FACTORS AFFECTING THE HEALTH STATUS OF THE PEOPLE OF LESOTHO

By

ABIEL MOETI

Submitted in fulfilment of the academic

requirements for the degree of

MASTER OF SCIENCE

in

Applied Statistics

in the

School of Statistics and Actuarial Science

University of KwaZulu-Natal

Pietermaritzburg

2007

Dedication

To my wife, the late 'Marethatbile Moeti (may her soul rest in peace), and my mother, 'Maselloane Moeti.

Declaration

The research work described in this thesis was carried out in the School of Statistics and Actuarial Science, University of KwaZulu-Natal, Pietermaritzburg Campus, under the supervision of Prof. Temesgen Zewotir and the co-supervision of Dr Principal Ndlovu.

The work represents original work by the author and has not otherwise been submitted in any form for any degree or diploma to any University. Where use of the work of others has been made it is duly acknowledged.

November, 2007.

Mr Abiel Moeti Date

Prof. Temesgen Zewotir Date

Dr Principal Ndlovu Date

Acknowledgement

I thank my supervisor and co-supervisor, Prof. Temesgen Zewotir and Dr Principal Ndlovu respectively for their tireless guidance, encouragement and supervision.

I also extend my sincere gratitude to The Consulate General of Ireland for funding my studies and thank its staff for their hospitality and professionalism.

My special thanks go to the Bureau of Statistics (Lesotho) for allowing me to undertake this study and for the data used in this thesis.

I thank the staff members in the School of Statistics and Actuarial Science, University of KwaZulu-Natal, Pietermaritzburg, for the conducive environment for studying they have provided. I thank the postgraduate students for their moral support which kept me focused on my work.

I thank my family and friends for their unfailing moral support.

Again, everyone who contributed greatly to the success of this thesis either directly or indirectly though not mentioned by name deserves my thanks.

Abstract

Lesotho, like any other country of the world, is faced with the task of improving the life of its inhabitants. The Government of Lesotho has taken steps to address this issue by embarking on National Vision 2020 and Millennium Development Goals. To facilitate this the government developed a poverty reduction strategy which recognised the improvement of health as one of the priority areas. For this priority to be effectively acted on, factors affecting health status of the people have to be identified. Thus, the objective of this research is to identify factors that affect the health status of the people of Lesotho and the direction of effect of these factors. To achieve this, generalized linear models, generalized linear mixed models, and survey logistic regression models are used. The data for this research come from 2002 Lesotho Core Welfare Indicators Questionnaire Survey. The response variable, namely the health status, is measured by the presence or absence of disease/injury. The first model fitted is the generalized linear model which is selected using a stepwise procedure. The same model selected for the generalized linear model is refitted using a generalized linear mixed model and a survey logistic regression model which accounts for the complexity of the survey design. Using the generalized linear model and generalized linear mixed model the following factors were found to be significantly affecting the health status of the people of Lesotho: district of residence, sex, marital status, age, ownership of dwelling, education, and the interaction of effects; sex by marital status, age by marital status, ownership of dwelling by marital status, education by ownership of dwelling and ownership of dwelling by household size. The analysis using the survey logistic model also lead to the same conclusions as the above two models as well as identified the following interaction effects as important for health status: education by type of toilet, fuel used for cooking by time taken to reach hospital/clinic, sex by household size, marital status by household size, and type of toilet by household size.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Health status is defined not only as absence of disease, but also as the complete physical, mental, and social well-being of people (Huebner et al., 2004). Poor health leads to reduced household saving, productivity and learning ability of people (World Health Organization, www.who.int). Thus, reduction in saving, productivity, and learning creates poverty in the households or worsens it if it already exists, and in turn this perpetuates poor health even further. This becomes a problem for the well-being of people globally, Lesotho included. Accordingly, the Government of Lesotho has embarked on Millennium Development Goals (MDGs) and National Vision 2020 initiatives in response to this issue of the well-being of the Basotho people. These initiatives which are aimed at improving the lives of the Basotho, identify improved access to health care and social welfare as constituting part in the top priority areas that need immediate attention. This priority also forms part of Poverty Reduction Strategy developed to facilitate attainment of the MDGs and National Vision 2020 goals. The electronic version of these documents is available on the Ministry of Finance website, www.finance.gov.ls. It is vital to identify specific focal areas of intervention to speed up (or guide) the process.

There are a number of potential factors important for the health status of the people of Lesotho. Lesotho is divided into ten administrative areas called districts (which will also be referred to as locations). All the ten districts comprise rural and urban areas. Inequality of socio-economic development among the districts and rural/urban areas is inevitable. How this is related to the health status of people is not well known. Intuitively, one would think that households heads in certain age groups are relatively more responsible when it comes to the welfare of their households' members. In that sense, it is important to identify the category of the people who need assistance.

It is known that traditionally females have the obligation to look after their household members, whilst males on the other hand feel more responsible for resource provision. This phenomenon itself prompts health consciousness in females. However, this does not exclude males from their responsibility for the well-being of their family members. It is also of interest to find out if the relationship found between sex (especially males) and health status (Zullig, Valois, and Drane, 2005) holds in Lesotho. Regarding marital status, Prior and Hayes (2001) claim that marriage provides some kind of a shield against behaviour related health risks, such as drinking, unhealthy diet, and promiscuous sex (which may lead to sexually transmitted diseases). They add that it also offers a supportive relationship and enhanced economic benefits to the households. How this shield and economic benefits impact on health status in Lesotho is to be established in this study.

Though the large household in the African setting serves as a form of social security, according to Howden-Chapman (2004) it leads to an increased risk of infectious diseases due to overcrowding. Therefore, from the health viewpoint, it is necessary to see if this security is worth having. The households that do not own their dwelling, especially those that rent, have a limited control of the environment around the dwelling. For instance, a household may be occupying a dwelling with hazardous particulate matter to human health from roofing materials such as asbestos (Howden-Chapman, 2004). There may be other harmful particulate matter, such as nitrogen dioxide and carbon monoxide, produced as the result of types of fuel used for cooking. According to Howden-Chapman (2004) exposure to this particulate matter can cause diseases, such as asthma. The fact that an owner-occupied dwelling gives some kind of financial benefit to the households and that occupiers are likely to have a better life compared to those who use rented dwellings (Howden-Chapman, 2004), is of interest to see if this holds true in Lesotho. Therefore, in the present study the relationship between health status and roofing material, type of fuel used for cooking and the ownership of dwelling will be assessed.

A number of studies on the relationship between education and health status have arrived at different conclusions. For instance, Hussain and Smith (1999) found that in Bangladesh, children with mothers who have high school or higher education are less likely to have diarrhea. Cooper and Kohlmann (2001) found that education has little effect on elder Americans' health status. On the other hand, Vingilis, Wade, and Adlaf (1998) found that education has a significant effect on school going adolescents' health. Because of the uniqueness of each country, it is important to find how this relationship works in Lesotho. Other potential factors that need to be investigated are the type of toilet facility, source of drinking water, the time taken to reach the nearest supply of drinking water, as well as the time taken to reach the nearest hospital/clinic.

Consequently, the objective of this study is to use statistical methods to identify important factors affecting the health status of the people of Lesotho. The household level data will be used to model health status. Since all the variables should be at the household level, variables such as age, sex, marital status, and education pertain to the household head.

It is, at this stage, critical to state how the outcomes from the study will benefit the inhabitants of Lesotho. Identification of important factors for health status and their direction of effect will help all stakeholders (including policy framers and decision makers, donors, individuals, etc.) know which areas need more attention or which policies need to be fast-tracked in an endeavour to achieve better health for all. This study will not only serve as a guide, but will also highlight areas of further research for an in-depth understanding of the health status pattern in Lesotho. If all stakeholders, including Government, act accordingly, improved health status will be realized, hence a positive move will be made towards attainment of MDGs and National Vision 2020 goals.

The thesis is organized as follows. In Chapter 2 the theory of generalized linear models is reviewed, as these models will also be used to model health status to achieve the research objective. In Chapter 3 the data to be modelled are introduced. The sampling method utilized for data collection, and the classification of the variables are discussed. Data are analysed in Chapter 4 using generalized linear models where all the factor effects are fixed. In Chapter 5 the random primary sampling units (PSUs) effects are incorporated into the model selected in Chapter 4, leading to mixed models. These models are referred to as generalized linear mixed models, which are extensions to the generalized linear models. Survey logistic regression models designed specifically for survey data are discussed and fitted in Chapter 6. In Chapter 7 conclusions, implications, and avenues for future research are presented.

Chapter 2

Generalized Linear Models

Recall from Chapter 1 that the objective of the thesis is to identify factors that affect the health status of the people of Lesotho. Health status is a binary response variable (disease or no disease) with Bernoulli distribution (a member of the exponential family). All the potential factors that affect health status will be assumed to have fixed effects and hence generalized linear models will be fitted to the data. In the sections that follow the theory of generalized linear models is reviewed.

2.1 General Linear Models

A general linear model for an n×1 response $\mathbf{y}_{n\times 1} = (y_1, y_2, \ldots, y_n)'$ is given by

$$\mathbf{y} \quad = \mathbf{X}\boldsymbol{\beta} + \epsilon \tag{2.1}$$

where $\mathbf{X}_{n\times(p+1)}$ is an $n\times(p+1)$ design matrix whose i^{th} row (i=1,2,. . .,n) is (1,$x_{i1}, x_{i2}, \ldots, x_{ip}$), $\boldsymbol{\beta}_{(p+1)\times 1} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is a (p+1) vector of parameters, and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)' \sim N_n(0, \sigma^2 I))$ is an n×1 vector of random errors.

The least-squares estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, which is also the maximum likelihood estimator

(MLE) if the independent errors assumptions hold, is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}) \tag{2.2}$$

where $(\mathbf{X}'\mathbf{X})^{-1}$ is the inverse of $\mathbf{X}'\mathbf{X}$ and if $(\mathbf{X}'\mathbf{X})$ is not of full rank this inverse is replaced by a generalized inverse $(\mathbf{X}'\mathbf{X})^{-}$.

The sampling distribution of $\hat{\boldsymbol{\beta}}$, $N_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$, is used to test the hypothesis about $\boldsymbol{\beta}$. In cases where the normal errors assumption and the central limit theorem conditions are not satisfied, the general linear models are not applicable, so generalized linear models are used instead to model the data.

2.2 Generalized Linear Models

The generalized linear models are used to model observations on a random variable having a distribution belonging to the exponential family. In addition to the accommodation of non-normal responses, they also allow modelling of the functions of the mean besides the mean itself (Agresti et al., 2000). The theory of these models is discussed in Lindsey (1999); McCullagh and Nelder (1989); Dobson (1990); Meyer and Laud (2002); Cantoni and Ronchetti (2001); and Schabenberger and Pierce (2002) among others. The model for the responses $y_1, y_2, \ldots, y_n$ is determined by specifying

(1) the distribution (belonging to the exponential family of distributions) of the responses

(2) the linear predictor (which is a constant linear combination of parameters and covariates) and

(3) the link function (which links the mean response and the linear predictor).

Responses $y_1, y_2, \ldots, y_n$ are assumed to be independent with the same distribution belonging to the exponential family of distributions. The canonical form of probability density (mass) function of y_i (i=1,2,. . .,n) is

$$f(y_i; \theta_i, \phi) = exp\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right) \ , \quad i = 1, 2, \ldots, n \tag{2.3}$$

where $a_i(\phi) = \frac{\phi}{w_i}$ is called a dispersion parameter (where w_i is the weight for the i^{th} observation), θ_i is a natural parameter, and $b(\theta_i)$ is a cumulant or normalizing function. E(y_i) and var(y_i) are related to θ_i and $a_i(\phi)$ in the following manner

$$E(y_i) = \frac{\partial}{\partial\theta_i} b(\theta_i) = b'(\theta_i) = \mu_i \ , \quad i = 1, 2, \ldots, n \tag{2.4}$$

$$var(y_i) = b''(\theta_i)a_i(\phi) = V(\mu_i)a_i(\phi) \ , \quad i = 1, 2, \ldots, n. \tag{2.5}$$

where V(μ_i) $= b''(\theta_i)$ is a variance function obtained by differentiating μ_i with respect to θ_i. The linear predictor is given by

$$\eta_i = \mathbf{x_i'}\boldsymbol{\beta} = (1, x_{i1}, x_{i2}, \ldots, x_{ip})\boldsymbol{\beta} \ , \quad i = 1, 2, \ldots, n \tag{2.6}$$

where $\boldsymbol{x}_i'$ is the i^{th} row of the design matrix mentioned in Section 2.1.

The link function (denoted by g) is a monotonic and differentiable function which links the mean response $\mu_i = E(y_i)$ and the linear predictor $\eta_i = \mathbf{x_i'}\boldsymbol{\beta}$ as follows

$$\eta_i = g(\mu_i) = \mathbf{x_i'}\boldsymbol{\beta} \ , \quad i = 1, 2, \ldots, n. \tag{2.7}$$

If θ_i equals η_i, the link function is called a canonical link function. Each member of the exponential family of distributions has a unique canonical link function. For example, the canonical link function for the Binomial (or Binary) data is the logit. The generalized linear model with logit link is referred to as the logistic regression model discussed in Section 2.5.

2.3 Estimation of Parameters

The method of maximum likelihood is the theoretical basis for parameter estimation in generalized linear models, where the mean response μ is related to linear predictors by the link function g (i.e. $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$). To specify the likelihood function, the distributional form of the response needs to be assumed (Collett, 2003), and should have the form given in (2.3) with the joint probability density (mass) function

$$\begin{aligned} f(\mathbf{y};\boldsymbol{\theta}) = L(\boldsymbol{\theta};\mathbf{y}) &= \textstyle\prod_{i=1}^{n} exp\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right) \\ &= exp\textstyle\sum_{i=1}^{N}\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right) \end{aligned} \tag{2.8}$$

Algebraically, the probability density (mass) function f$(\mathbf{y};\boldsymbol{\theta})$ and likelihood function L$(\boldsymbol{\theta};\mathbf{y})$ are the same. The only difference is the emphasis on the arguments: for the density (mass) function emphasis is on the random vector $\mathbf{y}$ given the fixed parameter vector $\boldsymbol{\theta}$; and for likelihood function emphasis is on the parameter vector $\boldsymbol{\theta}$ given the vector of observed values $\mathbf{y}$. The log-likelihood function of (2.8) is given by

$$\boldsymbol{\ell} = \log L(\boldsymbol{\theta};\mathbf{y}) = \sum_{i}^{n} \ell_i \quad \text{with} \quad \ell_i = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi). \tag{2.9}$$

The parameter estimates are obtained by differentiating the log-likelihood function with respect to each β_j, equating derivatives to zero, and then solving the system of equations simultaneously for the β_j. That is

$$\frac{\partial\boldsymbol{\ell}}{\partial\beta_j} = \sum_{i=1}^{n}\frac{\partial\ell_i}{\partial\beta_j} = 0 \quad , \qquad j = 0,1,2,\ldots,p. \tag{2.10}$$

Using the chain rule of differentiation, $\frac{\partial\ell_i}{\partial\beta_j}$ is obtained as

$$\frac{\partial\ell_i}{\partial\beta_j} = \frac{\partial\ell_i}{\partial\theta_i}\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_j}. \tag{2.11}$$

From (2.9) it can be seen that

$$\frac{\partial\ell_i}{\partial\theta_i} = \frac{y_i - b'(\theta_i)}{a_i(\phi)} = \frac{y_i - \mu_i}{a_i(\phi)} \qquad \text{since } \mu_i = b'(\theta_i) \text{ from (2.4)}$$

and

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = V(\mu_i) \quad or \quad \frac{\partial \theta_i}{\partial \mu_i} = [V(\mu_i)]^{-1} \text{ from (2.5).}$$

Recall that the linear predictor (2.7) is $\eta_i = g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{ip}$. Then

$$\frac{\partial \eta_i}{\partial \mu_i} = g'(\mu_i) \quad \text{or} \quad \frac{\partial \mu_i}{\partial \eta_i} = [g'(\mu_i)]^{-1}$$

and

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

Substituting for $\frac{\partial \ell_i}{\partial \theta_i}$, $\frac{\partial \theta_i}{\partial \mu_i}$, $\frac{\partial \mu_i}{\partial \eta_i}$, and $\frac{\partial \eta_i}{\partial \beta_j}$ in (2.11) gives

$$\begin{aligned}\frac{\partial \ell_i}{\partial \beta_j} &= \frac{y_i - \mu_i}{a_i(\phi)}[V(\mu_i)]^{-1}[g'(\mu_i)]^{-1}x_{ij} \\ &= \frac{(y_i - \mu_i)x_{ij}}{a_i(\phi)V(\mu_i)g'(\mu_i)} \\ &= \frac{(y_i - \mu_i)x_{ij}}{var(y_i)g'(\mu_i)} \ , \quad \text{since } var(y_i) = a_i(\phi)V(\mu_i).\end{aligned}$$

Therefore, the system of equations to be solved for the β_j's is

$$\frac{\partial \boldsymbol{\ell}}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{var(y_i)g'(\mu_i)} = 0 \ , \quad j = 0, 1, \ldots, p. \tag{2.12}$$

These equations are solved iteratively. That is, an initial solution of the equations denoted by $\hat{\boldsymbol{\beta}}^{(0)}$ is guessed and then updated until the iterative algorithm converges to the solution $\hat{\boldsymbol{\beta}}$, called the maximum likelihood estimate of $\boldsymbol{\beta}$. Iterative algorithms for solving (2.12) are available in most statistical packages, such as SAS, STATA, GenStat etc. The two most popular and widely used algorithms for maximum likelihood estimation are the Newton-Raphson and the Fisher's scoring algorithms. The Fisher's scoring method is equivalent to the iterative reweighted least-squares (Schabenberger and Pierce, 2002; and Kutner et al., 2005). The Newton-Raphson method solves maximum likelihood estimates iteratively using the standard least-squares methods (McCullagh and Nelder, 1989). Both methods basically give the same solutions.

The Newton-Raphson procedure is, for r ≥ 1 until convergence

$$\hat{\boldsymbol{\beta}}^{(r)} = \hat{\boldsymbol{\beta}}^{(r-1)} - \mathbf{H}^{-1}\mathbf{u}^{(r-1)} \tag{2.13}$$

where $\mathbf{H}$ (the Hessian matrix) and $\mathbf{u}$ (the gradient) are given by

$$\mathbf{H} = [h_{jp}] = \left[\frac{\partial^2 \ell}{\partial\beta_j \partial\beta_p}\right]_{\beta=\hat{\beta}^{(r-1)}} \tag{2.14}$$

$$\mathbf{u} = [u_j] = \left[\frac{\partial \ell}{\partial\beta_j}\right]_{\beta=\hat{\beta}^{(r-1)}} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{var(y_i)g'(\mu_i)}. \tag{2.15}$$

The Hessian matrix, $\mathbf{H}$, is also referred to as the observed information matrix.

The Fisher's scoring method uses the expected information matrix referred to as the Fisher's information matrix ($\mathcal{J}$). The $(j,p)^{th}$ element of $\mathcal{J}$ is given by

$$-E\left(\frac{\partial^2 \ell}{\partial\beta_j \partial\beta_p}\right)$$

evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(r-1)}$. The Fisher's information matrix, $\mathcal{J}$ at $\boldsymbol{\beta}$, has the following relationship with $\mathbf{u}$ at $\boldsymbol{\beta}$

$$\mathcal{J}_{jp} = E(u_j u_p) = E\left(\frac{\partial \ell}{\partial\beta_j}\frac{\partial \ell}{\partial\beta_p}\right) = -E\left(\frac{\partial^2 \ell}{\partial\beta_j \partial\beta_p}\right) \tag{2.16}$$

(Dobson, 1990). Therefore, $\mathcal{J}_{jp}$ is given by

$$\begin{aligned}
\mathcal{J}_{jp} = E(u_j u_p) &= E\left(\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{var(y_i)g'(\mu_i)} . \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ip}}{var(y_i)g'(\mu_i)}\right) \\
&= E\left(\sum_{i=1}^{n} \frac{(y_i - \mu_i)^2 x_{ij}x_{ip}}{[var(y_i)g'(\mu_i)]^2}\right) \\
&= \sum_{i=1}^{n} \frac{E(y_i - \mu_i)^2 x_{ij}x_{ip}}{[var(y_i)g'(\mu_i)]^2}.
\end{aligned}$$

Since E$(y_i - \mu_i)^2$ is var(y_i), $\mathcal{J}_{jp}$ becomes

$$\mathcal{J}_{jp} = \sum_{i=1}^{n} \frac{x_{ij}x_{ip}}{var(y_i)[g'(\mu_i)]^2}. \tag{2.17}$$

Hence, the Fisher's information matrix at $\boldsymbol{\beta}$ is

$$\mathcal{J} = \mathbf{X}'\mathbf{W}\mathbf{X} \tag{2.18}$$

where $\mathbf{X}$ is the design matrix in Section 2.1 and $\mathbf{W}$ evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(r-1)}$ is a diagonal weight matrix with i^{th} diagonal element given by

$$w_{ii} = \frac{1}{var(y_i)[g'(\mu_i)]^2}. \tag{2.19}$$

Substituting $\mathcal{J}$ evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(r-1)}$ for $\mathbf{H}$ in (2.13) gives for r ≥ 1 until convergence,

$$\hat{\boldsymbol{\beta}}^{(r)} = \hat{\boldsymbol{\beta}}^{(r-1)} + \mathcal{J}^{-1}\mathbf{u}^{(r-1)}. \tag{2.20}$$

Parameter estimates $\hat{\boldsymbol{\beta}}^{(r)}$ in (2.20) are also given by, for r ≥ 1 until convergence,

$$\hat{\boldsymbol{\beta}}^{(r)} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z} \tag{2.21}$$

evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(r-1)}$, where the i^{th} element of $\mathbf{z}$ is given by

$$z_i = \sum_j^p x_{ij}\hat{\beta}_j^{(r-1)} + (y_i - \mu_i)g'(\mu_i)$$

evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(r-1)}$. The asymptotic sampling distribution of $\hat{\boldsymbol{\beta}}$ is given by

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \mathcal{J}^{-1}).$$

This distribution can be used to:

1) Test the significance of each parameter estimate $\hat{\beta}_j$ using the test statistic $\frac{\hat{\beta}_j}{\sqrt{v_{jj}}}$ which has standard normal distribution leading to a Wald(z) test statistic $\frac{\hat{\beta}_j^2}{v_{jj}}$ (Vittinghoff et al., 2005), which has chi-square distribution with one degree of freedom where the v_{jj} are diagonal elements of $\mathcal{J}^{-1}$ evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(r-1)}$.

2) Calculate $(1-\alpha)100\%$ confidence intervals for each parameter β_j

$$\hat{\beta}_j \pm z_{(1-\frac{\alpha}{2})}\sqrt{v_{jj}}$$

where z_p is the $100p^{th}$ percentile of the standard normal distribution. Another candidate statistic that can be used, instead of the Wald statistic, is the likelihood-ratio based statistic (also known as profile likelihood statistic) for constructing confidence intervals of parameter estimates. Although these two are almost the same for large samples, the likelihood-ratio based statistic is preferred for generalized linear models because of its reliability (Vittinghoff et al., 2005).

3) Also examine correlation among parameter estimates ($\text{corr}(\hat{\beta}_j, \hat{\beta}_p) = \frac{v_{jp}}{\sqrt{v_{jj}v_{pp}}}$),

where $v_{jp}(j \neq p)$ are off-diagonal elements of $\mathcal{J}^{-1}$ evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(r-1)}$.

2.4 Model Selection and Diagnostics

2.4.1 Model Selection

There can be a number of models in the family of generalized linear models that describe a given data set. Therefore, it is indispensable to select the simplest reasonable model that adequately describes given data (Lindsey, 1999). The selection of variables that enter the model is done through three candidate procedures namely, forward, backward, or stepwise selection. Forward selection starts with the null model (no explanatory variables) and enters one explanatory variable at a time. Backward selection starts with a saturated model (with all explanatory variables) and drops one explanatory variable at a time. The stepwise selection procedure operates the same way as the forward selection. But the advantage it has over the forward selection is that the variables already in the model are considered for exclusion each time another variable enters the model. So, when there are many variables under consideration the stepwise is mostly preferred because it has an advantage of minimising the chances of keeping redundant variables and leaving out important variables in the model. In all the procedures, a variable that leads to a significant change in the deviance (measure of goodness-of-fit described in Section 2.4.2) when added to or dropped from the model (i.e. which leads to p-value less than specified significance level) is retained, otherwise it is dropped. The contribution of each variable to the deviance reduction is given by the type 1 and type 3 analysis of effects. The type 1 analysis of effects depends on the sequence in which variables enter the model, whilst type 3 analysis of effects considers the overall model and assesses the contribution of each variable to deviance reduction irrespective of the sequence in which variables enter the model. This method of model selection is referred to as deviance analysis and is used to test the model for the goodness-of-fit.

2.4.2 Model Checking

Goodness-of-fit Test

The log-likelihood-ratio (deviance) and the Pearson's chi-square statistics are the main tools used for assessing the goodness-of-fit of the fitted generalized linear model (Jiang, 2001; and Kutner et al., 2005). They measure the discrepancy of fit between the maximum log-likelihood achievable and the achieved log-likelihood by the fitted model. One can illustrate the use of these measures with the most commonly used measure (i.e. deviance), given by

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2\{\ell(\mathbf{y}; \mathbf{y}) - \ell(\hat{\boldsymbol{\mu}}; \mathbf{y})\} \tag{2.22}$$

where $\ell(\hat{\boldsymbol{\mu}}; \mathbf{y})$ is the log-likelihood under the current model and $\ell(\mathbf{y}; \mathbf{y})$ is the log-likelihood under the maximum achievable (saturated) model. The aim is to minimize D (abbreviation of D$(\mathbf{y}, \hat{\boldsymbol{\mu}})$) by maximizing $\ell(\hat{\boldsymbol{\mu}}; \mathbf{y})$. This deviance is scaled by the dispersion parameter ϕ. Let the scaled deviance be denoted by D^* and its relationship with D is given by

$$D^* = \frac{1}{\phi} D. \tag{2.23}$$

Sometimes ϕ is not known, so in that case it can be estimated by

$$\hat{\phi} = \frac{D}{n-p} \tag{2.24}$$

where n is the number of observations, and p the number of parameters. See Schabenberger and Pierce (2002) and Lindsey (1999) for more details on the handling of the dispersion parameter. D (or D*) has an asymptotic chi-square distribution with $n-p$ degrees of freedom (i.e. D$\sim$ χ^2_{n-p}) (Jiang, 2001; McCullagh and Nelder, 1989; and Der and Everitt, 2002). For this statistic to be used to test the goodness-of-fit of the current model its asymptotic properties should hold (Schabenberger and Pierce, 2002). The hypothesis about the goodness-of-fit of the model to the data is given by

H_0: model is adequate

H_1: model is not adequate

If the level of significance is α, the H_0 will be rejected if $D > \chi^2_{n-p,\alpha}$. Alternatively, if d is the observed value of D and if $P(\chi^2_{n-p} > d) = p - value < \alpha$, then H_0 is rejected. A simple rule of thumb that can be used is that the mean deviance (given by D divided by $n-p$) should be approximately equal to one for a satisfactory current model, especially if the distribution of the responses is Binomial or Poisson (Collett, 2003).

Outliers, Influential, and High-leverage points

An outlier is a datum point that differs from the general trend of the data and is not necessarily always influential (Lindsey, 1999). By 'influential' one means that a slight change or omission of an observation leads to a substantial effect on parameter estimates of the model. The magnitude of influence is measured by the leverage (denoted by h_{ii}), which is the i^{th} diagonal element of the hat-matrix (**H**) (Kutner et al., 2005; and

Lindsey, 1999). For generalized linear models this matrix is given by

$$\mathbf{H} = \mathbf{W}^{-\frac{1}{2}}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})\mathbf{X}'\mathbf{W}^{-\frac{1}{2}} \tag{2.25}$$

where $\mathbf{X}$ is the design matrix in (2.1) and $\mathbf{W}$ is the weight matrix in (2.18). The hat-matrix is invariant under nonsingular linear transformation (Rousseeuw and Leroy, 2003). The leverage, h_{ii}, always lies between 0 and 1 (inclusive). The index plot of h_{ii} (h_{ii} vs. observations) is usually used to detect influential data points. Observations with h_{ii} greater than $2p/n$ (where p is the number of parameters and n is the number of cases) are regarded as influential (Preisser and Garcia, 2005; Rousseeuw and Leroy, 2003; and Collett, 2003). Note that h_{ii} has the problem of masking effect (Krzanowski, 1998). This means that, the fact that it is based only on explanatory variables, it is unlikely to detect the influence that may be due to response variable values. The other commonly used measure, which is reliable for detection of observations with undue influence, is the Cook's distance measure (Williams, 1987) discussed below.

Cook's Distance (C_i)

Cook's Distance is used to measure the influence of the i^{th} observation on the estimates of the parameters (Kutner et al., 2005). This statistic, following the notation used in the SAS GUIDE (version 9.1) is given by

$$C_i = \frac{1}{p\ a(\hat{\phi})}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{(i)}})'(\mathbf{X}'\mathbf{W}\mathbf{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{(i)}}) \tag{2.26}$$

where $\mathbf{X}$ and $\mathbf{W}$ are as defined in (2.25), p is the number explanatory variables in the model, a(ϕ) is a scale parameter, $\hat{\boldsymbol{\beta}}$ is the vector of parameter estimates from the full data set, and $\hat{\boldsymbol{\beta}}_{\boldsymbol{(i)}}$ is a vector of parameter estimates obtained when the model is fitted without the i^{th} observation. C_i is approximated by, in terms of h_{ii},

$$C_i \simeq \frac{r_{p_i}^2 h_{ii}}{(1-h_{ii})^2} = \frac{r_{p_{is}}^2 h_{ii}}{(1-h_{ii})} \tag{2.27}$$

where $r_{p_i} = (1 - h_{ii})y_i$ is called the Pearson residual, and $r_{p_{is}} = \frac{r_{p_i}}{\sqrt{1-h_{ii}}}$ is called the standardized Pearson residual. A large C_i implies that the i^{th} observation has undue influence on the set of parameter estimates. The question is how large is a large C_i value. Generally, there is no clearly defined cut-off rule. However, the most widely used cut-off value is 1. But some authors examine just the index plot of C_i (C_i vs. observation i)and consider the data points departing from the rest as being influential. See Hammill and Preisser (2005); Rousseeuw and Leroy (2003); and Skovgaard and Ritz (2007) for more details.

Link Function

The choice of link function is fundamental. If it is not appropriate the resultant estimates will be wrong, and will lead to misleading conclusions. The appropriateness of the link function can be tested by refitting the model with the linear predictor obtained from the original model and the square of the linear predictor as explanatory variables (Vittignhoff et al., 2005). If the link function is appropriate, then the linear predictor will be statistically significant, and the squared linear predictor term insignificant. This means that, prediction given by the linear predictor is not improved by adding the squared linear predictor term which is basically used to evaluate the null hypothesis that the model is adequate. Alternatively, the original model can be estimated with an extra constructed variable, where for an adequate model the extra variable will be statistically insignificant (Williams, 1987). In both cases, if the constructed variables (squared linear predictor and extra variable) are significant, then either the link function is not appropriate or important factor(s) have been omitted in the model.

The appropriateness of the link function can also be checked graphically by plotting the residuals against the fitted values which for an appropriate link should not exhibit any systematic pattern. This plot can also be used to check the form of the linear predictor (Collet, 2003).

2.5 Logistic Regression Model

The logistic regression model is a member of generalized linear models used to model binary data. To illustrate, consider the i^{th} individual (i=1,2,...,n) characterised by $\boldsymbol{x_i'}$ a vector of appropriately coded values of the factors (explanatory variables) having 1 in the first column. That is, $\boldsymbol{x_i'}$ is the i^{th} row of the design matrix $\mathbf{X}_{n\times(p+1)}$. Let the response y_i be 1 if the outcome for the i^{th} individual is a success, and 0 otherwise. Furthermore, let $\pi_i = \mathrm{P}(y_i = 1)$ be the probability that the outcome for the i^{th} individual is a success. The logistic regression model when the canonical link function is used is given by

$$\mathrm{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x_i'\beta}\ , \quad i = 1, 2, \ldots, n \tag{2.28}$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters. The logit transformation ensures that the probabilities (π_i's) lie within the interval (0,1) for any values of $\boldsymbol{x_i'\beta}$ from -∞ to ∞. Alternatively, the logistic regression model is given by

$$\pi_i = (1 + exp(-\boldsymbol{x_i'\beta}))^{-1}\ , \quad i = 1, 2, \ldots, n. \tag{2.29}$$

The other competing non-canonical link functions for the binary data, which also force probabilities to fall within the range (0,1) for values of $\boldsymbol{x_i'\beta}$ from -∞ to ∞, are the probit and the complementary log-log functions. The probit regression model is given by

$$probit(\pi_i) = \Phi^{-1}(\pi_i) = \boldsymbol{x_i'\beta}\ , \quad i = 1, 2, \ldots, n \tag{2.30}$$

or, equivalently

$$\pi_i = \Phi(\boldsymbol{x_i'\beta})\ , \quad i = 1, 2, \ldots, n \tag{2.31}$$

where Φ is the standard normal cumulative distribution function. The complementary log-log regression model is given by

$$log(-log(1-\pi_i)) = \boldsymbol{x_i'\beta} \ , \quad i = 1, 2, \ldots, n \tag{2.32}$$

or, equivalently

$$\pi_i = 1 - exp(-exp(\boldsymbol{x_i'\beta})) \ , \quad i = 1, 2, \ldots, n. \tag{2.33}$$

To illustrate that probabilities are always between 0 and 1 for values of $\boldsymbol{x_i'\beta}$ from -∞ to ∞, under the three link functions, one uses x = $\boldsymbol{x_i'\beta}$ values with x ranging from -6 to 6. Figure 2.1 displays the graphs of the probabilities π vs. x for the three link functions. The figure shows that $\pi \longrightarrow 0$ and 1 as x $\longrightarrow$ -∞ and ∞, respectively, for the three models or link functions. Furthermore, the following characteristics are observed for the three link functions:

1. They are monotonic increasing functions which map (-∞, ∞) interval of x-values onto (0,1) interval of probabilities.

2. The logit and the probit functions are symmetric about x = 0 (or π = 0.5).

3. For x $\ll$ 0, the logit and the complementary log-log probabilities are approximately equal, and for x $\gg$ 0 the complementary log-log and the probit probabilities are approximately equal.

2.5.1 Estimation of Parameters

The probability mass function (p.m.f.) of the Binomial distribution is

$$\frac{m_i!}{y_i!(m_i-y_i)!}\pi_i^{y_i}(1-\pi_i)^{m_i-y_i}, \quad y_i = 0, 1, \ldots, m_i \ , \quad \text{and} \quad i = 1, \ldots, n.$$

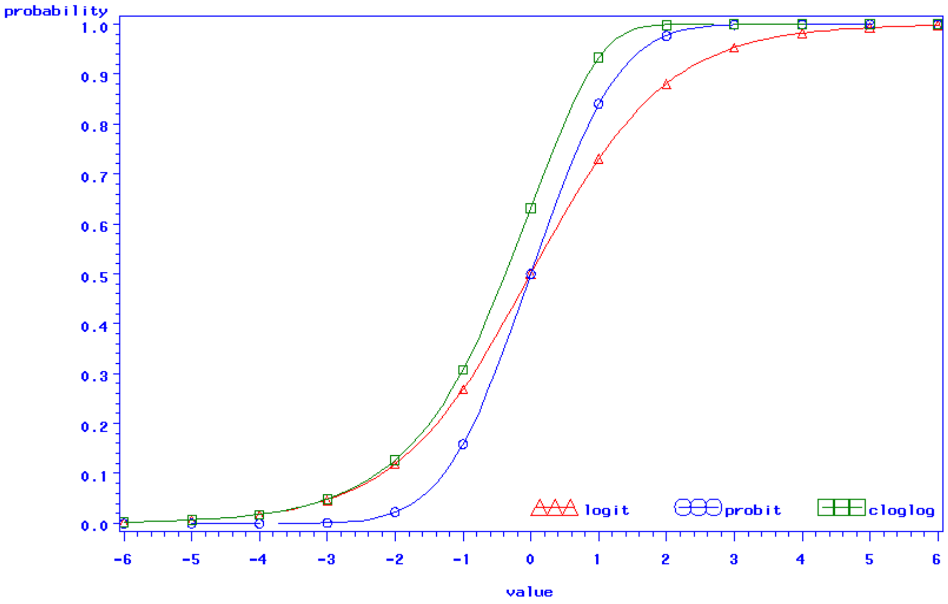


Figure 2.1: Comparison of link functions: probability (π) vs. x-value

The log-likelihood of π_i's or β is given by

$$\boldsymbol{\ell}(\boldsymbol{\mu};\boldsymbol{y}) = \ln\prod_{i=1}^{n}\frac{m_i!}{y_i!(m_i-y_i)!} + \sum_{i=1}^{n}\left[y_i\ln\left(\frac{\pi_i}{1-\pi_i}\right) + m_i\ln(1-\pi_i)\right] = \sum_{i=1}^{n}\ell_i \quad (2.34)$$

and the parameter estimates are obtained by solving the equations

$$\frac{\partial\boldsymbol{\ell}}{\partial\beta_j} = \sum_{i=1}^{n}\frac{\partial\ell_i}{\partial\beta_j} = 0\ , \qquad j = 1,2,\ldots,p \quad (2.35)$$

(see (2.11) to (2.12)). Note that for $m_i = 1$, $\text{var}(y_i) = \pi_i(1-\pi_i)$, $g'(\mu_i) = \frac{1}{\pi_i(1-\pi_i)}$, $a_i(\phi) = 1$, and $\mu_i = \pi_i$. Substituting for these in (2.12) obtains

$$\frac{\partial\boldsymbol{\ell}}{\partial\beta_j} = \sum_{i=1}^{n}\frac{(y_i-\pi_i)x_{ij}}{\pi_i(1-\pi_i)\frac{1}{\pi_i(1-\pi_i)}} = \sum_{i=1}^{n}(y_i-\pi_i)x_{ij} = 0. \quad (2.36)$$

Recall that these equations are solved iteratively using either the Newton-Raphson (2.13) and the Fisher's scoring (2.21) methods. When using (2.21) the weight matrix $\mathbf{W} = \text{diag}(\pi_i(1-\pi_i))$ and the i^{th} element of the constructed variable $\mathbf{z}$ is given by

$$z_i = \sum_{j}^{p} x_{ij}\beta_j^{(r-1)} + \frac{y_i-\pi_i}{\pi_i(1-\pi_i)}.$$

Inferences about $\boldsymbol{\beta}$ are made as described in Section 2.3.

2.5.2 Model Selection and Diagnostics

The same procedures discussed in Section 2.4 for model selection apply here. But, for ungrouped binary data, the deviance statistic D (or D^*) is used only to select variables and not as a measure of goodness-of-fit. For the goodness-of-fit measure, the Hosmer-Lemeshow goodness-of-fit test, proposed by Hosmer and Lemeshow (1989), discussed in the next section is used instead. Also discussed in the section is the inappropriateness of the deviance statistic as a measure of goodness-of-fit, and how the Hosmer-Lemeshow goodness-of-fit test is performed.

2.5.3 Model Checking

Goodness-of-fit Test

Recall that the deviance is given by (2.22) as

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2\{\ell(\mathbf{y};\mathbf{y}) - \ell(\hat{\boldsymbol{\mu}};\mathbf{y})\} \tag{2.37}$$

where $\ell(\hat{\boldsymbol{\mu}};\mathbf{y})$ is the log-likelihood under the current model and $\ell(\mathbf{y};\mathbf{y})$ is the log-likelihood under the maximum achievable (saturated) model. Suppose $Y_i \sim BIN(m_i, \pi_i)$, then $E(Y_i) = m_i\pi_i = \mu_i$. The likelihood function is

$$\prod_{i=1}^{n} f(y_i;\pi_i) = \prod_{i=1}^{n} \frac{m_i!}{y_i!(m_i - y_i)!}\pi_i^{y_i}(1-\pi_i)^{m_i - y_i}.$$

The log-likelihood is

$$\begin{aligned}
\ell(\boldsymbol{\mu};\mathbf{y}) &= \ln\prod_{i=1}^{n}\frac{m_i!}{y_i!(m_i-y_i)!} + \sum_{i=1}^{n} y_i\ln(\pi_i) + \sum_{i=1}^{n}(m_i - y_i)\ln(1-\pi_i) \\
&= \ln\prod_{i=1}^{n}\frac{m_i!}{y_i!(m_i-y_i)!} + \sum_{i=1}^{n} y_i\ln\left(\frac{m_i\pi_i}{m_i}\right) + \sum_{i=1}^{n}(m_i - y_i)\ln\left[\frac{m_i - m_i\pi_i}{m_i}\right] \\
&= \ln\prod_{i=1}^{n}\frac{m_i!}{y_i!(m_i-y_i)!} + \sum_{i=1}^{n} y_i\ln\left(\frac{\mu_i}{m_i}\right) + \sum_{i=1}^{n}(m_i - y_i)\ln\left[\frac{m_i - \mu_i}{m_i}\right].
\end{aligned}$$

Therefore, the log-likelihood for the fitted model is

$$\ell(\hat{\boldsymbol{\mu}};\mathbf{y}) = \ln\prod_{i=1}^{n}\frac{m_i!}{y_i!(m_i-y_i)!} + \sum_{i=1}^{n} y_i\ln\left(\frac{\hat{\mu}_i}{m_i}\right) + \sum_{i=1}^{n}(m_i-y_i)\ln\left[\frac{m_i-\hat{\mu}_i}{m_i}\right]$$

and that for the maximal (saturated) model ($\hat{\mu}_i = y_i$) is

$$\ell(\mathbf{y};\mathbf{y}) = \ln\prod_{i=1}^{n}\frac{m_i!}{y_i!(m_i-y_i)!} + \sum_{i=1}^{n} y_i\ln\left(\frac{y_i}{m_i}\right) + \sum_{i=1}^{n}(m_i-y_i)\ln\left[\frac{m_i-y_i}{m_i}\right].$$

Substituting $\ell(\hat{\mu};\mathbf{y})$ and $\ell(\mathbf{y};\mathbf{y})$ in (2.37) gives

$$\begin{aligned} D = -2\Bigg[& \ln\textstyle\prod_{i=1}^{n}\frac{m_i!}{y_i!(m_i-y_i)!} + \sum_{i=1}^{n} y_i\ln\left(\frac{\hat{\mu}_i}{m_i}\right) + \sum_{i=1}^{n}(m_i-y_i)\ln\left(\frac{m_i-\hat{\mu}_i}{m_i}\right) \\ - & \left\{\ln\textstyle\prod_{i=1}^{n}\frac{m_i!}{y_i!(m_i-y_i)!} + \sum_{i=1}^{n} y_i\ln\left(\frac{y_i}{m_i}\right) + \sum_{i=1}^{n}(m_i-y_i)\ln\left(\frac{m_i-y_i}{m_i}\right)\right\} \Bigg]. \end{aligned}$$

After rearrangement of terms, D becomes

$$\begin{aligned} D &= -2\left[\sum_{i=1}^{n} y_i\ln\left[\frac{\hat{\mu}_i}{m_i}\times\frac{m_i}{y_i}\right] + \sum_{i=1}^{n}(m_i-y_i)\ln\left[\frac{m_i-\hat{\mu}_i}{m_i}\times\frac{m_i}{m_i-y_i}\right]\right] \\ &= -2\sum_{i=1}^{n}\left[y_i\ln\left(\frac{\hat{\mu}_i}{y_i}\right) + (m_i-y_i)\ln\left(\frac{m_i-\hat{\mu}_i}{m_i-y_i}\right)\right] \\ &= 2\sum_{i=1}^{n}\left[y_i\ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i-y_i)\ln\left(\frac{m_i-y_i}{m_i-\hat{\mu}_i}\right)\right]. \end{aligned}$$

But, if $m_i = 1$, for all i, D becomes

$$\begin{aligned} D &= -2\sum_{i=1}^{n}\left[y_i\ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right] \\ &= -2\sum_{i=1}^{n}[y_i\ln(\hat{\pi}_i) + (1-y_i)\ln(1-\hat{\pi}_i)], \end{aligned}$$

since $y_i \ln y_i = 0$ and $(1-y_i)\ln(1-y_i) = 0$ if $y_i = 0$ or 1. After rearrangement of terms, D becomes

$$D = -2\sum_{i=1}^{n}\left[y_i\ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) + \ln(1-\hat{\pi}_i)\right]. \qquad (2.38)$$

Furthermore, since $\sum_{i=1}^{n} y_i\ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \sum_{i=1}^{n}\hat{\pi}_i\ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right)$, (2.38) becomes

$$D = -2\sum_{i=1}^{n}\left[\hat{\pi}_i\ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) + \ln(1-\hat{\pi}_i)\right] \qquad (2.39)$$

which is not informative about the relationship between $\hat{\pi}_i$ and the observed y_i values, since it is only a function of the estimated probabilities. A detailed discussion about this issue is given by Collett (2003). This discussion justifies that the deviance D cannot be used as a measure of goodness-of-fit of the model for ungrouped binary data (Krzanowski, 1998). However, it can still be used to identify important predictors as discussed above. The appropriate test of goodness-of-fit in this situation is the Hosmer-Lemeshow goodness-of-fit test (see Hosmer and Lemeshow, 1989).

Hosmer-Lemeshow Goodness-of-Fit Test

For this test, firstly the predicted probabilities ($\hat{\pi}_i$'s, $i = 1, 2, \ldots, n$) obtained using the current model being checked are used to form g groups with approximately n/g subjects (households). One grouping strategy (percentile strategy discussed by Hosmer and Lemeshow (1989)) is as follows:

(1) Group 1 subjects are approximately n/g subjects whose $\hat{\pi}_i$'s are less than or equal to the $\frac{100}{g}th$ percentile of all the $\hat{\pi}_i$'s.

(2) Group 10 subjects are approximately n/g subjects whose $\hat{\pi}_i$'s are more than $\left(1 - \frac{1}{g}\right)$ ×100th percentile of all the $\hat{\pi}_i$'s.

(3) For $j = 2, 3, \ldots, g-1$, group j subjects are approximately n/g subjects whose $\hat{\pi}_i$'s are greater than the $\frac{j-1}{g}$×100th percentile and less than or equal to the $\frac{j}{g}$ × 100th percentile of all the $\hat{\pi}_i$'s.

For large n (number of subjects/households) the frequently recommended g is 10 (see Vittinghoff et al., 2005; Dobson, 2002; and Hosmer and Lemeshow, 1989) in order for different analysts to get consistent conclusions.

Secondly, for each group the observed and expected frequencies of the responses y = 0 and y = 1 are determined as follows: For the j = 1, 2, …, g,

(1) The respective observed frequencies of the responses y = 1 and y = 0 are O_{1j} = number of subjects with responses y = 1 and $O_{0j} = n/g$ - O_{1j}.

(2) The respective expected frequencies of the responses y = 1 and y = 0 are E_{1j} = n/g × average of $\hat{\pi}_i$'s in group j and $E_{0j} = n/g$ - E_{1j}.

Finally, the Hosmer-Lemeshow statistic X^2_{HL} for testing the goodness-of-fit of the model is given by

$$X^2_{HL} = \sum_{j=1}^{g}\sum_{i=0}^{1}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}. \tag{2.40}$$

The statistic χ^2_{HL} has a chi-square distribution with $g-2$ degrees of freedom. Therefore, the statistic χ^2_{HL} is compared with the critical value of the chi-square distribution with $g-2$ degrees of freedom $(\chi^2_{(g-2,\alpha)})$ for checking goodness-of-fit of the model. Thus, if the χ^2_{HL} statistic is statistically significant then it indicates lack-of-fit of the model, whereas a non-significant one indicates goodness-of-fit of the model.

Outliers, Influential, and High-leverage points

Refer to Section 2.4 for the discussion on outliers, influential, and leverage points. Recall that hat-matrix (**H**) is given by (2.25) as

$$\mathbf{H} = \mathbf{W}^{-\frac{1}{2}}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})\mathbf{X}'\mathbf{W}^{-\frac{1}{2}} \tag{2.41}$$

and that its diagonal elements, h_{ii}, are used to measure the magnitude of the distance the i^{th} observation is from the rest. The data points with h_{ii} greater than $2p/n$ are considered as high leverage points and having potential influence on the model parameter estimates. Although index plot of h_{ii} can sometimes be useful in detecting potential influential data points (Collett, 2003), it cannot detect influence due to response values. That is because h_{ii} is based only on explanatory variables. Therefore, Cook's distance statistic discussed in Section 2.4 is better for effective detection of influential data points.

Link Function

The procedures discussed in Section 2.4.2 are employed here except the graphical one. For ungrouped binary data, the graph of the residuals vs. the fitted values always exhibits a systematic pattern even when the link is appropriate, hence is uninformative (Der and Everitt, 2002). Therefore, this approach cannot be used in the case of ungrouped binary data.

Validation of Predicted Probabilities

It is imperative to see to what degree the predicted probabilities agree with the outcomes. That is, one wants to have a reliable model that maximizes the chance and sensitivity of identifying the individuals who need justified intervention. In other words, one would like to reduce the proportion of the individuals that are incorrectly classified as having outcome of failure y = 0 (1-specificity) and hence deny those individuals the benefit of intervention. A cut-off value that minimizes the misclassification probabilities of the individuals should be specified. Table 2.1 is an example of how classification is done. In the table, y_i is the response of the i^{th} individual, and $\hat{y}_i$ is the predicted response of the i^{th} individual.

Sensitivity (probability of correctly classifying an individual with the outcome of success) is estimated as

$$Ss = \frac{a}{a+c}.$$

Table 2.1: Classification table

		Correct classification		
		y=1	y=0	Total
Predicted	$\hat{y}$=1	a	b	a+b
classification	$\hat{y}$=0	c	d	c+d
	Total	a+c	b+d	n

Specificity (probability of correctly classifying an individual with the outcome of failure) is estimated as

$$Sp = \frac{d}{b+d}.$$

False positive rate (Fpr) (probability of incorrectly classifying an individual with the outcome of failure) is estimated as

$$Fpr = \frac{b}{b+d}.$$

False negative rate (Fnr) (probability of incorrectly classifying an individual with the outcome of success) is estimated as

$$Fnr = \frac{c}{a+c}.$$

The Receiver Operating Characteristic (ROC) curve can be used to graphically display the predictive accuracy of the model (Vittinghoff et al., 2005). This graph is given in Figure 2.2. This is a plot of sensitivity against 1-specificity (in other words it is a plot of true positive rate against false positive rate) as shown in Figure 2.2. A curve along the 45^0 line (where area under the curve is 0.5) shows that classification is at random (Taylor and Krawchuk, 2005). The larger the deviation of this curve is from the 45^0 line to the left, the better is the prediction

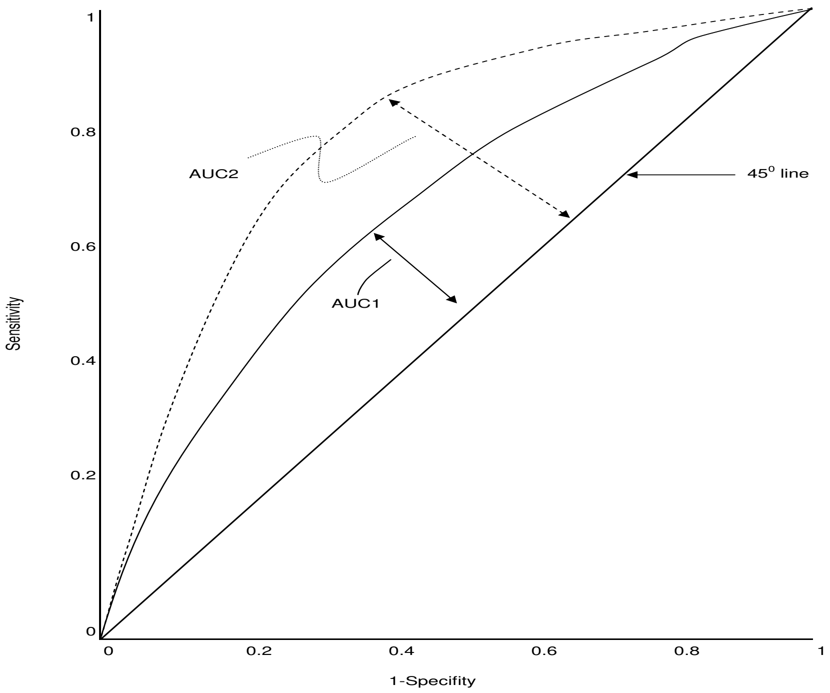


Figure 2.2: Sensitivity against 1-specificity: ROC curve

accuracy of the model. This means the prediction accuracy of the model can be measured by the total area under the ROC curve (AUC). The larger the AUC the better the accuracy of the model. For example, AUC1 < AUC2 in Figure 2.2 implies that the model with AUC2 has the better prediction accuracy than the model with AUC1. The AUC is also referred to as the probability that the predicted probability assigned to the event (y = 1) is higher that the non-event (y = 0) (Mason and Granam, 2002). From Taylor and Krawchuk (2005), the prediction accuracy of 0.6-0.7 suggests moderate prediction (or discrimination); of 0.7-0.8 suggests acceptable prediction; and of 0.8-0.9 suggests excellent prediction. If this measure is less than 0.6, then the prediction accuracy of the model is poor.

2.6 Interpretation of Logistic Regression Model Coefficients

2.6.1 Logit Model

For illustrative purposes, let one assume that one has a prediction model with three explanatory variables, x_1, x_2, and x_3 given by

$$\hat{\eta}_i = log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i1}x_{i2} + \hat{\beta}_4 x_{i3}\ , \quad i = 1, 2, \ldots, n. \quad (2.42)$$

The estimated coefficients are the log odds, where $\hat{\beta}_0$ is the estimate of the overall mean of the logits of the probabilities, and the other estimated coefficients are the estimated slopes associated with the respective variables of the model, controlling for others. Since there is an x_1x_2 interaction, the main effects of x_1 and x_2 are not interpreted. If x_1 and x_2 are categorical variables with two levels, $\hat{\beta}_3$ measures the change of the log odds when x_1 (or x_2) changes from reference level to the other, given that x_2 (or x_1) assumes a non-reference level controlling for x_3. The coefficient of x_3, measures the amount of change of the log odds for a unit change in x_3, controlling for x_1 and x_2. The most preferred way of interpreting logit model coefficients is by odds ratios. These are obtained by converting the log odds model to the odds model (i.e. taking the anti-log of $log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right)$ to get $\frac{\hat{\pi}_i}{1-\hat{\pi}_i}$, given by

$$\hat{O}_i = exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i1}x_{i2} + \hat{\beta}_4 x_{i3}).$$

Let one assume that all the variables are binary with values 0/1, where 0 is the reference level. Then, the odds when $x_1 = 1$ controlling for the other variables is given by

$$\hat{O}_i(1) = exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i2} + \hat{\beta}_4 x_{i3})$$

and the odds when $x_1 = 0$ is

$$\hat{O}_i(0) = exp(\hat{\beta}_0 + \hat{\beta}_2 x_{i2} + \hat{\beta}_4 x_{i3}).$$

Then, the odds ratio for the two ($x_1 = 1$ vs. $x_1 = 0$) is given by

$$\begin{aligned} \widehat{OR}(1) = \frac{\hat{O}_i(1)}{\hat{O}_i(0)} &= exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i2} + \hat{\beta}_4 x_{i3} - (\hat{\beta}_0 + \hat{\beta}_2 x_{i2} + \hat{\beta}_4 x_{i3})) \\ &= exp(\hat{\beta}_1 + \hat{\beta}_3 x_{i2}). \end{aligned}$$

Notice that this odds ratio is a function of x_2 which is interacted with x_1. If $x_2 = 1$ the ratio becomes $\exp(\hat{\beta}_1 + \hat{\beta}_3)$, and if $x_2 = 0$ the ratio is $\exp(\hat{\beta}_1)$. Thus, the ratio of the two gives $\exp(\hat{\beta}_3)$. This value is interpreted as the number of times the event $y_i = 1$ is more likely to occur when $x_2 = 1$ compared to when $x_2 = 0$, given that $x_1 = 1$ controlling for x_3, and *vice versa*. The odds ratio $\exp(\hat{\beta}_4)$ is interpreted as the number of times the event $y_i = 1$ is more likely to occur when $x_3 = 1$ compared to when $x_3 = 0$, controlling for x_1 and x_2.

Alternatively, the prediction model can be interpreted in terms of probabilities. Recall (2.29) or (2.42) in which the estimated probabilities are given by

$$\hat{\pi}_i = (1 + exp(-\hat{\eta}_i))^{-1}.$$

Therefore, the probability that $y_i = 1$, given that $x_1 = 1$, $x_2 = 1$, and $x_3 = 0$, is given by

$$P(y_i = 1 | x_{i1} = 1, x_{i2} = 1, x_{i3} = 0) = (1 + exp(-\hat{\beta}_0 - \hat{\beta}_1 - \hat{\beta}_2 - \hat{\beta}_3))^{-1}$$

and is interpreted as the probability that the event $y_i = 1$ will occur when $x_1 = 1$, $x_2 = 1$, and x_3 equals 0. For $x_1 = 0$, $x_2 = 0$, and $x_3 = 1$ the conditional probability for $y_i = 1$ is

$$P(y_i = 1 | x_{i1} = 0, x_{i2} = 0, x_{i3} = 1) = (1 + exp(-\hat{\beta}_0 - \hat{\beta}_3))^{-1}$$

interpreted as the probability that the event $y_i = 1$ will occur when $x_3 = 1$ and $x_1 = x_2 = 0$.

2.6.2 Probit Model

For probit link (2.42) becomes

$$\hat{\eta}_i = \Phi^{-1}(\hat{\pi}_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i1} x_{i2} + \hat{\beta}_4 x_{i3}\ , \quad i = 1, 2, \ldots, n. \qquad (2.43)$$

The coefficients are already in the Z score metric, hence can be interpreted directly. The change of x_3 from 0 to 1 increases the Z score (probit score) by $\hat{\beta}_4$. When x_1 changes from 0 to 1 the Z score changes by $\hat{\beta}_1 + \hat{\beta}_3 x_{i2}$, given by

$$\triangle\hat{\eta} = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i2} + \hat{\beta}_4 x_{i3} - (\hat{\beta}_0 + \hat{\beta}_2 x_{i2} + \hat{\beta}_4 x_{i3}) = \hat{\beta}_1 + \hat{\beta}_3 x_{i2}. \quad (2.44)$$

That means, when $x_2 = 0$ (2.44) equals $\hat{\beta}_1$ and when $x_2 = 1$ (2.44) equals $\hat{\beta}_1 + \hat{\beta}_1$. Hence the Z score changes by $\hat{\beta}_3$ when x_2 changes from 0 to 1 given that x_1 changed from 0 to 1.

The probit model can also be interpreted in terms of probabilities, which are widely understood compared to the Z scores, given by $\hat{\pi}_i = \Phi(\hat{\eta}_i)$. The probability that an event $y_i = 1$ will occur when $x_1 = 1$, $x_2 = 1$, and $x_3 = 0$ is

$$P(y_i = 1 | x_{i1} = 1, x_{i2} = 1, x_{i3} = 0) = \Phi(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3)$$

and the probability that $y_i = 1$ will occur when $x_1 = 0$, $x_2 = 0$, and $x_3 = 1$ is

$$P(y_i = 1 | x_{i1} = 0, x_{i2} = 0, x_{i3} = 1) = \Phi(\hat{\beta}_0 + \hat{\beta}_4).$$

2.6.3 Complementary Log-Log Model

Consider (2.42) with complementary log-log link,

$$\hat{\eta}_i = log(-log(1 - \hat{\pi}_i)) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i1} x_{i2} + \hat{\beta}_4 x_{i3}\ , \ i = 1, \ldots, n. \ (2.45)$$

The estimate $\hat{\beta}_4$ is interpreted as the amount of increase of the complementary log-log probability of $y_i = 1$, when x_3 changes from level 0 to level 1. When x_1 changes from

0 to 1 the complementary log-log probability of $y_i = 1$ changes by $\hat{\beta}_1 + \hat{\beta}_3 x_{i2}$, which is a function of x_2. When $x_2 = 0$, the change equals $\hat{\beta}_1$ and the change equals $\hat{\beta}_1 + \hat{\beta}_3$ when $x_2 = 1$. This means the complementary log-log probability of $y_i = 1$ will change by $\hat{\beta}_3$ when $x_1 = 1$, given that $x_2 = 1$ or when $x_2 = 1$, given that $x_1 = 1$.

Similarly, a complementary log-log model can be interpreted in terms of probabilities of $y_i = 1$, which are given by

$$\hat{\pi}_i = 1 - exp(-exp(\hat{\eta}_i)).$$

Considering the same examples, the probability that an event $y_i = 1$ will occur when $x_1 = 1$, $x_2 = 1$, and $x_3 = 0$ is

$$P(y_i = 1 | x_{i1} = 1, x_{i2} = 1, x_{i3} = 0) = 1 - exp(-exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3))$$

and when $x_1 = 0$, $x_2 = 0$, and $x_3 = 1$ the probability is

$$P(y_i = 1 | x_{i1} = 0, x_{i2} = 0, x_{i3} = 1) = 1 - exp(-exp(\hat{\beta}_0 + \hat{\beta}_4)).$$

Chapter 3

The Data

The data used in this thesis are from the Lesotho Core Welfare Indicators Questionnaire (CWIQ) Survey which was undertaken by the Bureau of Statistics, Lesotho. Because of the challenges that data collection is faced with, such as limited financial resources, this led to the development of a more complex survey design than the simple random sampling (SRS). The advantage of complex surveys has gone beyond cost saving, i.e. they obtain more reliable estimates compared to SRS which fail to take into account features of population leading to results that are statistically unrepresentative of the population. Consequently, employment of complex survey design such as stratified random sampling and cluster sampling increases the accuracy of the population level estimates, while keeping costs manageable (Villiant, Dorfman, and Royall, 2000; and Levy and Lemeshow, 1991).

Therefore, the two-stage sample design was used where Enumeration Areas (EAs) were the first-stage sampling units and the households were the second-stage sampling units. At the first stage, 25 EAs were randomly selected from each of ten districts in Lesotho, except for Maseru where an additional 10 EAs in the urban area were selected. A total of 260 EAs were selected. From each EA a random sample of 20 households was drawn, giving a total of 5200 households in the sample. The selected sample was distributed in the ratio 2:1 of rural and urban areas in each district. The overall response rate was 95.3%. See the final report of CWIQ survey prepared by the Bureau of Statistics (Demographic, Labour and Social Statistics Division) for more details.

The survey captured a vast range of variables. However, in this study the researcher considered only 14 variables, namely those which, from the literature and the writer's intuitions, have potential significant effects on health status as measured by the presence or absence of disease/injury. The 14 independent variables are (1) Urban/Rural, (2) District (or Location), (3) Age of household head, (4) Sex of household head, (5) Marital status of household head, (6) Education of household head, (7) Household size, (8) Ownership of dwelling, (9) Type of roofing material, (10) Type of source of drinking water, (11) Type of toilet, (12) Fuel used for cooking, (13) Time taken to reach the nearest supply of drinking water, and (14) Time taken to reach the nearest hospital/clinic.

Variables are coded in such a way that the reference level has the number 0. Urban/Rural is coded 1 = Rural and 2 = Urban, whilst District is coded 01 = Butha-Buthe, 02 = Leribe, 03 = Berea, 04 = Thaba-Tseka, 05 = Mafeteng, 06 = Mohale's Hoek, 07 = Quthing, 08 = Qacha's Nek, 09 = Mokhotlong, and 10 = Maseru. Sex is categorised into 1 = males and 2 = females. Marital status is categorised into 1 = not married and 2 = married which is similar to 'never married' versus 'ever married' used by Boniface et al. (2001), and 'non-married' versus 'married' used by Prior and Hayes (2001). Hussain and Smith (1999) categorised education into 'no education, primary, and post-primary education', whilst Boniface et al. (2001) categorised it into 'no qualification, GCSE, and A level+'. A similar three level classification is followed in this study where 1 = No education (including some primary), 2 = completed primary, and

3 = completed secondary.

Furthermore, two groups (1 = do not own dwelling, and 2 = own dwelling) follow the classification by Walker and Becker (2005) who used 'government or the private sector dwelling' and 'owning or purchasing dwellings'. Age is categorised into 1 = younger than 40 years, and 2 = 40 years or older. This follows the classification by Ghosh et al. (1998) and the present writer's own assumption that people reach the level of health consciousness at the age of 40.

The survey categorised 'time to reach the nearest source of drinking water and hospital' into 0-14, 15-29, 30-44, 45-59, 60-119, and 120+ minutes. The frequency distribution of the data showed that about 79.2% of households took 0-14 minutes to reach the nearest source of drinking water which is the first category. So the variable was reclassified into two groups 1 = 15 minutes or more, and 2 = less than 15 minutes. The cumulative frequency distribution of the time to reach the nearest clinic or hospital was also constructed, and showed that about 48.6 percent of households took less than 60 minutes to reach the nearest health clinic or hospital. Therefore, this variable was reclassified into 1 = 60 minutes or more, and 2 = less than 60 minutes.

The 'source of drinking water' is categorised into 1 = other (which includes unprotected well, river, rain, etc. other than protected well and tap/borehole), 2 = protected well, and 3 = tap/borehole. Since in Lesotho the average household consists of 5 people, household size (size refers to the number of people in the household) is categorised into 1 = more than 5 members, and 2 = less than or equal to 5 members. 'Roofing material, type of toilet, and fuel used for cooking' have 3 categories. 'Roofing' has 1 = other, 2 = thatch, and 3 = iron sheet categories; 'toilet' has 1 = none, 2 = other, and 3 = flush/pit latrine categories; and 'fuel' is categorised into 1 = firewood/charcoal, 2 = other, and 3 = kerosene/gas/electricity.

3.1 Preliminary Analysis

In this section some explanatory analysis of the data is presented. Table 3.1 displays the distribution of the number of households by each factor. Firstly, the distribution of households by each factor is discussed. A ratio of 2:1 for the number of households in rural to urban areas in the sample is in line with the ratio in the population in Lesotho. About 78% of the households owned their dwelling. The remaining 22% includes households who rented, used without paying rent, and used as temporary shelter. As expected, most of the households were headed by males, and these constituted about 64% of the households. The predominantly used roofing materials for the households' shelter were thatch (about 66% of the households) and other (about 30% of the households). Thatch is mainly used by rural poor households in Lesotho. About 65% of the households were headed by people who had some primary or no education, followed by 25% of those who had completed primary, and 10% completed secondary education. Again, 65% of the households were headed by people aged 40 years or older, and 35% were headed by people with less than 40 years of age.

Regarding marital status, about 58% of the households were headed by unmarried people. The majority of households had more than 5 members (68%). About 21% of the households used piped or borehole water, 10% of them used protected wells, and the rest (69%) used other sources of water. Most of the households (47%) used flush or pit latrine toilets, followed by those that did not have toilets (40%), and the rest (13%) used other types of toilets. About 96% of households in Lesotho used firewood, charcoal, kerosene, gas, or electricity as a source of fuel for cooking: with those that used firewood/charcoal constituting 56%, and those that used kerosene/gas/electricity constituting 40%. Supply of drinking water seemed to be accessible to the majority of households given by about 79% of the households that took less than 15 minutes to

Table 3.1: Summary of the data - Distribution of the number of households

Characteristic	n	%	No sick/injured	%
All	4954	100	2795	56
Household size				
6+ members	3392	68	1775	52
<6 members	1562	32	1020	65
Ownership of Dwelling				
Yes	3873	78	2335	60
No	1081	22	460	43
Education level of household head				
None and primary	3210	65	1970	61
Completed Primary	1236	25	608	49
Completed Secondary	508	10	217	43
Age of the household head				
< 40	1716	35	793	46
40+	3238	65	2002	62
Marital status of the household head				
not married	2883	58	1671	58
married	2071	42	1124	54
Sex of the household head				
Male	3151	64	1748	55
Female	1803	36	1047	58
District				
Butha-Buthe	485	10	250	52
Leribe	499	10	327	66
Berea	475	10	270	57
Maseru	631	13	322	51
Mafeteng	496	10	280	56
Mohale's Hoek	437	9	280	64
Quthing	482	10	298	62
Qacha's Nek	461	9	300	65
Mokhotlong	499	10	253	51
Thaba-Tseka	489	10	215	44

Table 3.1: Summary of the data (continues)

Characteristic	n	%	No sick/injured	%
All	4954	100	2795	56
Fuel used for cooking				
Firewood/Charcoal	2761	56	1671	61
Kerosene/gas/parafin	1984	40	990	50
Other	209	4	134	64
Type of toilet				
None	1977	40	1157	59
Flush/pit latrine	2353	47	1242	53
Other	624	13	396	63
Source of drinking water				
Protected well	481	10	262	55
Piped/borehole	1041	21	648	62
Other	3432	69	1884	55
Roofing material of dwelling				
Thatch	3265	66	1797	55
Iron sheets	213	4	119	56
Other	1476	30	879	60
Rural/Urban				
Rural	3230	65	1936	60
Urban	1724	35	859	50
Time taken to source of water				
< 15 minutes	3926	79	2178	55
15+ minutes	1028	21	617	60
Time take to hospital/clinic				
< 60 minutes	2409	49	1257	52
60+ minutes	2545	51	1538	60

reach the nearest supply of drinking water. This means, about 21% of the households were still taking 15 minutes or more to reach the nearest source of drinking. For accessibility of health services, the results show that about 51% of the households took 60 minutes or more to reach the nearest hospital/clinic. On average, each district has about 10% of the households who participated in the survey: Maseru having the highest with 13%, Mohale's Hoek and Qacha's Nek having the lowest with 9% each, and the rest having 10%.

Secondly, proportion of households with sick/injured members is discussed. On average, about 56% of the households had at least one sick/injured member. In the urban areas the distribution is balanced, that is, about 50% of households experienced illness/injury and 50% did not. A different distribution is observed in the rural areas where about 60% of the households experienced illness/injury. A similar incidence rate of illness/injury is observed for the households headed by males and females which stood at 55% and 58%, respectively. For the household size, about 52% of the households that had more than 5 members were unhealthy and about 65% of those with 5 or fewer members experienced disease/injury. The following are the categories which had less than 50% of unhealthy households: the households headed by people who completed primary education (49%), secondary education (43%), aged below 40 (46%), did not own dwelling (43%), and lived in Thaba-Tseka (44%). This basically means that over 50% of the households headed by people who completed primary education, completed secondary education, and are less than 40 years old did not experience disease/injury problems.

It is also observed in the categories that follow that more than 50% of the households experienced illness/injury problems: the households headed by people who had no education or some primary education (61%); the households that owned their dwelling (60%); the households headed by married (54%) and unmarried (58%) people; the households headed by people aged 40 years or older (62%); the households in all districts except those in Thaba-Tseka, with the households in Leribe having the highest proportion (66%); the households that used firewood/charcoal (61%), and those that used other type of fuel (64%); the households that had no toilet facilities (59%), used flush/pit latrine toilets (53%), and used other types of toilets (64%); households that used water from protected wells (55%), tap/borehole (62%), and other sources (55%); the households that used thatch (55%), iron sheets (56%), and other (60%) roofing material for households' shelter; the households that took less than 15 minutes to reach the nearest source of water (55%), and those that took 15 minutes or more (60%); and the households that took less than 60 minutes to reach the nearest hospital/clinic (52%), and those that took 60 minutes or more (60%).

Chapter 4

Fitting the Generalized Linear Model

Consider the i^{th} household (i = 1,2,. . . ,4954) characterised by $\boldsymbol{x_i'}$ which is an i^{th} row of the design matrix discussed in Section 2.5. Let the response $y_i = 1$ if disease/injury is present in the i^{th} household and be 0 otherwise. Again let $\pi_i = P(y_i = 1|\mathbf{x}_i')$ i.e. be the probability that disease/injury is present in the i^{th} household. The model discussed in Chapter 2 is fitted under the three link functions discussed in Section 2.5. For the logit link function the model will be called the Logit Model; for the probit link function, the model will be called the Probit Model; and for the complementary log-log link function the model will be called the Complementary log-log Model. The 14 variables discussed in Chapter 1 and Chapter 3 will be used together with their interaction terms as the independent variables in the model of the health status of the people of Lesotho. The health status measured by the presence or absence of disease/injury in the household is the dependent variable.

4.1 Logit Model

The logit model defined in (2.28) will be fitted to the data.

4.1.1 Model Selection

The stepwise selection procedure in SAS PROC LOGISTIC was used to select important factors affecting the health status. The factors with p-values less than 0.1 are given in Table 4.1 (see Section 2.4.1 for discussions on type 3 analysis of effects). All other factors excluded in the table have insignificant effects at the 10% significance level. Because of the disadvantage of dimensionality which saturated models suffer in terms of convergence of the estimation algorithm (Huang, 1998) only three factor interactions were allowed and the algorithm took more that 2 hours to converge. All the three factor interaction effects were insignificant, so are excluded in Table 4.1. Both first (which took 30 seconds to converge) and second order interaction models lead to the same model given in Table 4.1.

Table 4.1: Type 3 analysis of effects for the logit model

Effect	DF	Wald Chi-square	p-value
Location	9	90.7576	<.0001
Sex	1	8.4372	0.0037
Mstatus	1	5.4809	0.0192
Sex*Mstatus	1	17.8143	<.0001
Age	1	22.5970	<.0001
Mstatus*Age	1	10.8323	0.0010
Education	2	32.1698	<.0001
Dwelling	1	0.8186	0.3656
Mstatus*Dwelling	1	4.0541	0.0441
Education*Dwelling	2	14.3455	0.0008
HHsize	1	16.2029	<.0001
Dwelling*HHsize	1	4.5544	0.0328

The distribution of probabilities predicted by the logistic model with three competing link functions (i.e. logit, probit, and complementary log-log link functions) show that there are no extreme probability values (see Figure 4.1). Note that the complementary log-log model is preferred when there are many extreme values, and logit or probit models are preferred when there are few (or no) extreme values (Collett, 2003). Therefore, since there are no extreme values, the logit and the probit models were fitted to the data.

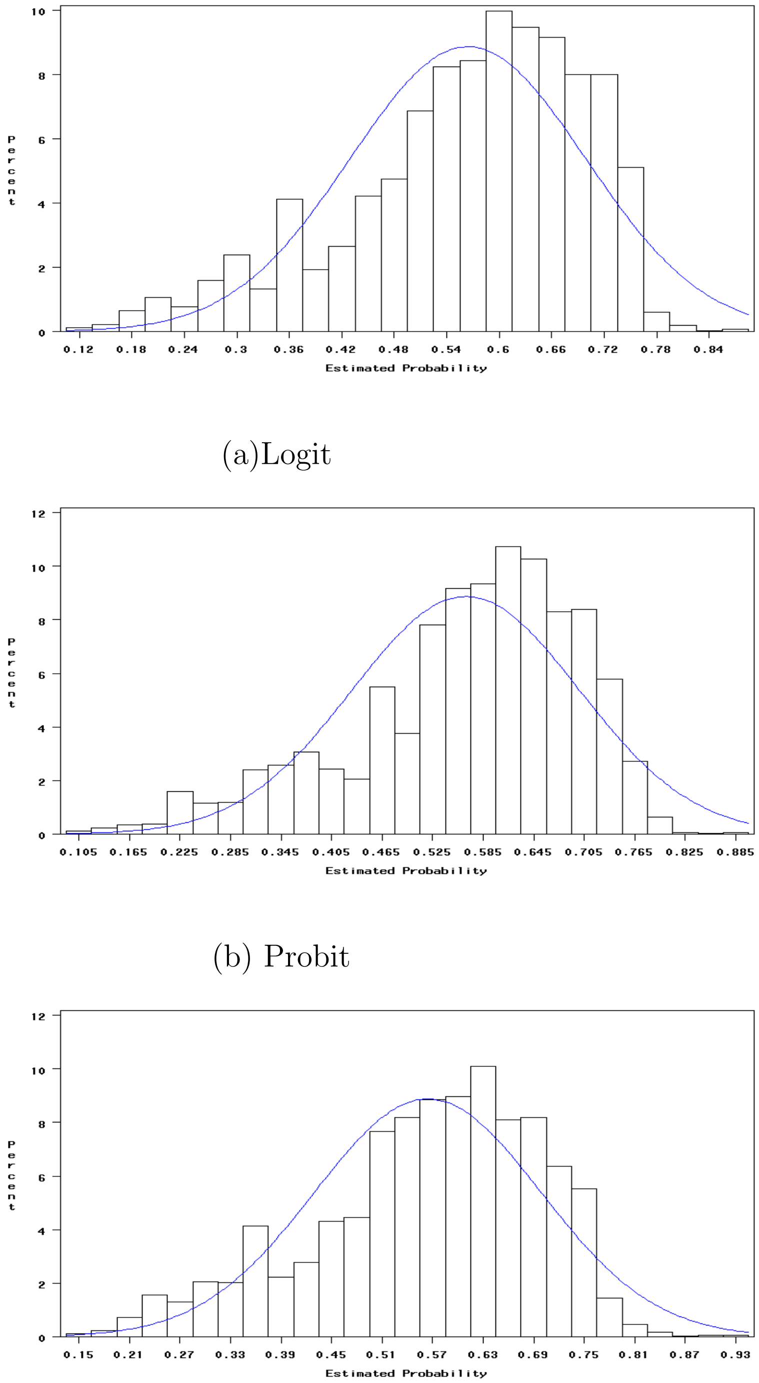
4.1.2 Model Checking

Goodness-of-Fit

The goodness-of-fit of the model can be tested using the Hosmer-Lemeshow test described in Section 2.5.3 and using 10 as the recommended number of groups. The observed and expected frequencies are given in Table 4.2. The goodness-of-fit statistic is 4.0195, with 8 degrees of freedom, and the corresponding p-value of 0.8554. The very

Table 4.2: Partition for the Hosmer-Lemeshow Goodness-of-Fit Test of the logit model

Group	Total	Event		Non-event	
		Observed	Expected	Observed	Expected
1	496	146	140.04	350	355.96
2	497	199	204.94	298	292.06
3	509	260	250.52	249	258.48
4	494	260	265.74	234	228.26
5	499	276	286.19	223	212.81
6	460	286	278.33	174	181.67
7	488	300	308.66	188	179.34
8	515	344	341.48	171	173.52
9	523	366	366.22	157	156.78
10	473	358	352.86	115	120.14



(a)Logit

(b) Probit

(c) Complementary Log-log

Figure 4.1: Probability distribution of predicted probabilities

large p-value for this test shows that the model fits the data well (i.e. the predicted probabilities agree with the observed values).

Link Function

The test for the appropriateness of the link function discussed in Section 2.5.3 is used here. The large p-value for the squared linear predictor and very small p-value for linear predictor variables in Table 4.3 suggest that the link is appropriate, agreeing with the goodness-of-fit test that the model fits the data well. The SAS procedure used for this test is given in Appendix A.1.2.

Table 4.3: Logit link function tests

Effect	DF	Chi-square	p-value
constant	1	0.06	0.7995
Linear predictor	1	330.02	<0.0001
Squared linear predictor	1	0.18	0.6673

Measure of Influence

As shown in Figure 4.2, none of the Cook's distance values for the fitted model is greater than 1, suggesting that there are no observations with undue influence on the parameter estimates. To verify this, the three observations with the largest Cook's distances were investigated further by refitting the model without each of them one at a time (referred to as single-case deletion). These observations are: number 3440 with Cook's distance = 0.002730291; number 3446 with Cook's distance = 0.002287120; and number 4250 with Cook's distance = 0.002183644. When these numbers were deleted (one at a time), the results were the same as those obtained from fitting the model to the full data set, confirming that the three observations do not have undue influence on the parameter estimates of the model.

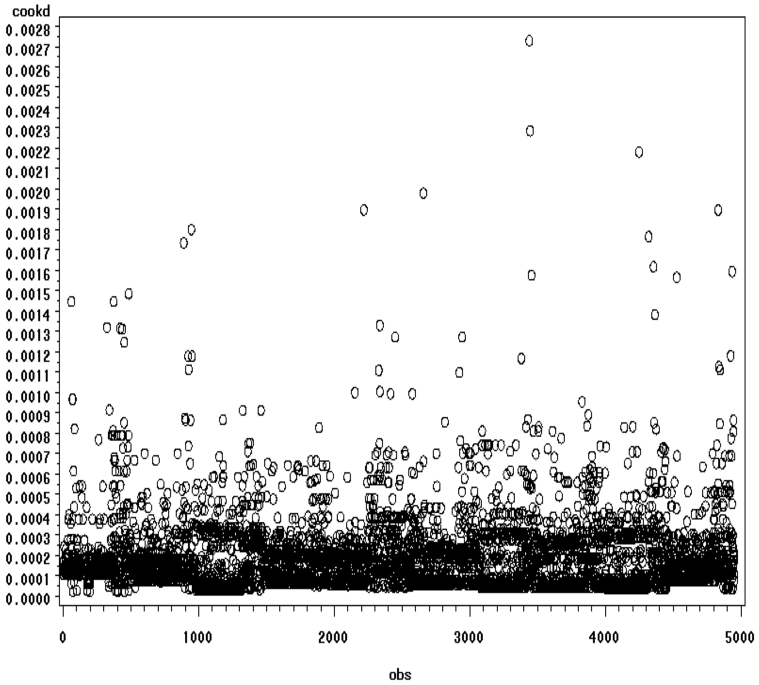


Figure 4.2: Index plot of Cook's distance for logit model: cookd = Cook's distance and obs = observation number

4.1.3 Prediction Accuracy of the Logit Model

The logit model is validated by checking its predictive accuracy, that is, checking by how often the model predicts unhealthy households as being unhealthy and healthy ones as healthy. Figure 4.3 displays the ROC curve of the fitted model. The area under the ROC curve is the proportion of the correctly predicted probabilities as was mentioned in Section 2.5.3. In this case about 65.21% of probabilities are predicted correctly, which is a moderate predictive accuracy (Taylor and Krawchuk, 2005). This measure is larger than the one obtained when the main effects model was fitted.

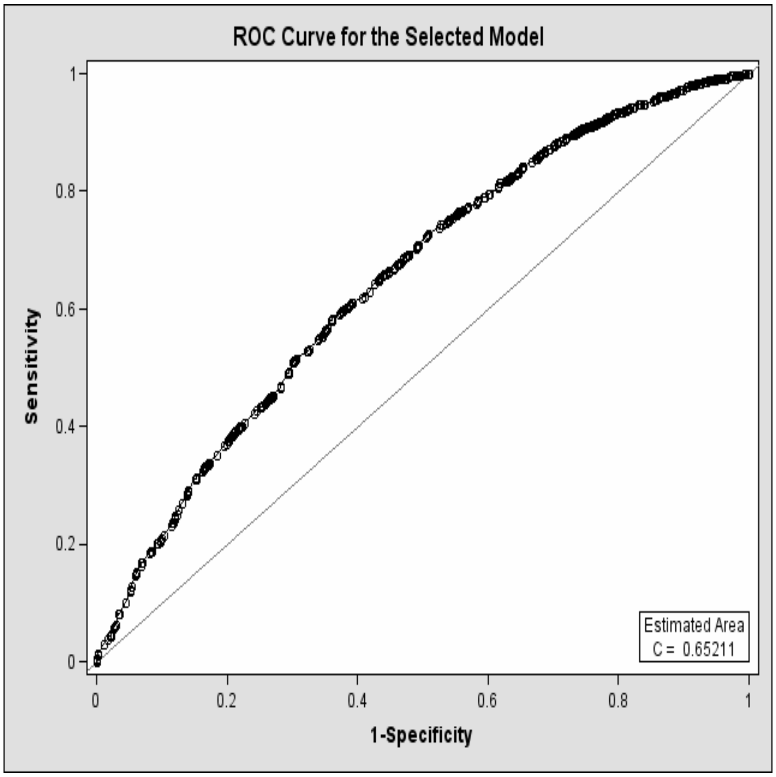


Figure 4.3: Sensitivity against 1-specificity of logit model with first order interaction

4.1.4 Interpretation of the Estimates of the Model Coefficients

Tables 4.5 and 4.6 contain the odds ratios of the incidence of disease/injury. The tables were constructed from the estimated model coefficients in Table 4.4. Note that the effects of Butha-Buthe, Berea, Mafeteng and Mokhotlong are not significant, which means that, controlling for other variables, the incidence of disease/injury is not different from that of Maseru (the reference district). But for other districts the incidence of disease/injury is significantly different from that of Maseru and the corresponding odds ratios are given in Table 4.5. The households in Leribe are 1.672 (between 1.355 and 2.065) times more likely to be ill/injured compared to those of Maseru. The values in brackets are 90% confidence limits of the parameters. The rate is slightly lower for the households in Mohale's Hoek, Quthing and Qacha's Nek which are 1.418 (between

Table 4.4: Parameter estimates under the logit model

Effect	Parameter	Estimates	Std errors	p-value	90% C. I.	
					Lower	Upper
Constant	$\hat{\beta}_0$	0.5978	0.3138	0.0568	0.0856	1.1197
Butha-Buthe	$\hat{\beta}_1$	-0.1978	0.1262	0.1171	-0.4055	0.0098
Leribe	$\hat{\beta}_2$	0.5140	0.1281	<.0001	0.3039	0.7253
Berea	$\hat{\beta}_3$	-0.0057	0.1274	0.9641	-0.2152	0.2040
Thaba-Tseka	$\hat{\beta}_4$	-0.4619	0.1264	0.0003	-0.6702	-0.2543
Mafeteng	$\hat{\beta}_5$	0.0492	0.1257	0.6956	-0.1575	0.2562
Mohale's Hoek	$\hat{\beta}_6$	0.3494	0.1332	0.0087	0.1308	0.5692
Quthing	$\hat{\beta}_7$	0.2922	0.1293	0.0238	0.0799	0.5055
Qacha's Nek	$\hat{\beta}_8$	0.3719	0.1313	0.0046	0.1565	0.5885
Mokhotlong	$\hat{\beta}_9$	-0.1816	0.1248	0.1456	-0.3869	0.0236
Sex	$\hat{\beta}_{10}$	-0.6770	0.1129	<.0001	-0.8634	-0.4917
Mstatus	$\hat{\beta}_{11}$	-0.5843	0.2131	0.0061	-0.9354	-0.2342
Sex*Mstatus	$\hat{\beta}_{12}$	0.8022	0.1901	<.0001	0.4894	1.1149
Age	$\hat{\beta}_{13}$	-0.6166	0.1258	<.0001	-0.8242	-0.4101
Mstatus*Age	$\hat{\beta}_{14}$	0.4932	0.1498	0.0010	0.2471	0.7401
Education1	$\hat{\beta}_{15}$	0.9009	0.1722	<.0001	0.6192	1.1860
Education2	$\hat{\beta}_{16}$	0.2449	0.1672	0.1430	-0.0290	0.5213
Dwelling	$\hat{\beta}_{17}$	0.0606	0.3376	0.8577	-0.4993	0.6127
Mstatus*Dwelling	$\hat{\beta}_{18}$	0.3365	0.1671	0.0441	0.0620	0.6119
Education1*Dwelling	$\hat{\beta}_{19}$	-0.7925	0.2245	0.0004	-1.1632	-0.4246
Education2*Dwelling	$\hat{\beta}_{20}$	-0.3578	0.2294	0.1187	-0.7363	0.0184
HHsize	$\hat{\beta}_{21}$	-0.8305	0.2592	0.0014	-1.2654	-0.4105
Dwelling*HHsize	$\hat{\beta}_{22}$	0.5717	0.2679	0.0328	0.1370	1.0203

1.140 and 1.767), 1.339 (between 1.083 and 1.658), and 1.45 (between 1.169 and 1.801) times more likely to be ill/injured, respectively. The households in Thaba-Tseka are 0.63 (between 0.512 and 0.775) times more likely to be ill/injured than those of Maseru.

Table 4.6 displays odds ratios corresponding to interaction effects. The incidence of disease/injury for the households headed by unmarried males is 2.23 (between 1.631 and 3.049) times that for the households headed by married males. Incidence of disease/injury of households headed by unmarried males is 2.23 (between 1.631 and 3.049)

Table 4.5: Odds ratios and their confidence limits for the residential area effects

Contrast	Odds ratios	90% CI of Odds ratios
LR vs MS	1.672	(1.355, 2.065)
MH vs MS	1.418	(1.140, 1.767)
QT vs MS	1.339	(1.083, 1.658)
QN vs MS	1.45	(1.169, 1.801)
TT vs MS	0.63	(0.512, 0.775)

Note: LR=Leribe, MS=Maseru, MH=Mohale's Hoek,

QT=Quthing, QN=Qacha's Nek, and TT=Thaba-Tseka

Table 4.6: Odds ratios and their confidence limits corresponding to interaction effects

Contrast	Odds ratios	90% CI of Odds ratios
M=1 vs M=2 / S=1	2.230	(1.631, 3.049)
S=1 vs S=2 / M=1	2.230	(1.631, 3.049)
A=1 vs A=2 / M=1	1.638	(1.280, 2.096)
D=1 vs D=2 / M=1	1.400	(1.064, 1.844)
M=1 vs M=2 / A=1	1.638	(1.280, 2.096)
M=1 vs M=2 / D=1	1.400	(1.064, 1.844)
E=1 vs E=3 / D=1	0.453	(0.312, 0.654)
H=1 vs H=2 / D=1	1.771	(1.147, 2.774)
D=1 vs D=2 / E=1	0.453	(0.312, 0.654)
D=1 vs D=2 / H=1	1.771	(1.147, 2.774)

Note: S=Sex, A=Age, (E=1)=No education, H=HHsize,

D=Dwelling, and M=Mstatus

times the one for households headed by unmarried females. The incidence for the households headed by young (less than 40 years old) unmarried people is 1.6375 (between 1.280 and 2.096) times the one for households headed by older unmarried people. Furthermore, households headed by unmarried people who do not own their dwellings are 1.4 (between 1.064 and 1.844) times more likely to be unhealthy than households of unmarried people who own their dwelling.

Furthermore, the households headed by young unmarried people are 1.6375 times more likely to be unhealthy than those headed by young married people. For education, one can interpret only coefficients associated with E=1 (have no education or have some primary) since the ones associated with E=2 (completed primary) are not statistically significant. The odds ratio of 0.4527 (between 0.312 and 0.654) corresponding to the interaction effect of education and ownership of dwelling implies that the households headed by people with no education (including some primary) who do not own their dwelling are less likely to be unhealthy than those headed by their counterparts who own their dwelling. That means, the incidence is higher for the households headed by uneducated people who own their dwelling than for those headed by uneducated people who do not own their dwelling.

The incidence of disease/injury for households headed by unmarried people who do not own dwelling is 1.4 (between 1.064 and 1.844) times the one for those headed by married people who do not own their dwelling. But the incidence for those headed by uneducated people who do not own their dwelling is 0.4527 (between 0.312 and 0.654) times the one for their educated counterparts. Large households not owning their dwelling are 1.7713 (between 1.147 and 2.774) times more likely to be unhealthy than small households not owning their dwelling.

Moreover, the large households (with more than 5 members) who do not own their dwelling are 1.7713 (between 1.147 and 2.774) times more likely to be unhealthy than large households who own their dwelling.

4.2 Probit Model

The Probit model is given by

$$\eta_i = \Phi^{-1}(\pi_i) = \boldsymbol{x_i'\beta} \ , \qquad i = 1, 2, \ldots, 4954. \tag{4.1}$$

4.2.1 Model Selection

Type 3 analysis of effects as discussed in Section 2.4.1 is given in Table 4.7 for the selected probit model. Note that this set of predictor variables is the same as the one selected for the logit model. The insignificant effect of ownership of dwelling is included because of the hierarchy principle of the models with interaction effects. But before making any inferences about the factor effects, (a) the goodness-of-fit of the model is tested, (b) the diagnostic test of the appropriateness of the link function is performed, and (c) the presence/absence of influence of the observations on the parameter estimates is checked.

Table 4.7: Type 3 analysis of effects for the probit model

Effect	DF	Wald Chi-square	p-value
Location	9	91.5309	<.0001
Sex	1	8.4143	0.0037
Mstatus	1	5.5497	0.0185
Sex*Mstatus	1	18.1149	<.0001
Age	1	22.7689	<.0001
Mstatus*Age	1	11.1042	0.0009
Education	2	32.3403	<.0001
Dwelling	1	0.7928	0.3732
Mstatus*Dwelling	1	4.2331	0.0396
Education*Dwelling	2	14.3556	0.0008
HHsize	1	16.6989	<.0001
Dwelling*HHsize	1	4.7417	0.0294

Table 4.8: Partition for the Hosmer-Lemeshow Goodness-of-Fit Test of the probit model

Group	Total	Event		Non-event	
		Observed	Expected	Observed	Expected
1	496	146	139.92	350	356.08
2	499	200	206.60	299	292.40
3	501	255	246.94	246	254.06
4	495	262	266.10	233	228.90
5	495	273	283.39	222	211.61
7	505	310	319.42	195	185.58
8	495	331	328.21	164	166.79
9	466	321	325.15	145	140.85

4.2.2 Model Checking

Goodness-of-Fit

Table 4.8 is used to perform the Hosmer-Lemeshow goodness-of-fit test for the probit model. The test statistic is 4.6951, with 8 degrees of freedom, and a corresponding p-value of 0.7896. The large p-value indicates that the probit model fits the data well.

Link Function

The test for the appropriateness of the link function discussed in Section 2.5.3 is used. The large p-value for the squared linear predictor and the very small p-value for linear predictor variables in Table 4.9 suggest that the probit link is appropriate. This confirms the theory that for moderate probability values π_i's, the logit and the probit link functions lead to the same model fit to the data.

Table 4.9: Probit link function test

Effect	DF	Chi-square	p-value
Constant	1	0.09	0.7703
Linear predictor	1	350.53	<0.0001
Squared linear predictor	1	0.24	0.6269

Measure of Influence

Figure 4.4 shows that there are no Cook's distance values greater than 1. This means that all the observations do not have undue influence on the parameter estimates of the model. This was confirmed by refitting the model without (one at a time) the three observations with the largest Cook's distance values. These were observation numbers: 3440 with Cook's distance = 0.00459336, 3446 with Cook's distance = 0.004248518, and 4250 with Cook's distance = 0.003768503. There was no change in the results of the model when these observations were deleted, each at a time, from that obtained from the model fitted to the full data set, indicating that there were no observations with undue influence on the parameter estimates of the model.

4.2.3 Predictive Accuracy

The predictive accuracy of this model is similar to that of the Logit model, standing at 65.2% as shown in Figure 4.5. Since the probit and logit models selected the same model, either of the two can be chosen. Therefore, inference based on the logit model will suffice.

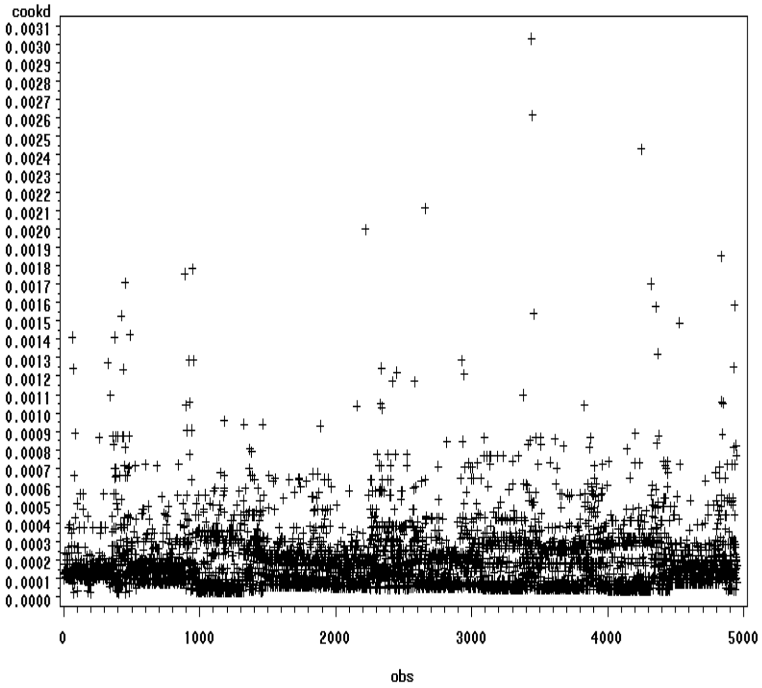


Figure 4.4: Index plot of Cook's distance for the probit model

4.3 Discussion

Location (districts), ownership of dwelling, household size, sex, marital status, education and age of the household head were found to be important factors affecting the health status of the Basotho people. The results suggest that the incidence of disease in the households is likely to reduced if households occupy their own dwelling, especially households with more than 5 members. Education is also correlated with good health status of the people. This is evidenced by a high incidence of disease/injury among the households headed by uneducated people, which could be associated with a number of factors, such as poverty, lack of information about health issues, and others. Poverty makes it hard for people to provide basic needs, such as food, medication, and clothing, for their household members. The importance of education in health status is also emphasized by Hussain and Smith (1999), who found that about 60% of the

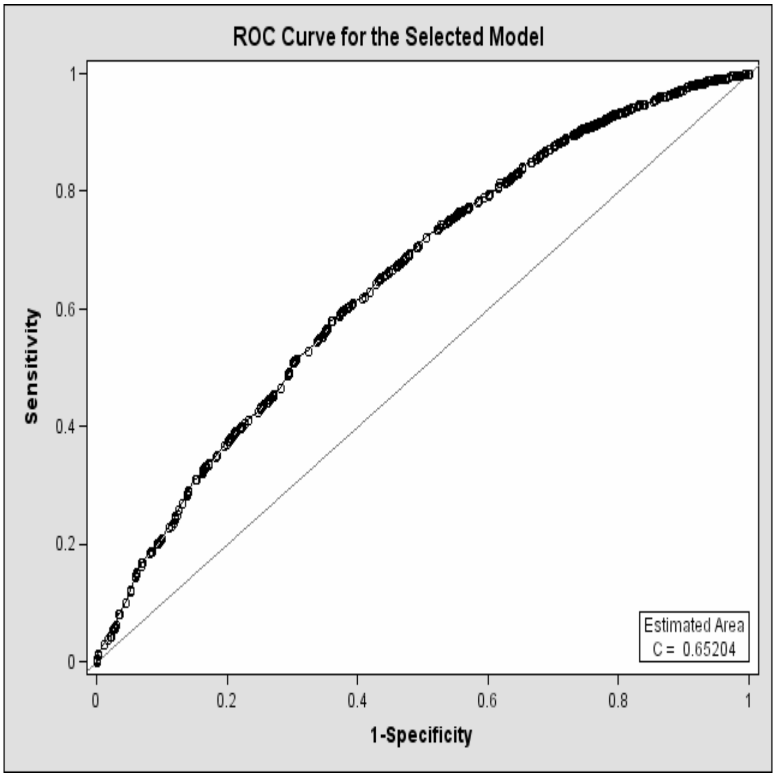


Figure 4.5: Sensitivity against 1-specificity for probit model

children whose mothers have secondary and higher education are less likely to have diarrhea than children of those who have no education.

The households not owning their dwelling, with more that 5 members, are more likely to be unhealthy compared to their counterparts who own their dwelling. Analogously, for those who do not own their dwelling, the incidence of disease is higher for large households than it is for small ones. This implies that small households (which have less risk of infectious diseases due to crowding) and owner-occupied dwellings (which leads to reduced environmental risk) are better off. This agrees with the findings of Howden-Chapman (2004) and Dedman et al. (2001). For age of the household heads, it is found that the incidence of diseases/injuries is higher for households headed by younger (< 40 years) unmarried people compared to older (40+ years) unmarried people. Households headed by young unmarried people are more likely to be unhealthy compared to their young married counterparts.

The incidence of diseases/injuries in Berea, Mafeteng, Butha-Buthe and Mokhotlong is not significantly different from the one observed in Maseru. Thaba-Tseka is the only district with a relatively lower significant incidence compared to Maseru. The districts in the southern part of Lesotho namely, Mohale's Hoek, Quthing, and Qacha's Nek and Leribe, which is in the north, have significantly higher incidence than that prevailed in Maseru. An in-depth research into the cause of this difference is necessary.

It is suggested that attention should be focused on Leribe, Mohale's Hoek, Quthing, and Qacha's Nek to address the problem of health status inequality that exists between Maseru and these areas. Attention should be focused on these areas even more so to improve existing socio-economic programmes such as education, health care, and social welfare, or to develop new ones. This suggests the need to develop some kind of system that can be utilized to foster owner-occupied dwellings, encourage a reasonable household size, and improve awareness campaigns on health issue to the entire community (especially males) of all age groups. It is suggested that strategies are developed for dealing effectively with the marriage issue that will encourage sustained marriages for the benefits that marriage offers.

4.4 Shortcomings of the Generalized Linear Model

Recall that Core Welfare Indicators Questionnaire Survey (CWIQ) data do not come from a simple random sample that generalized linear models assume. Thus, the generalized linear model employed above does not capture the structure induced by the sampling design. It does not allow estimation of random effects which account for the correlation of data resulting from homogeneity of outcomes within and heterogeneity among clusters from which data are collected (Berlin et al., 1999). The failure of this model to incorporate this phenomenon into the analysis may lead to biased variances of estimates and wrong inferences about those estimates (Zeger and Liang, 1986; Donald and Donner, 1987; and Rabe-Hesketh and Skrondal, 2006), although estimates themselves could be accurately calculated.

Since clusters, (PSUs), which are Enumeration areas were randomly selected, they can be incorporated into the model in a number of ways. For example, they can be incorporated in such a way that parameter estimates vary from cluster to cluster, leading to cluster-specific models. However, these models have a problem of size, i.e. they increase with the number of clusters. If this model is fitted to the CWIQ data, where there are 258 clusters, the resultant model will be very large. This problem of size can be solved by assuming that these clusters are random samples from the underlying population of clusters and by including clusters as a random effect in the model (Pendergast et al., 1996). The models restricted to random effects corresponding to clusters are referred to as random intercept models, where conditional independence within the cluster is commonly assumed (Pendergast et al., 1996). Because of the flexibility of these models they can also be fitted even when the assumption of conditional independence is questionable (Pendergast et al., 1996). Standard errors (or variances) of the parameter estimates for these models are calculated using re-sampling methods such as sample replicates, balanced repeated replication samples, jacknife samples, and Taylor series (expansion) methods.

According to Korn and Graubard (2002) parameters associated with simple marginal models tend to be the ones of most scientific interest. For models with random effects, maximum likelihood estimation methods require optimization of the marginal distribution of the data with respect to the fixed effects and the variance parameters (Pendergast et al., 1996). Pendergast et al. (1996) add that, since there is no closed form expression for the marginal distribution, numerical or Monte Carlo integration should be used to calculate the corresponding likelihood. When the response is binary, maximum likelihood estimation cannot be used for a fully parametric model (Waclawiw and Liang, 1993). Instead, the maximum pseudo-likelihood estimation method can be used which leads to a generally asymptotically consistent, but not efficient estimators (Pendergast et al., 1996; Christensen, Hobolth, and Jensen, 2005; Zhang, 2002; and Rabe-Hesketh and Skrondal, 2006), which can be obtained under suitable regularity conditions (Rabe-Hesketh and Skrondal, 2006). One of the conditions is the inclusion of design variables in the model as explanatory variables (Pfeffermann, 1993).

A generalized linear mixed model which allows for both random and fixed effects, especially the random intercept model, will be fitted to the data and discussed in Chapter 5. A survey logistic regression model designed specifically for data from surveys, will be fitted and shown in Chapter 6.

Chapter 5

Generalized Linear Mixed Models

The generalized linear models discussed in Chapter 2 may not be appropriate for the data discussed in Chapter 3 because they ignore the survey design in the sense that the random PSUs effect on health status is ignored. When the random PSUs effect is included in the analysis the models become generalized linear mixed models. In the sections that follow the theory of these models is reviewed and the models are also fitted to the data.

5.1 General Linear Mixed Models

To introduce generalized linear mixed models let one consider the situation where the vector of response variable **y** is normally distributed, given the vector of random effects **u**. This leads to the general linear mixed models which are extensions of general linear models discussed in Section 2.1 by including the vector of random effects **u**. The general linear mixed models have the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon} \tag{5.1}$$

where

$\mathbf{y}$ is an n×1 vector of responses

$\mathbf{X}$ is an n×(p+1) design matrix for fixed effects

$\boldsymbol{\beta}$ is a (p+1)×1 vector of unknown fixed effects parameters

$\mathbf{Z}$ is an n×q design matrix for random effects

$\mathbf{u}$ is a q×1 vector of unknown random effects parameters, assumed to have a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{G}$, i.e. $\mathbf{u}\sim N_q(\mathbf{0},\mathbf{G})$

ϵ is an n×1 vector of error terms which have a multivariate normal distribution, with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{R}$, i.e. $\boldsymbol{\epsilon} \sim N_n(\mathbf{0},\mathbf{R})$

Analogous to general linear models, general linear mixed models require that responses have normal distributions. Models that accommodate both normal and non-normal data which belong to exponential family of distributions are called generalized linear mixed models. The linear mixed models are special cases of the generalized linear mixed models.

5.2 Generalized Linear Mixed Models (GLMMs)

The GLMMs have the same features as generalized linear models. Recall that the linear predictor for the generalized linear models is $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$. When the random effects are included in the models one has GLMMs given by

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} \tag{5.2}$$

where η_i = g(μ_i), g is a link function, and $\boldsymbol{\mu} = E(\mathbf{y}|\mathbf{u})$. Parameter estimates of the model are obtained by partially differentiating (5.2) with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$, and iteratively solving the resulting estimating equations (Littell et al., 1996 & 2006). These equations are given by

$$\begin{bmatrix} \mathbf{X}'\mathbf{W}\mathbf{X} & \mathbf{X}'\mathbf{W}\mathbf{Z} \\ \mathbf{Z}'\mathbf{W}\mathbf{X} & \mathbf{Z}'\mathbf{W}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{W}\mathbf{y}^{*} \\ \mathbf{Z}'\mathbf{W}\mathbf{y}^{*} \end{bmatrix} \tag{5.3}$$

where

$\mathbf{y}^{*} = \hat{\boldsymbol{\eta}} + (\mathbf{y} - \hat{\boldsymbol{\mu}})\mathbf{D}^{-1}$ is referred to as the working dependent variate,

$\mathbf{W} = \mathbf{D}'\mathbf{R}^{-1}\mathbf{D}$,

$\mathbf{D}_{n\times n} = \frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\eta}} = [\frac{\partial\mu_i}{\partial\eta_j}]$ i = 1,2,. . .,n, j = 1,2,. . .,n,

$\mathbf{R} = \text{var}(\mathbf{y}|\mathbf{u}) = \mathbf{R}_{\mu}^{1/2}\mathbf{A}\mathbf{R}_{\mu}^{1/2}$, where $\mathbf{R}_{\mu}^{1/2}$ is a diagonal matrix of the square root of $\mathbf{R}_{\mu}$ whose i^{th} diagonal element is the variance function of the i^{th} response, and $\mathbf{A}$ is the scale parameter matrix whose i^{th} diagonal element is a(ϕ_i), and

$\mathbf{G} = \text{var}(\mathbf{u})$.

The expected value of the response vector, given the vector of random effect E($\mathbf{y}|\mathbf{u}$) = $\boldsymbol{\mu}$ obtained from the rearrangement of terms in (5.2) is given by

$$\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) = g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) \tag{5.4}$$

where $g^{-1}(.)$ is inverse link function.

Consider the application of GLMMs in the analysis of the CWIQ data where the effects of the PSUs, which were randomly selected, enter the model as random effects. The same variables selected in Chapter 4 will now be the fixed effects in the model. This gives a mixed model. Recall that the aim of this study is to model the health status of the people of Lesotho, where the response variable y is binary (1 = presence or 0 = absence of disease/injury). The distribution of y belongs to the exponential family of distributions required for generalized linear models (if the model has only fixed effects) and GLMMs (if both fixed and random effects are included in the model). This justifies the fitting of GLMMs to the CWIQ data to achieve the objectives of this research.

Since there is only one random factor effect (i.e. PSU), a special and simplest form of the GLMMs can be fitted (Pendergast et al., 1996). This model (called the random intercept model) is given by

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \ , \qquad \mathbf{u} \sim N_q(\mathbf{0}, \sigma_u^2\mathbf{I}) \tag{5.5}$$

where $\mathbf{X}$ and $\boldsymbol{\beta}$ are as discussed in (5.1) and $\mathbf{u}$ is a random vector of PSU effects whose i^{th} element represents the influence of the i^{th} PSU on household observations not captured by the observed covariates. This shows that even if the interest is in the estimation of fixed effects, random effects which characterise the degree of heterogeneity of the target population play an important role (Agresti et al., 2000). In the random intercept model, the random effect adjusts the overall intercept $\boldsymbol{\beta}_0$ in the model (Littell et al., 2006). Because $\boldsymbol{\beta}$, $\mathbf{u}$, and $\mathbf{G}$ are interrelated, their estimation must be carried out in a coherent and systematic manner (Waclawiw and Liang, 1993).

5.2.1 Estimation of the Model Parameter

For the GLMMs, the calculations of the likelihood function for making statistical inferences is sometimes not easy (Littell et al., 2006). According to the authors, obtaining the marginal distribution is not easy if the conditional distribution of $\mathbf{y}$, given $\mathbf{u}$, is not normal. Hence, one way of applying the linear mixed model is by using an approximated model where the estimation is done repeatedly until convergence (referred to as pseudo-likelihood (PL) approach) (Littell et al., 2006). Thus, since the likelihood function is not easy to construct for GLMMs where data are non-normal, the PL or the restricted pseudo-likelihood (REPL) proposed by Wolfinger and O'Connell (1993), and the penalized quasi-likelihood (PQL) proposed by Breslow and Clayton (1993) are used. The REPL is based on the assumption that the dispersion parameter ϕ is unknown, whilst for the PQL, ϕ is assumed to be fixed at 1 when modelling Binomial (or Binary) or Poisson data. For further comparison of the likelihood, the quasi-likelihood (QL) and the PL methods see Nelder and Lee (1992). The authors assert that these three methods of estimation are equivalent in the case of normal data.

Rabe-Hesketh and Skrondal (2006) report that the full maximum pseudo-likelihood estimation method is a better method for GLMMs than other competing methods. The method involves maximization of the log pseudo-likelihood function using optimization routines.

Recall that the estimating equations for GLMMs are solved iteratively to obtain parameter estimates. For binary response GLMMs, the terms in (5.3) are:

$\mathbf{y}^* = \hat{\boldsymbol{\eta}} + (\mathbf{y} - \hat{\boldsymbol{\pi}})\mathbf{D}^{-1}$ is referred to as a working (or pseudo) dependent variate

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

$\mathbf{W} = \mathbf{D}'\mathbf{R}^{-1}\mathbf{D}$

$\mathbf{D} = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}} = \mathrm{diag}[\pi_i(1-\pi_i)]$

$\mathbf{R} = \mathrm{var}(\mathbf{y}|\mathbf{u}) = \mathbf{R}_\mu^{1/2}\mathbf{A}\mathbf{R}_\mu^{1/2} = \mathbf{R}_\mu^{1/2}\mathbf{I}\mathbf{R}_\mu^{1/2}$

$\mathbf{R}_\mu = \mathrm{diag}[\pi_i(1-\pi_i)]$

$\mathbf{A} = \mathbf{I} =$ identity matrix

$\mathbf{G} = \mathrm{var}(\mathbf{u}){=}\mathbf{I}\sigma_u^2.$

The following features of the ungrouped binary conditional model are observed:

1. Conditional mean: $\mu_i = \pi_i = 1/\{1 + exp(-\eta)\}$;

2. Natural parameter: $\theta(\mu_i) = -log(\pi_i^{-1} - 1)$;

3. Variance function: $\mathrm{V}(\mu_i) = \pi_i(1 - \pi_i)$; and

4. Dispersion parameter: $\mathrm{a}(\phi_i) = 1$.

Numerical methods are used to obtain $\hat{u}_j$, which should follow a normal distribution with mean zero and variance $\hat{\sigma}_u^2$ for a correctly fitted model (Collett, 2003). This procedure is available in statistical packages such as SAS. Collett (2003) reports that the maximum likelihood estimation methods are not easy to use for calculating marginal parameter estimates, especially if the random vector component has more than one effect. That is, to obtain the marginal parameter estimates, the likelihood function has to be integrated over each of the random components. Possible methods that can be used include (among others) the QL, the PL or the Gibbs sampler based methods and their extensions. In this study, the PL based methods (especially residual or restricted PL) will be used which, according to Pendergast et al. (1996), leads to asymptotically consistent estimators. These methods are implemented in SAS GLIMMIX Procedure (see Littell et al. (1996 & 2006) and the SAS GLIMMIX Procedure Manual (2005) for more details). Estimates from (5.3) can be simplified as follows:
Profiled parameter (Fixed effects) estimates are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{y}^*$$

and the BLUP predictor of the random vector effect $\mathbf{u}$ is

$$\hat{\mathbf{u}} = \hat{\boldsymbol{G}}\mathbf{Z}'\mathbf{V}(\boldsymbol{\theta})^{-1}\hat{\mathbf{r}}$$

where $\hat{\mathbf{r}} = \mathbf{y}^* - (\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{y}^*$ for $\mathbf{y}^* = \hat{\boldsymbol{\eta}} + (\mathbf{y} - \hat{\boldsymbol{\pi}})\mathbf{D}^{-1}$, $\boldsymbol{\theta}$ is a q×1 vector of parameters containing all unknowns in $\mathbf{G}$ and $\mathbf{R}$ and $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{D}^{-1}\mathbf{R}_\mu^{1/2}\mathbf{A}\mathbf{R}_\mu^{1/2}\mathbf{D}^{-1}$, where $\mathbf{D} = (\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\eta}})_{\hat{\beta},\hat{u}}$. The parameter and random effect estimates are used to update the pseudo-response and weights which are in turn used to update parameter and random effect estimates. This process is continued until the convergence criterion is met: that is until the difference between parameters at two successive iterations is sufficiently small. When ϕ is different from 1, parameter estimates are also profiled from the log PL. The parameter ϕ in the model is estimated by

$$\hat{\phi} = \hat{\mathbf{r}}'\mathbf{V}^{-1}\hat{\mathbf{r}}/m \tag{5.6}$$

where $m = n$ (where n is the number of individuals used in the analysis) for MPL and $m = n - p$ (where p is the rank of $\mathbf{X}$) for RPL. Since ungrouped binary data is considered here, there is no problem of dispersion, so this issue is not discussed further.

According to Littell et al. (2006) the predictable functions, which are the primary tools of inference for GLMMs, are tested using the Wald statistic or F-statistic if the conditional variance $\mathbf{R}$ depends on a known or unkown scale parameter matrix $\mathbf{A}$. This is briefly discussed in Littell et al. (2006) and in more detail in the SAS GLIMMIX procedure manual.

5.2.2 Interpretation of the Results in Terms of the Least-squares Means Differences

Another form of inference about the parameters of the current fitted model is achieved by using the least-squares means differences of the response measured at different factor levels. This type of inference about parameters in the fitted model is more appropriate for general linear models (see Hsu and Peruggia, 1994; and Hsu, 1996). Littell et al. (2006) discussed this approach of inference in the context of generalized linear mixed models with examples for Binomial and Poisson data. According to Littell et al. (2006) least-squares means for factor levels are computed using estimable functions and they refer to them as 'least-squares means' (which are on the link scale). The least-squares means will be denoted by μ.

The factor least-squares means can be presented in tabular form or graphically.

The pairwise comparison of factor level least-squares means ($\mu_i - \mu_j$, for all $i \neq j$) and comparison of each factor level least-squares mean against the overall average of all factor levels least-squares means ($\mu_i - \bar{\mu}$, where $\bar{\mu}$ represents the overall least-squares mean) will be implemented in this study. For the pairwise comparison, the mean-mean scatter plot of least-squares means (see Hsu, 1996) called a 'Diffogram' in Littell et al. (2006) will be used. According to the authors, the Tukey-Kramer method of adjustment for multiplicity in the pairwise comparisons is preferable. For comparison of each factor level least-squares means against the average factor least-squares mean (which will be called 'Analysis of Means'), the Nelson method of adjustment for multiplicity is used (Littell et al., 2006; and Nelson, 1985 & 1993).

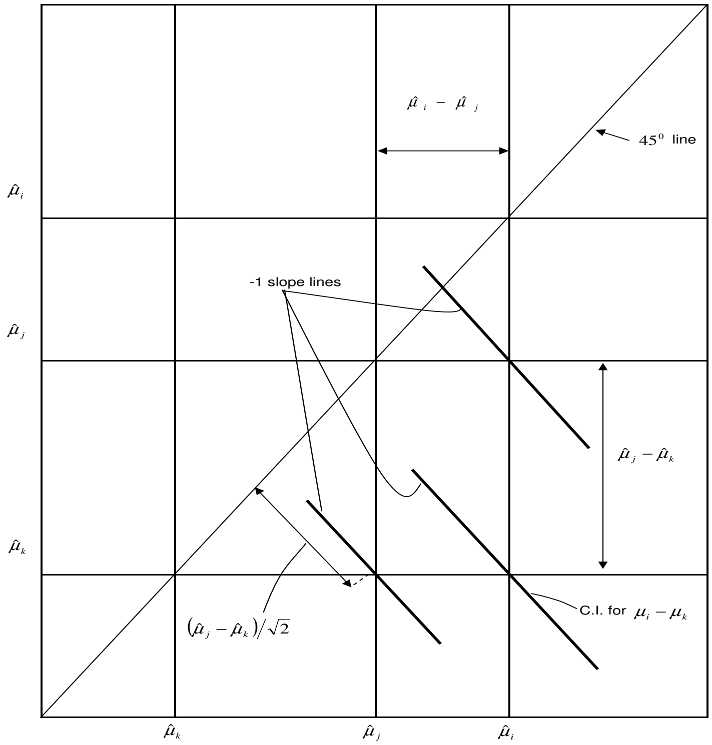


Figure 5.1: Diffogram

Figure 5.1 displays a Diffogram. The axes of Diffogram plot are the least-squares means, i.e. y-axis = $\hat{\mu}$ and x-axis = $\hat{\mu}$. The 45^0 line from the origin is a reference line which corresponds to the set of points satisfying $\hat{\mu}_i = \hat{\mu}_j$ for all i and j. The directional distance of any point from the 45^0 line is given by the difference of the two corresponding least-squares means divided by the square root of 2. For example, the directional distance of the point $(\hat{\mu}_j, \hat{\mu}_k)$ from the 45^0 line is given by $(\hat{\mu}_j - \hat{\mu}_k)/\sqrt{2}$, and of the point $(\hat{\mu}_i, \hat{\mu}_j)$ from 45^0 line is given by $(\hat{\mu}_i - \hat{\mu}_j)/\sqrt{2}$. The Tukey-Kramer's confidence interval for the difference between two least-squares means is represented by the length of the '-1 slope' lines (or lines perpendicular to the 45^0 line). For example, the Tukey-Kramer's confidence interval for $\hat{\mu}_i - \hat{\mu}_j$ is given by the length of the '-1 slope' line centered at the intersection of $\hat{\mu}_i$ and $\hat{\mu}_j$, and for $\hat{\mu}_i - \hat{\mu}_k$ is given by the length of the line centered at the intersection of $\hat{\mu}_i$ and $\hat{\mu}_k$.

The longer the '-1 slope' line the wider the confidence limits of the difference between least-squares means. If the difference between two least-squares means is significant, the corresponding line will not cross the 45^0 (reference) line, and vice versa. The difference between $\hat{\mu}_j$ and $\hat{\mu}_k$ is significant, whilst that between $\hat{\mu}_i$ and $\hat{\mu}_j$ is not significant. It should be noted that the '-1 slope' lines are adjusted for rotation and multiplicity and all the estimates are on the link scale. For more details on this discussion see Hsu and Peruggia (1994), Hsu (1996), and Littell et al. (2006).

The graphical presentation of the 'Analysis of Means' (where least-squares means for each factor level are compared against the average of all levels) has a different representation to that of the Diffogram. The x-axis here represents factor levels and the y-axis represents least-squares means (on the linked scale). The average of the least-squares means is given by the horizontal line in the center of the graph. The vertical lines from the horizontal line represent the magnitude of the difference of the factor levels least-squares means from the average least-squares means. On both sides of the horizontal line there are dashed horizontal step plots representing the lower decision limit (LDL) and upper decision limit (UDL). If the least-squares mean of the i^{th} level is significantly different from the average, the corresponding vertical line crosses one of the decision limits, and vice versa. This analysis of means is discussed in more detail in Nelson (1985 & 1993), and briefly in Littell et al. (2006), and the SAS GLIMMIX procedure manual (2005).

5.3 SAS GLIMMIX Procedure

The SAS procedure PROC GLIMMIX that accommodates all features of GLMMs was issued in November 2005. Before then, GLIMMIX MACRO had been used. This procedure combines both PROC GENMOD and PROC MIXED procedures. With this procedure subject-specific (conditional) and population-averaged (marginal) inferences can be made. The estimation of the parameters using this procedure follows likelihood-based techniques and the default is the pseudo-likelihood following the procedures of Wolfinger and O'Connell (1993), and Breslow and Clayton (1993). For the construction of Wald test statistics and confidence intervals for the estimates it relies on Taylor-series expansion techniques. Wald-type tests and the estimated variance-covariance matrix are used for hypotheses tests for the fixed effects. The following are the primary assumptions for this procedure as outlined in the SAS GLIMMIX manual:

1. If the model contains random effects, the distribution of the responses conditional on the random effects is known. The distribution can either be a member of the exponential family of distributions or one of the supplementary distributions provided by the procedure itself. But for the fixed effects model, the unconditional (marginal) distribution is assumed to be known for maximum likelihood estimation, whilst in the case of quasi-likelihood estimation, the first two moments are

known.

2. The conditional expected value of the response takes the form of a linear mixed model after a monotonic transformation (link function) is applied.

3. The objective function for the optimization is a function of either the actual log-likelihood, an approximation to the log-likelihood, or the log-likelihood of an approximated model.

This procedure, like any other procedure, has strengths and weaknesses. The major drawback from which it suffers is of having a doubly iterative fitting algorithm and the absence of a true log-likelihood.

The conditional binary response model given the random PSU effects will be fitted where the marginal covariance matrix is block-diagonal and the observations from the same PSU form the blocks. The residual PL, a default estimation technique in SAS PROC GLIMMIX for fitting GLMMs, will be used. Refer to Littell et al. (1996) for the containment method which will be used to determine degrees of freedom. Furthermore, the Dual Quasi-Newton method, the default optimization technique for GLMMs, will be used where only covariance parameters will be participating in the optimization. The objective function will be computed through the residual likelihood technique. For more details on the methods and techniques discussed above see SAS GLIMMIX procedure manual (2005).

5.4 Results

The type 3 test of fixed effects for the fitted model is given in Table 5.1. The F-statistic, used for the significance test for the fixed effects, show that all the effects are important in the fitted model when tested at the 10% level of significance. Only one effect (i.e. ownership of dwelling) is not significant which registered a p-value greater than 0.1, but due to hierarchical principle for the model with interaction effects which are significant, the main effect is retained in the model. The minus twice the residual log pseudo-likelihood of the fitted model is 21473.17, and the generalized chi-square statistic is 4913.84. The ratio of the generalized chi-square statistic to its degrees of freedom, which is a measure of the residual variability in the marginal distribution of the data, is $\frac{4913.84}{4683} = 1.05$. This is because ϕ is 1. This measure can also be used as a rule of thumb which asserts that the fitted model is satisfactory if the ratio is 1 (Collett, 2003). The variance of the random PSU effect on the logit scale is estimated as $\hat{\sigma}_u^2$=0.02956 as given in Table 5.2. The same variance is obtained when the PSU is nested within 'location' and 'urban/rural'.

Table 5.1: Type 3 tests of fixed effects

	Num	**Den**		
Effect	**DF**	**DF**	**F Value**	**Pr > F**
Location	9	4683	8.96	<.0001
Sex	1	4683	8.22	0.0042
Mstatus	1	4683	5.38	0.0204
Sex*Mstatus	1	4683	17.88	<.0001
Age	1	4683	22.41	<.0001
Mstatus*Age	1	4683	11.02	0.0009
Education	2	4683	15.72	<.0001
Dwelling	1	4683	0.77	0.3804
Mstatus*Dwelling	1	4683	4.19	0.0408
Education*Dwelling	2	4683	6.98	0.0009
HHsize	1	4683	15.94	<.0001
Dwelling*HHsize	1	4683	4.48	0.0344

Table 5.3 gives solutions for the fixed effects. Note that standard errors, when containment method for degrees of freedom is used, are the same as those for the Satterthwaite-based method (refer to Tables 5.3 for containment method and C.2 for

Table 5.2: Covariance parameter estimates

Covariance Parameter Estimates

Cov Parm	**Subject**	**Estimate**	**Standard Error**
Intercept	PSU	0.02956	0.02311
Intercept	PSU(URBRUR*LOCATION)	0.02956	0.02311

Asymptotic Correlation Matrix of Covariance Parameter Estimates

Cov Parm	Subject	CovP1
Intercept	PSU	1.0000

the Satterthwaite-based method). When adjustment for uncertainty in estimating **G** and **R** is made (see Table C.1), the standard errors are also not (significantly) different from the two discussed above. Since the model fitted is a random intercept model, and 0.6015 is the overall intercept of the model which is adjusted by a fairly small random intercept estimate of 0.02956. Figures 5.2 to 5.13 summarize all pairwise comparisons of the least-squares means analysis performing all pairwise differences and an analysis of means with multiplicity adjustments. The Diffogram displays a line for each comparison and the axes of the plot represent the scale of the least-squares means. The confidence limit for the least-squares means difference is reflected by the length of the line, which is adjusted for the rotation and also possibly for multiplicity. The 45^0 line is referred to as reference line of the plot. The lines cross this line if two

Table 5.3: Solutions for fixed effects

		Standard		
Effect	**Estimate**	**Error**	**t Value**	**Pr > |t|**
Constant	0.6015	0.3171	1.90	0.0590
Butha-Buthe	-0.1966	0.1343	-1.46	0.1433
Leribe	0.5162	0.1359	3.80	0.0001
Berea	-0.00446	0.1354	-0.03	0.9737
Thaba-Tseka	-0.4606	0.1344	-3.43	0.0006
Mafeteng	0.05094	0.1338	0.38	0.7034
Mohale's Hoek	0.3509	0.1410	2.49	0.0128
Quthing	0.2947	0.1372	2.15	0.0318
Qacha's Nek	0.3741	0.1391	2.69	0.0072
Mokhotlong	-0.1807	0.1329	-1.36	0.1738
Sex	-0.6768	0.1133	-5.97	<.0001
Mstatus	-0.5940	0.2140	-2.78	0.0055
Sex*Mstatus	0.8068	0.1908	4.23	<.0001
Age	-0.6196	0.1263	-4.90	<.0001
Mstatus*Age	0.4991	0.1504	3.32	0.0009
Education1	0.8953	0.1735	5.16	<.0001
Education2	0.2429	0.1684	1.44	0.1493
Dwelling	0.05312	0.3395	0.16	0.8757
Mstatus*Dwelling	0.3432	0.1677	2.05	0.0408
Education1*Dwelling	-0.7870	0.2258	-3.49	0.0005
Education2*Dwelling	-0.3570	0.2306	-1.55	0.1217
HHsize	-0.8275	0.2604	-3.18	0.0015
Dwelling*HHsize	0.5696	0.2691	2.12	0.0344

compared least-squares means are not significantly different.

Figure 5.11 displays a different test about least-squares means. Here factor levels are compared against an overall average and not against each other. The dashed horizontal step plots in the analysis of the means graph represent the upper and lower decision limits determined at the 90th percentile (i.e. UDL, LDL). If the level is significantly different from the average, then the corresponding vertical line crosses the decision limit, either the lower or upper. See Littell et al. (2006) for more details on the least-squares means analysis.

5.4.1 Interpretation of the Results

Coefficients for the fixed effects are interpreted the same way as in the ordinary logistic regression model, given in Section 4.1.4. Estimates in Table 5.3 are slightly lower than those given in Table 4.4, Section 4.1.4, due to the shrinkage of estimates when random effect of PSUs is accounted for. But the conclusions do not change, hence interpretation of the model coefficients will not explicitly be done here. Instead, another form of presentation based on least-squares means analysis is considered both tabularly and graphically. See the syntax in Appendix C used to perform this analysis. All the contrasts are on the logit scale $log(\hat{\pi}_i/(1-\hat{\pi}_i))$.

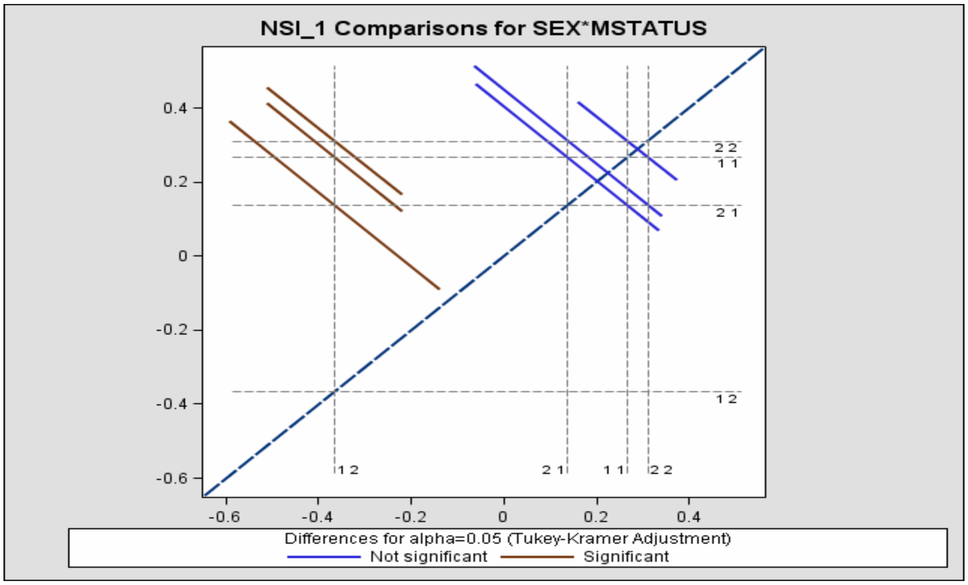


Figure 5.2: Diffogram for sex by marital status interaction effect: 11=unmarried males, 12=married males, 21=unmarried females, and 22=married females

Figure 5.2 illustrates adjusted comparison of sex by marital status interaction least-squares means for multiplicity. The lines that represent the significant difference between the least-squares means of the levels of the sex by marital status interaction

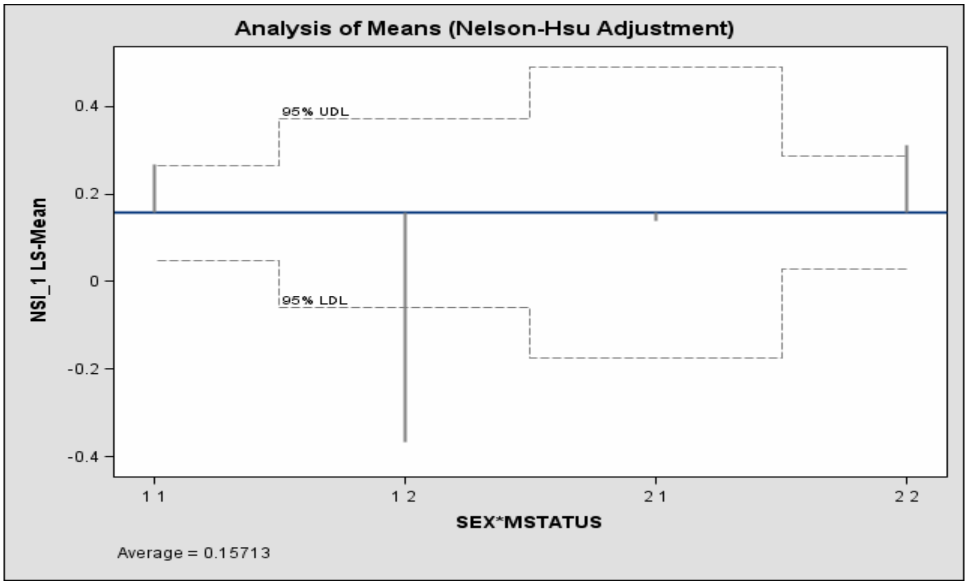


Figure 5.3: Analysis of means for sex by marital status interaction effect: 11=unmarried males, 12=married males, 21=unmarried females, and 22=married females

effect are the ones centered at the intersections of the lines 12-married males and 21-unmarried females, married males and 11-unmarried males, and lastly married males and 22-married females. Notice that the least-squares means difference given by the intersection of lines corresponding to married male and unmarried female heads has the widest confidence limits. According to this figure the prevalence of disease/injury on average for the households headed by married males is different from that which prevailed for those headed by unmarried females, unmarried males and married females. The lines that cross the 45^0 line show that the prevalence of disease/injury is not significant between corresponding categories. That is given by the households headed by unmarried females compared to those headed by unmarried males and married females as well as the comparison of the households headed by unmarried males and those headed by married females.

The average of sex by marital status interaction effect (on logit scale) is 0.15713 as given by Figure 5.3. From this figure one can see that the differences of means of all levels except that of unmarried females are significantly different from the average given by the vertical lines that cross the 90% LDL. This gives more insight into the reason why the difference depicted in Figure 5.2 is significant between the households headed by married males and those headed by other groups (i.e. unmarried females, unmarried males, and married females). This is given by a negative difference of households headed by married males effect from the average, whilst for other categories it is positive.

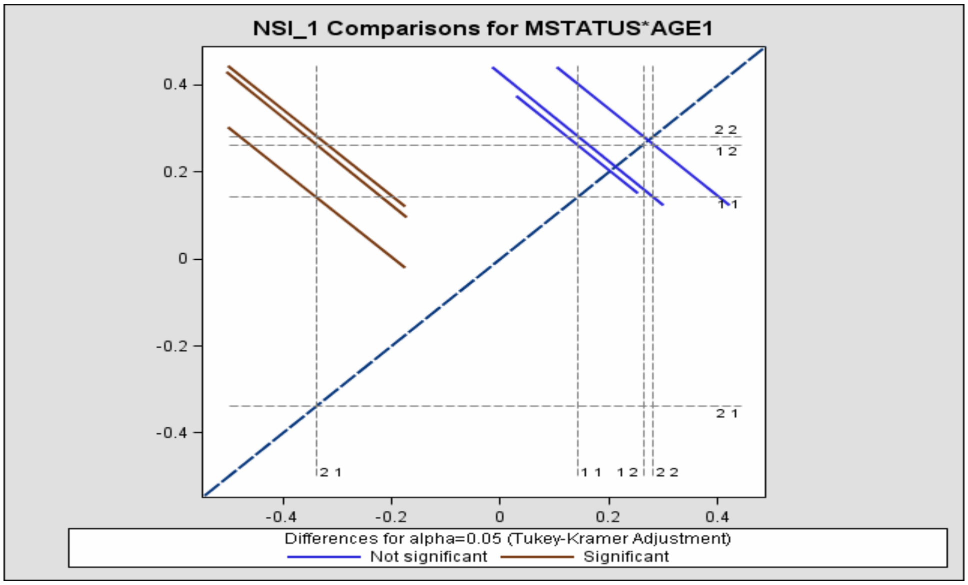


Figure 5.4: Diffogram for marital status by age interaction effect: 11=young unmarried, 12=older unmarried, 21=young married, and 22=older married

Moreover, Figure 5.4 shows that lines centered at the intersections 21-young married heads and 11-young unmarried heads, young married heads and 12-older unmarried heads, and older unmarried heads and 22-older married heads, represent significant differences of least-squares means of the levels of marital status by age interaction effects. The above difference of means suggests that the prevalence of disease/injury in the households headed by young married people is significantly different from that in households headed by young unmarried people, older unmarried people, and older married people. Recall that ‘young’ refers to people less that 40 years of age, and

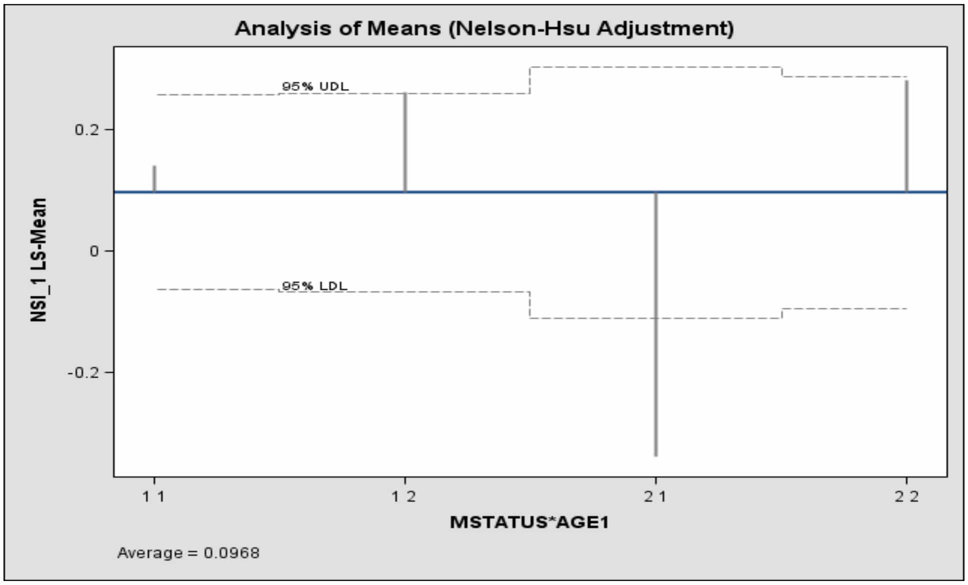


Figure 5.5: Analysis of means for marital status by age interaction effect: 11=young unmarried, 12=older unmarried, 21=young married, and 22=older married

'older' to people 40 years and above. The prevalence of disease/injury is not significantly different in the following groups: young unmarried and older unmarried; young unmarried and older married; and older unmarried and older married.

Furthermore, a clear significant difference of young married least-squares mean from the average is depicted in Figure 5.5 and Table C.3. The least-squares mean for young married people below the average and the least-squares means for the other levels above the average show why significant differences between least-squares mean for the young married people and least-squares means for other levels were observed in Figure 5.4. The values are given in Table C.3, where the least-squares mean estimate of young married heads is -0.3381, and the estimate of its difference from average is -0.4349. The difference of older unmarried heads least-squares mean from the average is significant.

For marital status and ownership of dwelling interaction effect, the observed significant difference of its levels is given by the line centered at the intersection 21-married people who do not own dwelling and 11-unmarried people who do not own dwelling.

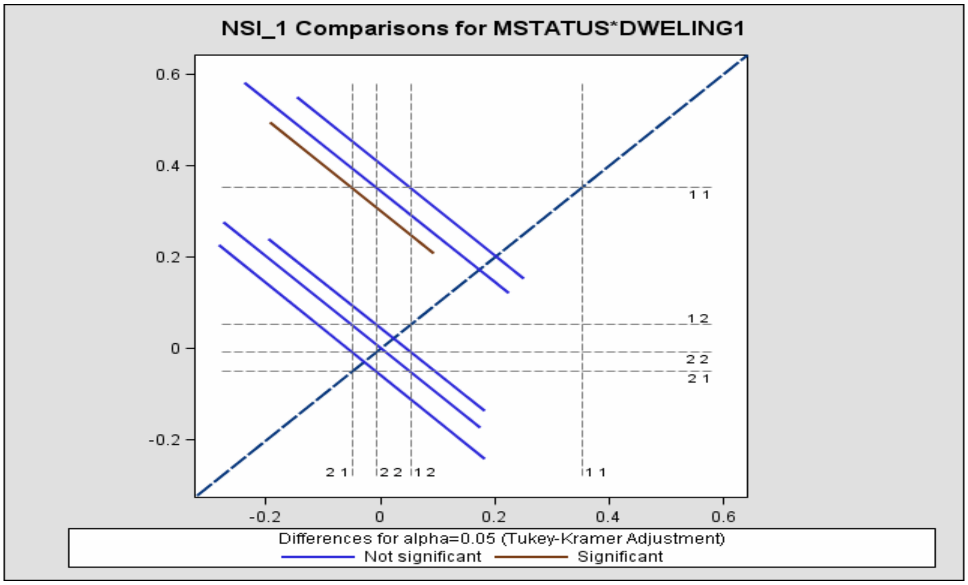


Figure 5.6: Diffogram for marital status by ownership of dwelling interaction effect: 11=unmarried people who do not own dwelling, 12=unmarried people who own dwelling, 21=married people who do not own dwelling, and 22=married people who own dwelling

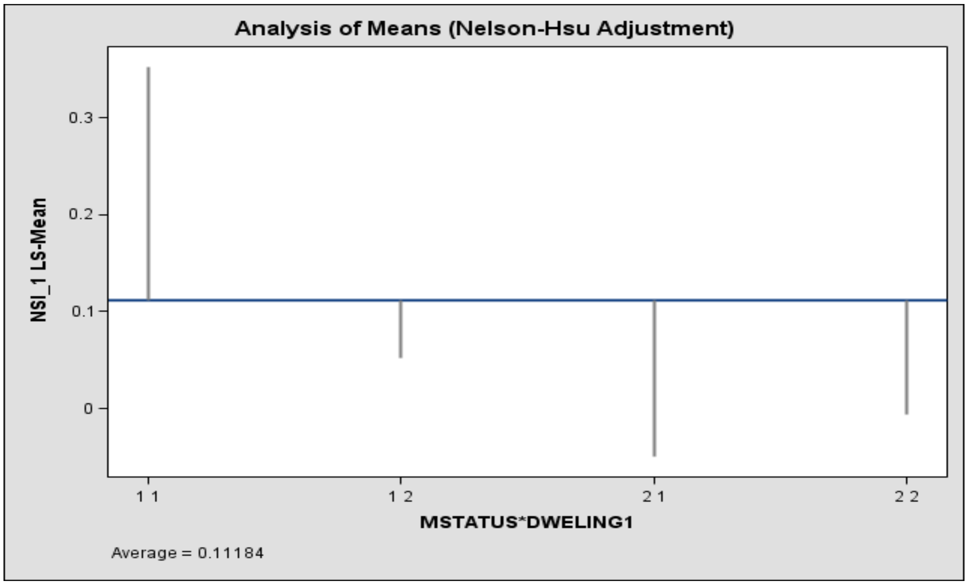


Figure 5.7: Analysis of means for marital status by ownership of dwelling interaction effect: 11=unmarried people who do not own dwelling, 12=unmarried people who own dwelling, 21=married people who do not own dwelling, and 22=married people who own dwelling

These are given in Figure 5.6 and Table C.4. The difference is not significant between all other levels i.e. 21-households headed by married people who do not own dwelling and 22-those headed by married people who own dwelling; households headed by married people who do not own dwelling and 12-those headed by unmarried people who own dwelling; households headed by married people who own dwelling and those headed by unmarried people who own dwelling; households headed by married people who own dwelling and those headed by unmarried people who do not own dwelling; households headed by unmarried people who own dwelling and those headed by unmarried people who own dwelling. Since adjustment could not be completed for this effect, decision limits were also not constructed. See the last column 'adj P' in Table C.4 which does not have values, and Figure 5.7 which does not have decision limits. But when using unadjusted value for inference, it can be seen that differences of unmarried people who do not own dwelling and married people who do not own dwelling levels from the average are significant. Least-squares mean for unmarried people who do not own dwelling is above the average and least-squares mean for married people who do not own dwelling is the furthest below the average.

There are four significantly different pairwise comparisons of least-squares means of levels of education by ownership of dwelling interaction effect on the health status of the households, which are represented by the lines centered at the intersections 32-households headed by people who completed secondary and own dwelling and 11-those who have some primary or no education and do not own dwelling; households headed by people who completed secondary and own dwelling and 12-those who have some primary or no education and own dwelling; 22-households headed by people who completed primary and own dwelling and those who have some primary or no education and own dwelling; as well as 21-households headed by people who completed primary and do not own dwelling and those who have some primary or no education and own

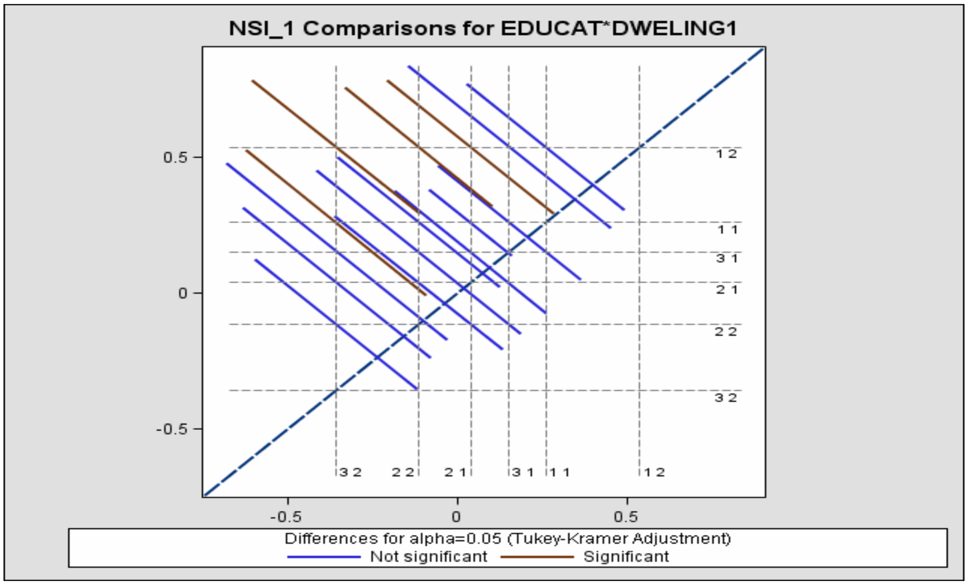


Figure 5.8: Diffogram for education by ownership of dwelling interaction effect: 11=no education and do not own dwelling, 12=no education and own dwelling, 21=completed primary and do not own dwelling, 22=completed primary and own dwelling, 31=completed secondary and do not own dwelling, and 32=completed secondary and own dwelling

dwelling. This is given in Figure 5.8. It indicates that the prevalence of disease/injury in the households owning dwelling and headed by educated (completed secondary) people is on average significantly different from that in households that do not own dwelling and are headed by uneducated (have some primary or no education) people. This significance is also observed in the following comparison groups: the households that own dwelling and are headed by people who completed secondary education and those that own dwelling but are headed by uneducated people; the households that own dwelling and are headed by uneducated people versus those that own dwelling and are headed by people who completed primary education; and the households that own dwelling and are headed by uneducated people versus those that do not own dwelling and are headed by people who completed primary.

In addition, the three vertical lines in Figure 5.9 corresponding to household heads

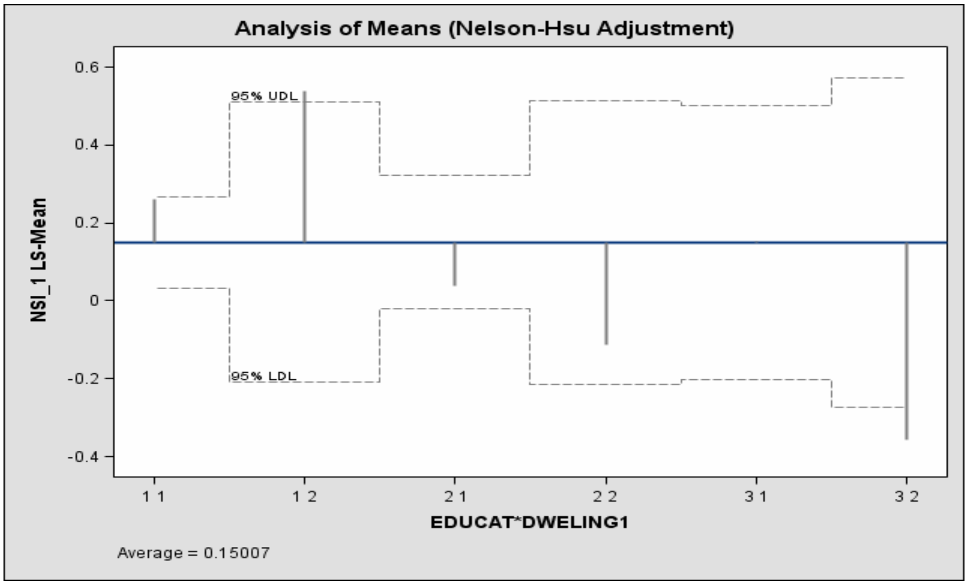


Figure 5.9: Analysis of means for education by ownership of dwelling interaction effect: 11=no education and do not own dwelling, 12=no education and own dwelling, 21=completed primary and do not own dwelling, 22=completed primary and own dwelling, 31=completed secondary and do not own dwelling, and 32=completed secondary and own dwelling

who have no education and own dwelling; household heads who have no education and do not own dwelling; and those who completed secondary and own dwelling, cross the decision limits. Household heads who have no education and own dwelling, and those who completed secondary and own dwelling are the most extreme, where one is above the average and the other below average. The significant difference of least-squares means of the two levels from the average is an indication of how important education and ownership of dwelling are for the well-being of the people. This importance is also stressed by Ulukanligil and Seyrek (2004) who assert that education should be the first thing to be done in any health programme aimed at improving the socio-economic development level of the community.

Two of the pairwise comparison of least-squares means of levels of ownership of dwelling by households size interaction effect on health status of the households are not

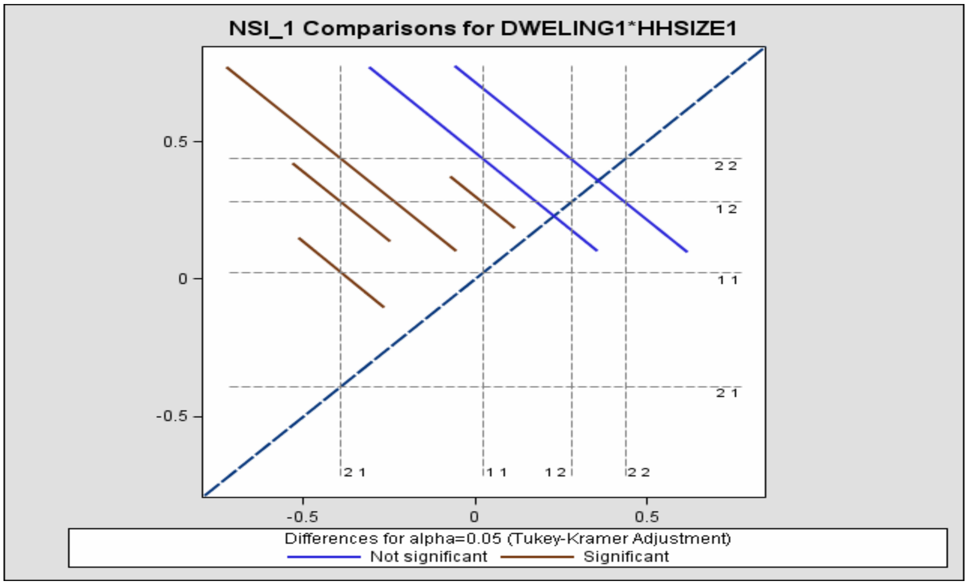


Figure 5.10: Diffogram for ownership of dwelling by household size interaction effect: 11=large household not owning dwelling, 12=small household not owning dwelling, 21=large household owning dwelling, and 22=small household owning dwelling

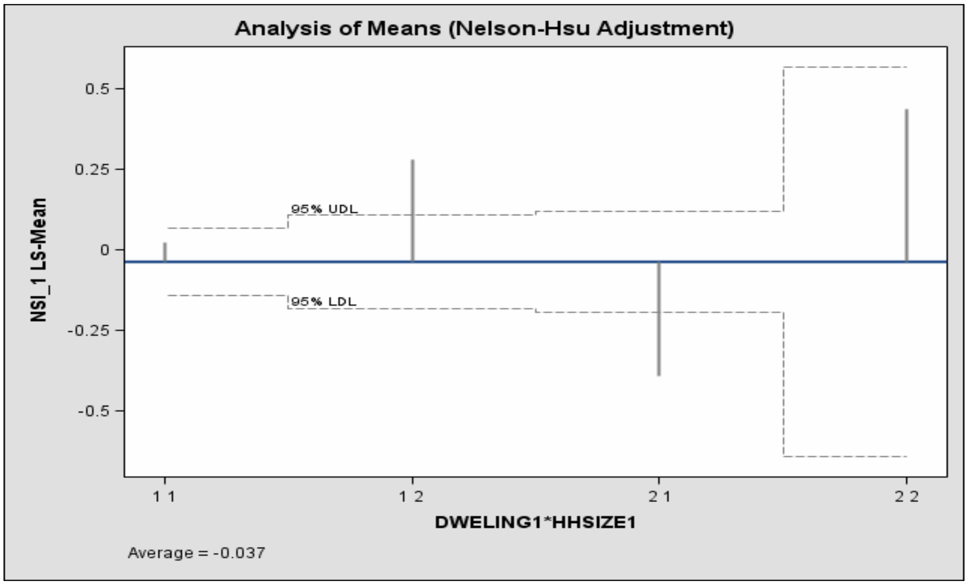


Figure 5.11: Analysis of means for ownership of dwelling by household size interaction effect: 11=large household not owning dwelling, 12=small household not owning dwelling, 21=large household owning dwelling, and 22=small household owning dwelling

significantly different, which are represented by the lines centered at the intersections 11-large households that do not own dwelling and 22-small ones that own dwelling, as well as 12-small households that do not own dwelling and small ones that own dwelling given in Figure 5.10. The significant difference is observed in the following contrasts: large households that own dwelling versus large ones that do not own dwelling; small households that do not own dwelling and small ones that own dwelling; and large households that do not own dwelling versus small ones that do not own dwelling.

Likewise, Figure 5.11 shows two extreme vertical lines that cross the decision limits corresponding to 12-small households that do not own dwelling and 21-large households that own dwelling. This also reflects how important ownership of dwelling and the family size is for the well-being of people.

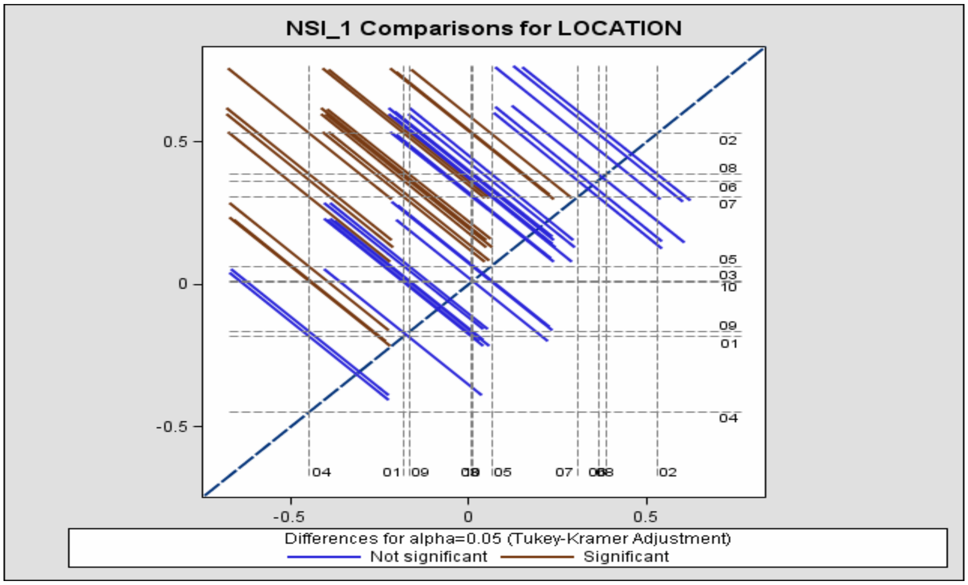


Figure 5.12: Diffogram for location effect: 01=Butha-Buthe, 02=Leribe, 03=Berea, 04=Thaba-Tseka, 05=Mafeteng, 06=Mohale's Hoek, 07=Quthing, 08=Qacha's Neck, 09=Mokhotlong, and 10=Maseru

Figure 5.12 portrays adjusted comparison of location least-squares means for multiplicity. Notice that lines centered at the intersections of locations 04-Thaba-Tseka and

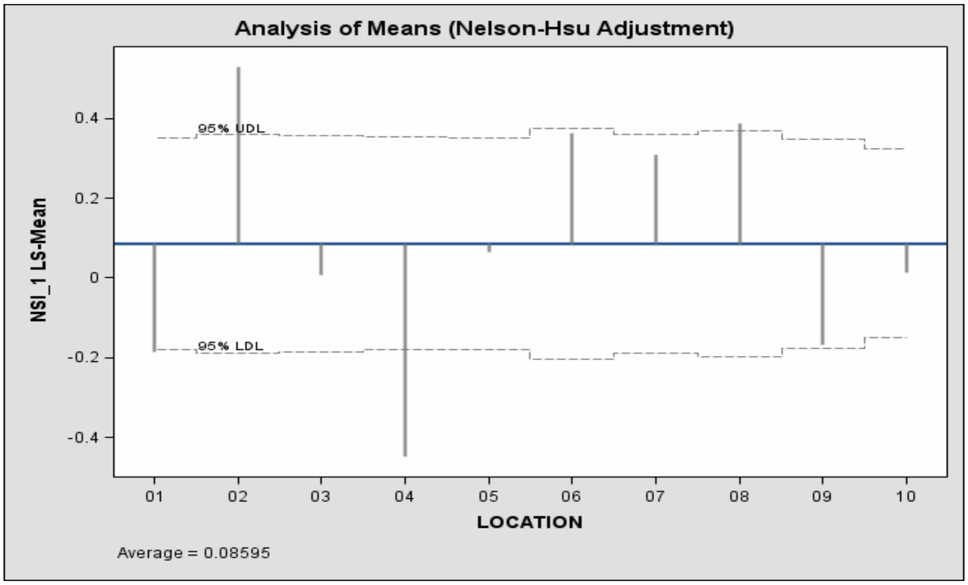


Figure 5.13: Analysis of means for location effect: 01=Butha-Buthe, 02=Leribe, 03=Berea, 04=Thaba-Tseka, 05=Mafeteng, 06=Mohale's Hoek, 07=Quthing, 08=Qacha's Neck, 09=Mokhotlong, and 10=Maseru

10-Maseru, Thaba-Tseka and 03-Berea, Thaba-Tseka and 05-Mafeteng, Thaba-Tseka and 07-Quthing, Thaba-Tseka and 06-Mohale's Hoek, Thaba-Tseka and 08-Qacha's Nek, Thaba-Tseka and 02-Leribe, 01-Butha-Buthe and Quthing, 09-Mokhotlong and Quthing, Butha-Buthe and Mohale's Hoek, Butha-Buthe and Qacha's Nek, Mokhotlong and Mohale's Hoek, Mokhotlong and Qacha's Nek, Butha-Buthe and Leribe, Mokhotlong and Leribe, Qacha's Nek and Leribe, and Mafeteng and Leribe are significantly different. This means that the incidence of disease/injury in Thaba-Tseka was significantly different from that in Maseru, Berea, Leribe, Mafeteng, Mohale's Hoek, Quthing, and Qacha's Nek. It is not significantly different from that in Mokhotlong and Butha-Buthe. The incidence experienced in Butha-Buthe was significantly different from that experienced in only three districts, namely Leribe, Mohale's Hoek, and Quthing. The incidence in these three districts was also different from that observed in Mokhotlong. In Mafeteng and Qacha's Nek the incidence appeared to be different from that in Leribe. The pairwise comparison of other districts other than the ones mentioned above do not have statistically significant differences.

The average location effect (on logit scale) is 0.08595, as given in Figure 5.13. Notice that vertical lines corresponding to Butha-Buthe, Leribe, Thaba-Tseka, Mohale's Hoek, Qacha's Nek, and Mokhotlong cross the decision limits, implying that they are significantly different from the average. The averages for Leribe and Thaba-Tseka are the most extreme on the opposite sides of the average. That means the least-squares mean for Leribe is greater than the average and the one for Thaba-Tseka is less than the average.

Estimates of these least-squares means can also be presented in tabular form, given in Tables C.3 to C.6 in the appendix.

Chapter 6

Survey Logistic Regression Models

6.1 Introduction

Logistic regression models used to analyse data from the complex sampling designs will be called survey logistic regression models in this study, to distinguish between them and the ordinary logistic regression models discussed in Chapter 2. Survey logistic regression models have the same theory as ordinary logistic regression models. The exception is that they account for the complexity of survey designs. When data are from simple random sampling, the survey logistic regression model and the ordinary logistic regression model are identical. In the present situation, PSUs are sampled in the first stage in each stratum (made up of districts and urban/rural). In the second stage households are sampled. So one specifies the response variable as y_{ijh} (i = 1,2,$\ldots, m_{hj}$; j = 1,2,$\ldots, n_h$; and h = 1,2,...,H) which equals 1 if disease/injury is present in i^{th} household within j^{th} PSU nested within h^{th} stratum, and 0 otherwise. Let π_{ijh} = p(y_{ijh} = 1) be the probability that the disease/injury is present in the i^{th} household within j^{th} PSU nested within h^{th} stratum. Thus the log-likelihood function in this case is given by

$$\ell(\boldsymbol{\beta};\mathbf{y}) = \sum_{h=1}^{H}\sum_{j=1}^{n_h}\sum_{i=1}^{m_{hj}}\left\{y_{ijh}\log\left(\frac{\pi_{ijh}}{1-\pi_{ijh}}\right) - \log\left(\frac{1}{1-\pi_{ijh}}\right)\right\} \tag{6.1}$$

and the survey logistic regression model is given by

$$\text{logit}(\pi_{ijh}) = \mathbf{x}'_{ijh}\boldsymbol{\beta}\ , \quad i = 1, 2, \ldots, m_{hj};\ j = 1, 2, \ldots, n_h;\ \textit{and}\ h = 1, 2, \ldots, H \tag{6.2}$$

where $\mathbf{x}_{ijh}$ is the row of the design matrix corresponding to the characteristics of the i^{th} household in the j^{th} PSU within h^{th} stratum, and $\boldsymbol{\beta}$ is a vector of unknown parameters of the model. If all design variables are included in the model as explanatory variables, the inference about the effects of the factors in the fitted model will be reliable (Pfeffermann, 1993).

6.2 Estimation of Parameters

Refer to Chapter 2 for discussion of method of maximum likelihood estimation used to estimate parameters of the model. Calculation of the standard errors of the parameter estimates, which are used to perform appropriate statistical tests on and construct confidence intervals for the parameters, when data come from complex design is complicated. The covariance matrix of parameter estimates is obtained through the Taylor expansion approximation procedure (Vittinghoff et al., 2005). This technique estimates variance from the variation among clusters and computes the overall variance estimate by pooling stratum variance estimates together. The discussion of this approximation is given in Chambless and Boyle (1983). There are other methods of variance estimation for complex survey data other than the Taylor expansion approximation (also known as linearisation method). These methods are called sample re-use methods. These are jacknife, sample replication, balanced repeated replication (BRR), and the bootstrap methods (see Vittinghoff, 2005); Lehtonen and Pahkinen, 1995; and Skinner, Holt, and

Smith, 1989). The jacknife, BRR and bootstrap methods are illustrated with examples in Lehtonen and Pahkinen (1995). But only the Taylor expansion approximation will be used here.

The degrees of freedom for the t-test statistics used for testing the significance of the parameters equals the number of clusters minus the number of strata in the sample survey design. This statistic can then be used to construct confidence intervals for the parameters, especially if n (the overall sample size) is small. When n is large, as is the case for the CWIQ data, the sampling distribution of the parameter estimators are approximated by a normal distribution. Hence, the Wald chi-square statistic can also be used to test for the significance of the parameters and to construct their confidence intervals (which are also called normal confidence intervals) given by

$$\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \sqrt{v_{jj}} \tag{6.3}$$

where $z_{1-\frac{\alpha}{2}}$ is the 100(1-$\frac{\alpha}{2}$)th percentile of the standard normal distribution, and v_{jj} is a variance of $\hat{\beta}_j$ given by diagonal elements of variance-covariance matrix of $\hat{\boldsymbol{\beta}}$. Note again that these intervals are on a logit scale, if the canonical link function is used. Untransformed confidence intervals are given by

$$exp(\beta_j \pm z_{1-\frac{\alpha}{2}} \sqrt{v_{jj}}). \tag{6.4}$$

Fortunately, the trouble of calculating estimates and their variance has been circumvented by implementation of the procedures in statistical packages that appropriately account for the complexity of survey designs. This procedure is implemented in packages such as SAS 9.1 and is called PROC SURVEYLOGISTIC. It was developed basically for fitting a linear logistic regression model for discrete response variables to survey data. When the data are from the simple random sampling method, PROC SURVEYLOGISTIC is identical to PROC LOGISTIC. PROC SURVEYLOGISTIC uses maximum likelihood estimation method and the Taylor expansion approximations mentioned above. This procedure will be used to fit the model (6.2). This procedure requires that for any cluster to be included in the calculation there should be at least two or more clusters in the stratum, otherwise the stratum will not make any contribution.

6.3 Model Selection

The same selection procedure discussed in Section 2.5 applies for survey logistic regression models. However, the selection procedures (i.e. forward, backward, and stepwise) are not yet included in PROC SURVEYLOGISTIC. The alternative is to start with the saturated model and observe the contribution of each effect to deviance reduction given by type 3 analysis of effects, then exclude one variable with insignificant effect (one at a time) and observe the contribution of the remaining effects to deviance reduction. Continue this process until the model has only significant effects.

Alternatively, the following criteria can be used to compare the goodness-of-fit of two nested models: The Akaike's information criterion (AIC) introduced by Akaike (1974), and the Schwarz Criterion (SC) (also known as Bayesian Information criterion (BIC)) introduced by Schwarz (1978). These methods are used to adjust (or impose stiffer penalties on) the likelihood ratio statistic -2logL which measures the deviation of the log-likelihood of the fitted model from the log-likelihood of the maximal possible model (Vittinghoff et al., 2005). The adjustment is necessary because the -2logL will always decrease as a new explanatory variable enters the model even if it is insignificant. The AIC is given by

$$AIC = -2logL + 2p \tag{6.5}$$

where p is the number of parameters in the model. This technique which tolerates violation of parametric model assumptions, can be used to compare multiple nested models, and it does not rely entirely on p-values for determining significance of explanatory variables (Alexander, Logan, and Paquet, 2006). Another criterion which adjusts the -2logL statistic for the number of parameters is SC given by

$$SC = -2logL + p\log(n) \tag{6.6}$$

where p is as explained above and n is the overall sample size. The smaller the value of the criteria, the better the goodness-of-fit of the model (Anderson, Burnham, and White, 2006; Caley and Home, 2002; and Buckland, Burnham, and Augustin, 1997). The model selected in Chapter 4 will be refitted accounting for the complexity of the survey design and will be compared with the one that will be selected through the cumbersome procedure proposed in this section.

6.4 Model Checking

6.4.1 Model Fit Test

The AIC and SC criteria will be used to test for the goodness-of-fit of the model. Since the criteria involve -2logL which is only used for variable selection in the case of ungrouped binary data, they are used as approximations. The Hosmer-Lemeshow goodness-of-fit statistic is used in the case of ungrouped binary data, is not yet implemented in the PROC SURVEYLOGISTIC.

6.4.2 Predictive Accuracy/Ability of the Model

The PROC SURVEYLOGISTIC, like other procedures used for fitting binary response models to data, produces statistics on the prediction ability of the model, such as

c, Sommer's D (SD), Goodman-Kruskal Gamma (GKG), and Kendall's Tau-a (KT). Following the SAS notation, these statistics are given as

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

where n is the total number of individuals in the data set, t is a total number of pairs given by $n(n-1)/2$, n_c is a number of concordant pairs (a pair of observations is concordant if a response y is 1 and the predicted probability is high), n_d is a number of discordant pairs (a pair of observations is discordant if the response y is 1 and the predicted probability is low), and tied pairs are given by '$t - n_c - n_d$'. See Agresti (1984) for more details. The widely employed statistic is c which is equal to area under the receiver operating characteristic (ROC) curve in the case of binary response models. Recall that the prediction accuracy is poor if c is between 0.5 and 0.6, moderate if between 0.6 and 0.7, acceptable if between 0.7 and 0.8 and excellent if greater than 0.8.

6.5 Results

6.5.1 Introduction

To begin, the model selected by the PROC LOGISTIC was refitted using the PROC SURVEYLOGISTIC to see if estimates would change when the complexity of the survey design is accounted for. Another model was selected using the PROC SURVEYLOGISTIC and compared to the former for goodness-of-fit. Table 6.3 compares estimates from ordinary logistic regression model and survey logistic regression model with the logit link function. Notice that estimated coefficients are the same from both procedures, but standard errors produced by PROC LOGISTIC are relatively small compared to those from the PROC SURVEYLOGISTIC. That means, when complexity of the survey design is ignored by invoking the procedure that assumes SRS, the variances are underestimated hence leading to inaccurate inferences. Again, not only the magnitude of effect is the same in both models, but also the direction of effect is the same.

6.5.2 Model Selection

Table 6.3 is for comparing estimates of the same model fitted using the two logistic procedures discussed above. Another candidate model supported by PROC SURVEYLOGISTIC is investigated. The largest model with significant effects is given in Tables 6.1 and was obtained through the PROC SURVEYLOGISTIC. This model has the smallest deviance (-2logL) amongst all the nested models with the first order interaction effects. The model selected in Chapter 4 was refitted using this procedure and is given in Table 6.3. Table 6.2 gives deviance analysis. The total deviance reduction for the model in Table 6.3 is 375.3615, with 22 degrees of freedom, and for the one in Table 6.1 is 421.8850, with 36 degrees of freedom which is very significant (with p-value < 0.0001) in both models. The AIC for the model in Table 6.3 is 6456.465 which is larger, by 18.524, than the one for the model in Table 6.1. On the other hand, the SC for the model in Table 6.3 is 6606.147 which is small compared to 6678.735 for the model in Table 6.1. Regarding suitability, the AIC suggests the model in Table 6.1, whilst SC suggests the model in Table 6.3. If the interest is in the order of the model, a model based SC is preferred; but if the interest is in consistent approximation and model fit, a model based on AIC is preferred (Buckland et al., 1997). This means, dimensionally, the model in Table 6.3 is the best, and when ignoring the order, the model in Table 6.1 is chosen over the model in Table 6.3. Rust et al. (1995) recommend SC for model comparison, selection, and probability estimation because of its simplicity and prediction accuracy which outperforms other criteria in terms of accuracy and consistency. The c statistic given in the same table suggests that both models have moderate prediction ability. Intuitively, which model to choose is based upon the research objectives and the most appealing model. Since the most parsimonious model is preferred, the model in Table 6.3 advocated by SC which is recommended by Rust et al. (1995) is chosen.

Table 6.1: Type 3 analysis of effects for model 2 using the PROC SURVEYLOGISTIC

Effect	DF	Wald Chi-Square	p-value
Location	9	89.6662	<.0001
Sex	1	0.1459	0.7025
Mstatus	1	12.8318	0.0003
Sex*Mstatus	1	15.9878	<.0001
Age	1	19.1262	<.0001
Mstatus*Age	1	5.4434	0.0196
Education	2	15.4142	0.0004
Dwelling	1	1.9525	0.1623
Education*Dwelling	15	15.8088	0.0004
Toilet	2	6.5768	0.0373
Education*Toilet	4	10.8161	0.0287
Fuel	2	12.5931	0.0018
TClinic	1	7.8929	0.0050
Fuel*TClinic	2	7.6423	0.0219
HHsize	1	4.4999	0.0339
Sex*HHsize	1	12.4799	0.0004
Mstatus*HHsize	1	9.9477	0.0016
Dwelling*HHsize	1	2.7987	0.0943
Toilet*HHsize	2	7.3706	0.0251

Table 6.2: Model fit statistics using the PROC SURVEYLOGISTIC

Criterion	Intercept only	Model 1†	Model 2‡
-2LogL	6785.826	6410.465	6363.941
AIC	6787.826	6456.465	6437.941
SC	6794.334	6606.147	6678.735
c		0.652	0.665

Note: † is given in Table 6.3, and ‡ is given in Table 6.1

6.5.3 Interpretation of Parameters

Since point estimates produced by the PROC SURVEYLOGISTIC are the same as those given by PROC LOGISTIC or PROC GENMOD, interpretation given in Section 4.1.4 applies here. The only difference is the confidence intervals for the coefficients, which are narrow for the ordinary logistic regression model due to underestimated standard errors of the coefficients in the ordinary logistic regression model.

It can be seen that the effects of Butha-Buthe, Berea, Mafeteng, and Mokhotlong are not significant, which means that, controlling for other variables, the incidence of disease/injury in these locations are not different from the one in Maseru (the reference location). The parameter for Leribe indicates that households in Leribe are 1.672 (between 1.296 and 2.157) times more likely to be ill/injured compared to the ones in Maseru. The rate is a little lower for the households in Mohales' Hoek, Quthing, and Qachas' Nek which are 1.418 (between 1.158 and 1.738), 1.339 (between 1.071 and 1.679) and 1.450 (between 1.133 and 1.858) times more likely to be ill/injured, respectively compared to the ones in Maseru. The households in Thaba-Tseka are 0.63 (between 0.510 and 0.779) times more likely to be ill/injured than those in Maseru (i.e. people in Thaba-Tseka are 0.37 less likely).

Table 6.5 summarizes the coefficient of the interaction terms. The positive sign of 'sex' and 'marital status' interaction effect indicates that households headed by un-

Table 6.3: Comparison of PROC LOGISTIC and PROC SURVEYLOGISTIC for fitting model 1

Effect	Proc logistic		Proc surveylogistic	
	Estimate	Std errors	Estimate	Std errors
Constant	0.5978***	0.3138	0.5978***	0.3273
Butha-Buthe	-0.1978	0.1262	-0.1978	0.1461
Leribe	0.5140*	0.1281	0.5140*	0.1548
Berea	-0.0057	0.1274	-0.0057	0.1413
Thaba-Tseka	-0.4619*	0.1264	-0.4619*	0.1286
Mafeteng	0.0492	0.1257	0.0492	0.1428
Mohale's Hoek	0.3494*	0.1332	0.3494*	0.1235
Quthing	0.2922**	0.1293	0.2922**	0.1362
Qachas'Nek	0.3719*	0.1313	0.3719**	0.1504
Mokhotlong	-0.1816	0.1248	-0.1816	0.1423
Sex	-0.6770*	0.1129	-0.6770*	0.1180
Mstatus	-0.5843*	0.2131	-0.5843*	0.2199
Sex*Mstatus	0.8022*	0.1901	0.8022*	0.1861
Age	-0.6166*	0.1258	-0.6166*	0.1151
Mstatus*Age	0.4932*	0.1498	0.4932*	0.1482
Education1	0.9009*	0.1722	0.9009*	0.1905
Education2	0.2449	0.1672	0.2449	0.1825
Dwelling	0.0606	0.3376	0.0606	0.3465
Mstatus*Dwelling	0.3365**	0.1671	0.3365**	0.1656
Education1*Dwelling	-0.7924*	0.2245	-0.7924*	0.2453
Education2*Dwelling	-0.3578	0.2294	-0.3578	0.2343
HHsize	-0.8305*	0.2592	-0.8305*	0.2385
Dwelling*HHsize	0.5717**	0.2679	0.5717**	0.2490

Note: * denotes significance at 1%, ** at 5%, and *** at 10%

married males are more likely to be unhealthy compared to those headed by unmarried females and that households headed by married males are less likely to be unhealthy compared to their unmarried counterparts. This is given by the ratio of log odds ratio of 0.8022 with the 90% confidence interval (between 0.496 and 1.108). The ratio of odds ratios is 2.230 (between 1.642 and 3.029). The coefficient for 'age' and 'marital status' interaction effect indicates that households headed by young unmarried people

Table 6.4: Odds ratios for location

Effect	Point Estimate	90% Confidence Interval	
		Lower	Upper
Butha-Buthe	0.821	0.645	1.040
Leribe	1.672	1.296	2.157
Berea	0.994	0.788	1.255
Thaba-Tseka	0.630	0.510	0.779
Mafeteng	1.050	0.831	1.329
Mohale's Hoek	1.418	1.158	1.738
Quthing	1.339	1.071	1.679
Qachas'Nek	1.450	1.133	1.858
Mokhotlong	0.834	0.660	1.054

Table 6.5: Odds ratios for interaction terms of sex, marital status, age, education, ownership of dwelling and household size

Effect	Ratio of log odds ratios	90% Confidence Interval		Ratio of odds ratios	90% Confidence Interval	
		Lower	Upper		Lower	Upper
S*M	0.8022	0.4961	1.1083	2.2304	1.6422	3.0293
M*A	0.4932	0.2494	0.7370	1.6375	1.2833	2.0896
M*D	0.3365	0.0641	0.6089	1.4000	1.0662	1.8384
E1*D	-0.7924	-1.1959	-0.3889	0.4528	0.3024	0.6778
D*H	0.5717	0.1621	0.9813	1.7713	1.1760	2.6679

Note: S=sex, A=age, D=dwelling, E1=no education, H=hhsize, and M=mstatus

are 1.6375 (between 1.2833 and 2.0896) times more likely to be unhealthy than young married ones. Similarly, households which do not own their dwelling and headed by unmarried people are 1.4 times more likely to be unhealthy compared to their married counterparts; and those who do not own their dwelling and are headed by unmarried people are 1.4 (between 1.0662 and 1.8384) times more likely to be unhealthy than their counterparts who own their dwelling. Furthermore, the households that do not own their dwelling and are headed by people with some primary or no education are 0.4528 (between 0.3024 and 0.6778) times more likely to be unhealthy compared to the ones which do not own their dwelling and are headed by people who completed secondary or have higher education. Similarly, 0.4528 suggests that households not owning their dwelling and headed by people with some primary or no education are less likely to be unhealthy compared to their counterparts who own their dwelling. Households without their own dwelling and having more than 5 members are 1.7713 (between 1.1760 and 2.6679) times more likely to have unhealthy members compared to their counterparts with 5 or fewer members and also more likely than their counterparts who own their dwelling. The large households which do not own their dwelling are more likely to be unhealthy compared to their counterparts in small households.

Recall that the model in Table 6.3 was chosen over the model in Table 6.1 on the basis of criteria discussed in Section 6.5.2. However, it may be informative if results of the model in Table 6.1 which identified more important factors for health status are interpreted. Table 6.6 displays results of this model in terms of odds ratios. Since the direction of effects of the factors identified by both models in Tables 6.3 and 6.6 are the same, the effects of the factors that appear in Table 6.6 (below the line) but not in Table 6.3 are interpreted. The odds ratios whose 90% C.I.s include 1 will not be interpreted because they are not significantly different from 1 at 10% significance level.

The results show that the households headed by uneducated people who use other types of toilets (other than flush or pit latrine) are 0.3286 (0.1518, 0.7113) times more likely to be unhealthy than their uneducated counterparts who use flush or pit latrine toilets, and more likely than those headed by educated people (who completed secondary education) who use other types of toilets. The incidence of disease/injury for the households headed by people who completed primary education and use other form of toilet (other than flush or pit latrine) is 0.2585 (0.1163, 0.5742) times that for the households headed by people who completed primary education and use flush or pit latrine toilets, and that for the households headed by people who completed secondary

Table 6.6: Odds ratios for model 2 in Table 6.1

		90% Confidence Interval	
Effect	Odds ratio	Lower	Upper
Constant	1.0535	0.4208	2.6377
Butha-Buthe	0.8406	0.6582	1.0738
Leribe	1.7255	1.3400	2.2218
Berea	1.0382	0.8168	1.3196
Thaba-Tseka	0.6590	0.5271	0.8240
Mafeteng	0.9847	0.7705	1.2583
Mohale's Hoek	1.4366	1.1681	1.7670
Quthing	1.3742	1.0921	1.7293
Qachas'Nek	1.5200	1.1742	1.9676
Mokhotlong	0.8538	0.6666	1.0937
Sex	1.1034	0.7222	1.6859
Mstatus	0.4024	0.2650	0.6113
Sex*Mstatus	2.1088	1.5515	2.8662
Age	0.6241	0.5227	0.7452
Mstatus*Age	1.3720	1.0978	1.7146
Education1	6.6240	2.9651	14.7981
Education2	4.4584	1.9788	10.0442
Dwelling	1.5417	0.9262	2.5664
Education1*Dwelling	0.4041	0.2670	0.6115
Education2*Dwelling	0.6635	0.4487	0.9809
HHsize	0.5434	0.3386	0.8720
Dwelling*HHsize	1.5204	1.0071	2.2958
Toilet1	1.3115	0.4562	3.7708
Toilet2	2.8137	1.2607	6.2789
Education1*Toilet1	0.5689	0.2005	1.6145
Education1*Toilet2	0.3286	0.1518	0.7113
Education2*Toilet1	0.3549	0.1174	1.0722
Education2*Toilet2	0.2585	0.1163	0.5742
Fuel1	0.5358	0.3982	0.7210
Fuel2	0.6147	0.4423	0.8544
Tclinic	0.4408	0.2729	0.7121
Fuel1*Tclinic	2.2140	1.3370	3.6664
Fuel2*Tclinic	1.8285	1.0892	3.0695
Sex*HHsize	0.4109	0.2716	0.6218
Mstatus*HHsize	2.1717	1.4493	3.2544
Toilet1*HHsize	1.2394	0.8783	1.7489
Toilet2*HHsize	0.8061	0.5692	1.1417

education and use other types of toilets.

In addition, the households which are far (which take 60 minutes or more to reach) from the hospital/clinic and use firewood/charcoal for cooking are 2.214 (1.337, 3.666) times more likely to be unhealthy than the households which are close (which less than 60 minutes to reach) to the hospital/clinic and use firewood/charcoal, and than the households which are far from the hospital/clinic and use kerosene/gas/electricity for cooking. But the households which are far from the hospital/clinic and use other type of fuel for cooking (other than firewood, charcoal, kerosene, gas, or electricity) are 1.8285 (1.0892, 3.0695) times more likely to be unhealthy than the households which are close to the hospital/clinic and use other type of fuel, and than the households which are far from the hospital/clinic and use kerosene/gas/electricity for cooking. The incidence of disease/injury for the large households headed by unmarried people is 2.1717 (1.4493, 3.2544) times that for the small households headed by unmarried people, and for large households headed by married people. The incidence for large households headed by males is 0.4109 (0.2716, 0.6218) times that for small households headed by females, and for large households headed by males.

Recall that the interaction effect of Education2*Dwelling was not significant in Table 6.3. But in Table 6.6 it is significant. Its odds ratio shows that households headed by people who completed primary education and do not own their dwelling are 0.6635 (0.4487, 0.9809) times more likely to be unhealthy compared to the households headed by their educated (completed secondary education) counterparts who own their dwelling, and compared to those headed by people who completed primary education and own their dwelling.

6.6 Shortcomings of the SURVEYLOGISTIC Procedure

Recall that for ungrouped binary data the likelihood ratio statistics cannot be used as a measure of goodness-of-fit, and hence the Hosmer-Lemeshow goodness-of-fit statistic is used instead. However, this statistic is not yet implemented in the SURVEYLOGISTIC procedure. Another drawback of this procedure is the absence of the 'output' option statement which facilitates further analysis of data, such as testing for appropriateness of the link function, outliers and influence detection. In the output about the model fit statistics the procedure provides the three above mentioned statistics. Therefore, the model is chosen through the use of the AIC and the SC criteria. Recall that both AIC and SC are statistics which introduce a penalty for a model having too many parameters. Since these statistics involve -2logL which is only used for variable selection in the case of ungrouped binary data, they are used as approximation for goodness-of-fit of the model.

Chapter 7

Conclusions

The objective of this study was to identify factors affecting the health status of the people of Lesotho. The identified factors will be used to guide policy and decision making to speed up the provision of a better life for all. Generalized linear models, generalized linear mixed models, and survey logistic regression models were used to identity these factors. To begin, a generalized linear model, called logistic regression model, which assumes simple random sampling was used. The highest second order interaction terms were allowed in the model. Due to the large number of variables, a stepwise selection procedure was adopted. District and the interaction terms sex by marital status, age by marital status, ownership of dwelling by marital status, education level by ownership of dwelling, ownership of dwelling by household size, and the main effects were significant except ownership of dwelling. However, due to the hierarchical principle of the models with interaction terms, ownership of dwelling was retained in the model. Model checks for goodness-of-fit, appropriateness of the link function, and influence were done, and all failed to reject the selected model. The selected model was refitted with the random PSUs effect incorporated which led to the generalized linear mixed model called the random intercept model. The survey logistic regression model that accounts for complexity of the design was also used to refit the model. These two models, which account for survey design, fitted the data well and the results from them given in Table 5.3 for generalized linear mixed model and Table 6.3 for survey logistic regression model lead to the same conclusions as the ones given by the generalized linear model in Table 6.3.

The incidence of disease/injury for the households headed by unmarried people who do not own their dwelling is higher than that for households headed by their counterparts who own their dwelling. A similar effect is observed for large households that do not own their dwelling versus those that own their dwelling. But for those headed by uneducated people who do not own their dwelling the incidence is low compared to their counterparts who own their dwelling. The disease/injury incidence for the households headed by uneducated people who do not own their dwelling is low compared to that of their educated counterparts. For the households headed by unmarried people who do not own their dwelling, the disease/injury incidence is higher than that for those headed by their married counterparts. A similar conclusion is drawn for unmarried males versus married males. Again, the disease/injury incidence is high for the large households that do not own their dwelling compared to the small households.

Moreover, the incidence of disease/injury for the households headed by young unmarried people is higher than that for older unmarried heads. The incidence is also found to be high for unmarried males compared to unmarried females. The districts in the southern part of Lesotho namely Mohale's Hoek, Quthing, and Qacha's Nek, and one in the northern part (i.e. Leribe) have a significantly higher incidence of disease/injury than that observed in Maseru. Only one district, Thaba-Tseka, has a significantly lower incidence compared to that observed in Maseru. The incidence observed in the other 4 districts is not significantly different from that observed in

Maseru.

The results in Table 6.6 suggest that the incidence of disease/injury for the households that do not own their dwelling and headed by people who completed primary education is low compared to that for the households that do not own their dwelling and headed by people who completed secondary education. The incidence is also low for the households that do not own their dwelling and headed by people who completed primary education compared to that for the households that own their dwelling and headed by people who completed primary education. This is the opposite of what one would expect, in the sense that the households headed by educated people who own their dwelling the incidence would be expected to be lower than for those headed by people with lower education level who do not own dwelling. The incidence of disease/injury for the households that use other types of toilets (other than flush/pit latrine) is low for the households headed by both uneducated people and people who completed primary education compared to that for the households headed by people who completed secondary education. Similarly, the incidence for the households headed by uneducated people and people who completed primary education is low for the households that use other types of toilets compared to those that use flush or pit latrine toilets.

The incidence of disease/injury for the households that use firewood/charcoal for cooking is high for the households that are far from the hospital/clinic (60 minutes or more walk away) compared to the households that are close (less than 60 minutes walk away) to the hospital/clinic. The incidence for the households that are far from the hospital/clinic is high for the households that use other type of fuel for cooking compared to the households that use kerosene/gas/electricity. A low incidence of disease/injury is also observed for the large households headed by males compared to that for large households headed by females. The results also show that the incidence of disease/injury is high for the large households headed by unmarried people compared to that for the small households headed by unmarried people, and for large ones headed by married people.

The findings of this study imply that the health status of the households is likely to improve: if household heads are married, especially males and those aged less than 40 years as well as those who do not own dwelling; if households own their dwelling, especially those that have more than 5 members, those headed by unmarried people, and those headed by people who completed secondary education; if household is not large so as to avoid problems of congestion and high dependency ratios i.e. households should comprise less than 6 members, particularly the households that do not own their dwelling; if households heads take good care of themselves so that they can be available for their households at mature age i.e. 40 years of age or more; if inequality in development that lead to unequal health facilities among districts is reduced by fast-tracking development in Leribe, Mohale's Hoek, Quthing, and Qachas Nek districts; if households headed by people who have no education or have completed primary education have basic toilet facilities (i.e. other form of toilet, other than flush/pit latrine); if the households which are far from the hospital/clinic use kerosene, gas, or electricity for cooking; if the hospitals or clinics are accessible to the households, more so to the households that use other types of fuel for cooking (other than kerosene, gas, or electricity); if large households are headed by males and also by married people; and if the households headed by females are small (i.e. have less 6 members).

This improvement could be achieved by creating an enabling environment for (1) the improvement of socio-economic development programmes, (2) the encouragement of owner-occupied dwelling, (3) well controlled household size (i.e. having a maximum of 5 members), (4) the improvement of awareness campaigns on health issues for the entire community (especially males and young heads), (5) the promotion of sustained marriage, (6) the improvement of hospitals/clinics accessibility to people, and (7) the financial empowerment of households to afford either kerosene, gas, or electricity.

The major limitation of the study is the data which could not allow analysis at the level of individual members of the household. The aggregated data do not capture all characteristics of each member of the household, such as education level. Individual member characteristics are likely to vary within and among households. Therefore, analysis at the individual level might give more insight into the diseases/injuries pattern than analysis at the aggregated (household) level.

There are avenues for further work on the subject. For instance, one could identity what were the major diseases/injuries contributing to poor health of the Basotho, especially those associated with factors found to be important for health status in this study. There are a number of ways in which this could be done. Each disease/injury (especially those that are chronic, acute, or highly prevalent) could be considered independently or separately and models for binary response analysis could be utilized. Alternatively, methods that coherently and systematically consider specific diseases/injuries or clusters of them in the analysis could be utilized. For example, multivariate models where a response has more than one binary component, each corresponding to a disease/injury category could be used (see Agresti and Liu, 1999). See also Knorr-Held and Best (2001) for shared component models used for simultaneous analysis of the spatial variation in two diseases. If the interest is to curb the burden of diseases/injuries, the focus could be on households with health problems that are chronic, acute, and widely prevalent.

Bibliography

Agresti, A. (1984). Analysis of ordinal categorical data. New York: Wiley.

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

Agresti, A. and Liu, I-M. (1999). Modeling a categorical variable allowing arbitrarily many category choices. *Biometrics*, **55(3)**, 936-943.

Akaike, H. (1974). A New Look at the Statistical Model Identification, IEEE Transactions on Automatic Control, **AC-19**, 716-723.

Alexander, S.M., Logan, T.B. and Paquet, P.C. (2006). Spatio-temporal co-occurrence of cougars (Felis concolor), wolves (Canis lupus) and their prey during winter: a comparison of two analytical methods. *Journal of Biogeography*, **33**, 2001-2012.

Anderson, Burnham, and White, (2006). AIC Model Selection in Overdispersed Capture-Recapture Data. *Ecology*, **76(6)**, 1780-1793.

Berlin, J.A., Kimmel, S.E., Ten Have, T.R. and Sammel, M.D. (1999). An empirical comparison of several clustered data approaches under confounding due to cluster effects in the analysis of complications of coronary angioplasty. *Biometrics*, **55(2)**, 470-476.

Boniface, D.R., Cottee, M.J., Neal, D. and Skinner, A. (2001). Social and demographic factors predictive of change over seven years in CHD-related behaviours in men aged 18 - 49years. *Journal of the Royal Institute of Public Health*, **115**, 246-252.

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

Buckland, S.T, Burnham, K.P. and Augustin, N.H. (1997). Model Selection: An Integral Part of Inference. *Biometrics*, **53(2)**, 603-618.

Caley, H. and Home, J. (2002). Estimating the force of infection; Mycobacterim bovis infection in feral ferrets mustela furo in New Zealand. *The Journal of Animal Ecology*, **71(1)**, 44-54.

Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, **96(455)**, 1022-1030.

Chambless, L.E. and Boyle, K.E. (1983). Logistic regression analysis for complex sample data. University of North Carolina, Chapel Hill, NC.

Christensen, O.F., Hobolth, A. and Jensen, J.L. (2005). Pseudo-likelihood analysis of codon substitution models with neighbor-dependent rates. *Journal of Computational Biology*, **12(9)**, 1166-1182.

Collett, D. (2003). Modelling Binary Data. 2nd edition, Chapman & Hall/CRC, New York.

Cooper, J.K. and Kohlmann, T. (2001). Factors associated with health status of older Americans. *Age and Ageing*, **30**, 495-501.

Dedman, D.J., Gunnell, D., Smith, G.D. and Frankel, S. (2001). Childhood housing conditions and later mortality in the Boyd Orr cohort. *Journal of Epidemioly and Community Health*, **55**, 10-15.

Der, G. and Everitt, S. (2002). A Handbook of Statistical Analysis: Using SAS. 2^{nd} edition, New York, Chapman and Hall/CRC.

Dobson, A.J. (1990). An Introduction to Generalized Linear Models. T.J. Press, Comwall.

Dobson, A.J. (2002). An Introduction to Generalized Linear Models. 2^{nd} edition, New York, Chapman and Hall/CRC.

Donald, A. and Donner, A. (1987). Analysis of data arising from a stratified design with the cluster as unit. *Statistics in Medicine*, **6**, 43-52.

Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998). Generalized linear models for small-area estimation. *Journal of the Ameriacn Statistical Association*, **93(441)**, 273-282.

Hammill, G.B. and Preisser, S.J. (2006). A SAS/IML software program for GEE and regression diagnostics. *Computational Statistics & Data Analysis*, **51(2)**, 1197-1212.

Hosmer, D.S. and Lemeshow, S. (1989). Applied Logistic Regression. New York: Wiley.

Howden-Chapman, P. (2004). Housing standards: a glossary of housing and health. *Journal of Epidemiology and Community Health*, **58**, 162-168.

Hsu, J. C. (1996). Multiple Comparisons: Theory and Methods. London: Chapman & Hall.

Hsu, C.J. and Peruggia, M. (1994). Graphical Representations of Tukey's Multiple Comparison Method. *Journal of Computational and Graphical Statistics*, **3(2)**, 143-161.

Huang, J.Z. (1998). Functional ANOVA models for generalized regression. *Journal of Multivariate Analysis*, **67**, 49-71.

Huebner, E.S., Valois, R.F., Suldo, S.M., Smith, L.C., Mcknight, C.G., Seligson, J.L. and Zullig, K.J. (2004). Perceived Quality of Life: A Neglected Component of

Adolescent Health Assessment and Intervention. *Journal of Adolescent Health*, **34**, 270-278.

Hussain, T.M. and Smith, J.F. (1999). Relationship between maternal work and other socioeconomic factors and child health in Bangladesh. *Public Health*, **113**, 299-302.

Jiang, J. (2001). A non-standard χ^2-test with application to generalized linear model diagnostics. *Statistics and Probability Letters*, **53**, 101-109.

Knorr-Held, L. and Best, G.N. (2001). A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **164(1)**, 73-85

Korn, L.E. and Graubard, I.B. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science*, **17(1)**, 73-96.

Krzanowski, W. (1998). Introduction to Statistical Modelling. Arnold, London, 163-207.

Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. (2005). Applied Linear Statistical Models. 5^{th} ed, New York, McGraw-Hill Irwin.

Lehtonen, R. and Pahkinen, E.J. (1995). Practical Methods for Design and Analysis of Complex Surveys. Chichester: John Wiley & Sons.

Lesotho core welfare indicators questionnaire (CWIQ) survey, 2002. Bureau of Statistics, Lesotho (Demographic, Labour and Social Statistics Division).

Levy, P.S. and Lemeshow, S. (1991). Sampling of populations: Methods and applications. John Wiley & Sons, INC.

Lindsey, J.K. (1999). Applying Generalized Linear Models. New York, Springer.

Littell, C.R., Milliken, A.G., Stroup, W.W. and Wolfinger, D.R. (1996). SAS System for Mixed Models. Cary, NC: SAS Institute Inc.

Littell, C.R., Milliken, A.G., Stroup, W.W., Wolfinger, D.R. and Schabenberger, O. (2006). SAS System for Mixed Models. 2nd ed. Cary, NC: SAS Institute Inc.

Mason, S.J. and Graham, N.E. (2002). Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Q. J. R. Meteorol. Soc.*, **128**, 2145-2166

McCullagh, P. and Nelder, J.A. (1989). Generalized linear models. 2nd ed, Chapman & Hall Ltd.

Meyer, M.C. and Laud, P.W. (2002). Predictive Variable Selection in Generalized Linear Models. *Journal of the American Statistical Association*, **97(459)**, 859-871.

Nelder, J.A. and Lee, Y. (1992). Likelihood, quasi-likelihood and pseudolikelihood: Some comparisons. *Journal of the Royal Statistics Society (B)*, **54(1)**, 273-284.

Nelson, P.R. (1985). Power Curves for the Analysis of Means, *Technometrics*, **27(1)**, 6573.

Nelson, P.R. (1993). Additional Uses for the Analysis of Means and Extended Tables of Critical Values, *Technometrics*, **35**, 6171.

Pendergast, J.F., Gange, S.J. Newton, M.A., Lindstrom, M.J., Palta, M. and Fisher, M.R. (1996). A survey of methods for analyzing clustered binary response data. *International Statistical Review*, **64**, 89-118.

Pfeffermann, D. (1993). The Role of Sampling Weights When Modeling Survey Data. *International Statistical Review*, **61(2)**, 317-37.

Preisser, S.J. and Garcia, I.D. (2005). Alternative computational formulae for generalized linear model diagnostics: identifying infuential observations with SAS software. *Computational Statistics & Data Analysis*, **48**, 755-764.

Prior, P.M. and Hayes, B.C. (2001). Marital status and bed occupancy in health and social care facilities in the United Kingdom. *Journal of the Royal Institute of Public Health*, **115**, 401-406.

Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **169(4)**, 805-827.

Rousseeuw, P.J. and Leroy, A.M. (2003). Robust Regression and Outlier Detection. A John Wiley and Sons, INC. Publication.

Rust, R.T., Semester, D., Brodie, R.J. and Nilikant, V. (1995). Model selection criteria: An investigation of relative accuracy, posterior probabilities, and combination of criteria. *Management Science*, **41(2)**, 322-333.

Schabenberger, O. and Pierce, F.J. (2002). Contemporary Statistical Models: for the Plant and Soil Science. CRC Press, New York.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461-464.

Skinner, C.J., Holt, D. and Smith, T.M.F. (1989). Analysis of Complex Survey. Chichester: John Wiley & Sons.

Skovgaard Ib.M. and Ritz C. (2007). Influence on tests with focus on linear models. *Journal of Statistical Planning and Inference*, **137**, 1979-1991.

Taylor, P.D. and Krawchuk, M.A. (2005). Scale and sensitivity of songbird occurrence to landscape structure in a harvested boreal forest. *Avian Conservation and Ecology*, **1(1)**, 5.

The Kingdom of Lesotho, Poverty Reduction Strategy 2004/2005 - 2006/2007.

Ulukanligil, M. and Seyrek, A. (2004). Demographic and socio-economic factors affecting the physical development, haemoglobin and parasitic infection status of schoolchildren in Sanliurfa province, Turkey. *Journal of the Royal Institute of Public Health*, **118**, 151-158.

Villiant, R., Dorfman, A.H. and Royall, R.M. (2000). Fininte population and inference: A prediction approach. John Wiley & Sons, INC.

Vingilis, E., Wade, T.J. and Adlaf, E., (1998). What factors predict students self-rated physical health? *Journal of Adolescence*, **21**, 83-97.

Vittinghoff, E., Glidden, D.V., Shiboski, S.C. and McCulloch, C.E (2005). Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models. Springer, New York.

Waclawiw, M.A. and Liang, K.Y. (1993). Prediction of random effects in the generalized linear model. *Journal of the American Statistical Association*, **88**, 171-178.

Walker, A.E. and Becker, N.G. (2005). Health inequalities across socio-economic groups: comparing geographic-area-based and individual-based indicators. *Journal of the Royal Institute of Public Health*, **119**, 1097-1104.

WHO, www.who.int/mdg/goals/goal1/poverty_*and*_health/en/

Williams, D.A. (1987). Generalized linear models diagnostics using the deviance and single-case deletions. *Applied Statistics*, **36**, 181-191.

Wolfinger, R. and O'Connell, M. (1993). Generalized Linear Mixed Models: A Pseudo-Likelihood Approach. *Journal of Statistical Computation and Simulation*, **4**, 233-243.

Zeger, S.L and Liang, K.-Y. (1986). Longitutinal data analysis using generalized linear models, *Biometrika*, **73**, 13-22.

Zhang, Y. (2002). A semiparametric pseudolikelihood estimation method for panel count data. *Biometrika*, **89(1)**, 39-48.

Zullig, K.J., Valois, R.F. and Drane, J.W. (2005). Adolescent distinctions between quality of life and self-rated health in quality of life research. *Health and Quality of Life Outcomes*, **3**, 64-72.

Appendix A

Procedures for the Generalized Linear Models

A.1 SAS Procedures

The SAS system was used to fit the logistic regression model discussed in Chapter 2 and fitted in Chapter 4. PROC LOGISTIC and PROC GENMOD were used to fit the model. The stepwise procedure implemented in PROC LOGISTIC was used to select the best model. The scale deviance was not specified because for ungrouped binary data the problem of dispersion does not hold. The logit, the probit, and the complementary log-log link functions were used.

A.1.1 Model Selection Using the PROC LOGISTIC

The following stepwise selection procedure was used:

```
proc logistic data=sasuser.Recodeddw;
class U L S M A E D R SW T F TS TC H / param=reference;
model y = U|L|S|M|A|E|D|R|SW|T|F|TS|TC|H@3/ link=logit alpha=0.1 selection=stepwise
lackfit;
run;
```

where 'Lackfit' performs Hosmer-Lemeshow goodness-of-fit test for ungrouped binary response data. U=urban/rural, L=location, S=sex, M=marital status, A=age, E=education, D=dwelling, R=roofing, SW=source of drinking water, T=toilet, F=fuel, TS=time taken to reach the nearest supply of drinking water, TC=time taken to reach the nearest hospital/clinic, and H=household size.

A.1.2 Model Fitting Using the PROC GENMOD

Variables used here are selected using stepwise procedure in the PROC LOGISTIC.

```
proc genmod data=sasuser.Recodeddw;
class L S M A E D H;
model y = L S M S*M A M*A E D M*D D*E H D*H/ dist=binomial link=logit
apha=0.1 lrci type3;
run;
```

A.1.3 Plots Using the PROC LOGISTIC

The plots were done in the PROC LOGISTIC by including the following statements:

```
proc logistic data=sasuser.Recodeddw;
class U L S M A E D R SW T F TS T H;
model y = U|L|S|M|A|E|D|R|SW|T|F|TS|TC|H@2/ link=logit alpha=0.1 selection=stepwise lackfit;
output out=output p=pred c=c xbeta=logit resdev=resdev;
run;

data output;
set output;
obs=_N_;
/* To approximate Cookd (=Cook's distance), divide c by the total number of parameters in the model */
cookd=c/23;
run;

/*statements below perform plots for the fitted model*/
ods html;
ods graphics on;
proc gplot data=output;
plot cookd*obs;
plot resdev*logit;
plot resdev*obs;
run;
ods graphics off;
ods html close;
```

where

cookd*obs; invokes index plot of Cook's distance

resdev*logit; invokes plot of residual deviance against linear predictor

resdev*obs; invokes index plot of deviance residual

A.1.4 ROC Curve for the Selected Model

The following codes were used to graphically present the prediction accuracy of the model:

```
ods html;
ods graphics on;
proc logistic data=Recodedd;
class L S M A E D H;
model y = L|S|M|A|E|D|H@2/ link=logit plcl outroc=roc1;
run;
ods graphics off;
ods html close;
```

Appendix B

SAS PROC SURVEYLOGISTIC

This procedure was used to fit a survey logistic regression model discussed and fitted in Chapter 6. The same variables selected by PROC LOGISTIC were used to fit the survey logistic regression model. The other sub-model selected using the alternative procedure discussed in Section 6.3 is given in Table 6.1.

```
PROC SURVEYLOGISTIC DATA = sasuser.Recodeddw;
STRATUM U L;
CLUSTER PSU;
CLASS U L S M A E D R SW T F TS T H / param=reference;
MODEL y=L S M S*M A A*M E D E*D M*D H D*H / LINK=LOGIT STB RSQ alpha=0.1;
RUN;
```

where PSU is a primary sampling unit.

Appendix C

SAS PROC GLIMMIX

PROC GLIMMIX was used to fit generalized linear mixed model (random intercept model) discussed and fitted in Chapter 5. The same variables selected in Chapter 4 (for logistic regression model) were used to fit the random intercept model.

```
ods html;
ods graphics on;
proc glimmix data=sasuser.Recodeddw;
class PSU U L S M A E D R SW T F TS T H;
model y = L S M S*M A A*M E D E*D M*D H D*H / dist=binary solution alpha=0.1;
lsmeans L S*M A*M E*D M*D D*H / plot=diffplot adjust=turkey alpha=0.1;
lsmeans L S*M A*M E*D M*D D*H / plot=anomplot adjust=nelson alpha=0.1;
random int / subject=PSU;
run;
ods graphics off;
ods html close;
```

The option ddfm=kenwardrover, which uses Satterthwaite-based degrees of freedom, is included to account for uncertainty that may exist when estimating $\mathbf{G}$ and $\mathbf{R}$ in the model i.e. accounting for underestimation of true sampling variability of $[\hat{\boldsymbol{\beta}}', \hat{\mathbf{u}}']'$, (GLIMMIX Procedure manual, 2005). This option is put in the model statement after solution. The results are given in Table C.1. Note that these results are not (significantly) different from the ones in Tables 5.2 and 5.3 where adjustment for uncertainty is not done, concurring with what the manual claims for well balanced data.

The Satterthwaite-based degrees of freedom can also be obtained without accounting for uncertainty in estimating $\mathbf{G}$ and $\mathbf{R}$ by replacing ddfm=kenwardroger option with ddfm=satterth option in the model statement. The results are given in Table C.2.

Table C.1: Accounting for uncertainty in estimating **G** and **R**

Covariance Parameter Estimates

			Standard
Cov Parm	**Subject**	**Estimate**	**Error**
Intercept	PSU	0.02956	0.02311

Solutions for Fixed Effects

		Standard			
Effect	**Estimate**	**Error**	**DF**	**t Value**	**Pr > |t|**
Int	0.6015	0.3172	4931	1.90	0.0580
Butha-Buthe	-0.1966	0.1343	245.9	-1.46	0.1445
Leribe	0.5162	0.1360	262.2	3.80	0.0002
Berea	-0.00446	0.1354	252.8	-0.03	0.9737
Thaba-Tseka	-0.4606	0.1345	248	-3.43	0.0007
Mafeteng	0.05094	0.1338	244.3	0.38	0.7037
Mohale's Hoek	0.3509	0.1410	285.3	2.49	0.0134
Quthing	0.2947	0.1372	266.8	2.15	0.0326
Qacha's Nek	0.3741	0.1391	281.3	2.69	0.0076
Mokhotlong	-0.1807	0.1329	238.2	-1.36	0.1750
Sex	-0.6768	0.1134	4931	-5.97	<.0001
Mstatus	-0.5940	0.2140	4931	-2.78	0.0055
Sex*Mstatus	0.8068	0.1909	4931	4.23	<.0001
Age	-0.6196	0.1264	4931	-4.90	<.0001
Mstatus*Age	0.4991	0.1504	4931	3.32	0.0009
Education1	0.8953	0.1736	4931	5.16	<.0001
Education2	0.2429	0.1685	4931	1.44	0.1496
Dwelling	0.05312	0.3396	4931	0.16	0.8757
Mstatus*Dwelling	0.3432	0.1678	4931	2.05	0.0409
Education1*Dwelling	-0.7870	0.2259	4931	-3.48	0.0005
Education2*Dwelling	-0.3570	0.2307	4931	-1.55	0.1219
HHsize	-0.8275	0.2605	4931	-3.18	0.0015
Dwelling*HHsize	0.5696	0.2692	4931	2.12	0.0344

Table C.2: Using Satterthwaite-based degrees of freedom without adjustment for uncertainty for estimating **G** and **R**

Covariance Parameter Estimates

			Standard
Cov Parm	**Subject**	**Estimate**	**Error**
Intercept	PSU	0.02956	0.02311

Solutions for Fixed Effects

		Standard			
Effect	**Estimate**	**Error**	**DF**	**t Value**	**Pr> |t|**
Int	0.6015	0.3171	4931	1.90	0.0579
Butha-Buthe	-0.1966	0.1343	245.9	-1.46	0.1445
Leribe	0.5162	0.1359	262.2	3.80	0.0002
Berea	-0.00446	0.1354	252.8	-0.03	0.9737
Thaba-Tseka	-0.4606	0.1344	248	-3.43	0.0007
Mafeteng	0.05094	0.1338	244.3	0.38	0.7037
Mohale's Hoek	0.3509	0.1410	285.3	2.49	0.0134
Quthing	0.2947	0.1372	266.8	2.15	0.0326
Qacha's Nek	0.3741	0.1391	281.3	2.69	0.0076
Mokhotlong	-0.1807	0.1329	238.2	-1.36	0.1750
Sex	-0.6768	0.1133	4931	-5.97	<.0001
Mstatus	-0.5940	0.2140	4931	-2.78	0.0055
Sex*Mstatus	0.8068	0.1908	4931	4.23	<.0001
Age	-0.6196	0.1263	4931	-4.90	<.0001
Mstatus*Age	0.4991	0.1504	4931	3.32	0.0009
Education1	0.8953	0.1735	4931	5.16	<.0001
Education2	0.2429	0.1684	4931	1.44	0.1493
Dwelling	0.05312	0.3395	4931	0.16	0.8757
Mstatus*Dwelling	0.3432	0.1677	4931	2.05	0.0408
Education1*Dwelling	-0.7870	0.2258	4931	-3.49	0.0005
Education2*Dwelling	-0.3570	0.2306	4931	-1.55	0.1217
HHsize	-0.8275	0.2604	4931	-3.18	0.0015
Dwelling*HHsize	0.5696	0.2691	4931	2.12	0.0344

Table C.3: Marital status by age interaction least-squares means

MSTATUS*AGE Least-Squares Means

			Standard			
MSTATUS	**AGE**	**Estimate**	**Error**	**DF**	**t Value**	**Pr> |t|**
1	1	0.1421	0.1083	4683	1.31	0.1897
1	2	0.2625	0.1076	4683	2.44	0.0148
2	1	-0.3381	0.1137	4683	-2.97	0.0030
2	2	0.2815	0.1074	4683	2.62	0.0088

Differences of MSTATUS*AGE Least-Squares Means

					Standard				
MSTATUS	**AGE**	**MSTATUS**	**AGE1**	**Estimate**	**Error**	**DF**	**t Value**	**Pr> |t|**	**Adj P**
1	1	1	2	-0.1204	0.08699	4683	-1.38	0.1663	0.5093
1	1	2	1	0.4801	0.1260	4683	3.81	0.0001	0.0008
1	1	2	2	-0.1394	0.1233	4683	-1.13	0.2584	0.6708
1	2	2	1	0.6006	0.1295	4683	4.64	<.0001	<.0001
1	2	2	2	-0.01900	0.1233	4683	-0.15	0.8775	0.9987
2	1	2	2	-0.6196	0.1263	4683	-4.90	<.0001	<.0001

Differences of MSTATUS*AGE Least-Squares Means

					Standard				
MSTATUS	**AGE**	**MSTATUS**	**AGE1**	**Estimate**	**Error**	**DF**	**t Value**	**Pr> |t|**	**Adj P**
1	1	Avg	Avg	0.04526	0.06494	4683	0.70	0.4858	0.8792
1	2	Avg	Avg	0.1657	0.06625	4683	2.50	0.0124	0.0456
2	1	Avg	Avg	-0.4349	0.08378	4683	-5.19	<.0001	<.0001
2	2	Avg	Avg	0.1847	0.07766	4683	2.38	0.0174	0.0630

Table C.4: Marital status by ownership of dwelling interaction least-squares means

MSTATUS*DWELLING Least-Squares Means

			Standard			
MSTATUS	**DWELLING**	**Estimate**	**Error**	**DF**	**t Value**	**Pr> |t|**
1	1	0.3522	0.09127	4683	3.86	0.0001
1	2	0.05237	0.1521	4683	0.34	0.7307
2	1	-0.05000	0.08909	4683	-0.56	0.5747
2	2	-0.00660	0.1581	4683	-0.04	0.9667

Differences of MSTATUS*DWELLING Least-Squares Means

					Standard				
MSTATUS	**DWELLING**	**MSTATUS**	**DWELLING**	**Estimate**	**Error**	**DF**	**t Value**	**Pr> |t|**	**Adj P**
1	1	1	2	0.2998	0.1546	4683	1.94	0.0525	0.2116
1	1	2	1	0.4022	0.1111	4683	3.62	0.0003	0.0017
1	1	2	2	0.3588	0.1790	4683	2.00	0.0451	0.1864
1	2	2	1	0.1024	0.1745	4683	0.59	0.5574	0.9361
1	2	2	2	0.05897	0.1466	4683	0.40	0.6875	0.9780
2	1	2	2	-0.04340	0.1814	4683	-0.24	0.8109	0.9952

Differences of MSTATUS*DWELLING Least-Squares Means

					Standard				
MSTATUS	**DWELLING**	**MSTATUS**	**DWELLING**	**Estimate**	**Error**	**DF**	**t Value**	**Pr> |t|**	**Adj P**
1	1	Avg	Avg	0.2403	0.06524	4683	3.68	0.0002	.
1	2	Avg	Avg	-0.05947	0.1221	4683	-0.49	0.6262	.
2	1	Avg	Avg	-0.1618	0.07035	4683	-2.30	0.0215	.
2	2	Avg	Avg	-0.1184	0.1373	4683	-0.86	0.3882	.

Table C.5: Education by ownership of dwelling interaction least-squares means

EDUCATION*DWELLING Least-Squares Means

			Standard			
Education	**Dwelling**	**Estimate**	**Error**	**DF**	**t Value**	**Pr> |t|**
1	1	0.2613	0.06501	4683	4.02	¡.0001
1	2	0.5388	0.1606	4683	3.35	0.0008
2	1	0.03894	0.08468	4683	0.46	0.6456
2	2	-0.1136	0.1581	4683	-0.72	0.4725
3	1	0.1530	0.1473	4683	1.04	0.2990
3	2	-0.3565	0.1807	4683	-1.97	0.0485

Differences of EDUCATION*DWELLING Least-Squares Means

					Standard				
Education	**Dwelling**	**Education**	**Dwelling**	**Estimate**	**Error**	**DF**	**t Value**	**Pr> |t|**	**Adj P**
1	1	1	2	-0.2774	0.1631	4683	-1.70	0.0891	0.5314
1	1	2	1	0.2224	0.08552	4683	2.60	0.0093	0.0973
1	1	2	2	0.3749	0.1672	4683	2.24	0.0250	0.2187
1	1	3	1	0.1083	0.1471	4683	0.74	0.4617	0.9775
1	1	3	2	0.6178	0.1872	4683	3.30	0.0010	0.0125
1	2	2	1	0.4998	0.1731	4683	2.89	0.0039	0.0450
1	2	2	2	0.6524	0.1534	4683	4.25	<.0001	0.0003
1	2	3	1	0.3857	0.2102	4683	1.83	0.0666	0.4433
1	2	3	2	0.8953	0.1735	4683	5.16	<.0001	<.0001
2	1	2	2	0.1525	0.1739	4683	0.88	0.3804	0.9520
2	1	3	1	-0.1141	0.1578	4683	-0.72	0.4699	0.9792
2	1	3	2	0.3955	0.1937	4683	2.04	0.0413	0.3190
2	2	3	1	-0.2666	0.2113	4683	-1.26	0.2072	0.8060
2	2	3	2	0.2429	0.1684	4683	1.44	0.1493	0.7011
3	1	3	2	0.5095	0.2277	4683	2.24	0.0253	0.2208

Differences of EDUCATION*DWELLING Least-Squares Means

					Standard				
Education	**Dwelling**	**Education**	**Dwelling**	**Estimate**	**Error**	**DF**	**t Value**	**Pr> |t|**	**Adj P**
1	1	Avg	Avg	0.1113	0.04472	4683	2.49	0.0129	0.0720
1	2	Avg	Avg	0.3887	0.1372	4683	2.83	0.0046	0.0267
2	1	Avg	Avg	-0.1111	0.06541	4683	-1.70	0.0894	0.4019
2	2	Avg	Avg	-0.2637	0.1393	4683	-1.89	0.0584	0.2843
3	1	Avg	Avg	0.002951	0.1343	4683	0.02	0.9825	1.0000
3	2	Avg	Avg	-0.5066	0.1612	4683	-3.14	0.0017	0.0099

Table C.6: Ownership of dwelling by household size interaction least-squares means

DWELLING*HHSIZE1 Least-Squares Means

			Standard			
DWELLING	**HHSIZE**	**Estimate**	**Error**	**DF**	**t Value**	**Pr> |t|**
1	1	0.02216	0.07155	4683	0.31	0.7568
1	2	0.2800	0.08749	4683	3.20	0.0014
2	1	-0.3909	0.07896	4683	-4.95	<.0001
2	2	0.4366	0.2551	4683	1.71	0.0870

Differences of DWELLING*HHSIZE Least-Squares Means

					Standard				
DWELLING	**HHSIZE**	**DWELLING**	**HHSIZE**	**Estimate**	**Error**	**DF**	**t Value**	**Pr> |t|**	**Adj P**
1	1	1	2	-0.2579	0.07321	4683	-3.52	0.0004	0.0024
1	1	2	1	0.4130	0.09784	4683	4.22	<.0001	0.0001
1	1	2	2	-0.4145	0.2597	4683	-1.60	0.1105	0.3808
1	2	2	1	0.6709	0.1107	4683	6.06	<.0001	<.0001
1	2	2	2	-0.1566	0.2634	4683	-0.59	0.5521	0.9337
2	1	2	2	-0.8275	0.2604	4683	-3.18	0.0015	0.0081

Differences of DWELLING*HHSIZE Least-Squares Means

					Standard				
DWELlING	**HHSIZE**	**DWELLING**	**HHSIZE**	**Estimate**	**Error**	**DF**	**t Value**	**Pr> |t|**	**Adj P**
1	1	Avg	Avg	0.05911	0.04306	4683	1.37	0.1699	0.4580
1	2	Avg	Avg	0.3170	0.05885	4683	5.39	<.0001	<.0001
2	1	Avg	Avg	-0.3539	0.06345	4683	-5.58	<.0001	<.0001
2	2	Avg	Avg	0.4736	0.2469	4683	1.92	0.0551	0.1771