MODELING THE SMOKING STATUS OF KENYA'S MALES IN THE PRESENCE OF MISSING DATA

A thesis submitted to the university of KwaZulu-Natal in fulfillment of the academic requirements for the degree of Master of Statistics in the Faculty of Science and Agriculture

By

Mrs Odette UMUGIRANEZA
School of Mathematical Sciences, Statistics and Computer Science
University of KwaZulu-Natal, Pietermaritzburg

2014

Contents

List of Figures

Abstract

The current research, modeling smoking status in Kenya's males in the presence of missing data has three objectives: The first objective of this study is to identify factors, associated with smoking which will lead to recommendations to the smoking policy in Kenya. The second objective is to apply the appropriate statistical models to model smoking status of Kenya males that incorporates missing data; Logistic regression as well as the generalized linear mixed model are used to model the smoking status. The third objective leads to comparison of the various statistical methods that handle monotone missing data and by their strengths and weaknesses. The following statistical methods for handling missing data are investigated. These are Last Observation Carried Forward (LOCF) and Multiple Imputation (MI) in order to handle the missingness. The missing data will be created by deleting randomly 20% and 30% of the data. The data used is KDHS 2008-2009, the response variable is the smoking status (smoker and non smoker) and the explanatory variables are region, marital status, religion, education, age group of the respondent, wealth index, size of household and access to mass media.

Declaration

The work described in this thesis was carried out in the School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, under the supervision of Dr Shaun Ramroop and the co-supervisor Professor Henry Mwambi. I, Odette Umugiraneza, declare that this thesis is my own work. It has not been submitted in any form for any degree or diploma to any other University and where use has been made of the work of others, it is acknowledged.

2014.

Student: _____ _____

Mrs Odette Umugiraneza Date

Supervisor: _____ _____

Dr.S.Ramroop Date

Co-Supervisor: _____ _____

Prof.H.Mwambi Date

Dedication

To my God almighty for his mercy that endures forever. To my lovely husband Faustin Habyarimana and our children B. Neree Blaise and B. S. Neri Bruce, for their patience and love.

Acknowledgment

I am grateful to my supervisor Dr Shaun Ramroop and my co-supervisor Professor Henry Mwambi for their guidance, encouragement, advice, assistance and social support in all steps of this research. My special thanks to the School of Mathematics, Statistics and Computer Science for their academic support and moral support during the period of my study. I am also indebted to a number of people who directly made contributions to this work. In particular, thanks are due to Faustin Habyarimana, Oscar Ngesa, Dawit, Artz Luwanda and all my officemates. My mother, brothers and sister have always been a pillar of strength, this is an opportunity to thank them for their sacrifices in ensuring my earlier studies.

Chapter 1

INTRODUCTION

1.1 Overview of smoking status

Smoking is considered as the effect of inhaling or exhaling the smoke of any tobacco product and includes the holding of, or control over, any device containing an ignited tobacco product (Tobacco Control Act, 2007). According to World Health Organization (WHO, 2011), tobacco use is defined as possibly gender-linked behavior with the greatest public health significance. Globally, being born male has been, and in many parts of the world continues to be, the greatest predictor of tobacco use. Male smoking prevalence and consumption of cigarettes greatly exceeds the figures for females in most regions of different countries .

According to the Center for Disease Control and Prevention (CDCP, 2008-2010), smoking status terms are defined as: never smoker (adults who never smoked a cigarette or smoked less than 100 cigarettes), former smoker (adults who smoked at least 100 cigarettes in their lifetime, but currently do not smoke), non smoker (adults who currently do not smoke cigarettes) and currently smoker (adults who smoke cigarettes every day or some days). Apart

from that, there exist other definitions of smoking status used in many studies. These include light smoker, intermittent smoker, and social smoker. Smoking is exponentially increasing worldwide. Estimates by WHO (2011) reveal an estimated 1 billion smokers in the world, and 6 trillion cigarettes are consumed each year. It has been found that the highest cigarette consumers are in the Western Pacific region (46%), followed by the European region (29%), African Region (8%), and the highest prevalence is from the upper-middle-income countries. Male tobacco use is concentrated among people with low incomes and less education in nearly all countries, where about 50% of men in developing countries smoke, compared with 35% in developed countries.

1.2 Factors influencing the smoking

Studies consider stress and associated distress or depression to be important factors for smoking initiation (Tyas and Pederson, 1998). Karen and Mary (1999), in their research on factors influencing adolescents to smoke, reported that adolescent smokers have been found to be more likely than adolescent nonsmokers to believe that cigarette smoking helps people to relax, reduce stress, and relieve boredom. Also, a study conducted by Bush et al.(2003), on the factors associated with smoking in Bangladesh and Pakistan in adults, found that gender, age, religion and tradition significantly influenced smoking attitudes. Other studies found that smoking status is influenced by several factors such as family, peer-group influence, exposure to alcohol and cigarettes, mass media content as well as the level of self-esteem; these are among the major risk factors contributing to cigarettes use (Jeanne et al., 2005). Australia Government Department of Ageing (2005) highlighted that

the factors influencing smoking are grouped in four categories:

- Socio-demographic factors (age, gender, parental, ethnicity, location family structure, personal income, and working hours)
- Environmental factors (parental smoking, siblings smoking, parental attitudes, family environment, attachment to family and friends, peer smoking, peers attitudes, availability of tobacco, and school environment).
- Behavioral factors (academic achievement, risk taking, rebeliousness and delinquent behaviors, participation in sports).
- Personal factors (genetic factors, first experience of smoking, attitudes to smoking, firm decision to smoke, knowledge of health effects of smoking, stress, coping, depression, religion).

In their study, Oteyo and Kariuki (2009) also stated that there is a correlation which exists between social influence that arise from parents, sibling, peer pressure (having close friends who smoke) and advertisement clips. Their study showed that students with low self-esteem as compared to those with high self-esteem, were more likely to use alcohol and cigarettes. CDCP (2008-2010) also conducted a study concerning smoking status by age, race, region, education, poverty status, and gender. United State Department of Health and Human Services (2012) also described the social, physical, environment and movies (advertising and marketing) as factors influencing peoples smoking behavior. Recent studies such as Emmanuel et al.(2007), focused on the prevalence and determinants of adolescent tobacco smoking in Addis Ababa Ethiopia, using bivariate and multivariate logistic regression to determine the associations between current smoking status and other variables (age, gender, parental smoking, best friends smokers, and perception that smoking

is harmful). Fred and Jade (2008) also studied the change in youth smoking from 1976 to 2008, using a time-series analysis. They reported that, cigarettes low prices were the factors in cigarettes use. Lefondre′ et al.(2002) made a research on modeling smoking history by demographic characteristics (age, ethnic, occupation, annual income), using logistic regression.
In Kenya, the studies carried out on smoking such as Kwamanga et al.(2003), on the risk factors of smoking among secondary school students in Nairobi (Kenya), reported that peer group, mass media, type of school, and age between 15-18 years old are the factors influencing smoking among students. Another research was conducted by Ogwell et al. (2003), and studied the use of tobacco in Kenya among primary school students aged from 12 to 17 years old, and place of residence (urban and non urban) using a generalized linear model and ANOVA analysis in the mean scores adjusted for age, gender, place of residence and parental education. They reported that peers' influences are among the strongest correlates of lifetime smoking prevalence.

1.3 Consequences of smoking

According to WHO (2009), the estimations made showed that tobacco use kills more that 5 million people annually in the world, and that most of them are found in low and middle-income countries. Secondary smoking (smoke transmitted after the cigarettes or tobacco use) is estimated to cause about 600 000 premature deaths per year worldwide, with 31% of them occurring among children and 64% among women. Secondary smoking occurs in all public places as long as smoking is permitted. It was estimated that 40% to 43% of people and children aged from 8 to 15 years old are exposed to secondary tobacco smoke at home.

Action of Smoking and Health (2007) in its research report, also indicated that tobacco is considered as the cause of creating poverty, where the poorest tend to smoke the most and 84% of smokers live in developing countries. Tobacco users are also exposed to the highest sickness burden such as dying prematurely due to cancer, heart attacks, respiratory diseases or other tobacco-related diseases, which contribute to family poverty. It has been shown that if there is no cessation of smoking, the calculated estimates reveal that by 2030 tobacco will be expected to be the single biggest cause of death worldwide, with an estimated 10 million people in the world who die of tobacco related causes. Approximately, 3 million of these will occur in the developed world and 7 million in developing countries. According to United State Department of Health and Human Services (2006), smoking damages human's health in many ways: smoking harms the immune system and increases the risk of infections and fractures, it causes dental diseases, sexual problems, eye diseases, as well as peptic ulcers. It has been found that if people smoke, their cells will not get the amount of oxygen needed to work properly. In addition, if a person is a smoker, his/her illnesses tend to stay longer and the individual is more likely to be absent from work.

In Kenya, WHO (2012) found that the consequences of smoking as well as noncommunicable diseases (NCDS) for which tobacco is a risk, currently account for more than 55% of the mortality in the country, and 50% of the public-hospital admissions. The environmental impact of tobacco-growing is also a concern, because wood from natural forests is being burned to cure tobacco leaves in order to manufacture cigarettes.

1.4 Prevention of smoking

According to the Health Canada (2009), a policy has been put in place where it has advised people to communicate with neighbors and the building owners to advocate for a smoke-free policy. These policies can govern a variety of spaces, including common areas, outdoor child playing areas, apartments, and blocks or floors of units. Phase-in policies where units occupied by smokers are converted to smoke-free areas when they leave are also an option. It is also noted that parents are advised to talk to their children about the consequences of smoking. WHO (2011) suggests the use of anti-tobacco mass media campaigns, such as TV and radio, because they reach a larger population quickly and efficiently. Increasing tobacco prices through higher taxes is the most effective intervention to reduce tobacco use and encourages smokers to quit. A comprehensive ban on all tobacco advertising, promotion and sponsorship could decrease tobacco consumption by about 7%, independent of other tobacco control interventions, with some countries experiencing a decline in consumption of smoking of up to 16%.

The Council of United State Government Health State Initiative (2007), also recommends using tobacco sales restrictions, tobacco promotion cessation, elimination exposure to secondary smoke and impact of tobacco disparities to prevent youth and young adult from beginning to smoke.

1.5 Smoking policy in Kenya

Kenya signed and ratified the World Health Organization Policy Framework Convention on Tobacco Control in June 2004, and domesticated the same through the Tobacco Control Act in 2007, which was enacted into law in October 2007, and came into force in July 2008. There are currently on-going

efforts to develop regulations to support the implementation and enforcement of the law. In addition, the country has developed a National Tobacco Control Action Plan (2010-2015), in which the priority areas of tobacco control for the country are identified. Tobacco Control in Kenya has mostly been done by the Ministry of Public Health and Sanitation (MOPHS), with minimal support from other relevant government departments and agencies. However, the MOPHS has partnered with stakeholders such as civil society to implement Tobacco Control programs in the country. Kenya is a tobacco-growing country and is also a regional hub for manufacturing tobacco products, but it has been involved in curbing the tobacco epidemic since 1992. It is also noted that the tobacco industry has a strong presence in Kenya, and this came out clearly in several interviews (WHO, 2012).

The Tobacco Control Act (2007) is the principal law governing tobacco control in Kenya. This comprehensive law defines key terms and covers topics including, but not limited to, restrictions on public smoking; tobacco advertising, promotion and sponsorship; and packaging and labeling of tobacco products. Other topics addressed by the law include: public education and information campaigns; sales to minors; and enforcement of the law. It grants powers, including implementation and enforcement authority, to individuals appointed under the Public Health Act. The Traffic Act provides a definition of public service vehicle, incorporated by the Tobacco Control Act with regards to smoke free provisions.

Kenyan tobacco control is efficient because of good policy and institutional frameworks as well as positive relationships between government and civil society. However, lack of effective enforcement of existing laws continues to present a major set of challenges; lack of coordination of efforts among activists seeking policy change, a weak capacity both within government and

civil society for enforcement of the Tobacco Control Law, it manifests a strong opposition from tobacco industry to tobacco control, low public awareness of the provisions of the law and lack of formal cessation services (Jeffery, 2011). In August 2007, Kenya passed the Tobacco Control Act which prohibits smoking in almost all public places (e.g. government buildings, hospitals, factories, and cinemas). It also prohibits tobacco related advertising, promotional giveaways, and children's candies and toys resembling tobacco are strictly prohibited. Smokers can only buy packets that hold a minimum of ten cigarettes.

The objective and purpose of this Act is to provide a legal framework for the control of the production, manufacture, sale, labeling, advertising, promotion, sponsorship and use of tobacco products, including exposure to tobacco smoke, in order to protect the health of the individual in light of conclusive scientific evidence implicating tobacco production, use and exposure to tobacco smoke and tobacco products, in the incidence of debilitating illness, disease, disability and death.

The Tobacco Control Act (2007) also orders Kenyan smokers to stay specifically in designated smoking areas, they must be places ventilated in such a manner as to ensure that air from the area is directly exhausted to the outside and does not re-circulate or drift to other areas within the public facility; separated, enclosed and sealed from the floor to the roof with a door; in which non-smoking individuals do not have to enter the area; for any purpose while smoking is occurring; and cleaned or maintained only when smoking is not occurring in the area, for more details see Tobacco Control Act (2007).

1.6 Current statistics of smoking in Kenya

Kenya Tobacco Situation Analysis Consortium (2008) reported that smoking is increasing in the Kenyan population since male smoking prevalence far exceeds that for women. Consecutive Global Youth Tobacco surveys (GYTS) showed a dramatic increase in youth smoking, particularly among teen-aged girls. The use of chewed tobacco products by school children also appears to be rising. In 2001, Kenya Global Youth Tobacco Survey reported that 19% of adult males and less than 2% of women use tobacco; while in 2007, the prevalence rate for youths has gone up from 13% (10% girls and 16% boys) to 15% (14.5% girls and 14.9% boys). On access and availability, the survey reported that 36.6% usually smoke at home, 17.9% buy cigarettes in a store while 67% of those who bought in a store were not refused purchase irrespective of their age (Kimosop et al. 2012).

The Global Youth Tobacco Survey of (2001) also indicated that 13% of school children (less than 15 years old) used tobacco products, 16% of boys and 10% of girls. By 2007, however, this figure had risen to 18%, with equal smoking rates for boys and girls. The GYTS (2007), also examined the exposure to second-hand smoke. It was indicated that 27% of Kenyan school children (aged between 12 and 15 years old) live in homes where others smoke in their presence, 58% are around others who smoke in places outside their homes and 19% have one or more parents who smoke. The research carried out in University of Nairobi (Kenya) found that 12% of the student were current smokers, 6.95% had never smoked and 26% were former smokers (Komu et al., 2009)

WHO (2011) reported that 23% males smoke while just 1% of adult females smoke in Kenya. According to a former Minister for Health, Mrs Charity Ngilu, the Kenyan health sector (public and private) spends over Ksh 18

billion (US dollars 240 million) annually treating tobacco-related illnesses. This compares to tobacco-related revenue paid by the tobacco industry to the Government of about Ksh 7 billion: US dollars 93 million.

The study carried out by Kwamanga et al.(2003) showed that 32.2% of the total population of the students had ever smoked cigarettes at one time in 2003. Of this, 82.7% were males and 17.3% females. The overall rate of ever-smoking among the study population by gender was 38.7% among males and 17.7% among females.

KNBS and ICF Macro (2010) reported that 18% smoke cigarettes, 1.0% chewing tobacco, 1.5% use snuff, 1.3% use pipes and 1.1% smoke other sources of nicotine.

1.7 Missing data

Missing data causes a big problem in statistics surveys because it reduces the precision of the calculated statistics. When there is missingness, the planned objective is not achieved completely, since the information is partially observed. De Leeuw et al.(2003) states that missingness arises when the questions are not applicable to all respondents (by design), skipped questions, questionnaires partially completed as well as the blanks sometimes characterized by refusal or do not know. The methods for analyzing missing data require assumptions about the nature of the data and about reasons for the missing observations that are often not acknowledged (Pigott, 2001). These methods will be developed in chapter 5.

1.8 Problem statement

As stated in section 1.2, studies conducted on smoking in Kenya, have the following limitations on age group where they have considered age between 12 and 18 years old, and those limitations will be addressed in our current research. In addition, they do not consider factors such as marital status, region, ethnicity, size of household, religion and wealth index. The present research will be focused on modeling smoking status of Kenyan males, and the effect of the following characteristics such as age of respondent, size of household, region, marital status, religion, access to mass media, education, wealth index, and ethnicity. In addition, the research will address the issue of missing data, and find suitable ways to handle the missing data.

1.9 Methodology

In the current study, suitable statistical model such as Generalized linear model (GLM) using Logistic regression and Generalized linear mixed model (GLMM), are used for modeling of the data. The missing values are handled using Last observation carried forward (LOCF) and Multiple imputation (MI). In addition statistical softwares such as SAS and SPSS will helped to explore and analyze the data in the study.

1.10 Objectives of the study

In our research, we considered the Kenya demographic health survey 2008-2009. This study has the following main objectives:

- Identify factors associated with smoking as well as leading to recommendation to the smoking policy in Kenya.

- Identify appropriate statistical models applied to smoking status of Kenya's males that incorporate missing data.
- Compare various statistical methods that can be used to handle monotone missing data by addressing their strengths and weaknesses.

The rest of this thesis is structured as follows: Chapter 2 is focused on exploratory data, where we use cross-tabulations chi-square statistics to check which variables are significant associated with smoking status. We also explore how smoking status is distributed by region, religion, marital status, age, size of household, ethnicity and access to mass media and wealth index. Chapter 3 is focused on the literature of GLM, particulary logistic regression and its application to smoking status.
Chapter 4 is focused on GLMM's literatures as well as its application to smoking status.
Chapter 5 deals with the missing data; we created 20% and 30% missingness and applied statistical technics such as Last Observation Carried Forward (LOCF) and Multiple Imputation (MI) to handle the missingness. We modeled the full datasets containing imputed values using logistic and generalized linear mixed model for comparative purposes.

Chapter 2

EXPLORATORY DATA ANALYSIS

2.1 Background of Demographic and Health Survey (DHS)

The Demographic and Health Survey (DHS) programme has been funded by Westinghouse Health Systems in 1984, by the United State Agency for International Development (USAID) and the United Nation Population Fund (UNFPA), assisted by Great Britain and Northern Ireland, the Netherlands and Japan Government. The main objective of DHS is to collect up-to-date information on basic demographic and health indicators, including housing characteristics, fertility, childhood, mortality, contraceptive knowledge and use, maternal and child health, nutritional status of mothers and children, and sexually transmitted infections (ICF International, 2012). The project has been designed to be implemented every five years, and its action is to provide information to decision makers with the skills and resources neces-

sary to conduct high-quality demographic and health surveys, improve data collection and analysis tools and methodology as well as to improve the dissemination and utilization of data (Rutstein and Rojas, 2006). The Household questionnaire's objective is to identify women aged from 15 to 49 years old and men aged 15 to 54 years old. Women's questionnaire is supposed to collect information on education, residential history, media exposure, reproductive history, health insurance, adult and maternal mortality, domestic violence, female genital cutting. However, the Men's questionnaire collected information similar to that collected in the Women's questionnaire, but it was shorter because it did not contain questions on reproductive history, maternal and child health, nutrition, maternal mortality, and domestic violence (KNBS and ICF Macro, 2010).

In Kenya, DHS are conducted every five years. The 2008-2009 KDHS is the fifth survey its main objective was to provide information that would address the planning, programme implementation, monitoring and evaluation needs of health, family planning and HIV and AIDS programmes (Rutstein and Rojas, 2006).

2.2 Description of data

The present study will be based on modeling the smoking status of Kenyan males from KDHS 2008-2009. The sample for 2008-2009 KDHS was drawn from a master sampling household frame. KDHS had objective to produce estimates at the national level, for urban and rural area separately and at the province level. A sample of 10000 household from 400 clusters was drawn from the whole country, with 267 clusters in rural area and 133 clusters in urban areas. 2008-2009 KDHS used three questionnaires such as a house-

hold questionnaire, a questionnaire for individual women aged from 15 to 49 and a questionnaire for individual men aged from 15 to 54. Data collection was carried out from November 2008 and ended up in February 2009. The questionnaires were administered throughout the country among the selected household and selected men and women. The survey reported a total 8767 women eligible whose only 8444 were interviewed and 3910 of men eligible, but only 3465 of men were successfully interviewed. The principle reason of non- response among both eligible men and women was the failure to find them at home after repeated visits to the household. The smoking status of males is determined by people who smoke cigarettes, pipes, tobacco, snuff and other sources of nicotine. In order to conduct a suitable analysis, the variables smoke cigarettes, pipes, tobacco, snuff, and other source of nicotine were pooled into one new outcome variable: smoking status which is a binary outcome. Table 2.1 represents that the percentage of male smokers is 20.9% and 78.8% of non smokers. As stated above, the variables smokes cigarettes, pipes, tobacco, snuff, and other type of tobacco define the dependent variable called **smoking status** and were coded as 1 (smoker) and 0 (non smoker), as shown in Table 2.2.

Table 2.1: Table of smoking status

Smoking status	Values	Percentage
smoker	724	20.9%
non smoker	2732	78.8%

Table 2.2 describes the different variables and coding which will be used in the study.

Table 2.2: Variables, level and coding

Variables	Level and coding
Identification	1,...,3465
Smoking status	1:smoker, 0:non smoker
Age	1:15-19, 2:20-24, 3:25-29, 4:30-34, 5:35-39, 6:40-44, 7:45-49, 8:50-54
Size of household	1:1 person, 2:2persons, 3:3persons, 4:4persons, 5:5persons, 6:6persons and above
Region	1:Nairobi, 2:Cent., 3:Coast, 4:East., 5:Nyanza, 6:L. Valley, 7:West., 8:North.
Educational	0:Non Education, 1: primary incomplete,2:primary complete,4:secondary+
Wealth Index	1:Poorer or poorest, 2:Middle or richer, 3:Richest
Marital status	0:Never married, 1:Married, 2:L.together, 3:divorced
Religion	1:Roman Catholic, 2:Protestant, 3:Muslim, 4:Non religion or other
Ethic group	1:Embu, 2:Kelenjin, 3:Kamba, 4:Kikuyu, 5:Kisii, 6:others
Access to mass media	0:Not at all, 1:Less than once a week, 2:At least once a week, 3:Almost every day

2.2.1 Cross-tabulations

The main reason for cross tabulation, is to assess the relationship between two or more categorical variables and to test the association of these variables using the chi-square test of independence. Table 2.3 indicates the strength of the relationship between the smoking status and other factors using the chi-square test of independence. The test reveals a significant relationship between age, region, marital status, religion, ethnicity, size of household and access to mass media to smoking status since their p-values are all less than .05.

Table 2.3: χ^2 test of model effects

Variable	χ^2 value	DF	P-value
Age	296.687	7	.000
Region	128.270	7	.000
Type of residence	.854	1	.355
Wealth index	6.372	2	.041
Education attainment	28.332	3	.000
Current marital status	198.063	3	.000
Size of household	36.511	5	.000
Religion	68.126	3	.000
Ethnicity	84.035	5	.000
Access to mass media	15.950	3	.001

2.2.2 Distribution of smoking status

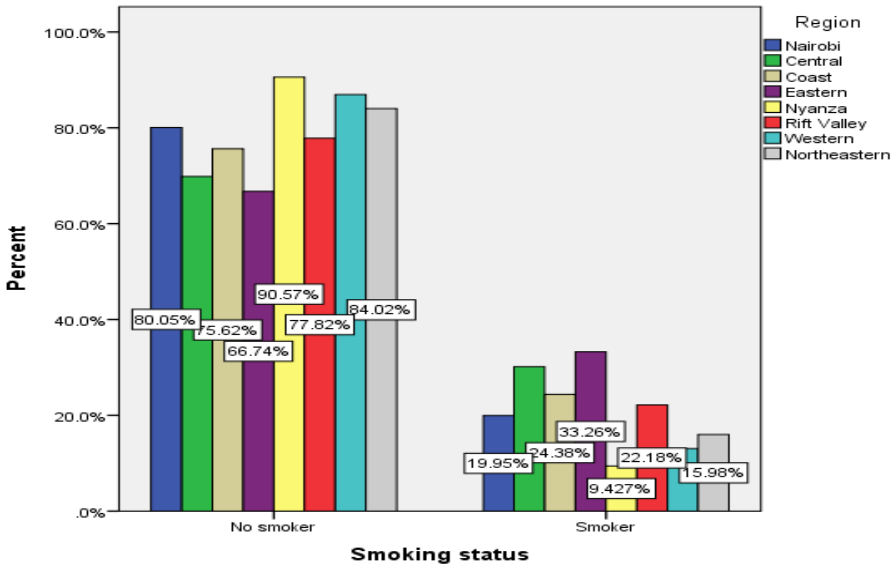


Figure 2.1: Distribution of smoking status by region

Clustered graphs help to describe graphically the data in order to visually inspect any trends in the data. Figure 2.1 displays the bar graph for smoking status and region; the graph shows the highest percentage of male smokers is located in Eastern Province (33.26%) followed by Central Province (30.15%).

Figure 2.2 displays the bar graph for smoking status and marital status; the graph shows that the highest percentage of male smokers are divorced (50.0%) followed by the married men (27.11%).

Figure 2.3 displays the bar graph for smoking and religion; the graph shows that the highest percentage of male smokers are found for men with non-religion (45.72%) followed by the Catholic men (30.25%).

Figure 2.4 displays the bar graph for smoking status and education; the

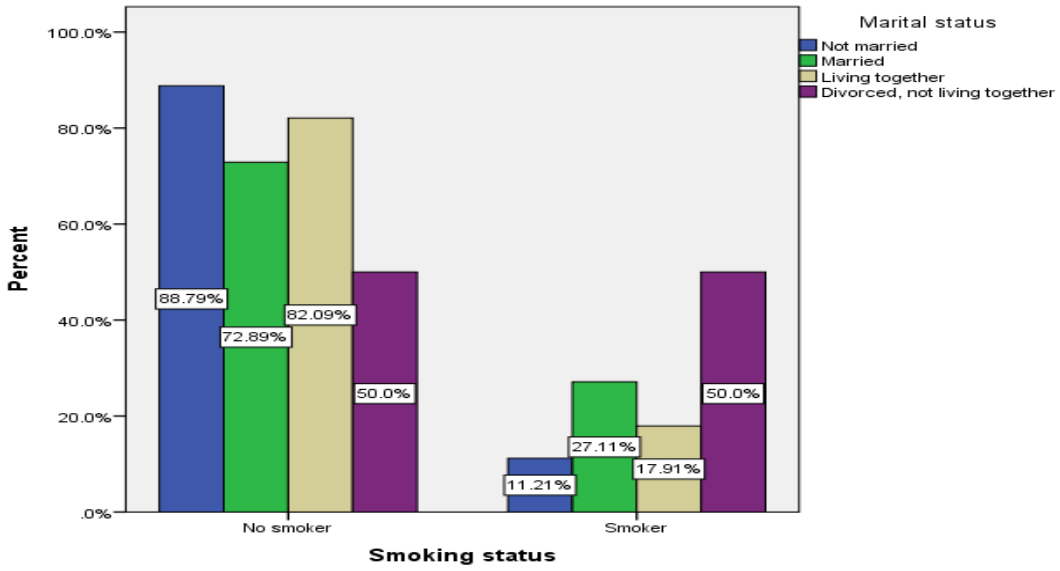


Figure 2.2: Distribution of smoking status by marital status

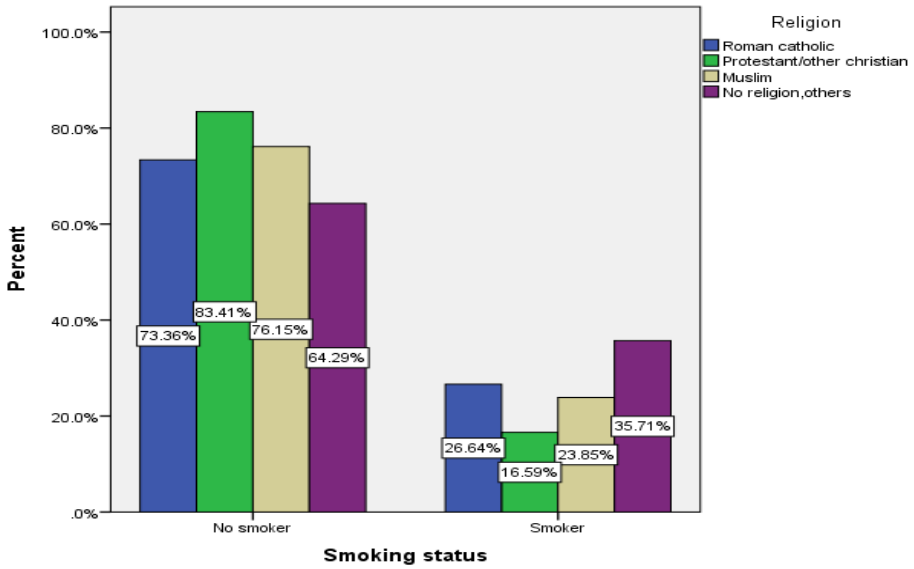


Figure 2.3: Distribution of smoking status by respondent's religion

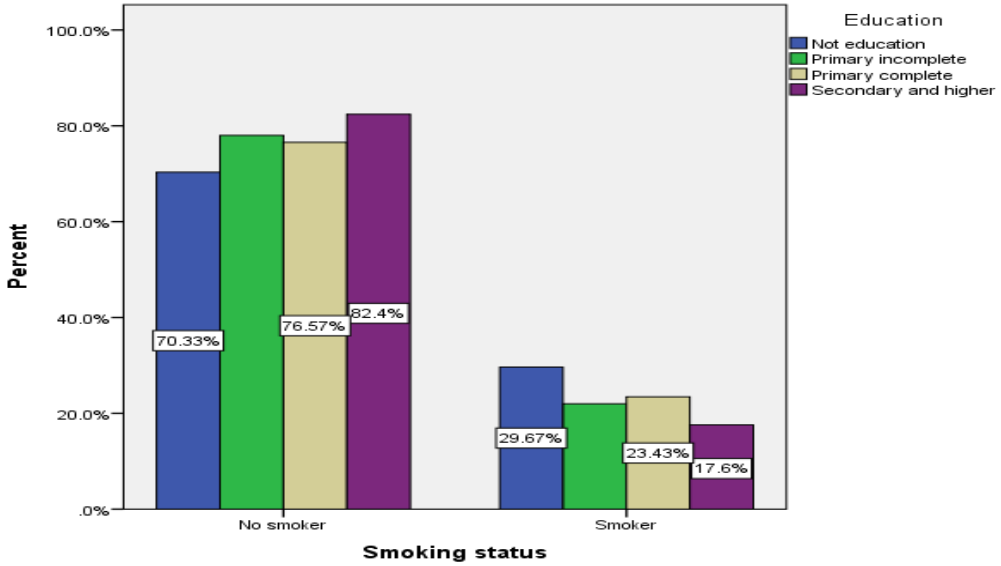


Figure 2.4: Distribution of smoking status by education

graph indicates that the highest percentage of smoking is found for men with non education (29.67 %), followed by the men with primary complete (23.43%).

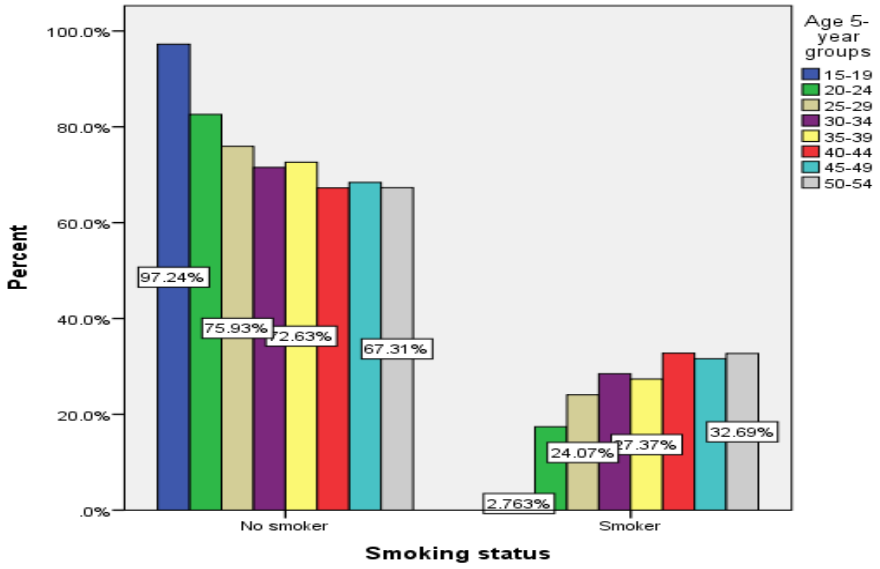


Figure 2.5: Distribution of smoking status by age of respondent

Figure 2.5 displays the bar graph for smoking status and age of respondent; the graph indicates the highest percentage of male smokers aged 40-44 (32.76%), followed by men aged 50-54 (32.69%).

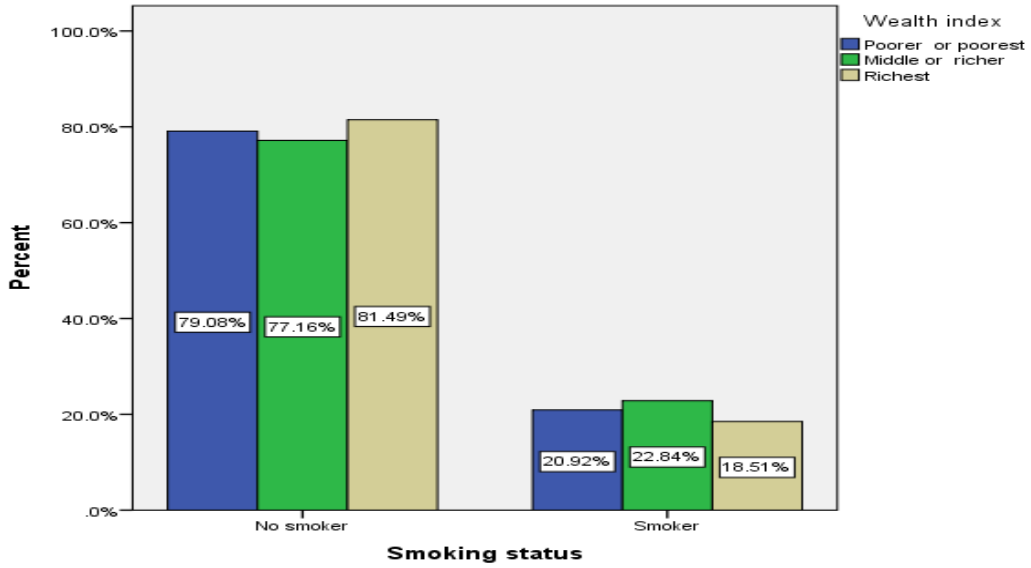


Figure 2.6: Distribution of smoking status by wealth index

Figure 2.6 displays the bar graph for smoking status and wealth index; the graph shows that the highest percentage of male smokers are middle or rich (22.84%) followed by the poor or poorest men (20.92%)

Figure 2.7 displays the bar graph for smoking status and ethnicity; the graph shows that the highest percentage of male smokers are found for men having Embu tribe (36.14%) followed by men who have Kikuyu (30.19%).

Figure 2.8 displays the bar graph for smoking status and size of household; the graph shows that the highest percentage of male smokers are found from a household of 1 person (29.89%) followed by a from a household of 5 persons (23.21%)

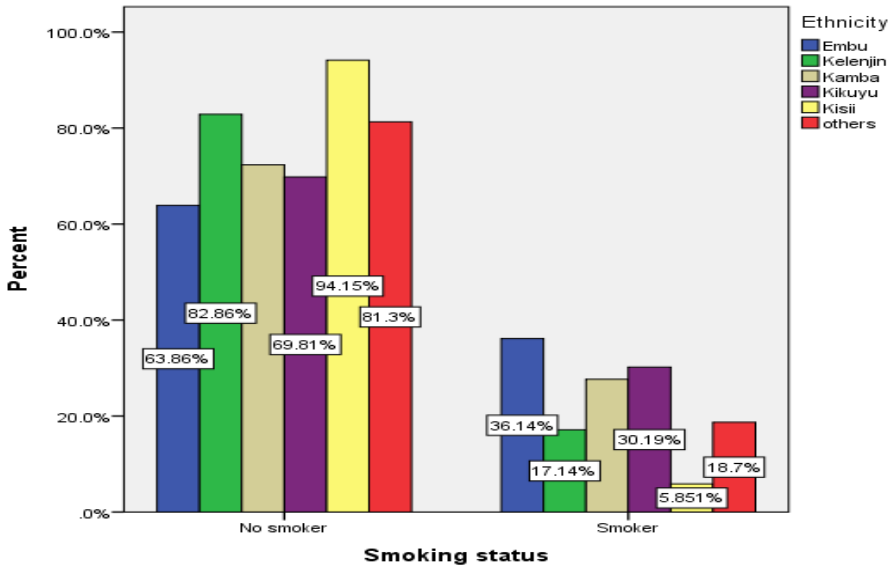


Figure 2.7: Distribution of smoking status by respondent's ethnicity

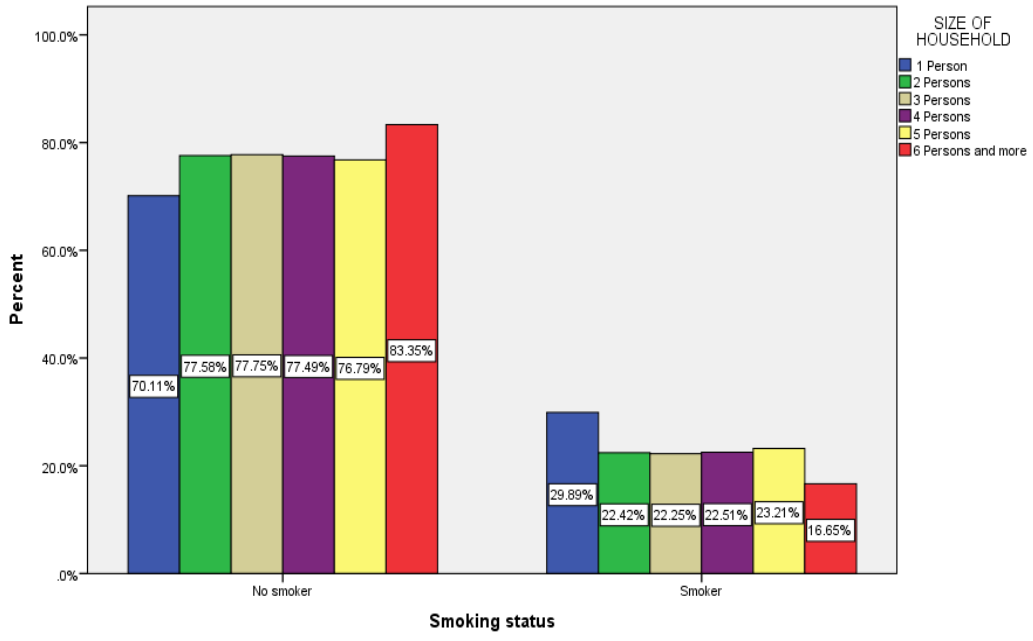


Figure 2.8: Distribution of smoking status by size of household

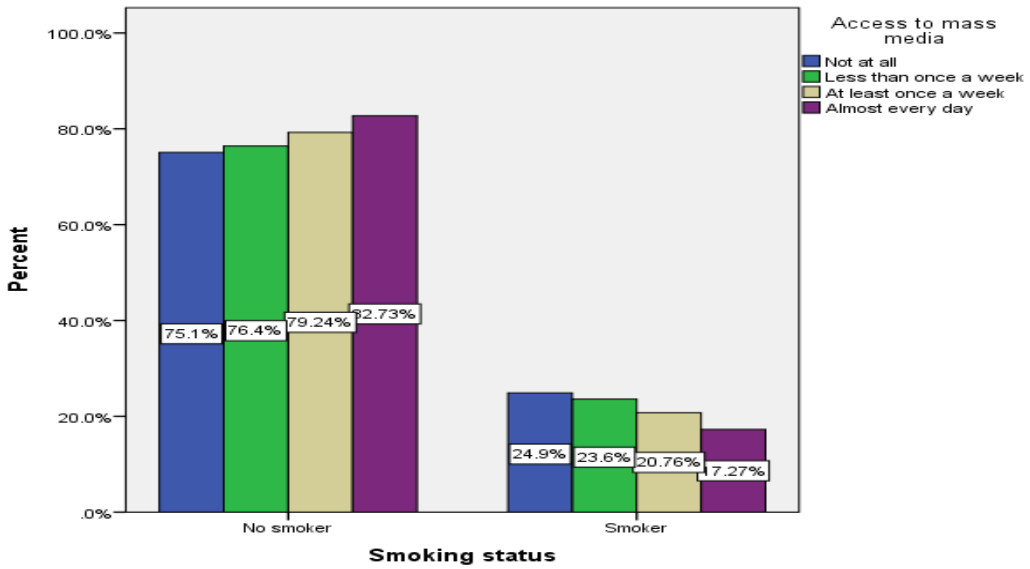


Figure 2.9: Distribution of smoking status by access to mass media

Figure 2.9 displays the bar graphs for smoking status and access to mass media; the graph indicates that the highest percentage of smoking is by men who do not have access to mass media (24.9%) followed by the men who have access less than once a week (23.6%), men who access at least once a week (20.76%) and smoking is lowest for the men who access to mass media almost every day (17.27%). Thus there is evidence of a clear trend of smoking with access to to mass media.

2.3 Summary

The present chapter was focused on the explanatory variables of smoking status and the distribution of the data. The chi-square test of independence was used to check the association between smoking status and the explanatory variables. The results shown in Table 2.3 revealed significant association between smoking status and the explanatory variables (region, marital status, religion, education, age, wealth index ethnicity, size and access to mass media), except the variable type of residence (p-value $>.05$).

Chapter 3

THE GENERALIZED LINEAR MODEL

3.1 Introduction

Nelder and Wedderburn (1972) developed the Generalized Linear model (GLM) as a generalization of classical linear model, in order to unify the seemingly different approaches in modeling data from distributions which do not necessarily follow the normal distribution. GLMs relate a linear model to a response via a link function; examples include familiar models like logistic regression, poisson regression, log-linear models and multinomial responses models for counts and some commonly used models in survival analysis (McCullagh and Nelder, 1989). GLM is applied in many domains such as astronomy, biology, medical and pharmaceutical sciences, just to mention a few.

3.2 Model structure

The GLM allows us to fit regression models for response variables that follow a general distribution belonging to the exponential family with an additional dispersion. The general exponential family includes all distributions such as continuous, discrete or mixed (Wedel and Kamakura, 2001). The probability function for each response is written as follows:

$$f(y_i;\theta_i,\phi) = exp\{\frac{y_i(\theta_i) - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\}, i = 1,2,...,n \qquad (3.1)$$

where ϕ is called the dispersion parameter and θ is a natural parameter. The function $a(\phi)$ has the form $a(\phi) = \frac{\phi}{w_i}$, and is called the dispersion parameter for a known weight w_i which usually equals 1 (Agresti, 2002). $b(\theta)$ and $c(y,\phi)$ are known functions. If the dispersion parameter is greater than 1 the model is said to be overdispersed and underdispersed when it is less than 1. Distributions such as Normal, Binomial, Poisson, Gamma, and Inverse Gaussian can be easily written in terms of the exponential family (Brown et al., 2003). The advantage of expressing diverse families of distributions in the common exponential form is that general family distributional properties can be applied to the individual cases. McCullagh and Nelder (1989) outline the three components of GLM as follows:

- The random component describes the conditional distribution of the response Y with explanatory variables, which is a member of exponential family such as normal, poisson, gamma, binomial.
- The systematic component involving the explanatory variables $x_1, x_2, ..., x_p$ used as a linear predictor
- The third component is the link function g, that links the predictor to the natural mean of the response variable Y.

The systematic component of a GLM relates a vector $(\zeta_1, ..., \zeta_N)$ to the explanatory variables through a linear predictor. Thus,

$$\zeta_i = \sum_l \beta_l x_{il}$$

where x_{il} denote the value $l(l = 1, 2, ..., p)$ for the subject i (Agresti, 2002). The link function between the distribution y_i and the linear predictor ζ_i is provided by the link function g. Thus μ_i relates to the explanatory variables by a linear predictor given by

$$\zeta_i = g(\mu_i) = \sum_l \beta_l x_{il}, i = 1, 2, ..., N$$

Thus, the generalized linear model is given by :

$$g(\mu_i) = x_i^{'}\beta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} \tag{3.2}$$

A link function describes how the mean $\mathrm{E}(Y) = \mu_i$ depends on the linear predictor $g(\mu_i) = \zeta_i$. The variance $\mathrm{Var}(Y)$, depends on the mean such that $\mathrm{Var}(Y) = \phi V(\mu_i)$ where the dispersion parameter ϕ is a constant and $V(\mu_i)$ is variance function. For normal data, the link $g(\mu_i) = \mu_i$ which is called the identity link function and the variance function $V(\mu_i) = 1$. For binomial data, $g(\mu_i) = logit(\mu_i) = log(\frac{\mu_i}{1-\mu_i})$ and the variance function
$V(\mu_i) = \mu_i(1 - \mu_i)$. The canonical link function becomes when

$$g(\mu_i) = \theta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}$$

where θ_i is the canonical parameter.

3.3 Estimation in generalized linear model

In generalized linear model, estimation proceeds by defining a measure of goodness of fit between the observed data and the fitted values generated

by the model. The parameter estimates are the values that minimize the goodness of fit criterion (McCullagh and Nelder, 1952). According to Agresti (2002), the density function or the probability distribution for the relation (3.2) is written as:

$$f(y_i;\theta_i,\phi) = exp\{\frac{y_i(\theta_i) - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\} \tag{3.3}$$

From equation (3.3), the log-likelihood function is written as:

$$L(\beta) = \sum_i logf(y_i;\theta_i,\phi_i) = \sum_i \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_i c(y_i,\phi) \tag{3.4}$$

The likelihood equations are

$$\partial L(\beta)/\partial\beta_l = \sum_i \frac{\partial L_i}{\partial\beta_l} = 0 \tag{3.5}$$

for all $l = 1, 2, ...p$. Differentiating the log likelihood function (3.4) we get:

$$\frac{\partial L_i}{\partial\beta_l} = \frac{\partial L_i}{\partial\theta_i}\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\zeta_i}\frac{\zeta_i}{\beta_l} \tag{3.6}$$

Since

$$\frac{\partial logL_i}{\partial\theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} \tag{3.7}$$

$$\frac{\partial^2 logL_i}{\partial\theta_i^2} = \frac{-b''(\theta_i)}{a(\phi)}$$

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

and

$$Var(Y_i) = b''(\theta_i)a(\phi)$$

Therefore equation (3.7) is given by

$$\frac{\partial logL_i}{\partial\theta_i} = \frac{(y_i - \mu_i)}{a(\phi)}$$

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{Var(Y_i)}{a(\phi)}$$

Also since the link function

$$\zeta_i = \sum_i \beta_l x_{il}$$

then

$$\frac{\partial \zeta_i}{\beta_l} = x_{il}.$$

Substituting all the above expressions into equation (3.6) we get:

$$\frac{\partial L_i}{\partial \beta_l} = \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{il} \partial \mu i}{var(Y_i) \partial \zeta_i} = 0 \tag{3.8}$$

$l = 1, 2, ..., p$, which is the function of μ_i; since $\mu_i = g^{-1}(\sum \beta_i X_{il})$. The likelihood function for the GLM also determines the asymptotic covariance matrix of the maximum likelihood estimator $\hat{\beta}$. This matrix is the inverse of the information matrix I, which has elements $-E[\frac{\partial^2 log L(\beta)}{\partial \beta_k \partial \beta_l}]$. Thus, the information matrix I from equation (3.8) is given by:

$$-E[\frac{\partial^2 L(\beta)}{\partial \beta_k \partial \beta_l}] = \sum_{i=1}^{n} \frac{(y_i - \mu_i) X_{ik} X_{il}}{var(Y_i)} (\frac{\partial \mu_i}{\partial \zeta_i})^2$$

which can be written as the form:

$$I = X'WX$$

where X is the design matrix and W is the diagonal matrix with diagonal elements

$$w_i = \frac{(\partial \mu_i / \partial \zeta_i)^2}{Var(Y_i)} \tag{3.9}$$

The asymptotic covariance of $\hat{\beta}$ which is the inverse of the information matrix is also given by

$$c\hat{o}v(\hat{\beta}) = \hat{I}^{-1} = (X'\hat{W}X)^{-1}$$

where $\hat{W}$ is the estimated W evaluated at $\hat{\beta}$. The asymptotic sampling distribution of $\hat{\beta}$ is normally distributed as follow :

$$\hat{\beta} \sim (\beta, N\hat{I}^{-1})$$

Equations (3.8) are nonlinear in $\hat{\beta}$ and are solved using iteratives methods such as Newton Raphson method, Fisher Scoring method and Iterated reweighted least square method (Agresti, 2002). Newton Raphson method is used to find $\hat{\beta}$ of β that maximize $L(\beta)$ by solving nonlinear likelihood equations by indicating the location where the function gets its maximization. Firstly an initial solution $\hat{\beta}^{(0)}$ is guessed and then updated until the iterative algorithm converges to a solution $\hat{\beta}$ which indicates the maximum likelihood estimate of β (Agresti, 2002). Let

$$u' = (\frac{\partial L(\beta)}{\partial \beta_1}, \frac{\partial L}{\partial \beta_2}, ...)$$

and let H denote the matrix having

$$h_{(il)} = \frac{\partial^2 L}{\partial \beta_i \beta_l}$$

Let u^t and $H^{(t)}$ be those terms evaluated at $\beta^{(t)}$, the guess t for $\hat{\beta}$.
At step t in iterative process ($t = 0, 2, 3, ...$), $L(\beta)$ is approximated near $\beta^{(t)}$ by the terms up to second order in its Taylor series expansion,

$$L(\beta) \approx L(\beta^{(t)}) + u^{(t)'}(\beta - \beta^{(t)}) + (\frac{1}{2})(\beta - \beta^{(t)})'H^{(t)}(\beta - \beta^{(t)}) \qquad (3.10)$$

By solving equation

$$\frac{\partial L(\beta)}{\partial \beta} \simeq u^t + H^t(\beta - \beta^t) = 0 \tag{3.11}$$

for β yields to next guess,

$$\beta^{(t+1)} = \beta^{(t)} - (H^t)^{-1}u^t \tag{3.12}$$

by assuming that $H^{(t)}$ is non singular.

The maximum likelihood estimator is the limit of $\beta^{(t)}$ as $t \to \infty$. Fisher scoring is also an iterative method; however it differs from Newton Raphson by using expected value of the second derivative matrix (Agresti, 2002).

Let $I^{(t)}$ denote the t approximation for the Newton Raphson method for the maximum of the expected information matrix, this means $I^{(t)}$ holds the elements -E(H) evaluated at $\beta^{(t)}$. Fisher scoring is given by

$$\beta^{(t+1)} = \beta^{(t)} + (I^{(t)})^{-1}u^{(t)} \tag{3.13}$$

Agresti (2002) describes the relation between Fisher scoring and weighted least square estimation in order to find maximum likelihood estimates. Iterative algorithm Fisher's method is then used to solve the score equations (3.13). At t iteration, the updated estimators are written as

$$\beta^{(t+1)} = (X'W^{(t)}X)^{-1}XW^{(t)}z^{(t)} \tag{3.14}$$

where

$$z_i^{(t)} = \zeta_i^{(t)} + (y_i - \mu_i^{(t)})\frac{\partial \zeta_i^{(t)}}{\partial \mu_i^{(t)}} \tag{3.15}$$

and

$$w_i^t = \frac{a(\phi)}{V(\mu_i^{(t)})}(g'(\mu_i^{(t)}))^2 \tag{3.16}$$

3.4 Goodness of fit in the generalized linear model

In statistical modeling, the goodness of fit describes how well a set of observations is fitted. Measures of goodness of fit consist of summarizing the discrepancy between the observed values and the expected values. In GLM two common statistics are useful to measure the goodness of fit. These include Pearson chi-square test and deviance. According to Smyth (2003), the Pearson goodness of fit statistic is the score test statistic used to test the fitted model against the saturated model. Pearson's chi-square statistic includes the test of independence in two way contingence of tables. Given a generalized linear model with response y_i, weights w_i, fitted mean $\hat{\mu}_i$, variance function $\sigma^2(\mu_i)$ and dispersion $\phi = 1$, the Pearson's goodness of fit is written as:

$$\chi^2 = \sum \frac{w_i(y_i - \hat{\mu}_i)}{\sigma^2(\hat{\mu}_i)}$$

where $\hat{\mu}_i$ is the expected value μ_i evaluated at the maximum likelihood estimator $\hat{\beta}$ The Pearson residual is obtained by scaling the response residual with $\sqrt{\hat{\sigma}^2(y_i)}$, it is called the response residual normalized with the estimated standard deviation for the observations. The Pearson residual is written as

$$r = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\sigma}^2(y_i)}}$$

where μ_i is the maximum likelihood estimate for μ_i, and $\hat{\sigma}^2(y_i) = \phi(\hat{\mu}_i)$ is the estimated variance of y_i. Another measure of goodness of fit in GLM is the deviance. It is used for testing the fit of the link function and the linear predictor to the data or testing the significance of a particular predictor variable(s) in the model. Then, let $L(\Psi, y)$ and $L(\psi, y)$ be likelihood of the saturated and the model of interest and $L(\Psi, y) \geq L(\psi, y)$. The deviance is

defined $D = 2(LogL_{\hat{\Psi}} - logL_{\hat{\psi}})$, where $\hat{\Psi}$ and $\hat{\psi}$ are the maximum likelihood estimations of the saturated and model interest respectively. Suppose that μ is the mean parameter and ϕ the dispersion parameter, then

$$D = \frac{-2[logL(\hat{\mu}, \phi.y) - logL(y, \phi, y)]}{\phi} \tag{3.17}$$

where $logL(\hat{\mu}, \phi, y)$ is the log-likelihood obtained by fitting model and $logL(y, \phi, y)$ is the log likelihood obtained in saturated model.
When $\phi = 1$, the deviance becomes

$$D = -2logL(\hat{\mu}, y) - logL(y, y) \tag{3.18}$$

In addition, the likelihood ratio is also used to test for the goodness of fit. The likelihood-ratio statistic $-2(L_0 - L_1)$ tests whether certain models parameters are zero by comparing the log likelihood L_1 for the fitted model F_1 with L_0 for a simpler model F_0. The likelihood ratio (LR) comparing two models is given by:

$$\begin{aligned} LR(F_0 \setminus F_1) &= -2(L_0 - L_1) \qquad (3.19) \\ &= -2(F_0 - L_S) - [-2(L_1 - L_S)] \\ &= LR(F_0) - LR(F_1) \end{aligned}$$

(3.20)

where L_S denote the maximized log likelihood for the saturated model (Agresti, 2002).

3.5 Model selection

Model selection is based on well-justified criterion of what is the best model. The criterion must be estimable from the data for each fitted model and the

criterion must fit into a general statistic inference framework (Burnham and Anderson, 2004). Akaike Information Criterion (AIC) is used to measure discrepancy or symmetric distance between the model which generates the data (full model) and the fitted (reduced) model (Akaike, 1973 and 1974). It is defined as:
$AIC = -2(ln(likelihood)) + 2K$, where likelihood is the probability of the data given a model and K is the number of the free parameters in the model. It is defined as :

$$AIC_c = -2(ln(likelihood)) + 2K * (n/n - K - 1)$$

where n is the sample size. To select the best model, Ashby (1992) states that the minimum AIC optimize predictability. However, it only means that the model is the best among competing models in the sense that it gives predictions closest to those by the correct model. Agresti (2007) also states that the optimal model is one that tends to have its fitted values closest to the outcome probability. Bayesian Information Criterion (BIC) was introduced by Schwarz (1978) as a competitor to the Akaike (1973,1974) Information Criterion. As reported by Schwarz (1978), the problem of selecting one of a number of the models of different dimensions is treated by finding its Bayes solution and evaluation the leading terms of its asymptotic expansion. Model selection based on BIC is roughly equivalent to the model selection based on Bayes factors information Criterion (BIC) is written as : $BIC = -2(ln(likelihood)) + Klogn$.

3.6 Logistic regression model

3.6.1 Introduction

Logistic regression is a statistical tool analysis of categorical variables. It describes the relationship between a categorical response variable and a set of explanatory variables (Archera et al., 2007). Moreover, logistic model is a flexible modeling tool that can simultaneously accommodate both categorical and continuous predictors as well as interactions among predictors (Steven et al., 2010). The response variable can be dichotomous (present or absent, success or failure), polytocomous (more than two response variables), the multiple-level response variables and nominally or ordinarily scaled (Stokes et al., 2000). Logistic regression model are frequently used in epidemiologic studies for estimating associations that demographic, behavioral, medical research, banking, marketing research, social research and risk factor variables have on a dichotomous outcome, such as disease being present versus absent (Archera et al., 2007; Stokes et al., 2000). Proc GENMOD is a procedure in *SAS* for analyzing generalized linear model of which logistic regression is a simple case. In this research we limit analysis to the logistic regression with two outcomes (binary logistic regression) (Stokes et al., 2000).

3.6.2 Binary logistic regression model structure

Let Y denote binary outcome response variable, for example Y may indicate the smoking status (smoker, non smoker), diagnostic of high blood pressure (present, absent), or results of final examination (success, failure). Those outcomes will be attributed values 0 and 1, hence the distribution resulting

is binomial. The response Y is defined as

$$Y = \begin{cases} 1 & \text{if the outcome is present} \\ 0 & \text{if the outcome is absent} \end{cases} \tag{3.21}$$

For binary response variable Y and the explanatory variable X, let the probability $\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$.
Logistic model is given by

$$logit[\pi(x)] = log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x \tag{3.22}$$

The link function is the logit and it is nonlinear in $\pi(x)$, but is presumed to be linear in the regression parameter β. The values that the random variable Y can take are 1 and 0 with the probabilities π and $1 - \pi$ respectively. The mean function $E(Y) = P(Y = 1)$ denoted by $\pi(Y)$ and the variance function $V(Y) = \pi(x)[1 - \pi(x)]$ is called variance of binomial distribution for single trial. We note the following :

- Logistic regression does not assume a linear relationship between the response variable and explanatory variables.
- The response variable must be a dichotomous (2 categories).
- The explanatory variables are not assumed to be interval, not normally distributed, not linearly related and not having equal variance within groups.
- The categories (groups) must be mutually exclusive and exhaustive.

3.6.3 Odds Ratio in logistic regression

According to Collett (2003), the odds of a success π is defined to be the ratio of the probability of success to the failure $(1 - \pi)$. Thus if π is the true

success probability, the odds of a success, the odds of success are defined as follows:

$$Odds = \frac{\pi}{1-\pi}$$

The odds by which the response variable being present among individual with $Y = 1$ and $Y = 0$, is defined as $(\pi)/[1-\pi(1)]$ and $(\pi)/[1-\pi(0)]$ respectively.

Agresti (2007), describes the odds ratio (OR) as the measure of association for 2×2 contingency tables. It occurs as a parameter in the most important type of the model for categorical data. The ratio of the odds for $Y = 1$ to the odds for $Y = 0$ denoted by Υ , is given by the following equation:

$$\Upsilon = \frac{\frac{\pi(1)}{[1-\pi(1)]}}{\frac{\pi(0)}{[1-\pi(0)]}}$$

The precision of the odds ratio is determined by the 95% confidence intervals. When the confidence interval is large, the level of precision of OR is low, whereas a small confidence interval indicates the higher precision of OR.

- $OR = 1$, no effect for the exposure to odds of outcome.
- $OR > 1$, exposure associated with higher odds of outcome.
- $OR < 1$, exposure associated with lower odds of outcome.

3.7 Estimation in logistic regression

As stated by Hosmer and Lemeshow (2000), equation (3.22) is fitted to a set of data by estimating the values of β_0 and β_1 the unknown parameters. Recall that in linear regression those values are chosen to minimize the sum of squares of the observed values of the response variable Y from the predicted values based upon the model. The general method of estimation that leads to

the least squares function under the linear regression model (when the error terms are normally distributed) is called the maximum likelihood. In logistic regression, in order to apply this method we have to find a function, called likelihood function. The maximum likelihood estimator of the parameters are chosen to be the values that maximize the likelihood function. Thus, the resulting estimators are those which agree most with the observed data. The likelihood function for the response Y coded by 0 and 1 is given by:

$$l(\beta) = \prod_{i=1}^{N} [\pi(x_i)]^{y_i} [1 - \pi(x_i]^{1-y_i} \tag{3.23}$$

where the probability of success of y_i is π^{y_i} and the probability of $1-y_i$ failures is $[1 - \pi(x_i)]^{1-y_i}$. The principle of maximum likelihood states that we use our estimate of β the value which maximizes the expression in equation(3.23) (Hosmer and Lemeshow, 2000). The resulting log-likelihood function from equation(3.23) is given by:

$$\begin{aligned} L(\beta) &= log[l(\beta)] \\ &= \sum_{i=1}^{n} y_i log[\pi(x_i)] + (1 - y_i) log[1 - \pi(x_i)] \end{aligned} \tag{3.24}$$

To find the value of β that maximizes $L(\beta)$, we differentiate $L(\beta)$ with respect to β_0 and β_1 and set the resulting expressions equal to zero. The equations are known as the likelihood equations written as:

$$\sum [y_i - \pi(x_i)] = 0 \tag{3.25}$$

and

$$\sum x_i [y_i - \pi(x_i)] = 0 \tag{3.26}$$

where $i = 1, 2, ..., n$. Equations (3.25) and (3.26) are nonlinear in β_0 and β_1, and require iterative methods, stated above, for their solution.

3.8 Model selection in logistic regression

As we discussed in section 3.5, Akaike Information criterion (AIC) is well known to select the best model. AIC judges a model by how close its fitted values tend to be to the true values. The optimal model is the one that tends to have its fitted values closest to the true outcome probabilities (Agresti, 2007). The selection process becomes harder as the number of explanatory variables increase because of the rapid increase in possible effects and interactions. Stepwise procedures that is backward elimination and forward selection, are more useful methods of variables selection. Forward selection adds terms sequentially until further additions do not improve the fit. At each stage it selects the term giving the greatest improvement in fit. However, backward eliminations begins with a complex model and sequentially removes terms. At each stage, it selects the term for which its removal has the least damaging effect on the model; for more details see Agresti (2002) and Hosmer and Lemeshow (2000).

3.9 Goodness of fit in logistic regression

According to Hosmer and Lemeshow (2000), the goodness of fit is assessed over the constellation of fitted values determined by the covariates in the model, not the total collection of covariates. The strategies for the fit of the model by comparing the observed model whose values are: $Y = y_1, y_2, ..., y_n$, and the predicted model which has the values $\hat{Y} = \hat{y}_1, \hat{y}_2, ..., \hat{y}_n$ are the following:

1. The summary measures distance between Y and $\hat{Y}$ are small.

2. The contribution of each pair $(y_i, \hat{y}_i)$, $i = 1, 2, ..., n$ to these summary

measures is not systematic and is small compared to the error structure of the model.

In logistic regression, there are several possible ways to measure the difference between the observed and the fitted values. As stated by Hosmer and Lemeshow (2000), Hosmer and Lemeshow test, Pearson chi-square statistic and deviance are used for assessing the goodness of fit. Hosmer and Lemeshow test procedure provides the probabilities estimates of the event of each subject by ascending order. Hosmer and Lemeshow test groups the data according to percentile of the estimated probabilities by using $G = 10$ groups, and the first group contains the $n_1 = \frac{n}{10}$ subjects whose estimated probabilities are less or equal to the first decile of the estimated probabilities. The p^{th} group contains the subjects whose estimated probabilities are between the $(p-1)^{th}$ and p^{th} of the estimated probabilities. For the k^{th} group, let c^k denote the number of covariates patterns and count the number of observed responses among c_k covariates patterns: $O_k = \sum_{J=1}^{c_k} y_i$, the average estimated probability in k^{th} decile is computed as: $\overline{\pi}_k = \sum_{J=1}^{c_k} \frac{m_j \hat{\pi}_j}{n_k}$. The Hosmer and Lemeshow test statistic is then

$$\hat{C} = \sum_{k=1}^{h} \frac{(O_k - n_k \overline{\pi})^2}{n_k \overline{\pi}(1-\overline{\pi})} \tag{3.27}$$

Let X^2_{HLS} denote the Hosmer and Lemeshow goodness of fit. X^2_{HLS} has $g-2$ degrees of freedom, and it is compared with the critical value of the chi-square distribution with $g-2$ df (χ^2_{g-2}, α) to check the goodness of fit. The interpretation is as follows: if X^2_{HLS} is significant, it implies that there is lack of fit, but if X^2_{HLS} statistic is insignificant then the model is well fitted and has goodness of fit. The Pearson Chi-square test is defined as follow:

$$\chi^2 = \frac{\sum_{j=1}^{M} (y_j - n_j \hat{\pi}_j)^2}{n_j(1-\hat{\pi}_j)} \tag{3.28}$$

and

$$Deviance = \sum_{j=1}^{J} d(y_j, \hat{\pi}_j)^2 \tag{3.29}$$

where $\hat{\pi}_j = \frac{e^{\hat{g}(x_j)}}{1+e^{\hat{g}(x_j)}}$ and $\hat{g}(x_j)$ is the estimated logit.

3.10 Logistic model diagnostic

The aim of diagnostic test is : (1) to identify outliers that deviate from the postulated model, (2) to identify influential observations that have a large effect on the statistical inference drawn from the postulated model and lastly (3) to validate the chosen statistical model (Pan et Fang., 2002).

Diagnostic test in logistic regression is accessed by ROC (Receiver operating characteristic) curve, Pearson residual, deviance residual, *Dfbeta*, hat matrix diagonal (leverage) and Cook's distance, just to mention a few. Pregibon (1981) suggested the use of index plots of several diagnostic statistics to identify influential observations and to quantify the effects on various aspects of the maximum likelihood fit. Agresti (2002, 2007) stated that diagnostics test are used to detect many medical conditions. He defined sensitivity as the true proportion of positive results that a test elicits when performed on subjects known to have a disease. And specificity as the true proportion of negative results that a test elicits when performed on subjects known to be disease free. Hence the sensitivity is given by : $P(Z = 1|X = 1)$ and specificity: $P(Z = 0|X = 0)$, where X denote the true state of a person, with categories 1: diseased and 0: not diseased and Z denote the outcome diagnostic test, with categories 1: positive and 0: negative. The higher the sensitivity and specificity the better the diagnostic test. Moreover, Agresti (2002, 2007) defined ROC curves as the popular ways evaluating diagnostic

tests. ROC curve usually has a concave shape connecting the points $(0, 0)$ and $(1, 1)$. The prediction is $\hat{Z} = 1$ when $\pi_i > \pi_0$ and $\hat{Z} = 0$ when $\pi_i \leq \pi_0$. The ROC is a plot of sensitivity as function of 1-specificity for the possible cut-off π_0. The ROC curve is more informative than the classification table, since it summarizes predictive power for all possible π_0. The higher area under the curve gives the better predictions.

According to Hosmer and Lemeshow (2000), Pearson residual and deviance residual are used to measure the difference between the observed response variable Y and its predicted (fitted) values $\hat{Y}$. Pearson residual is defined as follows:

$$r = (y_j, \hat{\pi}_j) = \frac{(y_j - m_j\hat{\pi}_j)}{\sqrt{m_j\hat{\pi}_j(1 - \hat{\pi}_j)}} \tag{3.30}$$

where m_j subjects share the j^{th} observation pattern, y_j of them experience the event of interest and π_j is defined above. And the deviance residual is defined as:

$$d(y_j, \hat{\pi}_j) = sign\{2[y_j log(\frac{y_j}{m_j\hat{\pi}_j}) + (m_j - y_j)log(\frac{m_j - y_j}{y_j(1 - \hat{p}_j)})]\}^{\frac{1}{2}} \tag{3.31}$$

where the $sign = 1$ if $y_j > m_j\hat{\pi}_j$ and $sign = -1$ if $y_j < m_j\hat{\pi}_j$.

Cook's distance and *DFbeta* are also measures used in model diagnostic. Cook's measure of distance helps to determine whether an outlier (either on the response variable or on the set of predictors) is influential (Stevens, 1984). *Dfbeta* accesses the effect of individual observation on estimated parameter of the fitted. A *Dfbeta* diagnostic is computed for each observation for each parameter estimate. It is the standardized difference in the parameter estimate due to deleting the corresponding observation and is useful in deleting observation that causes instability in the selected coefficient.

3.11 Results from fitting logistic regression

As we discussed in Chapter 1 section 1.8, the previous studies conducted on smoking in Kenya have limitations, which will be addressed in our current research. In addition these studies did not consider other characteristics of households such as marital status, ethnicity, religion, size of household and wealth index. The present analysis was focused on modeling smoking status of Kenya males across the whole country by all age categories (15 to 54 years old) using also the variables such as size of household, region, marital status, religion, access to mass media, education, sex of household head, wealth index, and ethnicity. The results are obtained using logistic regression which is the special case of the generalized linear model. The dependent variable is smoking status, and the explanatory variables which have been found to be significant via the cross tabulation and the chi-quare test. These variables are region, marital status, religion, education, age, ethnicity, wealth index, size of household and access to mass media. In addition two ways of interaction effects will be investigated as to their significance in the current research.

3.11.1 Model of smoking status

Recall that the main objective of the current research is to model the smoking status of Kenya males in the presence of missing values using KDHS 2008-2009. Let Y be the response variable which describes the smoking status for males, the relation 3.21 is written as:

$$Y = \begin{cases} 1 & \text{if the } i^{th}\text{man is smoking} \\ 0 & i^{th}\text{man is not smoking, i=1,...3465} \end{cases} \tag{3.32}$$

The logistic model is defined as follow:

$$\begin{aligned} logit(\pi) &= ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \\ &+ \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X7 + \beta_8 X_8 + \beta_9 X_9 \\ &+ \beta_{10} X_3 * X_4 + \beta_{11} X_7 * X_8 + \beta_{12} X_2 * X_3 \end{aligned} \tag{3.33}$$

where π is the probability that i^{th} man is smoking, $1-\pi$ is the probability of no smoking, X_1 is the region, X_2 is the marital status, X_3 is the religion, X_4 is the education, X_5 is the age, X_6 is the ethnicity, $X7$ is the wealth Index, X_8 the size of household, X_9 is the access to mass media, X_3* X_4 is the interaction between religion and education, X_5* X_7 is the interaction between age and wealth index and X_8* X_9 is the interaction between size and access to mass media. The goodness of fit, overdispersion, influential point and model predictive accuracy power are then used for model checking.

3.11.2 Interpretation

Table 3.1 indicates that the overall fit of the model is significant. We used the Hosmer and Lemeshow goodness of fit, Pearson chi-square, deviance scale, AIC, residual and influential observation to check the goodness of fit. The results shown in the Table 3.2, indicate that the scale parameter is .917, which shows that there is non overdispersion. Also the Pearson chi-square is 1.012 which is approximately equal to 1, which further confirms that there is non overdispersion.

The results shown in Table 3.3 denote that the linear predictor is significant (p-value=<.0001), and the squared linear predictor is not significant (p-value=.8889), which means that the link function is appropriate. This implies that the prediction given by the linear predictor is not improved by adding the square linear predictor term, and thus suggest the consistency of

the choice of the link function (Vittinghof et al., 2005, pp. 192-194). The logit link is used when the residual are not normally distributed and they cannot be constant across values of the predictor. Instead of a normal distribution of errors, we assume the errors are logistically distributed and the logit is the cumulated function which describes the distribution of the residuals.

The Hosmer and Lemeshow test shown in Table 3.4 has the logit of 5.320 with 8 d.f and the p-value of .723, which indicates a good fit of the model (p-value is greater than 0.05) to the data. To access influential observations we plotted the Cook's distance statistic. Cook (1977) proposed that the influence diagnostic must be larger than 1 for individual case to have an effect on the estimated coefficient. Figure 3.1 shows that the values of the Cook's distance are less than 1, which implies that there is no influence on parameters. This prove the adequacy of the model. As shown in Figure 3.2, the studentized Pearson residuals and deviances residuals are plotted against the estimated logit probability respectively. This implies that the existing outliers detected by the residuals plots are not influential. In addition Figure 3.3 reveals that the curve of logit is .795(79%), which implies the predictive adequacy and the consistency of the model to predict the male smokers.

Table 3.1: Testing Global Null Hypothesis:BETA=0

Test	χ^2	Df	P-alue
Likelihood Ratio	681.8185	76	<.0001
Score	600.9978	76	<.0001
Wald	442.7817	76	<.0001

Table 3.2: Criteria for assessing model goodness of fit

Criterion	**DF**	**Value**	**Value/DF**
Deviance	2790	2559.006	.917
Scaled deviance	2559.274	2761.006	.917
Pearson chi-square	2790	2823.677	1.012
Scaled Pearson chi-square	2790	2823.677	1.012
Log Likelihood	-1345.713		
AIC(small is better)	2845.425		
AICC(small is better)	2845.985		
BIC(small is better)	3318.722		

Table 3.3: Link test for logistic model

Parameter	DF	Estimate	Std	wald χ^2	P-value
Intercept	1	.0022	.0680	.0000	.9748
Linear predictor	1	1.0102	.0875	133.23	<.0001
Squared linear predictor	1	.0041	.0294	.0200	.8889

Table 3.4: Hosmer and Lemeshow Test

Chi^2	df	P-value
5.320	8	0.723

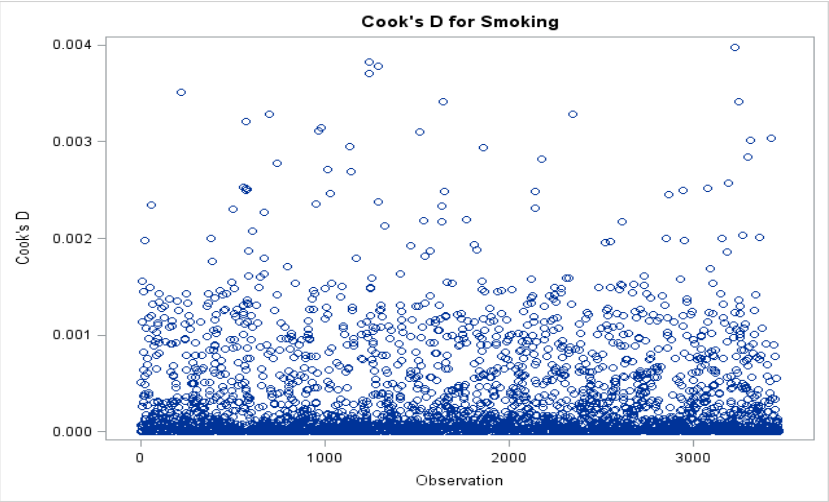


Figure 3.1: Cook's distance

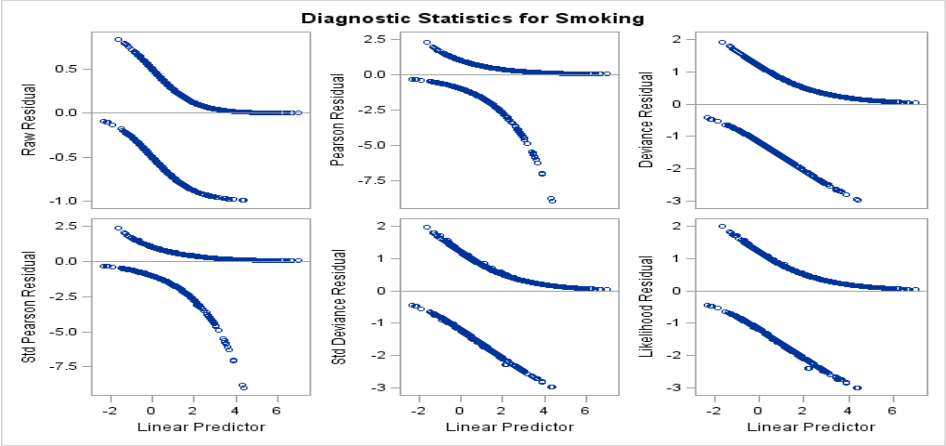


Figure 3.2: Residual plots

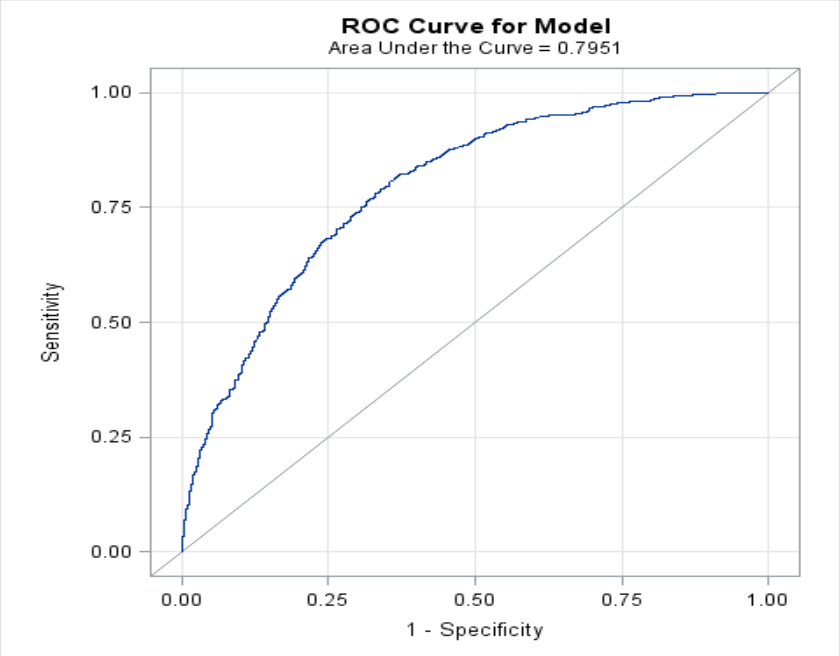


Figure 3.3: ROC curve

The results shown in Table 3.5 and 3.6 indicate the variables in the model of smoking status among different household characteristics in Kenya. We have found that the odds of smoking for a man from Eastern province is 1.570 (p-value=.049) times that of the odds of smoking for a man from Nairobi province. Further, the odds of a man from Nyanza province and Western province is .510 (p-value=.007) and .540 (p-value=.014) times respectively that of the odds of smoking for a man from Nairobi province. Marital status reveals that, the odds of smoking of a divorced man is 2.530 (p-value=.000) times that of the odds of smoking for an unmarried man. Religion shows that the odds of smoking for a Muslim man is .450 (p-value=.002) times that of the odds of smoking for a Catholic man. Age reveals that the odds of smoking of a man aged from 20 to 24, 30 to 34, 35 to 39, 40 to 44, 45 to 49 and 50 to 54 years old is 6.140 (p-value=.000), 13.100 (p-value=.000), 19.700 (p-value=.000), 17.400 (p-value=.000), 25.800 (p-value=.000), 25.900 (p-value=.000) and 18.700 (p-value=.000) times respectively that of the odds

of smoking for a man aged from 15 to 19 years old. Size of household reveals that a man from a household of 4 persons is 2.490 (p-value=.007) times more likely to be a smoker as compared to a man from a household of 1 person. The results also reveal that a man from a household 3 persons is .130 (p-value= .006) times less likely to be a smoker, as compared to a man from a household of 1 person. Ethnicity reveals that, the odds of smoking for a man from Kelenjin and Kisii tribe is .310 (p-value=.002) and .240 (p-value=.002) times respectively that of the odds of smoking for a man from Embu tribe.

Interaction effects

Several interaction effects were fitted in the model, but only three were found to be significant, that is interaction of religion by education, age by wealth index and size by access to mass media.

The interactions effects between religion and education, age and wealth index as well as the size of household and access to mass media, are shown in Figure 3.4, where the vertical axe indicates the logit of smoking over no smoking, and the horizontal axe indicates the interaction effects. The results show that the odds of smoking for a Muslim man with primary level of education and a Muslim man with secondary and higher is 10.100 (p-value=.000) and 9.490 (p-value=.000) times respectively that of the odds of smoking for a Catholic man with non education. The odds of smoking for a richest man aged from 25 to 29, 30 to 34, 40 to 44, 45 to 49, 50 to 54 years old is .480 (p-value=.027), .650 (p-value=.034), .360 (p-value=.014) and .350 (p-value=.034) times respectively that of the odds of smoking of a poor man aged from 15 to 19 years old. Further, the odds of smoking for a man from a household of 2 persons who has access to mass media less that once a week, at least once a week and almost every day is 12.200 (p-value=.024), 13.000

Table 3.5: Results from Logistic regression model

Indicator	Estimate	Std.Error	Wald	P-Value	Exp(B)	95% C.I of EXP(B)
Intercept	-1.97	.626	.9.892	.002*	.140	
Province (ref=Nairobi)						
Central	-.091	.226	.161	.688	.910	(563,1.381)
Coast	-.121	.211	.330	.566	.890	(.577,1.334)
Eastern	.453	.230	3.885	.049*	1.570	(1.003,2.508)
Nyanza	-.683	.254	7.251	.007*	.510	(.297,.809)
Rift Valley	.079	.233	.115	.735	1.080	(.672,1.689)
Western	-.608	.246	6.092	.014*	.540	(.306,.819)
Northeastern	-.477	.314	2.309	.129	.620	(.349,1.222)
Marital Status (ref=Not married)						
Married	.149	.159	.878	.349	1.160	(.810,1.543)
Living together	-.053	.387	.018	.892	.950	(.397,1.857)
Divorced or not living together	.924	.238	15.066	.000*	2.520	(1.505,3.858)
Religion (ref=Roman catholic)						
Protestant	-.701	.616	1.298	.255	.500	(.139,1.591)
Muslim	-1.451	.548	7.013	.001*	.230.	(076,.659)
Non religion and others	.277	.594	.217	.641	1.320	(.377,3.937)
Education (ref=Non education)						
Primary incomplete	.025	.479	.003	.959	1.030	(.390,2.594)
Primary complete	-.795	.480	2.743	.098	.450	(.177,1.176)
Secondary and higher	-.866	.474	3.337	.068	0.420	(.173,1.135)
Age of respondent(ref=15-19)						
20-24	1.815	0.331	30.099	.000*	6.1402	(2.142,11.369)
25-29	2.574	.345	55.651	.000*	13.100	(4.252, 24.041)
30-34	2.982	0.34	77.042	.000*	19.700	(6.753, 37.732)
35-39	2.856	.354	65.042	.000*	17.400	(5.747, 33.481)
40-44	3.250	.356	83.346	.000*	25.800	(8.593,50.266)
45-49	3.254	.382	72.586	.000*	25.900	(8.367,53.176)
50-54	2.926	.405	52.127	.000*	18.700	(5.911,40.517)

*P-values <.05

Table 3.6: Results from Logistic regression model

Indicator	Estimate	Std.Error	Wald	P-Value	Exp(B)	95% C.I of EXP(B)
Ethnicity (ref=Embu)						
Kelenjin	-1.188	.377	9.936	.002*	.310	(.145,.841)
Kamba	-.311	.308	1.025	.311	.730	(.390,1.327)
Kikuyu	.182	.327	.310	.577	1.200	(.623,2.264)
Kisii	-1.421	.453	9.832	.002*	.240	(.100,.791)
Others	-.587	.300	3.812	.051	.560	(.314,1.026)
Wealth Index (ref=poorer or poorest)						
Middle or rich	-0.125	.493	.064	.800	.880	(.336,2.319)
Richest	.001	.809	0.000	.999	1.000	(.205,2.287)
Size of household (ref=1 person)						
2 persons	-2.545	1.082	5.528	.019*	.080	(.007,.644)
3 persons	-2.038	1.093	3.479	.006*	.130	(.010,.915)
4 persons	.913	.508	3.223	.007*	2.490	(.519,4.107)
5 persons	.659	.676	.950	.330	1.930	(.322,7.174)
6 persons and above	-.251	.296	.720	.396	.780	(.209,1.597)
Access to mass media (ref=Not at all)						
Less than once a week	-.242	.519	.216	.642	.790	(.284,2.176)
At least once a weak	-.481	.525	.837	.360	.620	(.221,1.733)
Almost every day	.388	.509	.580	.446	.680	(.250,1.841)

* P-values <.05

(p-value=.020) and 9.000 (p-value=.042) times respectively that of the odds of smoking for a man from a household of 1 person who does not have access to mass media. The odds of smoking for a man from a household of 3 persons who has access to mass media less than a week is 9.390 (p-value=.043) times that of the odds of smoking for a man from a household of 1 person who does not have access to mass media. The odds of smoking for a man from a household of 4 persons who has access to mass media less than once a week, at least once a week and almost every day is .290 (p-value=.019), .260 (p-value=.014), and .250 (p-value=.009) times respectively that of the odds of smoking for a man from a household of 1 person who does not have access to mass media. The odds of smoking for a man from a household of 6 persons and above who has access to mass media almost every day is .510 (p-value=.047) times that of the odds of smoking for a man from a household

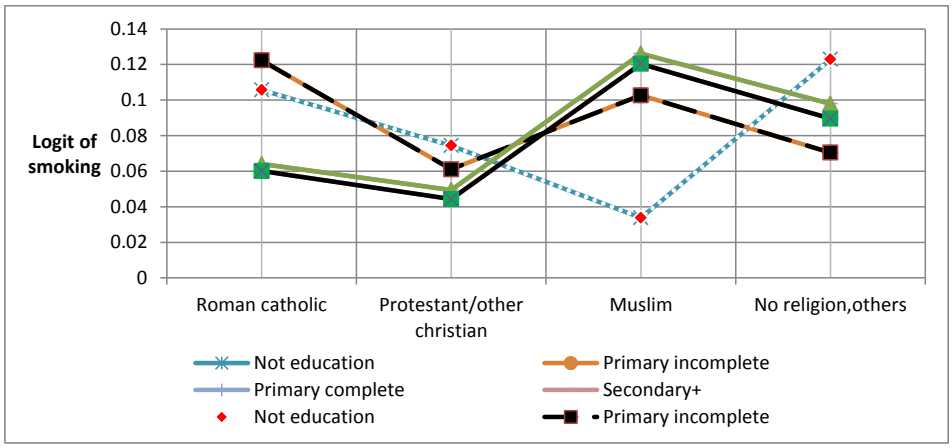
of 1 person who does not have access to mass media.

Table 3.7: Results from Logistic regression model

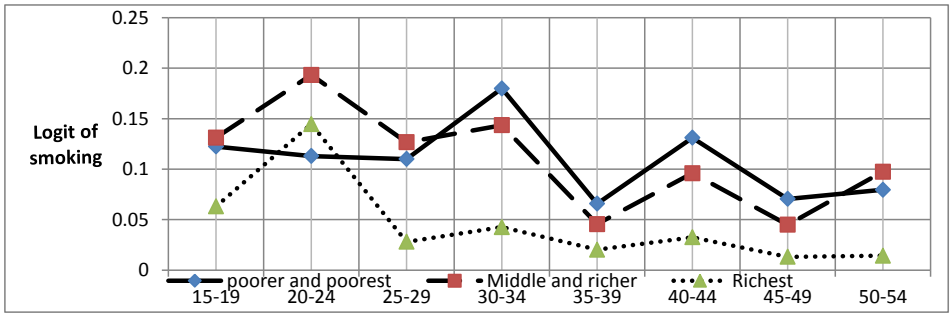
Indicator	Estimate	Std.Error	Wald	P-Value	Exp(B)	95% C.I of EXP(B)
Religion by Education(ref=Roman catholic by non education)						
Protestant by primary incomplete	-.148	.652	.051	.821	.860	(.240,3.096)
Protestant by primary complete	.44	.651	.457	.499	1.550	(.434,5.563)
Protestant by secondary and higher	.368	.640	.331	.565	1.450	(.412,5.069)
Muslim by primary incomplete	1.076	.620	3.014	.083	2.930	(.870,4.886)
Muslim by primary complete	2.314	.609	14.421	.000*	10.100	(3.064,20.388)
Muslim by secondary and higher	2.250	.585	14.776	.000*	9.490	(3.012,16.873)
Non religion by primary incomplete	-.970	.700	1.916	.166	.380	(.096,.997)
Non religion by primary complete	.655	.726	.815	.367	1.930	(.464,3.982)
Non religion by secondary and higher	328	.701	.218	.640	1.390	(.351,2.482)
Age by Wealth Index(ref=15-19 by poorer or poorest)						
20-24 by middle or rich	-.316	.293	1.164	.821	1.370	(.773,2.436)
20-24 by richest	.142	.340	.173	.677	1.150	(.591,2.244)
25-29 by middle or rich	-.080	.292	.076	.783	1.080	(.611,1.922)
25-29 by richest	-.727	.329	4.965	.027*	.480	(.254,.922)
30-34 by middle or rich	-.349	.278	1.578	.209	.710	(.410,1.216)
30-34 by richest	-.869	.310	7.846	.005*	.420	(.228,1.270)
35-39 by middle or Rich	-.468	.338	1.916	.166	.630	(.323,1.215)
35-39 by richest	.495	.323	2.353	.125	.610	(.324,1.147)
40-44 by middle or rich	-.432	.333	1.681	.195	.650	(.338,1.247)
40-44 by richest	-.772	.363	4.512	.034*	.460	(.227,.942)
45-49 by riddle or rich	.559	.349	2.558	.110	.570	(.288,1.134)
45-49 by richest	-1.013	.413	6.005	.014*	.360	(.161,.816)
50-54 by middle or rich	-.142	.383	.137	.711	1.150	(.544,2.443)
50-54 by richest	-1.059	.499	4.494	.034*	.350	(.130,.923)
Size by access to mass media (ref=1 person by not at all)						
2 persons by less than once a week	2.504	1.106	5.125	.024 *	12.200	(1.399,15.847)
2 persons by at least once a week	2.568	1.105	5.396	.020*	13.000	(1.494,17.738)
2 persons by almost every day	2.244	1.102	4.145	.042*	9.430	(1.087,12,863)
3 persons by less than once a week	2.240	1.106	4.100	.043*	9.390	(.734,11.852)
3 persons by at least once a week	1.863	1.108	2.827	.093	6.450	(.483,10.978)
3 persons by almost every day	1.454	1.114	1.705	.192	4.280	(.456,7.568)
4 persons by less than once a week	-1.239	.527	5.530	.019*	.290	(.088,.764)
4 persons by at least once a week	-1.35	.551	6.000	.014*	.260	(.087,.703)
4 persons by almost every day	-1.395	.532	6.881	.009*	.250	(.117,1.742)
5 persons by less than once a week	-.793	.688	1.329	.249	.450	(.110,1.678)
5 persons by at least once a week	-.846	.696	1.478	.224	.430	(.072,1.174)
5 persons by almost every day	-1.235	.712	3.011	.083	.290	(.078,1.358)
6 persons and above by less than once a week	-.136	.272	.251	.616	.870	(.513,1.486)
6 persons and above by at least once a week	-.230	.293	.620	.431	.790	(.563,1.896)
6 persons and above by almost every day	-.667	.339	3.941	.047*	.510	(.324,1.134)

*P-values <.05

Interaction between religion and education



Interaction between age and wealth index



Interaction between size and access to mass media

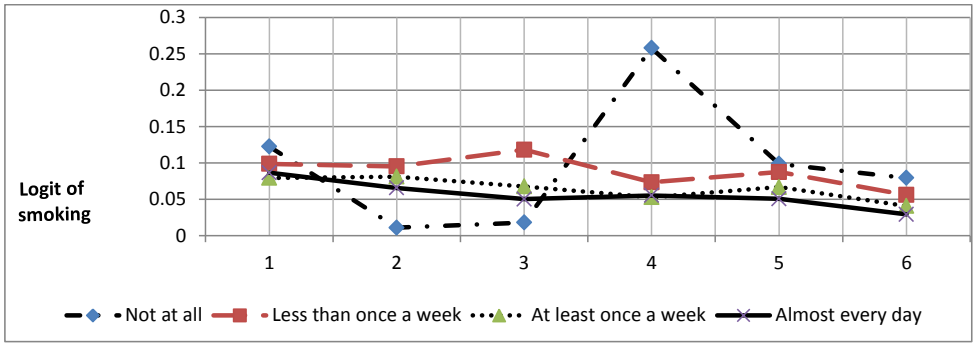


Figure 3.4: Interaction effects

Chapter 4

GENERALIZED LINEAR MIXED MODEL

4.1 Introduction

As stated by McCulloch and Searle (2001) and Bolker et al.(2008), generalized linear mixed model (GLMM) is an extension of the linear mixed model which deals with a non-normal distribution of the response variable and the linear predictor contains the random effects together with the fixed effects. GLMM provides a more flexible approach for analyzing non-normal data when random effects are present. The random effects describe the factors whose levels are sampled from a large population or whose interest lies in the variation among the levels rather than the specific effects of each level. However, the fixed effects are factors whose levels interest lies in the specific effects of each level, such as effects of the covariates, differences among treatments and interactions. Factors in the model are classified as either fixed or random effects depending on the choice of the scope of inference.

4.2 Model structure

Recall that the linear mixed model describes a statistical model containing both fixed effects and random effect for a continuous random variables. The model is represented as,

$$Y = X\beta + Zr + \varepsilon \tag{4.1}$$

where: Y is the $n \times 1$ vector of observations, β is a $p \times 1$ vector of coefficients of the fixed effects, r is a $q \times 1$ vector of random effects, ε is a $n \times 1$ vector of random error terms, X is the $n \times p$ design matrix for the fixed effects relating observations Y to β, Z is the $n \times q$ design matrix for the random effects relating observations Y to r. The vector r and ε are assumed to be uncorrelated random variables with zero means and covariance matrices G and R respectively. In section 3.1, we saw that generalized linear model generalizes the classical linear model is order to unify different approaches of modeling data from the distributions which do not necessary follow a normal distribution. In this section we discuss generalized linear mixed model, which incorporates the linear model, the generalized linear model and the random effects. GLMMs are the best tool for analyzing non-normal data that involve random effects (Bolker et al., 2008). Suppose that the response variables $y_1, y_2, .., y_n$ such that their conditional distribution, given the vector r, is a member of exponential family with the probability density function

$$f(y_i|r) = \exp\{\frac{y_i(\theta_i) - b(\theta_i)}{a(\phi)} - c(y_i, \phi)\} \tag{4.2}$$

where i=1,...,n and $b(.)$, $a_i(.)c_i(.)$ are known functions and ϕ is a dispersion parameter which may or not be known. The quantity θ_i is associated with the conditional mean $\mu_i = E(y_i|r)$. The linear predictor is formulated by the fixed and random effects

$$\zeta = x_i^{'}\beta + z_i^{'}r \tag{4.3}$$

where $x^{'}$ denote the i^{th} row of the model design matrix for the fixed effects; β denote the vector of the coefficients of the fixed effects; $z_i^{'}$ is the i^{th} row of model design matrix for the random effect r. The linear predictor is used to model the relationship between the fixed and random effects. Hence, the generalized linear mixed model is defined as:

$$\zeta = x_i^{'}\beta + z_i^{'}r \tag{4.4}$$

and the link function is given by

$$g(\mu_i) = x_i^{'}\beta + z_i^{'}r$$

Like generalized linear model, generalized linear mixed model includes a linear predictor ζ and a link function. In addition the condition mean μ_i depends on the linear predictor through an inverse link g(.), and the covariance matrix, R, depends on μ_i through a variance function. The mean of the GLMM is given by $E[y_i|r] = E(g^{-1}(x_i'\beta + z_i'r)) = g^{-1}(\zeta)$ where y_i represents the $n \times 1$ response vector, $x_i^{'}$ represents the $n \times p$ design matrix of rank k for the $p \times 1$ fixed effects coefficients β and $z_i^{'}$ the $n \times q$ design matrix for the $q \times 1$ random effect r. The variance is given by:

$$\begin{aligned} \text{var}(y_i) &= \text{var}(E[y_i|r]) + E[\text{var}(y_i|r)] \\ &= \text{var}(\mu_i) + E[\phi v(\mu_i] \\ &= \text{var}(g^{-1}[x_i^{'}\beta + z_i^{'}r]) + E[\phi v(g^{-1}(x_i^{'}\beta + z_i^{i}r))] \\ &= \frac{\phi^2}{a_{ij}} V(\frac{E(y_i)}{r}) \end{aligned} \tag{4.5}$$

where ϕ is a dispersion parameter and a_{ij} is a prior weight usually equal to 1.

The covariance

$$\begin{aligned} \text{Cov}(y_i, y_j) &= \text{cov}(E[y_i|r], E[y_j|r]) + E[\text{cov}(y_i, y_j|r)] \\ &= \text{cov}(g^{-1}[x_i^{'}\beta + z_i^{'}r], g^{-1}[x_j^{'}\beta + z_j^{'}r]) \end{aligned} \tag{4.6}$$

For example: GLMM for logistic regression is given by:

$$g(\pi_i) = x_i^{'}\beta + z_i^{'}r$$

4.3 Estimation in GLMM

Generalized linear mixed model likelihood function is expressed as an integral with respect to the random effects and does not have a closed form (Capanu et al, 2013). The argument is that likelihood function is difficult for computation when data is non-normal (Jiang, 2007). As a result, statisticians have proposed numerous approximation methods with different degrees of accuracy, complexity of implementation, and computation time (Capanu et al, 2013). These include pseudo and penalized quasi-likelihood (Schall, 1991, Wolfinger and O'Connell, 1993 and Breslow and Clayton, 1993), Laplace approximation (Raudenbush et al., 2000), Gaussian Hermite quadrature as well as Markov Chain Monte Carlo (Gilks et al., 1996). However, these techniques become increasingly more difficult to use when the random effects increases and the computations are intensive (Feddag and Mesbah, 2006). Likelihood methods such as the penalized quasi-likelihood approach have been shown to produce biased estimates especially for binary outcome clustered data with small cluster sizes (Capanu et al., 2013). Penalized Quasi-likelihood (PQL) estimates the fixed effects parameters by fitting a generalized linear model with the variance covariance matrix based on a linear mixed model fit. PQL estimates the variances and covariances by fitting a linear mixed model with unequal variance calculated from the previous generalized linear model fit. PQL is known to be flexible and is widely implemented. In addition, it computes the quasi-likelihood rather than a true likelihood. Another disadvantage is that PQL works poorly for poisson data when the mean per treatment combination is less than five, and for binomial data where the ex-

pected number of success and failures for each observation is less than five. For more details see (Breslow and Clayton 1993, Breslow and Lin, 1995).

Another form for estimation is the Laplace approximation. It is used to approximate the true GLMM likelihood rather than a quasi-likelihood by allowing the use of likelihood based inference (Raudenbush et al., 2000). Laplace approximation approximates the likelihood by assuming that the distribution of the likelihood is approximately normal, making the likelihood function quadratic on the log scale, and allowing the use of a second-order Taylor expansion. It is known to be more accurate and less flexible than the penalized quasi-likelihood (Bolker et al., 2008). Gaussian-Hermite Quadrature (GHQ) is also used to approximate the likelihood by picking optimal subdivisions at which to evaluate the integrand (Pinheiro and Chao, 2006). GHQ is defined as follow:

$$\int_{-\infty}^{+\infty} f(x)exp(-x^2)dx$$

and can be approximated by $\sum_{i=1}^{m} w_i f(x_i)$ and the constant weights w_i and evaluation points x_i can be calculated by polynomial of degree m.

Adaptive Gaussian Hermite Quadrature perform well for binary outcomes, but can be overwhelmed by problems with large numbers of random effects (Capanu et al., 2013). Adaptive Gaussian-Hermite quadrature incorporates information from an initial fit to increase precision. This method is also known to be accurate but slower than Laplace approximation, because its speed decreases rapidly with increasing numbers of the random effect. Usually Adaptive Gaussian Hermite is limited to 2-3 random-effects (Pinheiro and Chao, 2006). Another estimation technic is the Markov Chain Monte Carlo (MCMC). This method generates random samples from the distributions of parameters values for the fixed and random effect. MCMC is highly flexible, accurate and uses the arbitrary number of random effects, but on the

other hand it is known to be very slow and technically challenging (Bolker et al., 2008). MCMC technique produces a Markov chain and the sample path average of this Markov process for estimating characteristics of the distribution. For more details see Metropolis et al.(1953) and Hasting (1970).
As discussed by Wolfinger and O'Connell (1993), pseudo-likelihood is a linealization approach which proves better for a model with only one random effect. The parameter estimates are reached by solving iteratively the estimating equations, which upon convergence leads to new parameter estimates that are used to update the linealization. Pseudo-data are generated from the original data, and the likelihood function is approximated using Taylor's Series expansions available in the procedure. Pseudo quasi-likelihood and pseudo likelihood produce identical parameters estimates because the objective functions minimized by the two methods differ only by a constant. This linealization makes pseudo-likelihood method run much faster. However, this approach includes the absence of a true objective function for overall optimization process (Schabenberger, 2005). It is well known that the PQL estimates of the variance components are subject to bias especially for certain cases such as clustered binary with clusters of small size (Capanu et al., 2013).

4.4 Inference in GLMM

In linear model the approach of likelihood ratio (LR) and Wald-based tests are used for testing hypotheses about estimable and predictable functions (Walter, 2013). For GLMMs, LR tests are defined only when we estimate the model effects using integral approximation for example Laplace approximation and Gauss-Hermite quadrature. Likelihood ratio has limited value because of its computational intensity (Walter, 2013). As discussed by Bolker et al. (2008), LR is not recommended for testing the fixed effects in GLMM,

because it is unreliable for small to moderate sample size. Littell et al. (2006) reported that Wald Z and χ^2 are only appropriate for GLMMs without overdispersion. However Wald t and F tests account for the uncertainty in the estimate of overdispersion.

Model selection criteria are statistical instruments that serve the purpose of choosing a suitable statistical model from a candidate class. They are used to assign scores to each of the fitted candidate models in order to assist the data analyst in selecting a good model (Johnson and Omland, 2004). For GLMM model selection is critical and challenging, because it involves integration and computation of the variances components (Yang, 2007). The most commonly used method to select the fixed effects is to test for the significance of additional terms in ambled model (Pan and Lin, 2005). Other researchers proposed Akaike's Information Criterion (AIC) (Akaike, 1974), and the Bayesian Information Criterion (Schwarz, 1978). Other researchers proposed the model concordance correlation coefficient, which is a generalization of R^2 measure for linear model, to access the overall adequacy of the response function (Vonesh et al, 1996). In GLMM it is often difficult to compare the pseudo-values with the AIC statistics, because the AIC is based on the maximum likelihood computation for simple random sampling, whereas the pseudo-likelihood is based on complex survey design (Duijn et al., 2008). The use of AIC for mixed model in the analysis of clusters is inappropriate. AIC is appropriate for comparing model at subject specific level (Vaida and Branchard, 2005). When the model contains random effect the definition of AIC is not straightforward. The penalty term condition AIC is related to the effective degrees of freedom for a linear mixed model (Hodges and Sargent, 2001).

For model diagnostic in GLMM, residual plots are routinely used to assess the

adequacy of regression models for independent responses (Cook and Weisberg, 1994). It is often difficult to determine whether the observed pattern reflects model misspecification or random fluctuation (Pan and Lin, 2005). This kind of graphical assessment is even more challenging with dependent responses due to the correlativeness of the residuals. The developments of the model-checking procedures for GLMMs are challenging because of the existence of random effects which result in computational challenges (Pan and Lin, 2005). Furthermore, such plots are uninformative for binary data because all the points lie on one of two curves according to the two possible value of the response.

Least square means are also used for inference in GLMM. They estimate the marginal means over a balanced population. Least square means are computed on the link scale, that is, the scale on which the model effect is additive. For example, for logistic model for binary data, the least square means are predicted margins of the logits. Least square means computations are currently not supported for multinomial models. LSMEANS statement helps to compute least square means of the fixed effects (Schabenberger, 2005).

4.5 Results from fitting GLMM

The use of GLMMs can allow random effects to be properly specified and computed and errors can also be correlated. Also, GLMMs can allow the error terms to exhibit non-constant variability while allowing investigation into more than one source of variations (McCulloch and Searle, 2000). In this present study PROC GLIMMIX, procedure in SAS, is used to fit GLMM. As discussed by Schabenberger (2005), the magnitude of the variance component σ^2 depends on the metric of the random effects. If the solution of the variance component is near zero, then a rescaling of the random effect

data can help the optimization problem by removing the solution for the variance component away from the boundary of the parameter space. Moreover, Bolker et al. (2008) stated that zero variance component or singularity implies that the model is not well defined and convergence errors or zero variances could indicate insufficient data. (Schabenberger, 2005) also stated that asymptotic covariance matrix of the covariance parameter estimator $\hat{\theta}$ is computed based on the observed or expected Hessian matrix of the optimization procedure.

In this study, covariance parameter matrix as well as the asymptotic covariance estimates were used for the goodness of fit. We also used the -2 log likelihood and Pearson chi-square over its degree of freedom for assessing the model fit as well as checking of the goodness of fit. The likelihood is processed by Gaussian-Hermite quadrature which is one among methods used for estimation in GLMM. Tables 4.1 and 4.2, represent the covariance parameter estimates for the random effect Primary Sampling Unit (PSUT). We found that the estimate of the variance component of the random effect PSUT intercept is $\hat{V}ar_c$= .3223, and the estimated standard error of the estimated variance component is .1012, which is significantly different from zero. This suggests that the model is appropriate.

Table 4.1: Covariance parameter estimates

Covparm	**Subject**	**Estimate**	**Standard Error**
Intercept	PSUT	.3223	.1012

Table 4.3 represents statistics used to assess the model fit using the log Gaussian-Hermite quadrature and the Pearson chi-square test. The differ-

Table 4.2: Asymptotic covariance matrix

Covparm	**Subject**	**Std Error**
Intercept	PSUT (random effect)	.0103

ent values of quadrature points did not lead to considerable differences in the parameter estimates nor standards errors. The estimates and standards errors were similar with 5 and 10 points. The minus twice the residual log Gaussian-Hermite quadrature of the model is 2847.00 while the Pearson-chi square is 2999.22. The ratio of the Pearson chi-square statistics divided by its degree of freedom is .95. This ratio is close to 1, since $\phi = 1$; this indicates that the variability in the data has been properly modeled and indicates that there is no overdispersion. This indicates that the model is adequate.

Table 4.3: Goodness of fit

Criterion	Statistic
-2Log (smoking status)/random effect	2847.00
Pearson chi-square	2999.22
Pearson chi-square/DF	.95

Tables 4.4 and 4.5 show the parameter estimates as solutions for the fixed effects in the model as well as their standards errors. We found that a man from Nyanza province is 2.664 (p-value=.0045) times more likely to be a smoker as compared to a man from Nairobi province. Marital status reveals that a divorced man is 2.654 (p-value=.0001) times more likely to be a smoker as compared to an unmarried man. Ethnicity reveals that a man from Kisii

tribe and Kelinjin is 2.328 (p-value=.0001) and 2.122 (p-value=.0413) times respectively more likely to be smoker as compared to a man from Embu tribe. For size of household we found that a man from a household of 2 persons is 2.460 (p-value=.0034) times more likely to be a smoker as compared to a man from a household of 1 person. The results of age show that the odds of smoking for a man aged from 20 to 24, 25 to 29, 30 to 34, 35 to 39, 40 to 44, 45 to 49 and 50 to 54 years old is 5.380 (p-value=.0093), 4.669 (p-value=.0184), 6.419 (p-value=.0048), 7.300 (p-value=.0026), 8.743 (p-value=.0016), 6.620 (p-value=.0069) and 4.850 (p-value=.0340) times respectively that of the odds of smoking for a man aged from 15 to 19.

Interaction effects

Figure 4.5 displays the results for interaction effects. The results show that the odds of smoking for a Protestant man with primary education complete is 4.364 (p-value=.0096) times that of the odds of smoking for a non religion man with non education. The odds of smoking for a non religion man with primary education incomplete is .144 (p-value=.0045) times that of the odds of smoking for a non religion man with non education. The odds of smoking of a richest man aged from 50 to 54 years old is 5.467 (p-value=.0445) times that of the odds of smoking for a poorer or poorest man aged from 15 to 19 years old. Also, the odds of smoking for a man from a household of 3 persons who has access to mass media less than once a week is .048 (p-value=.0097) times that of the odds who does not have access to mass media.

4.5.1 Comparisons of least-square means

The analysis of means in PROC GLIMMIX does not compares the least squares means by contrasting them against each other as with all pairwise or

Table 4.4: Solution for the fixed effects

Effect	Estimate	Std.Error	t-value	P>t	EXP(B)	95% C.I of EXP(B)
Intercept	-3.9218	.7887	-4.97	<.0001		
Region (ref=Nairobi)						
Central	.4939	.3664	1.35	.1778	1.639	(1.023,3.487)
Coast	.4314	.3829	1.13	.2600	1.539	(.954,3.415)
Eastern	.4084	.3200	1.28	.2019	1.504	(.942,3.312)
Nyanza	.9798	.3445	2.84	.0045*	2.664	(1.245,4.542)
Rift Valley	-.2004	.3704	-.54	.0846	.818	(.345,1.978)
Western	.6391	.3704	1.73	.0846	1.895	(.945,3.487)
Northeastern	-0.1029	.3610	-.29	.7757	.9022	(.634,2.541)
Marital Status (ref=Not married)						
Married	.1819	.1675	1.09	.2777	1.200	(.678,2.154)
Living together	-.0171	.4094	-.04	.9667	.983	(.564,1.842)
Divorced or not living together	.9761	.2524	3.87	.0001*	2.654	(1.001,4.154)
Religion (ref=Roman catholic)						
Protestant	-.7714	.4044	-1.91	.0566	.462	(.214,1.201)
Muslim	-1.1272	.3906	-2.89	.0039*	.324	(.102,.945)
Non religion and others	.09615	.4497	.21	.8307	1.100	(.567,2.130)
Education (ref=Non education)						
Primary incomplete	-.5750	.5221	-1.10	.2709	.563	(.234,1.021)
Primary complete	-.3640	.5499	-.66	.2709	.695	(.324,1.214)
Secondary and higher	-.4785	.5810	.82	.4103	.619	(.334,1.124)
Age of the respondent (ref=15-19)						
20-24	1.6826	.6467	2.60	.0093*	5.380	(3.145,10.245)
25-29	1.5409	.6530	2.36	.0184*	4.669	(2.344,9.547)
30-34	1.8592	.6581	2.83	.0048*	6.419	(3.247,11.001)
35-39	1.9879	.6606	3.01	.0026 *	7.300	(4.285,13.879)
40-44	2.1683	.6864	3.16	.0016 *	8.743	(5.189,14.895)
45-49	1.8901	.6992	2.70	.0069*	6.620	(4.259,12.458)
50-54	1.5790	.7444	2.12	.0340*	4.850	(2.548,8.421)

*P-values <.05

Table 4.5: Solution for the fixed effects

Effect	Estimate	Std.Error	t-value	P>t	EXP(B)	95% C.I of EXP(B
Ethnicity (ref=Embu)						
Kelenjin	.7524	.3686	2.04	.0413*	2.122	(.1.025,4.560)
Kamba	-.6803	.2741	-2.48	.0131*	.506	(.235,1.230)
Kikuyu	.3554	.2154	1.65	.0990	1.427	(.852,2.814)
Kisii	.8450	.2013	4.20	<.0001*	2.328	(1.025,4.521)
Others	-.8209	.3726	-2.20	.0377*	.440	(.230,1.022)
Wealth Index (ref=poorer or poorest)						
Middle or rich	-0.3700	.7102	-.52	.6024	.691	(.313,1.235)
Richest	-.4815	.7031	-.68	.4935	.618	(.284,1.247)
Size of the household (ref=1 person)						
2 persons	.8998	.3069	2.93	.0034*	2.460	(1.025,4.158)
3 persons	.6160	.3300	1.87	.0621	1.852	(.954,3.894)
4 persons	.3265	.3345	.98	.3292	1.386	(.854,3.012)
5 persons	.4421	.3024	1.46	.1438	1.556	(.856,4.001)
6 persons and above	.3424	3415	1.00	.3161	1.408	(.845,4.132)
Access to mass media (ref=Not at all)						
Less than once a week	.6994	.3532	1.98	.0478*	2.013	(1.002,4.321)
At least once a weak	.5347	.2648	2.02	.0435*	1.707	(.901,3.546)
Almost every day	.4734	.2704	1.75	.0802	1.605	(.854, 3.054)

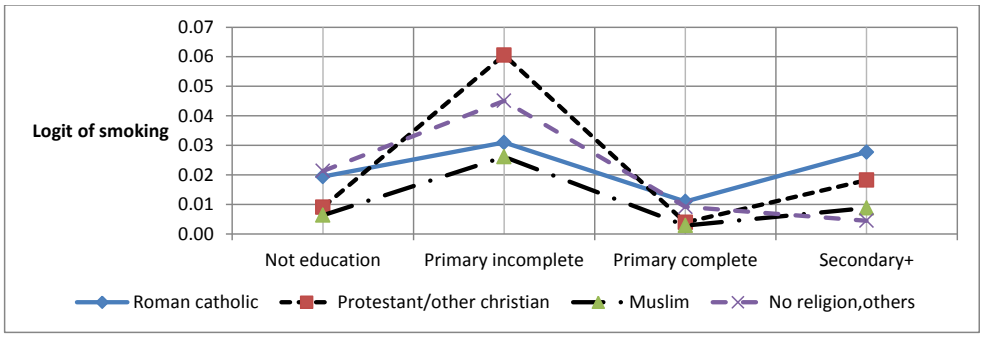
*P-values<.05

Table 4.6: Solution for the fixed effects

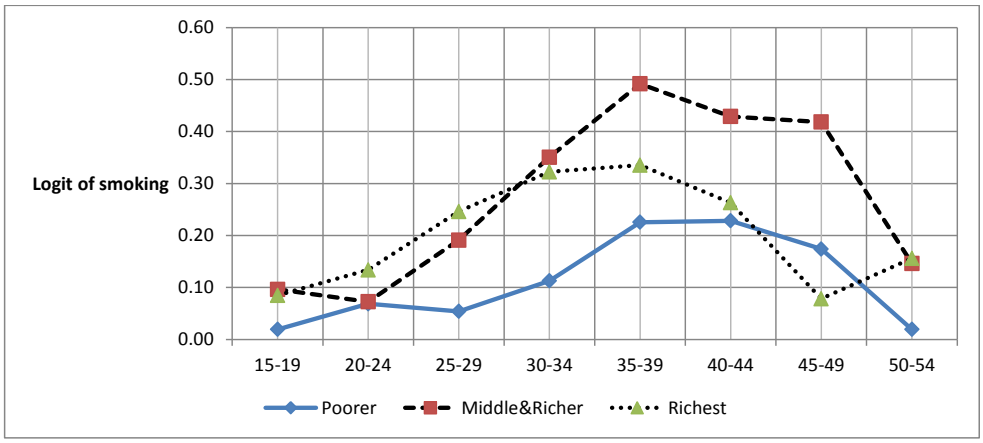
Effect	Estimate	Std.Error	t-value	P>t	EXP(B)	95% C.I of EXP(B)
Religion by Education (ref=Roman Catholic by non education)						
Protestant by primary incomplete	.3431	.7532	.46	.6488	1.409	(.856,2.536)
Protestant by primary complete	1.4733	.5686	2.59	.0096*	4.364	(2.365,8.248)
Protestant by secondary and higher	-.2670	.5934	-.45	.6528	.766	(.358,1.421)
Muslim by primary incomplete	-.0397	.7338	-.05	.9568	.961	(.563,1.463)
Muslim by primary complete	.9520	.5451	1.75	.0808	2.591	(1.203,5.258)
Muslim by secondary and higher	-.2286	.5699	-.40	.6884	.796	(.465,1.432)
Non religion by primary incomplete	-1.9371	.6813	-2.84	.0045*	.144	(.065,.563)
Non religion by primary complete	.2940	.6360	.46	.6439	1.342	(.786,2.312)
Non religion by secondary and higher	-.2826	.6466	-.44	.6621	.754	(.321,1.894)
Age by Wealth Index (ref=15-19 by poor or poorest)						
20-24 by middle or rich	.2019	.7721	.26	.7937	1.224	(.945, 2.031)
20-24 by richest	.5635	.7531	.75	.4544	1.757	(1.023,3.451)
25-29 by middle or rich	1.1014	.7706	1.43	.1530	3.010	(1.320,6.102)
25-29 by richest	1.2956	.7560	1.71	.0867	3.653	(.1.820,7.045)
30-34 by middle or rich	1.3174	.7620	1.73	.0839	3.734	(2.102,7.231)
30-34 by richest	1.0099	.7535	1.34	.1803	2.745	(1.001,4.532)
35-39 by middle or rich	1.0212	.7684	1.33	.1839	2.777	(1.546,4.875)
35-39 by richest	.6476	.7760	.83	.4041	1.911	(.864,3.148)
40-44 by middle or rich	1.2107	.7872	1.54	.1242	3.356	(1.235,7.234)
40-44 by richest	.7773	.7914	.98	.3261	2.156	(1.045,5.234)
45-49 by middle or rich	1.5389	.8129	1.89	.0584	4.660	(2.874,9.214)
45-49 by richest	.9807	.7988	1.23	.2196	2.666	(1.056,5.324)
50-54 by middle or rich	1.4583	.8620	1.69	.0908	4.299	(2.075,9.254)
50-54 by richest	1.6988	.8451	2.01	.0445*	5.467	(2.354,10.246)
Size by access to mass media (ref=1 person by not at all)						
2 persons by less than once a week	-.3116	.6157	-.51	.6128	.732	(.245,1.845)
2 persons by at least once a week	-.4392	.4181	-1.05	.2936	.6445	(.354,1.568)
2 persons by almost every day	-.5614	.4298	-1.31	.1916	.570	(.235,1.546)
3 persons by less than once a week	-3.0456	1.1760	-2.59	.0097*	.048	(.004,.532)
3 persons by at least once a week	-.2680	.4521	-.59	.5533	.765	(.125,1.547)
3 persons by almost every day	-.1148	.4532	-.25	.8001	.892	(.354,1.954)
4 persons by less than once a week	.-2.2324	1.1834	-1.89	.0593	.1073	(.003,.635)
4 persons by at least once a week	-.2534	.4314	.59	.5571	.776	(.235,1.352)
4 persons by almost every day	.06498	.4360	-.15	.8816	1.067	(.563,2.567)
5 persons by less than once a week	-.7308	.6390	1.14	.2529	.482	(.231,1.235)
5 persons by at least once a week	-.4340	.4005	-1.08	.2786	.648	(.341,1.537)
5 persons by almost every day	-4331	.4347	-1.00	.3192	.648	(.221.1.305)
6 persons and above by less than once a week	-.6743	.7998	.84	.3992	.510	(.213,1.324)
6 persons and above by at least once a week	-.1391	.4212	-.33	.7412	.870	(.316,2.012)
6 persons and above by almost every day	-.0748	.4381	-.17	.8643	.928	(.521,2.36)

*P-values <.05

Interaction between religion and education



Interaction between age and wealth index



Interaction between size and access to mass media

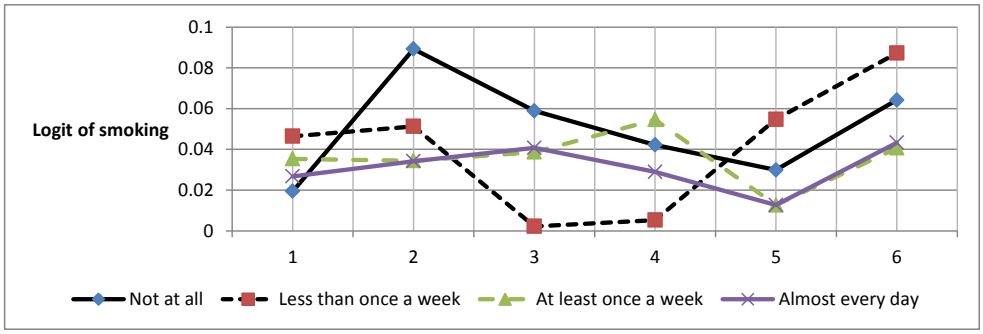


Figure 4.1: Interaction graphs

control differences. Instead, the least square means are compared against an average value. The significance level of the decision limits is determined from the ALPHA=Level in the LSMEANS statement. The reference is drawn at the average, then the vertical lines extend from this reference line upward or downward, depending on the magnitude of the least squares means compared to the reference value (SAS/STAT User's Guide, 2009). The dash upper and lower horizontal reference lines are upper and lower decision limits for tests against the control level. If the vertical line crosses the upper or lower decision limit, the corresponding least square mean is significantly different from the least square mean in the control group (SAS/STAT User's Guide, 2009).

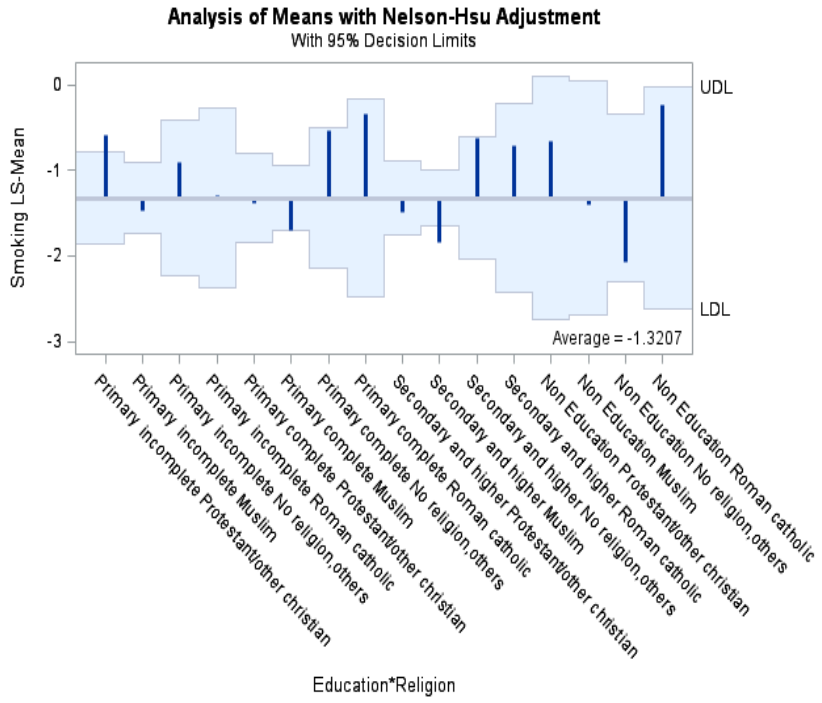


Figure 4.2: Least squares means for religion and education interaction effect

Figure 4.2 displays the interaction effect between the means for religion

interaction effects with education. The average religion by education interaction effect on the logit scale is -1.3207 (Figure 4.2). The differences between a Protestant man with primary education incomplete, a Muslim man with primary education complete and a Muslim man with non education are significantly different from the average, as shown by the vertical lines that cross the 95% decision limits in Figure 4.2. For a Protestant man with primary education incomplete, the least square means are greater that the average, whereas the least square for a Muslim man with primary education complete and a Muslim man with non education, are less than the average.

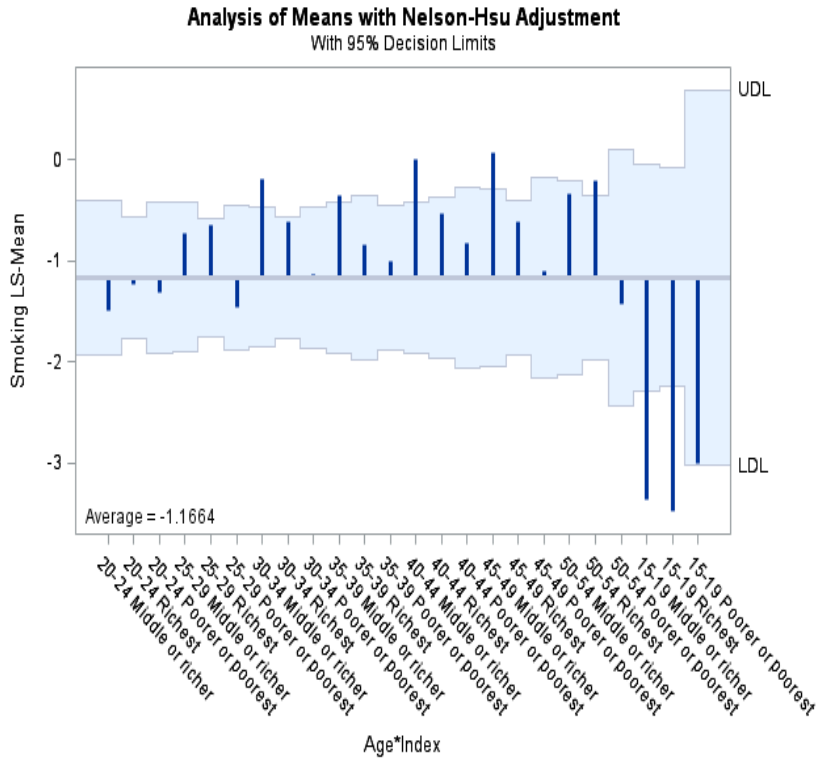


Figure 4.3: Least square means for age and wealth index interaction effect

Figure 4.5.1 displays the analysis of the mean for smoking status by age interaction effects with the wealth index. The average age by wealth index interaction effect on the logit scale is -1.1664. The differences between a

middle or rich man aged from 30 to 34 years old, a middle or rich man aged from 35 to 39 years old, a middle or rich man aged from 45 to 49 years, a richest man aged from 50 to 54 years old, a middle or rich man aged from 15 to 19 years old and a richest man aged from 15 to 19 years old are significant from the average as shown by the vertical lines that cross the 95% decision limits in Figure 4.5.1. The least square means are greater than the average for middle or rich man aged from 30 to 34 years old, a middle or rich man aged from 35 to 39 years old, a middle or rich man aged from 45 to 49 years old, a richest man aged from 50 to 54 years old, and less that the average for a middle or rich man aged from 15 to 19 years old and a richest man aged from 15 to 19 years old. Figure 4.5.1 displays the analysis of mean

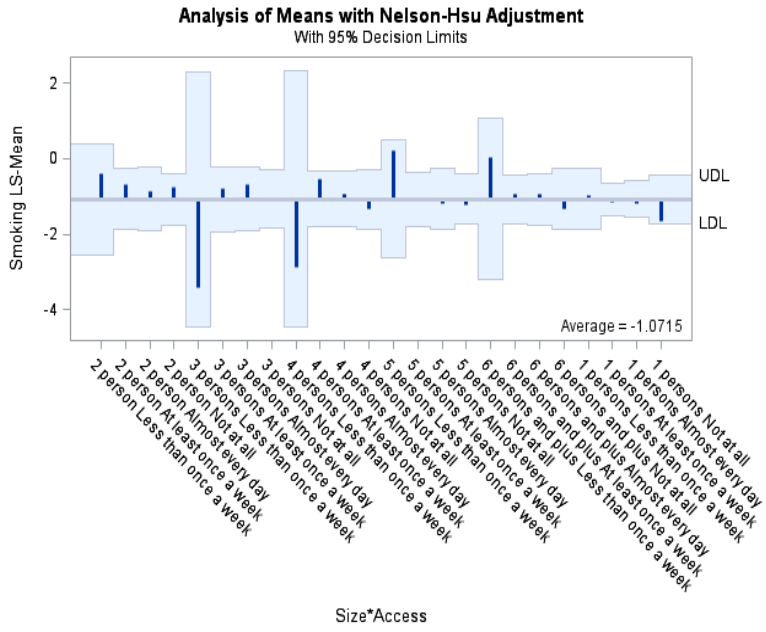


Figure 4.4: Least square means for size and access to mass media interaction effect

for smoking status interaction effect between size of household and access to

mass media. The average size by access to mass media on the logit scale is -1.0715. The differences of pairwise comparison on least square means for smoking status by size of household interaction with access to mass media, are not significant. It is also shown that all the vertical lines do not cross the 95% decision limit.

Chapter 5

MISSING DATA

5.1 Introduction

Missing data has been a serious issue in statistical studies. The argument is that missing data may reduce the precision of the calculated statistics because there is less information than originally planned (Hill, 1997). Missing data occurs when the values for one or more variables are missing from recorded observations (Mohan et al, 2013). In surveys statistics, a missing value is considered as a lack of response indicated by; do not know, refuse or unintelligible (Schafer and Graham, 2002). Statistical analysis in the presence of missing data has been an area of considerable interest, because ignoring the missing data often destroys the representativeness of the remaining sample and is likely to lead to biased parameter estimates. Pigott (2001) states that to avoid missing data is an optimal means to handle incomplete observations. In this section we deal with a monotone missing data and we apply statistical analysis such as last observation carried forward (LOCF) and multiple imputation (MI) which will help to handle this form of missingness.

5.2 Patterns of missing data

Data sets can be arranged in a rectangular or matrix form, where the rows correspond to observational units or participants and columns correspond to items or variables. With rectangular data, there are several important classes of overall missing data patterns (Schafer and Graham, 2002). Patterns of missing data as discussed by Little and Rubin (2002), illustrate the different ways the missing data are sorted using the missing data indicator matrix; let Y_{ij} denote $(n \times m)$ rectangular data set fully completed, having i^{th} row $y_i = (y_{i1}, ..., y_{im})$, where y_{ij} is the value of Y_j for subject i. We denote $Q = (q_{ij})$ the missing data matrix indicator, such that $q_{ij} = 1$ if y_{ij} is missing and $q_{ij} = 0$ if y_{ij} is present. The missing data pattern can then be defined by the matrix Q whose (i, j), the elements is Q_{ij}. The patterns of missing are illustrated in Figure 5.1. As shown from Figure 5.1(a) univariate non-response pattern describes univariate missing data where the missingness data is confined to a single variable. It appears in design experiments in which the response variable Y has missing values, but a set of factors $X_1, X_2, ..., X_m$ is fully observed.

However, for unit and item non-response pattern a set of variables $X_{j+1}, ..., X_m$ represents the missingness and the responses $Y_1, ..., Y_m$ are fully observed; the missing is present in more than one measured variable on the same set of subject. It can appear when questionnaires are partially completed because of refusal to answer, no contact or some other reason. To deal with this, weighting analysis and multiple imputation are useful. For example Figure 5.1(a) shows that for $m = 4$, j_2 is fully observed, hence j_3 and j_4 are missing. General pattern arises on particular items in questionnaires, and are represented by a haphazard pattern Figure 5.1(d). Pattern is said to be file matching if variables are never observed together. when we have datasets collected from

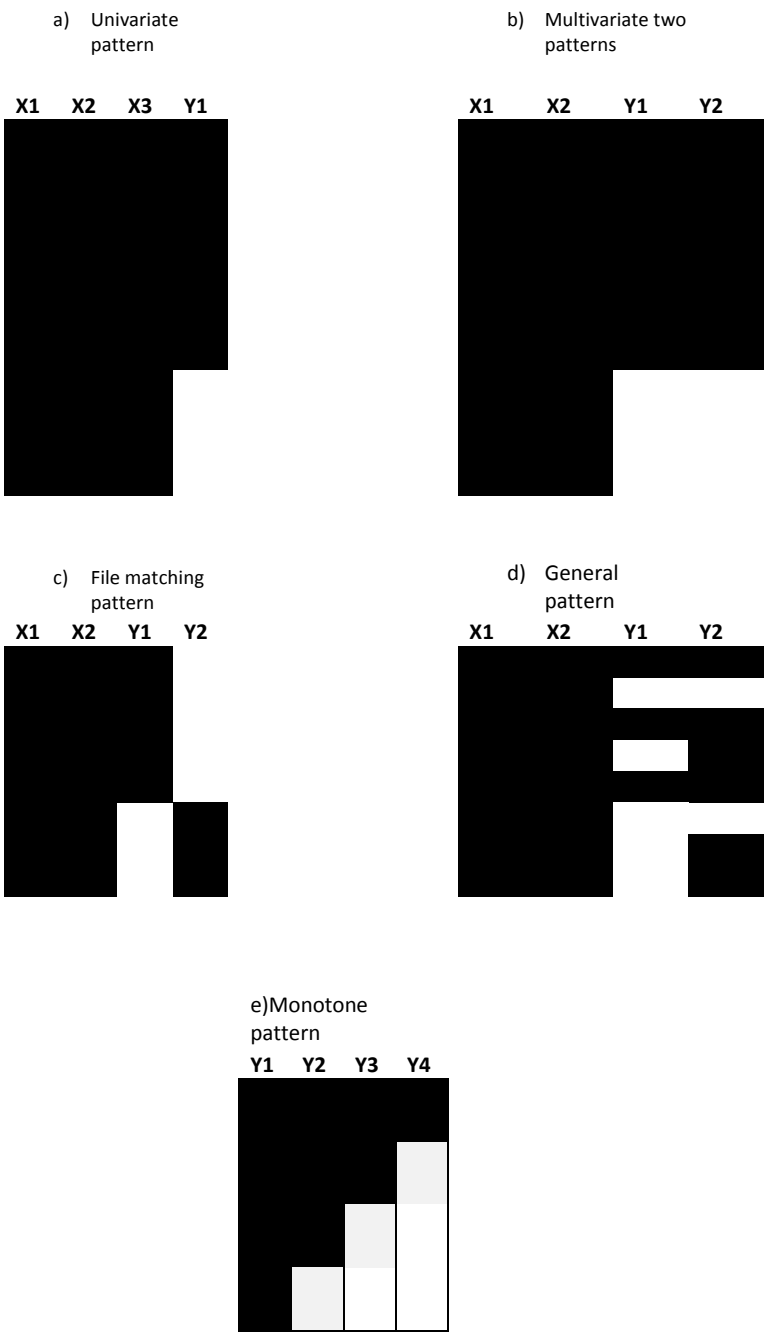


Figure 5.1: Missing data patterns

different studies and grouped together for analysis. Some of these datasets are partially observed and others are fully observed. As shown from Figure 5.1(c) file matching shows that dataset X_1 and X_2 are fully observed. On the other hand the dataset Y_1 is fully observed by the first data source but partially on the second, and Y_2 set is observed for the second data source but partially observed on the first data source. File matching fills the missingness of Y_1 and Y_2 values by matching units across files. Figure 5.1(e) also represents a monotone pattern. It exists when $Y_j, Y_{j+1}, ..., Y_m$ are missing while Y_{j-1} are fully observed ($j = 2$ and $m = 4$) as shown by the Figure 5.1(e). Monotone pattern usually appears in longitudinal studies where the subjects drop out prior to the end of the study, and non-returning. The strategy to handle this is to apply multiple imputation technique (Little and Rubin, 2002).

5.2.1 Notation

According to Little and Rubin (1987), the mechanisms of missing data are detailed as follow: Let $(Y) = (y_{ij})$ be an $n \times p$ data matrix $Y = (y_1, y_2, y_3, ...y_n)^T$, where $y_i = (y_{i1}, ..., y_{ip})^T$ is a random sample from a p-dimensional multivariate probability distribution $P(Y|\Phi, Y)$ governed by parameters Φ. We refer to the rows of Y as observations, given by y_i^T (i=1,2,...,n), and the column of Y as variables, denoted by $Y_j (j = 1, ...p)$. The missing indicator matrix $Q = (Q_{ij})$ is defined as follow:

$$Q_{ij} = \begin{cases} 1 & \text{if } \mathrm{Y}_{ij} \text{ is missing} \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

Defining $P(q_{ij} = 0|y_{ij}) = P(y_{ij}\text{observed}|y_{ij}) = P_{ij}$ the Q is subject to a probability distribution $P(Q|\xi, Y)$ governed by parameter ξ. Using this assumption, the joint probability of the response variable and the missingness

indicator variable will factorized as follow: $P(Y, Q|\xi, \Phi) = P(Y|\xi)P(Q|\Phi, Y)$, where $P(Y|\xi)$ is the marginal distribution of the response variable, and $P(Q|\Phi, Y)$ is the conditional distribution of the missingness given the response variable. Observed portion and missing portion, according to Little and Rubin (1987), will be represented by $Y_{obs} = (y_{ij}|q_{ij} = 0)$ and $Y_{miss} = (y_{ij}|q_{ij} = 1)$ respectively. For i^{th} observation, the observed portion and the missing portion of variable Y_j will noted as $y_i(obs)$ and $y_i(miss)$.

5.3 Mechanism of missing data

5.3.1 Missing Completely at Random (MCAR)

According to Little and Rubin (2002) missing data are classified into three categories based on the conditional distribution $P(Q|\Phi, Y)$. There is missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Missing is said to be completely at random if $P(Q|\Phi, (Y_{obs}, Y_{miss}) = f(Q|\Phi)$, for all Y. In this case the missingness does not depend on the values of Y, missing or observed. In surveys studies MCAR arise when the data are missing because of uncontrolled events in the course of data collection, which generates the nonresponse errors in recording in the data. This assumption occurs if the missing are missing by design. As an example, suppose $m = 2$, Y_1 describes age, and Y_2 is the income. If the probability that income is missing is the same for individuals, not influenced by their age or income, then the data are MCAR.

5.3.2 Missing at Random (MAR)

Missing data are said to be missing at random if $P(Q|\Phi, (Y_{obs}, Y_{miss})) = P(Q|\Phi, Y_{obs})$, for all Y_{miss}, Φ, where Y_{obs} denote the

observed components or entries of Y, and Y_{miss} denote the missing components. Data are said to be MAR if the missingness depends on the values prior to dropout, but not the value after dropout; this means :
$P(Q_i = j|y_{i1}, ..., y_{im}; \Phi) = P(Q_i = j|y_{i1}, ..., y_{i,m-1}, \Phi)$ for all $y_{i1}, ..., y_{im}$. Using the income age example MAR arises if the probability that income is missing varies according to age of respondent but invariant according to the income of the respondent with the same age.

5.3.3 Not Missing at Random (NMAR)

Missing data are said to be missing not at random if the missingness depends on observed and unobserved, that is: $P(Q|\Phi, (Y_{obs}, Y_{miss})) \neq P((Q|\phi), Y_{obs})$, where ϕ is unknown parameters, if the probability that Y_m depends on conditioning on other variables the data are NMAR. In the surveys NMAR arises when some questions are skipped for participants with certain characteristics. As example NMAR arises if the probability that income is recorded varies according to income for those with the same age.

5.4 Imputation methods for handling missing values

Imputation is a procedure where missing data are simulated (imputed) given the available information (Rubin, 1987). First, using a Multiple Imputation procedure provides a general purpose solution to statistical analysis with missing data and MI solution can provide more valid estimates of statistical quantities such as means, standard errors and regression coefficients. Two kinds of imputation that is, single imputation and multiple imputation are discussed (Rubin, 1987).

5.4.1 Single imputation

Single imputation is referred to as a direct replacement of subjects by new subjects from an identifiable source population based on observed subject characteristics. This approach may be feasible when the number of study variables is limited. Single imputation is represented through four forms i.e mean imputation, hot-deck imputation, last observation carried forward and regression imputation (Little and Rubin, 2002). Single imputation is easy to employ with a single value imputed for a missing value (Rubin, 1987). On the other hand, it does not reflect the extra-uncertainty and does not display variation due to missing data. In addition in single imputation the distributions of the surveys variables are compressed and relationships between variables may be distorted (Little and Rubin, 2002).

5.4.2 Mean imputation

Mean imputation replaces the missing data for a given feature (attribute) by the mean of all known values of that attribute in the class where the instance with missing attribute belongs. Let us consider that the value x_{ij} of the k^{th} class C_k, is missing, then it will be replaced by

$$\hat{x}_{ij} = \sum_{i:x_{ij} \in C_k} \frac{x_{ij}}{n_k}$$

where n_k represents the number of non-missing values in the j^{th} feature of the k^{th} class. The disadvantages of this method are that the sample size is overestimated, variance is underestimated, correlation is negatively biased and the distribution of new values is an incorrect representation of the population values because the shape of the distribution is distorted by adding values equal to the mean (Little and Rubin, 2002).

5.4.3 Hot-deck imputation

Hot-deck imputation is used when the missing attribute value is filled in with a value from an estimated distribution for the missing value from the current data. Hot-deck imputation is obtained in two stages: first, data are partitioned into clusters and second the data are replaced within a cluster by calculating the mean or mode of the attribute within a cluster. In random hot-deck, a missing value (the recipient) of an attribute is replaced by an observed value (the donor) or the attribute is chosen randomly. However for the cold-deck imputation method, the data source in which the imputed value is chosen must be different from the current data source (Little and Rubin, 2002).

5.4.4 Predicted mean

This method is used to create a predictive model for estimating the value that will used to substitute the missing value. Then missing data is used as the response attribute, and the remaining attributes are used as input the predictive model.

5.4.5 Last Observation Carried Forward (LOCF)

LOCF method is the last measured observation before the missing value is forwarded, and works best if there are a few missing values (Nakai, 2011). This method is not a good method for handling missing data. It is limited to less than 30% missing percentages unless σ^2 is large (Nakai, 2011). As discussed by Lane (2008) LOCF is considered as an imputation method that has been frequently used in the analysis of the response at an individual time-point. This method replaces the missing values at that time-point by the latest observed value. It is known to be biased and does not give a sta-

tistical estimate in the commonly understood sense. It is not consistent and is not an estimator for any population parameter. In some cases LOCF may provide an acceptably conservative approach and it is likely to be acceptable if measurements are expected to be relatively constant over time. LOCF provides no benefits, it creates unnecessary risk of generating bias or even false conclusions. Therefore this method is not a recommended method (Molnar et al., 2008).

5.5 Multiple imputation (MI)

Rubin and Little (1987, 2002) reported that the MI is a technique for non-response to replace each missing value by two or more plausible values so that each missing value is replaced by a vector of $m \geq 2$ imputed values. Multiple imputations will ensure that the uncertainty in the imputation is accounted for. In this method the missingness in a feature is filled in with values drawn randomly (with replacement) from a fitted distribution for that feature. Rubin (2004) describes the advantages for multiple imputation, namely, the ability to use complete data methods of analysis and the ability to incorporate the collector's knowledge. Normally there exist three extremely important advantages to multiple imputation over single imputation. First, when imputations are randomly drawn in an attempt to represent the distribution of the data, MI increases the efficiency of estimation. Second, when the multiple imputation represent repeated random draws under a model for non response, inferences are simply done by combining complete data inferences in a straightforward manner. The third distinct advantage is to generate repeated randomly drawn imputations under more than one model. It allows a straightforward study of the sensitivity of inferences to various models for non response simply by using complete-data methods repeatedly.

Rubin (2004) also states three obvious disadvantages of multiple imputation relative to single imputation. Firstly, multiple imputation needs more work as compared to single imputation. Secondly, more memory space is needed to store a multiply-imputed data set. Thirdly, more work is needed to analyze a multiple-imputed data set than a single-imputed data set (Rubin, 2004). Rubin (1987) highlights three steps of MI as follow:

- Sets of plausible values for missing observations are created that reflect uncertainty about the non-response model. Each of these sets of plausible values can be used to fill-in the missing values and create a complete data set.

- Each of these completed data sets can be analyzed using a complete data method.

- The m results are combined using methods that allow for uncertainty regarding the imputation to be taken into account.

As described by Rubin and Schenker (1986), the advantages of MI are that standard complete-data methods are used to analyze each complete data set; moreover, the ability to utilize data collector's knowledge in handling the missing values is not only retained but enhanced. MI allows data collectors to reflect their uncertainty as to which of values to impute. The disadvantages of MI include the time intensiveness imputing five to ten data sets, fitting models for each data set separately, and recombining the model results in one summary (Rubin and Schenker, 1986). MI can be implemented to generate the imputed data sets under three methods such as propensity score, regression method and MCMC. There also exists multivariate imputation techniques such as Joint modeling and Full conditional specification as discussed by Buuren (2007).

5.5.1 Combination for the inferences from imputed data sets

Given m imputed values, m different sets of the point and variance can be imputed for a parameter Q. Let $\hat{G}_k$ and $\hat{U}_k$ be the point and variance estimates from the k^{th} imputed data set, $k = 1, ..., m$. Then the point estimate for G for MI is the average of the m completed data estimates.

$$\overline{G} = \frac{1}{m}\sum_{k=1}^{m}\hat{G}_k \tag{5.2}$$

The uncertainly in $\overline{G}$ has two parts, the average within imputation variance

$$\overline{U} = \frac{1}{m}\sum_{k=1}^{m}U_k$$

and the between imputation variance

$$B = \frac{1}{m-1}\sum_{k=1}^{m}[\hat{G}_k - \overline{G}]^2 \tag{5.3}$$

The total variance is a modified sum of the two components such that

$$T = \overline{U} + (1 + m^{-1})B \tag{5.4}$$

and the $\sqrt{T = \overline{U} + (1 + m^{-1})B}$ is the overall standard error (Schafer and Graham, 2002).

5.5.2 Propensity Score

The propensity score is identified as a multiple imputation method to handle the missingnesss by generating the propensity scores for all observations, which is then used to estimate the probabilities that each observation is missing. Propensity score was developed by Rosenbaum and Rubin (1983) so that units with similar covariates could be matched between treatment groups and

unbiased estimates of the average treatment effect may be obtained. and Rosenbaum and Rubin (1983) defined a balancing score as a function $b(X)$ such that the covariates are independent of the response mechanism conditional on the balancing score. They state that the finest balancing score is the complete set of covariates X and the coarsest balancing score is the propensity score as the probability of the response given the covariates. The propensity score according to Rosenbaum and Rubin (1983), is the conditional probability of observing y_{ij}, given the previous history $y_{i1}, ..., y_{i,j-1}$ and is called the propensity score noted by sc_{ij}, denoted by and is defined as :

$$sc_{ij} = P(q_{ij} = 0|y_{i1}, ..., y_{i,j-1} \tag{5.5}$$

If the missing observations follow the monotone patten, the propensity score will be written as:

$$log(\frac{sc_{ij}}{1 - sc_{ij}}) = \alpha_0 + \alpha_1 y_{i1} + ... + \alpha_{j-1} y_{i,j-1} \tag{5.6}$$

where $\alpha_0, \alpha_1, ..., \alpha_{j-1}$ are the regression coefficients.

5.5.3 Data augmentation and Markov Chain Monte Carlo

Martin and Wing (1987) developed an algorithm using data augmentation referring to a scheme of augmenting the observed data so as to make it easy to analyze. The observed data y is augmented by the quantity z, which is referred as the latent data. Assuming that if y and z are both known, the problem is straightforward to analyze; that is the augmented data posterior $p(\omega|y, z)$ can be calculated. If, however, one can generate multiple values of z from the predictive distribution $p(z, y)$ (multiple imputation of z), then $p(\omega|y)$ can be approximately as the average of $p(\omega|y, z)$ over the imputed z's.

Then, the algorithm is based on calculating the posterior density

$$= \int_Z p(\omega|z,y)p(z|y)dz \tag{5.7}$$

where $p(\omega|y)$ denotes the posterior density of the parameter ω given the data y, $p(z|y)$ denotes the predictive density of the latent data z given y, and $p(\omega|z,y)$ denotes the conditional density of ω given the augmented data $x = (z,y)$. Firstly, the algorithm requires the generation of multiple values of latent data z by sampling from the conditional density of z given y. Secondly, the algorithm requires the computation (or sampling) of the posterior distribution of ω based on the augmentation data sets (process of posterior). The algorithm consists of iterating between the imputation and posterior steps; for more details see Martin and Wing (1987). MCMC is a set of tools used to make pseudo-random sample drawn from a target probability distribution (Gilks et al., 1996). A fundamental step in all Monte Carlo methods is to generate pseudo-random samples that follow a target probability distribution function (process of multiple imputation by Rubin (1986)). Let $P(Z) = f(Z)$, be the density of a random variable Z, $P(Z)$ is termed the target distribution. Instead of drawing directly from f, a sequence Z^1,Z^2,...,Z^m,... may be generated, where each variable in the sequence depends in some way on the proceeding ones and where the stationary distribution (that is the limiting marginal distribution of Z^m as $m \to \infty$), is the target of f. For sufficiently large m, Z^m is approximately a random drawn from. MCMC are those methods that allow samples of such pseudo-random quantities to be drawn from such target distribution (Thakuriah, 2010)

5.5.4 Regression method in MI

Yuan (2000) states that a regression method is a fitted model for each variable with missing data, using the remaining variables as covariates. Based on fitted model, a new regression is then drawn and is used to impute the missing values for each variable . The predicted value obtained from the regression replaces the missing value. Regression method for monotone pattern for variable Y_j with missing values, is defined as :

$$Y_j = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + ... + \alpha_k X_k \tag{5.8}$$

is fitted using observations with observed values for the variables Y_j and its covariates $X_1, X_2, ..., X_k$. The fitted model contains the regression parameters estimates $\hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1, ..., \hat{\alpha}_k)$ and the associated covariance matrix $\hat{\sigma}_j^2 V_j$, where V_j is the j^{th} diagonal element of the usual $(X'X)^{-1}$ matrix derived from the intercept and covariates $X_1, X_2, ..., X_k$.

Algorithm for generating imputed values for each imputation

To impute the missing values, the following steps are used:

1. New parameters $\alpha_* = (\alpha_{*0}, \alpha_{*1}, ..., \alpha_{*(k)})$ and $\hat{\sigma}_j^2$ are drawn from the posterior predictive distribution of the parameters. That is, they are simulated from $(\hat{\alpha}_0, \hat{\alpha}_1, ..., \hat{\alpha}_k, \hat{\sigma}^2$ and V_j. The variance is drawn as

$$\sigma_{*j}^2 = \hat{\sigma}_j^2 (n_j - k - 1)/g \tag{5.9}$$

   where g is a χ^2_{nj-k-1} random variate and n_j is the number of non-missing observations for Y_j. The regression coefficients are drawn as

$$\alpha_* = \hat{\alpha} + \sigma_{*j} V'_{hj} Z \tag{5.10}$$

   where $V'_{hj} V_{hj}$ and Z is a vector of k+1 independent random standard normal variates.

2. The missing values are then replaced by:

$$\alpha_{*0} + \alpha_{*1}x_1 + ... + \alpha_{*(k)}x_k + z_i\sigma_{*j}$$

where $x_1, ..., x_k$ are the values of the covariates and z_j is a simulated standard normal deviate.

5.6 Application

5.6.1 Methodology

In this section, we utilized a data set of 3465 of males observations from KHDS (2008-2009). We consider Y the response variable (smoking status) that is fully observed, and the explanatory variables X_1-region, X_2-marital status, X_3-religion, X_4-education, X_5-age, X_6-ethnicity, X_7-size of household and X_8-access to mass media have missing data. Recall that monotone pattern if formulated so that if X_j is missing the subsequent $X_{(j+1)}, ..., X_m$ are also missing (Little and Rubin, 2002). In order to achieve the third objective of this current research, we created a monotone pattern by dropping out 20% and 30% missingness while the variable Y was fully observed. Missing data are generated from the variables X_1 to X_8 assuming MAR mechanism (missing data are related to the observed data). The monotone data of missingness was created by randomly selecting values from the data matrix; the pattern was created by discarding values that lie next to the selected ones. This was done using SPSS (version 21). Afterwards, we filled in the missing values in $X_1, ...X_8$ to generate a complete data set using LOCF and MI. LOCF technique was used to replace the missing data by the last observed value and the dataset was analyzed using logistic regression and GLMM. For MI technique, we combined the results of five imputations using formulas that account for variation within and between imputed datasets using equations (5.2),(5.3)

and (5.4) in Satty and Mwambi (2012). We fitted the model using the full data set by imputation using PROC LOGISTIC and PROC GLIMMIX. We used PROC MIANALYZE to combine five logistics regression dataset in one logistic regression dataset and to combine the five GLMM datasets in one GLMM dataset. The results containing the parameters estimates, standards errors and the p-values from these models are presented in Table 5.1 and 5.2.

5.7 Results from fitting logistic regression and GLMM in the presence of missing data

5.7.1 Interpretation and comparison of the results of logistic regression applied in presence of missing data to the results of logistic regression found from original data (before creating missing values).

Table 5.1 shows the parameters estimates for logistic regression modeled after applying LOCF for 20% and 30% and for MI for 20% and 30% monotone missingness. In this section the parameters estimates help for interpretation and comparison of logistic model used for original data and logistic used after filling the missingness. Results show that for region, a man from Nyanza province is more likely to be a smoker. These results are different from what we found in original data where smoking was higher for a man from Eastern province as compared to Nairobi province. Results for marital status show that a divorced man or not living together with his partner is more likely to be a smoker as compared to an unmarried man. These results are the same as found from original data. For religion, the results show that a Protestant man is more likely to be a smoker as compared to a Catholic man. This is different from the original data where smoking was higher for non religion

man. For education, we found that a man with primary complete education is more likely to be a smoker. This is different from what we found in original data, where smoking was higher for a man with primary education incomplete as compared to a man with non-education. For age, the results show that a man aged from 40 to 44 years old is more likely to be a smoker. This is different from what we found in original data where smoking was higher for a man aged from 45 to 49 years old as compared to a man aged from 15 to 19 years old. Ethnicity reveals that smoking is higher for a man from Kisii and Kalenjin tribe. This is different from the original data where smoking was higher for a man from Kikuyu tribe as compared to a man from Embu tribe. For wealth index, we found that a poorer or poorest man is more likely to be a smoker. This is different from the original data where smoking was higher for a richest man as compared to a poorer or poorest man. For size of household, we found that a man from a household of 2 persons is more likely to be smoker. This is different from the original data where smoking was higher for a man from a household of 4 persons respectively as compared to a man from a household of 1 person. The results for interaction between religion and education show that a Muslim man with primary education complete is more likely to be a smoker as compared to a Catholic man with no education. The same results were found in original data. Interaction between age and wealth index show that a man classified in middle or rich and aged from 20 to 24 years old is more likely to be a smoker. This is different from the original data where smoking was higher for a richest man aged 30 to 34 years old as compared to a poorer or poorest man aged from 15 to 19 years old. Access to mass media reveals that a man from a household of 6 persons and above who has access to mass media less than once a week is more likely to be a smoker. This is different from the original data where

smoking was higher for a man from a household of 2 persons who has access to mass media less than once a week respectively as compared to a man from a household of 1 person who does not have access to mass media.

5.7.2 Interpretation and comparison of the results of GLMM applied in presence of missing data to the results of GLMM found from original data (before creating missingness).

The results for GLMM modeled after applying LOCF for 20%, 30% and MI for 20% and 30% monotone missingness show that for region, a man from Nyanza province is more likely to be a smoker as compared to a man from Nairobi province; these results are also found in the original data. For age, a man aged from 40 to 44 years old is more likely to be a smoker as compared to a man aged from 15 to 19 years old; these results are also found from original data. For ethnicity, a man from Kisii tribe is more likely to be a smoker as compared to a man from Embu tribe; the same results were found in original data. Access to mass media reveals that a man who has access to mass media at least once a week and less than once a week is more likely to be a smoker as compared to a man who does not have access to mass media. These results were also found in original data. Size of household reveals that smoking is higher for a man from a household of 2 persons as compared to a man from a household of 1 person; the same results were found in original data.

Interaction effect between religion and education reveals that a non-religion man with primary education incomplete is more likely to be smoker whereas from original data smoking was higher for a Protestant man with primary incomplete.

Interaction between age and wealth index reveals that a richest man aged from 50 to 54 is more likely to be a smoker as compared to a poorer or poorest man aged from 15 to 19 years old; these results are the same as we found in original data. Interaction between size and access to mass media reveals that a man from a household of 6 persons and above who has access to mass media less than once a week is more likely to be a smoker; the same results were found in the original data. From the above results we can state that the results found using logistic regression for 20% and 30% monotone missingness, are different from those found using logistic regression for original data. But the results found using GLMM for 20% and 30% monotone missingness are almost the the same.

5.7.3 Comparison between LOCF and MI

The results from Tables 5.1 and 5.2 also show the comparison of the results for the standard errors, and p-values for logistic regression and GLMM modeled after applying LOCF and MI for 20% and 30% monotone missingness respectively. Table 5.1 shows the results for comparison among the results of logistic regression modeled after applying LOCF for 20% and 30% as well the MI for 20% and 30% monotone missingness. We found that the standards errors of LOCF applied 20% missingness are smaller as compared to the standard errors for LOCF that applied 30% missingness. Hence LOCF for 20% is better than LOCF for 30% missingness. We have also found that MI for 30% performed better than MI for 20%. This was confirmed by Nakai (2011) who concluded that LOCF works better if there are few missing values, and MI performs better when the missing data is large (Nakai, 2011). In addition, we found that variables such as marital status, religion, education, age, ethnicity, wealth index, size of household, interaction between religion and

education, interaction between age and wealth index and size of household interaction and access to mass media provide the values of standard errors smaller for MI.

Table 5.2, shows the results for comparison of GLMM modeled after applying LOCF for 20% and 30% and MI for 20% and 30% monotone missingness. We have found that the standards errors of MI were smaller than the standards errors of LOCF for the variables marital status, religion, education, age, ethnicity, wealth index, size of household, interaction between religion and education, interaction between age and wealth index and size of household interaction and access to mass media. Several researchers, for example Baron et al.(2008), investigated the efficiency of the three methods of handling missing data, namely case-complete, LOCF and MI. They concluded that MI is the best method to minimize the bias (Baron et al., 2008). However the variables region and education did not perform well for MI, but worked well for LOCF.

Table 5.1: Results for logistic regression model in the presence of missing data

Logistic regression Model												
	LOCF20%			LOCF30%			MI20%			MI30%		
	Est.	Std.Error	sig.	Est.	Std.Error	sig.	Est.	Std.Error	sig.	Est.	Std.Error	Sig.
Intercept	-3.716	.364	.000	-3.722	.391	.000	-3.703	.376	000	-4.008	.403	.000
Region (ref=Nairobi)												
Central	.479	.279	.086	.552	.316	.081	.574	.303	.058	.611	.311	.050
Coast	.417	.292	.152	.482	.333	.148	.660	.316	.037	.659	.323	.041
Eastern	.439	.239	.066	.475	.275	.084	.548	.265	.038	.442	.272	.103
Nyanza	.987	.254	.000	.981	.301	.001	1.073	.285	.000	.885	.292	.002
Rift Valley	-.229	.284	.420	-.195	.325	.549	-.139	.309	.625	-.173	.312	.580
Western	.586	.280	.036	.663	.318	.037	.637	.303	.036	.590	.309	.056
Northeastern	-.158	.275	.566	-.096	.316	.762	-.111	.303	.715	-.142	.308	.646
Marital Status (ref=Not married)												
Married	.196	.150	.190	.228	.158	.149	.264	.152	.082	.283	.152	.062
Living together	-.036	.384	.925	.-018	.387	.963	.305	.309	.325	.-152	.309	.622
Divorced or not living together	.917	.233	.000	.959	.239	.000	.785	.225	.001	.968	.222	.000
Religion (ref=Roman catholic)												
Protestant	.250	.487	.607	.345	.475	.467	.184	.475	.699	.000	.462	.999
Muslim	-.242	.473	.608	.-324	.484	.503	.-059	.448	.895	.-525	.441	.234
Non religion and others	-1.164	.354	.001	-1.099	.372	.003	.-1.138	.352	.001	.-1.056	.365	.004
Education (ref=Non education)												
Primary incomplete	-.801	.531	.131	.-.819	.540	.129	.-.759	.529	.151	.-.812	.539	.132
Primary complete	.065	.551	.907	.369	.566	.515	.055	.551	.921	.290	.565	.607
Secondary and higher	-.322	.540	.551	-.262	.540	.627	-.339	.539	.530	.-326	.543	.549
Age of respondent (ref=15-19)												
20-24	1.779	.325	.000	1.741	.326	.000	1.638	.318	.000	.1.760	.327	.000
25-29	1.686	.332	.000	1.633	.336	.000	1.503	.326	.000	.1.632	.335	.000
30-34	1.832	.339	.000	1.798	.344	.000	1.719	.330	.000	1.871	.335	.000
35-39	2.085	.343	.000	2.030	.348	.000	1.899	.334	.000	.2.098	.345	.000
40-44	2.247	.384	.000	2.181	.389	.000	.1.979	.376	.000	2.180	.383	.000
45-49	1.955	.407	.000	1.829	.417	.000	.1.770	.395	.000	.2.039	.403	.000
50-54	1.641	.471	.001	1.584	.476	.001	1.506	.465	.001	1.718	.461	.000

Logistic regression Model												
	LOCF20%			LOCF30%			MI20%			MI30%		
	Est.	Std.Error	sig.	Est.	Std.Error	sig.	Est.	Std.Error	sig.	Est.	Std.Error	Sig.
Ethnicity (ref=Embu)												
Kelenjin	.502	.299	.094	.559	.300	.063	.424	.292	.146	.699	.293	.017
Kamba	-.5072	.235	.031	-.521	.222	.019	-.499	.224	.026	-.514	.228	.024
Kikuyu	.161	.190	.398	.220	.191	.250	.146	.184	.428	.241	.180	.182
Kisii	.702	.179	.000	.705	.179	.000	.556	.1700	.001	.575	.168	.001
Others	-.864	.344	.012	-.853	.345	.014	.-.780	.286	.006	-.434	.260	.095
Wealth Index (ref=Poorer or poorest)												
Middle or rich	-0.051	.684	.940	.-.056	.685	.935	-.249	.611	.683	.-.084	.685	.903
Richest	-.359	.688	.602	-.346	.688	.615	-.688	.623	.269	-.367	.688	.594
Size of household (ref=1 person)												
2 persons	.6246	.170	000	.562	.171	.001	726	.167	.000	598	169	.000
3 persons	.412	.174	.018	.413	.174	.018	.455	.172	008	.504	.169	.003
4 persons	.380	.164	.020	.355	.164	.030	.437	.158	.006	.368	.160	.021
5 persons	.205	.148	.166	.172	.153	.261	.170	.152	.263	.140	.151	352
6 persons and plus	.281	.151	.063	.265	.149	.076	.232	.147	.144	.259	.146	.076
Access to mass media (ref=Not at all)												
Less than once a week	.334	.231	.148	.367	.247	.138	.277	.388	.475	.447	.275	.104
At least once a week	.351	.143	.014	.347	.144	.016	.417	.247	.092	.386	.140	.006
Almost every day	.241	.135	.073	.240	.135	.074	.328	.253	.195	.195	.131	137

Logistic regression Model												
	LOCF20%			LOCF30%			MI20%			MI30%		
	Est.	Std.Error	sig.	Est.	Std.Error	sig.	Est.	Std.Error	sig.	Est.	Std.Error	Sig.
Religion by Education (ref=Roman Catholic by Non education)												
Protestant by primary incomplete	.043	.486	.929	.-.058	.472	.903	.121	.473	.798	.291	.460	.527
Protestant by primary complete	-.771	.487	.114	-.890	.464	.055	-.690	.475	.146	.-.505	.459	.271
Protestant by secondary and higher	-.943	.478	.049	-1.041	.464	.025	-.865	.465	.063	-.593	.452	.190
Muslim by primary incomplete	-.241	.453	.595	-.176	.466	.705	-.426	.429	.321	.038	.422	.928
Muslim by primary complete	-.543	.455	.233	-.482	.466	.301	-.715	.428	.095	.-.229	.421	.587
Muslim by secondary and higher	-.770	.453	.089	-.688	.464	.138	-.926	-.426	.030	-.393	.420	.349
Non religion by primary incomplete	1.153	.398	.004	.963	.416	.021	1.039	.394	.008	.927	.399	.020
Non religion by primary complete	1.488	.378	.000	1.436	.385	.000	1.316	.366	.000	1.322	.377	.001
Non religion by secondary and higher	1.320	.357	.000	1.298	.364	.000	1.253	.349	.000	1.376	.354	.000
Age by wealth index (ref=15-19 by Poorer or poorest)												
20-24 by middle or rich	-.091	.334	.786	-.078	.335	.815	-.020	.333	.953	-.081	.333	.807
20-24 by richest	.178	.284	.530	.206	.285	.470	.217	.282	.440	.228	.282	.418
25-29 by middle or rich	.778	.323	.016	.794	.327	.015	.841	.320	.009	.688	.327	.036
25-29 by richest	.797	.285	.005	.829	.286	.004	.892	.281	.002	.867	.282	.418
30-34 by middle or rich	1.073	.303	.000	1.054	.286	.000	1.062	.297	.000	.890	.303	.003
30-34 by richest	.653	.279	.019	.655	.281	.012	.593	.271	.029	.480	.277	.083
35-39 by middle or rich	.684	.319	.032	.729	.320	.023	.744	.313	.017	.524	.317	.098
35-39 by richest	.098	.328	.765	.075	.331	.821	.111	.319	.728	.001	.315	998
40-44 by middle or rich	.892	.356	.012	.898	.356	.012	1.014	.352	.004	.731	.351	.038
40-44 by richest	.396	.363	.275	.308	.369	.404	.382	.356	.284	.252	.357	.481
45-49 by middle or rich	1.261	.408	.002	1.415	.415	.001	1.341	.395	.001	1.004	.402	.013
45-49 by richest	.565	.382	.138	.639	.388	.099	.615	.368	.095	.334	.370	.387
50-54 by middle or rich	1.190	.492	.016	1.188	.496	.017	1.162	.483	.016	.922	.480	.055
50-54 by richest	1.259	.442	004	1.197	.460	.009	1.082	.454	.0171	1.139	.443	.010

Logistic regression Model												
	LOCF20%			LOCF30%			MI20%			MI30%		
	Est.	Std.Error	sig.	Est.	Std.Error	sig.	Est.	Std.Error	sig.	Est.	Std.Error	Sig.
Size by access to mass media (ref=1 persons by not at all)												
2 persons by less than once a week	.399	.692	.565	-.494	.755	.513	.491	.684	.473	-.274	.773	.723
2 persons by at least once a week	-.398	.396	.316	-.400	.397	.314	-.362	.390	353	-.434	.391	.267
2 persons by almost every day	-.530	.409	.195	-.530	.409	.195	-.363	.398	.362	-.579	.403	.152
3 persons by less than once a week	-2.462	1.161	.034	-2.157	.1.163	.064	-2.391	1.159	.039	-2.191	1.181	.064
3 persons by at least once a week	-.293	.427	.492	-.297	.427	.487	-.208	.423	.622	-.254	.413	.538
3 persons by almost every day	-.136	.429	.751	-.135	.430	.754	-.046	.422	.914	-.069	.416	.869
4 persons by less than once a week	-1.081	1.191	.364	-2.030	1.515	.999	.203	.863	.814	-2.0350	1.528	.999
4 persons by at least once a week	.256	.412	.488	.278	.412	.500	.395	.400	.324	.460	.404	.254
4 persons by almost every day	-.006	.417	.9897	-.002	.417	.997	-.098	.407	.809	.030	.409	.941
5 persons by less than once a week	.181	.554	.743	.567	.713	.426	.346	.710	.626	.487	.743	.512
5 persons by at least once a week	-.335	.380	.379	-.347	.381	.362	-.309	.376	.411	-.404	.372	.278
5 persons by almost every day	-.349	.414	.400	-.342	.414	.409	-.288	.403	.474	-.440	.403	.275
6 persons and above by less than once a week	.897	.827	.278	.737	.802	.358	.627	.787	.426	.490	.800	.540
6 persons and above by at least once a week	-033	.417	.936	-.026	.417	.951	.044	.407	.914	-.143	.405	725
6 persons and above by almost every day	-3.716	.364	.000	-3.722	.391	.000	-3.703	.376	.000	-4.008	.403	.000

Table 5.2: Results for generalized linear mixed model in the presence of missing data

GLMM												
	LOCF20%			LOCF30%			MI20%			MI30%		
	Est.	Std.Error	sig.	Est.	Std.Error	sig.	Est.	Std.Error	sig.	Est.	Std.Error	Sig.
Intercept	-3.596	.829	<.000	-3.633	.841	<.000	.-3.223	.776	<.000	-3.524	.835	<.000
Region (ref=Nairobi)												
Central	.689	.313	.028	.640	.339	.059	.723	.326	.027	.661	.333	.047
Coast	.598	.328	.068	.550	.355	.121	.778	.339	.022	.700	.345	.042
Eastern	.539	.269	.046	.489	.294	.097	.602	.285	.035	.484	.291	.097
Nyanza	1.098	.286	.000	.972	.321	.003	1.110	.306	.000	.868	.313	.006
Rift Valley	-.064	.319	.843	-.134	.346	.698	-.068	.329	.850	-.146	.332	.660
Western	.720	.314	.022	.710	.341	.037	.706	.326	.030	.631	.332	.057
Northeastern	-.008	.308	-.907	-.036	.335	.915	-.026	.323	.936	-.089	.328	.785
Marital Status (ref=Not married)												
Married	.264	.162	.103	.249	.166	.134	.291	.160	.069	.302	.159	.058
Living together	.009	.412	.982	.-022	.414	.959	.374	.330	.257	.-137	.329	.678
Divorced or not living together	1.032	.248	<.000	1.072	.250	<.000	.866	.237	.000	1.077	.234	<.000
Religion (ref=Roman catholic)												
Protestant	-1.020	.587	.082	-.016	.634	.981	-.290	.650	.655	-.363	.645	.574
Muslim	-.889	.647	.169	-.911	.661	.168	-.675	.624	.279	-1.117	.631	.077
Non religion and others	-1.638	.558	.003	-1.577	.574	.006	-1.645	.553	.003	.-1.553	.572	.007
Education (ref=Non education)												
Non education	reference											
Primary incomplete	-1.020	.587	.082	-1.004	.599	.094	-.975	.581	.093	-1.060	.599	.077
Primary complete	-.071	.604	.906	.145	.622	.816	-.054	.602	.928	.008	.620	.990
Secondary and higher	-.545	.592	.358	-.461	.598	.441	-.573	.588	.329	-.566	.600	.345
Age of respondent (ref=15-19)												
20-24	1.632	.645	.011	1.632	.644	.011	.1.273	.573	.026	.1.517	.636	.017
25-29	1.514	.649	.020	1.515	.649	.020	1.137	.578	.049	1.393	.640	.030
30-34	1.730	.655	.008	1.747	.655	.008	1.422	.582	.015	1.664	.644	.010
35-39	1.943	.657	.003	1.947	.657	.003	1.570	.585	.007	1.850	.647	.004
40-44	2.093	.682	.002	2.095	.683	.002	1.641	.612	.007	1.951	.670	.004
45-49	1.833	.696	.008	1.786	.699	.011	1.444	.625	.021	1.830	.682	.007
50-54	1.458	.741	.049	1.464	.741	.048	1.133	.677	.094	1.440	.723	.046

GLMM												
	LOCF20%			LOCF30%			MI20%			MI30%		
	Est.	Std.Error	sig.	Est.	Std.Error	sig.	Est.	Std.Error	sig.	Est.	Std.Error	Sig.
Ethnicity (ref=Embu)												
Kelenjin	.615	.321	.056	.674	.321	.036	.519	.313	.097	.796	.315	.012
Kamba	-.557	.254	.028	-.579	.245	.019	-.546	.242	.024	-.528	.245	.031
Kikuyu	.223	.202	.270	.274	.203	.177	.205	.195	.294	.308	.192	.110
Kisii	.768	.190	<.000	.771	.189	<.000	.595	.179	.001	.625	.176	.000
Others	-.872	.357	.015	-.831	.356	.020	-.747	.298	.012	-.391	.271	.150
Wealth Index (ref=Poorer or poorest)												
Middle and rich	-.091	.696	.896	-.096	.695	.890	-.309	.623	.620	-.169	.689	.807
Richest	-.356	.698	.610	-.350	.669	.616	-.720	.634	.256	-.406	.691	.557
Size of household (ref=1 person)												
2 persons	.856	.301	.005	.843	.301	.005	.881	.294	.003	.891	.297	.003
3 persons	.606	.321	.060	.321	.060	.604	.575	.317	.070	.658	.315	.037
4 persons	.268	.329	.415	.270	.329	.412	.288	.322	.370	.204	.326	.532
5 persons	.419	.297	.157	.424	.297	.153	.375	.292	.200	.414	.293	.159
6 persons and plus	.269	.335	.421	.274	.334	.154	.180	.327	.582	.312	.326	.339
Access to mass media (ref=Not at all)												
Less than once a week	.364	.418	.054	.552	.361	.127	.267	.415	.520	.667	.456	.143
At least once a weak	.511	.265	.054	.512	.266	.054	.428	.262	.102	.540	.264	.041
Almost every day	.420	.271	.122	.417	.271	.124	.329	.268	.218	.388	.270	151

GLMM												
	LOCF20%			LOCF30%			MI20%			MI30%		
	Est.	Std.Error	sig.	Est.	Std.Error	sig.	Est.	Std.Error	sig.	Est.	Std.Error	Sig.
Religion by education (ref=Roman Catholic by non education)												
Roman Catholic by non-education	reference											
Protestant by primary incomplete	1.003	.777	.197	-1.850	.774	.272	1.040	.763	.173	1.213	.764	.113
Protestant by primary complete	-.727	.788	.357	1.123	.787	.154	-.680	.779	.383	-.656	.777	.399
Protestant by secondary and higher	-362	.765	.636	-.584	.758	.441	-.288	.754	.703	-.162	.749	.829
Muslim by primary incomplete	.834	.749	.266	.955	.767	.441	-.288	.754	.703	-.162	.749	.829
Muslim by primary complete	-.376	.762	.622	-.504	.782	.520	-.589	.742	.428	-.195	.752	.796
Muslim by secondary and higher	-.044	.745	.953	-.010	.757	.990	-.207	.724	.775	-.242	.730	.740
Non religion by primary incomplete	2.140	.714	.003	1.982	.734	.007	2.011	.708	.005	1.968	.725	.007
Non religion by primary complete	1.561	.717	.030	1.355	.735	.066	1.387	.710	.051	1.316	.731	.072
Non religion by secondary and higher	1.919	.687	.005	1.887	.697	.007	1.884	.680	.006	1.984	.692	.004
Age by wealth index (ref=15-19 by poorer or poorest)												
20-24 by middle or rich	-.087	.763	.910	-.095	.763	.901	.194	.697	.781	.056	.756	.941
20-24 by richest	.512	.749	.494	.523	.749	.485	.910	.689	.187	.640	.742	.388
25-29 by middle or rich	.735	.759	.333	.771	.760	.311	.981	.692	.156	.819	.754	.277
25-29 by richest	1.183	.752	.116	1.186	.751	.114	1.622	.691	.019	1.333	.743	.073
30-34 by riddle or rich	1.042	.752	.166	1.007	.745	.176	1.246	.683	.068	1.082	.743	.145
30-34 by richest	.908	.748	.225	.897	.748	.230	1.209	.686	.078	.877	.7395	.236
35-39 by middle or rich	.688	.752	.363	.731	.757	.334	.951	.689	.168	.740	.748	.323
35-39 by richest	.440	.770	.568	.405	.771	.599	.808	.709	.254	.452	.759	.551
40-44 by middle or rich	.887	.775	.252	.885	.748	.254	1.212	.708	.087	.910	.764	.234
40-44 by richest	.703	.788	.372	.597	.789	.450	1.013	.728	.164	.677	.777	.383
45-49 by middle or rich	1.171	.803	.145	1.329	.805	.099	1.488	.734	.043	1.156	.791	.144
45-49 by richest	.815	.795	.305	.873	.797	.274	1.488	.733	.087	.725	.782	.354
50-54 by middle or rich	1.072	.858	.212	1.107	.859	.198	1.271	.794	.110	1.061	.841	.207
50-54 by richest	1.547	.837	.065	1.571	.842	.062	1.822	.787	.021	1.659	.826	.045

GLMM												
	LOCF20%			LOCF30%			MI20%			MI30%		
	Est.	Std.Error	sig.	Est.	Std.Error	sig.	Est.	Std.Error	sig.	Est.	Std.Error	Sig.
Size by access to mass media (ref=1 persons by not at all)												
2 persons by less than once a week	.446	.713	.532	-.515	.788	.514	.541	.706	.443	-.301	.804	.708
2 persons by at least once a week	-.389	.410	.342	-.395	.410	.336	-.358	.403	.375	-.424	.406	.296
2 persons by almost every day	-.556	.424	.190	-.559	.424	.187	-.394	.413	.341	-.607	.419	.147
3 persons by less than once a week	-2.541	1.187	.0324	-2.283	1.992	.057	-2.465	1.186	.038	-2.356	1.222	.054
3 persons by at least once a week	-.286	.442	.518	-.291	.442	.510	-.201	.438	.646	-.232	.429	.589
3 persons by almost every day	-.160	.445	.720	-.161	.445	.718	-.064	.439	.885	-.086	.433	.843
4 persons by less than once a week	-1.051	1.222	.390	-5.680	1.406	.585	.244	.894	.785	-5.943	1.611	.609
4 persons by at least once a week	.342	.425	.421	.333	.425	.434	.449	.413	.277	.532	.418	.203
4 persons by almost every day	.020	.430	.964	.208	.430	.961	-.087	.421	.837	.070	.423	.869
5 persons by less than once a week	.217	.584	.710	.542	.747	.468	.273	.746	.714	.444	.779	.568
5 persons by at least once a week	-.365	.393	.353	-.380	.393	.334	-.338	.389	.385	-.426	.386	.269
5 persons by almost every day	-.371	.427	.385	-.369	.427	.388	-.330	.416	.428	-.466	.416	.263
6 persons and above by less than once a week	.890	.856	.299	.741	.834	.374	.639	.816	.434	.493	.828	.552
6 persons and above by at least once a week	-.077	.413	.853	-.078	.412	.850	.016	.404	.970	-.085	.402	.833
6 persons and above by almost every day	-.025	.430	.953	-.202	.430	.963	.054	.419	.897	-.131	.419	.755

Chapter 6

CONCLUSION AND RECOMMENDATIONS

The main objective of this study was to identify factors associated with smoking and hence lead to recommendation to the smoking policy in Kenya. The second objective was to use the appropriate statistical models applied to smoking status of Kenyan males that incorporate missing data as well as the comparison of the various statistical methods that handle monotone missing data by addressing their strengths and weaknesses. Statistical models for modeling smoking status such as logistic regression and a generalized linear mixed model were utilized in this study. A generalized linear model in the form of logistic regression and generalized linear mixed model were fitted for the smoking status as the dependent variable and explanatory variables of region, marital status, religion, age, wealth index, ethnicity, size of household and access to mass media and three two ways interactions such religion by education, age by wealth index and size of household by access to mass media. The results obtained using logistic regression showed that smoking is higher among men from the Eastern province; this is in line with KDHS (2008-2009) findings. For marital status, divorced men or not living together with their partners were found to be the most likely smokers. Religion revealed that

smoking is higher for men with non religion. The results for age showed that, smoking is higher among the men belonging to the age group 45 and 49 years; this is in line with with KDHS (2008-2009) findings. For ethnicity, smoking is higher for men from Kikuyu tribe. Concerning a size of household, smoking is higher for a man from a household of 4 persons. And the results for access to mass media showed that smoking is higher for men who do not have access to mass media. The results obtained using GLMM showed that smoking is higher for Nyanza province, the divorced men, non religion men, men aged from 40 to 44 years old, men from Kisii tribe, poor men, men from a household of 2 persons, and men who have access to mass media less than once a week. From those results we recommend the that Government of Kenya improves the existing policy of prohibiting smoking in all public places, tobacco advertising, tobacco sales to minors, tobacco production, sponsorship and tobacco use. Our recommendations to the Government of Kenya is that to improve the policy might be by educating men aged between 35 and 55 years and men from 5 and 4 members in household about the dangers of smoking. We recommend the Government of Kenya to educate and advertise men from Kikuyu tribe about the dangers of smoking. We also recommend that the Government of Kenya improves the existing policy by targeting the provinces with a higher prevalence of smoking such as Eastern province. From our results, we found that smoking is higher for the separated men; this suggests a special target for separated partners, maybe through seminars on smoking. From our findings we also found that smoking decreases with increasing access to mass media; this suggests that the Government of Kenya should improve the availability of access to mass media to every citizen of Kenya in these provinces particularly and advertise the dangers of smoking. We have found that smoking is higher for men with primary or

non education; this suggests that every man citizen of Kenya should at least study to secondary school level. In this research, we have also compared the statistical methods that handle the missingness after creating 20% and 30% monotone missingness and discussed their strengths and weaknesses. We have found that MI is the best method for handling missing data; as discussed before it provided a small bias than LOCF. This present research was limited to logistic regression and generalized linear mixed model. For future research we could expand this work to longitudinal data by measuring smoking status repeatedly after the new policy interventions and then accessing the effect in Kenyan males. We could apply generalized estimating equation(GEE) and joint modeling approach. In addition, we could utilize statistical techniques that handle the missing data if present, such as inverse probability weighting and MI-inverse probability weighting .

Bibliography

[1] Action on Smoking and Health (2007). *Research report, tobacco*: Global trends, www.ash.org.uk (accessed 30 December 2012).

[2] Agresti, A. (2002). *Categorical Data Analysis*, Second Edition: New York: John Wiley and Sons.

[3] Agresti, A. (2007). *An Introduction to Categorical Data Analysis*, Second Edition, New York: John Wiley and Sons.

[4] Akaike, H. (1973). *Information theory as an extension of the maximum likelihood principle*: In Petrov BN, and Caski (eds). Second International Symposium on Information Theory. Akademiai Kiado, Budapest; 267-281.

[5] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transanction on Automatic control* **AC-19**, 716-723.

[6] Archera, K. J, Lemeshow, S. and Hosmer, D. W. (2007). Goodness-of-fit test for logistic regression model when data are collected using a complex sampling design. *Computational Statistics and data Analysis* **51**, 4450-4464.

[7] Ashby, F. G (1992). *Multidimensional Models of Perception and Cognition*. Hillsadal, N. J:Erlbaum.

[8] Australian Government Department of Health and Ageing: Youth Tobacco Prevention Literature review. *EUREKA Strategic Research*. Project number 3032.

[9] Baron, G., Philippe, R., Adeline, S. and Bruno, G. (2008). Missing Data in Randomized Controlled Trial of Rheumatoid Arthritis with Radiographic Outcomes: A Simulation Study. *Arthritis and Rheumatism* **59**, 25-34.

[10] Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Pulsen, J.R.; Stevens, M. H. H. and White, J. S. S. (2008). Generalized Linear Mixed Model : a practical guide for ecology and evolution. *Trends in Ecology and Evolution* **24**, 127-135.

[11] Breslow, N. E. and Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Model. *Journal of the American Statistical Association* **88**, 9-25.

[12] Breslow, N. E. and Lin, X. (1995). Bias Correction in Genaralized Linear Mixed Model with a Single Component of Dispersion. *Biometrika* **82**, 81-91.

[13] Brown, L.D., Tony Cai, T. and Dasgupta, A. (2003). Interval Estimation in Exponential Families. *Statistica Sinica* **13**, 19-49.

[14] Burnham, K. P. and Anderson, D. R. (2004). *Model Selection and Multiple Inference: A practical Information-Theoretic Approach*. 2nd ed. New York: Springer-Verlag.

[15] Bush, J., White, M., Kai, J., Rankia, J. and Bhopal, R. (2003). Understanding influences on smoking in Bangladesh and Pakistan adults: community based, qualitative study. *MBJ* **326**, 962

[16] Buuren, S. V. (2007). Multiple Imputation of Discrete and Continuous Data by Full Conditional Specification. *Statistical Methods in Medical Research* **16**, 219-242.

[17] Capanu, M., Mithat Gonen and Begg, C. B. (2013). An assessement of estimation methods for generalized linear mixed models with binary outcomes. *Statistics in Medicine* **32**, 4550-4566.

[18] Centers for Disease Control and Prevention (2008-2010). *Health behaviors of adults: United States, 2005-2007.* Series 10, Data from Survey, No. 245. March. U.S. Department of Health and Human Servies (PHS) 2010-1573. http://www.cdc.gov/nchs/data/ series/sr_sr_245.pdf (accessesd 10 December 2013).

[19] Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics* **19**, 15-18.

[20] Colett, D. (2003). *Modeling binary data.* Second edition. Chapman and Hall/CRC.

[21] Cook, R. D. and Weisberg, S. (1994). Ares Plots for Generalized Linear Model. *Computational Statistics and Data Analysis* **17**, 303-315.

[22] Council of United States Government Health State Initiative (2007). *Comprehensive Smoking Prevention Programs.*

[23] De Leuuw, E. D., Hox, J. and Husman, M. (2003). Prevention and treatment of item non response. *Journal of Official Statistics* **19**, 153-176.

[24] Duijn, M. A. J. V., Gile, K. J. and Handcock, M. S. (2008). *A framework for Comparison of Maximum Pseudo-Likelihood and Maximum-*

Likelihood of Exponential Family Random Graph. Working No.74 Center for Statistics and the Social Sciences University of washington.

[25] Emmanuel, R., Abdurahman, A. and Adamson, S. M. (2007). Determinants of adolescent tobacco smoking in Addis Ababa, Ethiopia. *MBC Public Health* **7**, 176.

[26] Feddag, M. L. and Mesbah, M. (2006). Approximate Estimation in Generalized Linear Mixed Model with Application to the Rasch Model. *Computers and Mathematics with Applications* **51**, 269-278.

[27] Fred, C. P. and Jade, A. (2008). Changes in Youth, 1976-2002: A Time-Series Analysis. *Youth and Society* **39**, 453-479.

[28] Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

[29] Government of Kenya. Tobacco Control Act 2007. Government Printers. http:kenyalaw.org /kenyalaw/klr_app/frames.php (accessed 12 January 2013).

[30] Hasting, W. K.(1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika* **57**, 97-109.

[31] Health Canada (2009). *Smoking and your body: Health effects of Smoking*.

[32] Hill, M. A. (1997). *Missing Value Analysis 7.5*. SPSS Inc.

[33] Hodges, J. S. and Sargent, D. J. (2001). Counting degrees of freedom in hierarchical and other richly parameterized models. *Biometrika* **88**, 367-379.

[34] Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*, Second Edition: John Wiley and Sons, Inc.

[35] ICF International (2012). *Demographic and Health Survey Sampling and Household Listing Manual*. MEASURE DHS, Calverton, Maryland, U.S.A.

[36] Jeanne, A. M., Van Loon, Marja, T., Paul, G. S. and John, O. (2005). Determinants of smoking status: Cross-sectional Data on Smoking Initiation and Cessation.*European Journal of Public Health* **15**, 256-261.

[37] Jeffrey, D. (2011). *Tobacco Control in Africa; people, politics and policies*. Anthem Press.

[38] Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and and their Applications*: New York.

[39] Johnson, J. B. and Omland, K. S. (2004). Model Selection in Ecology and Evolution. *Trends in Ecology and Evolution* **19**, 101-108.

[40] Karen, H. S. and Mary, A. A. (1999). Factors that Influence Adolescents to Smoke. *The Journal of Consumers Affairs* **33**, 321-357.

[41] Kenya National Bureau of Statistics (KNBS) and ICF Macro (2010). *Kenya Demographic and Health Survey 2008-2009*. Calverton, Maryland: KNBS and ICF Macro.

[42] Kenya Tobacco Situaton Analysis and consortium (2008). *Situation Analysis of Tobacco Control in Kenya*.

[43] Kimosop, V., Wanth, C. and Wanyonyi, E. (2012). International Institute for Legislative Affairs: Monitoring Implementation of Tobacco

Control in Kenya. Identifying gaps and opportunities. Available at www.ilakenya.org (accessed 15 June 2013).

[44] Komu, P., Dimba, E. A, Macigo, F. G. and Ogwell, A. E. (2009). Cigarrettes smoking and Oral Health among health care students. *East Africa Medical Journal* **86**, 178-182.

[45] Kwamanga, D. H, Odhiambo, J. A. and Amukoye, E. I. (2003). Prevalence and risk factors of smoking among secondary school students in Nairobi. *East African Medical Journal* **80**, 207-212.

[46] Lane, P. (2008). Handling drop-out in longitudinal clinical trial: a comparison of LOCF and MMRM approaches. *Pharmaceutical Statistics* **7**, 93-106.

[47] Lefondre′, K., Michel, A., Jack, S. and Bernard, R. (2002). Modeling Smoking History: A comparison of different approaches. *American Journal of Epidemiology* **156**, 813-823.

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

[49] Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*, Second Edition. New York: John Wiley and Sons.

[50] Littell, C. R., Milliken, A. G., Stroup, W.W., Wolfinger, D. R. and Schabenberger, O. (2006). *SAS System for Mixed Model*. Second Edition, Carry, NC, SAS Institute Inc.

[51] Martin, A. T. and Wing, H. W.(1987). The Calculation of Posterior Distribution by Data Augmentation. *Journal of American Statistical Association* **82**, 528-550.

[52] McCullagh, P. and Nelder, J. A. (1952a). *Generalized linear Model*. Chapman and Hall. New York: John Wiley.

[53] McCullagh, P. and Nelder, J. A. (1989b). *Generalized linear Model*, Second Edition. Chapman and Hall. New York: John Wiley

[54] McCulloch, C. E. and Searle, S.R. (2001). *Generalized Linear and Mixed Models*. John Wiley and Sons, Inc., New York.

[55] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of state calculations by fast computing machine. *Journal of Chemical Phyiscs* **21**, 1087-1092.

[56] Ministry of Health, Kenya. Global Youth Tobacco Survey (2001). Kenya GYTS Fact Sheet.

http:www.cdc.gov/tobacco/global/GYTS/afro/2001/Kenya_report (accessed 12 December 2012).

[57] Ministry of Health, Kenya. Global Youth Tobacco Survey (2007). Kenya GYTS Fact Sheet.

http:www.afro.who.int/index.php? (accessed 10 December 2012).

[58] Mohan, K., Pearl, J. and Tian, J. (2013). Graphical Models for Inference with missing data. Advances in neural Information Processing System **26**, 1277-1285.

[59] Molnar, F. J., Futto, B. and Furgusson, D. (2008). Does Analysis using "last observation carried forward" inroduce bias in mandia research? *Canadian Medical Journal* **179**, 751-753.

[60] Nakai, M. (2011). Analysis of imputation methods for missing Data in AR(1) longitudinal dataset. *International Journal of Mathematical Analysis* **5**, 2217-2227.

[61] Nelder, J. A. and Wedderbun, R. W. M. (1972). Generalized Linear Model. *Journal of the Royal Statistical Society, Series A* **135**, 370-384.

[62] Ogwell, A. E, Astrom, A. N. and Haugejorden, O. (2003). Socio-demographic factors of pupils who use tobacco in randomly-selected primary schools in Nairobi Province, Kenya. *East Africa Medical Journal* **80**, 235-241.

[63] Oteyo, J. and Kariuki, M. (2009). Extent to which selected factors contribute to alcohol and cigarette use among public day secondary schools male students: A case of Nakuru municipality, Kenya.*Educational Research and Review* **4**, 327-333.

[64] Pan, Z. and Lin, D. Y. (2005). Goodness of Fit methods for generalized linear mixed model. *Biometrika* **61**, 1000-1009.

[65] Pan, J. X and Fang, K. (2002). *Growth curve models and statistical diagnostics*. New york: science press. Springer series in statistics.

[66] Pigott, T. D. (2001). A Review of Methods for Missing Data. *Education Research and Evaluation* **7**, 353-383.

[67] Pinheiro, J. C. and Chao, E. C. (2006). Efficient Laplacian and Adaptative Gaussian quadrature algorithms for multilevel generalined linear mixed model. *Journal of Computational and Graphical Statistics.* **15**, 58-81.

[68] Pregibon, D. (1981). Logistic regression diagnostic. *Annals of Statistics* **9**, 705-724.

[69] Raudenbush, S. W., Yang, M. L. and Yosef, M. (2000). Maximum Likelihood for Generalized Linear Model with Nested Random Effects via High-Order Multivariate Laplace Approximation. *Journal of Computation and Graphical Statistics* **9**, 141-157.

[70] Rosenbaum, P. R. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observation Studies for Causal Effects. *Biometrika* **170**, 41-45.

[71] Rubin, D. B. and Schenker, N. (1986). Multiple imputaion for interval estimation from simple random samples with ignorable nonresponse. *Journal of American Statistical Association* **81**, 366-374.

[72] Rubin, D. B. (1987). *Multiple Imputation for non-response in Surveys.* New York, John Wiley and Sons.

[73] Rubin, D. B. (2004). *Multiple Imputation for non-response in Surveys*, Fourth Edition. New York: John Wiley and Sons.

[74] Rutstein, S. O. and Rojas, G. (2006). *Guide to DHS Stastistcs.* Demographic and Health Survey Methodology, ORC Macro.

[75] SAS Institute Inc. 2009. *SAS/SAT© 9.2. User's Guide*, Second Edition. Cary, NC: SAS Institute Inc.

[76] Satty, A. and Mwambi, H. (2012). Imputation methods for estimating regression parameters under a monotone missing covariates pattern: A comparative analysis. *South Africa Statistal Journal.* **46**, 327-356.

[77] Schafer, J. L. and Graham, J. W. (2002). Missing data: Our View of Stae of the Art. *Psychological Methods* **7**, 147-177.

[78] Schall, R.(1991). Estimation in generalized linear models with random effects. *Biometrika* **78**, 719-727.

[79] Schabenberger, O. (2005). Introducing the GLIMMIX Procedure for Generalized Linear Mixed Models: *SUGI Proccedings*, **30**.

[80] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461-464.

[81] Smyth, G. K. (2003). *Pearson Goodness of fit statistic as score test statistic*. In: Goldstein, D. R. editor. Science and Statistics: A testschrift for Terry Speed Hayward, CA: Institute of mathematical statistics. p.115-226. IMS Lecture Notes Monograph Series, 40.

[82] Stokes, M. E., Davis, C. S. and Koch, G. G. (2000). *Categorical Data Analysis using the SAS System*. Second Edition, Cary, N C. SAS Institute Inc.

[83] Steven, G.H., Brady, T. W. and Patricia, A. B. (2010). *Applied Survey Data Analysis. Statistical in the Social and Behavioral Science Series*. Chapman and Hall, CRC.

[84] Stevens, J. P. (1984). Outliers and Influential Data Points in regression Analysis. *Psychological Bulletin* **95**, 334-344.

[85] Thakuriah, V. (2010). Evaluation of Alternative Data Imputation Strategies: A case study of motor carrier safety. Transportation letters. *The International Journal of Transportation research* **2**, 199-216.

[86] Tyas, S. L. and Pederson, L. L. (1998). Psychosocial factors related to adolescent smoking: a critical review of the literature. *Tobacco Control* **7**, 409-420.

[87] United States Department of Health and Human Services (2006). The health consequences of smoking: A report of the Surgeon General. Center for Disease Control and Prevention. National Center for Chronic Disease and Health Promotion. Office on Smoking and Health . Available on World Wide Web at http://www.surgeongeneral.gov/library (accessed 27 December 2012).

[88] United States Department of Health and Human Services (2012). The health consequences of smoking: A report of the Surgeon General. Center for Disease Control and Prevention. National Center for Chronic Disease and Health Promotion. Office on Smoking and Health. Available on World Wide Web at http://www.surgeongeneral.gov/library (accessed 6 June 2013).

[89] Vaida, F. and Blanchard, S. (2005). Conditional Akaie information for mixed-models models. *Biometrika* **92**, 351-371.

[90] Vittinghof, E. Glidden, D. V, Shiboski, S. C and McCulloch, C. E. (2005). *Regression Method in Biostatistics: Linear,Logistic,Survival and repeated measures Models*: Springer, New York.

[91] Vonesh, E.F., Chinchilli, V. M. and Pu., K. (1996). Goodness of fit in generalized linear nonlinear mixed-effects models. *Biometrika* **52**, 275-587.

[92] Walter, W. S. (2013). *Generalized Linear Mixed Model. Modern Concepts, Methods and Applications*. Chapman and Hall/CRC, Taylor and Francis Group.

[93] Wedel, M. and Kamakura, W. A. (2001). Factors models with (mixed) observed and latent variables in the exponential family. *Psychometrika* **66**, 515-530.

[94] Wolfinger, R. and O'Connell, M. (1993). Generalized Linear Mixed Model: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* **48**, 233-243.

[95] World Health Organisation (2009). Report on global tobacco epidemic: Implementing smoke-free environmenets. Geneva, World Health Organisation http://www.who.int/tobacco/mpower/2009/en/index.html (accessed 12 December 2012).

http://www.who.int/tobacco/communications/events/wntd/2006 (accessed 15 June 2013).

[96] World Health Organisation (2009). WHO Report on global tobacco epidemic: Warning about the danger of tobacco. Geneva, World Health Organisation. http://www.who.int/tobacco/global-report/2011/en/index.html (accessed 20 December 2012).

[97] Yang, H. (2007). *Variable Selection Procedures for Generalized Linear Mixed Models in Longitudinal Data Analysis*. PHD thesis, North Carolina State University.

[98] World health Organisation (2011). Gender, Health, Tobacco and Equity. Available at http: www.who.int (accessed 15 December 2012).

[99] World Health Organisation (2012). Joint National Capacity Assessment on the Implementation of the Effective Tobacco Control Policies in Kenya. Available at http://www.who.int/tobacco/en (accessed 30 December 2012).

[100] Yuan, Y. C.(2000). Multiple imputation for missing data: Concepts and New Development. *SUGI Proccedings*, pp. 267-25.