IDENTIFYING FACTORS ASSOCIATED WITH SMOKING IN GAUTENG IN THE PRESENCE OF MISSING DATA

By

Siyanda Mabungane

208525160

A dissertation submitted in partial fulfilment of the requirements for the degree of

Master of Science

School of Mathematics, Statistics and Computer Science

Department of Statistics and Biometry

Supervisor: Dr. S. Ramroop

2014

DECLARATION

This research has not been previously accepted for any degree and is not being currently considered for any other degree at any other university. I declare that this Dissertation contains my own work except where specifically acknowledged.

This Dissertation is prepared in partial fulfilment of the requirement of the Master of Science at the School of Mathematics, Statistics and Computer Science, Department of Statistics and Biometry, University of KwaZulu-Natal, Pietermaritzburg South Africa.

Supervisor : Dr. S. Ramroop

Signature :

Date :

Student : Siyanda Mabungane

Signature :

Date :

Dedication

This thesis is dedicated to my brothers, Odwa, Lunga and Bonisile; and to my fiancee Snethemba Makhathini. For their endless love, support and encouragement

Acknowledgements

Dr. Shaun Ramroop has been the ideal thesis supervisor. His sage advice, insightful criticisms, and patient encouragement aided the writing of this thesis in innumerable ways. I would also like to thank him very much for believing in me and his support of this project was greatly needed and deeply appreciated.

Abstract

Smoking still remains one of the leading preventable causes of death in South Africa. It increases the chances of lung diseases such as emphysema, chronic bronchitis and many other diseases. The current research aims to model the smoking survey data which was part of the October 1996 omnibus smoking survey in Gauteng (South Africa). The surveyed variables were race, sex, marital status, socio-economic status, smoking status, age and education level. Generalized Linear Models (GLMs) and Generalized Linear Mixed Models (GLMMs) were used to model this data. Multiple Correspondence Analysis (MCA) was used to check for the relationships and correlation among the variables. Furthermore, the problem of missing data was addressed using the classical methods such as Last Observation Carried Forward (LOCF) as well as more modern advanced methods viz. Inverse Probability Weighting (IPW) and Multiple Imputation (MI).

The percentage of smokers was found to be lower than that of non-smokers amongst all the surveyed variables. Race, sex, age and socio-economic status were found to be significant when fitted with both GLMs and GLMMs. It was found that race and socio-economic status were closely correlated, education was closely correlated with race, education was closely correlated with socio-economic status, and age was closely correlated with marital status. MI and IPW estimators were found to be more consistent than the LOCF estimators. In spite of the effort by several health policy makers of trying to alert people about the dangers of smoking, there appears to be a lack of awareness that smoking causes tuberculosis (TB), lung cancer, stroke, throat and mouth cancer, as well as various other lung and heart diseases.

Contents

Chapter 1

1 Introduction and Literature Review

1.1 Introduction

Smoking is a major health problem worldwide; smoking also pollutes the environment. According to Morrison (2011) the use of tobacco in South Africa has seen a decline in the past decade. Smoking tobacco or using tobacco is known scientifically to be habit-forming. Tobacco use is the major cause of many diseases and kills a vast of people nationwide (Morrison, 2011). Most people start smoking when they are teenagers and some as a result of peer pressure by taking a few puffs which ultimately becomes habit forming. Smoking has serious health effects for both the smoker and people around them. Smokers below the age of 50 may increase their risk of a fatal heart attack as compared to nonsmokers (Edwards, 2004). Smokers also expose themselves to an increased risk of throat and lung cancer, osteoporosis, infertility, and many more health issues (Edwards, 2004). A study conducted by Groenewald et al. (2007) found that smoking caused between 41,632 and 46,656 deaths in South Africa. Research shows that smoking is the number one cause of many diseases in South Africa (Saloojee, 2006).

A report by Saloojee (2006) showed that about 4.83 million premature deaths in the world were associated with cigarette smoking, 2.41 million in lower income countries and 2.43 million in developed countries (Saloojee, 2006). According to Sitas et al. (2004), in 1998 an estimated 8% of adult deaths (21 500) was due to tobacco use. South Africa is expecting a growth in the proportion of deaths due to the use of tobacco. Tuberculosis, chronic obstructive

pulmonary disease (COPD), lung cancer and ischaemic heart disease (IHD) all as a consequence of tobacco smoking are the leading causes of mortality in South Africa. Sitas et al. (2004) reported that out of 100 people who die from tobacco related disease in South Africa 19 die of TB, 28 die of COPD, 10 die of cancer, 9 die of stroke or vascular disease and 9 of other conditions. Over the past 20 years the use of tobacco among the youth has seen a significant decline in rich societies (Mashita et al., 2011). The use of tobacco products nationwide is the major cause of chronic diseases morbidity and mortality. Most smokers start their smoking habit from childhood and adolescent stages (Mashita et al., 2011). Tobacco marketing and advertising has an influence on teenage smoking. Nowadays most teenagers are exposed to tobacco advertising and smoking in movies, magazines and other media reports. This exposure increases positive attitudes about smoking and makes them see a reason to smoke (Wellman, et al., 2006; Sargent et al., 2003). A study among the Ellisras (South Africa) rural children showed that the prevalence of tobacco use increases with increasing age among the boys; the usage of tobacco products was high among the older boys; and the results also revealed that girls did not smoke cigarette but only used homemade tobacco products such as pipe and snuff (Mashita et al., 2011). Parents, grandparents and television played a significant role in influencing smoking behaviour among the Ellisras rural children. The prevalence of snuff usage among the girls of Ellisras rural ranged from 0.7% to 4.1% for girls aged 11 to 18 years. For boys, the prevalence of tobacco usage ranged from 4.9% to 17% (Mashita et al., 2011).

The use of tobacco during pregnancy increases the risks of perinatal mortality and morbidity such as miscarriage, premature birth and low birth-weight (Hammoud et al, 2005). Smoking during pregnancy is associated with a

higher risk of respiratory infections, such as asthma and bronchitis in the infant (Jaakola and Gissler, 2004). Prabhu et al. (2010) found that mothers who smoke beyond their first trimester are likely to deliver smaller babies who are then at greater risk of having adverse respiratory outcomes in childhood as compared to expectant mothers who do not smoke.

The effects of smoking in pregnancy is also evident later in life, with low birth-weight being associated with coronary heart disease, type 2 diabetes, and being overweight in adulthood (Doherty et al, 2009). Parents who smoke place their children at a risk of having lung diseases, even the unborn babies are affected by the tobacco smoke. Women who quit smoking successfully during their pregnancy are more likely to have stable relationships and have a greater chance of getting married (Graham et al, 2006). Haslam and Lawrence (2004) reported that women who continue to smoke during pregnancy and after pregnancy are more likely to be poor, unemployed, have low education, live without a partner and have low social support (Haslam and Lawrence, 2004).

In South Africa the British American Tobacco makes up 85.7% share of the legal tobacco market. The illegal tobacco market has caused a vast decline in the legal market. From 2008 the legal sales of tobacco had dropped from a 25 billion to 21 billion in 2010 but the opposite happened with the illegal sales as they had gone up sharply from 3 billion to 6.3 billion in the period. Tobacco marketing and advertising affects and has an influence on teenage smoking. The other problem with tobacco consumption is not only that it causes great numbers of morbidity and mortality around the world, but it also creates high costs for health care systems. The treatment for tobacco-related diseases is costing the government a lot of money, and again is costing

the individuals and families who are dealing with the health consequences of being a tobacco consumer a lot of money too (Narula, 2011).
The health effects of families of tobacco consumers and consumers themselves can be intensified by poverty and poor shelter. Most often rich families have a lesser number of people consuming tobacco as compared to poor families. Poorer families are not only vulnerable to the effects of tobacco use but to the long term health risks associated with smoking such as Tuberculosis and other diseases (Esson and Leeder, 2004).

The South African tax on cigarettes has been very low. In 1994, the South African government announced that it was going to raise the tax on cigarettes to 50% of the retail price over the course of a few years (Van Walbeek, 2002). In order to be effective, tobacco tax increases should not create incentives for people to shift their tobacco consumption from one form to another. The tax increases should thus be similar for the various tobacco products (Van Walbeek, 2002). In 2005 the results that were carried out by the Medical Research Council (MRC) revealed that in the past years one out of ten individuals who died in South Africa died because they were smoking (Steyn et al., 2006). The results revealed that at least 8% of South African adults died because of tobacco use. From these results they found that lung cancer claimed more lives than the other diseases, it claimed 58% of lives; chronic bronchitis claimed 37%, with heart disease and tuberculosis claiming about 20% of lives (Steyn et al., 2006). Studies have shown that persons whose partners smoke have a 20% - 30% greater risk of having lung cancer than those whose partners do not smoke.

1.2 Literature Review

South Africa is one developing country classified as upper-middle income that has a high rate of smoking (22.9% overall) (Saloojee, 2006). In 2009, the South African population comprised of 28% of young adults, aged 16 -24 years and this was the highest proportion of adults in the country (Narula, 2009). In 1993 the prevalence of smokers in this age group was 23.7% and it was 17% in 2003 (Walbeek, 2005). However the prevalence of smokers decreased among the 16-19 years age group in 2001 compared to 1993. Over the past decade, the prevalence rates for adult daily cigarette smoking have decreased. According to the South African Advertising and Research Foundation surveys, the daily smoking rates of young adults and adults from the age of 15 years and older fell by a fifth decreasing from 30.2% in 1995 to 24.1% in 2004 (Van Walbeek, 2002). The South African Advertising and Research Foundation surveys found that an estimated 2.5 million smokers stopped smoking during the period 1995-2004 (Walbeek, 2002). Data from national surveys confirmed that about a fifth to a quarter of adults smoked cigarettes, i.e South African Social Attitude Survey found that 21.4% of adults smoked in 2003; this include a 35.8% of men and 8.1% of women (Ayo-Yusuf, 2004.), while the earlier South African Demographic and Health Survey reported a prevalence rate of 24.6% in 1998 (Steyn et al., 2006).

Laws and Legislation of Smoking in South Africa

Prior to the 1990s, the ruling government put minimal efforts into reducing the impact of smoking tobacco in South Africa. Smoking legislation was initially governed by the Tobacco Products Control Act 83 of 1993. However, many changes have been introduced through the Tobacco Prod-

ucts Control Amendment Act 23 of 2007 and the Tobacco Products Control Amendment Act 63 of 2008. After 1994 when the country was ruled by the new government, the department of health made tobacco control a priority. The anti-tobacco non-governmental organizations such as the National Council Against Smoking, the Heart Foundation of South Africa and the South African Cancer Association supported the decision. In 2000, South Africa became one of the first countries to ban smoking in the public places when its Tobacco Products Control Amendment Act was introduced. The Act prohibited smoking in restaurants, pubs, shopping centres and offices where there were no separate enclosed smoking rooms. Section 2 (1) (a) Act 23 of 2007 states that no person may smoke any tobacco product in any indoor, enclosed or partially enclosed areas. In 2009, the smoking laws were tightened even further when government even banned smoking in partially enclosed public places such as balconies, verandas, walkways and parking areas. The law also prohibited children under the age of 18 from buying tobacco products and entering smoking areas.

The aim of this research is to identify factors associated with smoking in the presence of missing data. The value that the study brings is recommendation on methods and techniques on how to handle missing data as well as recommendations to smoking policies in South Africa for possible revision based on the findings of this thesis. Furthermore we wish to:

1. Model the data using modern statistical methods.

2. Apply modern techniques for handling missing data and compare the results of these techniques to assess their strengths and weaknesses.

Chapter 2

2 Data description and Methodology

2.1 Data description

South African adults' smoking status, perceptions of the health effects of nicotine and cigarettes, attitudes toward smoking control and awareness of Government health warnings on the harmful effects of smoking were investigated. This was done by means of a series of interviewer administered questionnaires conducted by fieldworkers of the Human Sciences Research Council through the Omnibus surveys in February 1995, February 1996 and October 1996. The first survey preceded the implementation of the Tobacco Control Act whilst the second and third surveys followed. The data which were part of the October 1996 omnibus smoking survey in South Africa is used in this study. We use the data of the survey that was carried out in Gauteng only.

There are 343 values of the dependent variable (which is the smoking status variable in our case; this response variable will be described later on). The surveyed variables were race, sex, marital status, socio-economic status, smoking status, age and education level. The sampling methodology for the data is as follows: The study population consisted of South African residents of 18 years and older, in all 9 provinces. The population was stratified by province and type of area. The sample allocation to the resulting strata was done proportionally to the 1991 census figures except that disproportion was introduced to give 120 respondents per province and the minimum num-

ber of Indians in the sample was fixed at 120. Multistage cluster sampling with probability proportional to size was used to draw respondents, with the adjusted 1991 population census figures as a measure of size. Census enumeration areas and similar areas were used as the clusters.

There were 40 clusters in total in Gauteng. The clusters for which we have data are considered to be a random selection of clusters from all the clusters in Gauteng. A random selection of respondents was then drawn from the clusters, for example we have 4 respondents from cluster 1. All clusters were drawn with probability proportional to size, whilst households were drawn from the final clusters with equal probability. One respondent aged 18 years or older was selected from each household by applying a grid. For each selected respondent a sampling weight was calculated, using the stratification variables of province and type of area and by post-stratification for age, gender, education and race. As stated previously, the main aim of the survey was to investigate the prevalence of smoking and the factors associated with it through statistical modeling techniques in the presence of missing data.

2.2 Exploratory Data Analysis

The exploratory data analysis (EDA) approach is used to summarize the main characteristics of the data in the form of bar graphs and cross-tabulations without using a statistical model. For this data, analysis is done with a complete data set in SPSS version 21. The surveyed variables analyzed were race, sex, marital status, socio-economic status, smoking status, age and education level.

Cross Tabulation of Race by Smoking Status

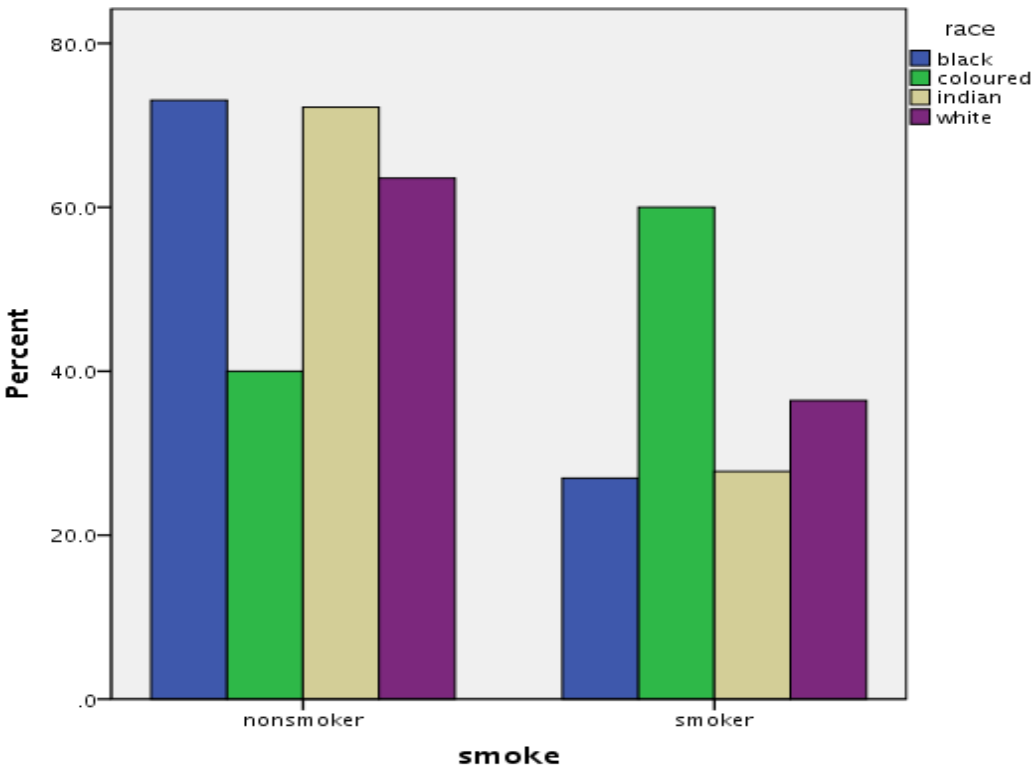


Figure 2.1: A Clustered Bar Graph of Race by Smoking Status

There is a significant association between race and smoking. We find that $\chi^2_3 = 8.217$, p-value is 0.042.

		Race				
Smoking Status		Blacks	Coloured	Indian	White	Total
Non-smoker	count	103	6	13	82	204
	% race	73.0%	40.0%	72.2%	63.6%	67.3%
Smoker	count	38	9	5	47	99
	% race	27.0%	60.0%	27.8%	36.4%	32.7%
Total	count	141	15	18	129	303
	% sex	100.0%	100.0%	100.0%	100.0%	100.0%

Table 2.1: Cross Tabulation of Race by Smoking Status

Figure 2.1 and Table 2.1 show that Blacks, Whites and Indians have a high percentage of non-smokers compared to Coloured. The Coloured race has a higher percentage of smokers than non-smokers. The sampled blacks have 73% non-smokers and 27% smokers; Indian 72.2% of non-smokers and 27.8% of smokers; Whites 63.6% of non-smokers and 36.4% of smokers; and the Coloured race with a high percentage of smokers compared to non-smokers, 40% of non-smokers and 60% of smokers.

Cross Tabulation of Gender and Smoking Status

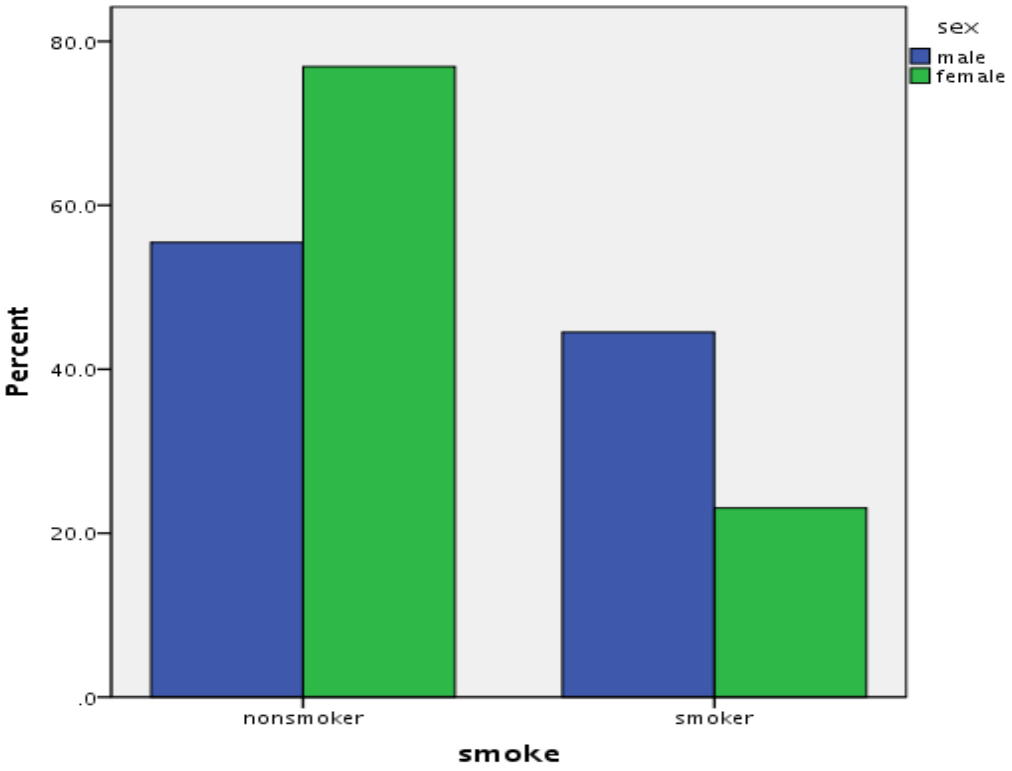


Figure 2.2: A Clustered Bar Graph of Gender by Smoking Status

There is a strong association between gender and smoking status with χ_1^2 = 15.822 and p-value is <0.001. Figure 2.2 shows that females have a high percentage of non-smokers than males.

		Sex		
Smoking Status		Males	Females	Total
Non-smoker	count	76	130	206
	% sex	55.5%	76.9%	67.3%
Smoker	count	61	39	100
	% sex	44.5%	23.1%	32.7%
Total	count	137	169	306
	% sex	100.0%	100.0%	100.0%

Table 2.2: Cross Tabulation of Gender and Smoking status

The table 2.2 shows that the proportion of smokers within males is 44.5% while for females the proportion of smokers is 23.1%. The proportion of non-smokers for males within non-smokers is 55.5% while for females the proportion of non-smokers is 76.9%.

Cross Tabulation of Age and Smoking Status

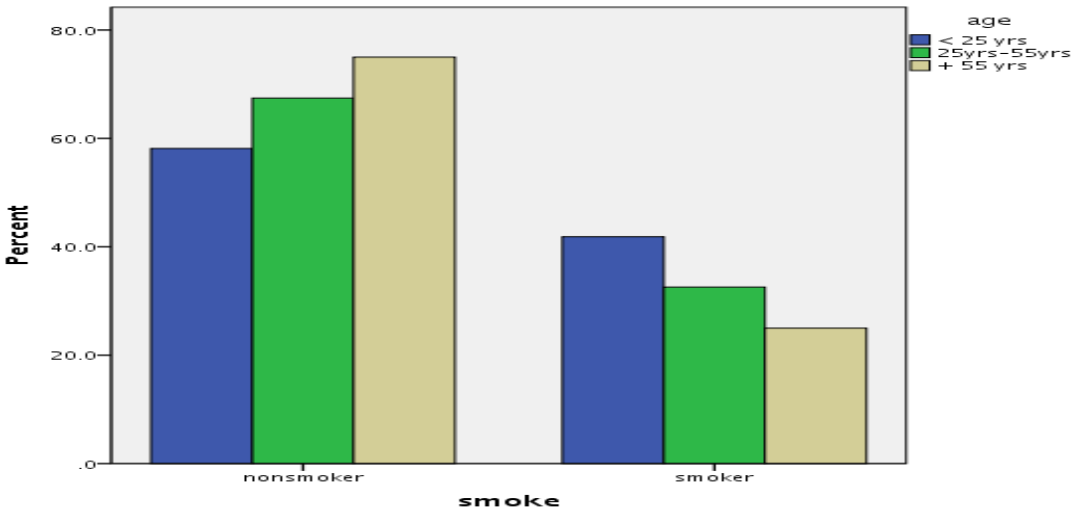


Figure 2.3: A clustered Bar Graph of Age by Smoking Status

There is a significant association between age and smoking. We find the χ_2^2 = 2.936 and a p-value of 0.230.

		Age			
Smoking Status		Less than 25 years	25-55 years	Over 55 years	Total
Non-smoker	count	25	145	36	206
	% age	58.1%	67.4%	75.0%	67.3%
Smoker	count	18	70	12	100
	% age	41.9%	32.6%	25.0%	32.7%
Total	count	43	215	48	306
	% sex	100.0%	100.0%	100.0%	100.0%

Table 2.3: Cross Tabulation of Age and Smoking Status

In Figure 2.3, the percentage of non-smokers is higher than that of smokers in all three age groups. From Table 2.3 we can see that again the percentage of non-smokers is higher than that of smokers in all three age groups. The 25 years age group or younger comprised 58.1% of non-smokers and 41.9% of smokers; those who are 25-55 years comprised 67.4% of non-smokers and 32.6% smokers; those of 55 years and above comprised 67.3% of non-smokers and 32.7% of smokers. The 25 years age group or younger had a higher percentage of smokers than the two other age groups; the 55 years and above had a very low percentage of smokers. Therefore we can conclude that people who are 55 years and above do not smoke frequently.

Cross Tabulation of Socio-economic Status and Smoking Status

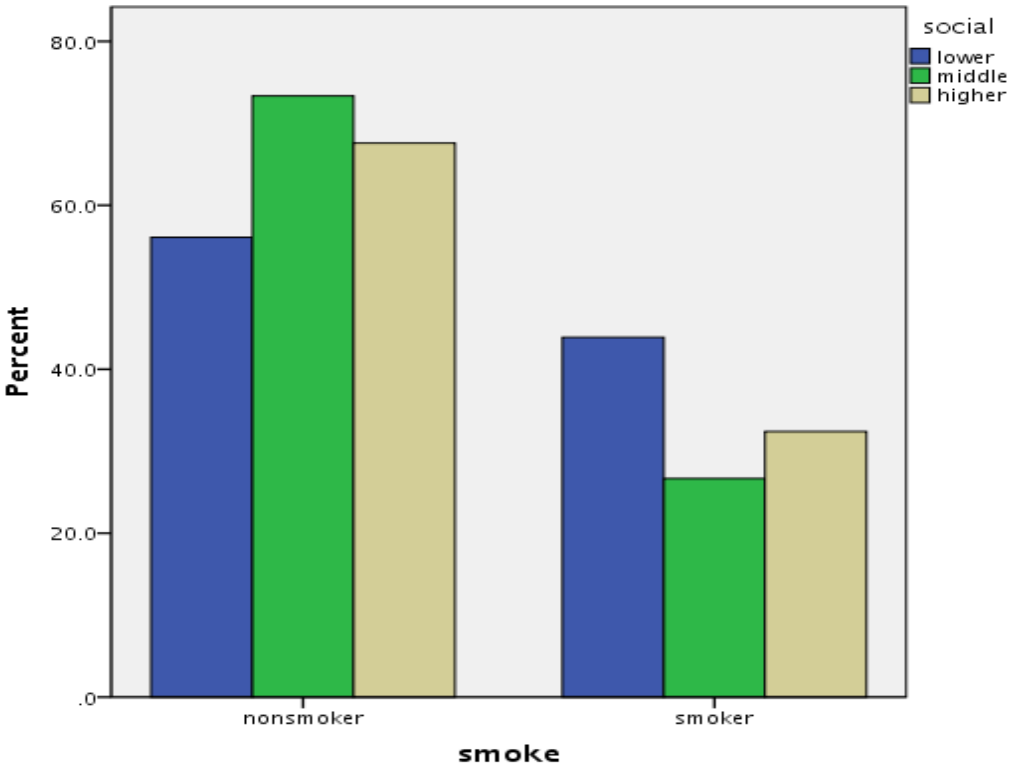


Figure 2.4: A Clustered Bar Graph of Socio-economic Status by Smoking Status

There is no significant association between socio economic status and smoking. We find that $\chi^2_2 = 3.833$ and p-value of 0.147.

Figure 2.4 shows us that the percentage level of non-smokers was higher than that of smokers in all three socio-economic status levels.

	Socio-economic Status				
Smoking Status		Lower	Middle	Higher	Total
Non-smoker	count	23	66	121	210
	% social	56.1%	73.3%	67.6%	67.7%
Smoker	count	18	24	58	100
	% social	43.9%	26.7%	32.4%	32.3%
Total	count	41	90	179	310
	% sex	100.0%	100.0%	100.0%	100.0%

Table 2.4: Cross Tabulation of Socio-economic status and Smoking Status

Table 2.4 shows that the lower level of socio-economic status had 56.1% of non-smokers and 43.9% of smokers; the middle level of socio-economic status had 73.3% of non-smokers and 26.7% of smokers and finally the higher level of socio-economic status had 67.7% of non-smokers and 32.3% of smokers. These results show that people in the lower class smoke more frequently than people in the other classes.

Cross Tabulation of Marital Status and Smoking Status

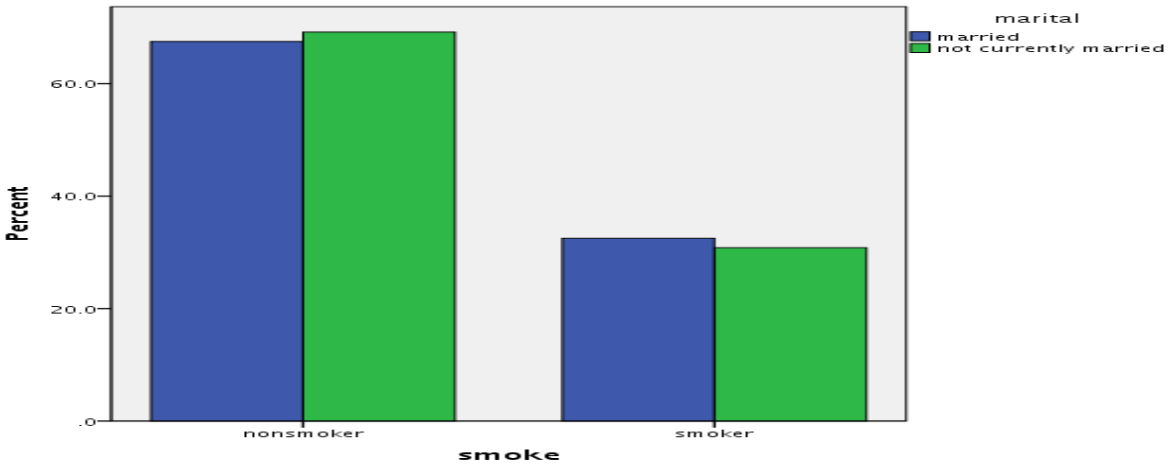


Figure 2.5: A Clustered Bar Graph of Marital Status by Smoking Status

There is no significant association between marital status and smoking. We find that $\chi_1^2 = 0.098$ and p-value is 0.754.

	Marital Status			
Smoking Status		Married	Not Currently Married	Total
Non-smoker	count	137	83	220
	% marital	67.5%	69.2%	68.1%
Smoker	count	66	37	103
	% marital	32.5%	30.8%	31.9%
Total	count	203	120	323
	% sex	100.0%	100.0%	100.0%

Table 2.5: Cross Tabulation of Marital Status and Smoking Status

From Figure 2.5 we can see that both the married and the not currently married group have a higher percentage of non-smokers compared to smokers. The not currently married group had a higher percentage of non-smokers than the married group, not currently married group 69.2% of non-smokers and 30.8% of smokers; married group 67.5% of non-smokers and 32.5% of smokers. From these results we can conclude that married people smoke cigarettes more frequently than those not currently married, however the difference is not statistically significant.

Cross Tabulation of Education and Smoking Status

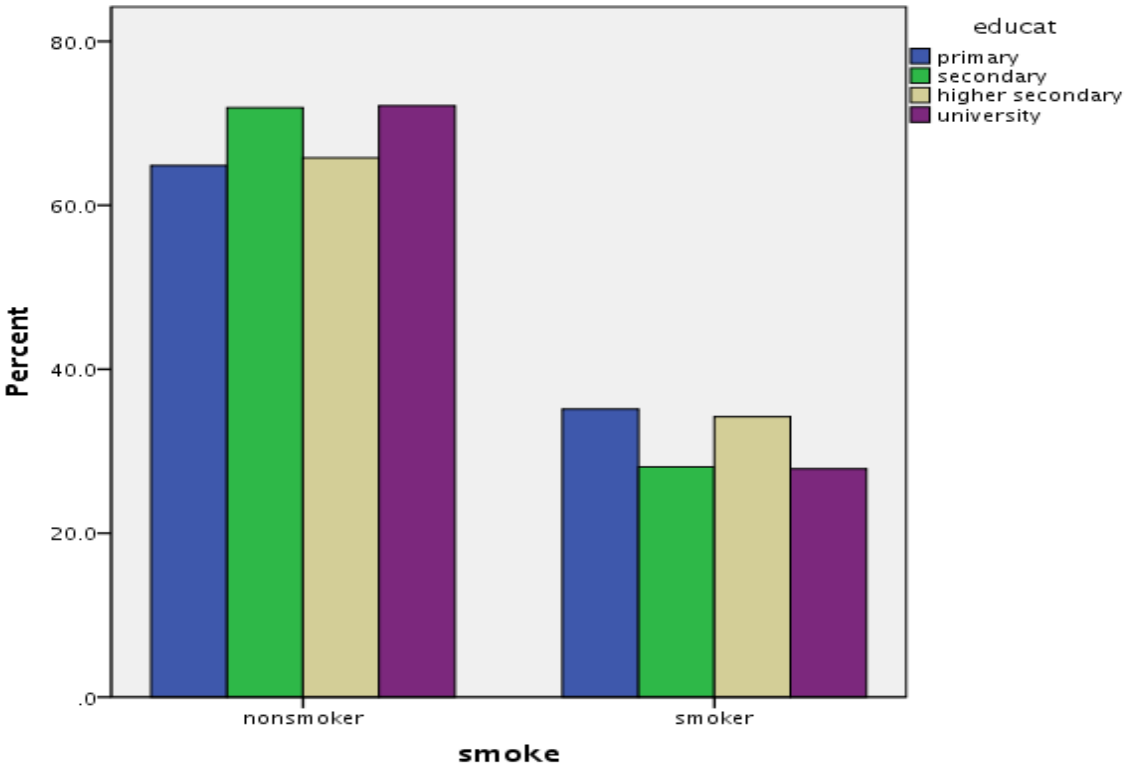


Figure 2.6: A Clustered Bar Graph of Education Status by Smoking Status

There is no significant association between education and smoking. We find that $\chi^2_3 = 3.833$ and p-value of 0.668.

	Educational level					
Smoking Status		Primary	Secondary	Higher Secondary	University	Total
Non-smoker	count	24	64	75	57	220
	% educational level	64.9%	71.9%	65.8%	72.2%	69.0%
Smoker	count	13	25	39	22	99
	% educational level	35.1%	28.1%	34.2%	27.8%	31.0%
Total	count	37	89	114	79	319
	% sex	100.0%	100.0%	100.0%	100.0%	100.0%

Table 2.6: Cross Tabulation of Education and Smoking Status

Figure 2.6 and Table 2.6 both show that non-smokers have a higher percentage than smokers in all the educational levels. The university level has the highest percentage of non-smokers and the primary level has the highest percentage of smokers, university level 72.2% of non-smokers and 27.8% of smokers; secondary level 71.9% of non-smokers and 28.1% of smokers; higher secondary level 65.8% of non-smokers and 34.2% of smokers; primary level 64.9% of non-smokers and 35.1% of smokers. The results in Table 4.6 show a small difference in the percentage of non-smokers between the university level and secondary level; also, the primary level and higher secondary level have a small difference in the percentage of smokers. From these results, we can conclude that the primary level uses cigarettes more often than the other educational levels, however the difference is not statistically significant. The primary and higher secondary group had a higher percentage of smokers when compared to the other two groups. We conclude that individuals who have only primary education smoke more frequently since they had a highest percentage of smokers and this also shows that these individuals find it difficult to read the messages on the boxes of tobacco products.

2.3 Overall discussion of exploratory data analysis

The results show that the percentage of non-smokers is higher than that of smokers in all the surveyed variables. Race compromised 63.7% of non-smokers and 36.3% of smokers and was also found to be significant with respect to smoking status. The association between sex and smoking status is not statistically significant at 5% level of significant. We find that sex comprised of 67.3% of non-smokers and 32.7% of smokers. The association between age and smoking status is not statistically significant at 5% level of significant, with 67.3 % of non-smokers and 32.7% of smokers, marital status has 68.1% of non-smokers and 31.9% of smokers and that smoking status and marital status are significantly different, we find that socio economic status was statistically related to smoking status and socio economic status comprised 67.7% of non-smokers and 33.3% of smokers, educational level had 69.0% of non-smokers and 31.0% of smokers and we find that there was statistically significance association between educational level and smoking status. From these results we can conclude that, even though the percentage of non-smokers is higher than that of smokers in all the surveyed variables. However, tobacco usage is a persistent problem and interventions are required.

2.4 Methodology

The likelihood techniques to handle missing data include generalized linear models (GLMs) and generalized linear mixed models (GLMMs). The Generalized Linear Model (GLM) is a generalization of the general linear model. A generalized linear model can be characterised by the following three components:

- Stochastic component:

  This component specifies the conditional distribution of the response variable, Y_i (for the ith of n independently sampled observations), and the Y_i are usually assumed to have independent normal distributions.

- Systematic component:

  This component is said to be a linear predictor. The covariates X_i combine linearly with the coefficients to form the linear predictor given by:

$$\eta_i = \alpha + \beta_1 + \beta_1 x_{i1} + \beta_2 x_{i2} ... + \beta_k x_{ik} \tag{2.1}$$

- Link between the random and systematic components:

  The link between the random and the systematic component is characterised by the link function g(.) which transforms the expectation of the response variable, $\mu_i = E(Y_i)$, to the linear predictor η_i.

$$g(\mu_i) = \eta_i \quad = \quad \alpha + \beta_1 + \beta_1 x_{i1} + \beta_2 x_{i2} ... + \beta_k x_{ik} \tag{2.2}$$

More details on the GLM is given in the subsequent chapters. The following methods and techniques will be used in this study to handle missing data:

2.5 Missing data mechanisms

Every researcher dealing with missing data would want to understand why the data is missing. There are several reasons for the data to be missing. Data could be missing because of bad weather in experimental studies, it could be missing because of patients dropping out in clinical studies or else it could be missing because some participants never wanted to answer certain questions or did not see some questions; or maybe the questionnaire was too long. There are three types of missing data mechanisms. We look at these mechanisms now.

- Missing Completely at Random

  If Y is the data vector composed of two parts, those completely observed and those potentially missing, meaning that Y=($Y_{observed}, Y_{missing}$) where $Y_{observed}$ is the observed data and $Y_{missing}$ is the missing data, then the data on Y are said to be Missing Completely at Random (MCAR) if the probability of missing data on Y is not related to the data of Y itself or to the values of any other variables in the data set (Allison, 2002). Missing data is said to be missing completely at random (MCAR) if the missingness does not depend on the observed values and again does not depend on missing values of all variables (Becker et al., 2007). This means that the missing value is not related to the values of any variables, whether missing or observed. An example of MCAR is one where survey respondents drop out of the survey because they are relocating for career reasons and again the data could be MCAR in a survey where respondents are followed over time if the survey respondents did not come for a survey because of the weather conditions.

- Missing at Random

The data is said to be missing at random (MAR) if the missingness depends only on the observed values of variables and does not depend on any missing values (Allison, 2001). An example of MAR is one where people who are depressed might be less likely to report their income, and thus reported income will be related to depression (Howell, 2007). Another example of MAR is when a psychologist is studying quality of life in a group of cancer patients and finds that elderly patients and patients with less education have a higher propensity to refuse to answer the quality of life questionnaire (Craig, 2010).

- Not Missing at Random

  An observation is said to be Not Missing at Random (NMAR) when the probability of an observation missing depends on unobserved measurements. A good example of NMAR is that of dropouts especially in clinical trials. For example, if we are studying mental health and people who have been diagnosed as depressed are less likely than others to report their mental status, the data are not missing at random. The other example is one where people with low incomes are less likely to report their income on a data collection form (Howell, 2007). Craig (2010) looks at an example where a number of cancer patients in a cancer trial become so ill (meaning that their quality of life became so poor) that they could no longer participate in the study.

2.6 Methods

(i) Complete case analysis (CC)

Complete case analysis is a common approach used in handling missing data. The problem with this approach is that it deletes all units with incomplete data from the analysis; for this reason this approach is regarded as being inefficient. This approach only uses subjects who have all variables observed, meaning that it excludes individuals with missing data (Seaman and White, 2013). Seaman and White (2013) report that estimates obtained from complete case analysis may be biased if the excluded individuals are systematically different from those individuals included. MCAR is the only missing data mechanism that allows for the use of complete case analysis (Becker et al., 2007). Much of the available software (e.g SAS and SPSS) uses complete case analysis as a default function.

(ii) Last Observation Carried Forward

The Last Observation Carried Forward (LOCF) is said to be the most common used imputation procedure or technique in handling missing data. It is a commonly used way of imputing data with dropouts. LOCF takes the last available response and substitutes the value into all subsequent missing values (Myers, 2000).

(iii) Multiple Imputation

Multiple imputations use the existing values from other variables to predict any missing values of the other variables. The predicted values are then substituted for the missing values; this will result in a new full data set known as the imputed data. This process is repeated a number of times resulting in multiple imputed data sets (Rubin, 1987).

(iv) Inverse Probability Weighting

Inverse Probability Weighting (IPW) estimation is a technique used with observed data and is also used to adjust for unequal sampling fractions (Seaman et al., 2012). The technique is also used to account for missing data when subjects with missing data cannot be included in the primary analysis (Hernan et al., 2006). IPW is one of several methods that can be used to reduce the bias from complete case analysis. The complete cases are thus weighted by the inverse of their probability of being a complete case (Seaman and White, 2013). The approach can again be used to correct for unequal sampling fractions. IPW is commonly applied to two scenarios where the outcomes are observed for only a portion of the sample. The first scenario is if missing data is due to survey nonresponse. The second scenario involves the estimation of the causal effect of a treatment or treatments in which only one of the possible potential outcomes for each study unit is observed (McCaffrey et al., 2011). For example, in studies of educational interventions, student achievement as measured by standardized tests is almost always used as the key covariate for removing hidden biases but standardized test scores often have substantial measurement errors for many students (McCaffrey et al.,2013).

Chapter 3

3 Generalized linear models

3.1 Introduction

Generalized linear models (GLM) extend the idea of conventional regression analysis. These models are suitable when the response variable is non-normally distributed along with exploratory variables that are categorical and continuous. Also, these models are a large class of statistical models used for relating responses to linear combination of predictor variables, they can also include interaction term (Dobson and Barnett, 2008). The parameters of the model provide a way to assess which exploratory variables are related to the response variables. The regression parameters are estimated using the maximum likelihood method (McCullagh and Nelder, 1989). GLMs include the Normal distribution as a special case where one uses the identity link. If the response variable is assumed to be nonlinear, then the link function is used to model the response variable.

3.1.1 The model

Gill (2001) proposes a familiar linear regression model in a matrix notation as follows

$$Y_i = X_i\beta + e_i \tag{3.1}$$

where i takes on i=1,2,...n

Y_i is the dependent variable.

β is a vector of unknown parameters.
e_i are said to be zero mean stochastic disturbances and they are assumed to be normally distributed with zero mean and have a constant variance σ^2 .
X_i is a vector of k independent variables.

For a linear model, Gill (2001) suggests that we need to make some relatively strict assumptions. The assumptions we then make relate to the Gauss-Markov theorem which is given in the following regression model:

$$Y_i = \alpha + \beta X_i + e_i \tag{3.2}$$

Assumptions are:

1. The relationship between each exploratory variable and the outcome variable is approximately linear in structure.

2. The residuals are independent with mean zero and constant variance

3. There is no correlation between any regressor variables and exploratory variables(Gill, 2001)

The general linear model assumes that a random variable the Y_i has a normal distribution with mean μ_i and variance σ^2 (Rodrigues, 2001).

$$Y_i \sim N(\mu_i, \sigma^2)$$

The expected value of Y_i is assumed to be a linear function of p-predictors that take on values $x_i' = (x_{1i}, x_{2i}, ..., x_{pi})$ for the i^{th} case, so that

$$\mu_i = x_i\beta$$

where β is a vector of unknown parameters.

The structure of generalized linear models (GLMs) consists of three components which are:

(i) A random component

The $Y_i's$ are usually assumed to be independent normally distributed with mean $E(y_i) = \mu_i$ and constant variance σ^2. i.e

y is iid N(μ_i, σ^2)

(ii) A linear predictor that is a linear function of regressors

The covariates X_i combine with the coefficients to form the linear predictorη_i (Nelder and Beker, 1972).

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_{i2} X_{i2}, ..., \beta_{ip} X_{ip} \tag{3.3}$$

(iii) The link function g(.)

The link function equates a function of the mean response μ_i to the linear predictor $\eta_i = x_i'\beta$. The linear predictor is a function of mean parameter μ_i via the link function according to the equation (4.4) below:

$$g(\mu_i) = \eta_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} +, ... + \beta x_{ip} \tag{3.4}$$

3.1.2 The link function

A link function is defined as the process of linking a transformation of the observed responses to the original data (O'Connell, 2006). Table 3.1 below shows common examples of link function and their inverses in Generalized Linear Models (GLMs)

Table 3.1: Table of link functions for Generalized Linear Models and their link functions (Nelder and Beker, 1972)

Link	$g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
Identity	μ_i	η_i
Log	$log_e(\mu_i)$	e^{η_i}
Inverse	μ_i^{-1}	η_i^{-1}
Inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
Logit	$log_e(\frac{\mu_i}{1-\mu_i})$	$\frac{1}{1+e^{-\eta_i}}$
Log-log	$log_e[-log_e(\mu_i)]$	$exp[-exp(\eta_i)]$
Complementary log-log	$log_e[-log_e(1-\mu_i)]$	$1 - exp[-exp(\eta_i)]$

3.2 The Exponential Family

Exponential family consists of a set of distributions for both continuous and discrete random variables. All Generalized Linear Models (GLMs) are based on this family of distributions. If we take a continuous random variable Y_i from a distribution that is a member of the exponential family, and it depends on a single parameter θ, then the probability density function for Y_i, $f(y_i|\theta)$, can be written as:

$$f(y_i|\theta) = exp\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\} \tag{3.5}$$

where θ_i and ϕ are parameters, $b(\theta_i), c(y_i, \phi), a(\phi))$ are known functions. If Y_i has a distribution in the exponential family then it can be shown that it's mean and variance are given by:

$$E(Y_i) = \mu_i = b'(\theta_i) \tag{3.6}$$

$$Var(Y_i) = \sigma^2 = b''(\theta_i)a(\phi) \tag{3.7}$$

where $b'(\theta_i)$ and $b''(\theta_i)$ are first and second derivatives of $b(\theta_i)$ respectively (Rodríguez, 2001). From the probability density function of the exponential family $f(y/\theta)$, we can show the special cases of Normal, Binomial and Poisson distributions.

3.2.1 Normal distribution

The normal distribution belongs to the exponential family. The density function of the normal distribution is given by:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{\frac{-1}{2\sigma^2}(y_i - \mu_i)^2\right\} \tag{3.8}$$

The exponential family is given by:

$$logf(y_i) = -\frac{1}{2}log(2\pi\sigma^2) - \left\{\frac{1}{\sigma^2}(y_i^2 + \mu_i^2 - 2y_i\mu_i)\right\} \tag{3.9}$$

$$= exp\left\{-\frac{1}{2}log(2\pi\sigma^2) - \left\{\frac{1}{\sigma^2}(y_i^2 + \mu_i^2 - 2y_i\mu_i)\right\}\right\} \tag{3.10}$$

$$= exp\left\{\frac{y_i\mu_i - \frac{\mu_i^2}{2}}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}log(2\pi\sigma^2)\right\} \tag{3.11}$$

where

$\theta_i = \mu_i$

$b(\theta_i) = \frac{1}{2}\theta^2$

$a(\phi) = \sigma^2$

The mean and the variance of the Normal distribution is given by:

$$E(y_i) = \mu \tag{3.12}$$

$$Var(y_i) = \sigma^2 \tag{3.13}$$

The deviance of normal distribution is given by the following equation:

$$D = 2\sum \left\{ y_i(y_i - \hat{\mu}_i) - \frac{1}{2}y_i^2 + \frac{1}{2}\hat{\mu_i^2} \right\} \tag{3.14}$$

$$D = 2\sum \left\{ \frac{1}{2}y_i^2 - y_i\mu_i + \frac{1}{2}\hat{\mu_i^2} \right\} \tag{3.15}$$

$$D = 2\sum \left\{ (y_i - \mu_i)^2 \right\} \tag{3.16}$$

3.2.2 Binomial distribution

We can verify that the binomial distribution $B(n_i, \pi i)$ belongs to the exponential family and πi is the probability of success. The density function of the binomial distribution is given by:

$$f(y_i) = \binom{n}{y_i} \pi^{y_i}(1 - \pi_i)^{n - y_i} \tag{3.17}$$

The exponential form of this distribution is given by:

$$logf(y_i) = exp\left\{log[\binom{n}{y_i}\pi^{y_i}(1-\pi_i)^{n-y_i}]\right\} \tag{3.18}$$

$$= \quad exp\left\{ylog\pi_i + (n-y_i)log(1-\pi_i) + log\binom{n}{y_i}\right\} \tag{3.19}$$

$$= \quad exp\left\{y_i log(\frac{\pi_i}{1-\pi_i}) + nlog(1-\pi_i) + log\binom{n}{y_i}\right\} \tag{3.20}$$

where

$\theta_i = log(\frac{\pi}{1-\pi})$

$b(\theta_i) = -nlog(1-\pi)$

$c(y_i, \phi) = log\binom{n}{y}$

$a(\phi) = 1$

The mean and the variance of Binomial distribution is given by:

$$E(y_i) = b'(\theta_i) = n\pi_i \tag{3.21}$$

$$Var(y_i) = a(\phi)b''(\theta_i) = n\pi_i(1-\pi_i) \tag{3.22}$$

The deviance of binomial distribution is given by the following equation:

$$\begin{aligned} D = 2\sum\left\{y_i log(\frac{y_i}{n}) + (n-y_i)log(\frac{n-y_i}{n})\right\} \\ -2\sum\left\{y_i log(\frac{\hat{\mu}_i}{n}) + (n-y_i)log(\frac{n-\hat{\mu}_i}{n})\right\} \end{aligned} \tag{3.23}$$

If we cancel out all the terms with log (n_i) and collect all the terms with y_i and terms with $(n_i - y_i)$ we get that:

$$D = 2\sum \left\{ y_i log(\frac{y_i}{\hat{\mu}_i}) + (n - y_i) log(\frac{n - y_i}{n}) \right\}] \tag{3.24}$$

If we simplify this equation we get;

$$D = 2\sum o_i log(\frac{o_i}{e_i}) \tag{3.25}$$

where o_i denotes observed outcome and e_i denotes the expected value under the model of interest and the sum is over both "successes" and "failures" for each i.

3.2.3 Poisson distribution

The Poison distribution is a member of exponential family. The canonical link of this distribution is the log link and the distribution has a natural parameter $\theta = log(\mu)$. The density function of the binomial distribution is given by:

$$f(y_i; \mu) = \frac{\mu^y e^{-\mu}}{y!} \tag{3.26}$$

The exponential form of this distribution is given by:

$$log(f(y_i;\mu)) = log[\frac{\mu^{y_i}e^{-\mu}}{y_i!}] \tag{3.27}$$

$$= \quad y_i log(\mu) - \mu - log(y!) \tag{3.28}$$

where

$\theta_i = log(\mu)$

$b(\theta_i) = -\mu$

$c(y_i, \phi) = -log(y!)$

$a(\phi) = 1$

The mean and the variance of the Poisson distribution is given by:

$$E(y_i) = b'(\theta_i) = \mu \tag{3.29}$$

$$Var(y_i) = a(\phi)b''(\theta_i) = \mu \tag{3.30}$$

The deviance of Poisson distribution is given by the following equation:

$$D = 2\sum \{y_i log(y_i) - y_i - log(y_i!) - y_i log(\hat{\mu}_i) + \hat{\mu}_i + log(y_i!)\} \tag{3.31}$$

If we cancel out all the terms with $(y_i!)$ and collect all the terms with y_i and terms with $(n_i - y_i)$ we get;

$$D = 2\sum \left\{ y_i log(\frac{y_i}{\hat{\mu}_i}) - (y_i - \hat{\mu}_i) \right\} \tag{3.32}$$

If we simplify this equation we get that:

$$D = 2\sum o_i log(\frac{o_i}{e_i}) \tag{3.33}$$

where o_i denotes observed output and e_i denotes the expected value under the model of interest. We note that the Poisson distribution and the Binomial distribution both have equivalent form of deviances when the equations are simplified; the second term goes to zero because $\sum y_i = \sum \hat{\mu}_i$.

Table 3.2: Examples of useful GLMs along with their link functions

Model	Distribution	Link function
Normal	Normal distribution	Identity link
Poisson	Poisson distribution	Log link
Binomial	Binomial distribution	Logit link

Inference on GLMs

The Wald, the score and the likelihood ratio inference methods are the common methods used in generalized linear models (Agresti, 1996). These methods are discussed in of logistic regression.

3.3 Logistic Regression- Special Case of GLM

3.3.1 Introduction

Regression methods have become an integral component of any data analysis concerned with describing the relationship between a response variable and one or more exploratory variables. It is often the case that the outcome variable is discrete, taking on two or many possible values (Hosmer and Lemeshow, 2000). Before we interrogate a study of logistic regression it is important to understand that the goal of any analysis such as logistic regression is the same as that of any model-building technique used in statistics i.e. to find the best fitting and most parsimonious model to describe the relationship between an outcome (dependent or response) variable and a set of independent (predictor or exploratory) variables. These independent variables are often called covariates or exploratory variables (Hosmer and Lemeshow, 2000). The most common form of modeling is the usual linear regression model where the outcome variable is assumed to be of continuous form and that it follows the Normal distribution (Hosmer and Lemeshow, 2000).

Logistic regression determines the impact of multiple independent variables presented simultaneously to predict membership of one or other of the two dependent variable categories.

Logistic regression assumes no linear relationship between the dependent and independent variables. It assumes that the dependent variable must be of two categories. With logistic regression the independent variable does not need to be of interval scale and normally distributed and does not need to have of equal variance within each category. A distinguishing feature of a logistic regression model from a linear regression model is that the outcome variable in

logistic regression is binary or dichotomous (Hosmer and Lemeshow, 2000). This difference in logistic regression and linear regression model is reflected both in the choice of a parameter model and in the assumptions. Once this difference is accounted for, the methods employed in an analysis using logistic regression follow the same general principles used in linear regression. Thus, the techniques used in linear regression analysis will motivate our approach to logistic regression (Hosmer and Lemeshow, 2000).
Logistic regression regresses a dichotomous dependent variable on a set of independent variables. There are two key purposes of logistic regression i.e

(i) To predict probability of success for a given set of covariates.

(ii) To supply knowledge of the relationships and strengths among the variables.

As previously stated, logistic regression allows one to predict a discrete outcome, such as group membership, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these. Generally, the dependent or response variable is dichotomous, such as presence/absence or success/failure. Discriminant analysis is also used to predict group membership with only two groups. However, discriminant analysis can only be used with continuous independent variables. Thus, in instances where the independent variables are categorical, or a mix of continuous and categorical, logistic regression is preferred.

3.3.2 Logistic Regression Models

The mathematical concept that underlies logistic regression is the logit, the natural logarithm of an odds ratio. The simplest example of a logit derives from a 2x2 contingency table. The dependent variable in logistic regression is usually dichotomous, that is, the dependent variable can take the value 1 with a probability of success θ, or the value 0 with probability of failure $(1-\theta)$. This type of variable is called a Bernoulli (or binary) variable. The logistic regression makes no assumption about the distribution of the independent variables. They do not have to be normally distributed, linearly related or of equal variance within each group. The relationship between the predictor and response variables is not a linear function in logistic regression, instead, the logistic regression function is used, which is the logit transformation of θ and θ is given as:

$$\pi = \frac{e^{(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_i x_i)}}{1+e^{(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_i x_i)}} \tag{3.34}$$

3.3.3 The Wald Test

A Wald test is used to test the statistical significance of each coefficient (β) in the model. A Wald test is based on a statistic Z given by:

$$Z = \frac{\hat{\beta}}{SE(\hat{\beta})} \tag{3.35}$$

where $\hat{\beta}$ is the parameter estimate on fitting the model and $SE(\hat{\beta})$ is the estimated standard error of the parameter at the value. The Z-value is squared, yielding a Wald statistic which approximately follows the chi-square distribution with df=1. Once the model is fitted, the test value is then compared

to a pre-determined critical value at a given level of significance; if the test statistic is found to be greater than the critical value then we can conclude that the exploratory variable is significant in the model. However, several authors have identified problems with the use of the Wald statistic. Agresti (1996) states that the likelihood-ratio test is more reliable for small sample sizes than the Wald test.

3.3.4 The Likelihood-Ratio Test

The likelihood-ratio test uses the ratio of the maximized value of the likelihood function for the simple model (L_0) over the maximized value of the likelihood function for the full model (L_1). The likelihood-ratio test statistic equals:

$$-2log\left(\frac{L_0}{L_1}\right) = -2[log(L_0) - log(1)] \tag{3.36}$$

$$= \quad -2[l_0 - l_1] \quad \sim \chi^2_k \tag{3.37}$$

where k is the degrees of freedom. This log transformation of the likelihood functions yields a chi-squared statistic. This is the recommended test statistic to use when building a model through backward stepwise elimination.

3.3.5 The Chi-Squared Test of Association

A chi-squared is used to investigate whether distributions of categorical variables differ from one another. When calculating the chi-square test statistic

we need to calculate the expected count two or more groups, population. Once the expected values have been computed, the chi-square test statistic is computed as:

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected} \tag{3.38}$$

where the square of the differences between the observed and expected values in each cell divided by the expected value are added across all the cells in the table. The distribution is chi squared distributed with (r-1)(c-1) degrees of freedom, where r represents the number of rows in the two-way table and c represents the number of columns.

3.3.6 Odds ratio

An odds ratio (OR) is said to be a measure of association between two binary data values from two groups. The odds ratio shows the strength of association between a predictor and the response of interest. The odds ratio compares the odds of an event between two groups. We can use the odds ratios when analyzing case-control studies, cross-sectional and cohort study. When a binary outcome variable is modeled using logistic regression, it is assumed that the logit transformation of the outcome variable has a linear relationship with the predictor variables.

A logit is defined as the logarithm of the odds. If p is the probability of an event, then (1 - p) is the probability of not observing the event, and the odds of the event are ($\frac{p}{1-p}$). Hence, the logit is given by:

$$log(p) = log(\frac{p}{1-p}) \tag{3.39}$$

The logit transform is most frequently used in logistic regression and for fitting linear models to categorical data (log-linear models). The precision of the odds ratio is found by calculating the 95% confidence interval (CI).
For a dichotomous independent random variable, x, coded as either zero or one, x=0 or x=1 the odds ratios OR is given as:

$$OR = \frac{p(1)/(1-\pi(1))}{p(0)/(1-p(0))} \tag{3.40}$$

One can give the values for a logistic regression model when the independent variable is dichotomous as follows (Hosmer and Lemeshow, 2000).

3.4 Application of Logistic Regression to data

3.4.1 Hosmer-Lemeshow Goodness of Fit Test

It is important that we evaluate the model before we draw conclusions and predict future outcomes. The goodness of fit test is an approach used for evaluating the quality of the model. Since our data is binary, the Hosmer-Lemeshow goodness of fit test for logistic regression is used to assess our model. This approach will help in identifying whether our model is correctly specified. For the model to pass the test, the p-value produced must be high (p-value above 0.05); if the p-value is low (p-value below 0.05) then the model fails the test and it is rejected. We performed the Hosmer-Lemeshow goodness of fit for our data and obtained the following results.

Table 3.3: Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	Pr >Chi-Square
2.7564	0.9066

The model shows no evidence of lack of fit based on the Hosmer and Lameshow test. Since the p-value of 0.9066 is greater than 0.5 this suggests that the model fitted the data well.

3.4.2 Specification Test

When applying a logistic regression to the data we make an assumption that the logit of the probability of success is a linear combination of the independent variable. Sometimes this assumption could be invalid, i.e the logit of P the probability of success or event of interest could not be a linear combination of the independent variable. In SAS 9.3, the logit function is by default the link function when fitting a logistic regression. To evaluate whether our logistic model has all the relevant predictors and if the linear combination of them is sufficient we must make use of the specification test (Vittinghoff, 2005). The idea behind the specification test is that if the model is correctly specified, the predictor labeled (logit) in our case will be statistically significant and the square of this predictor labeled (logit*logit) will not to be statistically significant. If the square of this predictor is statistically significant this could mean that we have left out relevant variables in our model. Table 3.4 below shows the results obtained using this test in SAS 9.3.

Table 3.4: The Table for Specification Test

Parameter	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	0.0167	0.1719	0.010	0.9225
logit	0.9087	0.2144	17.96	<0.0001
logit*logit	-0.0687	0.1117	0.3800	0.5383

We see in Table 3.4 that the predictor (logit) is statistically significant (with p-value = 0.0001) and the predictor variable logit*logit is not significant (with p-value = 0.0.5383). This confirms that we have chosen a meaningful predictor in our model.

3.4.3 The relationship between a categorical dependent variable and independent variables

The results below were obtained using PROC LOGISTIC in SAS 9.3. We used Logistic regression to measure the relationship between a categorical dependent variable (smoking status) and independent variables (race, sex, age, social, marital and education).

Table 3.5: The Analysis of the relationship between a categorical dependent variable and independent variables

Parameter	Estimate	Standard Error	Wald Chi-Square	Odds Ratio
Race		10512	0.015*	
Reference (White)				
Blacks	-0.905	4.325	0.038*	0.405
Coloureds	0.974	1.715	0.190	2.648
Indians	-0.589	0.978	0.323	0.405
Sex		19.56	0.0001*	
Reference (Female)				
Males	0.492	1.873	0.171	1.635
Age		6.274	0.043*	
Reference (>55 years)				
<25 years	1.487	5.835	0.016*	4.425
25-55 years	0.861	4.102	0.043*	2.365
Socio-economic status		7.104	0.029*	
Reference (Higher level)				
Lower level	0.318	0.221	0.683	1.374
Middle level	-1.559	5.133	0.023*	0.210
Marital status		0.710	0.790	
Reference (Not currently married)				
Married	0.001	0.000	0.997	1.001
Education		2.044	0.563	
Reference (University)				
Primary level	0.801	1.664	0.197	2.228
Higher secondary level	0.280	0.380	0.538	1.324
Secondary level	0.384	1.037	0.309	1.468
Sex * Social		10.266	0.006*	
Males * Middle level	2.420	9.769	0.002 *	11.24

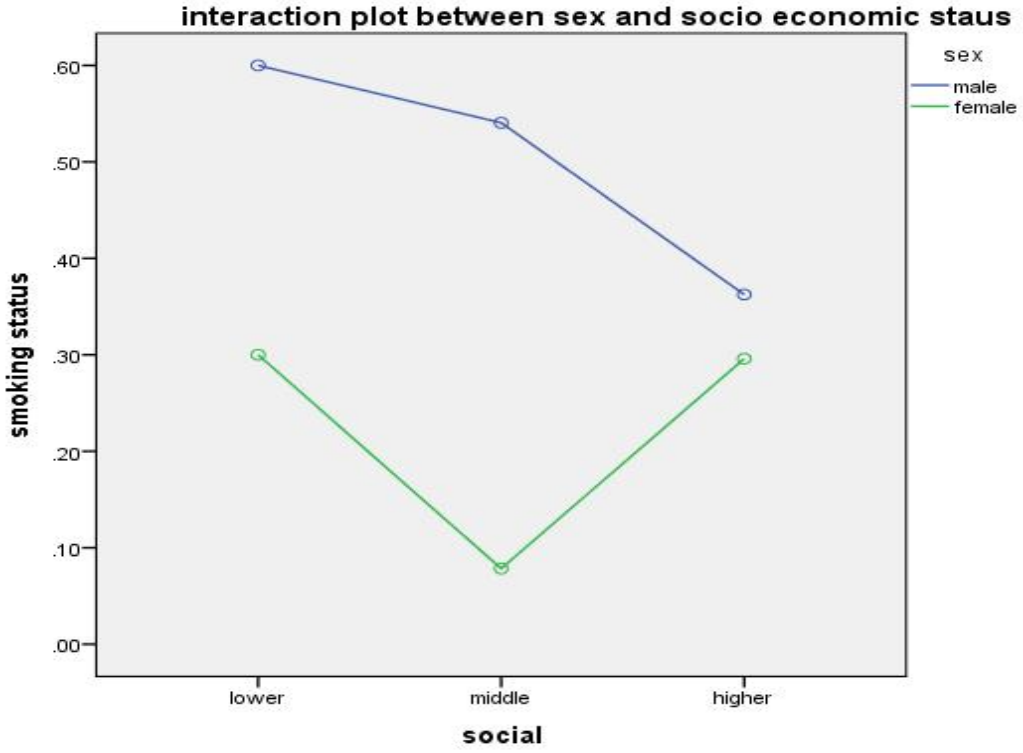
(*)→ significant at 5% level

In Table 3.5 above we find that race was significant at 5% significance level with respect to smoking status; we also find that Blacks are significantly different from Whites at the 5% significance level with respect to smoking status. Blacks are at a lower risk of smoking than Whites. We find that the risk of smoking for Blacks is 0.405 times less than the risk of smoking for Whites. Coloureds were not significantly different from Whites at the 5% significance level with respect to smoking status. Coloureds are at a higher risk of smoking than Whites. We find that the risk of smoking for Blacks is 0.405 times less than the risk of smoking for Whites. Sex was found to be significant at 5% significance level with respect to smoking status. Males were found to be 1.635 more likely to be smokers than females. Age also was significant at 5% significance level with respect to smoking status. The <25 years age group was found to be significantly different from >55 age group at the 5% significance level with respect to smoking status. The <25 years age group is at a higher risk of smoking as compared to the >55 years age group. The risk of smoking for the <25 years age is 4.425 times the risk of smoking for the >55 years age group. Also, the 25-55 years age group was significantly different from >55 years age group at the 5% significance level with respect to smoking status. The 25-55 years age group is at a higher risk of smoking than the >55 years age group. The risk of smoking for the 25-55 years age is 2.365 more times the risk of smoking for the>55 years age group.

The socio-economic status was also found to be significant at 5% significance level with respect to smoking status. The middle level of socio-economic status was found to be significantly different from higher level of socio-economic status at the 5% significance level with respect to smoking status. The middle level of socio-economic status was found to be at a lower risk of smoking

than the higher level of socio-economic status. The risk of smoking for the middle level of socio-economic status is 0.210 times less than the risk of smoking for the higher level of socio-economic status. The interaction between socio-economic status and sex was significant at 5% significance level with respect to smoking status. We also find that males in the middle level of socio-economic status are at a higher risk of smoking than females in the middle level of socio-economic status. The risk of smoking for males in the middle level of socio-economic status was found to be 11.247 times the females in the middle level of socio-economic status.

Figure 3.1: Interaction between sex and socio economic status



In an interaction plot, if the slopes of lines is not parallel then the interaction effect will be significant. In Figure 3.1 above, we see that the slope of the lines is not parallel and this tells us that the interaction between sex and socio-economic status is significant. The figure shows that females have a lower smoking status as compared to males. The difference between females in middle level and females in lower level is greater than the difference between males in middle level and males in lower level. The other noticeable difference is that females in higher level and in lower level have the same rate of smoking status but there is a greater difference between males in higher level and those in lower level. This means that males smoke more frequently than females. Females in middle level smoke less frequently compared to females in higher level and those in lower level. Males in higher level of smoking status smoke less frequently compared to males in middle and lower level, but males in lower level smoke more frequently than males in middle level and males in higher level.

Chapter 4

4 Generalized linear mixed models (GLMMs)

4.1 Introduction

Generalized linear mixed models (GLMMs) extend the conventional linear mixed models to allow for the response variables to follow non-normal distributions (Breslow et al., 1993). Generalized linear mixed models can be thought of as an extension of generalized linear models in which the linear predictor contains random effects in addition to the usual fixed effects (Breslow et al., 1993). McCulloch (2003) explains that the idea behind generalized linear mixed models (GLMMs) is that it incorporates the random effects into the linear predictor portion of a generalized linear model. The inclusion of random effects in the generalized linear model allows accommodating correlation in the context of broad class models for non-normally distributed data (McCulloch, 2003). Generalized linear mixed models also allow for departures from the Binomial and Poisson distributions with an additional parameter for extra Binomial or extra Poisson variation (Giovanini, 2008).

4.2 The Theory of Generalized Linear Mixed Model

Since the generalized linear mixed model (GLMM) is an extension to the generalized linear model (GLM) in which the linear predictor contains random effects in addition to the fixed effects it follows that it has an expectation given by the following equation:

$$E(\mathbf{y}|\mathbf{u}) = g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) = g^{-1}(\boldsymbol{\eta}) \tag{4.1}$$

$\mathbf{y}$ represents the (n $\times$ 1) response vector,
$\mathbf{X}$ the (n $\times$ p) design matrix of rank k for the (p $\times$ 1) fixed effects $\boldsymbol{\beta}$ and
$\mathbf{Z}$ the (n $\times$ q) design matrix for the (q $\times$ 1) random effects u.
The random effects u are assumed to be normally distributed with mean 0 and variance matrix $\mathbf{G}$. this means that $\mathbf{u} \sim \mathrm{N}(0, \mathbf{G})$.

$$E(\mathbf{u}) = 0 \tag{4.2}$$

and

$$Var(\mathbf{u}) = G \tag{4.3}$$

The linear predictor

The linear predictor η is obtained by combining the fixed and random effects and yields the following equation:

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} \tag{4.4}$$

There are several types of important statistical models within the class of generalized linear mixed models (GLMM). Some examples of these statistical models are :

- The linear models

A linear model is used to specify or model the linear relationship between a dependent variable or response variable and one or more predictor, independent or explanatory variables. In matrix notation, the regression equation can be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4.5}$$

- Generalized linear models

  As already mentioned above, generalized linear models are models suitable when the response variable is non-normally distributed along with exploratory variables that are categorical.

4.3 Advantages and Disadvantages of Generalized Linear Mixed Models

Generalized linear mixed models are powerful models since they are of wide applicability and practical importance. The main advantage of GLMMs is that they are more flexible in analyzing and accommodating non-normally distributed responses when random effects are present in the model (Bolker et al., 2008 and McCulloch, 2012). GLMMs are also useful when the objective of the study is to make inferences about individuals rather than the study population (Fitzmaurice et al., 2012).

The techniques for making inferences in GLMMs have advantages and disadvantages. Bolker et al. (2008) discuss these advantages and disadvantages in the context of the Wald test(Z,χ^2, t, F), the likelihood ratio test, information criteria and the deviance information criterion.

- Advantages

  The advantages of GLMMs are discussed extensively by Manning (2007)

and the reader is referred to this paper if he/she needs to investigate further. Since GLMMs are an extension of GLM, it is expected that GLMMs offer all the advantages of GLM (Mannimg, 2007). GLMMs can handle multinomial response variables and they can handle unbalanced data. GLMMs gives more information on the size and direction of effects (Manning, 2007). GLMMs can perform a combined analysis with all random effects at once and can handle missing data (Manning, 2007).

- Disadvantages

  The disadvantages of these techniques as discussed by Bolker et al. (2008) are such that, the Wald test has boundary issues, it is very poor for random effects and the t and f statistics needs residual *d.f.* The likelohood ratio test is not good for fixed effects with smaller sample sizes and is inappropriate for quasi-likelihood. The information criterion requires residual *d.f* estimate for AIC_c and there is no p-value produced.

4.4 Fitting a GLMM to the data using PROC GLIMMIX IN SAS

The GLIMMIX procedure is a procedure in SAS software which is designed for fitting Generalized Linear Mixed Models and Generalized Linear Models. It allows for non-normal data and random effects and also to include error terms that are not normally distributed (Gibbs, 2008). This procedure executes estimation and statistical inference for Generalized Linear Mixed Models (Gibbs, 2008). Since the data we are using is not continuous but binary, we apply the Glimmix procedure to the data. SAS 9.3 was the software

used to analyze the data.

Table 4.1: Type III Tests of Fixed Effects

Parameter	DF	F-value	Pr>F
Race	3	3.41	0.0181*
Sex	1	22.38	<0.0001*
Age	2	3.12	0.0458*
Social	2	3.08	0.0478 *
Marital	1	0.00	0.9941
Education	3	0.670	0.5706
Sex*Social	2	5.11	0.0067 *

(*)→ significant at 5% level

We can see from Table 4.1 that race, sex, age and socio economic status were significant at the 5% level of significance with respect to smoking status. Also, the interaction between sex and socio economic status was found to be significant at the 5% level of significance with respect to smoking status. The other variables namely marital status and educational level were not significant at the 5% level of significance with respect to smoking status.There is a significant effect of sex, social and an interaction between sex and social (Table 4.1 above).

Table 4.2: The Solutions for Fixed Effects

Parameter	Estimate	Odds Ratio	Standard Error	t-value	Pr>|t|
Reference (White)					
Blacks	-0.9400	0.404	0.4393	-2.14	0.0406 *
Coloureds	0.6128	2.607	0.6857	0.89	0.2015
Indians	-0.7660	0.550	0.6135	-1.25	0.3227
Reference (Female)					
Males	1.244	5.314	0.2821	4.41	<0.0001*
Reference (>55 years)					
<25 years	1.3538	4.427	0.6058	2.23	0.0167*
25-55 years	0.9744	2.363	0.4254	2.29	0.0446*
Reference (Higher level)					
Lower level	0.8719	2.379	0.5275	1.65	0.6499
Middle level	-0.1350	0.706	0.4463	-0.30	0.0247*
Reference (Not currently married)					
Married	0.0914	2.216	0.3403	0.270	0.9941
Reference (University)					
Primary level	0.9793	2.216	0.5993	1.63	0.2023
Higher secondary level	0.4888	1.320	0.4544	1.08	0.5440
Secondary level	0.6308	1.464	0.3806	1.66	0.3132
Males*Middle level	2.4197		0.7754	3.12	0.0020*

(*)→ significant at 5% level

From Table 4.2 above we find that the only significant variable in the Race category was Blacks. The odds ratio for Blacks is 0.404 as compared to the risk of smoking for Whites. This result shows that Blacks are at a lower risk of smoking than Whites. Coloureds have a higher odds ratio of smoking compared to the Blacks . Males were also found to be significant at the 5% level of significance. Sex is significant at 5% level of significance, males have higher odds ratio of smoking compared to females. The odds ratio of smoking for males is 5.134 times that of females. Age was also found to be significant

at 5% level of significance with respect to smoking status. Both age groups were found to be at a higher level of smoking than the >55 age group. The <25 years age group was found to be 2.363 times more likely to smoke than the >55. The odds of smoking for the 25-55 years age group were found to be 2.379 than the odds of smoking for the >55. We also found that the middle level of socio-economic status was significant at the 5% level of smoking status. It was found that the middle of socio-economic status was at a lower risk of smoking as compared to the higher level of socio-economic status. The risk of smoking for middle level of socio-economic status is 0.706 as compared to the risk of higher level of socio-economic status. We also found that the interaction between males and the middle level of socio-economic status was significant at 5% level of significance with respect to smoking status. All the other variables were found not to be significant at the 5% level of significance.

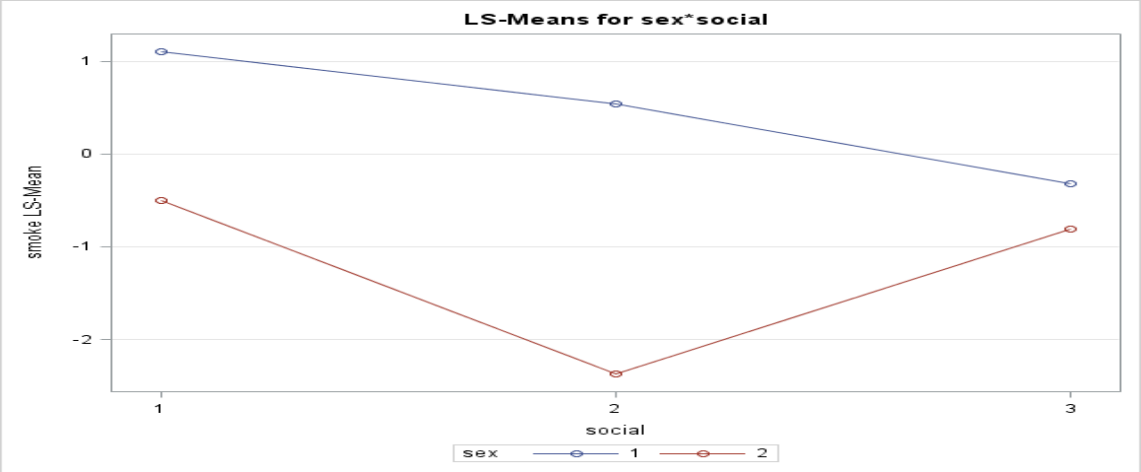


Figure 4.1: The interaction plot between males and the middle level of socio economic status

The interaction between the two effects is evident in the lack of parallelism in Figure 4.1 above. When lines of the effects are not parallel, they show

that the interaction was significant between the two effects.
For the above results we fitted a GLMM model using proc glimmix to investigate the effect that cluster as random effect has on the results.

4.5 Fitting a GLMM to the data using PROC NLMIXED

An alternative procedure in SAS 9.3 for fitting non-linear mixed models is the NLMIXED procedure. This procedure is used when models with both fixed effects and random effects are allowed to have a non-linear relationship with the response variable. The most common models fitted by the NLMIXED procedure are those with a conditional distribution for the response variable (Flom et al., 2007). The results below were obtained using SAS 9.3. The NLMIXED model can be given as follows:

$$\eta_{ij} = log(\frac{\pi}{1-\pi}) = \beta_0 + \pi_{ij} + \mu_i \tag{4.6}$$

where the notation p_j denotes the j^{th} treatment and the μ_i are assumed to be $iid\ N(0,\ \sigma_u^2)$.

Advantage of using NLMIXED than GLIMMIX

The difference between NLMIXED than GLIMMIX is in the estimation method used by each procedure. Both procedures approach parameter estimation as an optimization problem, which solves for an approximation of the marginal log likelihood (Flom et al., 2007). NLMIXED accomplishes this using an integral approximation through Gaussian quadrature, whereas GLIMMIX relies on approximation of a linear mixed model (linearization). Advantages of the NLMIXED method are that it is generally more accurate

and generates a true log-likelihood fit statistic that can be used to compare nested models. The method also permits greater flexibility to accommodate user-defined likelihood functions. GLIMMIX, in contrast, can produce potentially biased estimates for both fixed effects and covariance parameters, especially for binary data (Flom et al., 2007).

Table 4.3: The Parameter Estimates

Parameter	Estimate	Standard Error	t-value	Pr>$\|t\|$
Race	1.379	0.590	2.34	0.0067*
Sex	-4.478	1.504	-2.98	0.0003 *
Age	-2.660	1.194	-2.23	0.0.007*
Social	-1.872	1.077	-1.74	0.0025*
Marital	-0.946	1.205	-0.79	0.6228
Education	-0.683	0.554	-1.23	0.2214

(*)→ significant at 5% level

Table 4.3 lists the six parameters, their maximum likelihood estimates, standard errors, and inferential statistics. The output are coefficient estimates for Race, Sex, Age, Social, Marital and Education . Each of these parameters can be converted to an adjusted odds ratio by exponentiating it. The results indicate that Race, Sex,Age and Social are significant predictors of smoking status since the coefficients are significant at 5% level of significance.

The difference among the result of GLM, GLMM, PROC GLIMMIX and PROC NLMIXED

When GLM was fitted to the data using Logistic regression, race was significant at 5% significance level with respect to smoking status; sex was significant at 5% significance level with respect to smoking status; age was significant at 5% significance level with respect to smoking status, socio economic status was significant at 5% significance level with respect to smoking status; marital status was found not to be significant at 5% significance level with respect to smoking status and education was found not to be significant at 5% significance level with respect to smoking status. When GLMM was fitted to the data using proc GLIMMIX, race, sex, age and socio economic status were significant at the 5% level of significance with respect to smoking status. Also, the interaction between sex and socio economic status was found to be significant at the 5% level of significance with respect to smoking status. The other variables namely marital status and educational level were not significant at the 5% level of significance with respect to smoking status. When GLMM was fitted to the data using proc NLMIXED, Race, Sex,Age and Social are significant predictors of smoking status since the coefficients are significant at 5% level of significance. The results of these procedures are similar, there was no difference among the results of these procedures.

Chapter 5

5 Multiple Correspondence Analysis (MCA)

5.1 Introduction

Multiple correspondence analysis (MCA) is an analysis method for examining relationships among the categorical variables (Abdil and Valentin, 2007). This relationship is displayed graphically among the categories of several variables. MCA is commonly used when analyzing large data set. The technique is used to display complex relationships among the variables by means of plots. The MCA algorithm assigns to each category of each categorical variable a two dimensional coordinate (Du, 2003). The coordinates could be close or far apart from each other. The coordinates close to each other are said to have a closer association than those coordinates which are not close to each other. MCA is an extension of simple correspondence analysis (CA) in that it is applicable to a large set of categorical variables (Le Roux and Rouanet, 2010; Greenacre, 2007). This approach forms part of a family of descriptive methods that reveal patterning in complex data. MCA has special characteristics in describing these patterns. Each variable or unit of analysis is located at a point in a low-dimensional space. MCA has the ability to map several variables and individuals. This mapping allows for the visualization of complex structures between the variables and individuals.

Simple correspondence analysis is an exploratory tool used to analyze a contingency table with only two categorical variables whereas MCA extends this analysis from two categorical variables to several categorical variables (Du,

2003). R.A Fisher is among those who developed correspondence analysis but the technique was then popularized in the 1960s and 1970s by Jean-Paul Benzécri (Le Roux and Rouanet, 2010). The method, as stated before, is used to study the association between two or more qualitative variables (Greenacre, 2007). Since this method extends from a simple correspondence analysis, in order to perform the method one has to apply the simple correspondence algorithm to an indicator matrix (Greenacre, 2007). An indicator matrix is explained by Le Roux and Rouanet (2010) as an individual x variable matrix, where the individuals are represented by the rows and the dummy variables indicating the categories of the variable are represented by columns.

The analysis in multiple correspondence analysis is actually based on the inner product of the matrix called the Burt table, which is said to be the result of the inner product of an indicator matrix (Demosthenes et al., 2004). Since multiple correspondence analysis is an exploratory technique, it then follows that no statistical significance test should be applied to the result of correspondence analysis (Panagiotakos et al., 2004). The main idea behind multiple correspondence analysis is to produce a simplified version of the information in a large frequency table (Demosthenes et al., 2004). MCA is a technique used for the analysis of nominal categorical data set. This technique is used when analyzing a set of observations from a set of nominal variables (Abdi and Valentin, 2007). These nominal variables all have several levels and each level is taken as a binary variable (Abdi and Valentin, 2007). MCA is an exploratory technique but it is very powerful since it can provide the researcher with a key insight on the relationship among the variables.

Correspondence analysis is an exploratory technique used to analyze a simple two-way table and is normally generalized to the case of R categorical variables. When R exceeds two, that is when there are more than two categories to be analyzed, multiple correspondence analysis is then used to analyze the indicator matrix of the data set (Du, 2003). When CA is applied to the indicator matrix it provides factor scores for rows and columns but these scores, however, need to be rescaled for when using MCA (Abdi and Valentin, 2007). Let the indicator matrix be denoted by $\mathbf{X}$ and let it be given by

$$\mathbf{X} = (\mathbf{X}_1, ..., \mathbf{X}_R) \tag{5.1}$$

where $\mathbf{X}_R$ represents an n $\times m_r$ matrix (Du, 2003) so that the component of any Burt matrix be given as follows

$$\mathbf{B} = \mathbf{X}^T\mathbf{X} \tag{5.2}$$

where $\mathbf{B}$ is a symmetric matrix. The coordinates of categories found when using MCA are obtained using this Burt matrix.

Advantages and Disadvantages of Multiple Correspondence Analysis

Multiple correspondence analysis (MCA) provides more detailed results when compared to other exploratory techniques that can be used for the same purposes. The technique is well suitable for categorical data sets. MCA has the ability to produce displays whose rows and column geometries have similar interpretations which assist in detection of relationship among the variables. The technique has very flexible data requirements. The main disadvantage of this technique is that it is only suitable for categorical data set (i.e discrete

data set).

Assumption of Multiple Correspondence Analysis (MCA)

- MCA assumes that the nominal scaling level is the same for all variables (multiple).
- MCA assumes that the data set contains at least three valid cases.
- The data set for analysis must consist of only positive integer values.

Properties of Multiple Correspondence Analysis (MCA)

Multiple correspondence analysis (MCA) is a method that allows study of the association between two or more variables. This is the application of CA to cross-tabulations of more than two categorical variables and is generally defined in two practically equivalent ways (Greenacre and Pardo, 2005).

- MCA can be defined as the correspondence analysis of the individual response data in the format of an indicator matrix say $\mathbf{X}$ which codes individual responses in a 0/1 indicator form, where all response categories form the columns of the indicator matrix (Greenacre and Pardo, 2005); or
- It can be defined as the correspondence analysis of all cross-tabulations of variables joined in the Burt matrix $\mathbf{B}$ including also the diagonal cross-tabulations of each variable with itself (Greenacre and Pardo, 2005).

5.1.1 The application of Multiple Correspondence Analysis to the data

As already mentioned, Multiple Correspondence Analysis is a technique used to display graphically the relationship among the categories of several variables. Figure 5.1 below shows this relationship. The results below were obtained by fitting the data in SPSS 21.

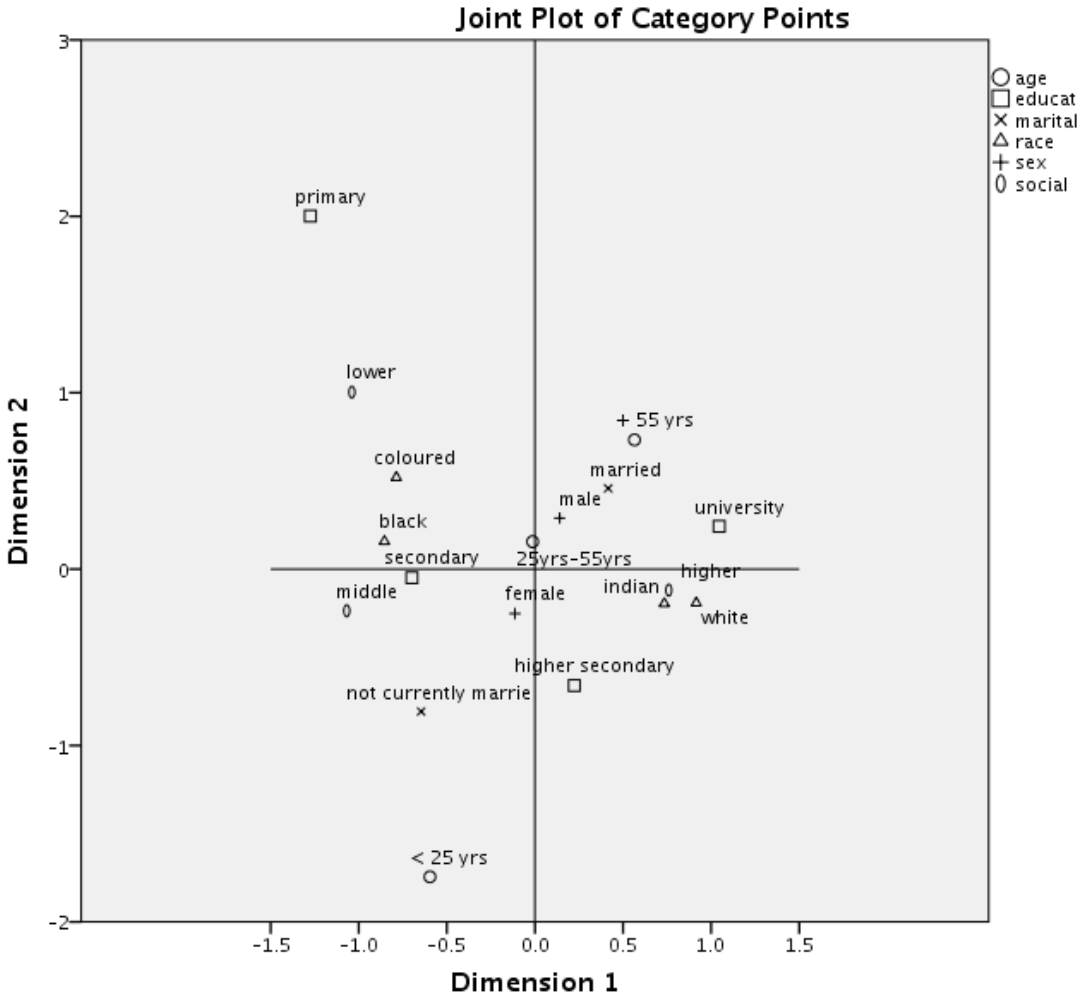


Figure 5.1: Joint Plot of Category Points

Figure 5.1 shows the association among the categories of the variables. The top right part of the plot shows that males, university, the married status group and the >55 years are associated. This means that married males who are >55 years old with a university qualification have a similar smoking status. In the bottom right part of the plot, Whites, Indians, the higher secondary level and the higher level of socio-economic status are associated. This shows that Indians and Whites of higher level of socio-economic status with higher secondary education have a similar smoking status. In the bottom left part of the plot, females, the <25 years, not currently married, secondary level and the middle level of socio economic status are also related. This means that not currently married females who are <25 years old of middle level of socio-economic status with a secondary qualification have a similar smoking status. The top left part of the plot shows that the categories primary level, lower level of socio economic status, Coloureds and Blacks are associated. The 25-55 years age group is on the centroid but more on the top left part of the plot than on the top right, so we conclude that it is also associated with categories on the top left of the plot. This means that Blacks and Coloureds who are 25-55 years old of lower level of socio economic status with a primary qualification have a similar smoking status.

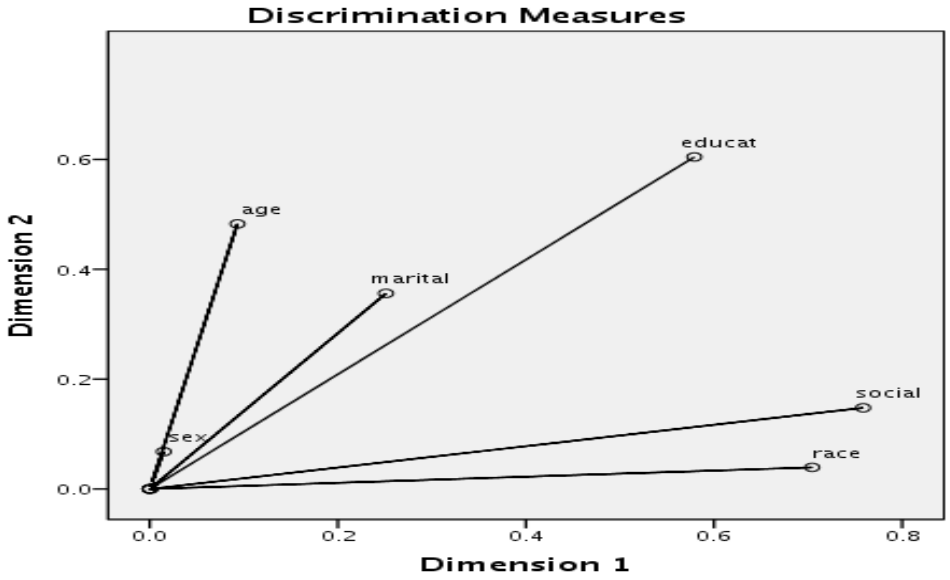


Figure 5.2

Figure 5.2 shows that race discriminates mostly along dimension one and contributes very little in the inertia of dimension two. Sex discriminates mostly along on dimension two and is contributing very little in the inertia of dimension one. Age contributes very little on the inertia of dimension one and discriminates mostly along dimension two. Socio economic status discriminates mostly along dimension one and contributes more on the inertia of dimension one and little on the inertia of dimension two. Marital status discriminates mostly on dimension two and has a smaller contribution on the inertia of dimension one. Education discriminates well on both dimension one and dimension two and has almost the same contribution in the inertia of dimension one and dimension two. the distance or length between two variables signifies homogeneity.

Chapter 6

6 Methods for Handling Missing Data

6.1 Introduction

Analyzing data without considering the effect of missing data could potentially lead to biased results due to reduced sample sizes (Andrew and Selamat, 2012). Most standard statistical procedures were developed to handle complete data sets only; these procedures can therefore produce very biased and less efficient estimates. One of the current solutions to handling the problem of missing data is multiple imputation. There are a number of imputation methods that can be used for handling the missing responses. These methods lead to valid inferences under certain conditions. We consider the following two methods

- Multiple Imputation (MI)
- Inverse Probability Weighting (IPW) and

These three methods are available in standard statistical software such as SAS and SPSS. Our original data set contained no missing observations, we then created the missing data by deleting some parts of our data. The mechanism used in creating the missing data was a random mechanism where parts of the data were deleted randomly. This mechanism allowed each member of the data to have an equal probability of being chosen. The missing observations will then be imputed using MI, IPW and LOCF. To the newly imputed data sets, we then apply survey logistic regression and PROC GLIMMIX.

6.1.1 The application of survey logistic procedure to the original data set.

When the original data set was fitted using the survey logistic procedure, we found an AIC value of 341.799.

Table 6.1: The Analysis of Effects

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr>ChiSq**
race	3	198.423	<.0001*
sex	1	97.1127	<.0001*
age	2	7.6257	0.0221*
social	2	3.9832	0.1365
marital	1	0.0041	0.9487
education	3	1.4886	0.6849
race*sex	3	251.4357	<.0001*

(*)→ significant at 5% level

From Table 6.1, we find that race, sex and age were significant at the 5% level of significance with respect to smoking status. Also, the interaction between race and sex was found to be significant at the 5% level of significance.

Table 6.2: The Analysis of Maximum Likelihood Estimates

Analysis of Maximum Likelihood Estimates				
Parameter	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr>ChiSq**
Blacks	0.5771	0.3705	2.4263	0.1193
Coloured	2.4358	0.5683	18.3716	<.0001*
Indian	-4.6602	0.4819	93.5337	<.0001*
Male	2.2571	0.229	97.1127	<.0001*
<25 years	0.5841	0.3271	3.1894	0.0741
25-55 years	0.1782	0.1778	1.0044	0.3162
Lower level	0.7044	0.3823	3.3946	0.0654
Middle level	-0.3821	0.2421	2.4899	0.1146
Married	-0.0141	0.2192	0.0041	0.9487
Primary level	0.3896	0.4031	0.9339	0.3338
Secondary level	-0.0517	0.2107	0.0602	0.8062
Higher secondary level	0.0534	0.2699	0.0391	0.8432
Males*Blacks	-1.0823	0.3122	12.0162	0.0005*

(*)→ significant at 5% level

Table 6.2 shows us that race was significant with respect to smoking status. Sex was also found to be significant with respect to smoking status. The interaction between Blacks and males was found to be significant at the 5% level of significance with respect to smoking status.

6.1.2 The application of GLIMMIX procedure to the original data set.

When the glimmix procedure was fitted to the original data set, we obtained an AIC value of 283.55

Table 6.3: The Analysis of Fixed Effects

Type III Tests of Fixed Effects			
Effect	**DF**	**F Value**	**Pr>F**
race	3	3.38	0.019*
sex	1	16.9	<.0001*
age	2	3.28	0.0393*
social	2	1.99	0.1391
marital	1	1.31	0.2529
education	3	1.04	0.3734
social*marital	2	3.7	0.0262*

(*)→ significant at 5% level

From Table 6.3, race, sex, age and the interaction between socio economic status and marital status were significant with respect to smoking status.

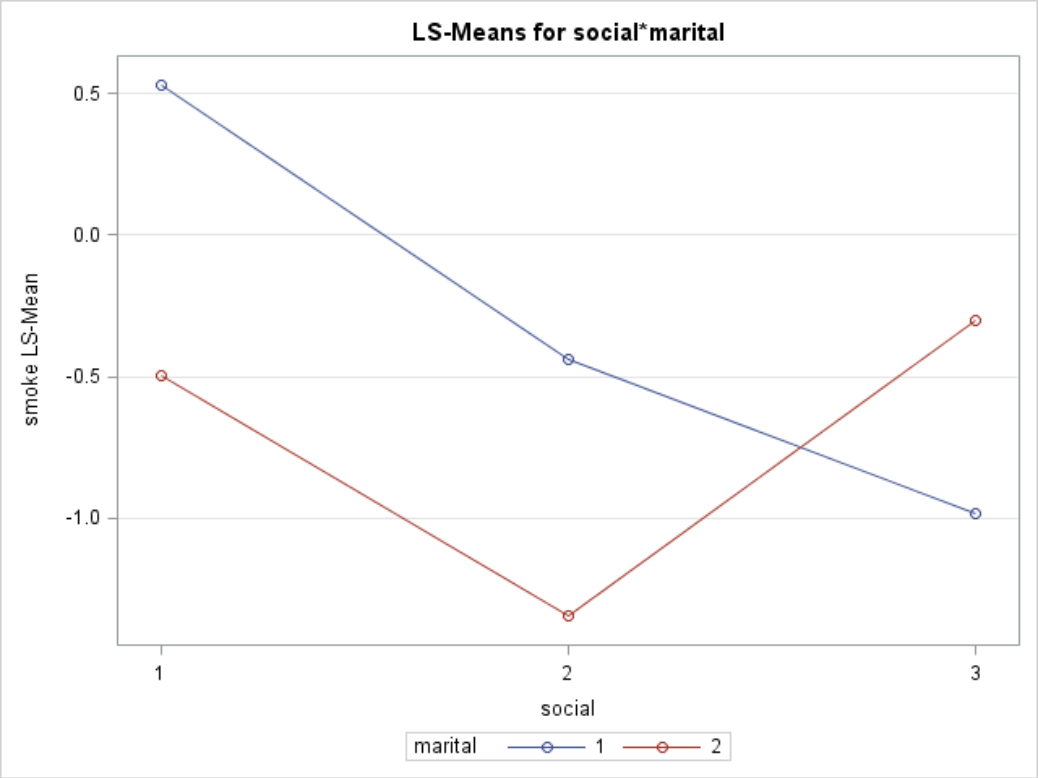
Table 6.4: The Solutions for Fixed Effects

Solutions for Fixed Effects					
Parameter	**Estimate**	**Odds Ratio**	**Standard Error**	**t–Value**	**Pr>|t|**
Blacks	-0.9926	0.371	0.4409	-2.25	0.0253*
Coloured	0.5294	1.698	0.6888	0.77	0.4429
Indian	-0.7608	0.467	0.6135	-1.24	0.2161
Male	1.1745	3.237	0.2857	4.11	<.0001*
<25 years	1.3763	3.96	0.6095	2.26	0.0248*
25-55 years	1.0215	2.777	0.4286	2.38	0.0179 *
Lower level	-0.1929	1.933	0.7406	-0.26	0.7947
Middle level	-1.0441	0.779	0.589	-1.77	0.0776
Married	-0.6794	1.518	0.432	-1.57	0.1171
Primary level	0.8414	2.32	0.6112	1.38	0.1699
Secondary level	0.5213	1.684	0.4557	1.14	0.2538
Higher secondary level	0.6227	1.864	0.3817	1.63	0.1041
Middle level*Married	1.5875		0.6862	2.31	0.0215*

(*)→ significant at 5% level

In Table 6.4 we find that race is significant, The odds ratio of Blacks is 0.371 times that of Whites; the odds ratio for Coloureds is 1.698 times that of whites and the odds ratio for Indians is 0.467 times that of Whites. Sex was found to be significant at 5% level of significance. The odds ratio of males is 3.237 times that of females. Age was found to be significant at 5% level of significance, the odds ratio of <25 years age group is 3.96 times that of >55 years age group. The odd ratio for 25-55 years age group is 2.777 times that of >55 years age group. Socio economic status was found not to be significant at 5% level of significance. The interaction was found to be significant at 5% level of significance.

Figure 6.1: The interaction plot for socio economic status and marital status



The interaction plot of socio-economic status and marital status shows that the interaction is significant since the lines are not parallel between the middle level and higher level of socio economic status and they intersect.

6.2 Multiple Imputation (MI)

6.2.1 Introduction

Missing values are observed predominantly in medical sciences and social sciences. The missing values can be imputed using a technique called multiple imputation. Multiple imputation is the general technique for handling missing data and is available in all the commonly used statistical packages (Stern et al., 2009). Rubin (1996) explains multiple imputation as the technique that was originally designed to handle missing data in public data bases. The use of multiple imputation was popularized by Donald B. Rubin in the early 1970s. This statistical technique is said to be flexible since it can be used appropriately in a number of situations (Rubin, 1996). Previously, the technique was used to generate public use data sets that were used by many different users. However, over the years the use of multiple imputation has attracted many researchers in many different fields (Rubin, 1996). Rubin (1996) emphasizes that the main purpose of multiple imputation is to provide a valid statistical inference. Multiple imputation replaces each missing data value with a set of plausible values that represent the uncertainty about the correct value to impute into the data. The multiple imputed data sets are then analyzed using standard statistical procedures for complete data.

To predict the missing observations, multiple imputation uses the existing values from other observations (Wayman, 2003). When missing values are imputed by the predicted, a new complete data set is formed. This new data set is called the imputed data set (Wayman, 2003). One data set is not enough, so the process is repeated a number of times to create multiple imputed data sets (Wayman, 2003). These imputed data sets are then an-

alyzed using standard procedures. Stern et al., (2009) discusses two stages that multiple imputation follows when generating a data set. In the first stage, a multiple data set is created. In the second stage, the standard statistical methods are used to fit the model of interest to each data of the imputed data set (Stern et al., 2009). Once the model is fitted to each of the data sets, the produced multiple results are then combined to an overall analysis (Wayman, 2003). The modeling strategy required by this technique is such that the user must model the distribution of each variable with the missing values in terms of the observed data (Stern et al., 2009). Stern et al., (2009) describes the multiple imputation as a technique with a very high potential to improve the validity of medical research.

6.2.2 How Does Multiple Imputation (MI) Work?

The first phase in multiple imputation is to generate values to use when imputing the missing data (Wayman, 2003). In order to generate the imputes, a model has to be identified. The model then uses the non-missing variables, often called predictor variables, in the data to create the imputes. The missing data are then filled in m times resulting in m complete data sets. The user decides on the number of imputed data sets to be created. Most researchers use m between 5 and 10. Once multiple imputation has created the imputed data sets then the m complete data sets are analyzed by using standard analysis or methods as mentioned above. Foe example the analysis can simple be logistic regression if the data is binary. These procedures could be any procedures used when analyzing data with no missing values except that the analysis has to be performed on each imputed data set separately.

6.2.3 How Does Multiple Imputation (MI) Combine The Data Set?

After analyzing each imputed data set, multiple imputation then combines this analysis to produce an overall set of estimates to produce valid statistical inferences about the parameters (Abdi and Valentin, 2007). The process by which the results are combined is the same for any complete data set used (Wayman, 2003). The rules established by Rubin (1987) are then used to combine the results (Wayman, 2003). The rules state that in order to combine the estimates of the parameter of interest, one has to average the individual's estimates produced by the analysis of each imputed data set (Wayman, 2003). This statement can be generalized as follows, for a single parameter of interest θ,

$$\overline{\theta} = \frac{1}{M} \sum_{i=1}^{m} \hat{\theta}_i \tag{6.1}$$

where m is the number of imputed data set, $\hat{\theta}_i$ is the point estimate and θ is the parameter of interest. The total variance for the overall MI estimate is given by

$$T = \overline{U} + (1 + \frac{1}{m})B \tag{6.2}$$

where the within imputation variance is given by

$$\overline{U} = \frac{1}{M} \sum_{i=1}^{m} U_i \tag{6.3}$$

and the between imputation variance is given by

$$B = \frac{1}{M-1}\sum_{i=1}^{m}(\hat{\theta}_i - \overline{\theta})^2 \tag{6.4}$$

The MI procedure in SAS 9.3 uses PROC MI and PROC MIANALYZE procedures when creating and analyzing data sets. PROC MI is used for generating the multiple data sets. These data sets will then be analyzed using a standard procedure such as PROC GLMMIX. PROC MIANALYZE is then used for making statistical inference, thus combining the results. The data was imputed using multiple imputation then Survey logistic procedure and the GLIMMIX procedure were applied to the imputed data and the results below were obtained.

The missingness in our data set is then assumed to be MCAR since the missing data was created by deleting some parts of our data set. This means there is no relationship between the missingness of the data and any values, observed or missing.

6.2.4 The application of GLIMMIX procedure to the multiple imputed data set.

The PROC GLIMMIX procedure was used to analyze the imputed data and the results below were obtained. The generalized Chi-square value found when using this technique was 316.82.

Table 6.5: The Tests of Fixed Effects

Type III Tests of Fixed Effects			
Effect	**DF**	**F Value**	**Pr>F**
race	3	2.12	0.0976
sex	1	18.95	<.0001*
age	2	3.21	0.0419 *
social	2	1.16	0.3137
marital	1	2.12	0.1468
education	3	0.26	0.8531
social*marital	2	3.73	0.0253*

(*)→ significant at 5% level

Table 6.5 shows the analysis of race, sex, age, marital, education and social and the interaction between social and marital status. We find that sex, age and the interaction between social and marital status were significant at 5% level with respect to smoking status.

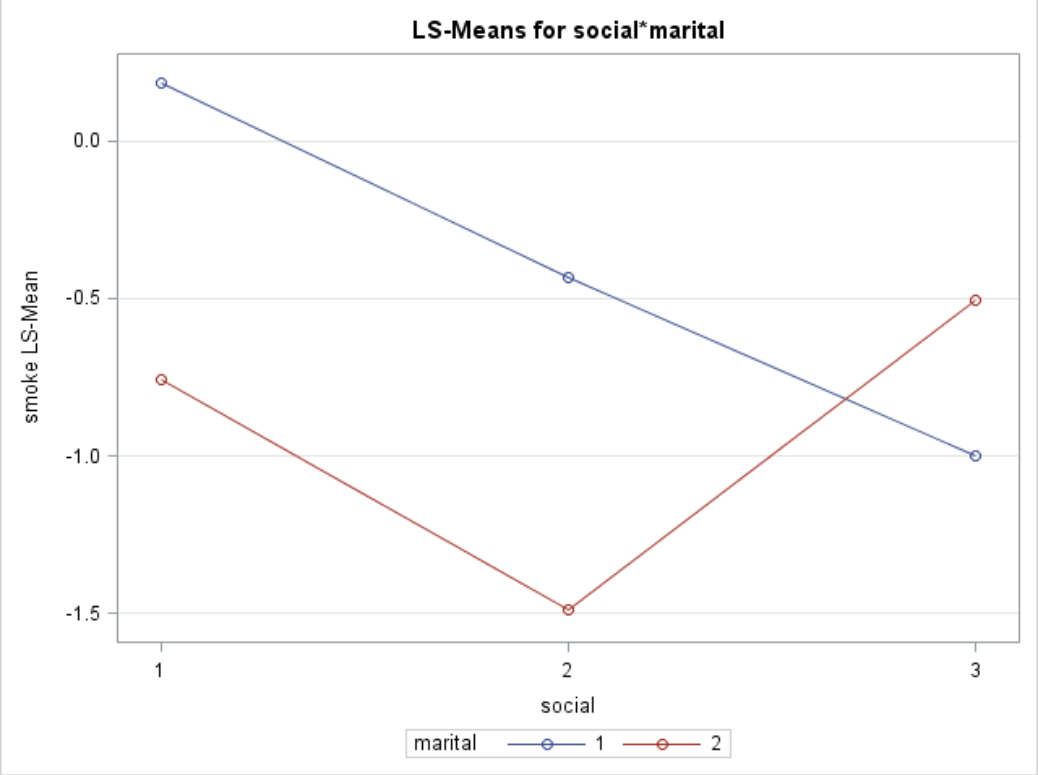
Table 6.6: The Solutions for Fixed Effects

Solutions for Fixed Effects					
Parameter	**Estimate**	**Odds Ratio**	**Standard Error**	**t Value**	**Pr>|t|**
Blacks	-0.7416	0.476	0.4145	-1.79	0.0746
Coloured	0.1564	1.169	0.6566	0.24	0.8119
Indian	-0.9476	0.388	0.6239	-1.52	0.1299
Male	1.1547	3.173	0.2652	4.35	<.0001*
<25 years	1.3323	3.79	0.5461	2.44	0.0153*
25-55 years	0.7759	2.173	0.3787	2.05	0.0414 *
Lower level	-0.2528	1.595	0.7153	-0.35	0.7241
Middle level	-0.982	0.812	0.5347	-1.84	0.0673
Married	-0.4947	1.651	0.3934	-1.26	0.2096
Primary level	0.2461	1.279	0.5584	0.44	0.6597
Secondary level	0.321	1.378	0.4155	0.77	0.4405
Higher secondary level	0.2757	1.317	0.3455	0.8	0.4255
Middle level*Married	1.5481		0.6282	2.46	0.0143*

(*)→ significant at 5% level

In Table 6.6 we find that sex and age were both significant at 5% level with respect to smoking status. Males were also found to be significant at 5% level with respect to smoking status. We found that the risk of smoking for males was 3.173 as compared to the risk of smoking for females. The <25 years age group and the 25-55 years age group were also found to be significant at the 5% level with respect to smoking status. We found that both the age groups were at a higher risk of smoking as compared to the >55 years age group with risks of 3.79 and 2.173 respectively. We also found the interaction between the middle level of socio-economic status and the married group to be significant at 5% level with respect to smoking status.

Figure 6.2: The interaction plot between socio-economic status and marital status



The lines in the plot intersect proving that the interaction between socio-economic status and marital status is significant.

6.3 Inverse Probability Weighting (IPW)

6.3.1 Introduction

As previously mentioned, missing data is encountered in almost all studies. The problem with missing data is that it is an issue of concern with standard analyses that are restricted to subjects with complete data (Vansteelandt et al., 2010). The issue with missing data is that the missing values in the data set can result in biased conclusions (Vansteelandt et al., 2010). Seaman and White (2013) proposed the use of Inverse Probability Weighting (IPW) for correcting the bias caused by missing data. IPW is valid under the missing at random (MAR) mechanism but requires that the dropout model be specified in terms of observed outcomes and/or variates (Satty, 2012). In sample surveys where there are unequal sampling fractions, inverse probability weighting is used to adjust for these sampling fractions (Seaman and White, 2013). Over the past decade, the use of inverse probability weighting has been appraised and affirmed by many researchers (Vansteelandt et al., 2010). Inverse probability weighting is particularly useful for missing data where the response variable is binary and the missingness is monotone.

McCaffrey et al. (2013) mentions that some of the inverse probability weighting estimators possess a property of double robustness. These estimators are consistent and are asymptotically normal if the model for the mean or for the response or for the treatment is correctly specified (McCaffrey et al., 2013). There is a condition regarding the consistent and normality of the inverse probability weighting estimators.(McCaffrey et al., 2013). The condition is such that the estimators will only hold when the treatment or response is independent of the outcome of interest on a set of observed co-

variates (McCaffrey et al., 2013). McCaffrey et al. (2013) further explains that the literature on inverse probability weighting only considers the covariates that are free of measurement error but the use of inverse probability weighting estimation is commonly used when covariates are measured with error. Vansteelandt et al. (2010) points out a problem with the practical usefulness of the inverse probability weighting methods since the literature of this topic is not easily accessible simply because the inverse probability weighting estimators are not very stable in the presence of influential weights.

Unlike in simple surveys where weights are known, the weights in IPW are estimated using the observed data. To illustrate how these weights are estimated we follow an example given by Satty (2012). Suppose that our complete data was as follows:

Group	1	2	3
Response	5 5 5	6 6 6	7 7 7

The average response for this data is 6, this average is not biased since the data is complete. However, if now the data has missing values as follows:

Group	1	2	2
Response	5 ? ?	6 6 6	? 7 7

from the above data, the average now changes to $\frac{37}{6}$, which is biased. The probability of response in group 1 can be calculated as $\frac{1}{3}$, 1 in group 2 and as $\frac{2}{3}$ in group 3. The weighted average can be calculated as follows when each observation is weighted by 1/[probability of observed response]. Therefore, the weighted average is calculated as:

$$\frac{5\times\frac{3}{1}+(6+6+6)\times 1+(7+7)\times\frac{3}{2}}{\frac{3}{1}+1+1+1+\frac{3}{2}+\frac{3}{2}} = 6$$

The weighted average response is similar to the average response of 6, this suggest that IPW has corrected the bias caused by the missing observation.

The data was imputed using inverse probability weighting in SAS 9.3 using a dropout macro. The Survey logistic procedure and the GLIMMIX procedure were applied to the data and the results are shown below.

6.3.2 The application of GLIMMIX procedure to the IPW data set

The generalized Chi-square value we found when we fitted the IPW data set was 319.26.

Table 6.7: The Tests of Fixed Effects

Type III Tests of Fixed Effects			
Effect	**DF**	**F Value**	**Pr>F**
race	3	2.59	0.0533
sex	1	14.17	0.0002 *
age	2	2.34	0.0985
social	2	1.02	0.3619
marital	1	0.24	0.6256
education	3	0.7	0.5531
social*marital	2	3.03	0.049 *

(*)→ significant at 5% level

In Table 6.7 above only two variables and the interaction were found to be significant at 5% level with respect to smoking status. We found that sex, age and the interaction between socio-economic status and marital status were significant at 5% level.

Table 6.8: The Solutions for Fixed Effects

Solutions for Fixed Effects					
Parameter	**Estimate**	**Odds Ratio**	**Standard Error**	**t Value**	**Pr>|t|**
Blacks	-0.7975	0.45	0.4398	-1.81	0.0708
Coloured	0.5795	1.785	0.6923	0.84	0.4032
Indian	-0.7682	0.464	0.5727	-1.34	0.1809
Male	0.9914	2.695	0.2634	3.76	0.0002 *
<25 years	1.0151	2.76	0.542	1.87	0.0621
25-55 years	0.7805	2.182	0.3857	2.02	0.044 *
Lower level	-0.0546	1.82	0.709	-0.08	0.9387
Middle level	-0.706	0.989	0.5236	-1.35	0.1786
Married	-0.736	1.177	0.3935	-1.87	0.0624
Primary level	0.09241	1.097	0.56	0.17	0.8691
Secondary level	0.271	1.311	0.4102	0.66	0.5094
Higher secondary level	0.4741	1.607	0.351	1.35	0.1779
Middle level*Married	1.389		0.6272	2.21	0.0276 *

(*)→ significant at 5% level

From Table 6.8 we can see that only males, the 25-55 years age group and the interaction between middle level of socio-economic status and marital status were significant at 5% level with respect to smoking status. We found that males are at a higher risk of smoking as compared to females. The risk of smoking for males is 2.695 more as compared to the risk of smoking for females. Again, we found that the 25-55 years age group was at a higher risk of smoking compared to the >55 years age group. The 25-55 years age group was found to be 2.182 more likely to smoke as compared to the >55 years age group.

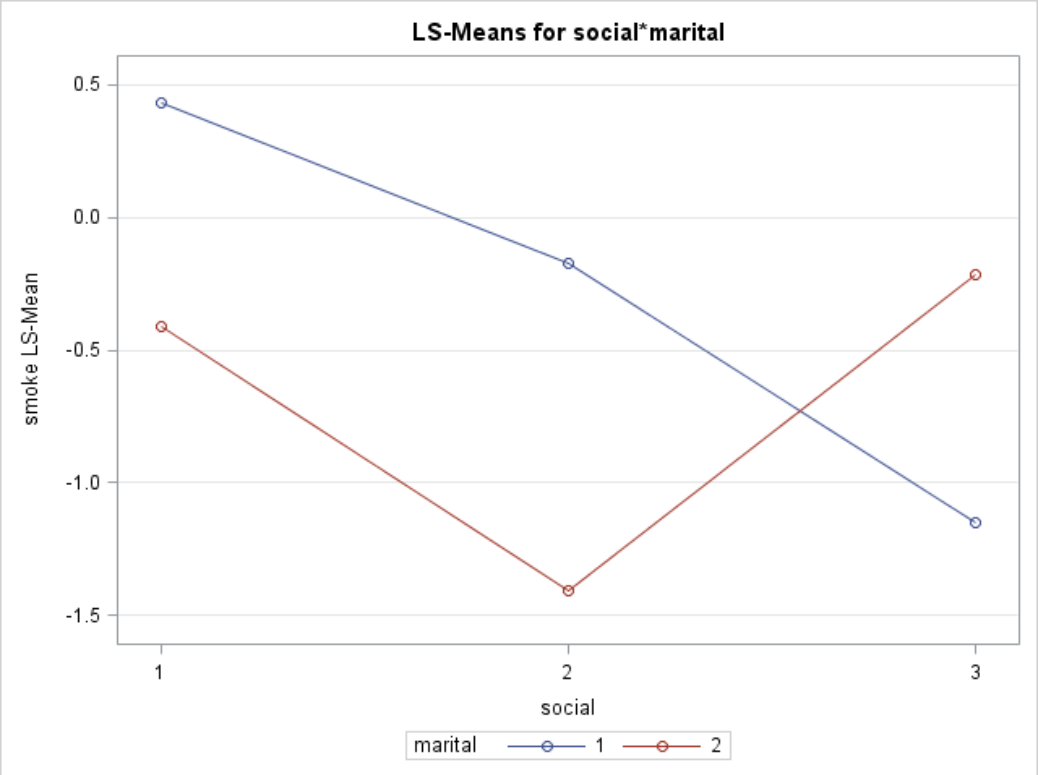


Figure 6.3: The interaction plot between the socio-economic status and marital status

Figure 6.3 shows the interaction between the socio-economic status and marital status. We can see that the lines in figure 6.3 intersect proving that the interaction between these variables is significant.

6.4 Comparison of Multiple Imputation and Inverse Probability Weighting

The most commonly used methods of handling missing data are multiple imputation, inverse probability weighting and last observation carried forward. We have considered these three techniques to fill in (impute the missing values) the missingness that was present in our data. PROC GLIMMIX was used to analyze the data and the results are shown in the tables below. Our main aim was to compare multiple imputation (MI)and inverse probability weighting (IPW) to assess which method best handled the missing values in our data set.

Table 6.9: The Tabulation of Multiple Imputation and Inverse Probability Weighting for the Solutions of Fixed Effects using proc GLIMMIX

Solutions for Fixed Effects						
Original data set			**MI data set**		**IPW data set**	
Generalized Chi-square	**283.55**		**316.82**		**319.26**	
Parameter	**Standard Error**	**Pr >|t|**	**Standard Error**	**Pr >|t|**	**Standard Error**	**Pr >|t|**
Reference (Whites)						
Blacks	0.4409	0.0253 *	0.4145	0.0746	0.4398	0.0708
Coloured	0.6888	0.4429	0.6566	0.8119	0.6923	0.4032
Indian	0.6135	0.2161	0.6239	0.1299	0.5727	0.1809
Reference (Females)						
Male	0.2857	<.0001*	0.2652	<.0001 *	0.2634	0.0002 *
Reference (>55 years)						
<25 years	0.6095	0.0248 *	0.5461	0.0153*	0.542	0.0621
25-55 years	0.4286	0.0179 *	0.3787	0.0414 *	0.3857	0.044*
Reference (Higher level)						
Lower level	0.7406	0.7947	0.7153	0.7241	0.709	0.9387
Middle level	0.589	0.0776	0.5347	0.0673	0.5236	0.1786
Reference (Not currently married)						
Married	0.432	0.1171	0.3934	0.2096	0.3935	0.0624
Reference (Tertiary level)						
Primary level	0.6112	0.1699	0.5584	0.6597	0.56	0.8691
Secondary level	0.4557	0.2538	0.4155	0.4405	0.4102	0.5094
Higher secondary level	0.3817	0.1041	0.3455	0.4255	0.4398	0.1779
Middle level*Married	0.6862	0.0215 *	0.6282	0.0143*	0.6923	0.0276 *

(*)→ significant at 5% level

Table 6.9 also shows the summary of results which were found when the original data set, Multiple Imputed data set and the data set imputed using Inverse Probability Weighting were fitted using the GLIMMIX procedure. Looking at the generalized Chi-square values, we find that Multiple Imputation (MI) has produced a smaller value of 316.82 and Inverse Probability Weighting has a value of 319.26. Looking at these values we can see that MI produced smaller value when compared IPW and this proves that MI more consistent than IPW. The standard errors also suggest that MI is more consistent than IPW. Taking the lower level of socio-economic status for example, we see that MI produced a higher standard error estimate than IPW. This standard error estimate is very close to that produced by the original data set showing that IPW is a very consistent method. Thus we can say that MI is a more consistent methods than IPW.

Chapter 7

7 Conclusion

Smoking continues to remain a health hazard globally. It is the highest cause of mortality thus far. The current research aims to model the complex survey data which were part of the October 1996 omnibus smoking survey in South Africa. The research sought to model the data using generalized linear models (GLMs) and generalized linear mixed models (GLMMs) using logistic regression and PROC GLIMMIX in SAS 9.3. The objective of the thesis was to identify significant variables associated with smoking status as this will enable recommendations to the South African smoking policy.

In Chapter 2 an exploratory analysis of the data was carried out and it found that there was a higher percentage of non-smokers than smokers in all the surveyed variables. Doctors and clinicians continue to emphasize the use of effective tobacco counseling and medication treatments to their patients, and health systems and government officials perpetually help assist clinicians in making such effective treatments available.

Chapter 3 centres on the application of logistic regression to the data. Logistic regression reveals that race was significant at 5 % level with respect to it's influence on smoking status. Prevention programmes need to be designed to prevent the movement of smoking habits from parents to children. Sex was also significant at 5 % level with respect to it's influence on smoking status. More counseling programmes need to implemented for males since they smoke more frequent than females. Age was significant at 5 % level

with respect to it's influence on smoking status. For most people smoking started at adolescence stage most likely because of peer pressure and other factors so prevention and treatment programmes need to be implemented for adolescents in particular. The socio economic-status was also significant at 5 % level with respect to it's influence on smoking status.

In Chapter 4 we fitted a generalized linear mixed model (GLMM) using PROC GLIMMIX. GLMM is suitable for the data since it can incorporate the random effects-clusters in our case. When PROC GLIMMIX was applied to the data, race, sex, age and socio-economic status were again significant at 5 % level with respect to their influence on smoking status. These results are similar to those we found in Chapter 3. We also found that the interaction between males and middle level socio-economic status was significant with respect to smoking status. The government must implement tobacco related educational programmes for the public, targeting mostly young boys and girls from the lower and middle level of socio-economic status. These programmes should force the locals and pubs not to sell tobacco to young boys as this will stop the usage of tobacco smoking among teenage boys and will reduce the rate of crime in our societies.

In Chapter 5 we fitted multiple correspondence analysis to the data to check for the association and correlation between the variables. We found that there was a very strong correlation between race and socio-economic status, Yu and Zhang (2012) also found a strong correlation between race and socio-economic status. This suggests that the intervention programes should be for everyone. Another strong correlation was between race and education. It is not easy to say educated White or Black people smoke more frequently

than not educated White or Black people or *vice versa*. So, educated or not educated people of all races should be targeted and educated about the dangers of smoking. There was also a strong correlation between education and socio-economic status. Age was found to be closely correlated with marital status. It is important that we never assume that a correlation means that a change in one variable causes a change in another variable but use correlation to provide general indications.

In Chapter 6 the issue of missing data was addressed using Multiple Imputation(MI)and Inverse Probability Weighting(IPW). We then fitted a GLMM to the data using PROC GLIMMIX. This chapter was aimed at comparing the results of these techniques (MI and IPW) to assess their strengths and weaknesses. When PROC GLIMMIX was fitted to the data imputed using MI we found that race and sex were significant at 5% level with respect to their influence on smoking status. Only sex was significant at 5% level with respect to it's influence on smoking status when PROC GLIMMIX was fitted to the data imputed using IPW. These results show that MI fitted the data well when compared to IPW. We therefore recommend the use of MI for handling missing data. Our findings clearly demonstrate that Multiple Imputation is the most consistent and efficient method for handling missing data.

As is the case with most research, there are limitations. The limitations for this research were sample size, we created the missingness manually and there was a limited number of explanatory variables involved in the survey. As seen, IPW estimators can be more consistent than MI estimators, Seamen et al (2012) suggested that these estimators show the potential for

being a competitive and attractive tool for tackling missing data. It is not very easy to perform this method so there is a need for necessary software development. The South African government must implement educational programmes through advertising that will emphasize the dangers of smoking tobacco and its related diseases. South Africa's smoking policy should make regular updates of their tobacco control strategies, their plans and their programmes. We also recommend that South Africa should ensure effective enforcement of bans on direct advertising of tobacco. The South African government should also raise tobacco taxes since higher tobacco taxes are the only effective way to encourage people to reduce the use of tobacco and to quit smoking and this will lead to a decline in tobacco related diseases.

This paper has evaluated two modeling techniques, viz GLMMs and GLMs, that handle binary data. More research is needed to evaluate other modeling techniques for binary data like the Bahadur model. This model was first introduced by Bahadur in 1961. The model was formed for modeling binary data and can work well with a clustered data set. More research is also needed for testing other missing data techniques like the pattern mixture models (PMM). These models are used when the data is not missing at random (MNAR) like in a repeated measures data where the model is used to incorporate the non ignorable missing values.

Appendices

Coding for the data

proc glimmix data=mabunganeRAW;

class race sex age social marital educat secluster;

model smoke (descending) = race sex age social marital educat / dist=binary

solution

oddsratio ;

random secluster;

run;

proc nlmixed data=mabunganeRAW qpoints=15;

parms beta0=0.0112 beta1=0.06901 beta2=-0.6790

beta3=-0.1462 beta4=-0.3709 beta5=-0.02982 beta6=-0.03361;

eta = u + beta0 + beta1*race + beta2*sex + beta3*age + beta4*social +

beta5*marital + beta6*educat; expsmoke = exp(smoke);

mu = exp(eta) / (1 + exp (eta));

model smoke binary (mu);

random u normal(0, s2u) subject=id;

predict smoke out=smoke;

run;

References

[1] Abdi, H., and Valentin, D. (2007). Multiple correspondence analysis. *Encyclopedia of Measurement and Statistics*, 651-657. Thousand Oaks (CA): Sage.

[2] Agresti, A. (1996). *An introduction to categorical data analysis*. John Wiley and Sons. Inc., New York.

[3] Allison, P. D. (2002). Missing data: Quantitative applications in the social sciences. British Journal of Mathematical and Statistical Psychology, 55(1), 193-196.

[4] Andrew, B., and Selamat, A. (2012). Systematic Literature Review of Missing Data Imputation Techniques for Effort Prediction. *International Proceedings of Computer Science & Information Technology*, 45.

[5] Ayo-Yusuf, O.A. Control of Traditional Tobacco Products? Usage in South Africa: The Outstanding Challenge. (Personal communication) 2004.

[6] Bahadur, R.R. (1961). A Representation of the Joint Distribution of Responses to n Dichotomous Items. In studies in Item Analysis and Prediction 158-176 H. Solomon (ed.). Stanford University Press

[7] Bartlett,J. and Carpenter, J. (2012). '*Missing Data Concepts*'. Accessed 15 April 2013 from http://www.bristol.ac.uk/cmm/learning/module-samples/14-concepts-example.pdf

[8] Becker, K. M., & Whyte, J. J. (2007). Clinical evaluation of medical devices: principles and case studies. Springer.

[9] Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, **24(3)**, 127-135.

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

[11] Carpenter, J. R., Kenward, M. G., and Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **169(3)**, 571-584.

[12] Dalton, M. A., Sargent, J. D., Beach, M. L., Titus-Ernstoff, L., Gibson, J. J., Ahrens, M. B., ... & Heatherton, T. F. (2003). Effect of viewing smoking in movies on adolescent smoking initiation: a cohort study. The Lancet, 362(9380), 281-285.

[13] Dobson , A. J., and Barnett, A. C. (2008). An Introduction to Generalized Linear Models (third edition) Chapman & Hall/CRC;

[14] Doherty, S. P., Grabowski, J., Hoffman, C., Ng, S. P., and Zelikoff, J. T. (2009). Early life insult from cigarette smoke may be predictive of chronic diseases later in life. *Biomarkers*, **14**, 97-101.

[15] Du, Y. (2003). Multiple Correspondence Analysis in Marketing Research.

[16] Edwards, R. (2004). The problem of tobacco smoking. Bmj, 328(7433), 217-219.

[17] Eliason, S.E. (1993). *Maximum Likelihood Estimation: Logic and Practice*, Issue 96, SAGE, USA.

[18] Esson, K. M., and Leeder, S. R. (2004). *The Millennium Development Goals and tobacco control: an opportunity for global partnership*. World Health Organization.

[19] Fiore, M. (2008). *Treating tobacco use and dependence: 2008 update: Clinical practice guideline*. DIANE Publishing, USA.

[20] Flom, P. L., McMahon, J. M., & Pouget, E. R. (2007). Using PROC NLMIXED and PROC GLMMIX to analyze dyadic data with a dichotomous dependent variable. In SAS Global Forum.

[21] Gibbs, P. (2008). *An Introduction to Generalized Linear Mixed Models Using SAS* PROC GLIMMIX. Accessed 26 April 2013 from http://collaboratory.ucr.edu/files/glimmix.pdf.

[22] Gibbons, R. D., and Hedeker, D. (1997). Random effects probit and logistic regression models for three-level data. *Biometrics*, 1527-1537.

[23] Gill, J. (2001). *Generalized Linear Models*: Thousand Oaks, CA. SAGE Publications.

[24] Giovanini, J. (2008). *Generalized Linear Mixed Models with Censored Covariates*. ProQuest Publications, USA.

[25] Graham, H., Francis, B., Inskip, H. M., and Harman, J. (2006). Socioeconomic lifecourse influences on women's smoking status in early adulthood. *Journal of Epidemiology and Community Health*, **60(3)**, 228-233.

[26] Greenacre, M., (2007). *Correspondence Analysis in Practice*, Second Edition. London: Chapman and Hall/CRC.

[27] Greenacre, M., and Pardo, R. (2005). *Multiple correspondence analysis of a subset of response categories.*

[28] Groenewald, P., Vos, T., Norman, R., Laubscher, R., Van Walbeek, C., Saloojee, Y., Sitas, F., Bradshaw, D., and The South African Comparative Risk Assessment Collaborating Group,. (2007). Estimating the burden of disease attributable to smoking in South Africa in 2000: original article. *South African Medical Journal*, **97(8)**, 674-681.

[29] Hammoud, A. O., Bujold, E., Sorokin, Y., Schild, C., Krapp, M., and Baumann, P. (2005). Smoking in pregnancy revisited: findings from a large population-based study. *American Journal of obstetrics and gynecology*, **192(6)**, 1856-1862.

[30] Hamer, R., and Simpson, P. (2009). Last observation carried forward versus mixed models in the analysis of psychiatric clinical trials. *American Journal of Psychiatry*, **166(6)**, 639-641.

[31] Haslam, C., and Lawrence, W. (2004). Health-related behavior and beliefs of pregnant smokers. *Health Psychology*, **23(5)**, 486.

[32] Hernán, M. A., and Robins, J. M. (2006). Estimating causal effects from epidemiological data. *Journal of epidemiology and community health*, **60(7)**, 578-586.

[33] Hosmer, D.W. and Lemeshow, S. (2000). *Applied logistic regression.* 2nd edition, John Wiley & Sons. New York.

[34] Howell, D. C. (2007). The treatment of missing data. The Sage handbook of social science methodology, 208-224.

[35] Jaakkola, J. J., and Gissler, M. (2004). Maternal smoking in pregnancy, fetal development, and childhood asthma. *American Journal of Public Health*, **94(1)**, 136.

[36] Laugesen, M., Scragg, R., Wellman, R. J., & DiFranza, J. R. (2007). R-rated film viewing and adolescent smoking. Preventive medicine, 45(6), 454-459.

[37] Le Roux, B. and Rouanet, H.. (2010). *Multiple Correspondence Analysis*, SAGE Publications. USA.

[38] Lee, Y., and Nelder, J. A. (2004). Conditional and marginal models: another view. *Statistical Science*, **19(2)**, 219-238.

[39] Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data*: Wiley, New York

[40] Manning, C. (2007). Generalized Linear Mixed Models (illustrated with R on Bresnan et al.?s datives data). Unpublished handout. http://nlp. stanford. edu/ manning/courses/ling289/GLMM. pdf.

[41] Mashita, R. J., Themane, M. J., Monyeki, K. D., and Kemper, H. C. (2011). Current smoking behaviour among rural South African children: Ellisras Longitudinal Study. *BMC pediatrics*, **11(1)**, 58.

[42] McCaffrey, D. F., Lockwood, J. R., and Setodji, C. M. (2013). Inverse probability weighting with error-prone covariates. *Biometrika*, **100(3)**, 671-680.

[43] McCullagh, P., & Nelder, J. A. (1989). Generalized linear models.

[44] Mechanic, L. E., Millikan, R. C., Player, J., de Cotret, A. R., Winkel, S., Worley, K., Heard,K., Chiu-Kit T., and Keku, T. (2006). Polymorphisms in nucleotide excision repair genes, smoking and breast cancer in African Americans and whites: a population-based case-control study. *Carcinogenesis*, **27(7)**, 1377-1385.

[45] Millar, R.B., (2011), *Maximum Likelihood Estimation and Inference: With Examples in R, SAS and ADMB*. John Wiley & Sons. New York, United Kingdom.

[46] Molnar, F. J., Hutton, B., and Fergusson, D. (2008). Does analysis using "last observation carried forward" introduce bias in dementia research?. *Canadian Medical Association Journal*, **179(8)**, 751-753.

[47] Morrison, R. A. (2011). Parental, Peer, and Tobacco Marketing Influences on Adolescent Smoking in South Africa.*Public Health Theses*.

[48] Myers, W. R. (2000). Handling missing data in clinical trials: an overview. *Drug Information Journal*, **34(2)**, 525-533.

[49] Narula, S. K. (2011). *South African College Students' Attitudes Regarding Smoke-Free Policies in Public Spaces, Private Spaces, and on Campus* (Doctoral dissertation, Emory University).

[50] Nelder, J. A., & Baker, R. J. (1972). Generalized linear models. John Wiley & Sons, Inc..

[51] O'Connell, A.A. (2006). *Logistic Regression Models for Ordinal Response Variables*. Thousand Oaks, CA: SAGE.

[52] Paul, D. A. (2001). *Missing Data, Quantitative Application in the Social Sciences*, SAGE publications, USA.

[53] Panagiotakos, D. B., and Pitsavos, C. (2004). Interpretation of epidemiological data using multiple correspondence analysis and log-linear models. *Journal of Data Science*, **2(1)**, 75-86.

[54] Prabhu, N., Smith, N., Campbell, D., Craig, L. C., Seaton, A., Helms, P. J., Graham, D. and Turner, S. W. (2010). First trimester maternal tobacco smoking habits and fetal growth. *Thorax*, **65(3)**, 235-240.

[55] QMIN(2006-02-08) Transformations-1.1 accessed 18 May 2013 from http://psych.colorado.edu/ carey/Courses/PSYC5741/ handouts/Transformations.pdf.

[56] Rodríguez, G., Generalized Linear Model Theory Revised November 2001

[57] Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, **91(434)**, 473-489.

[58] Rubin, D. B., (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons, Inc.

[59] Rubin, L. H., Witkiewitz, K., Andre, J. S., and Reilly, S. (2007). Methods for handling missing data in the behavioral neurosciences: don?t throw the baby rat out with the bath water. *Journal of Undergraduate Neuroscience Education*, **5(2)**, A71.

[60] Satty, A. H. (2012). Comparative approaches to handling missing data, with particular focus on multiple imputation for both cross-sectional and longitudinal models

[61] Sahai, H. and Ageel,M.I.(2000). *The Analysis of Variance: Fixed, Random, and Mixed Models*, Birkhäuser Boston.

[62] Saloojee Y. (2006). Chronic Disease of lifestyle in South Africa since 1995-2005. Accessed March 17 2013 from http://www.mrc.ac.za/chronic/cdlchapter5.pdf.

[63] Seaman, S. R., White, I. R., Copas, A. J., and Li, L. (2012). Combining Multiple Imputation and Inverse Probability Weighting. *Biometrics*, **68(1)**, 129-137.

[64] Seaman, S. R., and White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. Statistical Methods in Medical Research, 22(3), 278-295.

[65] Sitas, F., Urban, M., Bradshaw, D., Kielkowski, D., Bah, S., and Peto, R. (2004). Tobacco attributable deaths in South Africa. *Tobacco Control*, **13(4)**, 396-399.

[66] Sterne,J.A.C., White,I.R., Carlin,J.B., Spratt, Kenward,M.G., Wood,A.M., Carpenter,J.R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ: *British Medical Journal*, 338.

[67] Streiner, D., and Geddes, J. (2001). Intention to treat analysis in clinical trials when there are missing data. *Evidence Based Mental Health*, **4(3)**, 70-71.

[68] Stroup, W.W. (2012). *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*, CRC Press. USA.

[69] Steyn, K., Fourie, J., & Temple, N. (2006). Chronic diseases of lifestyle in South Africa: 1995?2005. Cape Town: South African Medical Research Council, 33-47.

[70] Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, **19(3)**, 227.

[71] Tabachnick, B.G. and Fidell, L.S. (1996). *Using Multivariate Statistics*, 3rd edition. Harper Collins.

[72] Vansteelandt, S., Carpenter, J., and Kenward, M. G. (2010). Analysis of incomplete data using inverse probability weighting and doubly robust estimators. Methodology: *European Journal of Research Methods for the Behavioral and Social Sciences*, **6(1)**, 37-48.

[73] Vittinghoff, E., Shiboski, S., McCulloch, C. E. (2005). Regression methods in biostatistics (pp. 98-109). New York:: Springer.

[74] Vaseghi, S.V. (2000). *Advanced Digital Signal Processing and Noise Reduction*, Second Edition. John Wiley & Sons, Chichester, U.K.

[75] Van Walbeek, C. (2002). Recent trends in smoking prevalence in South Africa: some evidence from AMPS data. *South African Medical Journal*, **92(6**), 468-472.

[76] Walls, T.A. (2005). *Models for Intensive Longitudinal Data*, Oxford University Press.

[77] Wayman, J. C. (2003). Multiple imputation for missing data: What is it and how can I use it. In *Annual Meeting of the American Educational Research Association, Chicago, IL* (pp. 2-16).

[78] Xu, H. (2009). LOCF Method and Application in Clinical Data Analysis. Accessed 5 August 2013 from http://www.nesug.org/Proceedings/nesug09/po/po12.pdf.

[79] Yu, T. T., and Zhang, M. M. (2012). Income Inequality Among Races: *A Statistical Analysis. Southwestern Economic Proceedings*, **32(3)**, 31-40.