FLEXIBLE STATISTICAL MODELING OF DEATHS BY DIARRHOEA IN SOUTH AFRICA

BY

SIZWE VINCENT MBONA

Submitted in fulfilment of the academic

requirements for the degree of

MASTER OF SCIENCE

in

Statistics

in the

School of Statistics and Actuarial Science

University of KwaZulu-Natal

Pietermaritzburg

2013

Dedication

To my parents, Ntombi Florina Mbona & Msolwa Mbona, and my loving childrens, Mpume (may her soul rest in peace), Lizwi and Noxolo

Declaration

The research work described in this thesis was carried out in the School of Statistics and Actuarial Science, university of Kwa-Zulu-Natal, Pietermaritzburg Campus, under the supervision of Dr. Shaun Ramroop and Prof. Henry Mwambi

I, Sizwe Vincent Mbona, declare that this thesis is my own. It has not been submitted in any form for any degree or diploma to any other University. Where use has been made of the work of others, it is duly acknowledged.

July, 2013.

--------------------------------- ------------------------------------

Mr Sizwe V. Mbona Date

---------------------------------- --------------------------------------

Dr. Shaun Ramroop Date

------------------------------------ ---------------------------------------

Prof. Henry Mwambi Date

Acknowledgement

I am deeply indebted to my supervisor Dr. Shaun Ramroop for his directional guidance, assistance, advice, patience and encouragement throughout the project. I also want to take this opportunity to extend my gratitude to other members of the Statistics department who have assisted me in my studies thus far.

My special thanks go to my parents, sisters and brothers; I am extremely thankful for your support, understanding, and encouragement in my quest for academic growth. Further thanks go to my friends and colleagues who were always there to offer the support and encouragement which made my study possible.

I wish to express my warm and sincere thanks to Professor Henry Mwambi for reading the first draft of the thesis and also for his unlimited encouragement, support and advice.

To the Almighty God, who makes everything possible, I give thanks.

Abstract

The purpose of this study is to investigate and understand data which are grouped into categories. Various statistical methods was studied for categorical binary responses to investigate the causes of death from diarrhoea in South Africa. Data collected included death type, sex, marital status, province of birth, province of death, place of death, province of residence, education status, smoking status and pregnancy status. The objective of this thesis is to investigate which of the above explanatory variables was most affected by diarrhoea in South Africa.

To achieve this objective, different sample survey data analysis techniques are investigated. This includes sketching bar graphs and using several statistical methods namely, logistic regression, surveylogistic, generalised linear model, generalised linear mixed model, and generalised additive model. In the selection of the fixed effects, a bar graph is applied to the response variable individual profile graphs. A logistic regression model is used to identify which of the explanatory variables are more affected by diarrhoea. Statistical applications are conducted in SAS (Statistical Analysis Software).

Hosmer and Lemeshow (2000) propose a statistic that they show, through simulation, is distributed as chi-square when there is no replication in any of the subpopulations. Due to the similarity of the Hosmer and Lemeshow test for logistic regression, Parzen and Lipsitz (1999) suggest using 10 risk score groups. Nevertheless, based on simulation results, May and Hosmer (2004) show that, for all samples or samples with a large percentage of censored observations, the test rejects the null hypothesis too often. They suggest that the number of groups be chosen such that G=integer of {maximum of 12 and minimum of 10}. Lemeshow et al. (2004) state that the observations are firstly sorted in increasing order of their estimated event probability.

Table of Contents

List of Tables

List of Figures

Chapter 1

1 Introduction

Diarrhoea is defined as the frequent excretion usually of liquid or unformed stools. Diarrhoea is characterised by a change in bowel habit, either an increase in the number of stools per day or an increase in the fluid content of the stool. It is termed acute if it lasts for less than two weeks or chronic if it lasts for more than four weeks. Diarrhoea can occur in virtually any person regardless of age and general state of health. Furthermore diarrhoea can range from mild discomfort to a severe and life threatening illness due to the risk of dehydration. Generally it is self-limiting and may not require any intervention. Intervention may be considered necessary by patient because of their beliefs and attitude towards normal bowel function (Hogue, 2000). The Ministry of Health`s Standard Treatment Guidelines (2004) defines diarrhoea as passing frequent, loose, watery stools three or more times in a day. An increase in stool water excretion above 150ml to 200ml every 24-hours is an objective parameter for acute diarrhoea (Hogue, 2000). The term diarrhoea means different things to different people. Many patients and doctors think of diarrhoea in terms of increased stools. Diarrhoea means having frequent stools of more than four in a day sometimes accompanied with pain or cramps, fever and or vomiting with nausea and chills. The most severe symptom in many patients is the urgency of defaecation, and faecal inconsistence is a common event in acute and chronic diarrhoeal illness (Haslett et al. 1999).

1.1 Causes of Diarrhoea

Diarrhoea can vary in severity. It presents itself as an abnormality in the digestive process, producing an increase in the wateriness, volume, or frequency of bowel movements over a given duration. Normally, when food passes through the colon, fluids are readily absorbed and only semi-solid stools remain. Diarrhoea is the reverse of this process and may occur due to a variety of causes. The most common causes of diarrhoea are:

- Bacterial infections
- Viral infections
- Parasitical functional bowel disorders
- Intestinal diseases
- Food intolerances and sensitivities
- Reactions to medicines

Diarrhoea can be treated by replacing lost fluids and electrolytes to prevent dehydration. Depending on the nature of the problem, medication may also be needed to stop the diarrhoea or to treat an infection.

1.2 Signs and Symptoms

The symptoms of diarrhoea (generally) includes a frequent need to defecate, weight loss, abdominal pain, cramping, bloating, vomiting, and a general feeling of being ill. Some infections that cause diarrhoea can also cause fever, chills, or bloody stools. Diarrhoea can cause dehydration and further loss of electrolytes through dehydration affects the amount of water in the body, muscle activity, and other important functions. Dehydration is very dangerous in children, older adults, and people with weakened immune systems and must be treated promptly to avoid serious health problems such as organ damage, shock, or a-comatose sleep-like state in which a person is not conscious. This means co-infection with a disease as HIV/AIDS can accelerate the effects of diarrhoea due to enhanced immune weakening. Incorrect mixing and dispensing of milk replacer by a malfunctioning automatic feeder has been reported as a cause of child/infant diarrhoea (Lane, 1987). More than two and a half million children under the age of five succumb to diarrhoea and dehydration each year (Kosek et al, 2003). Most cases of diarrhoea are of short duration, although they may recur multiple times. Reviews of research studies have determined that mothers tend to overstate the numbers of current or recent episodes of diarrhoea slightly, whereas they dramatically understate the number of events that occurred more than two or three days in the past (Boerma et al, 1991). Epidemiologic studies ideally should inquire about diarrhoea events occurring no more than three days prior to the study in order to be most accurate. Epidemiologists commonly divide diarrhoea into acute cases consisting of three or more loose watery stools in less than 24 hours, and persistent cases lasting for 14 days or more. Persistent diarrhoea causes more than half of all deaths from diarrhoea in many developed counties such as the United States, France, Germany, the United Kingdom, and South Korea (Victoria et al, 1993).

Bryce et al (2005) reported that of the estimated total 10.6 million deaths among children younger than five years worldwide, 42 percent occur in the World Health Organization (WHO) African region. However mortality rates among these children have declined globally from 146 per 1,000 in 1970 to 79 per 1,000 in 2003 (WHO 2005), and although this showed the most marked downward trend in diarrhoea the African region also showed the smallest reduction in mortality rates. During the 1990s, the decline of under-five mortality rates in 29 countries of the world stagnated, and in 14 countries rates went down but then increased again. Most of these countries are from the African region (WHO 2005). The global estimates of the number of deaths due to diarrhoea have shown a steady decline, from 4.6 million in the 1980s (Snyder and Merson, 1982) to 2.5 million in the year 2000 (Kosek, Bern, and Guerrant, 2003).

According to Child Health Research Specific project Report (1998), diarrhoea is one of the top causes of childhood mortality in Sub Sahara African and has been estimated to be responsible for 25 to 75% of all childhood illness in Africa. In addition, episodes of diarrhoea leads to about 14% of outpatient visits, 16% of hospital admissions and accounts for an average of 35 days of illness per year in children less than five years old. The report also stated that unlike the decline in mortality rates, diarrhoea incidence does not appear to have changed substantially over the last decade. A study “Review of Diarrhoeal Disease Cases Admitted to a Busy Referral Hospital in Ghana” (Baffoe-Bonnie et al. 1998) indicated that children less than 5 years of age make up 84% of all child admissions and 56.5% of them being infants below one year.

1.3 Statement of the problem

There has been a drastic increase death rate in South Africa due to diarrhoea during the period under study (1998-2005). For example, an increasing number of children, the elderly and even adults, are dying from diarrhoea in South Africa, making it the third leading cause of death in the country. A recent CSIR study has reported a steady increase in diarrhoeal death statistics in South Africa over the last 12 years with a significant increase in diarrhoea-related deaths for adults aged between 45 and 64.

Diarrhoea as underlying natural cause of death for all ages has increased from being the 10th leading cause in 1998 to become the third-leading underlying natural cause of death for two consecutive years in 2004 and 2005. While the highest death rates were recorded for the vulnerable age groups (children under five and the elderly). Researches were surprised to note an increase in adult deaths due to diarrhoea for each province. The trend was particularly visible for the age groups 45 to 64. According to CSIR researchers, diarrhoea statistics are an important indicator of the health of a community.

A report of WHO reveals that each child in at region has five episodes of diarrhoea per year and that 800,000 children die each year from diarrhoea illness and dehydration (WHO, 1996). The main causes of death among children under five year of age are acute respiratory infection (17%) and diarrhoeal disease (16%), and children infected with human immunodeficiency virus (HIV) have greater morbidity and mortality related to these conditions (WHO, 2008). One of the most dominant risk factors associated with childhood mortality and frequent cause of faltering growth of children is the existence of diarrhoeal episode, killing nearly two million children under age of five every year (Vesikari, 1997). WHO (1996) showed that the factors associated with diarrhoea were age of children, quality of water, availability of sanitation facilities, housing conditions, level of education, economic status of households, place of residence, feeding practices and personal or domestic hygiene to name a few. South African Demographic and Health Survey (SADHS, 1998) also showed that substantial racial disparities in the prevalence of diarrhoea exist, with the black: white rate ratio amounting to 6:5 in 1998. This was further confirmed in a study by Choi (2003), reporting that one of the reasons for this disparity is that black and colored populations were forced, under apartheid period, to reside in poor townships and distant rural areas with no any piped water and sanitation services. This it is probable that racial disparity in access to such essential necessities contributed to higher level of the prevalence of diarrhoea amoung them.

Gouws et al. (2005) identified gender disparities favoring males in diarrhoeal treatment practices among the wealthier, non-slum city corporation households in Dhaka and Chittagong where the prevalence of diarrhoeal illness was lowest and the occurrence of prolonged diarrhoea was greatest within urban slum households affecting one-quarter of the children identified. The provinces are the lowest level of geographic information available for analysis in publicity released data, which precludes district level analysis. Population group information is available for about 75% of the records. However, there are very high levels of under recording of socio-demographic details, such as education level and occupation (Statistics South Africa, 2007).

Due to the information were have about diarrhoea and some of the factors that are related with on it to cause death. This gives me some good grounds to choose my exploratory variables in my study.

1.4 Objectives of the study

The study was conducted at several explanatory variables namely: sex, marital status, province of birth, province of death, place of death, province of resident, education status, smoking status and pregnancy status. The reason for including these explanatory variables in the thesis was to check how diarrhoea was related with them as previous studies shows that other explanatory variables such as age (as stated in the statement of problem above) were related with diarrhoea, so in this thesis age was not taken into account because previous studies has shown that diarrhoea was more affecting children under age of five year but this variable can be included in future research studies.

The aim of this study is therefore to check on how diarrhoea was affecting people in South Africa in 2007 by

- Determining whether diarrhoea was related to factors such as gender, marital status, smoking status and pregnancy status by fitting various statistical models.
- To identify and compare the mortality trends for the different provinces and places.
- Ascertain the advantages and disadvantages of the various statistical models that are fitted to the data.
- To make recommendations to the health policies of South Africa based on the results of the analysis.

1.5 Structure of the study

The study will be subdivided into six chapters. Following this introductory chapter, chapter two presents' techniques necessary to gain insight into data via exploratory data analysis, particularly by analysing data sets with graphs. Chapter three reviews the theoretical aspect of generalised linear model (GLM) and analysis of the data using PROC GENMOD in SAS. The fourth chapter presents a special case of GLM i.e. the Logistic regression model. Chapter five focuses on the theory of generalised linear mixed models (GLMM), analysis and interpretation of results. The last chapter (chapter six) concludes the study by providing key findings and suggestions for further research.

Chapter 2

Exploratory Data analysis

2.1 Description of data

South African people were surveyed by means of a series of interviewer-administered questionnaires conducted by fieldworkers of Stats SA in 2007. The survey was carried out in South Africa and the population was stratified by type of province. The surveyed variables were sex, marital status, province of birth, province of death, place of death, province of residence, pregnancy status of the deceased, smoking status of the deceased and education status of the deceased. The study population consisted of permanent South African residents only. The aim of the survey was to identify the causes of death in South Africa. There were many response variables (causes) like TB, diarrhoea, and cancer to name a few. This project focused on the causes of death from diarrhoea in South Africa. The dependent variable was diarrhoea and the explanatory variables were sex, marital status, province of birth, province of death, place of death, province of residence, pregnancy status, smoking status and education status.

The socio-demographic variables categories were encoded as sex (male and female), marital status (single, civil marriage, living as married, widowed , religious law marriage, divorced, and customary marriage); province of birth, death, and residence(Western Cape, Eastern Cape, Northern Cape, Free State, KwaZulu Natal, North West, Gauteng, Mpumalanga, and Limpopo); place of death (hospital, emergency room, nursing home, and home), pregnancy status (yes and no), smoking status (yes and no), and education status (none, Grade 1, Grade 2, Grade 3 up to Grade 12, and university).The explanatory variables (levels and codes) were summarised in the following table.

Table 2.1: The explanatory variables with numerical codes as used in the analysis.

Explanatory variable	**Survey Code**
Sex (2 levels)	1=male, 2=female
Marital status (7 levels)	1=single, 2=civil marriage, 3=living as married 4=widowed, 5=religious law marriage, 6=divorced, 7=customary marriage
Province of birth (9 levels)	1=Western cape, 2=Eastern cape, 3=Northern cape, 4=Free state, 5=KwaZulu Natal, 6=North West, 7=Gauteng, 8=Mpumalanga, 9=Limpopo
Province of death (9 levels)	1=Western Cape, 2=Eastern Cape, 3=Northern Cape, 4=Free State, 5=KwaZulu Natal, 6=North West, 7=Gauteng, 8=Mpumalanga, 9=Limpopo
Province of residence (9 levels)	1=Western Cape, 2=Eastern Cape, 3=Northern Cape, 4=Free State, 5=KwaZulu Natal, 6=North West, 7=Gauteng, 8=Mpumalanga, 9=Limpopo
Place of death (5 levels)	1=Hospital, 2=Emergency room, 3=dead on arrival, 4=Nursing home, 5=Home
Education (14 levels)	0=None, 1= Grade 1, 2= Grade 2, 3= Grade 3 4= Grade 4, 5= Grade 5, 6= Grade 6, 7= Grade 7 8= Grade 8, 9= Grade 9, 10= Grade 10, 11= Grade11, 12= Grade 12. 13=University
Smoking (2 levels)	1=Yes, 2=No
Pregnancy (2 levels)	1=Yes, 2=No

2.2 Exploratory data Analysis

Exploratory Data analysis or “EDA” is an approach to analysing data sets to summarise main characteristics in easy to understand form, often with visual graphs without having formulated a hypothesis. Any method of looking at data that does not include formal statistical modeling and inference falls under the term exploratory data analysis. It was promoted by John Tukey (1977) to encourage statisticians visually to examine their data sets, to formulate hypotheses that could be tested on new data-sets. EDA is a critical first step in analysing the data from an experiment (Tukey, 1977). Here are the main reasons we use EDA:

- Detection of mistakes
- Checking of assumptions, and
- Determining relationships among the explanatory variables.

2.2.1 Graphs

We consider pie charts and bar graphs as a form of graphical display of the data to check for visual differences between groups of the same variable, for example, the differences in provinces with respect to the dependent variable (diarrhoea) can give us insight and inference about the differences of occurrence. The dependent variable is a binary random variable with two levels, namely those “affected” are those who died of diarrhoea, and those “unaffected” are those who died of diseases other than diarrhoea. The independent variables like sex, marital status, province of birth, province of death, place of death, province of residence, pregnancy of the deceased, smoking status of the deceased, and education of the deceased were graphed on the x-axis. The death rate (proportion) was calculated as a percentage in all the graphs. The results are as follows:

Figure 2.1: Graph of Sex vs Diarrhoea

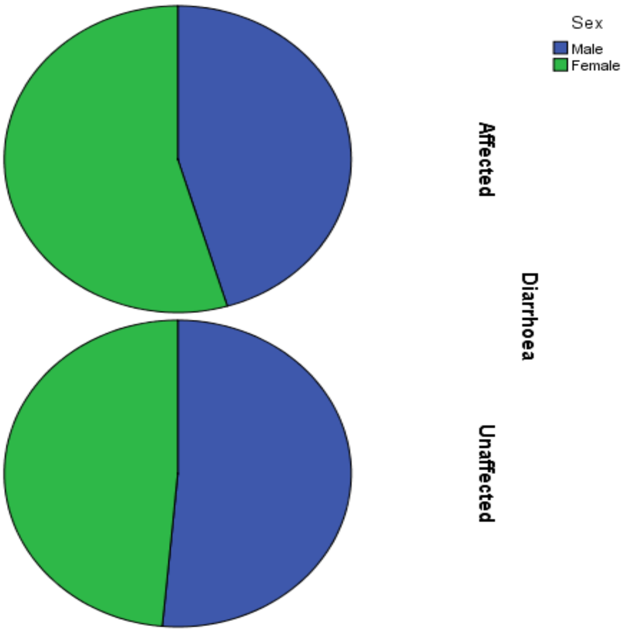

Figure 2.2: Graph of Marital status of deceased vs Diarrhoea

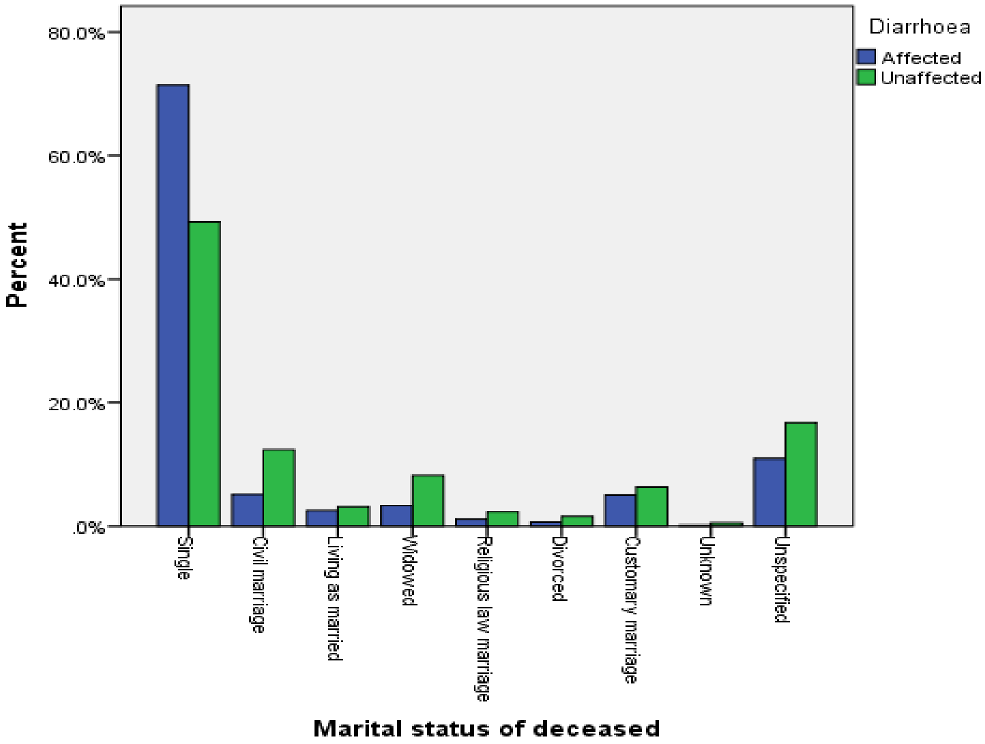

Figure 2.3: Graph of Province of Birth vs Diarrhoea

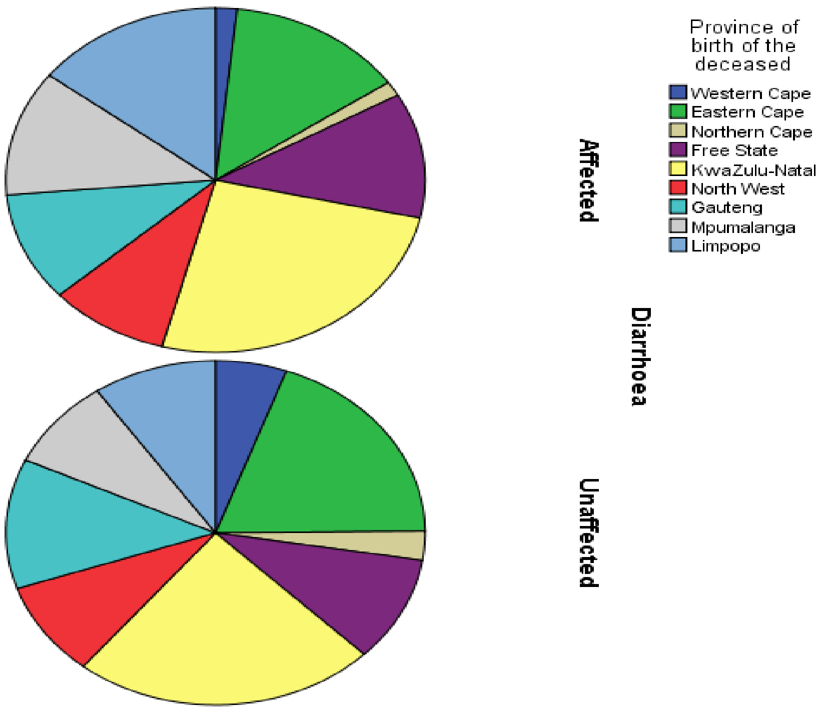

Figure 2.4: Graph of Province of Death vs Diarrhoea

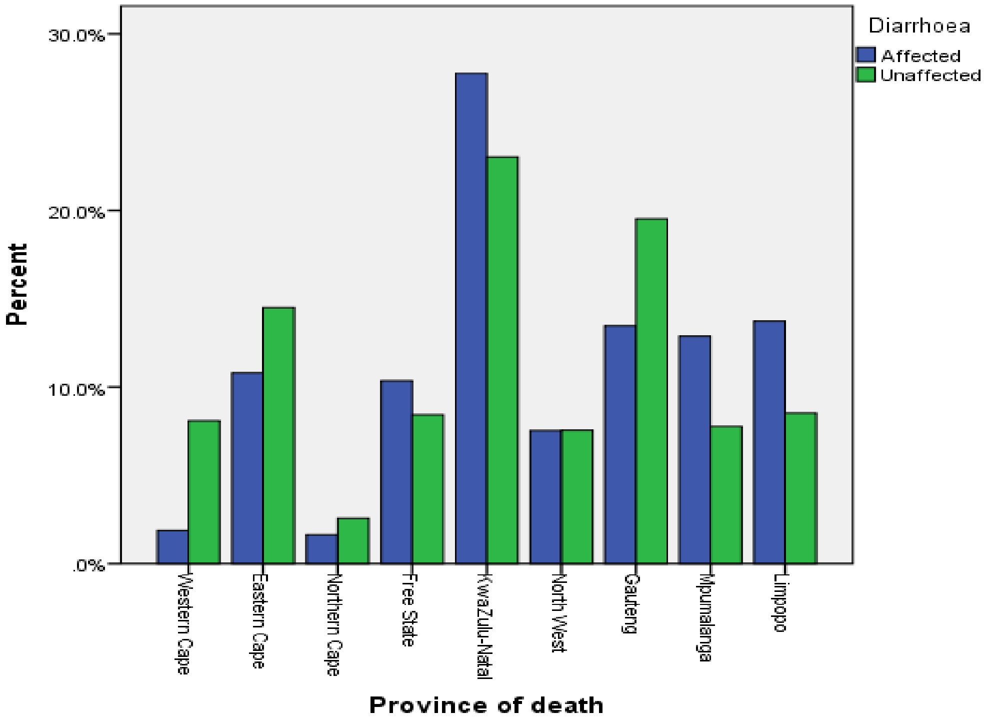

Figure 2.5: Graph of Province of Residence vs Diarrhoea

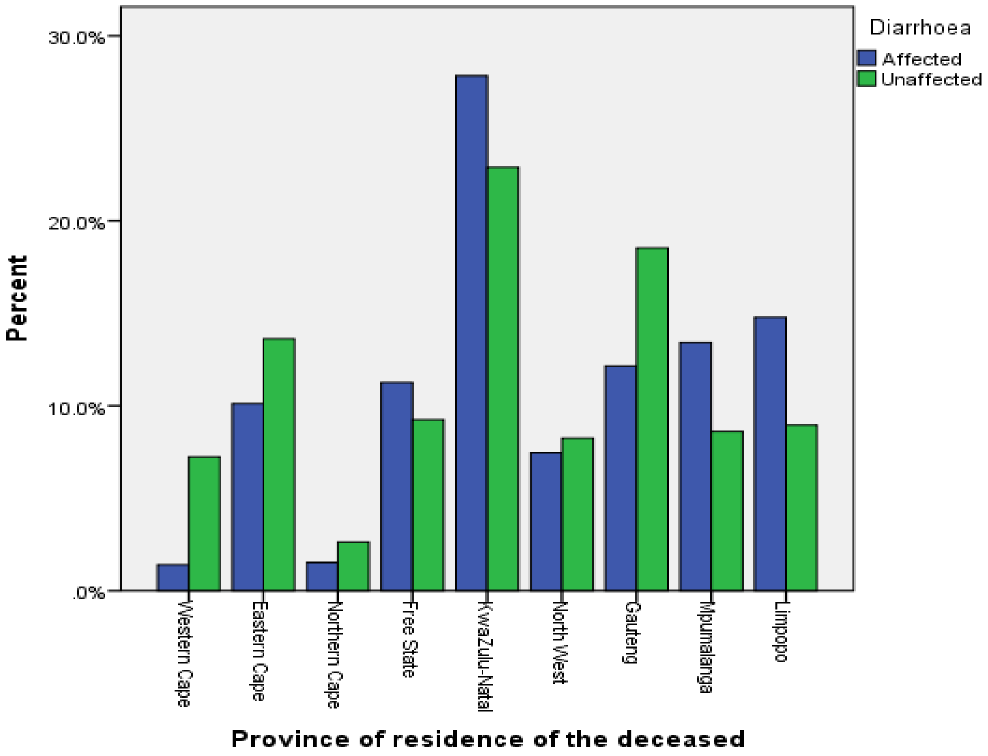

Figure 2.6: Graph of Place of Death vs Diarrhoea

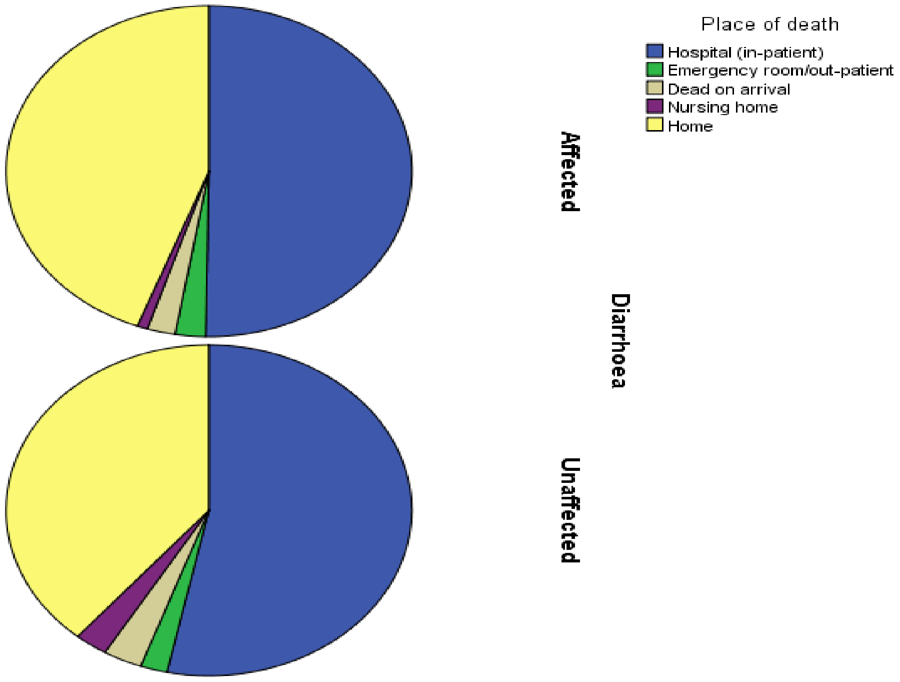

Figure 2.7: Graph of Education of the deceased vs Diarrhoea

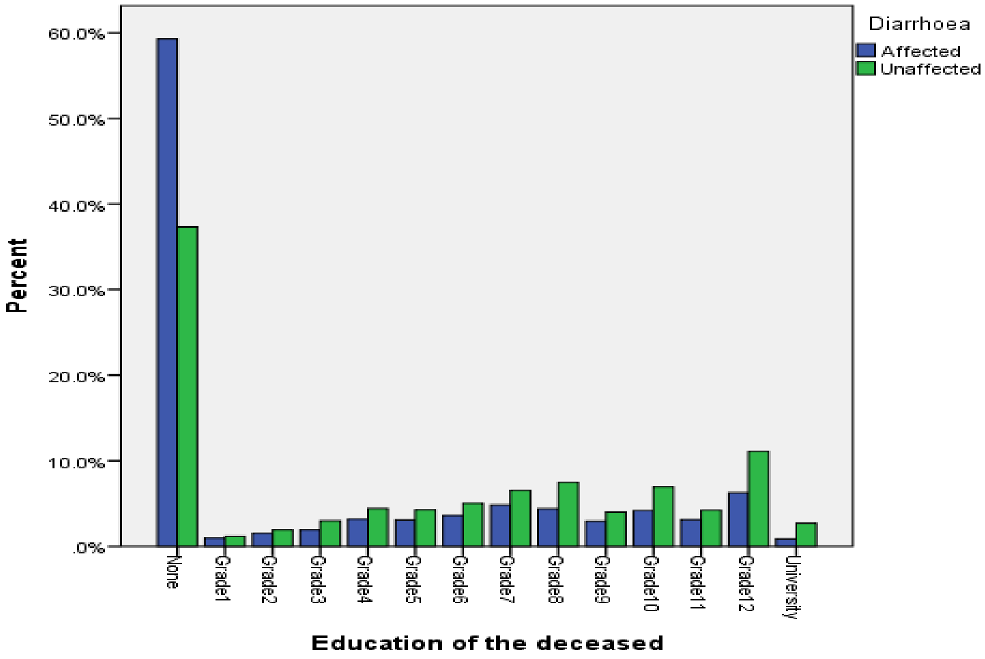

Figure 2.8: Graph of Smoking status of the deceased vs Diarrhoea

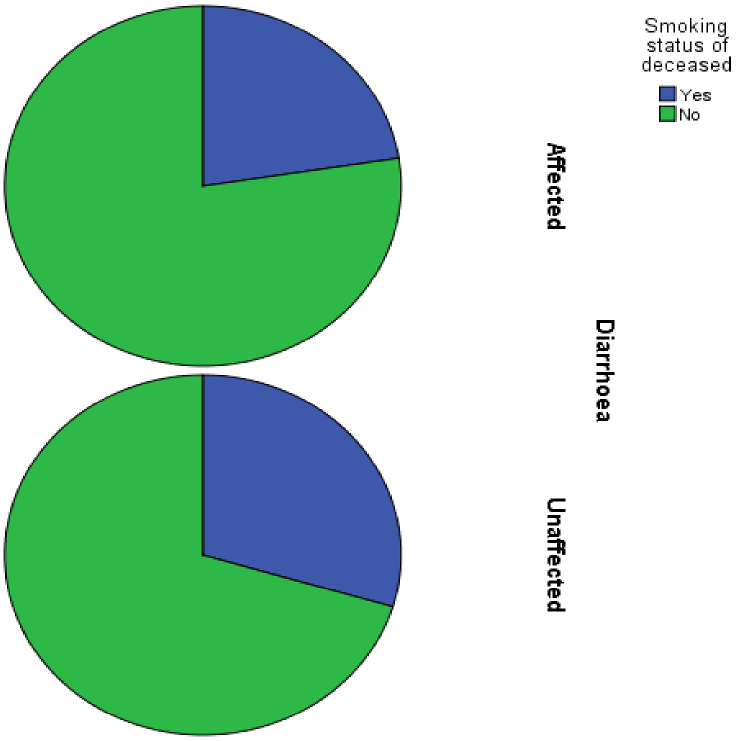

Figure 2.9: Graph of Pregnancy status of the deceased vs Diarrhoea

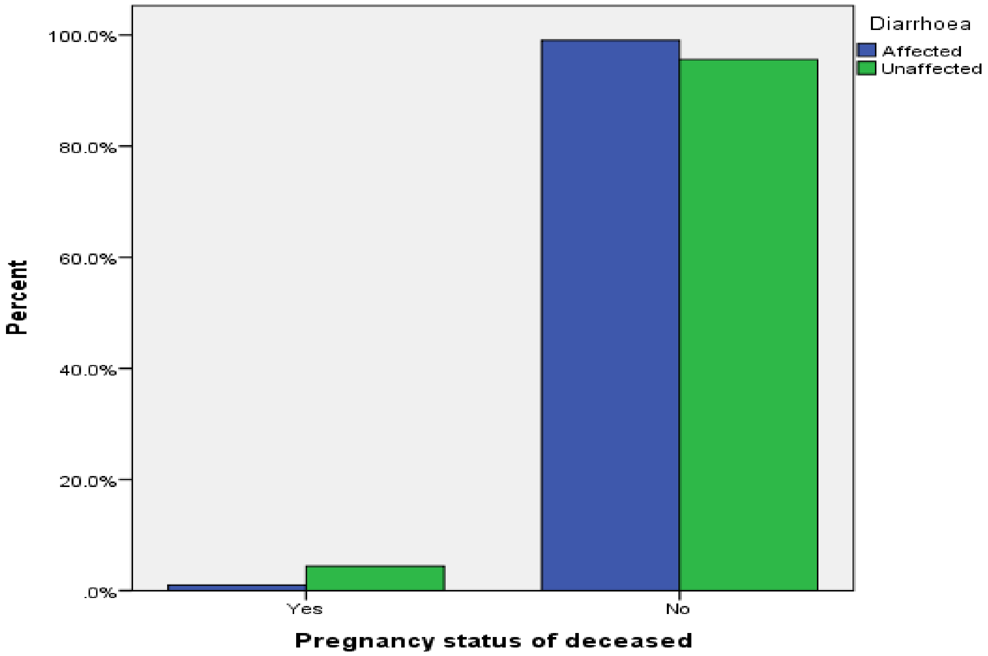

Figure 2.1: Shows that more females were affected by diarrhoea than males; more females died of diarrhoea than males.

Figure 2.2: This graph indicates that diarrhoea was affecting more people who were single with a death rate of about 71 percent. The proportion of death from diarrhoea from other marital status groups was very low (less than 20 percent). This clearly shows that single people were at a very high risk of dying from diarrhoea.

Figure 2.3: The graph shows that most of the people who had died because of diarrhoea were born in KwaZulu-Natal compared to other provinces. The proportion of those who died in KwaZulu-Natal province was large as compared to other provinces. People from other provinces such as the Eastern Cape, the Free State, the North West, Gauteng, Mpumalanga, and Limpopo were dying at a lower rate. The proportion of those dying of diarrhoea in other provinces compared to KZN, particularly the Western and Northern Cape was relatively very low.

Figure 2.4: This graph indicates that KwaZulu-Natal was the leading province for death from diarrhoea followed by Gauteng. The proportion of death in KwaZulu-Natal was roughly 27 percent. In the Eastern Cape and Free State, people were also dying (because of diarrhoea) at a proportion of about 11 percent.

Figure 2.5: This graph reveals that people who were living in KwaZulu-Natal were dying more than people living in other provinces and that the proportion was around 26 percent. People who were staying in Western and Northern Cape were dying at a lower proportion than the rest of the provinces. But the death rate in most of the provinces was roughly 10 percent.

Figure 2.6: This graph depicts that people were dying mostly in hospitals and home. In other places, the proportion was very low.

Figure 2.7: This graph shows that diarrhoea was affecting people who were not educated or who were still not yet at school in 2007, and that the rate of dying was about 59 percent.

Figure 2.8: This graph shows two groups of smoking status: smokers and non-smokers. The graph shows that people who were non-smokers were more affected by diarrhoea than smokers.

Figure 2.9: The graph shows that non-pregnant women are the one`s who were most affected by diarrhoea than those who were pregnant.

2.2.2 Cross tabulation of the data

Cross tabulation was considered as part of exploratory data analysis. The main reason for doing this was to compare the death rate for each independent variable with diarrhoea. The output from SPSS using cross tabulations yielded the following table:

Table 2.2: Descriptive results of Diarrhoea

Variables	Levels	Diarrhoea			
		no. of people affected	% no. of people affected	no. of people not infected	% no. of people unaffected
Sex	Male	12266	2	301872	49.2
	Female	14761	2.4	285172	46.4
	TOTAL	**27027**	**4.4**	**587044**	**95.6**
Marital status of deceased	Single	19341	3.8	289725	56.6
	Civil marriage	1392	0.3	72512	14.2
	Living as married	667	0.1	18238	3.6
	Widowed	892	0.2	47984	9.4
	Religious law married	285	0.1	13383	2.6
	Divorced	156	0	8823	1.7
	Customary married	1358	0.3	36743	7.2
	TOTAL	**24091**	**4.7**	**487408**	**95.3**
Province of birth of the deceased	Western Cape	352	0.1	23822	5.2
	Eastern Cape	2930	0.6	83730	18.4
	Northern Cape	293	0.1	11792	2.6
	Free State	2510	0.6	42961	9.5
	KwaZulu-Natal	5451	1.2	101352	22.3
	North West	1975	0.4	38508	8.5
	Gauteng	2187	0.5	52850	11.6
	Mpumalanga	2515	0.6	36776	8.1
	Limpopo	3104	0.7	41363	9.1
	TOTAL	**21317**	**4.7**	**433154**	**95.3**
Province of death	Western Cape	509	0.1	47582	7.7
	Eastern Cape	2920	0.5	85280	13.9

Table 2.2 (Continue)

Variables	Levels	Diarrhoea			
		no. of people affected	% no. of people affected	no. of people not infected	% no. of people unaffected
	Free State	2801	0.5	49540	8.1
	KwaZulu-Natal	7513	1.2	135348	22
	North West	2035	0.3	44296	7.2
	Gauteng	3646	0.6	114803	18.7
	Mpumalanga	3488	0.6	45680	7.4
	Limpopo	3714	0.6	50112	8.2
	TOTAL	**27070**	**4.4**	**587663**	**95.6**
Place of death	Hospital(in-patient)	11968	2.4	251994	50.8
	Emergency room/out-patient	575	0.1	10097	2
	Dead on arrival	534	0.1	14720	3
	Nursing home	199	0.3	12431	2.5
	Home	10539	2.1	183311	36.9
	TOTAL	**23815**	**4.8**	**472553**	**95.2**
Province of residence	Western Cape	351	0.1	37654	6.9
	Eastern Cape	2506	0.5	70866	13
	Northern Cape	379	0.1	13681	2.5
	Free State	2791	0.5	48133	8.8
	KwaZulu-Natal	6892	1.3	119111	21.9
	North West	1849	0.3	42919	7.9
	Gauteng	3010	0.6	96452	17.7
	Mpumalanga	3320	0.6	44905	8.2
	Limpopo	3663	0.7	46556	8.5
	TOTAL	**24761**	**4.5**	**520277**	**95.5**
Education of the deceased	None	10972	3.4	112699	35.2
	Grade 1	178	0.1	3528	1.1
	Grade 2	281	0.1	5811	1.8
	Grade 3	254	0.2	4356	8.3

Table 2.2 (Continue)

Variables	Levels	Diarrhoea			
		no. of people affected	% no. of people affected	no. of people not infected	% no. of people unaffected
	Grade 4	585	0.2	13293	4.1
	Grade 5	565	0.2	12849	4
	Grade 6	664	0.2	15188	4.7
	Grade 7	893	0.3	19739	6.2
	Grade 8	809	0.3	22626	7.1
	Grade 9	539	0.2	12012	3.7
	Grade 10	774	0.2	20976	6.5
	Grade 11	573	0.2	12735	4
	Grade 12	1157	0.4	33537	10.5
	University	158	0.3	8157	2.5
	TOTAL	**18505**	**5.8**	**302084**	**94.2**
Smoking status of deceased	Yes	1771	0.8	60357	28.5
	No	6084	2.9	143264	67.7
	TOTAL	**7855**	**3.7**	**203621**	**96.3**
Pregnancy status of deceased	Yes	28	0.1	2253	4.2
	No	2807	5.2	48619	90.5
	TOTAL	**2835**	**5.3**	**50872**	**94.7**

The results show that females have a higher percentage of death than males with respect to diarrhoea. Diarrhoea affected more single people than married or widowed people. The highest death rate was for people who were born in KwaZulu-Natal compared to other provinces and the death rate for those who were born in the Northern Cape was the lowest. In KwaZulu-Natal more people were dying from diarrhoea than in other provinces and in the Northern Cape the death rate was the lowest. The results reveal that people were dying in hospitals and homes with estimated numbers of 50.3% and 44.3% respectively. People living in KwaZulu-Natal were more highly affected by diarrhoea than those living in other provinces. People who were not educated or non-scholars, were more likely to die from diarrhea, compared to university students. The results show that diarrhoea was affecting non-smoking more than smokers. People who were not pregnant were more likely to die from diarrhoea the percentage was very high (95.8%).

Chapter 3

Generalised Linear Model

3.1 Introduction

The generalised linear models are a family of important models for categorical as well as continuous responses in statistics. Thus we define the generalised linear model as a flexible generalisation of ordinary least squares regression. Generalised linear models (GLMs) attempt to accommodate variance heterogeneity and asymmetric, non-normal behaviour by offering a range of distributional types that cover at least the more common mean-variance relationships. GLMs are useful for non-normal data, such as binary data. Nelder and Wedderburn (1972) formulated the linear models as a way of unifying various other statistical models, including linear regression, logistic regression, and Poisson regression. The first unifying treatment by Nelder and Wedderburn (1972) demonstrated that a range of the results from applied statistical work can be greatly enhanced by this further development of the general theory. The approaches that were taken emphasise the theoretical foundations of the generalised linear model.

The GLMs generalise linear regression by allowing the linear model to be related to the response variable via a link function and allowing the magnitude of the variance of each measurement to be a function of its predicted value. Generalised linear models accommodate responses that violate the linear model assumptions through two mechanisms: a link function and a variance function, where the link function defines the relationship between the systematic component of the data and the outcome variable in such a way that asymptotic normality and constancy of variance are no longer required (Nelder and Wedderburn 1972). Generalised linear models do not differ in any important way from regular linear models in terms of the process of model specification except that a link function is included to accommodate noncontinuous and possibly bounded outcome variables. Therefore, all of the admonitions about the dangers of the data mining, inverse probability misinterpretation, and probabilistic theory confirmation, apply (Gill, 1999; Greenwald, 1975; Leamer, 1978; Lindsay, 1995; Miller, 1990; Rozeboom, 1960). However, it is still important to be able to assume uncorrelated observations. The variance function expresses the variance as a function of the predicted response, thereby accommodating responses with non-constant variances (such as binary responses). It is also important to be aware that a single data set can lead to many perfectly plausible model specifications and subsequent substantive conclusions (Raftery and Sylvia, 1995).

In linear models there is a set of restrictive assumptions one of which is that the target (dependent variable y) is normally distributed conditioned on the value of predictors with a constant variance regardless of the predicted response value. But the advantages of linear models and their restrictions are computational simplicity, an interpretable model form, and the ability to compute certain diagnostic information about the quality of the fit. Generalised linear models relax these restrictions which are often violated in practice. The binary (yes/no or 0/1) responses do not have the same variance across classes. Furthermore, the sum of terms in a linear model can typically have large ranges encompassing very negative and very positive values. For the binary response example, we would like the response to be a probability in the range [0, 1].

3.2 Exponential Family of Distribution

The development of the generalized linear model theory is based upon the exponential family of distribution (Gill et al. 1999). In the generalised linear model, the random component consists of a response variable Y with independent observations (y_1, y_2,….,y_N) from a distribution in the natural exponential family (Nelder and Wedderburn, 1972). Barndorff-Nielsen (1978) shows that exponential family probability functions have all of their moments. Fisher (1934) developed the idea that many commonly applied probability mass functions and probability density functions are really just a special case of a more general classification he called the exponential family. We will assume that the observations come from a distribution in the exponential family with probability density function

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

Where a, b and c are arbitrary functions and φ is a scale parameter. The functions a and c are such that $a(\varphi) = \frac{\varphi}{w_i}$ *and* $c = c\left(y_i, \frac{\varphi}{w_i}\right)$, where w_i is a known weight for each observation, usually one.

Here φ_i and φ are natural parameters and $a_i(\varphi), b(\theta_i)$ and $c(y_i, \varphi)$ are known functions. The parameters θ_i and φ are essentially location and scale parameters. The scale parameter φ in (3.1) is typically estimated by an appropriate moment estimator involving the Pearson χ^2 Statistic (McCullagh and Nelder, 1989). It can be shown that if Y_i has a distribution in the exponential family then it has mean and variance.

$$E(Y_i) = \mu_i = b'(\theta_i) \tag{3.2}$$

$$Var(Y_i) = {\theta_i}^2 = b''(\theta_i)a_i(\varphi) \tag{3.3}$$

where $b'(\theta_i)$ and $b''(\theta_i)$ with respect to θ_i. $b''(\theta_i)$ is known as the variance function. When you substitute $a_i(\varphi) = \frac{\varphi}{w_i}$ into Eq. (3.3) the variance has the simper form:

$$var(Y_i) = {\theta_i}^2 = \frac{\varphi b''(\theta_i)}{w_i}$$

The standard reference for generalised linear models is McCullagh and Nelder (1989). They stated that GLMs are one such family of models and generally suitable for discrete repeated measurements in the context of correlated data, while a clear exposition is also given by Firth and Harris (1991). The generalised linear model can be seen as an extension of linear multiple regression for a single dependent variable, and understanding the multiple regression is fundamental to understanding the general linear model (Diggle et al., 2002). The exponential family just defined includes as special cases the normal, binomial, Poisson, exponential, gamma, and inverse Gaussian distributions.

3.3 Normal Distribution

This type of distribution is also called a Gaussian distribution and is considered the most prominent probability distribution in statistics as the outcome of the Central Limit Theorem, which states that under mild conditions the sum of a large number of random variables is distributed approximately normally. For this reason, the normal distribution is commonly encountered in practice, and is used throughout statistics as a simple model for complex phenomena.

The density of the normal distribution is given by

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[\frac{-1}{2}\frac{(y_i - \mu_i)^2}{\sigma^2}\right], i = 1, 2, \dots, n \qquad (3.4)$$

where parameter μ is the mean (location of the peak) and σ^2 is the variance (the measure of the width of the distribution). The distribution with $\mu = 0$ and $\sigma = 1$ is called the standard normal.

Expanding the term $(y_i - \mu_i)^2 = {y_i}^2 + {\mu_i}^2 - 2y_i\mu_i$ the density in (3.4) can be expressed as

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

So that θ_i=μ_i , $\varphi = \sigma^2$ and $a_i(\varphi) = \varphi,\ b(\theta_i) = \frac{1}{2}\theta_i{}^2$ (where$\theta_i = \mu_i$)

and $c(y_i, \varphi) = -\frac{1}{2}\left(\frac{y_i{}^2}{\sigma^2} + \log(2\pi\sigma^2)\right)$.

3.4 The Binomial Distribution for GLM

We consider the response variable y_i is binary (taking on only two values that for convenience we code as one or zero). First we verify that the binomial distribution $B(n_i, \pi_i)$ belongs to the exponential family of Nelder and Wedderburn (1972). The binomial probability distribution function is given by

$$f_i(y_i) = \binom{n_i}{y_i}\pi_i{}^{y_i}(1-\pi_i)^{n_i - y_i} \tag{3.5}$$

Where π_i is the probability of successes $(or\ p(y_i = \pi_i))$
Taking logs we find that

$$logf_i(y_i) = y_i\log(\pi_i) + (n_i - y_i)\log(1-\pi_i) + log\binom{n_i}{y_i}$$

$$= y_i log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i\log(1-\pi_i) + log\binom{n_i}{y_i}. \tag{3.6}$$

This expression has the general exponential form

$$logf_i(y_i) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\varphi)} + c(y_i, \varphi)$$

Where the canonical parameter θ_i is the logit of π_i, i.e

$$\theta_i = log\left(\frac{\pi_i}{1-\pi_i}\right). \tag{3.7}$$

Solving for π_i we therefore get

$$\pi_i = \frac{e^{\theta_i}}{1-e^{\theta_i}} \text{ and } \quad 1 - \pi_i = \frac{1}{1+e^{\theta_i}}$$

Rewriting the second term in expression (3.6) as a function of θ_i

$\log(1-\pi_i) = -\log(1-e^{\theta_i})$,

we can identify the cumulant function $b(\theta_i)$ as

$b(\theta_i) = n_i \log(1+e^{\theta_i})$.

The remaining term is a function of y_i but not π_i, leading to

$$c(y_i, \varphi) = log\binom{n_i}{y_i}.$$

We may now set $a_i(\varphi) = \varphi$ and $\varphi = 1$. Finally, we verify the mean and variance. By taking the first and second derivatives of $b(\theta_i)$ with respect to θ_i we find that

$$\mu_i = b'(\theta_i) = n_i \frac{e^{\theta_i}}{1+e^{\theta_i}} = n_i \pi_i$$

and

$$\theta_i{}^2 = a_i(\varphi)b''(\theta_i) = n_i \frac{e^{\theta_i}}{(1+e^{\theta_i})^2} = n_i\pi_i(1-\pi_i)$$

in agreement with what we knew. McCullagh and Nelder (1989) work with the proportion $w_i = \frac{y_i}{n_i}$, which takes values 0 and 1. Note that the mean and variance depend on the underlying probability π_i. This shows that any factor that affects the probability will alter not just the mean but also the variance of the observation. This suggests that a linear model which allows the predictors to affect the mean but assumes that the variance is constant will not be adequate for the analysis of binary data.

3.5 The Poisson Distribution

This type of distribution is a discrete distribution which takes on the values Y=0,1,2,… and is used as a model for the number of events (such as the number of deaths in South Africa) in a specific time period. This distribution was first introduced by Simon Denis Poisson (1781-1840). The Poisson distribution is determined by one parameter, lambda. The Poisson distribution is given by

$$f_i(y_i) = \frac{e^{-\lambda_i}\lambda_i{}^{y_i}}{y_i!} \tag{3.8}$$

where
e is the base of the natural logarithm
y is the number of occurrences of an event
y_i! is the factorial of y_i
λ_i is the positive real number.

We now verify that this distribution belongs to the exponential family as defined by Nelder and Wedderburn (1972). Taking logs we find

$$logf_i(y_i) = y_i \log(\lambda_i) - \lambda_i - \log(y_i!) \tag{3.9}$$

The coefficient of x_i shows us immediately that the canonical parameter is

$$\theta_i = \log(\lambda_i)$$

and therefore that the canonical link is the log-link.

Solving for λ_i we obtain the inverse link

$$\lambda_i = e^{\theta_i}$$

and we can write the second term in expression (3.9) as

$$b(\theta_i) = e^{\theta_i}$$

The remaining term is a function of y_i only, so we identify

$$c(y_i, \varphi) = -\log(y_i!)$$

Finally, note that we can take $a_i(\varphi) = \varphi$ and $\varphi = 1$, just as we did in the binomial case. Lastly, we verify the mean and variance. By taking the first and second derivative of $b(\theta_i)$ with respect to θ_i we have

$$\lambda_i = b'(\theta_i) = e^{\theta_i} = \lambda_i$$

and

$$a_i(\varphi)b''(\theta_i) = e^{\theta_i} = \lambda_i.$$

As is the case, the mean and variance are equal. This means that the parameter λ_i is equal to the expected number of occurrences during the given interval, i.e ($\lambda_i = E(y_i)$). But λ_i is not only the mean of occurrences, but also its variance

$$\sigma_{y_i}{}^2 = E(y_i{}^2) - E(y_i)^2 = E(y_i) = \lambda_i$$

3.6 The Gamma Distribution

This type of distribution is defined as a two- parameter family of continuous probability distributions, namely a scale parameter β and a shape parameter γ. The gamma distribution also represents the sum of n exponentially distributed random variables where the scale parameter is the mean of the exponential distribution and the shape parameter represents the number of variables. This is apparent when the profile of an exponential distribution with mean set to one is compared to a gamma distribution with a shape parameter of one and a mean of one.

The gamma distribution has a probability density function that can be expressed in terms of the gamma function parameterised in terms of a shape parameter γ and scale parameter β, where both γ and β are positive values. The equation defining the probability density function of a gamma distributed random variable x is

$$f(y) = \frac{\left(\frac{y-\mu}{\beta}\right)^{\alpha-1} exp\left(\frac{-(x-\mu)}{\beta}\right)}{\beta_{\Gamma}(\alpha)} \qquad ,y \geq \mu; \;\; \alpha, \beta > 0 \tag{3.10}$$

where

α is the shape parameter

μ is the location parameter

β is the scale parameter and

Γ is the gamma function which has the formula

$$\Gamma(\alpha) = \int_0^\infty e^{\alpha-1} e^{-t} dt$$

In the special case where $\mu = 0$ and β=1, the distribution is called the standard gamma distribution and the equation for the standard gamma distribution reduces to

$$f(y) = \frac{y^{\gamma-1} e^{y}}{\Gamma(\alpha)}, \qquad y \geq 0;\ \alpha > 0$$

3.7 Link Function for Generalised Linear Model

Using the link function in generalised linear models, we can transform any predicted curve to conform to different assumptions about the form of the relationship and the error distribution (Nelder and Wedderburn, 1972). The link function provides the relationship between the linear predictor and the mean of the distribution function. One of the difficult things to grasp about GLMs is the relationship between the values of the response variable (as measured in the data and predicted by the model in fitted values) and the linear predictor. The thing to remember is that the link function relates the mean value of y to its linear predictor (Crawley, 2007). In symbols, this means that

$$\boldsymbol{\eta} = g(\mu) = \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{\beta}$$

This function must be monotonic and differentiable.

Table 3.1 Some Common Link Functions and their Inverses

Link	$\boldsymbol{\eta_i} = \boldsymbol{g}(\mu_{\boldsymbol{i}})$	$\mu_{\boldsymbol{i}} = \boldsymbol{g}^{\boldsymbol{-1}}{}_{(\boldsymbol{\eta_i})}$
Identity	μ_i	η_i
Log	$\log_e \mu_i$	e^{η_i}
Inverse	$\mu_i{}^{-1}$	$\eta_i{}^{-1}$
Inverse-square	$\mu_i{}^{-2}$	$\eta_i{}^{-1/2}$
Square-root	$\sqrt{\mu_i}$	$\eta_i{}^2$
Logit	$\log_e \frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+e^{-\eta_i}}$
Probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
Log-log	$-\log_e[-\log_e(\mu_i)]$	$exp[-\exp(-\eta_i)]$
Complementary log-log	$\log_e[-\log_e(1-\mu_i)]$	$1 - exp[-\exp(\eta_i)]$

μ_i is the expected value of the response; η_i is the linear predictor; $\Phi(.)$ is the cumulative distribution function of the standard-normal distribution.

Because the link function is invertible, we can also write

$$\mu = g^{-1}(\boldsymbol{\eta}) = g^{-1}(\boldsymbol{X^T\beta})$$

and, thus, the GLM may be thought of as a linear model for a transformation of the expected response or as a nonlinear regression model for the response. The inverse link $g^{-1}(.)$ is also called the mean function.

The canonical link function is the function which transforms the mean to a canonical location parameter of the exponential dispersion family member (Lindsey, 1974).

Example:

Normal: $g(\mu) = \mu$

Inverse Gaussian: $g(\mu) = \mu^{-2}$

Gamma: $g(\mu) = \mu^{-1}$

Poisson: $g(\mu) = \log(\mu)$

Binomial: $g(\mu) = log\left(\frac{\mu}{1-\mu}\right)$

The canonical link and the variance function are related by $V(\mu) = \frac{1}{g'(\mu)}$. Note that while canonical links are often used, this is not always the case. The only requirements of the link function are that it should be monotonic and differentiable and should range the whole real line (-∞, ∞).

3.8 Similarities and Differences of GLMs

A GLM consists of three components. First is the random component, which is the response variable and its probability distribution. The probability distribution must be from exponential family of distributions, which includes normal, binomial, Poisson, gamma and negative binomial. If the response variable is a continuous variable, its probability distribution might be normal; If the response variable is binary (e.g. alive or dead), the probability distribution might be binomial; If the response variable represents counts, then the probability distribution might be Poisson. Probability distributions from the exponential family can be defined by the natural parameter, a function of the mean, and the dispersion parameter, a function of the variance that is required to produce standard errors for estimates of the mean (Hilbe, 1994).

3.9 Likelihood and Log-likelihood Equations

The maximum likelihood estimation represents the backbone of statistical estimation. This means that the parameters can be estimated using maximum likelihood (Fisher, 1921). In order to fit a generalised linear model, we need to estimate the parameters **β** in the linear predictor. The probability density function for a random variable, y, conditioned on a set of parameters, $\boldsymbol{\theta}$, is defined as $f(\boldsymbol{y}|\boldsymbol{\theta})$. The likelihood function/joint density for the n independent observations $y_1, y_2, \dots \dots, y_n$ is the product of the individual densities:

$$f(y_1, y_2, \dots \dots, y_n|\boldsymbol{\theta}) = \Pi(f(y_i|\boldsymbol{\theta})) = L(\boldsymbol{\theta}|\boldsymbol{y}). \tag{3.11}$$

This likelihood function is defined as a function of the unknown parameter vector, θ, where y is used to indicate the collection of sample data. It is usually simpler to work with the log of the likelihood function, so the log-likelihood is given by

$$L(\boldsymbol{\theta}, \boldsymbol{y}) = \log(\boldsymbol{\theta}, \boldsymbol{y}) = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\varphi)} + c(y_i, \varphi) \right\}$$

To emphasise our interest in the parameters, given the observed data, we denote this function $L(\boldsymbol{\theta}|\boldsymbol{data}) = L(\boldsymbol{\theta}|\boldsymbol{y})$.

It is well known that if the likelihood function has the exponential family form, maximum likelihood estimates of the regression parameter can often be found using the method of weighted least square (Nelder and Wedderburn, 1972; Bradley, 1973; Wedderburn, 1974; and Jennrich and Moore, 1975). Modified and conditional likelihood sometimes have the required exponential form. Thus the method of weighted least squares can be used to find the maximum likelihood estimates even in cases where the likelihood function does not have the exponential family form (Jorgensen, 1983).

The quasi-likelihood estimation is a one way of allowing for overdispersion. It is often used with models for count data or grouped binary data. The quasi-likelihood models can be fitted using a straightforward extension of the algorithms used to fit generalised linear models (Wedderburn, 1974).

Given the vector of random variables **Y** with mean **μ** and covariance matrix $\sigma^2\boldsymbol{V}(\mu)$ the log quasi-likelihood considered as a function of **μ**, is given by the system of partial differential equations

$$\frac{\partial \boldsymbol{l}(\mu,\boldsymbol{y})}{\partial \mu} = \boldsymbol{V}^{-}(\mu)(\boldsymbol{y}-\mu)$$

which extends Wedderburn`s (1974) definition.

Solving for $l(\mu;\boldsymbol{y})$ we have

$$l(\mu;\boldsymbol{y}) = \boldsymbol{y}^{\boldsymbol{T}}(\boldsymbol{\theta}) - b(\boldsymbol{\theta}) - c(\boldsymbol{y},\boldsymbol{\sigma})$$

where $c(\boldsymbol{y},\sigma)$ is entirely arbitrary.

The variance of $\boldsymbol{Y}$ is $Var(\boldsymbol{y}) = \frac{\mu(1-\mu)}{m}$ for the binomial distribution and $Var(\boldsymbol{y}) = \boldsymbol{\mu}$ for the Poisson distribution. Overdispersion occurs when the variance of $\boldsymbol{Y}$ exceeds the $Var(\boldsymbol{y})$ above. That is $\sigma^2\boldsymbol{V}(\mu)$, where $\sigma > 1$. With overdispersion, methods based on quasi-likelihood can be used to estimate the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$. A quasi-likelihood function

$$Q(\boldsymbol{\mu};\boldsymbol{y}) = \int_{y}^{\mu} \frac{y-t}{\sigma^2 V(\mu)}\,dt$$

is specified by its associated variance function.

3.10 Maximum Likelihood Estimation in the GLM

Recall expression (3.11) that the likelihood function is given by

$$l(\beta,\varphi) = \sum_{i=1}^{n}\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\varphi)} + c(y_i,\varphi)\right] \tag{3.12}$$

where

$$\theta_i = \theta(\boldsymbol{x_i}^{\boldsymbol{T}}\boldsymbol{\beta}) = \theta(\eta_i),$$

and **θ** is a known, monotone function and subsumes all of the θ_i.

The ML estimation method selects as estimates the value of the parameters **β** that maximize the likelihood, i.e expression (3.12).

In any GLM, the likelihood function depends on **β** only through η_i, so

$$\frac{\partial}{\partial\beta_j} l(\boldsymbol{\beta}, \boldsymbol{\varphi}) = \sum_{i=1}^{n} \frac{\partial l_i}{\partial\beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial\varphi_i} . \frac{\partial\theta_i}{\partial\eta_i} . x_{ij}$$

We define $\Delta_i = \frac{\partial\theta_i}{\partial\eta_i}$, which is sometimes called link adjustment. If the canonical link is used, then $\Delta_i \equiv 1$, since in this case $\theta_i = \eta_i$.

We now have

$$\frac{\partial l_i}{\partial\theta_i} = \frac{y_i - b'(\theta_i)}{\varphi} = \frac{y_i - \mu_i}{\varphi}.$$

The likelihood equations may thus be written as

$$\frac{\partial}{\partial\beta_j} l(\boldsymbol{\beta}, \boldsymbol{\varphi}) = \sum_{i=1}^{n} (y_i - \mu_i)\Delta_i x_{ij} = 0, \quad \text{j=1,2,.....,p.} \tag{3.13}$$

These equations can be written in matrix form as

$$\boldsymbol{X}^{\boldsymbol{T}}\Delta(\boldsymbol{y} - \boldsymbol{\mu}) = 0$$

where X is $n \times p$ with element x_{ij} in the ith row and jth column, Δ is an $n \times n$ diagonal matrix with diagonal elements $\Delta_1, \Delta_2, \ldots.., \Delta_n$, $\boldsymbol{y} = (y_1, y_2, \ldots\ldots, y_n)^T$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots\ldots, \mu_n)^T$.

If we let $\boldsymbol{S} = (\boldsymbol{y} - \boldsymbol{\mu})$, the equations are

$$\boldsymbol{X}^{\boldsymbol{T}}\Delta\boldsymbol{S} = \boldsymbol{0} \tag{3.14}$$

In general, the equations in (3.13) are nonlinear in **β.**

To obtain the MLE of **β,** called $\widehat{\boldsymbol{\beta}}$, we can use Newton-Raphson. The Newton-Raphson procedure may be expressed as follows:

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^{t} - \left[\frac{\boldsymbol{\partial}(\boldsymbol{X}^{\boldsymbol{T}}\Delta\boldsymbol{S})}{\boldsymbol{\partial\beta}}\right]^{-\mathbf{1}} \times (\boldsymbol{X}^{\boldsymbol{T}}\Delta\boldsymbol{S})|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{t}}$$

according to Harville (1977), this leads to

$$\frac{\partial}{\partial\beta_j}\boldsymbol{X}^{\boldsymbol{T}}\Delta\boldsymbol{S} = X^{T}\left[\Delta\frac{\partial\boldsymbol{S}}{\partial\beta_j} + \frac{\partial\Delta}{\partial\beta_j\boldsymbol{S}}\right]$$

Using the definition of S, i.e $\boldsymbol{S} = \boldsymbol{y} - \boldsymbol{b}'(\boldsymbol{\theta}_{\boldsymbol{i}})$ as $\boldsymbol{\mu}_{\boldsymbol{i}} = \boldsymbol{b}'(\boldsymbol{\theta}_{\boldsymbol{i}})$

$$\frac{\partial\boldsymbol{S}}{\partial\beta_j} = -\big(b''(\theta_1)\theta'(\eta_1)x_{1j}, \ldots.., b''(\theta_n)\theta'(\eta_n)x_{nj}\big)^{T} = -\boldsymbol{V}\Delta\big(x_{1j}, \ldots.., x_{nj}\big)^{T},$$

where V is the $n \times n$ diagonal matrix with diagonal elements $b''(\theta_i),\ i = 1,2, \ldots.., n.$

It follows that

$$\boldsymbol{X}^{\boldsymbol{T}}\Delta\frac{\boldsymbol{\partial S}}{\boldsymbol{\partial\beta}} = -\boldsymbol{X}^{\boldsymbol{T}}\Delta\boldsymbol{V}\Delta\boldsymbol{X}$$

and also,

$\frac{\partial\Delta}{\partial\beta_j}$= diagonal matrix.

Therefore

$$\frac{\partial\Delta}{\partial\beta_j}\boldsymbol{S} = \dot{\Delta}\boldsymbol{H}\big(x_{1j}, \ldots\ldots.., x_{nj}\big)^{T},$$

where $\dot{\Delta}$ and **H** are diagonals. Then

$$\frac{\boldsymbol{\partial}}{\boldsymbol{\partial\beta}}\boldsymbol{X}^{\boldsymbol{T}}\Delta\boldsymbol{S} = -\boldsymbol{X}^{\boldsymbol{T}}\big(\Delta\boldsymbol{V}\Delta - \dot{\Delta}\boldsymbol{H}\big)\boldsymbol{X}. \qquad (3.15)$$

So the Newton-Raphson iterations may be written as

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^{t} + \left[\boldsymbol{X}^{\boldsymbol{T}}(\Delta\boldsymbol{V}\Delta - \dot{\Delta}\boldsymbol{H})\boldsymbol{X}\right]^{-1} \times (\boldsymbol{X}^{\boldsymbol{T}}\Delta\boldsymbol{S})|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{t}}.$$

When using a canonical link, $\Delta = \boldsymbol{I}$ and $\dot{\Delta} = \boldsymbol{0}$, in which case

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^{t} + (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{S}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{t}}.$$

The Hessian in Fisher scoring is

$$-\boldsymbol{X}^{\boldsymbol{T}}(\Delta\boldsymbol{V}\Delta - \dot{\Delta})\boldsymbol{X}$$

which has expectation $-\boldsymbol{X}^{\boldsymbol{T}}\Delta\boldsymbol{V}\Delta\boldsymbol{X}$ since **E (H) =0**

The Fisher Information matrix for **β** is

$$\boldsymbol{I}(\boldsymbol{\beta}) = \frac{-(-\boldsymbol{X}^{\boldsymbol{T}}\Delta\boldsymbol{V}\Delta\boldsymbol{X})}{\boldsymbol{\varphi}} = \frac{\boldsymbol{X}^{\boldsymbol{T}}\Delta\boldsymbol{V}\Delta\boldsymbol{X}}{\boldsymbol{\varphi}}$$

Now the Fisher scoring algorithm is

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^{t} + (\boldsymbol{X}^{\boldsymbol{T}}\Delta\boldsymbol{V}\Delta\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\Delta\boldsymbol{S}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{t}}$$

Note that the Fisher scoring algorithm and the Newton-Raphson are identical when the canonical link is used. In the canonical link case, the Newton-Raphson scheme is an iteratively reweighted least squares (IRLS) algorithm.

Define

$$\boldsymbol{Z}^{t} = \boldsymbol{X}\boldsymbol{\beta}^{t} + \boldsymbol{V}^{-1}\boldsymbol{S}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{t}}$$

Since

$$\boldsymbol{\beta}^{t} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}\boldsymbol{X})^{-1}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}\boldsymbol{X})\boldsymbol{\beta}^{t}$$

and

$$(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{S} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{V}\boldsymbol{V}^{-1}\boldsymbol{S})$$

we have

$$\boldsymbol{\beta}^{t+1} = (\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{V})|_{\boldsymbol{\beta}=\boldsymbol{\beta}^t}\boldsymbol{Z}^t = (\boldsymbol{X}^T\boldsymbol{V}^t\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{V}^t)\boldsymbol{Z}^t$$

where

$$\boldsymbol{V}^t = \boldsymbol{V}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^t}$$

Therefore $\boldsymbol{\beta}^{t+1}$ corresponds to a weighted least squares regression of $\boldsymbol{Z}^t$ on **X** with weight matrix $\boldsymbol{V}^t$.

$$\sum_{i=1}^{n} \boldsymbol{v}_i(\boldsymbol{y}_i - \boldsymbol{x}_i{}^T)^2.$$

the solution is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{y},$$

where $\boldsymbol{V}$ is the diagonal matrix with diagonal entries $v_1, v_2, \dots, v_n$.

3.11 Application of GLM to the dataset

The fitted model has as its explanatory variables sex, marital status of deceased, province of birth of deceased, province of death of deceased, place of death of deceased , province of residence of deceased, pregnancy of deceased, smoking status of deceased, and education status of deceased. In this application the model was fitted using SAS PRO GENMOD which is an in built procedure is SAS version 9.1 or version 9.2 capable of fitting both generalised linear models and logistic regression models.

By default, PROC GENMOD uses a base line parameterisation for categorical variables where the last category of each variable is used as the reference category. The output from SAS using PROC GENMOD result in the following table:

Table 3.2: Analysis of parameter estimates using PROC GENMOD

Analysis of Parameter Estimates								
	Parameter		DF	estimate	Standard Error	Wald 95% Confidence		Pr>ChiSq
	Intercept		1	-12.03	0.0107	-0.0506	-0.008	0.0054*
Sex	Male	1	1	-0.0314	0.0019	0.0151	-0.0078	0.3242
	Female (reference)	2	-	-	-	-	-	-
Marital Status	Single	1	1	0.0327	0.0037	0.0256	0.0399	<.0001*
	Civil marriage	2	1	-0.0079	0.0042	-0.0162	0.0004	0.0622
	Living as married	3	1	-0.0011	0.0059	-0.0128	0.0105	0.8518
	Widowed	4	1	0.0132	0.0046	-0.0222	-0.0041	0.0044*
	Religious law marriage	5	1	-0.0091	0.0065	-0.0218	0.0036	0.162
	Divorced	6	1	-0.0054	0.0079	-0.0209	0.0101	0.4942
	Customary marriage (reference)	7	-	-	-	-	-	-
Province of birth	Western Cape	1	1	-0.0031	0.0072	-0.0109	0.0171	0.6648
	Eastern Cape	2	1	-0.0069	0.0063	-0.0192	0.0055	0.2757
	Northern Cape	3	1	0.0014	0.0099	-0.018	0.0207	0.8881
	Free State	4	1	-0.007	0.007	-0.0067	0.0207	0.3187
	KwaZulu-Natal	5	1	-0.0061	0.0057	-0.0173	0.005	0.0287*
	North West	6	1	0.0104	0.0066	-0.0025	0.0232	0.114
	Gauteng	7	1	-0.0084	0.0056	-0.0025	0.0193	0.1295
	Mpumalanga	8	1	0.0132	0.0059	-0.0247	-0.0017	0.0245*
	Limpopo (reference)	9	-	-	-	-	-	-

Table 3.2 (Continue)

Analysis of Parameter Estimates								
	Parameter		DF	estimate	Standard Error	Wald 95% Confidence		Pr>ChiSq
Province of death	Western Cape	1	1	-0.0325	0.0088	-0.0497	-0.0152	0.9734
	Eastern Cape	2	1	0.013	0.0081	-0.0289	0.003	0.1104
	Northern Cape	3	1	-0.0338	0.0123	-0.0579	-0.0097	0.0059*
	Free State	4	1	-0.0055	0.0105	-0.0261	0.0151	0.6008
	KwaZulu-Natal	5	1	0.0003	0.0076	-0.0151	0.0146	0.0002*
	North West	6	1	-0.0105	0.0082	-0.0266	0.0056	0.1998
	Gauteng	7	1	-0.023	0.007	-0.0368	-0.0093	0.001*
	Mpumalanga	8	1	-0.0088	0.0073	-0.0055	0.0231	0.2262
	Limpopo (reference)	9	-	-	-	-	-	-
Province of residence	Western Cape	1	1	-0.0258	0.009	-0.0434	-0.0082	0.3525
	Eastern Cape	2	1	0.0131	0.0079	-0.0287	0.0024	0.0979
	Northern Cape	3	1	-0.0132	0.0126	-0.0379	0.0116	0.2965
	Free State	4	1	-0.0061	0.0104	-0.0265	0.0143	0.558
	KwaZulu-Natal	5	1	0.007	0.0075	-0.0218	0.0078	0.0041*
	North West	6	1	-0.0305	0.008	-0.0462	-0.0148	0.0001*
	Gauteng	7	1	-0.0174	0.0072	-0.0316	-0.0032	0.0165*
	Mpumalanga	8	1	-0.0027	0.0075	-0.012	0.0174	0.7168
	Limpopo (reference)	9	-	-	-	-	-	-
Place of death	Hospital	1	1	-0.0032	0.002	-0.0071	0.0007	0.0181*
	Emergency room	2	1	-0.0113	0.0066	-0.0016	0.0242	0.0855
	Dead on arrival	3	1	0.0039	0.0058	-0.0152	0.0074	0.5015
	Nursing home	4	1	-0.0121	0.0062	-0.0243	0	0.0507
	Home (reference)	5	-	-	-	-	-	-

Table 3.2 (Continue)

Analysis of Parameter Estimates								
	Parameter		DF	estimate	Standard Error	Wald 95% Confidence		Pr>ChiSq
Education	None	0	1	-0.0239	0.0074	0.0094	0.0384	0.0013*
	Grade 1	1	1	-0.0201	0.0127	-0.0048	0.0449	0.1131
	Grade 2	2	1	0.0222	0.0113	0.0001	0.0443	0.0488*
	Grade 3	3	1	-0.0117	0.0099	-0.0077	0.0312	0.2378
	Grade 4	4	1	-0.0242	0.0091	0.0062	0.0421	0.0082*
	Grade 5	5	1	-0.0099	0.0092	-0.0082	0.0279	0.2833
	Grade 6	6	1	-0.0082	0.0089	-0.0093	0.0256	0.3587
	Grade 7	7	1	0.0144	0.0085	-0.0023	0.0312	0.0916
	Grade 8	8	1	-0.0036	0.0084	-0.0129	0.0201	0.6683
	Grade 9	9	1	0.0053	0.0093	-0.0129	0.0235	0.5665
	Grade 10	10	1	-0.0082	0.0085	-0.0085	0.0248	0.3361
	Grade 11	11	1	-0.0011	0.0091	-0.019	0.0167	0.9017
	Grade 12	12	1	0.0029	0.0081	-0.0129	0.0187	0.7193
	University (reference)	13	-	-	-	-	-	-
Smoking	Yes	1	1	0.0058	0.0033	-0.0121	0.0006	0.0779
	No (reference)	2	-	-	-	-	-	-
Pregnancy	Yes	1	1	-0.0423	0.0136	-0.069	-0.0156	0.0695
	No (reference)	2	-	-	-	-	-	-

We find that single people are significant than customary married people at 5% level of significant, therefore single people were more likely to die from diarrhoea than customary married people. The result also reveals that people who are widowed versus customary married people are significant at 5% level. This means that single and widowed people are more likely to die from diarrhoea than customary marriage people. We find that KwaZulu-Natal province versus Limpopo province was significant at 5% level of significance, which clearly shows that people who are born in KwaZulu-Natal province are more likely to die from diarrhoea than those who are born in Limpopo province.

For province of death, we find that KwaZulu-Natal Province versus Limpopo province was significant at 5% level of significance; this shows that people are more likely to die from diarrhoea in the KwaZulu-Natal than in Limpopo province. The result also reveals that the Northern Cape province was more significant than Limpopo province at 5%, which tells us that people in the Northern Cape province are more likely to die from diarrhoea than people in Limpopo province. We also find that Gauteng province versus Limpopo province was significant at 5% level, which shows that people are more likely to die from diarrhoea in Gauteng than in Limpopo province. The results show us KwaZulu-Natal residents are significant compared to Limpopo residents, which means that KwaZulu-Natal residents are more likely to die from diarrhoea than Limpopo residents. The result also reveals that North West residents versus Limpopo residents are significant at 5% level. This clearly shows that North West residents are more likely to die from diarrhoea than Limpopo residents. We also find that Gauteng residents versus Limpopo residents are significant at 5% level, which shows that Gauteng residents are more likely to die from diarrhoea than Limpopo residents.

We find that hospital versus home are significant at 5% level of significance, which means that people were more dying in hospitals as compared to those who were home. The results show that uneducated people are significant at 5% level as compared to university students. We also find that Grade 2 students versus university students are significant at 5% level. The results further reveal that Grade 4 students versus university students are significant at 5% level. This clearly tells us that uneducated people, Grade 2 and Grade 4 students, are more likely to die from diarrhoea than university students.

Chapter 4

Logistic Regression

4.1 Introduction

This chapter explains the motivation for the use of Logistic regression for the analysis of binary response data. When there are several explanatory variables, multiple regression is used. However, in our case the response is not of the continuous type. Instead, the response is simply a binary response e.g. alive or dead. In this chapter we look at binary response data and its analysis via logistic regression on how does one model relate between explanatory variables and a binary response variable.

Logistic regression (also called logistic modeling or the logit model) is a statistical technique that allows group membership to be predicted from predictor or independent variables, regardless of whether the predictor variables are continuous, discrete, or a combination of both. The logistic regression model was introduced by Cox (1970) to describe the dependency of a binary variable on a set of continuous variables. Logistic regression is a type of predictive model that can be used when the target variable is a categorical variable with two categories. It falls within a broader class of models called generalised linear models (GLMs), developed and addressed quite exhaustively by McCullagh and Nelder (1986).

These types of methods (regression) have become an integral component of any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables, and it is often the case that the outcome variable is discrete, taking on two or more possible values (Hosmer, and Lemeshow, 2005). According to Larsen (2008), dichotomous response variables (that is, response variables that have only two possible outcomes) cannot be assumed to be normally distributed, thus the most common method to use for analysing data with dichotomous response variables is arguably logistic regression. Berkson (1944, 1953) in connection with the analysis of so-called bio assays introduced the statistical model underlying logistic regression and described logistic regression as a form of statistical modeling that is often appropriate for categorical outcome variables.

As previously stated the response variable is usually dichotomous, but it may be polytomous, that is, have more than two response levels in which case multinomial logistic regression is appropriate. In logistic regression, it is assumed that the explanatory variables are independent. If the explanatory variables are not independent one has to take the dependencies into account. Suggestions for such models are due to Conolly and Liang (1988). Logistic regression has been given a very comprehensive history and methodology by Imrey et al. (1981). It is applicable to multilevel responses and the responses may be ordinal or nominal. For ordinal response outcomes, one can model functions called cumulative logits by performing ordered logistic regression using the proportional odds model, and for nominal response outcome one forms generalised logits and performs a logistic analysis (McCullagh 1980).

4.2 Binary response logistic regression

The linear logistic model assumes a dichotomous dependent variable Y with probability π of a positive outcome or success. When there are only a few or no repeated observations at the various levels X_j of the independent variable, as is often the case in observational studies, we estimate the logistic response function from the individual Y_i observations and the dependent variable which is binary, taking on the values 1 and 0 with probabilities π and 1-π, respectively. In other words, Y is a Bernoulli random variable with parameter $E(Y) = \pi$ (Neter et al. 1989).

Given a single predictor variable X, the simple logistic regression model is given by

$$E(Y_i) = \pi_i = \frac{\exp(\beta_0+\beta_1 X_i)}{1+\exp(\beta_0+\beta_1 X_i)} \qquad (i = 1,2,\ldots.,n) \tag{4.1}$$

an equivalent form of (3.1) is given by:

$$E(Y_i) = \pi_i = [1 + \exp(-\beta_0 - \beta_1 X_i)]^{-1}$$

where Y_i are independent Bernoulli random variables with expected values $E(Y_i) = \pi_i$, β_0 and β_1 are regression parameters which need to be estimated. The X observations are assumed to be known constants. Alternatively, if the X is random, $E(Y_i)$ is viewed as a conditional mean, given the value of X_i.

A major problem with the linear probability model is that probabilities are bounded by 0 and 1, but linear functions are inherently unbounded (Allison, 1999). The solution is to transform the probability so that it is no longer bounded. Transforming the probability to odds removes the upper bound. If we then take the logarithm of the odds, we also remove the lower bound. When we finally set the result equal to a linear function of the explanatory variables, we get the logit model. As discussed, a major problem with the linear probability model is that probabilities are bounded by 0 and 1, but linear functions are inherently unbounded. The solution is to transform the probability so that it is no longer bounded. Note that model is very precisely the logit model because with simple algebra it can be shown that:

$$\log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 X_i$$

a transformation of π_i that is central to our study of logistic regression is the logit transformation. This transformation is defined, in terms of π_i as:

$$\pi_i^{'} = \ln\left[\frac{\pi_i}{1-\pi_i}\right] \qquad (4.2)$$

we obtain from (3.1):

$$\pi_i^{'} = \beta_0 + \beta_1 X_i \qquad (4.3)$$

The ratio $\frac{\pi_i}{1-\pi_i}$ in the logit transformation is called the odds. The transformed response function (4.2) is referred to as the logit response function, and $\pi_i^{'}$ is called the logit or log-odds. The importance of this transformation is that $\pi_i^{'}$ has many of the desirable properties of a linear regression model.

In the logistic regression we assume that an observation of the outcome variable may be expressed as

$$y = \pi + \mathcal{E}$$

where $\mathcal{E}$ is called the error term and may assume one of two possible values.

Here if $y = 1$ then $\mathcal{E} = 1 - \pi$ with probability π, and if $y = 0$ then $\mathcal{E} = -\pi$ with probability $1 - \pi$. Thus, $\mathcal{E}$ has a distribution with mean zero and variance equal to $\pi[1 - \pi]$.

4.3 Likelihood function

We shall use the method of maximum likelihood to estimate the parameters of the logistic regression function since this method is well suited to deal with the problems associated with the observations Y_i being binary (Neter et al. 1989). Before doing that, we first need to develop the joint probability function of the sample proportion. Since each Y_i observation is an ordinary Bernoulli random variable, where:

$P(Y_i = 1) = \pi_i$ and $P(Y_i = 0) = 1 - \pi_i$

we can represent its probability distribution as follows:

$$f_i(Y_i) = {\pi_i}^{Y_i}(1 - \pi_i)^{1-Y_i}\,, Y_i = 0, 1\,;\;\; i = 1, 2, \dots, n$$

Note that $f_i(1) = \pi_i$ and $f_i(0) = 1 - \pi_i$. Hence, $f_i(Y_i)$ is simply the probability that $Y_i = 1$ or 0. Suppose Y_i, $i = 1, \dots, n$ are independent observations. Then the joint probability distribution is:

$$(\pi, y) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} {\pi_i}^{Y_i}\,(1 - \pi_i)^{1-Y_i} \qquad (4.4)$$

Where $\pi = \pi_1, \dots, \pi_n$ and $y = y_1, \dots, y_n$

The logarithm of the likelihood function is given by:

$$l(\pi, y) = \log L(\pi, y) = \log_e \prod_{i=1}^{n} {\pi_i}^{Y_i} (1 - \pi_i)^{1-Y_i} = \sum_{i=1}^{n} \left[Y_i \log_e \left(\frac{\pi_i}{1-\pi_i} \right) \right] + \sum_{i=1}^{n} \log_e(1 - \pi_i) \qquad (4.5)$$

In the method of the maximum likelihood, we can maximize either the likelihood function or the logarithm of the likelihood function because both they lead to the same answer for the parameters. We will use the logarithm of the likelihood function as it is the easiest.

Using (4.3), it follows that

$$\log_e \left(\frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 X_i$$

Similarly, for the multiple logistic regression we have

$$l(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_i \beta X_i y_i - \sum_i \log(1 + \exp(\beta X_i))$$

Where

$\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ **,** $\boldsymbol{y} = (y_0, y_1, \dots, y_n)^T$

And

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdot & \cdot & \cdot & . & . & X_{1p} \\ 1 & X_{21} & X_{22} & \cdot & \cdot & \cdot & . & . & X_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ . & . & . & . & . & . & \cdot & \cdot & . \\ 1 & X_{n1} & X_{n2} & . & . & . & . & . & X_{np} \end{bmatrix}$$

4.4 Maximum-Likelihood Estimation of the Logistic Regression Model

The existence of maximum likelihood estimates for the logistic regression model depends on the configuration of the data points in the data set (Albert and Anderson, 1984; Santner and Duffy, 1986; So, 1995). There are three mutually exclusive and exhaustive categories for the configuration of data points in a data set:

- Complete Separation
- Quasi-Complete Separation
- Overlap

A binary response logistic regression model is considered. Unconditional maximum likelihood estimation (asymptotic inference) is used when matched data are not considered, provided that the total number of variables in the model is not too large relative to the number of observations (Kleinbaum, 1994). This method of inference is based on maximizing the likelihood function for parameter estimation using the unconditional formula (Kleinbaum, 1994). This is the usual large-sample asymptotic method used by most of the current statistical software packages such as SAS and Genstat (Kleinbaum, 1994; Mehta and Patel, 1995). As previously stated, the existence and uniqueness of maximum likelihood parameter estimates for the logistic regression model depends on the pattern of the data points in the observation space (Albert and Anderson, 1984; Santer and Duffy, 1986; So, 1993). Fox (2005) used the Newton-Raphson method as a common iterative approach to estimating a logistic-regression model. Fox (2005) firstly chose initial estimates of the regression coefficients, such as b_0 = 0. At each iteration t, update the coefficients:

$$\boldsymbol{b_t} = \boldsymbol{b_{t-1}} + \left(\boldsymbol{X}'\boldsymbol{V_{t-1}}\boldsymbol{X}\right)^{\boldsymbol{-1}}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{p_{t-1}}) \tag{4.6}$$

where

$\boldsymbol{X}$ is the model matrix, with $x_i^{'}$ as ith row;

$\boldsymbol{y}$ is the response vector (containing 0`s and 1`s); and

$\boldsymbol{p_{t-1}}$ is the vector of fitted response probabilities from the previous iteration, the ith entry of which is

$$p_{i,t-1} = \frac{1}{1 + \exp(-x_i{}' b_{t-1})}$$

$\boldsymbol{V_{t-1}}$ is a diagonal matrix, with diagonal entries $p_{i,t-1}\left(1 - p_{i,t-1}\right)$.

We repeat (4.6) until $\boldsymbol{b_t}$ is close enough to $\boldsymbol{b_{t-1}}$. The estimated asymptotic covariance matrix of the coefficients is given by $\left(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}\right)^{-\mathbf{1}}$.

4.5 Goodness of Fit

The validity for all regression models needs to be examined before it is accepted for use and this is usually done using the residuals of the model. We now examine the aptness of the logistic regression model and hence examine whether the estimated response function for the data is monotonic and sigmoidal in shape. Residuals are very important in assessing the adequacy of the fitted model in linear regression analysis. When the dependent variable is binary, each residual can take on only two values, $1 - \hat{\pi}_i$ or 0$-\hat{\pi}_i$ as for the one for the logistic regression models since they are not nearly informative. Therefore, the residuals cannot be expected to provide much direct information about the adequacy of a fitted logistic model. By grouping the data, it is possible to examine the goodness of fit of the logistic response function (Neter et al, 1989).

4.5.1 The Hosmer-Lemeshow Goodness of Fit Test

Hosmer and Lemeshow (2000) propose a statistic that they show, through simulation, is distributed as chi-square when there is no replication in any of the subpopulations. They also state that this test is available only for binary response models. Due to the similarity of the

Hosmer and Lemeshow test for logistic regression, Parzen and Lipsitz (1999) suggest using 10 risk score groups. Nevertheless, based on simulation results, May and Hosmer (2004) show that, for all samples or samples with a large percentage of censored observations, the test rejects the null hypothesis too often. May and Hosmer suggest that the number of groups be chosen such that G=integer of {maximum of 12 and minimum of 10}. The event is the response level specified in the response variable, or the response level that is not specified, or, if neither of these options are specified, then the event is the response level identified in the "Response Profiles". The observations are then divided into approximately 10 groups according to the total following scheme (Parzen and Lipsitz, 1999).

Let N be the total number of subjects. Let M be the target number of subjects for each group given by

M= [0.1×N+0.5].

The Hosmer–Lemeshow test statistic is given by:

$$H = \sum_{g=1}^{n} \frac{\left(O_g - E_g\right)^2}{N_g \pi_g (1 - \pi_g)}.$$

Here O_g, E_g, N_g, and π_g denote the observed events, expected events, observations, predicted risk for the g^{th} risk decile group, and n is the number of groups. The test statistic asymptotically follows a X^2 distribution with *n-2* degrees of freedom. The number of risk groups may be adjusted depending on how many fitted risks are determined by the model. This helps to avoid singular decile groups (Agresti, 2002).

4.6 Odds and Odds Ratios

The concept of odds first arose in gambling. The odds ratio is a measure of effect size, describing the strength of association or non-independence between two binary data values. It is used as a descriptive statistic and plays an important role in logistic regression (Scott, 2010).

Unlike other measures of association for paired binary data such as relative risk, the odds ratio treats the two variables being compared symmetrically, and can be estimated using some type of non-random samples. Scott (2010) also states that logistic regression is one way to generalise the odds ratio beyond two binary variables.

For an event with a given probability value, p, the corresponding odds is a numerical value given by

$$odds\ (event) = \frac{p}{1-p}$$

The odds ratio is a comparative measure of two odds relative to different events. For two probabilities, $p_A = \Pr(event\ A\ occurs)$ and $p_B = \Pr(event\ B\ occurs)$ the corresponding odds of A occurring relative to B occurring is

$$odds\ ratio\ (A\ vs\ B) = \frac{odds\ (A)}{odds\ (B)} = \frac{p_A/(1-p_A)}{p_B/(1-p_B)}$$

4.7 Application to the data set

The most popular SAS procedure for doing ML estimation of the logistic regression model is PROC LOGISTIC. SAS has several other procedures such as Genmod, Catmod (to name a few) that will also do this but we focus on PROC LOGISTIC in this chapter. Logistic regression was applied to the data. In this application the model (logit) was fitted using SAS proc LOGISTIC which is an in built procedure in SAS version 9.1 or version 9.2 capable of fitting both generalised linear models and logistic regression models. The main explanatory variables that were considered are sex, type of death, marital status of deceased, province of birth of deceased, province of death of deceased, place of death of deceased, province of residence of deceased, education status of deceased, smoking status of deceased, and pregnancy status of deceased.

Table 4.1: Type 3 Effects for explanatory variables.

Type 3 Analysis of effects			
Effect	DF	Wald Chi-Square	Pr>chisq
Sex	2	35.9348	<.0001*
Marital status	7	453.9235	<.0001*
Province of birth	9	29.7085	0.0005*
Province of death	9	55.9411	<.0001*
Place of death	5	61.769	<.0001*
Province of residence	9	31.8392	0.0002*
Education status	13	73.5731	<.0001*
Smoking status	2	32.3212	<.0001*
Pregnancy status	2	6.985	0.0304*

The type 3 statistics show that all the explanatory variables (sex, marital status, province of birth, province of death, place of death, province of residence, education status, smoking status, and pregnancy status) were significant at the 5% level.

Table 4.2: Analysis of parameter estimates using PROC LOGISTIC

Analysis of Parameter Estimates								
	Parameter		DF	estimate	Standard Error	Wald Chi-Square	Pr>ChiSq	Odds Ratio
	Intercept		1	-11.7204	59.4466	0.0389	0.0032*	
Sex	Male	1	1	-0.0322	0.155	0.0431	0.8356	1.269
	Female (reference)	2	-	-	-	-	-	-
Marital Status	Single	1	1	0.7745	0.0472	269.041	<.0001*	1.528
	Civil marriage	2	1	-0.2371	0.0777	9.3028	0.0023*	1.452
	Living as married	3	1	-0.0877	0.1125	0.6078	0.4356	1.05
	Widowed	4	1	0.3415	0.0924	13.6555	0.0002*	1.612
	Religious law marriage	5	1	-0.2472	0.1509	2.6819	0.1015	1.467
	Divorced	6	1	-0.178	0.1968	0.8175	0.3659	1.369
	Customary marriage (reference)	7	-	-	-	-	-	-
Province of birth	Western Cape	1	1	-0.2038	0.1718	1.4063	0.2357	0.784
	Eastern Cape	2	1	-0.2044	0.0984	4.3169	0.0877	1.179
	Northern Cape	3	1	0.0106	0.228	0.0022	0.9629	0.951
	Free State	4	1	-0.0891	0.1107	0.6476	0.421	0.879
	KwaZulu-Natal	5	1	-0.1643	0.074	4.9325	0.0264*	1.733
	North West	6	1	0.263	0.1087	5.8553	0.6681	0.739
	Gauteng	7	1	-0.1938	0.0815	5.651	0.0174*	1.492
	Mpumalanga	8	1	0.2731	0.0895	9.3079	0.0023*	1.263
	Limpopo (reference)	9	-	-	-	-	-	-

Table 4.2 (Continue)

Analysis of Parameter Estimates								
	Parameter		DF	estimate	Standard Error	Wald Chi-Square	Pr>ChiSq	Odds Ratio
Province of Death	Western Cape	1	1	-0.7264	0.2244	10.4775	0.982	0.991
	Eastern Cape	2	1	0.1468	0.1584	0.8585	0.3541	1.302
	Northern Cape	3	1	-0.5499	0.2934	3.5129	0.0609	2.614
	Free State	4	1	-0.3004	0.2021	2.2098	0.1371	1.117
	KwaZulu-Natal	5	1	0.4197	0.1418	8.7639	0.0031*	3.119
	North West	6	1	-0.1809	0.1582	1.3075	0.2529	1.259
	Gauteng	7	1	-0.1363	0.1383	0.971	0.0003*	1.729
	Mpumalanga	8	1	-0.5469	0.1527	12.8341	0.3244	0.873
	Limpopo (Reference)	9	-	-	-	-	-	-
Province of residence	Western Cape	1	1	-0.5573	0.2106	7.001	0.1102	1.122
	Eastern Cape	2	1	0.0636	0.1163	0.2989	0.5846	1.226
	Northern Cape	3	1	-0.1078	0.2713	0.1579	0.6911	1.173
	Free State	4	1	-0.1862	0.172	1.1718	0.279	1.085
	KwaZulu-Natal	5	1	0.1524	0.0954	2.551	0.0081*	2.281
	North West	6	1	-0.3668	0.1209	9.2086	0.0024*	1.886
	Gauteng	7	1	-0.0795	0.1016	0.6115	0.0056*	0.955
	Mpumalanga	8	1	-0.3136	0.1132	7.6813	0.4342	1.415
	Limpopo (reference)	9	-	-	-	-	-	-
Place of death	Hospital	1	1	-0.08	0.05	2.5614	0.0015*	1.051
	Emergency room	2	1	-0.3519	0.1108	10.0898	0.1095	0.801
	Dead on arrival	3	1	0.0973	0.1136	0.7349	0.3913	1.033
	Nursing home	4	1	-0.3297	0.1618	4.1535	0.0415*	1.584
	Home (reference)	5	-	-	-	-	-	-

Table 4.2 (Continue)

Analysis of Parameter Estimates								
	Parameter		DF	estimate	Standard Error	Wald Chi-Square	Pr>ChiSq	Odds Ratio
Education	None	0	1	-0.3003	0.0477	39.6752	<.0001*	0.841
	Grade 1	1	1	-0.2236	0.1965	1.2946	0.2552	0.506
	Grade 2	2	1	0.2834	0.161	3.0995	0.0264*	0.0783
	Grade 3	3	1	-0.0592	0.1443	0.1683	0.6816	0.596
	Grade 4	4	1	-0.3083	0.1093	7.9596	0.0048*	0.665
	Grade 5	5	1	-0.0148	0.124	0.0142	0.9052	0.624
	Grade 6	6	1	-0.0446	0.1183	0.1421	0.7062	0.662
	Grade 7	7	1	0.1067	0.0968	1.2151	0.2703	0.569
	Grade 8	8	1	-0.2022	0.11	3.3794	0.066	0.775
	Grade 9	9	1	0.0928	0.1287	0.5197	0.471	0.694
	Grade 10	10	1	-0.0608	0.1038	0.3437	0.5577	0.673
	Grade 11	11	1	-0.2394	0.1275	3.527	0.0604	0.804
	Grade 12	12	1	0.1991	0.0897	4.9265	0.0783	0.477
	University (reference)	13	-	-	-	-	-	-
Smoking	Yes	1	1	0.1921	0.0502	14.6482	0.1382	0.205
	No (reference)	2	-	-	-	-	-	-
Pregnancy	Yes	1	1	-0.7917	0.3028	6.8338	0.1742	0.324
	No (reference)	2	-	-	-	-	-	-

The common practice of interpreting logistic regression estimates is through odds ratios. PROC LOGISTIC also calculates the odds ratio which by default uses the corner point parameterisation for categorical variables where the last category of each variable is used as the reference category. The odds ratios results from SAS using PROC LOGISTIC are presented in the last column of Table 4.2.

The odds ratio of 1.528 implies that single people are 1.528 times more likely than those in a customary marriage to die of diarrhoea. Those in civil marriages are 1.452 times more likely to die from diarrhoea as compared to those that are in customary marriages. We also find that widowed people are 1.612 times more likely to die from diarrhoea compared to those that are in customary marriages. The results show that people who were born in KwaZulu-Natal, Gauteng, and Mpumalanga are 1.733, 1.492 and 1.263 times, respectively, more likely to die from diarrhoea as those who were born in Limpopo. Respondents were 3.119 times more likely to die in KwaZulu-Natal Province than the Limpopo province and are 1.729 times more likely to die in Gauteng province as compared with Limpopo province.

People who are resident in KwaZulu-Natal province are 2.281 times more likely to die from diarrhoea than those who are Limpopo residents. The results also reveal that North West residents are 1.886 times more likely to die from diarrhoea than Limpopo residents. We also find that people who live in Gauteng province are 1.415 times more likely to die from diarrhoea than people who live in Limpopo province. People who are in hospital are 1.051 times more likely to die from diarrhoea than those who are at home. Furthermore the results reveal that people who are in nursing homes are 1.584 times more likely to die from diarrhoea than people who are at home. Uneducated people are 0.841 times more likely to die from diarrhoea than those who have university levels of education. Students who are in grade 2 are 0.772 more likely to die from diarrhoea when compared to those who are at university institutions. Students who are also in grade 4 are 0.465 times more likely to die from diarrhoea compared to those who are at university institutions.

According to a Medical Research Council (MRC) published data that indicates, in South Africa, the under-five mortality rate was 57 per 1000 live births in 2010, translating to around 58 000 children dying in that one year. A Medical Research Council (MRC) review of vital registration data from various sources reveals that in 2007, the majority of registered child deaths in South Africa were infants (76%), with 22% of these deaths occurring in the first month of life. Of the 61335 under-five deaths registered in 2007, diarrhoea accounted for 21% of deaths and lower respiratory infections for 16%.

4.8 SURVEYLOGISTIC

The SURVEYLOGISTIC procedure is similar to the logistic procedure and the other regression procedures such as generalised linear model (Nelder and Wedderburn, 1972) in the SAS system. PROC SURVEYLOGISTIC is developed based on PROC LOGISTIC for logistic regression with survey data. Logistic regression analysis investigates the relationship between discrete responses and sets of explanatory variables. PROC SURVEYLOGISTIC is designed to handle sample data and thus incorporates the sample designs, including designs with stratifications, clustering, and unequal weighting into the analysis, and fits linear logistic regression models for discrete response survey data by the method of maximum likelihood.

The maximum likelihood estimation of the regression coefficients is carried out with either the Fisher-scoring algorithm or the Newton-Raphson algorithm. Variances of the regression parameters and odds ratios are computed using a Taylor expansion approximation (Binder, 1983 and Morel, 1989). The SURVEYLOGISTIC procedure enables one to specify class variables as explanatory variables in the model by using the same syntax for main effects and interactions as in the GLM and logistic procedures.

We now consider PROC SURVEYLOGISTIC, which is designed to handle complex sample surveys with stratifications, clustering, and unequal weighting. On the cluster statement, we simply name the variable that contains the ID numbers for the persons (which are the clusters in this case). If the cluster statements were omitted, we would get the same results that we just saw in Tables 3 and 4. Results with the cluster statement are shown in Table 5 and 6 respectively.

Table 4.3: Type 3 Effects for explanatory variables.

Type 3 Analysis of effects			
Effect	DF	Wald Chi-Square	Pr>chisq
Sex	2	35.0809	<.0001*
Marital status	7	1115.5596	<.0001*
Province of birth	9	75863.7503	<.0001*
Province of death	9	42588.145	<.0001*
Place of death	5	176.9916	<.0001*
Province of residence	9	263.3377	<.0001*
Education status	9	302.8382	<.0001*
Smoking status	2	39.2016	<.0001*
Pregnancy status	2	9.7373	0.0077*

When using the PROC SURVEYLOGISTIC method type 3 statistics shows that all the explanatory variables (sex, marital status, province of birth, province of death, place of death, province of residence, education status, smoking status, and pregnancy status) were significant at the 5% level. Thus all the explanatory variables do influence diarrhea by death.

Table 4.4: Analysis of parameter estimates using PROC SURVEYLOGISTIC.

Analysis of Parameter Estimates								
	Parameter		DF	estimate	Standard Error	Wald Chi-Square	Pr>ChiSq	Odds Ratio
	Intercept		1	-17.8551	0.6799	689.708	<0.001*	
Sex	Male	1	1	-0.0423	0.0463	26.3808	0.7852	1.269
	Female (reference)	2	-	-	-	-	-	-
Marital Status	Single	1	1	0.6383	0.1002	40.5458	<.0001*	1.528
	Civil marriage	2	1	-0.3733	0.1607	5.3972	0.0202*	1.452
	Living as married	3	1	-0.0485	0.1474	0.1083	0.7421	1.05
	Widowed	4	1	0.4777	0.163	8.5902	0.0034*	1.612
	Religious law marriage	5	1	-0.3834	0.2101	3.3286	0.0681	1.467
	Divorced	6	1	-0.3142	0.2875	1.1944	0.2744	1.369

Table 4.4 (Continue)

Analysis of Parameter Estimates								
	Parameter		DF	estimate	Standard Error	Wald Chi-Square	Pr>ChiSq	Odds Ratio
	Customary marriage (reference)	7	-	-	-	-	-	-
Province of birth	Western Cape	1	1	-0.2434	0.135	3.2487	0.0715	0.784
	Eastern Cape	2	1	-0.1648	0.1285	1.6465	0.1994	1.179
	Northern Cape	3	1	0.0502	0.3971	0.016	0.8994	0.951
	Free State	4	1	-0.1286	0.2206	0.3401	0.5598	0.879
	KwaZulu-Natal	5	1	-0.1247	0.0352	12.5832	0.0004*	1.733
	North West	6	1	0.3025	0.1156	6.8491	0.2537	0.739
	Gauteng	7	1	-0.2333	0.0754	9.5792	0.002*	1.492
	Mpumalanga	8	1	0.2335	0.0657	12.6399	0.0004*	1.263
	Limpopo (reference)	9	-	-	-	-	-	-
Province of death	Western Cape	1	1	-1.1374	0.1993	32.5692	0.982	0.991
	Eastern Cape	2	1	0.1312	0.1146	5.3102	0.3541	1.302
	Northern Cape	3	1	-0.9609	0.3021	10.1187	0.0609	2.614
	Free State	4	1	-0.3102	0.2348	0.2219	0.1376	1.117
	KwaZulu-Natal	5	1	0.0087	0.0639	0.0185	0.0031*	3.119
	North West	6	1	-0.2301	0.1744	1.7414	0.287	1.259
	Gauteng	7	1	-0.5473	0.0995	30.2852	0.0003*	1.729
	Mpumalanga	8	1	-0.1359	0.2309	0.3464	0.5561	0.873
	Limpopo (reference)	9	-	-	-	-	-	-
Province of residence	Western Cape	1	1	-0.8247	0.117	49.6787	0.1102	1.122
	Eastern Cape	2	1	0.0238	0.141	2.0901	0.1483	1.226
	Northern Cape	3	1	-0.1596	0.2527	0.3991	0.5276	1.173
	Free State	4	1	-0.0812	0.2593	0.0981	0.7541	1.085
	KwaZulu-Natal	5	1	0.115	0.0665	2.9873	0.0081*	2.281
	North West	6	1	-0.6343	0.1121	31.9994	<.0001*	1.886
	Gauteng	7	1	-0.3469	0.0378	84.2136	0.8633	0.955
	Mpumalanga	8	1	-0.0462	0.2684	0.0297	<.0001*	1.415
	Limpopo (reference)	9	-	-	-	-	-	-

Table 4.3 (Continue)

Analysis of Parameter Estimates								
	Parameter		DF	estimate	Standard Error	Wald Chi-Square	Pr>ChiSq	Odds Ratio
Place of death	Hospital	1	1	-0.0499	0.0893	0.3132	0.0015*	1.051
	Emergency room	2	1	-0.222	0.1439	2.3803	0.1229	0.801
	Dead on arrival	3	1	0.0326	0.0994	0.1075	0.743	1.033
	Nursing home	4	1	-0.4596	0.2326	3.9065	0.0481*	1.584
	Home (reference)	5	-	-	-	-	-	-
Education	None	0	1	-0.7578	0.3159	5.7531	0.0165*	0.841
	Grade 1	1	1	-0.681	0.49	1.9317	0.1646	0.506
	Grade 2	2	1	0.7409	0.3809	3.7828	0.0264*	0.477
	Grade 3	3	1	-0.5167	0.3716	1.9333	0.1644	0.596
	Grade 4	4	1	-0.7658	0.2919	6.881	0.0087*	0.665
	Grade 5	5	1	-0.4723	0.2342	4.066	0.9052	0.624
	Grade 6	6	1	-0.4129	0.3017	1.8734	0.1711	0.662
	Grade 7	7	1	0.5642	0.3142	3.2248	0.0725	0.569
	Grade 8	8	1	-0.2553	0.2234	1.3067	0.253	0.775
	Grade 9	9	1	0.3647	0.3151	1.3393	0.2471	0.694
	Grade 10	10	1	-0.3966	0.3213	1.5238	0.217	0.673
	Grade 11	11	1	-0.2181	0.3842	0.3222	0.5703	0.804
	Grade 12	12	1	0.2584	0.2603	0.9854	0.3209	0.772
	University (reference)	13	-	-	-	-	-	-
Smoking	Yes	1	1	0.1865	0.101	3.4076	0.0649	0.205
	No (reference)	2	-	-	-	-	-	-
Pregnancy	Yes	1	1	-1.2011	0.3849	9.7372	0.1743	0.324
	No (reference)	2	-	-	-	-	-	-

The intercept of the model fitted was significant at 5% level and the estimate of the model was 17.8551. We find that males are 1.269 times more likely to die of diarrhoea than females. This is reflected by the results that show males were more significant than females at 5% level of significance. Single, civil marriage and widowed people were respectively 1.528, 1.452 and 1.612 times more likely than customary marriage people to die because of diarrhoea. Those people who were born in KwaZulu-Natal, Gauteng and Mpumalanga provinces were respectively 1.733, 1.492 and 1.263 times more likely to die from diarrhoea than those who were born in Limpopo province.

More people were dying of diarrhoea the most in KwaZulu-Natal and in Gauteng province than in Limpopo province. KwaZulu-Natal and Gauteng residents were respectively 3.119 and 1.729 times more likely to die from diarrhoea than Limpopo residents. This is reflected by the results which show that KwaZulu-Natal and Gauteng residents were significant at 5% level than those people who were Limpopo residents.

We find that people who were at hospital and nursing home were 1.051 and 1.584 times more likely than home people to die from diarrhoea. The results reveal that uneducated people were more affected by diarrhoea than university students. Grade 2 and grade 4 students were significant as compared to university students at 5% level of significance. Therefore uneducated people, grade 2 and grade 4 students were respectively 0.841, 0.477 and 0.465 times more likely than university students to die from diarrhoea.

Chapter 5

Generalised Linear Mixed Models

5.1 Introduction

In the context of statistical modeling, a generalised linear mixed model (GLMM) is an extension of the linear mixed model in which the linear predictor contains random effects in addition to the usual fixed effects (hence mixed models). These random effects are usually assumed to have a normal distribution (Breslow and Clayton, 1993). This allows the modeling of correlated, possibly non-normally distributed data with flexible accommodation of covariates. Thus the GLMM allows the response or dependent variable to be non-normal in nature.

As discussed in Chapter 3, generalised linear models were formulated by Nelder and Wedderburn (1972) as a way of unifying various other statistical models, including linear regression, logistic regression, and Poisson regression. The generalised linear model allows the model to be related to the response variable via a link function by allowing the magnitude of the variance of each measurement to be a function of its predicted value. The generalised linear mixed model focuses more on the inverse link function rather than the link function to model the relationship between the linear predictor and the conditional mean and also includes nonlinear mixed models (Nelder and Wedderburn, 1972).

In the 1950s, Charles Roy Henderson provided best linear unbiased estimates (BLUE) of fixed effects and best linear unbiased predictions (BLUP) of random effects. It must be stressed again that generalised linear mixed model extends generalised linear models (GLMs) by the inclusion of random effects, and is commonly used for analysis of correlated non-normal over dispersed data. Their broad application to various disciplines, such as longitudinal studies and small area estimation, has been described extensively by Breslow and Clayton (1993). Unfortunately, a full likelihood analysis in GLMMs is often hampered by the need for numerical integration. Several approximate inference procedures have hence been proposed. These include Laplace approximations of the integrated likelihood (Liu and Pierce 1993; Solomon and Cox 1992) and penalised quasi-likelihood procedures (PQL) (Breslow and Clayton 1993; Schall 1991). Compared to the more complicated Laplace approximations, a key feature of the PQL approach is that it is easily implemented by iteratively fitting a linear mixed model to a modified dependent variable. The recently developed SAS macro GLIMMIX using the MIXED procedure (Wolfinger 1993) provides easy access to this PQL method. Good examples of the applications of GLMMs include their use in studying age-specific reproductive success in barn owls (Atlwegg et al. 2007), snow petrels (Angelier et al. 2007), brown bears (Zedrosser et al. 2007), and mountain goats (Côté, Festa-Bianchet 2001), just to name a few of the many applications of the GLMMs.

5.2 The Formal Definition of the Generalised Linear Mixed Model

The GLMM can be defined by

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{e} \tag{5.1}$$

As in the GLM, y is the response/dependent variable, e is the error/residual term and **μ** is the vector of expected means of the observations and is linked to the model parameters by a link function, g:

$$\boldsymbol{g}(\boldsymbol{\mu}) = \boldsymbol{X\alpha} + \boldsymbol{Zb}. \tag{5.2}$$

$\boldsymbol{X}$ and $\boldsymbol{Z}$ are the fixed and random effects design matrices, and **α** and $\boldsymbol{b}$ are the vectors of fixed and random effects parameters. The random effects**,** $\boldsymbol{b}$, can again be assumed to follow a normal distribution:

$$\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{G})$$

and $\boldsymbol{G}$ is the variance-covariance matrix. The variance matrix for the response vector $\boldsymbol{y}$ can be written

$$var(\boldsymbol{y}) = \boldsymbol{V} = var(\boldsymbol{e}) + \boldsymbol{R},$$

where $\boldsymbol{R}$ is the residual variance matrix, $var(\boldsymbol{e})$.

However, $\boldsymbol{V}$ is not as easily specified as it was for normal data where $\boldsymbol{V} = \boldsymbol{ZGZ'} + \boldsymbol{R}$**.** This is because **μ** is now not a linear function of the fixed effects $\boldsymbol{\alpha}$.

5.3 The Three Specifications in Generalised Linear Mixed Model

We define, firstly,

Linear Predictor, η

As with the linear mixed model, the fixed and random effects are combined to form a linear predictor

$$\eta = \boldsymbol{X\beta} + \boldsymbol{Zu}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are $n \times p$, $n \times q$ design matrix for fixed effects parameters **(β**) and random parameters ($\boldsymbol{u}$) respectively.

With the linear mixed model the model for the vector of observations $\boldsymbol{y}$ is obtained by adding a vector of residuals, $\boldsymbol{\varepsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{R})$ as follows:

$$\boldsymbol{y} = \eta + \boldsymbol{\varepsilon} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon}$$

Equivalently, the residual variability can be modeled as

$$\boldsymbol{y}|\boldsymbol{u} \sim \boldsymbol{N}(\eta, \boldsymbol{R})$$

Unless y has a normal distribution, the formulation using $\boldsymbol{\varepsilon}$ is clumsy. Therefore, a generalised linear mixed model uses a second approach to model the residual variability. The relationship between the linear predictor and the vector of observations in the generalised linear mixed model is modeled as

$$\boldsymbol{y}|\boldsymbol{u}\sim\boldsymbol{N}(\boldsymbol{h}(\eta),\boldsymbol{R})$$

where the notation, $\boldsymbol{y}|\boldsymbol{u}\sim\boldsymbol{N}(\boldsymbol{h}(\eta),\boldsymbol{R})$, specifies that the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{u}$ has mean $\boldsymbol{h}(\eta)$, and variance, $\mathbf{R}$.

The conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{u}$ will be referred to as the error distribution. The choice of which fixed and random effects to include in the model will follow the same considerations as for a linear mixed model. It is important to note that the effect of the linear predictor is expressed through an inverse link function. Except for the identity link function, $\boldsymbol{h}(\eta)=\eta$, the effect of a one unit change in η_i will not correspond to a one unit change in the conditional mean; that is, predicted progeny difference will depend on the progeny`s environment through $\boldsymbol{h}(\eta)$ (Breslow and Clayton. 1993).

Link function

The second specification of a generalised linear mixed model is the selection of a link function $g(.)$ which converts the expected value $\boldsymbol{\mu}$ of the outcome variable $\boldsymbol{Y}$ to the linear predictor η**.**

$$\boldsymbol{g}(\boldsymbol{\mu})=\eta$$

Here, the expected value of the outcome is conditional on the random effects [i.e, $\boldsymbol{\mu}=\boldsymbol{E}(\boldsymbol{Y}|\boldsymbol{u})$**].**

The natural link is so-called the logistic or logit link:

$$\eta=\mathbf{log}(\boldsymbol{\mu}/(\mathbf{1}-\boldsymbol{\mu}))$$

where

$$\boldsymbol{\mu}=\boldsymbol{e}^{\eta}/(\mathbf{1}+\boldsymbol{e}^{\eta})$$

But others are in common usage such as the probit link

$\eta=\boldsymbol{\Phi}^{-\mathbf{1}}(\boldsymbol{\mu}),\boldsymbol{\mu}=\boldsymbol{\Phi}(\eta)$

where $\varPhi$ is the standard normal distribution function. The variance function has the form $\boldsymbol{v}(\boldsymbol{\mu}) = \boldsymbol{\mu}(\mathbf{1} - \boldsymbol{\mu})$ and the scale parameter is known, $\phi = 1$.

Inverse Link Function

The inverse link function is used to map the value of the linear predictor for observation i, η_i, to the conditional mean for observation i, μ_i. For many traits the inverse link function is one to one, that is both μ_i and η_i are scalars. For threshold models, $\boldsymbol{\mu_i}$ is a $t \times 1$ vector, where t is the number of ordinal levels. For growth curve models, $\boldsymbol{\mu_i}$ is an $n_i \times 1$ vector and η_i is a $p \times 1$ vector.

For the linear mixed model, the inverse link function is the identity function $\boldsymbol{h}(\eta) = \eta_{\boldsymbol{i}}$. For zero/one traits a logit link function $\eta_{\boldsymbol{i}} = \mathbf{ln}(\boldsymbol{\mu_i}/[\mathbf{1} - \boldsymbol{\mu_i}])$ is often used, the corresponding inverse link function is $\boldsymbol{\mu_i} = \frac{e^{\eta_i}}{1+e^{\eta_i}}$. The logit link function, unlike the identity link function, will always yield estimated means in the range of zero to one. However, the effect of a one unit change in the linear predictor is not constant. For most univariate link functions, link and inverse link functions are increasing monotonic functions. In other words, an increase in the linear predictor results in an increase in the conditional mean, but not at a constant rate. Selection of inverse link functions is typically based on the error distribution. Table 5.1 lists a number of common distributions along with their link functions.

Table 5.1: Common link functions and variance functions for various distributions

Distribution	**Link**	**Inverse Link**	$\boldsymbol{v(\mu)}$
Normal	Identity	η	1
Binomial/n	Logit	$e^{\eta}/(1+e^{\eta})$	$\mu(1-\mu)/n$
	Probit	$\Phi(\eta)$	
Poisson	Log	e^{η}	μ
Gamma	Inverse	$1/\eta$	μ^2
	Log	e^{η}	

5.4 The Likelihood and Quasi-likelihood Functions

The method of fitting the GLMMs is based on maximising the likelihood function for the model parameters. However, a difficulty with this is that true likelihood functions can only be defined for random effects and random coefficient models. Obtaining consistent and efficient estimators for the regression and the overdispersion parameters in GLMMs has proven to be difficult. For GLMMs with multidimensional random effects, Schall (1991) and Breslow and Clayton (1993), among others, use an approach analogous to the best linear unbiased prediction (BLUP), where random effects are treated as fixed effects (Henderson, 1963). In the following section we will specify the likelihood function for random effects and random coefficients models, define a quasi-likelihood function for covariance pattern model, and then give a general form of the quasi-likelihood function that is appropriate for all types of mixed model.

5.5 The Likelihood Function for Random Effects and Random Coefficients Models

For these models we can obtain a true likelihood function from the product of the likelihoods based on $\boldsymbol{y}|\boldsymbol{b}$ and $\boldsymbol{b}$. A true likelihood function is possible because the distributions of $\boldsymbol{y}|\boldsymbol{b}$ and $\boldsymbol{b}$ are known (binomial, Poisson, etc), hence likelihood functions can be formed from them. The likelihood for the fixed effects**, α**, and the variance parameters in the $\boldsymbol{G}$ matrix, $\boldsymbol{\gamma_G}$, can be written

$$\boldsymbol{L(\alpha, \gamma_G; y) = L(\alpha; y|b) L(\ \gamma_G; b)}. \tag{5.3}$$

Now $\boldsymbol{b}$ is assumed to have a multivariate normal distribution, $\boldsymbol{b \sim N(0, G)}$, so substituting the multivariate normal density for $\boldsymbol{L(\ \gamma_G; b)}$ we have

$$\boldsymbol{L(\alpha, \gamma_G; y) \propto \mathrm{L(\alpha;y|}b)|G|^{-1/2} \exp(-1/2\, \beta' G^{-1} b)}.$$

The $\boldsymbol{y}|\boldsymbol{b}$ are independent because we have assumed uncorrelated residuals ($\boldsymbol{R}$ is diagonal) and therefor $\boldsymbol{L(\alpha}; \boldsymbol{y}|\boldsymbol{b})$ is simply defined using the assumed distribution of $\boldsymbol{y}|\boldsymbol{b}$. This can be expressed using the same form obtained in Chapter 3 for the GLMs:

$$\begin{aligned}\boldsymbol{L(\alpha; y|b)} &= \mathbf{exp}\left[\boldsymbol{y' A^{-1} \theta - b(\theta)^{1/2'} A^{-1} b(\theta)^{1/2} + K}\right] \\ &\propto \mathbf{exp}\left[\boldsymbol{y' A^{-1} \theta - b(\theta)^{1/2'} A^{-1} b(\theta)^{1/2}}\right],\end{aligned}$$

where

$\boldsymbol{\theta = X\alpha + Z\beta}$,

$\boldsymbol{A} = diag\{a_i\}$, where a_i are constant terms

$\boldsymbol{b(\theta)} = (b(\theta_1), b(\theta_2), \dots.., b(\theta_n)\ )'$, where $\boldsymbol{b}$ is the function used in the general distribution form and

$\boldsymbol{K}$=constant.

The overall likelihood for α and γ_G can then be expressed as

$$\boldsymbol{L(\alpha,\gamma_G;y)} \propto \mathbf{exp}\left[\boldsymbol{y}'\boldsymbol{A}^{-1}\boldsymbol{\theta} - \boldsymbol{b(\theta)}^{1/2'}\boldsymbol{A}^{-1}\boldsymbol{b(\theta)}^{1/2}\right]|\boldsymbol{G}|^{-1/2}\,\mathbf{exp}(-1/2\,\boldsymbol{\beta}'\boldsymbol{G}^{-1}\boldsymbol{b}),$$

and the log likelihood as

$$\mathbf{log}\{\boldsymbol{L(\alpha,\gamma_G;y)}\} = \boldsymbol{y}'\boldsymbol{A}^{-1}\boldsymbol{\theta} - \boldsymbol{b(\theta)}^{1/2'}\boldsymbol{A}^{-1}\boldsymbol{b(\theta)}^{1/2} - 1/2\,\boldsymbol{log}|\boldsymbol{G}| - 1/2\,\boldsymbol{\beta}'\boldsymbol{G}^{-1}\boldsymbol{b} + \boldsymbol{K}. \qquad (5.4)$$

5.6 Likelihood and Maximum Likelihood

For a generalised linear mixed model the distribution of the response vector $\boldsymbol{y}$ depends on a vector quantity η which is related to vector regression variables through the equation

$$\eta = \boldsymbol{X\alpha} + \boldsymbol{Zb}$$

as in the previous section. Let $f(\boldsymbol{y};\boldsymbol{\alpha} \mid \boldsymbol{b})$ be the probability (density) function of $\boldsymbol{y}$ conditional on fixed $\boldsymbol{b}$. The log-likelihood of the observation vector $\boldsymbol{y} = (Y_1, Y_2, \dots, Y_n)$ conditional on fixed **b** is

$$\boldsymbol{l_1} = \mathbf{ln}\,\boldsymbol{f}(\boldsymbol{y};\boldsymbol{\alpha} \mid \boldsymbol{b}).$$

The likelihood of the random component vector $\boldsymbol{b}$ is

$$\boldsymbol{l_2} = \boldsymbol{constant} - \frac{1}{2}\sum_{j=1}^{k}\{\boldsymbol{v_j}\,\mathbf{ln}(2\boldsymbol{\pi\sigma_j}^2) + \boldsymbol{\sigma_j}^{-2}\boldsymbol{b_j}'\boldsymbol{A_j}^{-1}\boldsymbol{b_j}\}$$

And the joint log-likelihood of $\boldsymbol{y}$ and $\boldsymbol{b}$ is $\boldsymbol{l} = \boldsymbol{l_1} + \boldsymbol{l_2}$. The derivatives of $\boldsymbol{l}$ are

$$\partial \boldsymbol{l}/_{\partial \boldsymbol{\alpha}} = \partial \boldsymbol{l_1}/_{\partial \boldsymbol{\alpha}'}$$

$$\partial \boldsymbol{l}/_{\partial \boldsymbol{b_j}} = \partial \boldsymbol{l_1}/_{\partial \boldsymbol{b_j}} - \boldsymbol{\sigma_j}^{-2}\boldsymbol{A_j}^{-1}\boldsymbol{b_j},\ \ j = 1, 2, \dots, k.$$

The second-order derivatives of $\boldsymbol{l}$ which involve at least one $\boldsymbol{\alpha}$ are the same as the second-order derivatives of $\boldsymbol{l_1}$ while

$$\boldsymbol{\partial^2 l}\big/_{\boldsymbol{\partial b_j b_j}'} = \boldsymbol{\partial^2 l_1}\big/_{\boldsymbol{\partial b_j b_j}'} - \boldsymbol{\sigma_j}^{-2}\boldsymbol{A_j}^{-1},$$

$$\boldsymbol{\partial^2 l}\big/_{\boldsymbol{\partial b_j b_j}'} = \boldsymbol{\partial^2 l_1}\big/_{\boldsymbol{\partial b_j b_j}'}.$$

Simulated Maximum Likelihood for GLMM

Geyer and Thompson (1992) and Gelfand and Carlin (1993) suggest simulation to estimate the GLMM likelihood which can then be numerically maximised.

$$£(\boldsymbol{\alpha}, \boldsymbol{D}|\boldsymbol{y}) = \int \boldsymbol{f(y|\alpha, b)f(b|D)db}$$

$$= \int \frac{\boldsymbol{f(y|\alpha,b)f(b|D)}}{\boldsymbol{g(b)}}\boldsymbol{g(b)db}$$

$$\approx \frac{\mathbf{1}}{\boldsymbol{M}}\sum_{\boldsymbol{k}=\mathbf{1}}^{\boldsymbol{M}} \frac{\boldsymbol{f(y|\alpha,b^{(k)})f(b^{(k)}|D)}}{\boldsymbol{g(b^{(k)})}}$$

where $\boldsymbol{b}^{(k)}$ is drawn from the importance sampling distribution $g(\boldsymbol{b})$. To maximise the likelihood, we need to evaluate it at different values of $(\boldsymbol{\alpha}, \boldsymbol{D})$.

5.7 Marginal and Penalized Quasi-Likelihood (MQL and PQL)

We consider two of the several versions based on the approximation of the data (Molenberghs and Verbeke, 2005). We first consider the decomposition of Y_{ij} as

$$Y_{ij} = \mu_{ij} + \in_{ij} = h\left(x_{ij}'\beta + z_{ij}'b_i\right) + \in_{ij} \qquad (5.5)$$

where $h = g^{-1}(x_{ij}{'}\beta + z_{ij}{'}b_i)$ is the inverse link function. Assume the error terms follow a distribution with variance equal to $Var(Y_{ij}|b_i) = \phi v(\mu_{ij})$ where $v(.)$ is the variance function in the exponential family.

Then under the natural or canonical link function it follows that

$$v(\mu_{ij}) = h{'}(x_{ij}\beta + z_{ij}b_i) \tag{5.6}$$

where the derivative is with respect to μ_{ij}.

For illustration consider binary outcomes with the logit canonical link function.

We then have

$$\mu_{ij} = P(Y_{ij} = 1|b_i) = \pi_{ij} = \frac{\exp(x_{ij}{'}\beta + z_{ij}{'}b_i)}{1+\exp(x_{ij}{'}\beta + z_{ij}{'}b_i)} = h(x_{ij}{'}\beta + z_{ij}{'}b_i) \tag{5.7}$$

Note that from (5.5) $\in_{ij} = Y_{ij} - \mu_{ij}$. Since $\mu_{ij} = \pi_{ij}$ and $Y_{ij} = 1$ or 0. This implies that $\in_{ij} = 1 - \pi_{ij}$ with probability π_{ij} and $\in_{ij} = -\pi_{ij}$ with probability $1 - \pi_{ij}$ hence $E(\in_{ij}) = 0$ and $Var(\in_{ij}) = \pi_{ij}(1 - \pi_{ij})$. Note that π_{ij} is the conditional mean of Y_{ij} given b_i. Estimation then proceeds by using Taylor`s linear approximation to $h(x_{ij}{'}\boldsymbol{\beta} + z_{ij}{'}b_i)$ about $\hat{\theta} = (\hat{\beta}, \widehat{b_\iota})'$.

Penalized Quasi-Likelihood (PQL)

We first discuss a linear Taylor expansion of (5.5) around current estimates $\widehat{\boldsymbol{\beta}}$ of the fixed effect and $\widehat{b_\iota}$ of the random effect assuming canonical or natural link. This gives

$$\begin{aligned} Y_{ij} &\approx h(x_{ij}{'}\hat{\beta} + z_{ij}{'}\widehat{b_\iota}) \\ &+h{'}(x_{ij}{'}\hat{\beta} + z_{ij}{'}\widehat{b_\iota})x_{ij}{'}(\beta - \hat{\beta}) \\ &+h{'}(x_{ij}{'}\hat{\beta} + z_{ij}{'}\widehat{b_\iota})z_{ij}{'}(b_i - \widehat{b_\iota}) + \in_{ij} \\ &= \hat{\mu}_{ij} = v(\hat{\mu}_{ij})x_{ij}{'}(\beta - \hat{\beta}) + v(\hat{\mu}_{ij})z_{ij}{'}(b_i - \widehat{b_\iota}) + \in_{ij} \end{aligned}$$

where $\hat{\mu}_{ij}$ equals its current predictor $h(x_{ij}{'}\hat{\beta} + z_{ij}{'}\widehat{b_\iota})$ of the conditional mean $E(Y_{ij}|b_i)$.

In vector form this becomes

$$\boldsymbol{Y_i} \approx \widehat{\boldsymbol{\mu}}_i + \widehat{\boldsymbol{V}}_i\boldsymbol{X}_i(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + \widehat{\boldsymbol{V}}_i\boldsymbol{Z}_i(\boldsymbol{b}_i - \widehat{\boldsymbol{b}_\iota}) + \in_i \qquad (5.8)$$

for appropriate design matrixes $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$ and with $\boldsymbol{V_i}$ equal to the diagonal matrix with diagonal entries equal to $v(\hat{\mu}_{ij})$.

Re-ordering terms in the above equation yields

$$\boldsymbol{Y_i}^* = \widehat{\boldsymbol{V}_i}^{-1}(\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_i) + \boldsymbol{X}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\widehat{\boldsymbol{b}_\iota} \approx \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \in_i^* \qquad (5.9)$$

for $\in_i^* = \widehat{\boldsymbol{V}_i}^{-1} \in_i$, which still have a mean of zero.

The modified response $\boldsymbol{Y_i}^*$ allows us to approximate the problem as a linear mixed model. The approximate linear mixed model in (5.8) is used to obtain update estimates for **β,**$\boldsymbol{D}$, $\boldsymbol{\phi}$ using readily available procedures fitting linear mixed models. The resulting estimates are called penalised quasi-likelihood estimates (PQL) because they are obtained by optimising a quasi-likelihood function which only involves first and second order conditional moments, augmented with a penalty term on the random effects. We refer to Breslow and Clayton (1993) and Wolfinger and O´Connell (1993) for its implementation and programming in SAS.

Marginal Quasi-Likelihood (MQL)

This is an alternative approximation method very similar to the PQL method but it is also based on a linear Taylor expansion of the mean μ_{ij} of (5.5) around the current estimates $\widehat{\boldsymbol{\beta}}$ for the fixed effects but around $b_i = 0$ for random predictor. This gives a similar expression as that derived under PQL method with $b_i = 0$. The current predictor $\hat{\mu}_{ij}$ is now of the form $h(x_{ij}{}'\hat{\beta})$ instead of $h(x_{ij}{}'\hat{\beta} + z_{ij}{}'\widehat{b_\iota})$ as was the case under the PQL method. Re-ordering the terms gives $\boldsymbol{Y_i}^* = \widehat{\boldsymbol{V}_i}^{-1}(\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_i) + \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}$ which also satisfies the approximate linear mixed model

$$\boldsymbol{Y_i}^* \approx \boldsymbol{X}_i\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{V}_i}^{-1}(\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_i) \approx \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \in_i^*$$

similar to (5.9).

The resulting estimates are called marginal quasi-likelihood (MQL). They are obtained by optimising a quasi-likelihood function which only involve first and second order moments but now evaluated at the marginal linear predictor $\boldsymbol{x_i}'\boldsymbol{\beta}$ rather than the conditional linear predictor $\boldsymbol{x_i}'\boldsymbol{\beta} + \boldsymbol{z_i}'\boldsymbol{b_i}$.

5.8 Comparison of (PQL) and (MQL) Methods

The quasi-likelihood (Wedderburn, 1974; McCullagh, 1983) is useful especially in modeling overdispersed count data or overdispersed binomial data (that is, greater variability in the data than would be expected from the statistical model used), in which case the likelihood approach can be complicated. The commonly known penalised quasi-likelihood (PQL) procedure, as synthesised and popularised by Breslow and Clayton (1993), offers a means for approximate inference in generalised linear mixed models (GLMMs). The PQL method of estimation and inference, which is known to be relatively straight forward to implement, has been explored for its potential use in small area disease risk predictions and inference in the context of Bayesian disease mapping (Breslow and Clayton, (1993), Leroux et al., (1999), Macnab et al., (2004), Dean et al., (2004), Ainsworth and Dean, (2006) and Ugarte et al., (2008). The difference between PQL and MQL is that MQL does not incorporate the random effects in the linearisation process but both methods have the same key idea and will ideally have similar properties. The MQL estimation performs well if the random effects variance is very small. Rodriguez and Goldman (1995) show that both PQL and MQL may be seriously biased when applied to binary response data. Their simulations reveal that the fixed effects and variance components suffer from substantial, if not severe, attenuated measurements per cluster while with an increasing number of measurements per subject, MQL remains biased and PQL becomes consistent. Breslow and Lin (1995) suggest the inclusion of bias correction terms while Kuk (1995) suggest the use of iterative bootstrap. Goldstein and Rasbash (1996) show that one of the ways to improve the accuracy of the approximations is to include a second order term in the Taylor series expansion. They call these methods PQL2 and MQL2. Goldstein and Rasbash (1996) state that MQL2 performs slightly better than MQL but PQL2 is substantially better than PQL. Within the PQL and MQL methods, the linear mixed model approximation can be based on maximum likelihood estimation (ML) or restricted maximum likelihood estimation (REML) resulting in slightly different results.

5.9 SAS Software for Fitting Generalised Linear Mixed Model

The models can be fitted via a number of available statistical software programmes such as SAS, Genstat, and many more. In the current work we focus on SAS applications since this software will be used to analyse our data. SAS PROC GLIMMIX is capable for fitting statistical models to data with both random and fixed effects and where the response is not necessarily normally distributed. The GLIMMIX procedure generalises the MIXED and GENMOD procedures in two important ways. First, the response can have a non-normal distribution. The MIXED procedure assumes that the response is normally (Gaussian) distributed. Second, the GLIMMIX procedure incorporates random effects in the model and so allows for subject-specific (conditional) and population-average (marginal) inference. The GENMOD procedure allows only for marginal inference. PROC GLIMMIX performs both estimations and statistical inference for generalised linear mixed models. The GLIMMIX procedure can also fit models for non-normal data with hierarchical random effects, provided that the random effects have a normal distribution. The default estimation method in PROC GLIMMIX for models containing random effects is a technique known as restricted Pseudo-likelihood (RPL) estimation (Wolfinger and O`Connell, 1993).

To fit GLMMs via Gaussian and adaptive Gaussian quadrature methods in SAS, PROC NLMIXED is used. The NLMIXED procedure fits nonlinear mixed models where the conditional mean function is a general nonlinear function. The class of generalised linear mixed models is a special case of nonlinear mixed models; hence some of the models we can fit with PROC NLMIXED can also be fitted with the GLIMMIX procedure. The NLMIXED procedure relies by default on approximating the marginal log likelihood through adaptive Gaussian quadrature. In the GLIMMIX procedure, maximum likelihood estimation by adaptive Gaussian quadrature is available with the METHOD=QUAD option in the GLIMMIX statement. The default estimate methods thus differ between the NLMIXED and GLIMMIX procedures, because adaptive quadrature is possible for only a subset of the models available with the GLIMMIX procedure. If one chooses METHOD=QUAD in the PROC GLIMMIC statement for the generalised linear mixed model, the GLIMMIX procedure performs maximum likelihood estimation based on Laplace approximation of the marginal log likelihood. PROC NLMIXED computes derivatives of the adaptive Gaussian quadrature approximation and the default method used is dual quasi-Newton optimisation. The main advantage of NLMIXED is that the user is given a high degree of flexibility in the way the model is specified and parameterised. In the current application both PROC GLIMMIX and PROC NLMIXED are used.

The following results were obtained using PROC GLIMMIX where the household number was chosen as a random variable. The reason for using the random intercept in the model is because random intercept allow the overall level of the linear predictor to vary between clusters and above the variability explained by the covariates.

Table 5.2: Type 3 effects for explanatory variables.

Type 3 Analysis of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Sex	2	65472	17.97	<.0001*
Marital status	7	65472	64.85	<.0001*
Province of birth	9	65472	3.35	<.0001*
Province of death	9	65472	6.21	<.0001*
Place of death	5	65472	12.35	<.0001*
Province of residence	9	65472	3.54	0.0002*
Education status	13	65472	5.66	<.0001*
Smoking status	2	65472	16.16	<.0001*
Pregnancy status	2	65472	3.49	0.0304*

Table 5.2 reflects Type 3 analysis for fixed effects. The fixed effects parameter estimates and the type 3 analysis for the fixed effects result indicate that all the explanatory variables were significant at 5% significance level.

Table 5.3: Analysis of parameter estimates using PROC GLIMMIX

Analysis of Parameter Estimates								
	Parameter		DF	Estimate	Standard Error	t Value	Pr>ltl	Odds Ratio
	Intercept		65472	-15.878	44.191	-0.36	<.0001*	
Sex	Male	1	65472	-0.0238	0.0403	-5.91	0.0635	0.788
	Female (reference)	2	-	-	-	-	-	-
Marital status	Single	1	65472	0.6386	0.0825	7.74	0.0635	1.893
	Civil marriage	2	65472	-0.3728	0.1093	-3.41	0.0006*	0.688
	Living as married	3	65472	-0.0485	0.1434	-0.34	0.7352	0.62
	Widowed	4	65472	0.4774	0.123	-3.88	0.0001*	0.953
	Religious law married	5	65472	-0.383	0.185	-2.07	0.0783	0.682
	Divorced	6	65472	-0.3137	0.2358	-1.33	0.1833	0.73
	Customary marriage (reference)	7	-	-	-	-	-	-
Province of birth	Western Cape	1	65472	-0.2431	0.2078	1.17	0.2399	1.274
	Eastern Cape	2	65472	-0.1654	0.1361	-1.22	0.2255	0.848
	Northern Cape	3	65472	0.0490	0.2685	0.18	0.8517	1.052
	Free State	4	65472	-0.1284	0.1463	0.88	0.3786	1.137
	KwaZulu-Natal	5	65472	-0.1251	0.1128	-1.11	0.0355*	1.353
	North West	6	65472	0.3018	0.1439	2.1	0.269	0.883
	Gauteng	7	65472	-0.2329	0.1159	2.01	0.0442*	1.263
	Mpumalanga	8	65472	0.2335	0.1128	-2.07	0.0387*	1.392
	Limpopo (reference)	9	-	-	-	-	-	-
Province of death	Western Cape	1	65472	-1.1373	0.2519	-4.51	0.9544	1.009
	Eastern Cape	2	65472	0.2642	0.1781	-1.48	0.1379	0.768
	Northern Cape	3	65472	-0.9607	0.3309	-2.9	0.0037*	2.614
	Free State	4	65472	-0.1104	0.2272	-0.49	0.6262	0.895
	KwaZulu-Natal	5	65472	0.0898	0.1571	0.06	0.0289*	3.119
	North West	6	65472	-0.2299	0.1796	-1.28	0.2004	0.794
	Gauteng	7	65472	-0.5471	0.1528	-3.58	0.0003*	1.729
	Mpumalanga	8	65472	-0.1357	0.1398	0.97	0.3317	1.146
	Limpopo (reference)	9	-	-	-	-	-	-

Table 5.3 (Continue)

Analysis of Parameter Estimates								
	Parameter		DF	Estimate	Standard Error	t Value	Pr>ltl	Odds Ratio
Province of residence	Western Cape	1	65472	-0.8247	0.2627	-3.14	0.459	1.122
	Eastern Cape	2	65472	0.2032	0.1741	-1.17	0.2416	1.226
	Northern Cape	3	65472	-0.1587	0.3263	-0.49	0.6244	1.173
	Free State	4	65472	-0.0812	0.2232	-0.36	0.7157	1.085
	KwaZulu-Natal	5	65472	0.1149	0.1554	-0.74	0.0017*	2.281
	North West	6	65472	-0.6338	0.177	-3.58	0.0003*	1.886
	Gauteng	7	65472	-0.3468	0.1583	-2.19	0.0284*	1.415
	Mpumalanga	8	65472	-0.0459	0.1442	0.32	0.7485	0.955
	Limpopo (reference)	9	-	-	-	-	-	-
Place of death	Hospital	1	65472	-0.0499	0.0411	-1.21	0.0173*	1.584
	Emergency room	2	65472	-0.2223	0.1281	1.73	0.0832	0.801
	Dead on arrival	3	65472	0.0327	0.1316	-0.25	0.804	1.033
	Nursing home	4	65472	-0.4598	0.193	-2.38	0.2241	1.051
	Home (reference)	5	-	-	-	-	-	-
Education	None	0	65472	-0.7595	0.2256	3.37	0.0008*	0.854
	Grade 1	1	65472	-0.6822	0.3055	2.23	0.0713	0.506
	Grade 2	2	65472	0.742	0.2802	2.65	0.0082*	0.621
	Grade 3	3	65472	-0.5188	0.2693	1.93	0.055	0.596
	Grade 4	4	65472	-0.7679	0.2493	3.08	0.0021*	0.745
	Grade 5	5	65472	-0.4741	0.2572	1.84	0.0662	0.224
	Grade 6	6	65472	-0.4147	0.254	1.63	0.104	0.362

Table 5.3 (Continue)

Analysis of Parameter Estimates								
	Parameter		DF	Estimate	Standard Error	t Value	Pr>ltl	Odds Ratio
Education	Grade 7	7	65472	0.5657	0.2431	2.33	0.414	0.469
	Grade 8	8	65472	-0.2571	0.2494	1.03	0.306	0.575
	Grade 9	9	65472	0.3665	0.2596	1.41	0.1601	0.194
	Grade 10	10	65472	-0.3982	0.2462	1.62	0.1071	0.373
	Grade 11	11	65472	-0.2198	0.2588	0.85	0.3996	0.204
	Grade 12	12	65472	0.2599	0.2394	1.09	0.2805	0.372
	University (reference)	13	-	-	-	-	-	-
Smoking	Yes	1	65472	0.1866	0.0812	-2.3	0.0911	0.205
	No (reference)	2	-	-	-	-	-	-
Pregnancy	Yes	1	65472	-1.2015	0.4561	-2.63	0.0855	0.324
	No (reference)	2	-	-	-	-	-	-

Table 5.3 reveals to us that the intercept for the model fitted was significant at the 5% significance level since the p-value was smaller than 0.05. Single, civil marriage and widowed people were significant at 5% level as compared to customary married people. Therefore single, civil married and widowed people were more likely to die from diarrhoea than customary married people. The result also tells us that single, civil married and widowed people were respectively 1.893, 0.688 and 0.953 times more likely to die from diarrhoea than customary married people. People who were born in KwaZulu-Natal, Gauteng and Mpumalanga province were more likely to die from diarrhoea then people born in Limpopo province. This is reflected by the results which show that people in KwaZulu-Natal, Gauteng and Mpumalanga province were respectively 1.353, 1.263 and 1.392 times more likely to die from diarrhoea than people in Limpopo province. We find that people were more likely to die from diarrhoea in the following provinces: KwaZulu-Natal, Northern Cape and Gauteng province, as compared to Limpopo

province. This is also reflected by the results since KwaZulu-Natal, Northern Cape and Gauteng provinces were significant at 5% levels of significance as compared to Limpopo province. We also find that the odds ratio for KwaZulu-Natal, Northern Cape and Gauteng provinces were 3.119, 2.614 and 1.729 respectively.

The result reveals that (for province of residence) KwaZulu-Natal, North West and Gauteng provinces were significant at 5% levels when compared to Limpopo province, which means that KwaZulu-Natal, North West and Gauteng residents were respectively 2.281, 1.886 and 1.415 times more likely than Limpopo residents to die from diarrhoea. There is a significant difference at the 5% level between hospital and home deaths with respect to diarrhoea. People in hospitals are 1.584 times more likely to die from diarrhoea at 5% level of significance than people in homes. We find that uneducated people, Grade 2 and Grade 4 students were significant at 5% levels when compared to university students. This tell us that uneducated people, Grade 2 and Grade 4 students were respectively 0.854, 0.621 and 0.745 times more likely to die from diarrhoea compared to university students.

5.9.1 Direct Estimation via PROC NLIMIXED

The result for fitting the model using proc NLMIXED assuming a random intercept model (allows the overall level of the linear predictor to vary between clusters and above the variability explained by the covariates) are given in the Tables 5.4 and 5.5 respectively.

Table 5.4: Type 3 effects for explanatory variables

Type 3 Analysis of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Sex	2	65472	17.97	<.0001*
Marital status	7	65472	64.84	<.0001*
Province of birth	9	65472	3.3	<.0001*
Province of death	9	65472	6.22	<.0001*
Place of death	5	65472	12.35	<.0001*
Province of residence	9	65472	3.54	0.0002*
Education status	13	65472	5.66	<.0001*
Smoking status	2	65472	16.16	<.0001*
Pregnancy status	2	65472	3.48	0.0304*

Table 5.5: Analysis of parameter estimates using PROC NLMIXED

Analysis of Parameter Estimates							
	Parameter		DF	estimate	Standard Error	t Value	Pr>ItI
	Intercept		65472	-15.878	44.191	-0.36	<.0001*
Sex	Male	1	65472	-0.0238	0.0403	-5.91	0.4281
	Female (reference)	2	-	-	-	-	-
Marital status	Single	1	65472	0.6386	0.0825	7.74	<.0001*
	Civil marriage	2	65472	-0.3728	0.1093	-3.41	0.0006*
	Living as married	3	65472	-0.0485	0.1434	-0.34	0.7352
	Widowed	4	65472	0.4774	0.123	-3.88	0.0001*
	Religious law married	5	65472	-0.383	0.185	-2.07	0.0783
	Divorced	6	65472	-0.3137	0.2358	-1.33	0.1833
	Customary marriage	7	-	-	-	-	-

Table 5.5 (Continue)

Analysis of Parameter Estimates							
	Parameter		DF	estimate	Standard Error	t Value	Pr>ltl
Province of birth	Western Cape	1	65472	-0.2431	0.2078	1.17	0.2399
	Eastern Cape	2	65472	-0.1654	0.1361	-1.22	0.2255
	Northern Cape	3	65472	0.04901	0.2685	0.18	0.8517
	Free State	4	65472	-0.1284	0.1463	0.88	0.3786
	KwaZulu-Natal	5	65472	-0.1251	0.1128	2.1	0.0355*
	North West	6	65472	0.3018	0.1439	-1.11	0.269
	Gauteng	7	65472	-0.2329	0.1159	2.01	0.0442*
	Mpumalanga	8	65472	0.2335	0.1128	-2.07	0.0387*
	Limpopo (reference)	9	-	-	-	-	-
Province of death	Western Cape	1	65472	-1.1373	0.2519	0.06	0.9544
	Eastern Cape	2	65472	0.2642	0.1781	-1.48	0.1379
	Northern Cape	3	65472	-0.9607	0.3309	-2.9	0.0037*
	Free State	4	65472	-0.1104	0.2272	-0.49	0.6262
	KwaZulu-Natal	5	65472	0.00898	0.1571	-4.51	<.0001*
	North West	6	65472	-0.2299	0.1796	-1.28	0.2004
	Gauteng	7	65472	-0.5471	0.1528	-3.58	0.0003*
	Mpumalanga	8	65472	-0.1357	0.1398	0.97	0.3317
	Limpopo (reference)	9	-	-	-	-	-
Province of residence	Western Cape	1	65472	-0.8247	0.2627	-0.74	0.459
	Eastern Cape	2	65472	0.2032	0.1741	-1.17	0.2416
	Northern Cape	3	65472	-0.1587	0.3263	-0.49	0.6244
	Free State	4	65472	-0.0812	0.2232	-0.36	0.7157
	KwaZulu-Natal	5	65472	0.1149	0.1554	-3.14	0.0017*
	North West	6	65472	-0.6338	0.177	-3.58	0.0003*
	Gauteng	7	65472	-0.3468	0.1583	-2.19	0.0284*
	Mpumalanga	8	65472	-0.0465	0.1442	0.32	0.7485
	Limpopo (reference)	9	-	-	-	-	-

Table 5.5 (Continue)

Analysis of Parameter Estimates							
	Parameter		DF	estimate	Standard Error	t Value	Pr>ltl
Place of death	Hospital	1	65472	-0.0499	0.0411	-2.38	0.0173*
	Emergency room	2	65472	-0.2223	0.1281	1.73	0.0832
	Dead on arrival	3	65472	0.0327	0.1316	-0.25	0.804
	Nursing home	4	65472	-0.4598	0.193	-1.21	0.2241
	Home (reference)	5	-	-	-	-	-
Education	None	0	65472	-0.7595	0.2256	3.37	0.0008*
	Grade 1	1	65472	-0.6822	0.3055	2.23	0.0625
	Grade 2	2	65472	0.742	0.2802	2.65	0.0082*
	Grade 3	3	65472	-0.5188	0.2693	1.93	0.055
	Grade 4	4	65472	-0.7679	0.2493	3.08	0.0021*
	Grade 5	5	65472	-0.4741	0.2572	1.84	0.0662
	Grade 6	6	65472	-0.4147	0.254	1.63	0.104
	Grade 7	7	65472	0.5657	0.2431	2.33	0.0202
	Grade 8	8	65472	-0.2571	0.2494	1.03	0.306
	Grade 9	9	65472	0.3665	0.2596	1.41	0.1601
	Grade 10	10	65472	-0.3982	0.2462	1.62	0.1071
	Grade 11	11	65472	-0.2198	0.2588	0.85	0.3996
	Grade 12	12	65472	0.2599	0.2394	1.09	0.2805
	University (reference)	13	-	-	-	-	-
Smoking	Yes	1	65472	0.1866	0.0812	-2.3	0.0912
	No (reference)	2	0	0	0	0	0
Pregnancy	Yes	1	65472	-1.2015	0.4561	-2.63	0.0855
	No (reference)	2	-	-	-	-	-

In this particular case the result indicates not much difference in the parameter estimates. The fixed effects parameter estimates and the Type 3 analysis for the fixed effects result show not much difference from the ones obtained using proc GLIMMIX in table 5.2. The standard errors are approximately the same as those obtained using the Proc GLIMMIX procedure in the random intercept model in Table 5.3.

Chapter 6

Discussion and Conclusion

This dissertation has presented the chosen topics on aspects of categorical data analysis in an organised manner and the areas of graphs and contingency tables have been adequately addressed. Statistical methods aimed in modeling categorical data have been concerned with statistical methods for categorical binary data which is frequently encountered in applied statistics. In this thesis, we attempted to give more insight into the different categorical approaches when one has a binary outcome. These methodologies have been demonstrated with analyses of a practical data set with a binary outcome.

Although many approaches to the analysis of categorical data have been studied, most are restricted to the setting in which the response variable is binary. Bar graphs were plotted to analyse the data set and drawing of cross-tabulation was also done to analyse the data set. We used generalised linear models (GLMs), logistic regression, surveylogistic, and generalised linear mixed models (GLMMs) to analyse the data set. We find that the results using PROC LOGISTIC were quite similar to those found using PROC SURVEYLOGISTIC when the cluster statement was not included in the PROC SURVEYLOGISTIC procedure. It must also be said that SAS allows good flexibility in using PROC GLIMMIX and PROC NLMIXED to fit these models. The PROC NLMIXED procedure took a much longer time to converge and gave us quite similar parameter estimates as the ones found using PROC GLIMIX. The reason for PROC NLMIXED taking longer was because the method is computationally more intensive.

The graphs showed us that diarrhoea was frequently affecting females, single people, and pregnant people since their proportion of diarrheal deaths was very high. In the logistic regression results, we found that the explanatory variables such as sex, smoking status of deceased and pregnancy status of deceased were not affected by diarrhoea and the other explanatory variables (marital status, province of birth, province of death, place of death, province of residence and education of deceased) were affected. PROC GENMOD results showed us the same interpretation as the one we had in logistic regression.

The results show that single and widowed people were more likely than customary married people to die from diarrhoea since their odds ratios were high. We also found that people who were born in KwaZulu-Natal, Gauteng and Mpumalanga province were more likely to die from diarrhoea than those who born in Limpopo province. KwaZulu-Natal province was found to have a high proportion of deaths compared to other provinces. This is because the odds ratio for KwaZulu-Natal (as a province of death of deceased) was large. People who were KwaZulu-Natal and Gauteng residents were found to be more likely to die from diarrhoea than Limpopo residents. Furthermore, the results reveal that hospital people are more likely to die from diarrhoea than home people. We found that uneducated people were at higher risk of dying from diarrhoea than university students.

In general, the explanatory variables (marital status, province of birth, province of death, place of death, province of residence and education of deceased) were found to have a significant effect on diarrhoea.

The recommendation I will make to public health planners, policy makers and practitioners is to go and educate all single people, uneducated people and also those people who are staying especially in KwaZulu-Natal about the danger of diarrhoea so that they will know about it and be able to prevent things that can cause diarrhoea. And also tell them ways to cure it once someone is affected. To educate people, several methods can be used like having workshops in townships and rural areas so that all public people would know about diarrhoea and also giving them pamphlets may help to those who are educated.

Future research could look at other statistical methods such as structural equation modeling which is a form of path analysis concerned with inter-dependence between variables. Another area of research is joint modeling of diarrhoea and other dependent variables such as TB.

References

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley: New York.

Albert, A. and Anderson, J.A. (1984). On the existence of maximum likelihood estimates in logistic regression models, *Biometrika,* **71**: 1-10.

Allison, P. (1999). Logistic Regression Using the SAS System. Cary, NC: SAS Institute.

Angelier, J.-P. (2007). Economie des Industries de Rèseau, PUG, Collection Eco+, Grenoble.

Altwegg, R., Schaub M., and A. Roulin (2007). Age-specific fitness components and their temporal variation in the barn owl. *American Naturalist,* **169:** 47-61.

Ainsworth, L.M. and Dean, C. B. (2006). Approximate Inference for Disease Mapping. *Computational Statistics & Data Analysis*, **50:** 2552-2570.

Baffoe-Bonnie B; Addo-Yobbo E; Plange-Rhule J. (1998) Five-year review of diarrheal disease cases admitted to a busy referral hospital in Ghana. Available at http://www.cmj.hr/1998/39/3/9740650.htm (Accessed 17/06/08).

Barndorff-Nielsen, O. (1978), Information and Exponential Families in Statistical Theory, (John Wiley & Sons, Chichester).

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association*. **39:** 357-365.

Berkson, J. (1953). A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function. *Journal of the American Statistical Association.* **48:** 565-599.

Boerma, J.T., Black R.E., Sommerfelt, A.E., Rutstein, S.O., and Bicego, G.T. (1991). Accuracy and Completeness of Mother`s Recall of Diarrhea Occurrence in Pre-School in Demographic and Health Surveys. *International Journal of Epidemiology*, **20:** 1073-80.

Bradley, E.L. (1973). The equivalence of maximum likelihood and weighted least square estimates in the exponential family. *Journal of the American Statistical Association.* **68:** 199-200.

Bryce J., Boschi-Pinto C., Shibuya K., Black R.E. (2005). The Child Health Epidemiology Reference Group. WHO Estimates of the Causes of Death in Children. *Lancet*. **365:** 1147-52.

Bjöck, A., (1996). *Numerical methods for least square problems*. SIAM, Philadelphia.

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

Breslow, N.E., and Lin, X. (1995). Bias correction in generalized liner models with a single component of dispersion. *Biometrika,* **82:** 81-91

Binder, D.A. (1983), On the Variance of Asymptotically Normal Estimators from Complex Surveys, *International Statistical review*, **51:** 279-292.

Charles, R.H. (1950). Estimation of variance and covariance components. *Biometrics* **9:** 226-52.

Choi, S.Y.P. (2003). Mechanisms of racial inequalities in prevalence of diarrhoea in South Africa. *J. Health Populat. Nutrit*, **21(3):** 264.

Cox, D.R. (1970). *Analysis of binary data*. London: Chapman and Hall.

Cox, D.R. and Wermuth, N. (1992). Response models for mixed binary and quantitative variables*. Biometrika*, **79:** 441-461.

Coxe, S., West, S. and Aiken, L. (2009). The analysis of Count Data: A gentle Introduction to Poisson Regression and Its Alternatives. *Journal of Personality Assessment* **91:** 121-136.

Connolly, M.A., and Liang, K.Y. (1988). Conditional logistic regression models for correlated binary data*. Biometrika*, **75:** 501-506.

Côtè, S.D., and Festa-Bianchet, M. (2001). Offspring sex ratio in relation to maternal age and social rank in mountain goats (Oreamnos americanus). *Behavioral Ecotogy and Sociobiology* **49:** 260-265.

Crawley, M.J. (2007). *The R Book*. J Wiley & Son Ltd, West Sussex.

Dean, A.J., Bell, J., Christie, M.J., Mattick, R.P. (2004). Depressive symptoms during buprenorphine vs. Methadone maintenance: findings from a randomized, controlled trial in opioid dependence. *European Psychiatry*, **19:** 510-3.

Diggle, P.J., Heagert, P., Liang K.Y. and Zeger, S.L. (2002*). Analysis of Longitudinal Data,* (2nd ed.). Oxford New York.

Fisher, R.A. (1934). Two new Properties of mathematical likelihood. Proceedings of the Royal Society A, **144:** 285-307.

Fisher, R.A. (1921). On the ‘Probable Error’ of a Coefficient of Correlation Deduced from a Small Sample. *Metron,* **1:** 3-32.

Firth, D., and Harris, I.R. (1991). Quasi-likelihood for multiplicative random effects. *Biometrika,* **78:** 545-555.

Fox, C.S., Evans, J.C., Larson, M.G., Lloyd-Jones, D.M., O`Donnell, C.J., Sorlie, P.D., Manolio, T.A., Kannel, W.B., Levy, D. (2005): A comparison of death certificate out-of-hospital coronary heart diease death with physician-adjudicated sudden cardiac death. *Am J cardiol,* **95**(7): 856-859. (PubMed ID Number: 15781015; Abstract)

Geyer, C.J., and Thompson, E.A. (1992). Constrained Monte Carlo Maximum Likelihood for Dependent Data (with discussion). *Journal of Royal Statistical Society, Series B*, **54:** 657-699.

Greenwald, A.G. (1975). Significance, nonsignificance and interpretation of an ESP experiment*. Journal of Experimental Social Psychology,* **11:** 180-191.

Gelfand, A.E. and Carlin, B.P. (1993). Maximum Likelihood Estimation for Constrained or Missing Data Models, *Canadian Journal of Statistics,* **21:** 303-312.

Goldstein, H. and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, A.* **159:** 505-13.

Gouws, E., Bryce, J., Pariyo, G., Schellenberg, J.A., Amaral, J. and Habicht, J.P. (2005). Measuring the quality of child health care at first-level facilities: *Social Science and Medicine*, **61:** 613-625.

Gill, J. (1999). The Insignificance of Null Hypothesis Significance Testing, *Political Research Quarterly,* **52:** 647-674.

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

Haslett, C., Chilvers, E., Hunter, J.A.A., Boon, N.A. (1999). Principles and Practice of Medicine (18th ed.), Edin burgh and Cambridge.

Hilbe, J. (1994). Generalized linear models. *American Statistician,* **48:** 255-265.

Hosmer, D.W. and Lemeshow, S. (1999). *Applied Survival Analysis Regression Modeling of Time to Event Data*. New York: John Wiley and Sons.

Hosmer, D.W., Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). John Wiley & Sons, New York, USA.

Hogue, V.W. (2000). Constipation and diarrhea. In: Casebook for Text book of Therapeutics: Drug and Disease Management, 571-588. Philadelphia: Lippincott Williams & Wilkins

Henderson, C.R. (1963). Selection index and expected genetic advance. NAS-NRC Publ. 982.

Imrey, P.B., Koch, G.G. and Stokes M.E. (1981). Categorical data analysis: Some reflections on the log linear model and logistic regression. Part I: Historical and methodological overview. *International Statistical Review.* **49:** 265-283.

Jennrich, R.I. and Moore R.H. (1975). Maximum likelihood estimation by means of nonlinear least squares. *Proc. Statistical Computing Section, American Statistical Association.* 57-65.

Jorgensen, B. (1983). Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. *Biometrika,***70:** 19-28.

Kosek, M., Bern, C. and Guerrant, R. (2003). The Global Burden of Diarrheal Disease, As Estimated from Studies Published Between 1992 and 2000. *Bulletin of the World Health Organization;* **81:** 197-204.

Kleinbaum, D.G. (1994). *Logistic Regression*: A self-learning Text. New York, Springer.

Kosek, M., Bern, C. and Guerrant, R. L. (2003). The global burden of diarrhoeal disease, as estimated from studies published between 1992 and 2000. *Bull World Health Organ,* **81:** 197-204.

Kuk, A.Y.C. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the American Statistical Association. B.* **57**: 395-407.

Larsen, R. (2008). *Logistic Regression for Metadata*: Cheshire takes on Adhoc-TEL CLEF: 38-41

Leamer, E.E. (1978). *Specification Searches*: Ad Hoc Inference with Non experimental Data. New York: John Wiley & Sons.

Leroux, R., Fung, J. and Barbeau, H. (1999). Adaptation of the walking pattern to uphill walking in normal and spinal-cord injured subjects. *Experimental Brain Research*, **126**: 359-368.

Lindsey, J.K. (1974). Construction and Comparison of Statistical models. *Journal of the Royal Statistical Society,* **B36:** 418-425.

Lindsay, B.G. (1995). Mixture Models: Theory, Geometry and Applications, NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5. Alexandria, Virginia: Institute of Mathematical Statistics and the American Statistical Association.

Liu, Q. and Pierce, D.A. (1993). Heterogeneity in Mantel-Haenszel-type models. *Biometrika,* **80:** 543-556.

Macnab, Y.C., Farrell, P.J., Gustafson, P. and Wen, S. (2004). Estimation in Bayesian disease mapping. *Biometrics,* **60:** 865-873.

May, S. and Hosmer, D.W. (2004). A cautionary note on the use of the Gronnesby and Borgan goodness-of-fit test for the Cox proportional hazard model*. Lifetime Data analysis,* **10:** 283-291.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

McCullagh, P. (1980). Regression models for ordinal data (with discussion*). J. Roy. Statistical Association.* **B 42:** 109-142.

McCullagh, P. (1983). Quasi-likelihood functions. *Annnals of Statistic*, **11**: 59-67.

Melta, C.R. and N.R. Patel (1995). Exact logistic regression: Theory and examples. *Statistical Medicine,* **14:** 2143-2160.

Miller, A.J. (1990). *Subset Selection in Regression*. London: Chapman and Hall.

Morel, J.G. (1989). Logistic Regression under Complex Survey Designs, *Survey Methodology*, **15:** 203-223.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

Nelder, J.A. and Weddernurn, R.W. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society.* **135:** 370-384.

Neter, J., Wasserman, W., and Kutner, M.H. (1989). *Applied linear regression models* (2nd ed.). Homewood, IL: Irwin.

Parzen, M. and Lipsitz, S.R (1999). A global goodness-of-fit statistic for Cox regression models. *Biometrics,* **55:** 580-584.

Raftery, A.E., and Sylvia, R. (1995). Model Selection for Generalized Linear Models via GLIB, with Application to Epidemiology. *In Bayesian Biostatistics* edited by D.A. Berry and D.K. Stangl. New York: Dekker, forth coming.

Rodriguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses, *Journal of the Royal Statistical Society, Series A.* **158:** 73-89.

Rozeboom, W.W (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin,* **57:** 416-428.

SADHS. (1998). Retrived from: http://www.mrc.ac.za/bod/dhsfin1.pdf.

Santner, T.J. and Duffy, E.D. (1986). A note on A. Albert and J. Anderson`s conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika,* **73:** 755-758.

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

Scott, S.C. (2010). Statistical assessment of ordinal outcomes in comparative studies. *Journal Clinical Edidemiol*. **50:** 45-55.

Snyder J.D. and Merson M.H. (1982). The Magnitude of the Global Problem of Acute Diarrheal Disease: A Review of Active Surveilance Data. *Bulletin of the World Health Organization,* **60:** 605-13.

So, Y. (1995). *A tutorial on logistic regression*. SUGI Proceedings.

Solomon, P.J. and Cox, D.R. (1992). Nonlinear component of variance models, *Biometrika,* **79:** 1-11.

Statistics South Africa. (2007). Mortality and causes of death Statistical release in South Africa, 2005. Findings from death notification. P0309.3. Pretoria: Statistics South Africa. http://www.Stats SA.gov.za/publications/P03093/P03093.pdf (Accessed on 22 Jun 2007).

Tukey, J. (1977). *Exploratory Data Analysis,* Addison-Weslet, Reading, MA. Introduces many new descriptive and analytical methods. Not extremely easy to read.

Ugarte, M.D., Militino, A.F. and Goicoa, T. (2008). Prediction error estimatiors in empirical Bayes disease mapping. *Environmetrics*, **19:** 287-300.

Vesikari, T. (1997). Rotavirus vaccines against diarrhoeal diseases. *Lancet*, **350** (9090): 1538-1541.

Victoria C.G., Huttly S.R., Fuchs S.C., Barros F.C., Garenne M., Leroy O., Fontaine O., Bean J.P., Gauveau V. and Chowdury H.R. (1993). International Differences in Clinical Patterns of Diarrheal Deaths: A Comparison of Children from Brazil, Senegal, Bangladesh, and India. *Journal of Diarrheal Diseases Research,* **11**(1): 25-29.

Weddernburn, R.W.M (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton Method. *Biometrika,* **61:** 439-447.

WHO. (1996). Childhood Diseases in Africa. Fact Sheet N 109, Retrieved from: https://apps.who.int/inf-fs/en/fact 109. Html.

WHO. (2008). *Essential prevention and care interventions for adults and adolescents living with HIV in resource-limited settings*. Geneva, World Health Organization. Available at: http://www.who.int/hiv/pub/plliv/interventions/en/index.html.

Wolfinger, R. and O`Connell, M. (1993). Generalized linear Mixed Models: a pseudo likelihood Approach, *Journal of Statistical computation and simulation*, **48:** 233-243.

Yule, G.U. (1927). On a method of investigating periodicities in disturbed series, with special reference to Wolfer`s sunspot numbers, *Philosophical Transactions of the Royal Society, Series A,* **226:** 267-98.

Zedrosser, A., Stoen, O.G., Sæbo, S. and Swenson, J.E (2007). Should I stay or should I go? Natal dispersal in the brown bear. *Animal Behaviour,* **74:** 369-376.