable

The response (dependent) variable is tuberculosis which is dichotomous (yes or no). The response variable is coded as "0" if a doctor or nurse has never told the respondent that they have tuberculosis, and "1" if the response was told that he has tuberculosis. The prevalence of tuberculosis among men is 5.29% in South Africa, this result is from the current data set.

2.2.2 Explanatory variable

Current age, weight, health, alcohol consumption, region, place of residence, ethnicity, wealth index, usual or visitor, marital status, working status, times away from home, smoking status, education level, and chronic disease conditions are included as covariates.

2.2.3 Variable creation

The variable chronic disease was created using these variables high blood pressure, heart attack, cancer, stroke, high blood cholesterol, diabetes, chronic bronchitis, and asthma. If the respondent has at least one of these diseases, then he is categorized as someone who has a chronic disease; otherwise, he has no chronic illness.

Table 2. 1 Variable description

Code	*Label*	*Description*
mv012	current age	age of respondent
mv101	region	1 = "Western Cape, 2=Eastern Cape, 3=Northern Cape, 4=Free State, 5=Kwazulu-Natal, 6=North West, 7=Gauteng, 8=Mpumalanga, 9=Limpopo
mv102	type of place of residence	1=urban, 2=rural
mv106	education level	0=no education, 1=primary, 2=secondary, 3=higher
mv131	Ethnicity	1=Black/African, 2=White, 3=Colored, 4=Indian/Asian, 999=Other
mv135	usual resident or visitor	1=usual resident, 2=visitor
Mv136	No. of household members	Integer greater then 0
mv464aa	smoking status	0=do not smoke, 1=everyday, 2= sometimes
sm916	alcohol	0=no, 1=yes
	chronic disease	0=no, 1=yes
sm901	health	1=poor, 2=average, 3=good, 4=excellent
sm902	weight	1=underweight, 2=normal, 3=overweight, 4=obese, 8=don't know
mv714	Working status	0=no, 1=yes
mv501	marital status	0=never in union, 1=married, 2=living with partner, 3=widowed, 4=divorced, 5=separated
mv190	wealth index	1=poorest, 2=poor, 3=middle, 4=richer, 5=richest
mv167	No. of times away from Home	Integer greater or equal to zero

Table 2. 2 Exploratory analysis

Characteristics	*Frequency*	*Percent(%)*
Type of place of residence		
Urban	2315	55.34
Rural	1868	44.66
Education level		
No education	283	6.77
Primary	782	18.69
Secondary	2729	65.24
Higher	389	9.30
Alcohol		
Yes	2552	61.01
No	1631	38.99
Chronic Disease		
Yes	3412	81.57
No	771	18.43
Health		
Poor	350	8.37
Average	1159	27.71
Good	2016	48.20
Excellent	658	15.73
Weight		
Underweight	460	0.11
Normal	3378	80.76
Overweight	294	7.03
Obese	12	0.29
Don't know	39	0.93
Working Status		
Yes	1744	41.69
No	2439	58.31
Marital Status		
Never in union	2279	54.48
Married	1204	28.78
Living with partner	414	9.90
Widowed	110	2.63
Divorced	56	1.34
Separated	120	2.87
Smoking Status		
Do not smoke	2634	62.97
Every day	1282	30.65
Someday	267	6.38
Wealth index		
Poorest	850	20.32
Poor	947	22.64
Middle	968	23.14
Richer	787	18.81
Richest	631	15.08
Ethnicity		
Black/African	3547	84.8
White	193	4.61
Coloured	379	9.06
Indian/Asian	62	1.48
Other	2	0.05
Region		

Western Cape	279	6.67
Eastern Cape	553	13.22
Northern Cape	353	8.44
Free States	384	9.18
Kwazulu-Natal	589	14.08
North West	504	12.05
Gauteng	464	11.09
Mpumalanga	511	12.22
Limpopo	546	13.05
Total	4183	100

The intention of this particular investigation is to determine the factors affecting tuberculosis in South Africa. The prevalence of adult men with tuberculosis in South Africa in 2016 was 5.29%. Table 2. 2 shows that 31.08% (n=1300) of men were aged 15-24. The respondents with age between 25-34, 35-44, 45-54, and above 55 years accounted for 22.14%, 16.04%, 11.69%, and 19.05% respectively. Thus, most adult men in SA were between 15 and 24 years old. The results also show that more men were from urban areas (55.34%) than rural areas. The majority of adult men in South Africa had only a secondary level of education (65.24%), and 6.77% of South African men had no education. Furthermore, more than half of men were unemployed (58.31%). We observe that majority of SA men consumed Alcohol (61.01%). Also, 81.57% had at least one chronic diseases. Table 2. 2 also shows that most men had good health (48,20%), only a few respondents claimed they had poor health (8.37%). Table 2. 2 suggests that most respondents had a normal weight (80,76%), and most men stayed at home (59.48%). The results show 54.48% never had in marriage union, only 28.78% were married, and 1.34% were divorced. We also observe that most respondents were non-smokers (62.97), only a few smoke someday (6.38%). Only a few respondents were the richest. A majority were in the middle wealth index with 23.14%. The majority of men were black/African (84.8%). The 14.08% of adult men are from Kwazulu-Natal, and Western Cape was less represented with only 6.67% men. Figure 2. 1 indicates that prevalence of TB was high among men with high blood pressure, followed by men with asthma. The

prevalence of tuberculosis among men who had diabetes and high blood cholesterol did not differ. Men who had cancer showed a lower prevalence of tuberculosis.

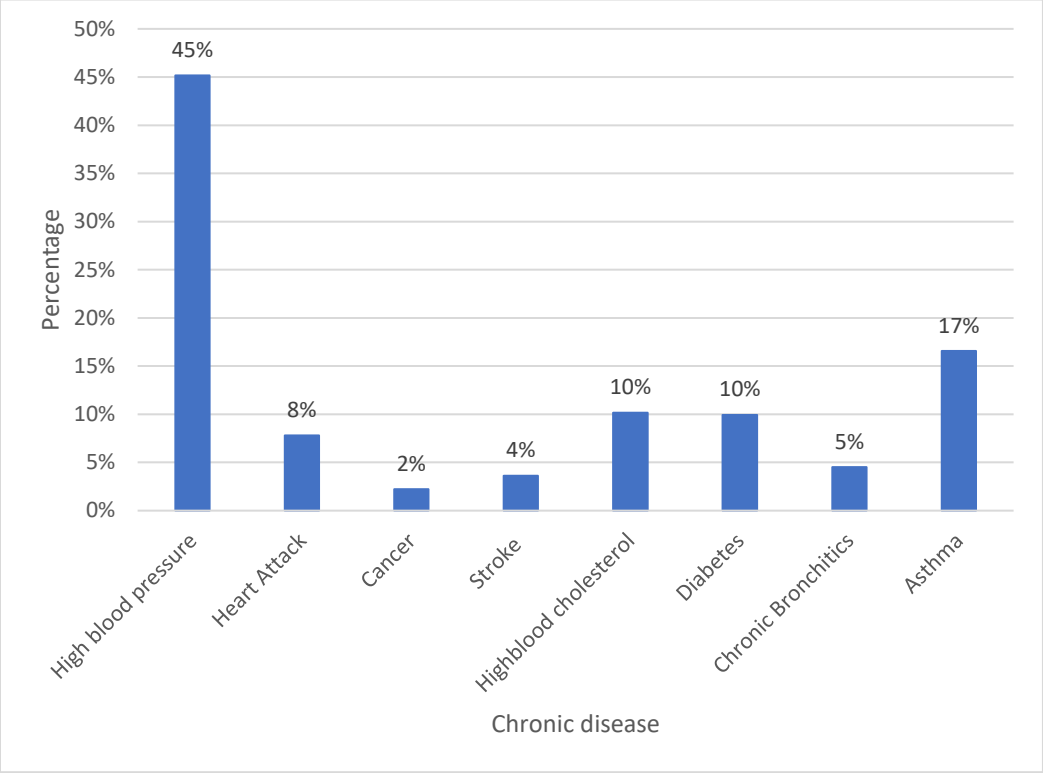

Figure 2. 1 Bar graph of men with chronic disease who have tuberculosis

Figure 2. 2 breaks down the risk of having tuberculosis according to the province of residence. The men from Eastern Cape had the highest percentage of having tuberculosis, 21.3%, followed by men from KwaZulu-Natal, 14.3%. Not much difference was observed in the percentage of having tuberculosis between Free State and Northern Cape province 12.5% and 12.1%, respectively. In contrast, men in North West and Western Cape had an equal percentage of 8.5% of having tuberculosis. The Gauteng province had the lowest percentage of men having tuberculosis. From Figure 2.3, the prevalence of having tuberculosis was highest for non-smokers, 56.3%. Followed by smoking every day, 37.5%, and smoking sometimes had the lowest prevalence of 6.3%. Therefore, this suggests that adult

men in South Africa who are non-smoker showed a high prevalence of TB than smokers.

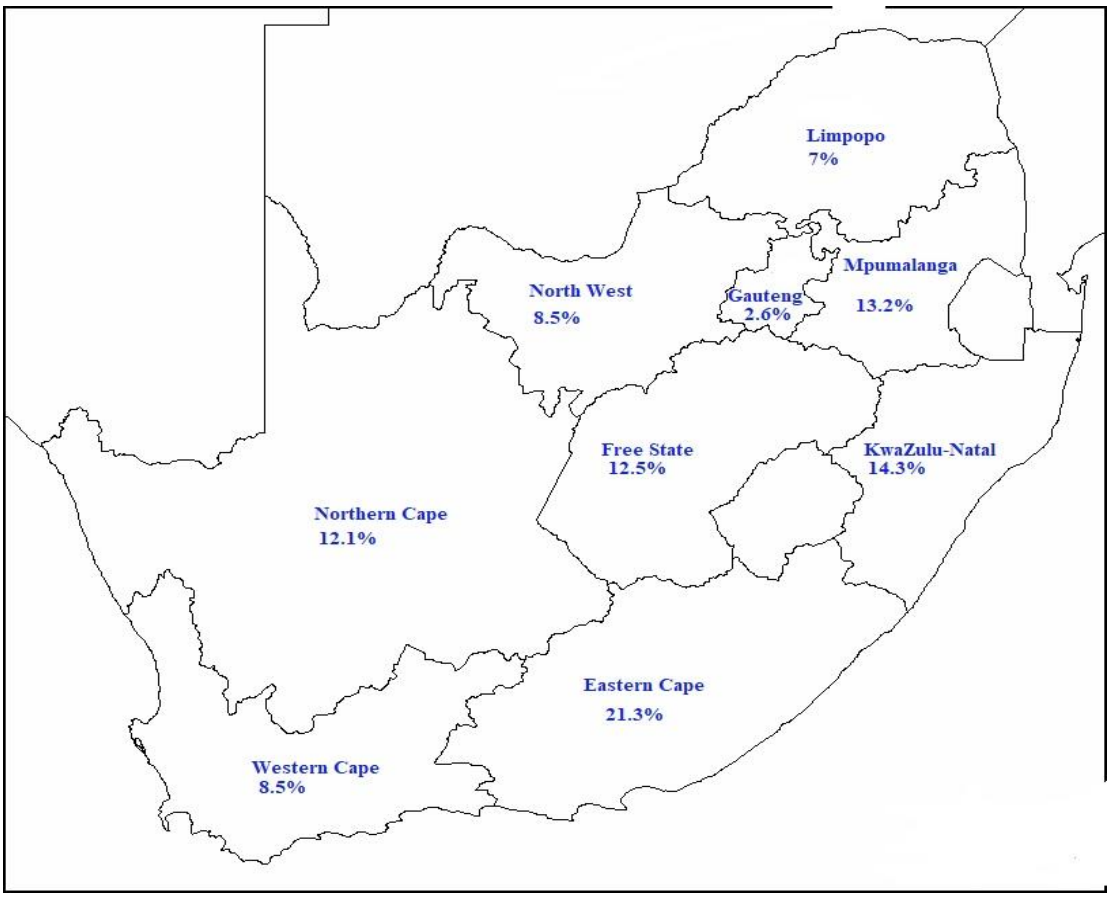

Figure 2. 2 Risk of tuberculosis according to the province

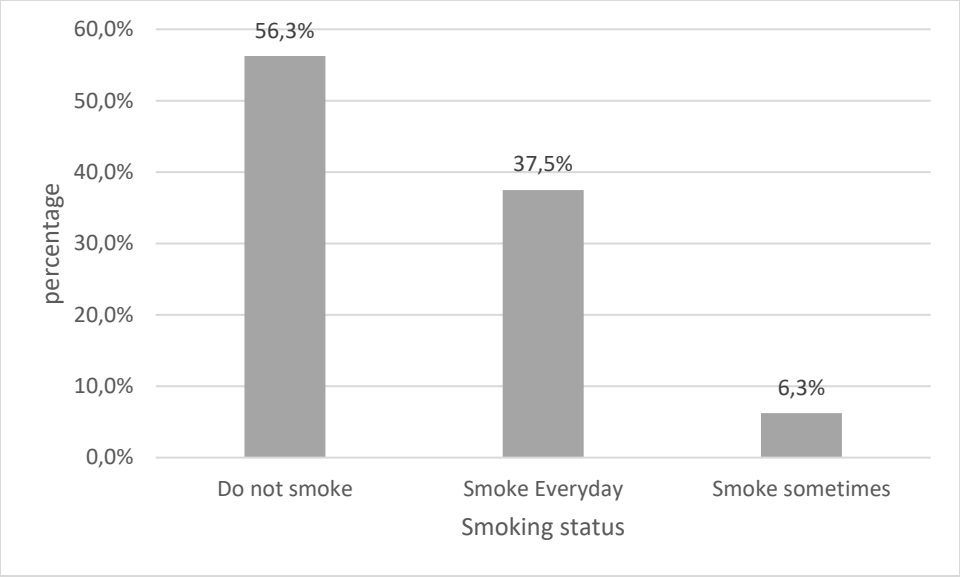

Figure 2. 3 The prevalence of TB according to smoking status

2.2.4 Chi-square Test of Association

A cross-tabulation technique was used to test for association between the explanatory variables and tuberculosis among males in the 2016 SADHS data. From Table 2. 3, we deduce that the predictor variables with a p-value less than 0.05 level of significance were significantly associated with the response variable (tuberculosis status). Table 2. 3 shows an association between tuberculosis status among men and current age, region, ethnicity, smoking status, chronic disease, health, weight, marital status, wealth index, number of times away from home, and education level. No association was found between tuberculosis and type of place of residence, usual resident or visitor, alcohol, and working status.

Table 2. 3 Cross-tabulation of tuberculosis and explanatory variables

Covariate	*DF*	*Chi-square*	*P-value*
Current age	4	84.4669	<.0001
Region	8	54.6532	<.0001
Type of place of residence	1	2.1163	0.1457
Ethnicity	4	14.8976	0.0049
Usual resident or visitor	1	0.1225	0.7264
Smoking status	2	6.5506	0.0378
Alcohol	1	1.0728	0.3003
Chronic Disease	1	24.9169	<.0001
Health	3	88.9494	<.0001
Weight	4	41.6510	<.0001
Working Status	1	0.6633	0.4154
Marital Status	5	50.4946	<.0001
Wealth Index	4	18.1604	0.0011
Times away from Home	1	118.3069	<.0001
Education Level	3	57.6805	<.0001

2.2.5 Summary

Exploratory analysis was used to describe the data. The chi-square technique allows for analyzing the association between predictors and the response variable before using model-based approaches. According to Table 2. 2, the descriptive analysis shows that most of the men were aged from 15-24, and also the majority of them were black/ African. There were slightly more men from urban than rural areas. Overall majority of men had normal weight with good health, but also most were unemployed. Ten out of fifteen factors were found to be associated with the presence of tuberculosis in men. Furthermore, the highest prevalence of TB is amongst men with High blood pressure, men from Eastern Cape, and non-smokers.

Chapter Three

Generalized Linear Models

3.1 Introduction

The primary goal of this study, as stated in Chapter One, is to use the 2016 South African DHS data to identify variables associated with the chance of contracting tuberculosis. A model can mathematically describe the relationship between a response variable and a set of explanatory variables. The general linear model (GLM) assumes that the response variable follows a normal distribution and linear independence with constant variance. While the Generalized Linear Models are the extension or generalization of a GLM that permits the response variable to have any other distribution other than the normal distribution, but that distribution should belong to the exponential family. The Generalized Linear Model includes logistic regression for a binary dependent variable, exponential and gamma models for survival times, multiple regression for a continuous response, log-linear models for categorical data, Poisson regression or negative binomial regression for count data, and gamma regression for variance models (Olsson, 2002). Logistic regression is a famous mathematical modeling method commonly used to model a dichotomous disease outcome (Kleinbaum et al., 2002).

3.1.1 The Linear Model

The Generalized Linear Model is a generalization of the linear model. given by:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}, \quad (3.1)$$

where **y** is the response variable, $\boldsymbol{X}$ a design matrix of covariates, $\boldsymbol{\beta}$ is the vector of coefficients, and $\boldsymbol{e}$ is the vector of error terms.

Let $\eta = \boldsymbol{X\beta}$ denote the linear predictor part of the model. Instead of modeling the mean directly as a function of the linear predictor $\boldsymbol{X\beta}$, the model is specified by function g (μ), so the model becomes

$$g(\mu) = \boldsymbol{\eta} = \boldsymbol{X\beta} \quad (3.2)$$

$g(.)$ is the link function.

3.1.2 Exponential Family

The exponential family is the set of distributions that include discrete and continuous random variables. It can be written in the following format.

$$f_y(\boldsymbol{y}; \boldsymbol{\theta}, \phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right], \quad (3.3)$$

where θ_i is known as a conical parameter, ϕ is the dispersion parameter, $a(\phi)$ and $b(\theta_i)$ are known functions and $c(y_i, \phi)$ is a function of y_i. The mean μ = E(y) = $b'(\theta)$ and variance, $var(y) = \phi b''(\theta)$ (McCullagh & Nelder, 1983).

3.1.3 Components of Generalized Linear Models

The Generalized Linear Model is specified by using three components (Agresti, 1990):

- ***Random component***

Identifies the probability distribution of the response.

- ***Systematic component***

This component that relates a vector $\boldsymbol{\eta} = (\eta_1, \ldots . \eta_N)'$ to set of explanatory variables through a linear predictor model.

$$\boldsymbol{\eta} = \boldsymbol{X\beta} \; or \; \eta_i = \beta_0 + \beta_1 x_{1i} + \ \ldots + \beta_p x_{pi} \qquad (3.4)$$

- ***Link function***

The link function, given by $g(\mu_i)$ is a monotonic differentiable function that describes the link between random and systematic components. Thus, GLM is defined as:

$$g(\mu_i) = \ \eta_i = \ \sum_j \beta_j x_{ij}, \ \ i = 1, \ldots N \qquad (3.5)$$

The function $g(\mu) = \mu$ gives the identity link where $\eta_i = \ \mu_i$ specifying a linear model for the mean response.

3.1.4 Odds ratio

Probability refers to the chance that an event will occur. Probability is some number ranging from 0 to 1 (Melesse et al., 2016). Probability is the ratio that compares the favorable outcome of an event to the total number of outcomes (favorable + unfavorable). Odds are the ratio of the probability of one outcome to another (Power & Xie, 1999). There are two ways to define odds: Odds in favor of a certain event and odds against a specific event. The ratio of favorable outcomes gives odds in favor of a specific event to the number of unfavorable effects. The odds in favor of an occurrence are a ratio of the likelihood of the event occurring to the probability

of the event failing. Odds against a specific event are calculated by dividing the number of unfavorable events by the number of positive occurrences (Grima, 1965). The odds are constrained between zero to infinity. The association between odds and probabilities can be defined as (Olsson, 2002):

$$Odds\ in\ favor\ of\ the\ event = \frac{p(event\ occur)}{p(event\ fails\ to\ occurs)} = \frac{p(success)}{p(failure)}$$

$$= \frac{P}{1-p}$$

The odds ratio (OR) is a ratio of the two odds that compares the odds for those who have the risk factor (X=1) to the odds for those who do not have the risk factor (Hosmer & Lemeshow, 1989; Czepiel, 2002). The odds ratio is employed to measure the relationship between the response variable and explanatory variables (Wilber & Fu, 2010; Anon., 2020). In a logistic regression model for a continuous variable, the odds ratio shows how the odds change with a one-unit increase in that variable while other variables are constant. For categorical variables, the odds ratio is the change in the odds of each category's particular event compared to the odds of an event for the reference category. Thus, a comparison between two groups for the occurrence of the same event, for example tuberculosis (event) in men compared to tuberculosis (event) in women can be made by calculating the odds ratio.

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

An OR = 1 indicates that the explanatory variable does not affect the odds of the event. An odds ratio of more than one suggests that the explanatory variable level in the numerator is associated with higher odds of the outcome. In contrast, odds

less than one indicate that explanatory variable level or exposure in the numerator is associated with lower odds of the outcome.

3.2 Logistic Regression

The logistic regression is a particular case of a Generalized Linear Model applied in different studies, especially epidemiology (Park H, 2013; Peng et al., 2002). The logistic regression model is intended to depict a probability that is always a number between 0 and 1 (Melesse et al., 2016). The binary response is commonly studied in many fields. Logistic regression is a technique used to investigate the association between a binary response variable and a set of predictor variables. The response can be a binary or ordinal response (Agresti, 2002; Hosmer & Lemeshow, 2000).

The simple logistic regression

The residuals are assumed to have a binomial distribution in a logistic regression model for a binary response variable (Quinn & Keough, 2002). In the model for a single explanatory variable X, for instance, chronic disease, and one dichotomous outcome variable Y, for example, having tuberculosis, the logistic model predicts the logit of P(Y=1) from the independent variable X. The binary response can be modeled as:

$$y_i = \begin{cases} 1 - if\ event\ occurs\ (e.g\ has\ tuberculosis) \\ 0 - if\ no\ event\ occurs\ (e.g\ has\ no\ tuberculosis) \end{cases}$$

The odds of “yes = 1” (i.e., has tuberculosis) would be:

$\frac{\pi(x)}{1-\pi(x)} = \frac{probability\ of\ a\ "yes(e.g\ has\ tuberculosis)"}{probability\ of\ a\ "no(e.g\ has\ no\ tuberculosis)"}$,

where,

$$\pi(x) = p(Y\ is\ an\ event|X = x)$$

The logit is the natural logarithm (*ln*) of odds of p = P(Y=1).

The simple logistic regression model in the form of the logit link function has the form:

$$logit(\pi(x)) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \ln(odds) = \alpha + \beta x$$

An alternative formula has the form:

$\pi(x) = \frac{exp(\alpha+\beta x)}{1+exp\ (\alpha+\beta x)}$,

where, $\pi(x)$ = probability that an adult male will have tuberculosis when the chronic disease takes a value of x. Where α and β are parameters to be estimated, such that α is the intercept or constant and β is a slope, X can be categorical or continuous.

Multiple logistic regression

Multiple logistic regression model relates a single binary variable outcome with more than one independent or p explanatory variable.

The multiple logistic regression in the form of a logit link function is given by

$$logit(\pi(x)) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \sum_{i=1}^{p}\beta_i x_i$$

Alternative form of multiple logistic regression is given by:

$$\pi(x) = \frac{exp(\alpha + \sum_{i=1}^{p} \beta_i x_i)}{1 + exp(\alpha + \sum_{i=1}^{p} \beta_i x_i)}$$

The logistic regression estimates the unknown parameter by applying the maximum likelihood (ML) estimation method. ML estimates the value of parameters for which the probability of being observed is maximum. The logistic regression assumes the binomial distribution as the probability distribution of the binary response variable. Logistic regression assumes a response variable is randomly determined. Also, logistic regression assumes no outliers in data and no strong inter-correlation between the predictors (Tabachnick & Fidell, 2007).

3.2.1 Estimation of the parameters

For risk factors of developing tuberculosis, the response variable (Y_i) has a Bernoulli distribution that is each adult man has "tuberculosis or not." Thus, probability is either π if y_i = 1 or $1 - \pi$, if $y_i = 0$. The likelihood function is then given as:

$$L(\beta|y_i) = \prod_{i=1}^{n} \pi(x)^{y_i} (1 - \pi(x))^{1-y_i} \qquad (3.6)$$

Then we take the log of the likelihood function in (3.12) then log-likelihood function becomes:

$$\ell(\beta|y_i) = \sum_{i=1}^{n} [y_i \log(\pi(x)) + (1 - y_i) \log(1 - \pi(x))]$$

$$= \sum_{i=1}^{n} \log(1 - \pi(x)) + \sum_{i=1}^{n} y_i \log\left(\frac{\pi(x)}{1 - \pi(x)}\right)$$

$$= \sum_{i=1}^{n} \log(1 - \pi(x)) + \sum_{i=1}^{n} y_i(\alpha + \boldsymbol{x}_i^T \beta)$$

$$= \sum_{i=1}^{n} y_i(\alpha + \boldsymbol{x}_i^T \beta) - \sum_{i=1}^{n} \log\left(1 - \frac{exp\ (\alpha + \beta x)}{1 + exp\ (\alpha + \beta x)}\right)$$

We then differentiate the log-likelihood with respect to α and β to get the following score functions.

$$U(\pi(x), y_i)_\alpha = \frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^{n} \left\{ y_i - \left[\frac{\exp(\alpha + \boldsymbol{x}_i^T \beta)}{1 + \exp(\alpha + \boldsymbol{x}_i^T \beta)} \right] \right\} = \sum_{i=1}^{n} (y_i - \pi(x))$$

and

$$U(\pi(x), y_i)_\beta = \sum_{i=1}^{n} x_{ij} \left[\frac{\exp(\alpha + \boldsymbol{x}_i^T \beta)}{1 + \exp(\alpha + \boldsymbol{x}_i^T \beta)} \right] + \sum_{i=1}^{n} y_i x_{ij}$$

$$= \sum_{i=1}^{n} x_{ij} [y_i - \pi(x)]$$

This score function can be solved by equating it to zero to get the maximum likelihood estimates for β. Thus, the maximum likelihood estimation approach is used to estimate the unknown parameters while fitting a logistic regression model.

3.2.2 Assessing the model fit

The statistical model is a good fit if it fits the observation well. The minimum discrepancy between the expected values and the observed values represents a good model. The two statistical methods used in assessing the model's fit are deviance and Pearson's chi-square statistics.

Deviance measures the discrepancy of fit between the maximum log-likelihood of the saturated model and the fitted models' log-likelihood. Deviance is defined as (Ongoma, 2017):

$$D = [2l(y, \phi; y) - 2l(\hat{\mu}, \phi, y)] \quad (3.7)$$

The Pearson's goodness of fit statistic measures residual variation is given by (Dobson A. J., Barnett S. G, 2008):

$$\chi_p^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}, \quad (3.8)$$

where o_i is the observed value y_i and e_i is fitted value "expected" from a model.

3.2.3 Model Selection.

Akaike's Information Criterion

The Akaike's information criterion (AIC) is the measure introduced by Akaike (1974) to summarize the information in the model. The AIC compares models (including non-nested) and chooses the best one, the lower the AIC, the better the model. The AIC is given by:

$$AIC = -2logLikelihood + 2p, \quad (3.9)$$

where p is the number of model parameters.

Bayesian Information Criterion (BIC)

Schwarz (1978)'s BIC is an alternative to AIC for comparing models. The model with smaller BIC is preferred. The BIC is given by

$$BIC = -2loglikelihood + klog(n) \quad (3.10)$$

Here, the sample size is n, and the number of parameters evaluated is k. BIC produces more severe penalization on the likelihood of estimating more parameters (Allison, 2012).

Receiver Operating Characteristic (ROC) Curve

ROC curve refers to a standard procedure for measuring classifier performance on the possible cut-off point between actual positive and false-positive error rates (Swets, 1988). Sensitivity is defined as the proportion of true positives classified as positive (probability that a model classifies a man as having TB, given he is truly infected). Specificity is a probability of actual negative which are predicted negative (true negative), i.e., the chance that model classifies a man as he does not have TB given he is truly not infected (Neovius et al., 2004). ROC curve is a plot of sensitivity versus 1-specificity for possible cut-off (π_o) ((Agresti, 2007, 2019). The ROC is the same as the concordance index (Agresti, 2002). In general, ROC estimates the prediction accuracy of the model (Šimundić, 2009): that is, how frequently the predicted probabilities agree with an outcome of interest.

3.3 Fitting logistic regression model

In this investigation, the probability of men developing tuberculosis is modeled as a function of explanatory variables listed in Chapter Two. The binary variable, namely tuberculosis, is our dependent variable. PROC LOGISTIC was used to fit data in SAS 9.4. From Table 3. ***1***, the full model (intercept and covariates) has a lower AIC than a reduced model (intercept only); this suggests that the full model explains the data better.

Table 3. 1 Model Fit Statistics for a logistics regression model

Criterion	Intercept Only	Intercept and covariates
AIC	2014.659	1811.951
SC	2020.998	2078.180
-2 Log L	2012.659	1727.951

The likelihood ratio statistic has a P-value of <0.0001 and is 284.7077. The score test had a value of 293.8380 with a P-value of <0.0001, while the Wald test had a value of 232.0157 with a P-value of <0.0001 Table 3. 2. All three tests have P-values less than 0.05, indicating that the overall logistic regression model is significant. It indicates that predictors have a major role in predicting the likelihood of TB in men. Moreover, the Hosmer and Lemeshow test for goodness of fit statistics with the value of 8.6511 and *P-value* = 0.3726 supports the evidence that the logistic regression model fitted is a good fit for the data. Logistic regression models should also be validated for their predicted probability (Melesse et al., 2016). The c-statistic measures the predictive accuracy; for this study, the c-statistic is 0.785, which indicates that 78.5% of the probabilities are predicted correctly. It shows a strong correlation between anticipated and actual probabilities. The concordant rate was 78.5%; this tells us how well the logistic regression model agrees with the observed outcome. The Gamma statistics value was 0.570, which indicates that there is no perfect association. The Somers'D statistics was 0.570 supported that there were not all pairs concordant and may be used to compare the model. The Pearson statistic with P-value=0.9986 and deviance with P-value = 1.0000 goodness fit were employed to assess the sufficiency of the logistic regression model to explain the DHS data. The fact that both Pearson and deviance are near to one indicates that the data is well fitted.

Table 3. 2 Overall model evaluation for logistic regression

Model Evaluation	Chi-square	D.F	P-value
Overall significance			
Likelihood Ratio	284.7077	41	<.0001
Score	293.8380	41	<.0001
Wald	232.0157	41	<.0001
Goodness of fit			
Hosmer and Lemeshow	8.6511	8	0.3726
Deviance	1725.1788	4115	1.0000
Pearson	3848.3965	4115	0.9986
Association of Predicted Probabilities and Observed Response			
Percent Concordant	78.5	Somers'D	0.570
Percent Discordant	21.5	Gamma	0.570
Percent Tied	0.0	Tau-a	0.069
Pairs	1063792	c	0.785

Table 3. ***3*** illustrates the hypothesis testing for each variable in the logistic regression model individually based on the multiple degrees of freedom test for the overall effect of the categorical variables. At a 5% significance level, the results show that the categorical variables: chronic disease, current age, region, ethnicity, times away from home, health, weight, and marital status were found to have a statistically significant effect on the probability of having tuberculosis. However, smoking status and the number of household members were found to have no significant effect on the probability of having tuberculosis based on the 2016 South Africa DHS data. Moreover, the significant interaction terms were: current age and chronic disease, smoking status, and the number of household members.

Table 3. 3 Type 3 Analysis of Effects

Main effect	**DF**	**Wald χ^2**	**P-value**
Chronic disease	1	23.0969	<.0001
Current age	1	3.9798	0.0460
Region	8	62.4891	<.0001
Education Level	3	10.9959	0.0117
Ethnicity	4	14.2790	0.0065
Times away from home	1	4.4939	0.0340
Wealth index	4	6.3030	0.1776
Marital status	5	15.5431	0.0083
Health	3	15.4820	0.0014
Weight	4	9.9279	0.0417
Smoking status	2	3.7474	0.1536
Number of Household members	1	0.2057	0.6502
Interaction effect			
Current age and chronic disease	1	22.0492	<.0001
No of household members and smoking status	2	9.1535	0.0103
Current age and times away from home	1	2.5279	0.1119

Table 3. ***4*** shows the significant interaction effect between chronic disease and current age, and the number of household members, and smoking status. Thus, the effect of the current age and chronic disease was found to be associated with decrease in having tuberculosis for men (P-value<0.0001). The effect of the number of household members and smoking status was associated with decrease in developing tuberculosis for the men. Other than the significant interaction effect, Table 3. 4 also shows that region (North West and Gauteng), education level (no education and higher), ethnicity (Indian/Asian and other), wealth index (poorest, poor and richest), marital status (married, living with a partner, and separated), health (average and excellent), weight, and smoking status were found to have no significant effect on the probability of having tuberculosis.

The probability of developing/having tuberculosis of a man as a function of the covariates is estimated by

$$\hat{P} = \frac{1}{1 + e^{-5.6284} + \sum_{i=1}^{n} \hat{\beta}_i x_i}$$

$\hat{\beta}$'s are the estimated coefficients of x_i's variables, which have a significant effect on the response. In β-coefficients, the +/- sign indicates their relationship is positive or negative with having tuberculosis. Odds are the exponent's power to the estimates. The odds ratio of 1.023 for the age suggests that with a one-unit increase in age, the odds of having tuberculosis increase by 2.3% (P-value=0.0364). Chronic diseases have a significant effect (P-value<0.0001) on the probability of having tuberculosis with an odds ratio of 8.102. Men who live in the Western Cape (OR=6.559 with P-value<0.0001) had higher odds of having tuberculosis compared to men who live in Limpopo, followed by the Northern Cape (OR=4.31 with p-value<.0001), Free States (OR=4.25 with p-value<0.0001), Eastern Cape (OR=4.162 with P-value<.0001), Mpumalanga (OR=2.887 with p-value=0.0007), and the KwaZulu-Natal (OR=2.882 with p-value=0.0006) when compared to Limpopo.

Furthermore, white and colored men have lower odds of having tuberculosis than black men, with odds of 0.093 and 0.49, respectively. Divorced men (OR=3.254) were at higher odds of having TB than men who were never in a union, followed by widowed (OR=2.243). Moreover, a man with poor health had higher odds of having tuberculosis than a man with good health (OR=0.122). Overweight men are associated with decrease in the risk of having tuberculosis compared to underweight men, with an odds ratio of 0.426. This implies that overweight men are (1-0.4226) %=57.74% less likely to have TB than underweight men. The odds ratio of 1.12 for a number of household members indicates that the odds of having tuberculosis increase significantly by 12% with a one-unit increase in household members.

Table 3. 4 Effects on response for indicator and significant two-way interaction effects

Indicator	DF	Estimate	SE	P-value	OR
Intercept	1	-5.6284	0.4296	<.0001	0.004
Chronic disease (ref = NO)					
Yes	1	2.0921	0.4353	<.0001	8.102
Current age	1	0.0318	0.0061	<.0001	1.032
Region (ref = Limpopo)					
Western Cape	1	1.8809	0.3859	<.0001	6.559
Eastern Cape	1	1.4259	0.2903	<.0001	4.162
Northern Cape	1	1.4610	0.3356	<.0001	4.310
Free State	1	1.4468	0.3256	<.0001	4.250
Kwazulu-Natal	1	1.0584	0.3084	0.0006	2.882
North West	1	0.3574	0.3373	0.2894	1.430
Gauteng	1	-0.5263	0.4658	0.2585	0.591
Mpumalanga	1	1.0601	0.3114	0.0007	2.887
Education Level (ref = Secondary)					
No education	1	-0.0788	0.2581	0.7602	0.924
Primary	1	0.4308	0.1675	0.0101	1.539
Higher	1	0.4899	0.2807	0.0809	1.632
Ethnicity (ref = Black/African)					
White	1	-2.3755	0.7607	0.0018	0.093
Colored	1	-0.7127	0.2828	0.0117	0.490
Indian/Asian	1	-13.9226	464.70	0.9761	0
Other	1	-13.9182	2936.8	0.9962	0
Times away from home	1	0.0289	0.0136	0.0340	1.029
Wealth Index (ref = Middle)					
Poorest	1	0.0415	0.1965	0.8328	1.042
Poor	1	-0.0174	0.1944	0.9285	0.983
Richer	1	-0.5068	0.2229	0.023	0.602
Richest	1	-0.1481	0.2705	0.584	0.862
Marital status (ref = Never in a union)					
Married	1	0.0257	0.1981	0.8967	1.026
Living with partner	1	0.2416	0.2232	0.2791	1.273
Widowed	1	0.8077	0.3277	0.0137	2.243
Divorced	1	1.1799	0.4196	0.0049	3.254
Separated	1	0.4508	0.3488	0.1962	1.570
Health (ref = Good)					
Poor	1	0.7799	0.2157	0.0003	2.181
Average	1	0.2190	0.1668	0.1892	1.245
Excellent	1	-0.2330	0.2541	0.3592	0.792
Weight (ref =Normal)					
Normal	1	-0.3377	0.1849	0.0678	0.713
Overweight	1	-0.8526	0.3755	0.0232	0.426
Obese	1	1.2451	0.8417	0.1390	3.473
Don't know	1	-0.3059	0.6551	0.6406	0.736
Smoking Status (ref = Everyday)					
Do not smoke	1	0.4582	0.2468	0.0634	1.581
Sometimes	1	0.5350	0.5107	0.2949	1.707
Number of household memebers	1	0.1132	0.0331	0.0006	1.120
Significant interaction effect					
Chronic disease and age (ref = No)					
Having chronic disease and current age	1	-0.0395	0.0084	<.0001	0.961
No of household members and smoking status(ref=Everyday)					
No of household members and do not smoke	1	-0.1306	0.0447	0.0035	0.878

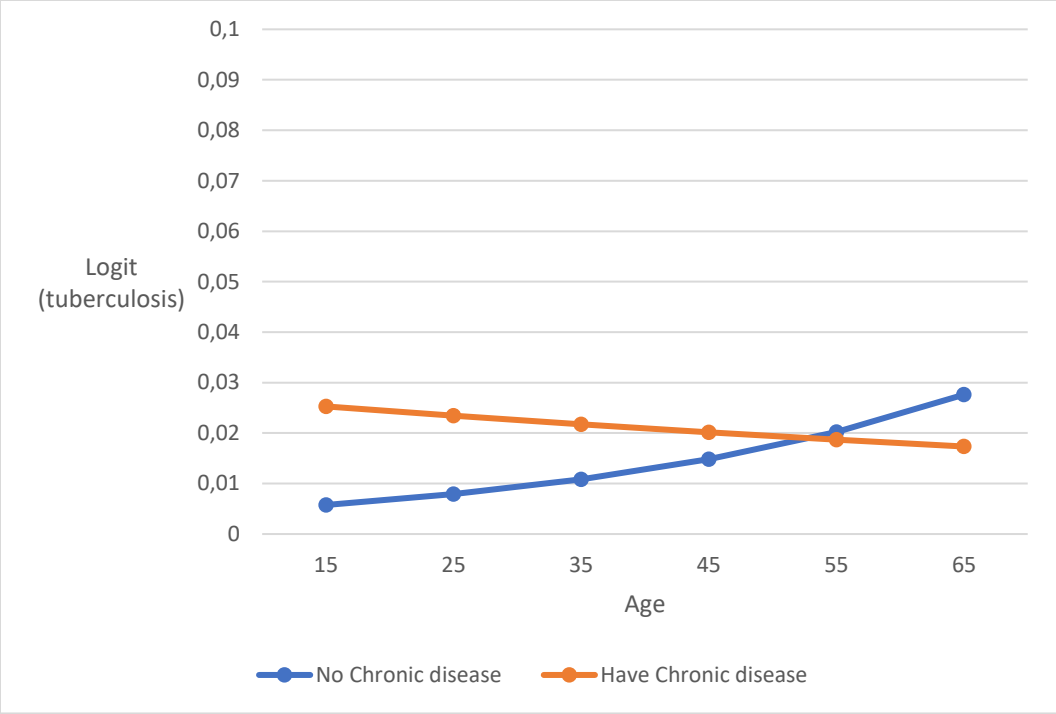

Figure 3. 1 Interaction effect for the current age and chronic disease for logistic regression

Figure 3. **1** indicates that men aged between 15 and 54 who have the chronic disease have higher chances of having tuberculosis than men who do not have a chronic disease. At the same time, men above 55 years who have chronic disease are less likely to have the TB than men without chronic disease. Figure 3. 2 shows that as the number of household members increases, the probability of having tuberculosis increases.

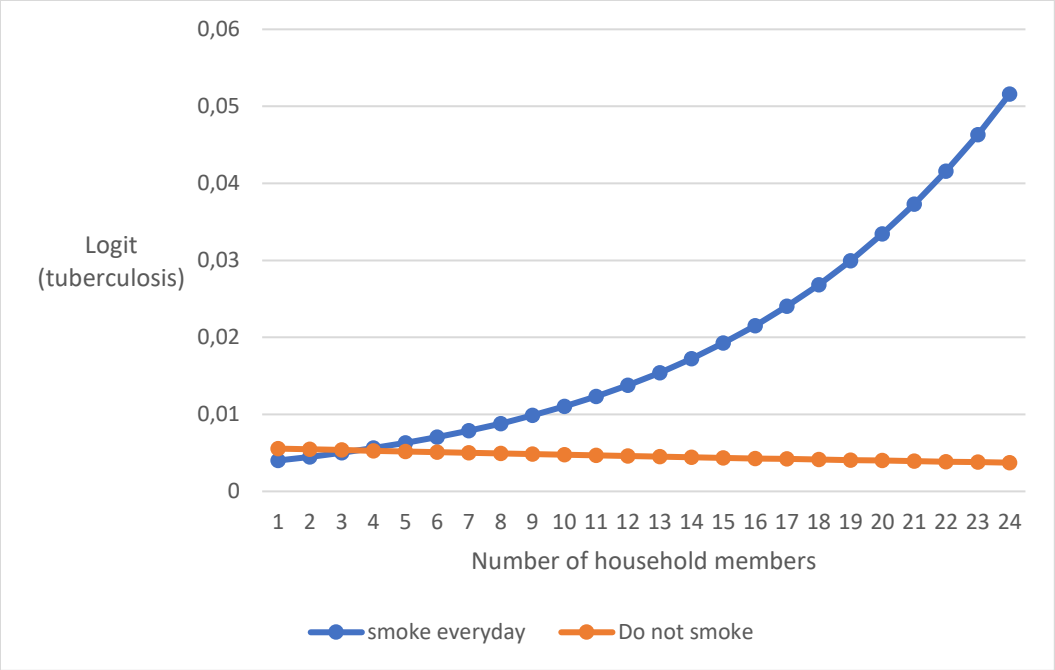

Figure 3. 2 Interaction for number of household members and smoking status for Logistic regression

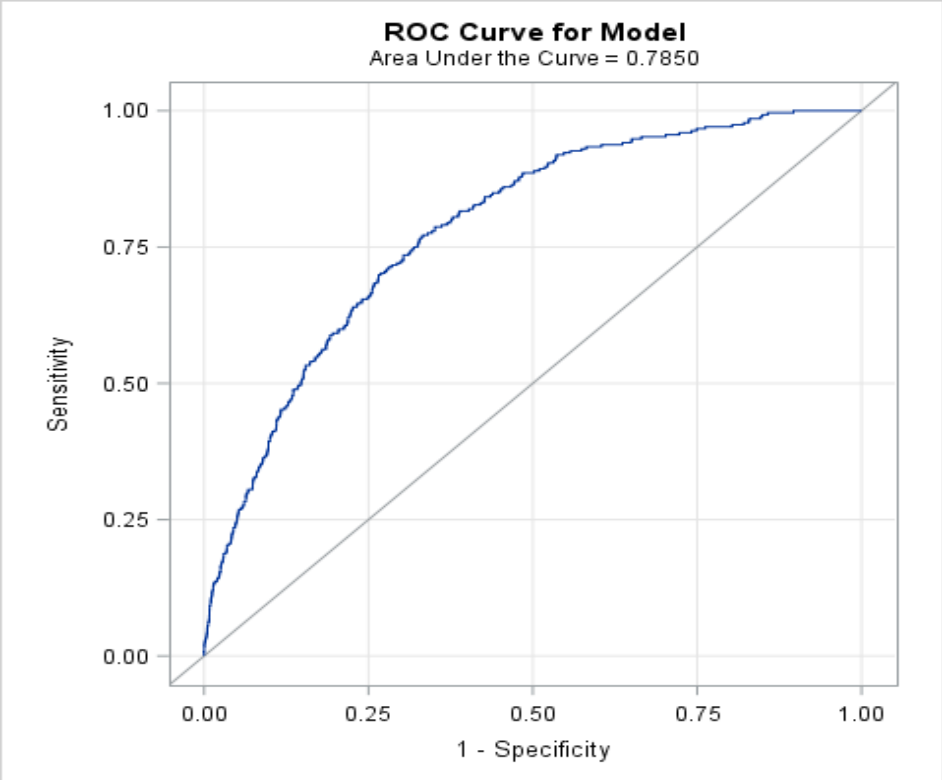

Figure 3. 3 ROC curve for a logistic regression model

The model's Area Under the Curve (AUC) of the model is higher if AUC is closer to 1; this implies that the model's prediction accuracy is excellent. In our case, the AUC= 0.785 in Figure 3. ***3*** suggests that the model's prediction accuracy is moderate.

3.4 Survey Logistic Regression

The common statistical analysis methods, such as the logistic regression models, assume that data is collected in a finite population by simple random sampling (SRS). SRS assumes all population members have an equal chance to be included in the sample. Survey data are derived from a limited population for a sample that is stratified by a variable rather than a simple random sample (Anthony, 2002). Survey logistic regression is a design-based statistical technique that is an extension of the classical logistic regression (Heering et al., 2010). Survey logistic regression models the association between binary responses to a set of explanatory variables by taking into account the complex sampling design (Moeti, 2007; Rao & Scott, 1981; Lu & Yang, 2012). Survey logistic regression model is helpful since it includes the effect of sampling design in the analysis results to accurate (or adjusted) estimates of standard errors and variability (Kish, 1965; Skinner et al., 1989). The survey logistic regression model differs from logistic regression in that it incorporates design complexity (Ayele et al., 2012). In the absence of sampling design, the standard error will probably be underestimated, resulting in statistically significant results; when they are substantial, this might result in skewed estimates

and confidence ranges (Vittinghoff et al., 2011). The advantages of sampling surveys are: feasiblity, quality, and accurate population estimates (Cochran, 1977).

3.4.1 The survey logistic regression model

Let us consider the survey logistic regression for a binary dependent variable Y_{hijp}, j =1,….,m_p; i = 1,….,n_{pj}; h = 1,….H_{pji},; p=1,2,…p which is 1 if the event occurred in the hth individual within the ith household, within jth cluster primary sample units (mv021) nested within pth stratum and 0 otherwise. Assume that the probability of having tuberculosis is $\pi_{hijp} = P(Y_{hij} = 1| X_{hijp})$. The survey logistic model is then given by

$$logit(\pi_{hijp}) = \boldsymbol{x}_{hijp}^{T}\beta \qquad (3.13)$$

and

$$\pi_{hijp} = \frac{\exp(\boldsymbol{x}_{hijp}^{T}\beta)}{1+\exp(\boldsymbol{x}_{hijp}^{T}\beta)},$$

where $\boldsymbol{x}_{hijp}$ is covariate matrix, and β are unknown regression coefficients to be estimated. Survey logistic regression estimate the unknown parameters β by applying the Pseudo-maximum likelihood, which includes sampling design and sampling weights in the estimation of β (Hosmer & Lemeshow, 2000; Pfeffermann, 1993).

3.4.2 Parameter estimation

In the complex survey design, the assumption that independent samples are equally likely to be selected is violated. The standard errors and model estimates are estimated by including the complexity of the sample design (clustering,

stratification, and the use of probability weights) (Siller & Tompkins, 2006). Based on (Kish & Frankel, 1974), stratification in sample design reduces variance when there is a negative correlation, whereas clustering increases the correlation among elements.

Pseudo-maximum likelihood estimation (PMLE)

The pseudo-maximum likelihood estimation is commonly used to obtain unknown parameters in complex survey data. The pseudo-maximum of a single observation for survey logistic regression model is given by

$$\pi_{hjip}^{W_{hjip}Y_{hjip}}(1-\pi_{hjip})^{(1-W_{hjip}Y_{hjip})}$$

Then the pseudo-maximum likelihood function with weight W_{hjip} for n-observations (Archer et al., 2007) is given by:

$$L(\beta|W_{hjip}Y_{hjip}) = \prod_{h=1}^{H}\prod_{j=1}^{n'_h}\prod_{i=1}^{n'_{hjp}}\pi_{hjip}^{W_{hjip}Y_{hjip}}(1-\pi_{hjip})^{(1-W_{hjip}Y_{hjip})} \quad (3.14)$$

The basic concept behind this technique is to create a function that approximates the likelihood function of a sampled finite population with a likelihood function produced by the observed sample and known sampling weights. The pseudo-log-likelihood is then given by

$$\ell(\beta|W_{hjip}Y_{hjip}) = \sum_{h=1}^{H}\sum_{j=1}^{n'_h}\sum_{i=1}^{n'_{hj}}\{(W_{hjip}Y_{hjip})\ln\left(\frac{\pi_{hjip}}{1-\pi_{hjip}}\right)+\ln(1-\pi_{hjip})\}$$

thus,

$$\ell\left(\beta\middle|W_{hji}Y_{hji}\right) = \sum_{h=1}^{H}\sum_{j=1}^{n'_h}\sum_{i=1}^{n'_{hj}}\{\left(W_{hjip}Y_{hjip}\right)\boldsymbol{x}_{hijp}^{T}\beta - \ln\left(1+\exp\left(\boldsymbol{x}_{hijp}^{T}\beta\right)\right)\}$$

then we take the derivative of pseudo-log-likelihood and equate it to zero to obtain the estimates of β.

$$\boldsymbol{S}(\boldsymbol{\beta}) = \frac{\partial\ell\left(\beta\middle|W_{hjip}Y_{hjip}\right)}{\partial\beta} = 0$$

Thus,

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{h=1}^{H}\sum_{j=1}^{n'_h}\sum_{i=1}^{n'_{hj}}\left(W_{hjip}\left(Y_{hji} - \frac{\exp\left(\boldsymbol{x}_{hijp}^{T}\beta\right)}{1+\exp\left(\boldsymbol{x}_{hijp}^{T}\beta\right)}\right)\boldsymbol{x}_{hijp}^{T}\right) = 0$$

Newton- Raphson and Fisher scoring iterative methods will solve unknown parameter β (SAS Institute Inc., 2015).

3.4.3 Variance estimation in survey logistic

The linearization method and replication methods such as Jackknife repeated replication, balanced repeated replication, bootstrap, and Taylor linearization are used to obtain the variance estimators under complex survey design (Binder, 1983; Cochran, 1964; Efron, 1980; Rust, 1985; Kolenikov, 2010; Lu, 2004; Rao & Wu, 1998; Wolter, 1985; Woodruff, 1971).

Binder (1983) proposed using the Taylor series approximation method to derive the variance estimates. Because the parameter estimates β are given by

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

The first-order Taylor expansion of $\boldsymbol{S}(\widehat{\boldsymbol{\beta}})$ at $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ the population value is given by

$$0 = \boldsymbol{S}(\widehat{\boldsymbol{\beta}}) \simeq S(\beta) + \frac{\partial S(\beta)}{\partial \beta}(\hat{\beta} - \beta)$$

Resulting to,

$$S(\beta) \simeq \frac{\partial S(\beta)}{\partial \beta}(\hat{\beta} - \beta)$$

Then take variance both sides the following results is obtained

$$Var[S(\hat{\beta})] = \left[\frac{\partial S(\beta)}{\partial \beta}\right] var(\hat{\beta}) \left[\frac{\partial S(\beta)}{\partial \beta}\right]'$$

or can be written as

$$var(\hat{\beta}) = \left[\frac{\partial S(\beta)}{\partial \beta}\right]^{-1} Var[S(\hat{\beta})] \left[\frac{\partial S(\beta)}{\partial \beta}\right]^{-1}$$

$$= [I(\hat{\beta})]^{-1} Var[S(\hat{\beta})] [I(\hat{\beta})]^{-1},$$

where $I(\hat{\beta})$ is information matrix and $Var[S(\hat{\beta})]$ is a variance-covariance matrix for p+1 estimating equations given by

$$Var[S(\hat{\beta})] = \sum_{j}^{H} (1 - f_h) \frac{n_h'}{n_h' - 1} \sum_{j=1}^{n_h'} (X_{hj\ldots} - \bar{X}_{h\ldots})(X_{hj\ldots} - \bar{X}_{h\ldots})'$$

the quantity $(1 - f_h)$ is a correction factor for a finite population where $f_h = \frac{n_h'}{n_h}$.

The specific mean in the stratum is given as $\bar{X}_{h\ldots} = \sum_{j=1}^{n_h'} X_{hjp\ldots}$ where $X_{hjp\ldots} = \sum_{j=1}^{n_h'} W_{hjip} \pi_{hjip} (1 - \pi_{hjip})$.

A SAS procedure PROC SURVEYLOGISTIC produces the covariance matrix of a parameter by using the Taylor expansion approximation procedure (Vittinghoff et al., 2011).

3.4.5 Model selection and check for survey logistic

Model Selection

SAS PROC SURVEYLOGISTIC (version 9.4) currently does not have the forward, backward, and stepwise variables selection option. This means we should examine variables by selecting each variable manually and observing its effects at a time by using Type 3 analysis effect, then excluding those variables with no significant contribution and monitoring the effect of the remaining variables (performing a univariate analysis between the response variable and the independent variable).

Model Checking

In survey logistic, we compare two models using AIC and BIC just like logistic regression. The survey logistic in SAS does not generate plots or Hosmer-Lemeshow statistics (Lumley & Scott, 2015).

Prediction/Accuracy

The SAS PROC SURVEYLOGISTIC generates the Kendall's tau-a, Gamma, and Somer'D statistics which are used to measure the association between covariates that are used in the model. The predicted probabilities are compared with actual outcomes using these statistics (Peng & So, 2002). Kendall's tau-a, Gamma, and

Somers'D result from concordant and discordant (Sheskin, 2000). Kendall's statistics are between -1 to +1. Where -1 indicates complete disagreement and +1 complete agreement among the rankings (Sheskin, 2000) and is defined as

$\hat{\tau} = \frac{n_c - n_d}{\left[\frac{n(n-1)}{2}\right]}$,

where n_c and n_d are numbers of concordant and discordant pairs, respectively, $\frac{n(n-1)}{2}$ is the total number of possible pairs of ranks.

Gamma statistics is defined as

$$G = \frac{n_c - n_d}{n_c + n_d}$$

The Somers'D is given by

$$Somers'D = \frac{n_c - n_d}{n_c + n_d + T_y}$$

T_y is the number of pairs tied ranks on the response variable.

The Gamma statistics and Somers'D are the measures of association

3.4.5 Design effect

Stratification, clustering, and weighting of selection of cases determine the sample variance of a survey statistic. Clustering and weighting reduce precision, while stratification improves precision (Dowd & Duggan, 2001). The precision of the parameter estimates depends on the sample size and sampling design. In survey sampling, a method called design effect determines the sample design in estimation

and analysis (Park & Lee, 2001). The design effect is then defined by (Dowd & Duggan, 2001):

$$deff = \frac{var(complex\ design)}{var(simple\ random\ sample)}$$

Because *deft* is less variable than *deff*, $deft = \sqrt{deff}$ is preferred for determining design effect. *deft* indicates how much the sample's standard error and confidence interval increase. If *deft* = 1, it implies that sample design has no influence on standard error. Assume that the sample design with *deft*>1 increases the standard error of the estimations. The value of *deft*<1 sample design decreases the estimate's standard error.

3.4.6 Survey logistic regression model application

PROC SURVEYLOGISTIC SAS 9.4 was used to fit a survey logistic regression model to the data, which considers the survey design's complexity. The univariate analysis was used to test each predictor variable's significance with a response, and variables that were not statistically significant into the model were removed. The two-way interactions among the explanatory variables were explored to avoid possible confounding effects.

Table 3. ***5*** shows AIC, SC, and -2logL of the model with intercept and covariates are smaller than a model with the only intercept, suggesting that the model contains intercept and covariates better explain the data.

Table 3. 5 Survey logistic regression model fit statistics

Criterion	Intercept only	Intercept and covariates
AIC	1489.852	1266.538
SC	1496.039	1526.395
-2 Log L	1487.852	1182.538

The likelihood ratio, score tests, and Wald test are statistically significant at a 5% level of significance. This means there is a significant contribution of covariates in the prediction of having tuberculosis. Table 3. ***6*** shows that 79.8% of the probabilities are predicted correctly, suggesting a good association between the predicted probabilities and actual. The concordant is 79.4%, Gamma is 60%, and Somers'D is 59.5%.

Table 3. 6 Survey logistic regression model evaluation

Model evaluation				
Overall Significance	F-value	Num DF	Den DF	Pr>F
Likelihood Ratio	6.96	34.75	23772	<.0001
Score	6.22	41	625	<.0001
Wald	60.80	41	6250	<.0001
Association of predicted probability with observed				
Percent Concordant	79.40	Somers'D	0.595	
Percent Discordant	19.90	Gamma	0.600	
Percent Tied	0.80	Tau-a	0.069	
Pairs	752402	c	0.798	

Table 3. ***7*** shows that all the interaction effects are significantly associated with a probability of having tuberculosis at a 5% level of significance. Furthermore, it indicates that education level, wealth index, health, weight, and the number of household members are not statistically significantly associated with the probability

of having tuberculosis among adult men, and all other main effects were statistically significant.

Table 3. 7 Type 3 Analysis of effect for survey logistic regression model

Main effect	F-Value	Num DF	Den DF	P-value
Chronic disease	24.73	1	665	<.0001
Current age	10.59	1	665	0.0012
Region	4.86	8	658	<.0001
Education Level	2.21	3	663	0.0862
Ethnicity	325.14	4	662	<.0001
Times away from home	6.59	1	665	0.0105
Wealth index	0.41	4	662	0.8037
Marital status	2.90	5	661	0.0134
Health	0.95	3	663	0.4138
Weight	1.55	4	662	0.1847
Smoking status	4.82	2	664	0.0083
Number of Household members	1.13	1	665	0.2892
Interaction effect				
Current age and chronic disease	17.57	1	665	<.0001
No of household members and smoking status	5.58	2	664	0.0039
Current age and times away from home	4.85	1	665	0.0280

The output from Table 3. 8 shows that the effect of age on tuberculosis depends on presence of chronic disease. This implies that as for those with no chronic disease, the presence of tuberculosis increases with age. The corresponding odds ratio is 0.937. The odds of having tuberculosis for men with chronic disease was 0.937 times lower compared to someone without chronic disease at the same age. The number of household members increases with the odds of having tuberculosis for men who do not smoke, and those who smoke sometimes the presence of tuberculosis decrease with increase in number of households, compared to those who smoke every day with an odds ratio of 0.827 and 0.756, respectively. Table 3. 8 also shows that the risk of having tuberculosis decreases with an increase in age and number of household members with OR = 0.999. The main effect of the men with chronic disease was a statistically significantly higher risk of having tuberculosis than those without the chronic disease (OR=24.989, p-value<0.0001). Age has a associated with increase inthe risk of having tuberculosis. This implies a

one-unit increase in age; the odds of having tuberculosis increases by 6.7% =(1.067-1)%.

All the regions that have significant association with tuberculosis have a positive significant effect risk of having tuberculosis. Men from Western Cape have higher odds than Limpopo (OR=6.343, p-value=0.0005), followed by men from Northern Cape compared to men from Limpopo (OR=5.434, p-value=0.0012). The odds of having tuberculosis for men from the Eastern Cape is 3.975 times higher than men from Limpopo (p-value =0.0012, for men from KwaZulu-Natal, is 3.004 times higher than men from Limpopo with p-value= 0.0103. Furthermore, the odds of having tuberculosis for men from Mpumalanga is 2.385 times higher than in Limpopo men.

Table 3. 8 Survey logistic regression analysis of maximum likelihood

Indicator	Estimate	S.E	P-value	OR
Intercept	-6.3473	0.6317	<.0001	0.0020
Chronic disease (ref = NO)				
Yes	3.2184	0.6471	<.0001	24.989
Current age	0.0646	0.0109	<.0001	1.0670
Region (ref = Limpopo)				
Western Cape	1.8474	0.5281	0.0005	6.3430
Eastern Cape	1.3801	0.4258	0.0012	3.975
Northern Cape	1.6928	0.4650	0.0003	5.434
Free State	1.3921	0.4464	0.0019	4.023
Kwazulu-Natal	1.1000	0.4278	0.0103	3.004
North West	0.3787	0.4787	0.4292	1.460
Gauteng	-0.8337	0.5851	0.1547	0.434
Mpumalanga	0.8692	0.4297	0.0435	2.385
Education Level (ref = Secondary)				
No education	0.4745	0.3401	0.1634	1.607
Primary	0.3808	0.2312	0.1000	1.464
Higher	0.6558	0.3409	0.0548	1.927
Ethnicity (ref = Black/African)				
White	-2.6057	0.9553	0.0065	0.074
Colored	-1.1217	0.4245	0.0084	0.326
Indian/Asian	-14.3155	0.4362	<.0001	0.000
Other	-15.0754	1.1363	<.0001	0.000
Times away from home	0.0437	0.0170	0.0105	1.045
Wealth Index (ref = Middle)				
Poorest	-0.0291	0.2970	0.9221	0.971
Poor	-0.0459	0.2871	0.8731	0.955
Richer	-0.1743	0.3044	0.5672	0.840
Richest	0.2703	0.3345	0.4194	1.310
Marital status (ref = Never in a union)				
Married	-0.0011	0.2480	0.9965	0.999
Living with partner	0.3022	0.3045	0.3213	1.353
Widowed	1.3966	0.6509	0.0323	4.041
Divorced	1.2704	0.4762	0.0078	3.562
Separated	0.6590	0.4233	0.1200	1.933
Health (ref = Good)				
Poor	0.4489	0.3173	0.1577	1.567
Average	0.0624	0.2177	0.7744	1.064
Excellent	-0.1349	0.3280	0.6809	0.874
Weight (ref =Underweight)				
Normal	-0.5296	0.2776	0.0568	0.589
Overweight	-1.1039	0.4789	0.0215	0.332
Obese	-0.5436	1.2255	0.6575	0.581
Don't know	-0.4199	0.9872	0.6708	0.657
Smoking Status (ref = Everyday)				
Do not smoke	0.8552	0.3173	0.0072	2.352
Sometimes	1.3504	0.6731	0.0452	3.859
Number of household members	0.1076	0.0430	0.0126	1.114
Interaction effect				
Chronic disease and age (ref = No)				
Having chronic disease and current age	-0.0654	0.0156	<.0001	0.937
No of household members and smoking status(ref=Everyday)				
No of household members and do not smoke	-0.1900	0.0638	0.0030	0.827
No of household members and Sometimes	-0.2798	0.1356	0.0394	0.756
Current age and times away from home	-0.0010	0.0005	0.0280	0.999

Table 3. 8 also suggests that the increase in the number of times away from home for adult men increases the risk of having TB by (1.045-1) %=4.5%. The risk of having tuberculosis for men who are widowed is 4.041 times higher than men who are never in a marriage union, followed by for men who are divorced, which is 3.562 times higher than those men never in a union. The risk of having tuberculosis for overweight men is (1-0.332) %=66.8% less likely than underweight men; furthermore, as the number of household members increases, the risk of tuberculosis increases by (1.114-1)%=11.4%.

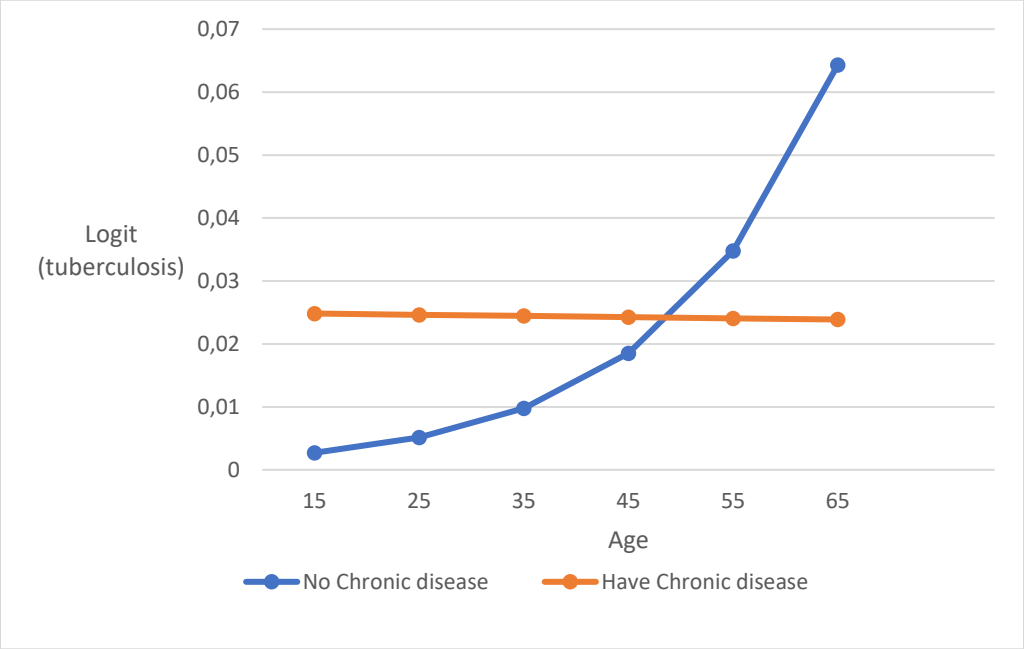

Figure 3. 4 Interaction effect for the current age and chronic disease for survey logistic regression

Figure 3. **4** indicates men aged between 15 and 46 with chronic disease have higher chances of having tuberculosis than men who do not have a chronic disease. At the same time, men above 47 years who have chronic disease are less likely to have the chronic disease than men without chronic disease. Figure 3. 5 shows that as the number of household members increases, the probability of having tuberculosis

increases for daily smokers. Whereas the number of household members increases, the risk of having tuberculosis decreases for men who do not smoke and for men who smoke sometimes.

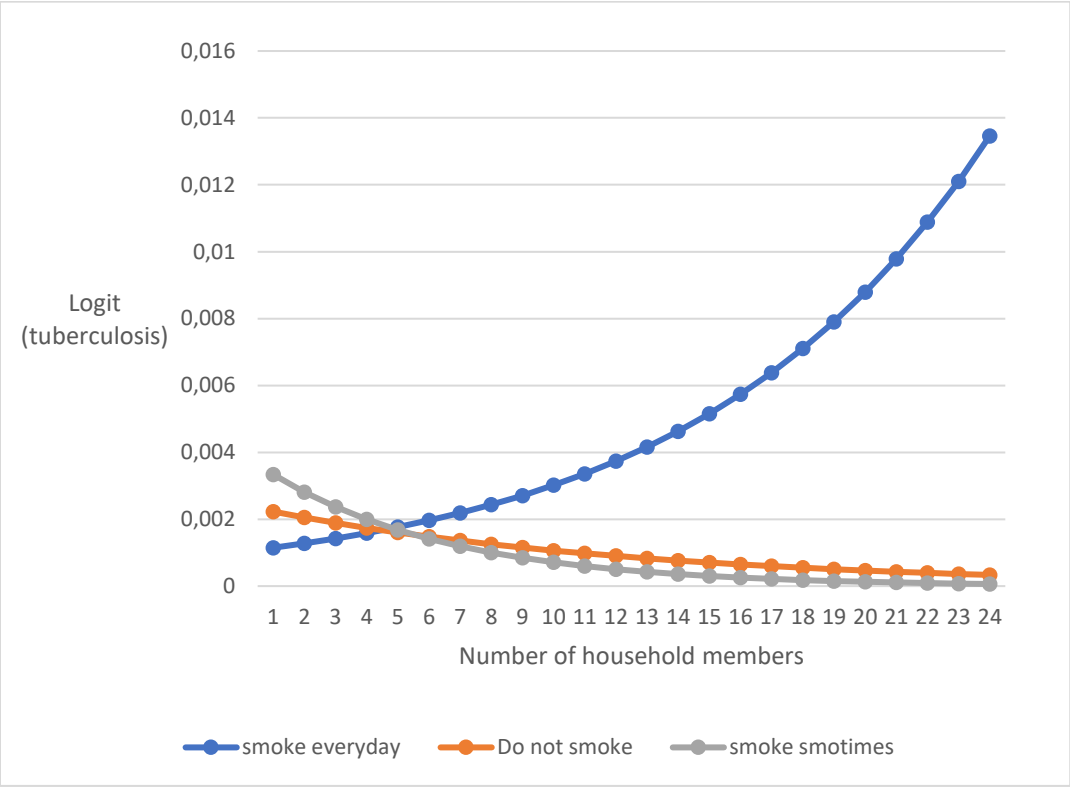

Figure 3. 5 Interaction for number of household members and smoking status for SLR

3.5 Comparison of Results from Logistic Regression and Survey Logistic Regression

Table 3. 9 Design Effect

Indicator	Estimate	P-value	V(CSD)	V(SRS)	deff	deft
Intercept	-6.8769	<.0001	0.4123	0.1846	2.2340	1.4947
Chronic disease (ref = NO)						
Yes	3.2184	<.0001	0.4187	0.1895	2.2099	1.4867
Current age	0.0646	<.0001	0.0001	3.71E-05	3.2035	1.7898
Region (ref = Limpopo)						
Western Cape	1.8474	0.0005	0.2789	0.1489	1.8728	1.3685
Eastern Cape	1.3801	0.0012	0.1813	0.0843	2.1514	1.4668
Northern Cape	1.6928	0.0003	0.2162	0.1126	1.9198	1.3856
Free State	1.3921	0.0019	0.1993	0.1060	1.8797	1.3710
Kwazulu-Natal	1.1000	0.0103	0.1830	0.0951	1.9242	1.3872
Mpumalanga	0.8692	0.0435	0.1846	0.0969	1.9041	1.3799
Ethnicity (ref = Black/African)						
White	-2.6057	0.0065	0.9126	0.5787	1.5771	1.2558
Colored	-1.1217	0.0084	0.1802	0.0799	2.2532	1.5011
Times away from home	0.0437	0.0105	0.0003	0.0002	1.5625	1.2500
Marital status (ref = Never in a union)						
Widowed	1.3966	0.0323	0.4237	0.1074	3.9453	1.9863
Divorced	1.2704	0.0078	0.2268	0.1761	1.2880	1.1349
Weight (ref = Underweight)						
Overweight	-1.1039	0.0215	0.2293	0.1400	1.6266	1.2754
Smoking Status (ref = Everyday)						
Do not smoke	0.8552	0.0072	0.1007	0.0609	1.6529	1.2857
Sometimes	1.3504	0.0452	0.4531	0.2608	1.7371	1.3179
Number of household members	0.1076	0.0126	0.0018	0.0011	1.6876	1.2992
Interaction						
Chronic disease and age (ref = No)						
Having chronic disease and current age	-0.0654	<.0001	0.0002	7.06E-05	3.4490	1.8571
No of household members and smoking status(ref=Everyday)						
No of household members and do not smoke	-0.1900	0.0030	0.0041	0.0019	2.0372	1.4273

Table 3. 9 is used to compare the standard error and confidence interval of PROC SURVEYLOGISTIC and PROC LOGISTIC based on the DEFT and DEFF. DEFT is a square root of the DEFF. The effect of chronic disease, which is associated with increase in tuberculosis, has a DEFT value of 1.4867 and a DEFF value of 2.2340. The standard error and confidence interval are 1.4867 times larger than in an SRS. The effect of current age, which is associated with increase in tuberculosis, has a DEFT value of 1.7898 and a DEFF value of 3.2035. The standard error and confidence interval are 1.7898 times larger than in an SRS. The effect of the Western Cape, which is associated with increase in the risk of having tuberculosis, has a DEFT value of 1.3685 and a DEFF value of 1.8728. The standard error and

confidence interval are 1.3685 times larger than they would be for SRS. The effect of men from the Eastern Cape with a associated with increase inthe risk of having TB has a DEFT value of 1.4668 and a DEFF value of 2.1514. The standard error and confidence interval are 1.4668 times larger than they would be for SRS. The effect of men from the Northern Cape with a associated with increase inthe risk of having tuberculosis has a DEFT value of 1.3856 and a DEFF value of 1.9198. The standard error and confidence interval are 1.3856 times larger than they would be for SRS. Men from the Free States, which is associated with increase in the risk of having tuberculosis, have a DEFT value of 1.371 and a DEFF value of 1.8797. The standard error and confidence interval are 1.371 times larger than they would be for SRS. The effect of men from KwaZulu-Natal, which is associated with increase in the risk of having tuberculosis, have a DEFT value of 1.3872 and a DEFF value of 1.9242. The standard error and confidence interval are 1.3872 times larger than they would be for SRS. Men from Mpumalanga, which is associated with increase in the risk of having tuberculosis, has a DEFT value of 1.3799 and a DEFF value of 1.9041. The standard error and confidence interval are 1.3799 times larger than they would be for SRS. The effect of white men is associated with decrease in the risk of having TB, has a DEFT value of 1.2558 and a DEFF value of 1.5771. The standard error and confidence interval are 1.2558 times larger than they would be for SRS. The effect of colored men is associated with decrease in having tuberculosis, with a DEFT value of 1.5011 and a DEFF value of 2.2532. The standard error and confidence interval are 1.5011 times larger than they would be for simple random sampling. The effect of widowed, which is associated with increase in the risk of having tuberculosis, has a DEFT value of 1.9863 and a DEFF value of 3.9453. The standard error and confidence interval are 1.9863 times larger than they would be for simple random sampling. The effect of divorced men, which is associated with increase in the risk of having tuberculosis, has a DEFT value of 1.1349 and a DEFF value of 1.2880. The standard error and confidence interval are 1.1349 times larger

than they would be for simple random sampling. The overweight, which is associated with decrease in the risk of having tuberculosis, has a DEFT value of 1.2759 and a DEFF value of 1.6266. The standard error and confidence interval are 1.2759 times larger than they would be for simple random sampling. The effect of do not smoke, which is associated with increase in the risk of having tuberculosis, has a DEFT value of 1.2857 and a DEFF value of 1.6529. The standard error and confidence interval are 1.2857 times larger than they would be for simple random sampling. The effect of widowed, which is associated with increase in the risk of having tuberculosis, has a DEFT value of 1.9863 and a DEFF value of 3.9453. The standard error and confidence interval are 1.9863 times larger than they would be for simple random sampling. Men smoking sometimes, which is associated with increase in the risk of having tuberculosis, has a DEFT value of 1.3179 and a DEFF value of 1.7371. The standard error and confidence interval are 1.3179 times larger than they would be for simple random sampling. The number of household members is associated with increase in having tuberculosis, with a DEFT value of 1.2992 and a DEFF value of 1.6876. The standard error and confidence interval are 1.2992 times larger than they would be for simple random sampling. Depending on whether a man has a chronic disease that is associated with decrease in the risk of having tuberculosis, the current age has a DEFT value of 1.8571 and a DEFF value of 3.4490. The standard error and confidence interval are 1.8571 times larger than they would be for simple random sampling. The effect of a number of household members depending on whether men do not smoke, which is associated with decrease in the risk of having tuberculosis, has a DEFT value of 1.4273 and a DEFF value of 2.0372. The standard error and confidence interval are 1.4273 times larger than they would be for simple random sampling.

Furthermore, the design effect is above one, implying that the variance has been underestimated when using logistic models compared to those calculated from complex designs. As a result, survey logistic regression has large standard errors.

Therefore, survey logistic regression models are suitable for this study because they consider survey design features.

Summary

Chapter three presents the GLM, which is employed in the analysis of a binary response. Logistic regression and survey logistic regression were fitted using the 206 SADHS data to analyze risk factors associated with tuberculosis. The logistic regression model assumes that the data was obtained using an SRS. This is not always the case, especial for survey data. In a contract survey logistic regression model incorporates the complexity of the survey design. The logistic regression and survey logistic model finding revealed that the critical risk factors associated with tuberculosis are chronic disease, age, region, education, marital status, times away from home, and the number of household members. For both models number of times away from home was associated with increase in the risk of tuberculosis. Whites and coloreds are less likely to have tuberculosis than Africans. Men from all other regions except Gauteng and North West have higher chances of having tuberculosis than those from Limpopo. Furthermore, both models suggest that poor health increases the risk of having tuberculosis; this indicates a need to educate our young adult men to live a healthy lifestyle. In survey logistic regression model, overweight men were less likely to have tuberculosis.

Moreover, from both models, all the interaction effects were associated with decrease in TB. The interaction term that we studied was the effect of chronic disease and age, smoking status, and the number of household members, age, and the number of times away from home. The design effect used to compare both models suggests that the logistic variance was underestimated compared to a complex survey. The design effect discourages using a logistic regression model for

this dataset; thus, the survey logistic model is suitable for this study. The model fitted based on survey logistic regression is better since it accounts for the complexity of the design and also relaxes the logistic regression assumes that the observations are independent. However, survey logistic regression has its own limitation. The lack of the Hosmer-Lemeshow test is a drawback of this approach. It may not be possible to determine where a model is a good fit or not. Furthermore, this method is only suitable for non-grouped data. GLMM will be used in chapter four to account for variability due to correlation among elements from the same cluster.

Chapter Four

The Generalized Linear Mixed Models

4.1 Introduction

The previous chapter used the Generalized Linear Models (logistic regression and survey logistic regression) to investigate the risk factors associated with tuberculosis in adult men. This chapter provides us with another method for modeling tuberculosis. The generalized linear mixed effect model (GLMM) is an extension of the Generalized Linear Model (GLM) (Nelder & Wedderburn, 1972), which allows modeling non-normal and non-linear data that includes random and fixed effects to incorporate correlations (Breslow & Clayton, 1993). McCulloch & Searle (2003) describes GLMMs as incorporating random effect into the linear predictors' portion of a GLM. The GLMM is a vital model used for inference of the population heterogeneity and problems of over-dispersion, which is used in other research such as epidemiology, ecology, etc. Generalized linear mixed models are advancements of linear mixed models (McCulloch et al., 2008; Verbeke & Molenberghs, 2009; Agresti et al., 2000; Antonio & Beirlant, 2007; Laird & Ware, 1982). GLMM combines LMM, and GLM features, handle a range of response distributions and data when observations are sampled in a group structure instead of independently (Molenberghs & Verbeke, 2006; Waagepetersen, 2007). Different types of responses can be modeled using GLM, such as a binary (McCullagh & Nelder, 1989).

4.2 GLMM Model

GLMM assumes that responses Y_{ij} of y_i are conditionally independent if the probability density of the response is a member of the exponential family, which is given as:

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

The mean for a conditional response $y_{ij}|\gamma_i$ is modeled using a linear predictor as

$$g\left(E(y_{ij}|\gamma_i)\right) = g(E(\boldsymbol{Y}|\boldsymbol{\gamma}))$$
$$= \boldsymbol{X\beta} + \boldsymbol{Z\gamma},$$

where Y is the (n×1) vector of observed data, γ is a (q×1) random effect coefficients vector, γ~N(0, G), and g(.) is a link function. **X** (n×p) is the design matrix for fixed effect coefficients. **G** is a variance-covariance for a random effect, **Z** design matrix for a random effect, and β is fixed effect regression coefficients where *i* represents observations and j cluster.

GLMM parameters can be estimated using a Bayesian approach or classical approach comprising maximum likelihood methods (McCulloch & Neuhaus, 2005). In this research, classical approach will be used to model contributing factors for tuberculosis.

4.3 Estimation

4.3.1 Maximum likelihood

The maximum likelihood technique is used to estimate parameters in a parametric model in order to maximize the likelihood function of the observed data (Searle et al., 2006). GLMM's are derived by integrating the distribution of the random effects (Molenberghs & Verbeke, 2006; Bolker et al., 2009). The likelihood is given by

$$L(\boldsymbol{\beta}, G, \emptyset) = \prod_{i=1}^{N} f_i(\boldsymbol{Y}_i|\beta, \boldsymbol{G}, \emptyset)$$

$$= \prod_{i=1}^{N} \int f_i(Y_i|\beta, \boldsymbol{G}, \emptyset).f(\boldsymbol{\gamma_i}, \boldsymbol{G})\boldsymbol{d}\,\boldsymbol{\gamma_i}$$

4.3.2 Approximation of the integral

Laplace approximation is a widely used method that approximates likelihood function, which is based on integrating the integrand (Jiang, 2007), integrals of the form

$$\int e^{-Q(x)}dx$$

Q(x) is a known unimodal function, and x is a q-dimensional vector of variables (Tuerlinckx et al., 2006). An approximate expression for Q(x) can be obtained by the second-order Taylor series expansion around x.

$$q(x) \approx q(\bar{x}) + \frac{1}{2}(x - \bar{x})'q''(\bar{x})(x - \bar{x}) + \cdots$$

Thus,

$$\int \exp\big(Q(x)\big)\,dx \;\approx (2\pi)^{\frac{q}{2}}\big|Q''(\hat{x})\big|^{-\frac{1}{2}}\exp\,(-Q'(\hat{x}))$$

Since γ~N(0, G), the integral can be expressed as

$$Q(\gamma) = \emptyset^{-1} \sum_{j=1}^{n_i} [y_{ij}(\boldsymbol{X\beta} + \boldsymbol{Z\gamma}) - b(\boldsymbol{X\beta} + \boldsymbol{Z\gamma})] - \frac{1}{2}\gamma' \boldsymbol{G^{-1}}\gamma$$

Higher-order terms in the Taylor expansion can improve the Laplace approximation accuracy (Raudenbush, et al., 2000).

4.3.3 Model Selection

An analysis of the significance of fixed effects parameters of a GLMM is performed as H_0: $\boldsymbol{C\beta} = 0$, C is a matrix of constants of row rank d (Tuerlinckx et al., 2006). The likelihood ratio test, the Wald test, or the score test are used to test a hypothesis. Test statistics are also employed to assess the significance of the random effect, which equals testing whether the corresponding variance components in G are zero. The test statistics follow a mixture of χ^2-distributions (Self & Liang, 1987; Stram & Lee, 1994; Zhang & Lin, 2008). Another model selection is Akaike Information Criterion (AIC), just like GLM.

4.3.4 Under-dispersion and Over-dispersion

Over-dispersion means that there exists more variability in the data than expected. Populations are mostly heterogeneous, which makes the over-dispersion to be common in data analysis. This dispersion occurs when the observed variance is greater than the variance of the theoretical model. Whereas under-dispersion occurs when the data exhibit less variance than expected.

4.4 Application of GLMM

PROC GLIMMIX was used in SAS 9.4 to make inferences for GLMM to the SADHS 2016 data. A random _residual_ statement was used since the distribution of the model is binomial, and the variance function of the binomial is $a(\mu) = \mu(1 - \mu)$ (SAS Institute Inc, 2020). The maximum likelihood estimation technique was used in fitting the GLMM. The same indicators and two-way interaction effects fitted in logistic and survey logistic regression were also incorporated into the GLMM. The model fit is summarized in Table 4. 1, where – 2 log L is 1727. It also observed that the Pearson chi-square value is 0.93. This suggests there is no under-dispersion or over-dispersion in the data, implying that residual variability has been adequately modeled.

Table 4. 1 Model Fit

-2 log L(SM1105 | r. effects)	1727.95
Pearson Chi-Square	3855.91
Pearson Chi-Square / DF	0.93

shows the type 3 analysis effect for GLMM. All the two-way interaction effects were significant at a 5% level of significance. The main effect of wealth index, smoking status, and the number of household members were the predictor variables that were not significantly associated with the risk of having tuberculosis. All other covariates were statistically significantly associated with the risk of having tuberculosis.

Table 4. 2 Type 3 Analysis Effect for GLMM

Main effect	**Num DF**	**Den DF**	**F-value**	**Pr>F**
Chronic disease	1	4141	24.80	<.0001
Current age	1	4141	4.27	0.0388
Region	8	4141	8.39	<.0001
Education Level	3	4141	3.94	0.0081
Ethnicity	4	4141	3.83	0.0041
Times away from home	1	4141	4.83	0.0281
Wealth index	4	4141	1.69	0.1488
Marital status	5	4141	3.34	0.0052
Health	3	4141	5.54	0.0009
Weight	4	4141	2.67	0.0308
Smoking status	2	4141	2.01	0.1338
Number of Household members	1	4141	0.22	0.6384
Interaction effect				
Current age and chronic disease	1	3435	23.68	<.0001
No of household members and smoking status	2	3435	4.92	0.0074
Current age and times away from home	1	3435	2.71	0.0995

Table 4.2. The corresponding odds ratio is 8.1019 for chronic disease. This implies that a man with chronic disease is 8.1019 more likely to have tuberculosis than those without chronic disease. A one-unit increase in a man's age increases the risk of having tuberculosis by (1.0325-1) %=3.25%. Under region Western Cape, Eastern Cape, Northern Cape, Free States, KwaZulu-Natal, and Mpumalanga were found to be associated with increase in the risk of having tuberculosis compared to Limpopo with the corresponding odds ratio 6.5594, 4.1616, 4.3103, 4.2495, 2.8818, and 2.8867 respectively. This implies that men from all the provinces mentioned are more likely to have TB than men from Limpopo.

Men with primary education (OR =1.5385) are more likely to have tuberculosis than men with secondary education. The corresponding odds ratio for White and Colored men are 0.0930 and 0.4903, respectively. This implies that whites are (1-0.0930) =90.7% less likely to have tuberculosis than blacks, and coloreds are (1-0.4903)

%=50,97% less likely to have tuberculosis than black men. The number of times being away from home is associated with increase in the risk of having tuberculosis. This implies that the number of times being away from home the odds of having TB by (1.0293-1)%= 2.93%.

Furthermore, richer men from South Africa have statistically associated with decrease in the TB risk than men with a middle wealth index. The corresponding odds ratio is 0.6024. This implies that richer men are (1-0.6024)*100%=39.76% times less likely to be at risk of having tuberculosis than the middle wealth index. The main effect for divorce and widowed men is associated with increase in the risk of having TB compared to those who were never in marriage union, with an odds ratio of 3.2540 and 2.2427, respectively. This implies divorce men are 3.2540 times more likely to have TB than men who were never in a marriage union, and widowed men are 2.2427 times more likely to be at risk of having tuberculosis than men who were never in a marriage union. Men with poor health are associated with increase in the risk of having tuberculosis compared to men with good health. The corresponding odds ratio is 2.1813. This implies the odds of having tuberculosis for men with poor health status is 2.1813 times more likely compared to men with good health. Overweight is associated with decrease in the risk of having TB compared to underweight with OR =0.4263. This implies that overweight men are (1-0.4262)*100%=57.38% less likely to have TB compared to underweight men. The number of households has an associated with increased in the risk of tuberculosis as the number of household members increases the odds of having tuberculosis for men increases by (1.1199-1) % = 11.99%.

Table 4.3 also shows that the covariates that were not found to be significant were region (North West & Gauteng), an education level (no education), race (Indian & others), wealth index (poorest, poor, and richest), weight, those who smoke sometimes, health (average and excellent).

Table 4. 3 Fixed Effect Solution for GLMM

Indicator	Estimate	S.E	OR	P-Value
Intercept	-5.2907	0.4136	<.0001	0.0050
Chronic disease (ref = NO)				
Yes	2.0921	0.4201	<.0001	8.1019
Current age	0.0318	0.0059	<.0001	1.0323
Region (ref = Limpopo)				
Western Cape	1.8809	0.3724	<.0001	6.5594
Eastern Cape	1.4259	0.2801	<.0001	4.1616
Northern Cape	1.4610	0.3239	<.0001	4.3103
Free State	1.4468	0.3142	<.0001	4.2495
Kwazulu-Natal	1.0584	0.2975	0.0004	2.8818
North West	0.3574	0.3255	0.2723	1.4296
Gauteng	-0.5263	0.4495	0.2416	0.5908
Mpumalanga	1.0601	0.3005	0.0004	2.8867
Education Level (ref = Secondary)				
No education	-0.0788	0.2490	0.7518	0.9242
Primary	0.4308	0.1617	0.0077	1.5385
Higher	0.4899	0.2709	0.0706	1.6322
Ethnicity (ref = Black/African)				
White	-2.3755	0.7341	0.0012	0.0930
Colored	-0.7127	0.2729	0.0090	0.4903
Indian/Asian	-13.9902	463.87	0.9759	0.0000
Other	-13.8134	26789.27	0.9959	0.0000
Times away from home	0.0289	0.0132	0.0281	1.0293
Wealth Index (ref = Middle)				
Poorest	0.0414	0.1896	0.8268	1.0424
Poor	-0.0174	0.1876	0.9259	0.9827
Richer	-0.5068	2151	0.0185	0.6024
Richest	-0.1481	0.2610	0.5705	0.8623
Marital status (ref = Never in union)				
Married	0.0257	0.1912	0.8930	1.0261
Living with partner	0.2416	0.2154	0.2620	1.2733
Widowed	0.8077	0.3162	0.0107	2.2427
Divorced	1.1799	0.4049	0.0036	3.2540
Separated	0.4508	0.3365	0.1805	1.5696
Health (ref = Good)				
Poor	0.7799	0.2082	0.0002	2.1813
Average	0.2190	0.1609	1737	1.2448
Excellent	-0.2330	0.2452	0.3421	0.7922
Weight (ref =Normal)				
Underweight	-0.3377	0.1784	0.0585	0.7134
Overweight	-0.8526	0.3623	0.0187	0.4263
Obese	1.2451	0.8122	0.1253	3.4733
Don't know	-0.3059	0.6321	0.6285	0.7365
Smoking Status (ref = Everyday)				
Do not smoke	0.4582	0.2382	0.0545	1.5812
Sometimes	0.5350	0.4928	0.2778	1.7074
Number of household members	0.1132	0.0319	0.0004	1.1199
Interaction effect				
Chronic disease and age (ref = No)				
Having chronic disease and current age	-0.0395	0.0081	<.0001	0.9613
No of household members and smoking status(ref=Everyday)				
No of household members and do not smoke	-0.1306	0.0432	0.0025	0.8776
No of household members and Sometimes	-0.1561	0.1074	0.1463	0.8555
Current age and times away from home	-0.0006	0.0003	0.0995	0.9994

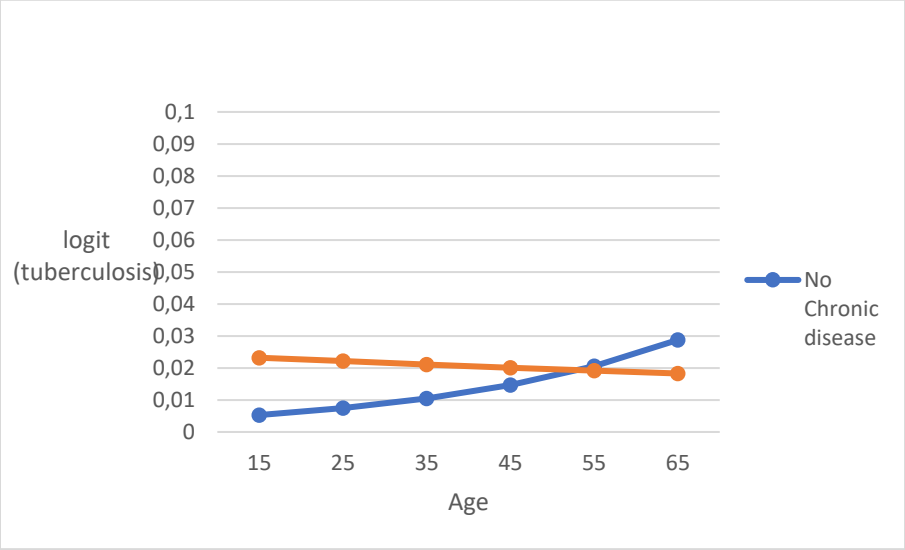

Figure 4. 1 Interaction effect for the current age and chronic disease for GLMM

Figure 4. 1 indicates that men aged between 15 and 54 who have the chronic disease have higher chances of having tuberculosis than men who do not have a chronic disease. In contrast, men above 55 years with chronic disease are less likely to have TB than men without chronic disease. Figure 4.2 shows that as the number of household members' increases, the probability of having tuberculosis increases for daily smokers.

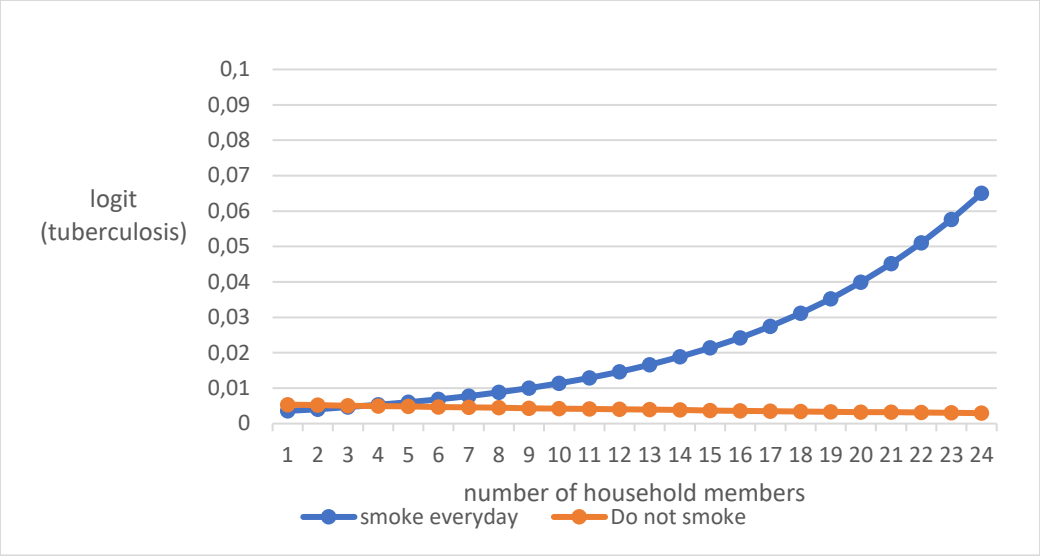

Figure 4. 2 Interaction for number of household members and smoking status for GLM

Summary

GLMM was employed to examine the risk factors of tuberculosis among South African men in 2016. GLMM extends the survey logistic regression by incorporating the random effect, which takes into account the variation that primary sampling units may have. However, the results from GLMM were consistent with the finding from the survey logistic regression. The next chapter we will fit GAMM with a semi-parametric model. This model will study the parametric effect and non-parametric effects.

Chapter Five

Generalized Additive Mixed Models

Statistical methods discussed in the previous chapters like GLM (logistic regression model and survey logistic regression) and GLMM assume a linear form for the covariate effects. The effect that a province has on the likelihood of developing tuberculosis is fixed for this model. On the other hand, neighboring provinces may have a similar effect compared to non-neighboring due to spatial autocorrelation. In this chapter, we will study an extension of the generalized linear mixed model, which is the generalized additive mixed model, a model in which parametric fixed effects can be modeled non-parametrically by incorporating smooth additive function (Hastie & Tibshirani, 1986; Hastie & TibshiraniR., 1990; Chen, 2000; Lin & Zhang, 1999; Breslow & Clayton, 1993). The GAMM explores the non-linear relationship between the dependent variable and covariates. P-spline and B-spline approximations approximate unknown smooth functions represent nonlinear effects in GAMM (Eilers & Marx, 1996).

5.1 Additive model

The association between the response and the predictors may not always be linear. When there is no such linearity, we need additive models. The additive is the generalization of linear regression models. The significant feature additive over linear model is that additive models involve a sum of smooth function of covariates in the predictor effect (Hastie & Tibshirani, 1986). The additive model (AM), suggested by (Friedman & Stuetzle, 1981) and developed by (Hastie & Tibshirani, 1990), is described as follows,

given n points, $\{(X_i, Y_i) : i = 1,2,\ldots,n\}$ then

$Y_i = \alpha + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i,$

where Y_i is a response variable, x_{ij} are the explanatory variables, α is the model intercept, ε_i is the error term and $f_j(\cdot)$ are the unknown smoothing functions (Hastie & Tibshirani, 1987). To obtain the best fit model for our data, we generally want a smooth function approximation.

5.1.2 Smoothing

Statistical smoothing is used to approximate the curve $f_j(\cdot)$, usually referred to as a mean of the outcome variables near point k.

Types of smoothing

Running mean smoother

In the running mean smoother, the smooth is estimated by averaging the observations around point x_i, n_i, in the neighborhood (Buja et al., 1989).

$$f(x_i) = \sum_{i \in N(x_i)} \frac{y_i}{x_i}$$

Running smoothers is linked with biased functions.

Running line smoother

The running line considers the bias problem in the running mean smoother by fitting a smooth curve to data points using least-squares in asymmetric nearest neighborhood.

$f(x_i) = \hat{\beta}_0 + \hat{\beta}_i x_i,$

where $\hat{\beta}_0 \; and \; \hat{\beta}_i$ are ordinary least squares.

Kernel smoothers

Kernel smoother calculates the estimates for each target value by using a defined set of local weights. In general, when moving away from a target point, a kernel smoother uses weights that decrease smoothly. To calculate an estimate of x_0, the weight at the j^{th} given point is defined by

$$S_{0j} = \frac{c_0}{\lambda} d\left(\frac{x_0 - x_j}{\lambda}\right) \qquad (5.\ 1)$$

where *d(t)* is an even decreasing function in t, λ is the bandwidth and c_0 is the constant.

Common choices of $d(\cdot)$ are:

- Epanechnikov kernel (Hardle, 1990)

$$d(t) = \begin{cases} \frac{3}{4}(1 - t^2), & \text{for } |t| \leq 1; \\ 0 & \text{otherwise} \end{cases}$$

- Gaussian Kernel (Buja et al., 1989) (Buja, et al., 1989)

$$d_\lambda(x_0, x_j) = \exp\left(-\left(\frac{x_0 - x_j}{\lambda}\right)^2\right)$$

The kernel smooth is given by

$$s(x_0) = \frac{\sum_{i=1}^{n} d\left(\frac{x_0 - x_j}{\lambda}\right) y_i}{\sum_{i=1}^{n} d\left(\frac{x_0 - x_j}{\lambda}\right)}$$

Natural cubic spline

Suppose we have a model defined as:

$$Y_i = g(x_i) + e_i \qquad (5.2)$$

$g(x_i)$ is a nonparametric smooth function that must be estimated. This smooth function defines the regression function of y on x. The penalized sum of squares (PSS) can be minimized by:

$s_\lambda(g) = \sum_{i=1}^{n}(Y_i - g(x_i))^2 + \lambda \int_a^b (g''(x))^2 dx,$

where λ is a fixed constant and $a \leq x_1 \leq \cdots \leq x_n \leq b$ (Hastie & Tibshirani, 1986) for detailed information.

P-splines

B-spline can be used as an alternative method of representing cubic splines. It is essential to choose well the smoothing parameter in spline smoothing. Cross-validation involves removing one data point at a time and choosing λ at which the missing data point is most accurately predicted by the remaining data (Silverman, 1985). Suppose g_λ^{-1} is the smoothing spline derived from all data points excluding (x_i, y_i), the cross-validation choice of λ is the value that minimizes the following cross-validation score (Ongoma, 2017).

$$cv(\lambda) = n^{-1} \sum_{i=1}^{n} [Y_i - g_\lambda(x_i)]^2$$

The average squared errors at design points $x_1, \cdots, x_n$ are minimized by generalized cross-validation (GCV) as

$$GCV(\lambda) = \frac{n^{-1}RSS(\lambda)}{[1 - n^{-1}trA(\lambda)]^2}$$

$RSS(\lambda) = \sum_{i=1}^{n}[Y_i - g_\lambda(x_i)]^2$ is the residual sum of squares. The function $n^{-1}trA(\lambda)$ is the mean value of the matrix $A(\lambda) = n^{-1}G(x_i, x_j)$ where $G(\cdot)$ is the weight function that relies on the design points and smoothing parameter (Ongoma, 2017; Hastie & Tibshirani, 1990).

5.2 Generalized Additive model

The generalized additive models (GAM) follow from additive models as the extension of the GLM that incorporates an additive term in the linear predictor, and the response may belong to any exponential family distribution (Hastie & Tibshirani, 1990). Let $y_i$be the response variable whose distribution belongs to the exponential family, and then the generalized additive model is defined by:

$g(\mu_i) = \eta_i = \boldsymbol{X_i^*\theta} + \sum_j f_j(x_j)$,

where $g(\mu)$ is a one-to-one function, $\boldsymbol{X_i^*}$ is the i^{th} row matrix of the model, $\boldsymbol{\theta}$ is the parametric estimates vector. $f_j(\cdot)$ are the smooth functions. Smooth functions allow the flexible specification of the dependence of the response on the covariates, then a parametric relationship.

Application of generalized additive model

The final result from GAM can be represented as follows.

$$g(\mu_j) = \beta_0 + \beta_1 Region_j + \beta_2 Education_j + \beta_3 Ethnicity_j + \beta_4 Wealth_j + \beta_5 Marital\ status_j + \beta_6 Health_j + \beta_7 Weight_j + s_1(Age_j) + s_2(number\ of\ household\ members_j) + s_3(No.of\ times\ away\ from\ home_j) + s_4(Age_j) * (No.of\ times\ away\ from\ home_j) + \varepsilon_{0j}$$

Table 5. 1 The parameter estimates for tuberculosis among adult men for fixed effect of GAM

Indicator	Estimate	S.E	OR	P-Value
Intercept	-3.4900	0.3620	0.0305	<2e-16***
Chronic disease (ref = NO)				
Yes	0.1336	0.1676	1.1429	0.4256
Region (ref = Limpopo)				
Western Cape	1.8240	0.3830	6.1966	2.02e-06***
Eastern Cape	1.3370	0.2882	3.8076	3.47e-06***
Northern Cape	1.3670	0.3344	3.9236	4.33e-05***
Free State	1.3480	0.3239	3.8497	3.16e-05***
Kwazulu-Natal	1.0130	0.3064	2.7539	0.0052**
North West	0.2724	0.3358	1.3131	0.4117
Gauteng	-0.5993	0.4640	0.5492	0.1965
Mpumalanga	0.8621	0.3087	2.3681	0.0052**
Education Level (ref = Secondary)				
No education	0.0998	0.2532	1.1049	0.6935
Primary	0.4319	0.1676	1.5402	0.0099**
Higher	0.4616	0.2818	1.5866	0.1015
Ethnicity (ref = Black/African)				
White	-2.0630	0.7549	0.1271	0.0063**
Colored	-0.7521	0.2831	0.4714	0.0079**
Indian/Asian	-41.8900	8.52E+06	0.0000	0.9999
Other	-41.7400	4.75E+07	0.0000	0.9999
Wealth Index (ref = Middle)				
Poorest	0.0570	0.1977	1.0587	0.7730
Poor	-0.0225	0.1952	0.9777	0.9081
Richer	-0.5619	0.2241	0.5701	0.0122*
Richest	-0.2764	0.2721	0.7585	0.3097
Marital status (ref = Never in the union)				
Married	-0.2697	0.2004	0.7636	0.1785
Living with partner	-0.0627	0.2339	0.9393	0.7888

Widowed	0.7312	0.3230	2.0776	0.0236
Divorced	0.9768	0.4247	2.6559	0.0214*
Separated	0.3105	0.3465	1.3641	0.3701
Health (ref = Good)				
Poor	0.7386	0.2157	2.0930	0.0006***
Average	0.2135	0.1688	1.2380	0.2006
Excellent	-0.1821	0.2555	0.8335	0.4761
Weight (ref =Normal)				
Underweight	-0.3579	0.1854	0.6991	0.0535
Overweight	-0.8785	0.3765	0.4154	0.0196*
Obese	1.2880	0.8400	3.6255	0.1252
Don't know	-0.3458	0.6593	0.7077	0.5999
Smoking Status (ref = Everyday)				
Do not smoke	-0.0011	1472	0.9989	0.9364
Sometimes	-0.1636	0.2930	0.8491	0.5775

Table 5. ***3*** indicates that under region Western Cape, Eastern Cape, Northern Cape, Free States, KwaZulu-Natal, and Mpumalanga were associated with increase in the risk of having tuberculosis when all these regions are compared to Limpopo. The corresponding odds ratio is 6.1966, 3.8076, 3.9236, 3.8497, 2.7539, and 2.3681, respectively. This implies that men from all the provinces mentioned are more likely to have TB than men from Limpopo.

Men with primary education 8.10 (OR =1.5402, p-value=0.009952) are more likely to have tuberculosis than men with secondary education. White and Coloreds are associated with decrease in the risk of having tuberculosis compared to black men. The corresponding odds ratio are 0.1271 and 0.4714, respectively. This implies that whites are (1-0.1271) %=87.29% less likely to have tuberculosis than blacks, and coloreds are (1-0.4717)%=52.83% less likely to have tuberculosis than black men.

Furthermore, richer men from South Africa are 0.5701 times less likely to be at risk of having tuberculosis than the middle wealth index. The parametric effect for divorce and widowed men is associated with increase in the risk of having TB compared to those who were never in union men, with an odds ratio of 2.6559 and

2.0776, respectively. This implies divorce men are 2.6559 times more likely to have TB than men who were never in a marriage union, and widowed men are 2.0776 times more likely to be at risk of having tuberculosis than men who were never in a union. Men with poor health are associated with increase in the risk of having tuberculosis compared to men with good health. The corresponding odds ratio is 2.0930. This implies the odds of having tuberculosis for men with poor health status is 2.0930 times more likely compared to men with good health. Overweight is associated with decrease in the risk of having TB compared to underweight with OR =0.4154. This implies that overweight men are (1-0.4154) %=58.46% less likely to have TB compared to underweight men.

Table 5. 2 Approximate significance of smooth terms for GAM

Smooth terms	**Edf**	**F-value**	**P-value**
S(number of household member)	5.0280	15.8690	0.0163*
s(current age)	1.9790	6.8440	0.0569
s(number of times away from home)	1.6950	0.9410	0.5763
s(age,no of times away from home)	3.0710	8.4170	0.0001***

Table 5. 4 the statistic test 15.869 with 5.028 degrees of freedom (p-value=0.016316) against the assumption that the number of household members is linearly associated with TB risk. The statistic test is 8.417 with 3.071 degrees of freedom with high significance (p-value=0.000119) against the assumption that the interaction of age and number of times away from home is linearly associated with the risk of tuberculosis. Figure 5. 2 shows the smooth term and confidence interval. The number of household members is unreliably estimated for large household sizes. This is because such household with more than 10 members are rare. The risk of having tuberculosis increases with the number of household members until ten members.

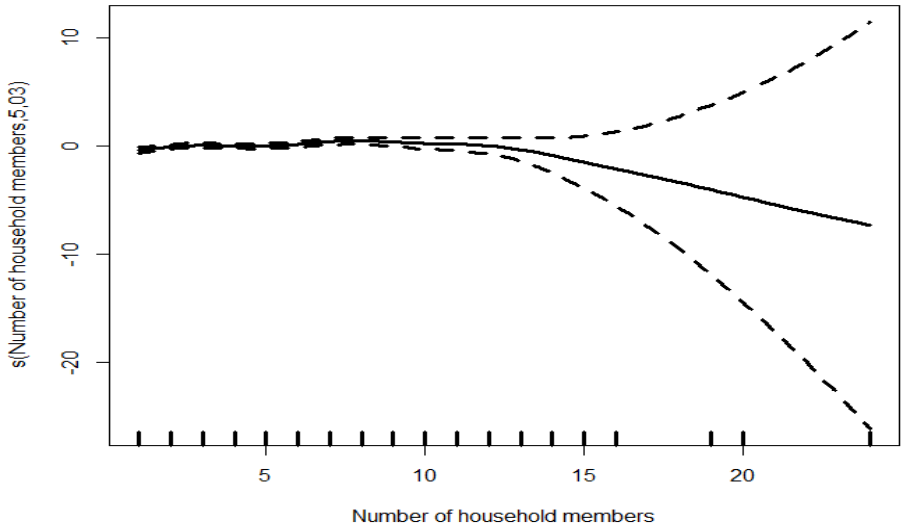

Figure 5. 1 Smoothing components for TB with number of household members

5.3 Generalized additive mixed model

The generalized additive mixed model (GAMM) is an extension of the GLMM statistical method. Similar to GLM and GLMM, GAMM can model the probability of a k^{th} adult man residing in household j and province i nested within p^{th} stratum developing tuberculosis as $P(Y_{ijkp} = 1) = \pi_{ijkp}$. The GAMM has the following structure

$$logit(\pi_{ijkp}) = \boldsymbol{x}'_{ijkp}\beta + \sum_{r=1}^{P} f_r(z_{ijkp}) + f_{spat}(s_i), \qquad (5.1)$$

where β vector of fixed effect of the predictors, $f_r(.)$ is a smooth arbitrary function of the predictors and $f_{spat}(s_i)$ is the non-linear spatial effect.

5.4 Estimation

Smoothing spline estimators and linear mixed models are closely related (Green & Silverman, 1993; Verbyla et al., 1999; Wang, 1998). Using penalized splines (P-splines) with B-splines basis functions, smooth functions f_r are estimated (Eilers & Marx, 1996). A spline is defined as a linear combination of $M_r = n_r + v$, B-spline basis function B_{rm} and regression coefficient of α_{rm} as

$$f_r(z_r) = \sum_{m=1}^{M_r} \alpha_{rm} B_{rm}(z_r)$$

Approximating the smooth function involves choosing the number of knots (where knots are $z_r^{\{min\}} < \xi_{ro} < \ldots. \xi_{rnr} < z_r^{\{max\}}$ equally spaced). The choice of the number of knots is that too many knots can lead to curves that over-fit the data, resulting in too rough functions. However, too few results do not always capture the variability in the data (Fahrmeir et al., 2004). Twenty to forty equally spaced knots ensure flexibility (ref). Thus, the penalized likelihood estimation is given as

$$P(\lambda_r) = \frac{1}{2}\lambda_r \sum_{m=v+1}^{M_r} (\Delta^v \alpha_{rm})^2,$$

where λ_r is the smooth parameter and Δ^v is the differencing operator.

Spatial effects represent the effect of geographic properties, including spatial autocorrelation. The spatial effect is divided into two spatially correlated and spatially uncorrelated. For the spatially correlated which assume conditional distributions is Gaussian, given by:

$$f_{str}(s_i)|f_{str}(s_j), i \neq j, \sim N\left(\frac{1}{n_{s_i}} \sum_{s_j \in N_i} f_{str}(s_i), \frac{1}{n_{s_i}\tau_{str}^2}\right),$$

where n_{s_i} number of neighbors of region s_i and conditional mean $f_{str}(s_i)$ is a mean of the function evaluations $f_{str}(s_j)$ of neighboring regions. Variance component *tau* accounts for spatial variability between regions, and it is used to capture the degree to which the spatial structure explains variation. The uncorrelated spatial effect is included in GAMM as an independently and identically random distribution effect, i.e

$$f_{unstr}(s_i) \sim N\left(0, \frac{1}{\tau^2_{unstr}}\right)$$

The overview of unstructured and structured spatial effects can be found here (Schabenberger & Gotway, 2017; Kneib et al., 2008).

Restricted maximum likelihood (REML)

The REML approach is a modification of ML pioneered by (Patterson & Thompson, 1971). REML is used as an alternative to provide unbiased estimates of variance components by incorporating the loss of degrees of freedom from an estimation of β. REML is used to transform the data so that fixed effects are removed and then uses transformed data to estimate the variance components (Melesse, 2014). The overview of restricted maximum likelihood methods for GAMM is given by (Wood, 2017).

5.5 Application of GAMM

Introduction

GLM and GLMM were applied in the previous chapters. (Mzolo, 2009) used a similar model based on parametric models in a comparable study. This chapter introduces GAMM to fit the dataset in order to determine the risk factors for tuberculosis in adult men. GAMM combines non-parametric and parametric regression features to fit nonlinear and non-normal data. The primary goal of the GAMM study is to model the effects of current age, number of household members, number of times away from home, and the interaction effect of current age and number of times away from home non-parametrically while other covariates remain parametric. To fit the model, the R-studio will be used.

5.5.1 Results and Interpretation

The R mgcv package in R version 4.2.0 was used to fit the data. The primary sampling units (mv021) variable from the data measured the random effect. The final GAMM was selected as

$$g(\mu_j) = \beta_0 + \beta_1 Region_j + \beta_2 Education_j + \beta_3 Ethnicity_j + \beta_4 Wealth_j +$$
$$\beta_5 Marital\ status_j + \beta_6 Health_j + \beta_7 Weight_j + s_1(Age_j) +$$
$$s_2(number\ of\ household\ members_j) +$$
$$s_3(No.of\ times\ away\ from\ home_j) + s_4(Age_j *$$
$$No.of\ times\ away\ from\ home_j) + \varepsilon_{0j},$$

where $g(\mu_j)$ is the logit link function, $\beta's$ are parametric coefficients, $S's$ are smooth functions, and $\varepsilon_{0j}$is the random effect. Additive models are often estimated

using kernel smoothers, locally-weighted running line smoothers, and cubic smoothing splines (Härdle, 1990; Hastie & Tibshirani, 1990; Ruppert et al., 2003). Table 5. ***3*** represents the parametric estimates for the generalized additive mixed model.

Table 5. 3 The parameter estimates for tuberculosis among adult men for fixed effect of GAMM

Parameter	Estimate	S.E	OR	P-Value
Intercept	-1.6960	0.3751	0.1834	<0.001***
Chronic disease (ref = NO)				
Yes	0.1292	0.1667	1.1379	0.4385
Region (ref = Limpopo)				
Western Cape	-0.4710	0.3256	0.6244	0.1481
Eastern Cape	-0.4261	0.3213	0.6531	0.1848
Northern Cape	-0.4518	0.3403	0.6365	0.1844
Free State	-0.8032	0.3411	0.4478	0.0186*
Kwazulu-Natal	-1.5280	0.3677	0.2169	<0.001***
North West	-2.4100	0.4834	0.0898	<0.001***
Gauteng	-0.9428	0.3426	0.3895	0.00595**
Mpumalanga	-1.8080	0.3924	0.1639	<0.001***
Education Level (ref = Secondary)				
No education	0.1211	0.2612	1.1287	0.6430
Primary	0.4424	0.1716	1.5564	0.0099**
Higher	0.4367	0.2865	1.5476	0.1275
Ethnicity (ref = Black/African)				
White	-2.0240	0.7579	0.1321	0.0076**
Colored	-0.7247	0.3005	0.4845	0.0159*
Indian/Asian	-18.9200	0.7261	6.07e-9	0.0092**
Other	-18.7900	0.4353	6.91e-9	<0.001***
Wealth Index (ref = Middle)				
Poorest	0.0415	0.2026	1.0423	0.8378
Poor	-0.0255	0.1994	0.9749	0.8984
Richer	-0.5507	0.2289	0.5765	0.0162*
Richest	-0.2520	0.2792	0.7772	0.3666
Marital status (ref = Never in a union)				
Married	-0.2408	0.2037	0.7860	0.2373
Living with partner	-0.3147	0.2341	0.7300	0.8931
Widowed	0.7293	0.3865	2.0736	0.0592
Divorced	0.9733	0.4368	2.6467	0.0259*
Separated	0.3137	0.3627	1.3685	0.3871
Health (ref = Good)				
Poor	0.7517	0.2173	2.1206	0.0005***
Average	0.2088	0.1675	1.2322	0.2128

Excellent	-0.1693	0.2577	0.8443	0.5113
Weight (ref =Normal)				
Underweight	-0.3628	0.1863	0.6957	0.0516
Overweight	-0.8774	0.3803	0.4159	0.0211*
Obese	1.2170	0.8478	3.3770	0.1513
Don't know	-0.3244	0.6600	0.7229	0.6231
Smoking Status (ref = Everyday)				
Do not smoke	-6.334e-04	0.1488	0.9994	0.9966
Sometimes	-0.1298	0.3039	0.8783	0.6693

Table 5. ***3*** indicates that under region Free States, KwaZulu-Natal, North West, Gauteng, and Mpumalanga were associated with decrease in the risk of having tuberculosis when all these regions are compared to Limpopo. The corresponding odds ratio 0.4478, 0.2169, 0.0898, 0.3895, and 0.1639 respectively. This implies that men from all the provinces mentioned are more likely to have TB than men from Limpopo.

Men with primary education (OR =1.5564) are more likely to have tuberculosis than men with secondary education. White and Colored men have a are associated with decrease in the risk of having tuberculosis compared to black men. The corresponding odds ratio are 0.1321 and 0.4845, respectively. This implies that whites are (1-0.1323) %=86.77% less likely to have tuberculosis than blacks, and coloreds are (1-0.4845) %=51.55% less likely to have tuberculosis than black men.

Furthermore, richer men from South Africa have associated with decrease in TB risk compared to men with a middle wealth index. The corresponding odds ratio is 0.5763. This implies that richer men are (1-0.5763) *100%=42.37% times less likely to be at risk of having tuberculosis than the middle wealth index. The parametric effect for divorce is associated with increase in the risk of having TB compared to those who were never in union men, with an odds ratio of 2.6453 and 2.0732, respectively. This implies divorce men are 2.6467 times more likely to have TB than men who were never in a union. Men with poor health are associated with increase

in the risk of having tuberculosis compared to men with good health. The corresponding odds ratio is 2.1206. This implies the odds of having tuberculosis for men with poor health status is 2.1206 times more likely compared to men with good health. Overweight is associated with decrease in the risk of having TB compared to underweight with OR =0.4159. This implies that overweight men are (1-0.4159) *100= 58.41% less likely to have TB compared to underweight men.

Table 5. 4 Approximate significance of smooth terms for GAMM

Smooth terms	**Edf**	**F-value**	**P-value**
S(number of household member)	2.0690	4.1350	0.0183*
S(sampling units)	9.636e-10	1.0000	0.2419
s(current age)	1.0220	0.4770	0.4903
s(number of times away from home)	1.0200	0.2570	0.6166
s(age,no of times away from home)	6.3890	1.7120	3.47e-08***

Table 5. 4 the statistic test 4.1350 with 2.0690 degrees of freedom (p-value=0.0183) against the assumption that the number of household members is linearly associated with TB risk. The statistic test is 1.7120 with 6.3890 degrees of freedom with high significance (p-value=3.47e-08) against the assumption that the interaction of age and number of times away from home is linearly associated with the risk of tuberculosis. Figure 5. 2 shows the smooth term and confidence interval. The number of household members has some quadratic effect, the risk of having tuberculosis increases with the number of household members until eight members.

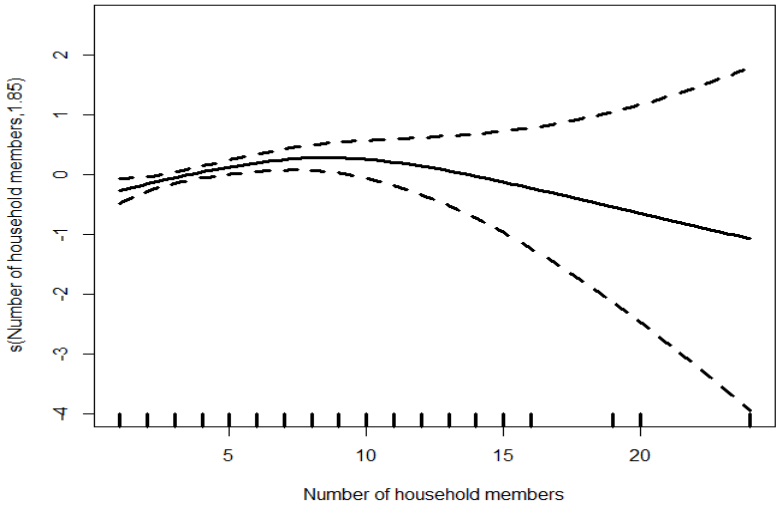

Figure 5. 2 Smoothing components for tuberculosis with number of household members

The Figure 5.3 shows the normal Q-Q plot of studentized residuals. Since the data is large, we are satisfied with the fact that the distributions in each plot follow the distributional assumption.

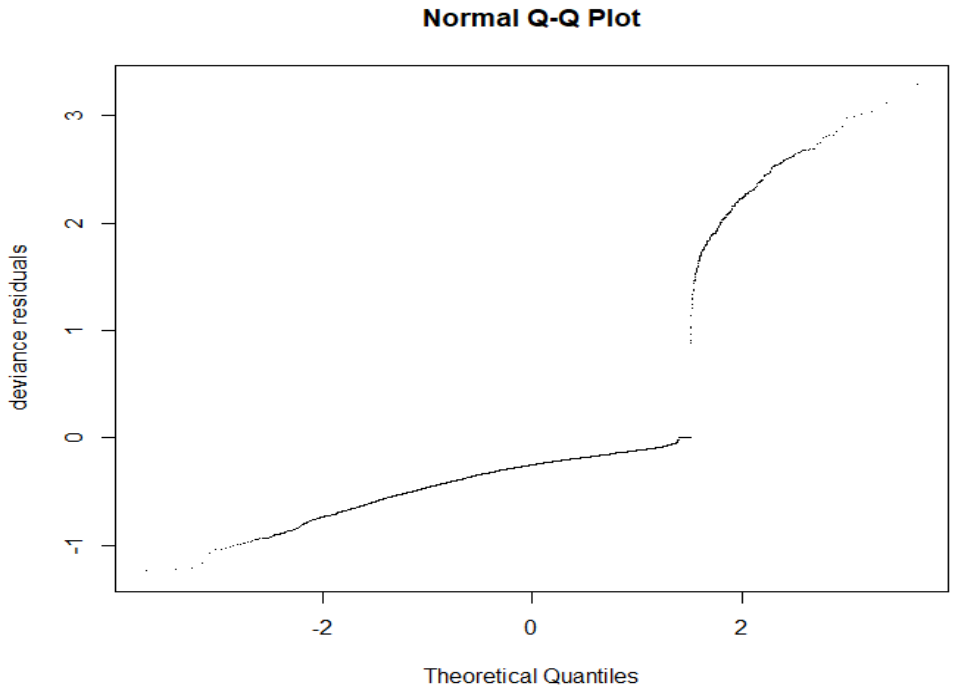

Figure 5. 3 Q-Q plots of conditional studentized residuals

Comparing GAM and GAMM

Table 5. 5 Comparison of GAM and GAMM

	GAM			GAMM		
Indicator	**Estimate**	**S.E**	**P-Value**	**Estimate**	**S.E**	**P-Value**
Intercept	-3.4900	0.3620	<0.001***	-1.6960	0.3751	<0.001***
Chronic disease (ref = NO)						
Yes	0.1336	0.1676	0.4256	0.1292	0.1667	0.4385
Region (ref = Limpopo)						
Western Cape	1.8240	0.3830	2.02e-06***	-0.4710	0.3256	0.1481
Eastern Cape	1.3370	0.2882	3.47e-06***	-0.4261	0.3213	0.1848
Northern Cape	1.3670	0.3344	4.33e-05***	-0.4518	0.3403	0.1844
Free State	1.3480	0.3239	3.16e-05***	-0.8032	0.3411	0.0186*
Kwazulu-Natal	1.0130	0.3064	0.0052**	-1.5280	0.3677	<0.001***
North West	0.2724	0.3358	0.4117	-2.4100	0.4834	<0.001***
Gauteng	-0.5993	0.4640	0.1965	-0.9043	0.3426	0.00595**
Mpumalanga	0.8621	0.3087	0.0052**	-1.8080	0.3924	<0.001***
Education Level (ref = Secondary)						
No education	0.0998	0.2532	0.6935	0.1211	0.2612	0.6430
Primary	0.4319	0.1676	0.0099**	0.4424	0.1716	0.0099**
Higher	0.4616	0.2818	0.1015	0.4367	0.2865	0.1275
Ethnicity (ref = Black/African)						
White	-2.0630	0.7549	0.0063**	-2.0240	0.7579	0.0076**
Colored	-0.7521	0.2831	0.0079**	-0.7247	0.3005	0.0159*
Indian/Asian	-41.8900	8.52E+06	0.9999	18.9200	7.261	0.0092**
Other	-41.7400	4.75E+07	0.9999	18.7900	4.353	<0.001***
Wealth Index (ref = Middle)						
Poorest	0.05703	0.1977	0.7729	0.04147	0.2026	0.8378
Poor	-0.02253	0.1952	0.9081	-0.0254	0.1994	0.8984
Richer	-0.5619	0.2241	0.0122*	-0.5507	0.2289	0.0162*
Richest	-0.2764	0.2721	0.3097	-0.2520	0.2792	0.3666
Marital status (ref = Never in union)						
Married	-0.2697	0.2004	0.1785	-0.2408	0.2037	0.2373
Living with partner	-0.0627	0.2339	0.7888	-0.3147	0.2341	0.8931
Widowed	0.7312	0.3230	0.0236*	0.7293	0.3865	0.0592
Divorced	0.9768	0.4247	0.0214*	0.9733	0.4368	0.0259*
Separated	0.3105	0.3465	0.3701	0.3137	0.3627	0.3871
Health (ref = Good)						
Poor	0.7386	0.2157	0.0006***	0.7517	0.2173	0.0005***
Average	0.2135	0.1688	0.2006	0.2088	0.1675	0.2128
Excellent	-0.1821	0.2555	0.4761	-0.1693	0.2577	0.5113
Weight (ref =Normal)						
Underweight	-0.3579	0.1854	0.0535	-0.3628	0.1863	0.0516
Overweight	-0.8785	0.3765	0.0196*	-0.8774	0.3803	0.0211*
Obese	1.28800	0.8400	0.1252	1.2170	0.8478	0.1513
Don't know	-0.3458	0.6593	0.5999	-0.3244	0.6600	0.6231
Smoking Status (ref = Everyday)						
Do not smoke	-0.0011	1472	0.9364	-0.0006	0.1480	0.9966
Sometimes	-0.1636	0.2930	0.5775	-0.1298	0.3039	0.6693

The findings from the generalized additive and generalized additive mixed models show some similarity in the results in terms of the p-value. Not all the significant variables in GAM are also significant in GAMM. The standard error of white men increased by 0.39%. The standard error of divorced was increased by 2.8%. Free States, KwaZulu-Natal and Mpumalanga regions standard errors were increased by 5.3%, 20% and 27.2% respectively. The standard errors of significant parameters in the GAMM are higher than the corresponding standard errors of the significant parameters in the GAM, more variables are significant in GAM. This suggests that the GAM may lead to false precisions and estimates. Thus, GAMM takes into account the correlation between the observations is better than GAM.

Summary

The generalized additive mixed model was used to identify risk factors associated with tuberculosis. The effect of age, number of household members, number of times away from home, and interaction effect of age and number of times away from home were analyzed non-parametrically while other covariates were modeled parametrically. The findings from GAMM validate the findings from previous models.

Chapter Six

Discussion and Conclusion

The main purpose of this investigation was to identify risk factors associated with tuberculosis among men in South Africa. This may assist policymakers in making a decision, which will help prevent the increased number of men at risk, thus reducing the number of men at risk with tuberculosis. Different statistical models were used to examine tuberculosis risk factors using the 2016 SADHS. Statistics South Africa and South Africa Medical Research Council were responsible organizations for the survey. The statistical methods used were parametric and semi-parametric. The parametric methods used include logistic regression, survey logistic regression, and generalized linear mixed model. The semi-parametric method used for analysis was the generalized additive mixed model (GAMM).

The response variable is the binary status of tuberculosis in men. The overall percentage of men at risk with tuberculosis is 5.26%. Most men resided in urban areas, were unemployed, drank alcohol, had good health, average weight, and were never in a marriage union. Most men had secondary education, and they were non-smoker. The majority of men were blacks and were from KwaZulu-Natal. Furthermore, the majority of men with tuberculosis are from Eastern Cape Province. Most of the respondents with tuberculosis do not smoke and have high blood pressure.

Chapter three presents the binary logistic regression (without complex design) and survey logistic regression (with complex design). These models have also been used in determining the risk factors associated with tuberculosis. Furthermore, as part of

the modeling process, two-way interaction effects were incorporated. The binary outcome variable is modeled using a logistic regression model, also known as a logit. An analysis of the goodness of fit of the binary logistic regression model was performed using Hosmer and Lemeshow, which indicates how well the model fits the data. Design effects were used to compare parameter estimates obtained from the logistic regression model and survey logistic regression. The design effect values were above one. The design effect values suggest that the variance of logistic regression, which assumes a simple random sampling as a sampling method, was underestimated.

The parameters estimate for survey logistic regression (that incorporates the complexity of the survey design) are different from the estimates obtained when simple random sampling was assumed. However, some parameters were closer to one another.

From the result, it can be noted that only variables that are significant from both models are: chronic disease, current age, region, race, number of times away from home, marital status, weight, and interaction effect of chronic disease and age, and interaction effect of the number of household members and smoking status. We observe that the risk of having tuberculosis increases with a unit increase in age, but infection can occur at any age. Also, chronic disease plays a role in the progression of tuberculosis; the higher body mass index (such as for overweight) plays a role in progression of risk of having tuberculosis. These results are supported by (Buskin et al., 1994; Shetty et al., 2006).

Similar findings with (Shetty et al., 2006) on variable smoking status and alcohol, both of these factors were not statistically significantly associated with the risk of having tuberculosis. Whites and Coloreds have a low risk of having tuberculosis compared to Africans. Both models suggest that the risk of having tuberculosis

increases with the increase in household members and the number of times away from home. This implies that men who are always outdoors are more likely to have tuberculosis. If the men can reduce the number of times being away from their home, their rate of having tuberculosis can also decrease.

Chapter four presents us with GLMM. GLMM considers the possible correlation between observations from the same cluster by using a random effect. While all three methods provide similar results, they assume a linear relationship between response variables and covariates. This may not be the case from the same covariates. In chapter five, the semi-parametric generalized additive mixed model was also fitted to the data.

GAMM is an extension of GLMM, which explores the non-linear relationship between a response variable and predictor variables. Unlike the other three methods, the variables chronic disease, current age, and the number of times away from home did not statistically significantly influence the risk of having tuberculosis.

The findings from this study suggest that policymakers need to focus on significant factors to develop strategies that will reduce the risk of having tuberculosis among adult men in South Africa. Also, discouraging men from being underweight or having a lower body mass index might reduce tuberculosis. The government needs to implement programs targeting these regions; Eastern Cape, Western Cape, Northern Cape, Free States, KwaZulu-Natal, and Mpumalanga, which are highly affected by tuberculosis infection.

In this thesis we investigated and examine risk factors for TB in adult men in South Africa. Hence future research is needed; this study will be extended by considering spatial modeling of the provinces to investigate tuberculosis patterns in each province. The percentage of missing values in this study was less than 5%. In the

future, we will use some techniques to impute missing values, such as multiple imputations.

References

Park H., 2013. An Introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean Academy of Nursing,* 43(2), pp. 154-164.

Agresti , A., Booth, J. G., Hobert, J. P. & Caffo, B., 2000. Random-effects modeling of categorical response data. *Sociological Methodology,* 30(1), pp. 27-80.

Agresti, A., 1990. *Categorical data analysis.* s.l.:Wiley.

Agresti, A., 2002. *Categorical Data Analysis, Second Edition: Wiley Series in Probability and Statistics.*. s.l.:s.n.

Agresti, A., 2007. *An Introduction to Categorical Data Analysis (Wiley Series in Probability and Statistics).* second edition ed. s.l.:Wiley-Interscience.

Agresti, A., 2019. *An introduction to categorical data analysis.*. THIRD EDITION ed. s.l.:John Wiley & Sons.

Akaike, H., 1974. A new loook at the statistical model identification. *IEEE transactions on automatic control.*

Allison, P. D., 2012. *logistic regression using SAS: Theory and application.* s.l.:SAS institute.

Anon., 2020. *psych scene hub.* [Online] Available at: https://psychscenehub.com/psychpedia/odds-ratio-2/

Anthony, B., 2002. Perfoming logistic regression on survey data with new survey logistic procedure. *In Proceedings of the twenty-seventh annual SAS users group international conference,* 258(27).

Antonio, K. & Beirlant, J., 2007. Actuarial statistics with generalized linear mixed models. *Insurance: Mathenatics and Economics,* 40(1), pp. 58-79.

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

Ayele, D. G., Zewotir, T. T. & Mwambi, H. G., 2012. Prevalence and risk factors of malaria in Ethiopia.. *Malaria Journal,* 11(1), pp. 1-9.

Binder, D. A., 1983. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review/Revue internationale de statistique,* pp. 279-292.

Bolker, B. M. et al., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution,* 24(3), pp. 127-135.

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

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

Buja, A., Hastie, T. & Tibshirani, R., 1989. Linear smoothers and additive models. *The Annals of Statistics,* 17(2), pp. 453-510.

Buskin, S. E., Gale, J. L., Weiss, N. S. & Nolan, C. M., 1994. Tuberculosis risk factoors in Adults in King County, Washington. 1988 through 1990. *American Journal of Public Health,* 11(84), pp. 1750-1756.

Chen, C., 2000. Generalized additive mixed models. *Communication in Statistics-Theory and Methods,* 29(5-6), pp. 1257-1271.

Cochran, W. G., 1964. *Sampling techniques.* New York: John Wiley and sons.

Cochran, W. G., 1977. *Sampling Techniques.* 3nd ed. New York: John Wiley and sons.

Coker, R. et al., 2006. Risk factors for pulmonary tuberculosis in Russia: case-control study. *Bmj,* 332(7533), pp. 85-87.

Cramer, J., 2002. The origins of Logistic regression. *Logit models from economic and other fields,* p. chapter 9.

Czepiel, A. S., 2002. *Maxium likelihood estimation of logistic regression models; theory and implementation.* [Online]
Available at: Available at czep. net/stat/mlelr. pdf.

Dobson A. J., Barnett S. G, 2008. *An introduction to Generaalized Linear Models.* London New York: Chapman & hall CRC press.

Dowd, A. C. & Duggan , M. B., 2001. Computing variances from data with complex sampling designs: A comparison of Stata and Spss. *North America Stata Users Group,* p. 12.

Dubos, R. J. & Dubos, J., 1987. *The white plague: tuberculosis, man and society*.. s.l.:Rutgers University Press.

Efron, B., 1980. *The jacknife, the bootstrap and other resampling plans,* California: Standford University.

Eilers, P. H. & Marx, B. D., 1996. Flexible smoothing with B-splines and penalties. *Statistical science,* 11(2), pp. 89-121.

Fahrmeir, L., Kneib, T. & Lang, S., 2004. Penalized structure additive regression for space-time data: a Bayesian perspective.. *Statistica Sinica,* pp. 731-761.

Friedman, J. H. & Stuetzle, W., 1981. Projection pursuit regression.. *Journal of the American Statistical Association,* 76(376), pp. 817-823.

Green, P. J. & Silverman, B. W., 1993. *Nonparametric regression and Generalized Linear Models.* London: CRC Press.

Grima, J., 1965. *gccaz.* [Online] Available at: https://web.gccaz.edu/~johwd63181/MAT142/chapter_4/problems/section%204.2.pdf

Hardle, W., 1990. *Applied nonparametric regression.* New York, USA.: Cambridge University press.

Härdle, W., 1990. Applied nonparametric regression. *Cambridge university press,* Volume 19.

Hastie, T. J. & Tibshirani, R. J., 1990. *Generalized Additive Models.* 1st ed. s.l.:Chapman and Hall/CRC.

Hastie, T. & TibshiraniR., 1990. *Generalized Additive Models.* USA: Chapman and Hall/CRC.

Hastie, T. & Tibshirani, R., 1986. Generalized Addictive models. 1(3), pp. 297-310.

Hastie, T. & Tibshirani, R., 1987. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398), pp. 371-386.

Heering, S. G., West, B. T. & Berlgund, P. A., 2010. *Applied Survey Data Analysis.* 1 ed. s.l.:Chapman and Hall/CRC.

Hosmer, D. W. & Lemeshow, S., 1989. *Applied Logistic regression.* USA: John Wiley and Sons.

Hosmer, D. W. & Lemeshow, S., 2000. *Special topics.* s.l.:Wiley online library.

Hussain, M. S., 2020. *Prevalence of tuberculosis.* [Online] Available at: https://www.barnesandnoble.com/w/prevalence-rate-of-tuberculosis-muhammad-shabbir-hussain/1137783988?A=9786200431073

Jiang, J., 2007. *Linear and Generalized Linear Mixed Models and Their Application.* s.l.:SpringerScience and Business Media.

Kish, L., 1965. *Survey Sampling.* New York: John Wiley and sons.

Kish, L. & Frankel, M. R., 1974. Inference from complex samples. *Journal of the Royal Statistical Society. Series B (Methodological),* 36(1), pp. 1-22.

Kleinbaum, D. G., Klein, M. & Pryor, E. R., 2002. *Logistic regression: A self-learning texxt.* s.l.:s.n.

Kneib, T., Muller, J. & Hothorn, T., 2008. Spatial smooothing techniques for the assessment of habitat suitability.. *Environmental and Ecological Statistics.*, 15(3), pp. 343-364.

Kolenikov, S., 2010. Resampling variance estimation for complex survey data. *The Stata Journal* , 10(2), pp. 165-199.

Laird, N. M. & Ware, J. H., 1982. Random-effect models for longitudinal data.. *Biometrics,* 38(4), pp. 963-974.

Lehohla, P., 2013. *Stats sa republic of south africa.* [Online] Available at: http://www.statssa.gov.za/?p=1335#:~:text=Tuberculosis%20maintained%20its%20rank%20as,deaths%20that%20occurred%20in%202010.
[Accessed 11 April 2013].

Lehohla, P., 2013. *use of health facilities and level of selected health conditions in South Africa: findings from general household survery 2011.* Pretoria: Statistics South Africa.

Lienhardt, C., 2001. From exposure to disease: the role of environmental factors in susceptibility to and developpment of tuberculosis. *Epidemiolgic reviews,* 2(23), pp. 288-301.

Lienhardt, C. et al., 2005. Investigation of the risk factors for tuberculosis: a case-control study in 3 countries in West Africa. *International journal of epidemiology,* 4(34), pp. 914-923.

Lin, X. & Zhang, D., 1999. Inference in generalized additive mixed modelsby using smoothing splines.. *Journal of the royal statistical society: Series b (statistical methodology),* 61(2), pp. 381-400.

Lumley, T. & Scott, A., 2015. AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology,* 3(1), pp. 1-18.

Lu, M. & Yang, W., 2012. Multivariate logistic regression analysis of complex survey data with application to BRFSS data. *Journal of Data Science,* Volume 10, pp. 157-173.

Lu, W. W., 2004. *Confidentiality and variance estimation in complex survey,* s.l.: s.n.

McCullagh, P. & Nelder, J. A., 1983. *Generalized Linear Models.* s.l.:Springer US.

McCullagh, P. & Nelder, J. A., 1989. *Generalized Linear Models.* s.l.:Cyclic Redundancy Check press.

McCulloch, C. E. & Searle, S. R., 2003. *Generalized, Linear and Mixed Models.* New York: John Wiley & sons.

McCulloch, C., Searle, S. & N., N., 2008. *Generalized Linear and Mixed Models.* New Jersey, USA: John Wiley and sons.

McCulloch, P. & Neuhaus, J., 2005. *Generalized Linear Mixed Models.* s.l.:John Wiley & sons.

Melesse, S. F., 2014. *Covariates and latents in growth modelling (doctoral dissertation),* s.l.: s.n.

Melesse, S., Sobratee, N. & Worken, T., 2016. Application to logistic regression statistical technique to evaluate tomato quality subjected to different pre- and post-harvest treatement. *Biological Agriculture & Horticulture,* 32(4), pp. 277-287.

Melesse, S., Sobratee, N. & Workneh, T. S., 2016. Application of logistic regression statistical technique to evaluate tomato quality subjected to different pre-and post-harvest treatments.. *Biological Agriculture & Horticulture.,* 32(4), pp. 277-287.

Moeti, A., 2007. Factors affecting the health status of lesotho.

Molenberghs, G. & Verbeke, G., 2006. *Models for discrete longitudinal data.* s.l.:Springer Science & Business Media.

Mzolo, T., 2009. Estimating risk determinants of HIV and TB in South Africa (Doctoral dissertation).

Narasimhan, P., Wood, J., Maclntyre, C. R. & Mathai, D., 2013. risk factors for TB. *Pulmonary medicine.*

Narayanan, P. R., Kolappan, C., Subramani, R. & Gopi, P. G., 2007. selected biolgical and behavioural risk factors associated with pulmonary tuberculosis. *The international Journal of Tuberculosis and Lung diseases,* 9(11), pp. 999-1003.

National Department of Health , 2019. *South Africa Demographic and Health Survey 2016.* [Online]
Available at: https://dhsprogram.com/pubs/pdf/FR337/FR337.pdf

Nelder, J. A. & Wedderburn, R. W., 1972. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (general),* 135(3), pp. 370-384.

Neovius, M. G., Linné, Y., Barkeling, B. S. & Rossner, S. O., 2004. Sensitivity and specificity of classification systems for fatness in adolescents. *The American journal of clinical nutrition,* 80(3), pp. 597-603.

Olsson, U., 2002. *Generalized Linear Models: An Applied Approach.* s.l.:Studentlitteratur.

Ongoma, D. N., 2017. *Prevalence and risk factors associated with malaria infection in children under the age of fourteen years in Kenya (Masters dissertation)*

P. McCullagh J.A. Nelder FRS, 1983. *Generalized Linear Models monographs on Statistics and Applied probability.* London New York: Chapman and Hall.

Park, I. & Lee, H., 2001. The design effect: do we know all about it.. *In Proceedings of the Annual Meeting of the American Statistical Association,* August.pp. 5-9.

Patterson, H. D. & Thompson, R., 1971. Recovery of inter-block information when block sizes are unequal.. *Biometrika,* 58(3), pp. 545-554.

Peng, C. J. & So, T. H., 2002. Logistic regression analysis and reporting. *A primer. Understanding Statistics: Statistical issues in Pyschology, Education, and the Social Science,* 1(1), pp. 31-70.

Peng, C. Y., Lee, K. L. & Ingersoll, G. M., 2002. An introduction to logistic regression analysis and reporting. *The Journal of Educational Research,* 96(1), pp. 3-14.

Pfeffermann, D., 1993. The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique,* pp. 317-337.

Power, D. & Xie, Y., 1999. *Statistical Methods for Categorical Data Analysis.* New York: Emerald Group Publishing.

Quinn, G. P. & Keough, M. J., 2002. *Experimental design and data analysis for biologists.* Uk: Chambridge in University Press.

Rafferty, A., n.d. [Online]
Available at: https://www.ukdataservice.ac.uk/media/440347/rafferty.pdf

Rao, J. N. & Scott, A. J., 1981. The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association,* pp. 221-230.

Rao, J. N. & Wu, C. F., 1998. Resampling inference with complex survey data. *Journal of the american statistics association,* 83(401), pp. 231-241.

Raudenbush , S. W., Yang, M. L. & Yosef, M., 2000. Maximum likelihood for Generalized Linear Models with nested random effect via higher-order, multivariate Laplace approximation. *Journal of Computation and Graphical Statistics,* 9(1), pp. 141-157.

Richard Coker, Martin Mckee, Rifat Atun, Boika Dimitrava, Ekaterina Dodonova, Sergei Kuznetsow, Francis Drobniewski, 2005. *risk factors for pulmonary tuberculosis.* Russia: BMJ.

Ruppert, D., Wand, M. P. & Carroll, R. J., 2003. *Semiparametric regression.* 1st ed. s.l.:Cambridge University Press.

Rust, K., 1985. Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics,* 1(4), p. 381.

SAS Institute Inc., 2015. *SAS/STAT® 14.1 User's guide.,* Cary, NC, USA: SAS Institute.

SAS Institute Inc, 2020. *SAS/STAT 15.2® User's guild The Glimmix Procedure.* [Online]
Available at: https://documentation.sas.com/api/docsets/statug/15.2/content/glimmix.pdf?locale=en#nameddest=statug_glimmix_overview

Schabenberger, O. & Gotway, C. A., 2017. *Statistical methods for spatial data analysis.* s.l.:CRC press.

Schackman, G., 2001. Sample size and design effect. *Albany Chapter of the American Statisstical Association.*.

Schwarz, G., 1978. Estimating the dimenision of a model. In: *The annals od statistics*. s.l.: Institute of Mathematical Statistics, pp. 461-464.

Searle, S. R., Casella, G. & McCulloch, C. E., 2006. *Variance Compponents.* New Jersey: John Wiley and Sons.

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

Shalizi C., n.d. *Carnegie Mellon University Statistics and Data Science.* [Online] Available at: https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf

Sheskin, D. J., 2000. *Handbook of parameteic and nonparametric statistical procedures.* 2nd ed. Boca Raton, Florida: Chapman Hall/CRC.

Shetty, N., Shemko, M., Vaz, M. & D'souza, G., 2006. An epidemiological evalution of risk factors for tuberculosis in South India: a matched case control study. *the international Journal of Tuberculosis and Lung Disease,* 1(10), pp. 80-86.

Siller, B. A. & Tompkins, L., 2006. The big four: Analyzing complex sample survey data using SAS, SPSS, STATA, and SUDAAN. *In proceedings of the thirty-first annual SAS® Users Group international conference,* pp. 26-29.

Silverman, B. W., 1985. Some aspects of the spline smoothing approach to non-parametric regression curve fitting.. *Journal of the Royal Statistical Society. Series B (Methodological),* 47(1), pp. 1-52.

Šimundić, A. M., 2009. Measures of Diagnostic Accuracy: Basic Definitions. *EJIFCC,* 19(4), pp. 203-211.

Skinner, C., Holt, D. & Smith, T., 1989. *Analysis of Complex Survey.* New York: John Wiley and sons.

StatsSA, 2013. *STATSA.* [Online] Available at: http://www.statssa.gov.za/?p=1023 [Accessed 10 July 2013].

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

Swets, J. A., 1988. Measuring the accuracy of diagnostic systems.. *Science,* 240(4857), pp. 1285-1293.

Tabachnick, B. G. & Fidell, L. S., 2007. *Using multivariate statistics.* Pearson Education ed. s.l.:Allyn & Bacon.

Tekkel, M., Rahu, M., Loit, H. M. & Baburin, A., 2002. Risk factors for pulmonary tuberculosis in Estonia. *The International Journal of Tuberculosis and Lung Disease,* 10(9), pp. 887-894.

Tuerlinckx, F., Rijimen, F., Verbeke, G. & Boeck, P. D., 2006. Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology,* 59(2), pp. 225-255.

Verbeke, G. & Molenberghs, G., 2009. *Linear Mixed models for longitudinal data.* USA: Springer series in Statistics .

Verbyla, A. P., Cullis, B. R., Kenward, M. G. & Welham, S. J., 1999. The analysis of designed experiments and longitudinal data by using smoothing splines.. *Journal of the Royal Statistical society: Series C (Applied Statistics),* 48(3), pp. 269-311.

Vittinghoff, E., Glidden, D. V., Shiboski, S. C. & McCulloch, C. E., 2011. *Regression methods in biostatistics: linear, survival, and reepeated measure models.* s.l.:Springer Science & business Media.

Vittinghoff, n.d.

Waagepetersen, R., 2007. *Computation of the lilkelihood function for GLMMs.* [Online]

Available at: https://people.math.aau.dk/~rw/Undervisning/Topics/Handouts/6.hand.pdf

Wang, Y., 1998. Mixed effect smoothing spline analysis of variance. *Journal of royal statistics society,* 60(1), pp. 159-174.

WHO, 2018. Tuberculosis in women. *world health organization,* p. 1.

WHO, 2020. *Global tuberculosis report 2020: executive summary.,* s.l.: World Health Organization.

WHO, 2020. *World health organization.* [Online] Available at: https://www.who.int/news-room/fact-sheets/detail/tuberculosis [Accessed 14 October 2020].

Wilber, S. T. & Fu, R., 2010. Risk ratios and odds ratios for common events in cross-sectional and cohort studies. *Academic Emergency Medicine,* 17(6), pp. 649-651.

Wolter, K. M., 1985. *Introduction to variance estimation.* New York: Springer-Verlag.

Woodruff, R. S., 1971. A simple method for approximating the variance of a complicated estimate. *Journal American Statistics Association,* 66(334), pp. 411-414.

Wood, S. N., 2017. *Generalized Additive models: An Introduction with R.* s.l.:Tylor and Fransis Group.

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

Appendix A

Logistic regression SAS code

```
proc import out=data2
datafile="E:\Bsc stream M\mz1.xlsx" dbms=excel
Replace;
run;
ods graphics on;
proc logistic data=data2 plots=all;
class chronicdisease(ref="0") mv101(ref="9") mv106(ref="2") mv131(ref='1')
 mv190(ref="3") mv501(ref="0") sm901(ref="3") sm902(ref="1")
mv463aa(ref="1")/ param=glm;
 model sm1105(event="1")=chronicdisease mv012 mv101 mv106 mv131 mv167
 mv190 mv501 sm901 sm902  mv463aa mv136
 mv012*chronicdisease  mv136*mv463aa
mv012*mv167
 / LINK=LOGIT EXPB CL selection=none scale=none RSQUARE LACKFIT
AGGREGATE=(chronicdisease mv012 mv101 mv106 mv131 mv167
 mv190 mv501 sm901 sm902 mv463aa mv136);
output out=work.outdata p= pred;
RUN ; ods graphics
off;
```

Survery logistic regression SAS code

```
/*final model*/
proc import out=data2
datafile="E:\Bsc stream M\mz1.xlsx" dbms=excel
```

```
Replace;
run;
ods graphics on ;
proc surveylogistic data=data2 ;
stratum mv022/list;
cluster mv021;
weight wgt;
class chronicdisease(ref="0")  mv101(ref="9") mv106(ref="2") mv131(ref='1')
 mv190(ref="3") mv501(ref="0") sm901(ref="3") sm902(ref="1")
mv463aa(ref="1")/ param=glm;
 model sm1105(event="1")=chronicdisease mv012 mv101 mv106 mv131 mv167
 mv190 mv501 sm901 sm902  mv463aa mv136
 mv012*chronicdisease  mv136*mv463aa
mv012*mv167
/ LINK=LOGIT clodds expb stb RSQUARE ; RUN ; ods graphics off;
/*End!!*/
```

Generalized linear mixed model SAS code

```
proc import out=data
datafile="E:\Bsc stream M\mz1.xlsx" dbms=excel
Replace;
run;
ods graphics on;
proc glimmix data=WORK.data plots=pearsonpanel;
class chronicdisease(ref="0") mv001 mv101(ref="9") mv106(ref="2")
  mv131(ref='1')
 mv190(ref="3") mv501(ref="0") sm901(ref="3") sm902(ref="1")
```

```
mv463aa(ref="1");
 model sm1105(Event='1') =chronicdisease mv012 mv101 mv106 mv131 mv167
 mv190 mv501 sm901 sm902  mv463aa mv136
 mv012*chronicdisease  mv136*mv463aa
mv012*mv167 /
dist=binary solution link=logit;
random _residual_;
run;
```

Generalized additive mode R code

```
mz5<- read.csv("E:/Bsc stream M/mz5.csv")
library(mgcv)
ga3=gam(SM1105~s(MV136)+s(MV012)+factor(CHRONICDISEASE)+
factor(MV1011)+factor(MV1062)+factor(MV1312)+
factor(MV1902)+factor(MV5012)+factor(SM9012)+factor(SM9021)
+factor(MV463AA2)+s(MV167)+s(MV012,MV167),family=binomial(link=logit),
data=mz5)
summary(ga3)
plot(ga3,col="black",lwd=3, xlab = "Number of household members", ylab ="s(Number of
household members,5,03)")
```

Generalized additive mixed mode R code

```
mz5<- read.csv("E:/Bsc stream M/mz5.csv")
library(mgcv)
ga7=gamm(mz5$SM1105~s(MV136)+s(MV012)+factor(CHRONICDISEASE)+
s(mv021,bs=”re”)+factor(MV1011)+factor(MV1062)+factor(MV1312)+factor(MV1902)
+factor(MV5012)+ factor(SM9012)+factor(SM9021) +factor(MV463AA2)
+s(MV167)+s(MV012,MV167), niterPQL=1000, family=binomial(link=logit), data=mz5)
```

```
summary(ga7$gam)
gam.check(ga7$gam,pch=19, cex=.3)
plot(ga7$gam,col="black",lwd=3, xlab = "Number of household members", ylab
="s(Number of household members,1.85)" )
```