UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

GENERALIZED LINEAR MIXED MODEL AND GENERALIZED ESTIMATING EQUATION FOR BINARY LONGITUDINAL DATA

by

Sandra Moepeng Sepato

(BSc. Honours, Statistics)

Submitted in fulfilment of the requirements for the degree of

Magister Scientiae

in the

Department of Statistics

Faculty of Natural Sciences and Agriculture Sciences

University of Pretoria

Pretoria March 2014

Declaration

I, Sandra Moepeng Sepato declare that the dissertation, which I hereby submit for the degree Magister Scientiae in Mathematical Statistics at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

March, 2014.

Student :

_____
Sandra Moepeng Sepato

Supervisor :

_____
Legesse Kassa Debusho (Ph)

Acknowledgements

I would like to express my deep gratitude to my supervisory, Dr. Legesse Kassa Debusho, for all his guidance, support and motivation, without whom accomplishment of this work would not have been made possible. The time spent collaborating and discussing has been extremely instructive. I would also like to acknowledge Dr. Gudeta Sileshi for making the data used in the study available for us.

I am grateful to my family for the support and love they have given me throughout my study years. I would like also to thank my employer for the time they have given me to commit to my studies.

Finally, I gratefully acknowledge the financial support from National Research Foundation of South Africa through my supervisor research grant.

Summary

The most common analysis used for binary data is generalised linear model (GLM) with either a binomial or bernoulli distribution using either a logit, probit, complementary log-log or other type of link functions. However, such analyses violate the independence assumption if the binary data are measured repeatedly over time at the same subject or site. Failure to take into account the correlation can lead to incorrect estimation of regression parameters and the estimates are less efficient, particularly when the correlations are large. Therefore, to obtain the most efficient estimates that are also unbiased the methods that incorporate correlations (McCullagh and Nelder, 1989) should be used. Two of the statistical methodologies that can be used to account for this correlation for the longitudinal data are the generalized linear mixed models (GLMMs) and generalized estimating equation (GEE).

The GLMM method is based on extending the fixed effects GLM to include random effects and covariance patterns. Unlike the GLM and GLMM methods, the GEE method is based on the quasi-likelihood theory and no assumption is made about the distribution of response observations (Liang and Zeger, 1986). The main objective of the study is to investigate the statistical properties and limitations of these three approaches, i.e. GLM, GLMMs and GEE for analyzing longitudinal data through use of a binary data from an entomology study. The results reaffirms the point made by these authors that misspecification of working correlation in GEE approach would still give consistent regression parameter estimates. Further, the results of this study suggest that even with small correlation, ignoring a random effects in a binary model can lead to inconsistent estimation.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Objective of the Study

Binary data have a unique property that each observation in the data takes only one of two values and these values may represent presence or absence of a certain characteristic(s); or each observation in the data is classified into one of two categories, such as success or failure and defective or non-defective. Researchers in various fields usually encounter such type of data. For example, in ecology presence or absence of a particular species of insect, and in medicine a patient in a clinical trial to compare two different treatments may or may not experience relief from symptoms.

The most common analysis used for binary data is generalised linear model (GLM) with either a binomial or bernoulli distribution (Collett, 2003) using either a logit, probit, complementary log-log or other type of link functions. However, such analyses violate the independence assumption if the binary data are measured repeatedly over time at the same subject or site, i.e. when the data are longitudinal in nature. Statistical methods that assume independence among observations result in optimistic estimates of uncertainty when applied to correlated data, which are ubiquitous in applied ecological research (Fieberg et al., 2009). The generalized linear mixed models (GLMMs) could be used to take into ac-

count the correlation between observations within subject. As in the case of GLM, GLMM assumes specific probability distribution for the observations (data) and the introduction of a random intercept, that takes into account for within subject correlation, assumes a compound symmetry structure. The generalized estimating equations (GEE) approach, which is developed by Liang and Zeger (1986), is also used to analyze the longitudinal data (Vens and Ziegler, 2012). Unlike GLM and GLMM, the GEE methodology avoids the assumption of simultaneous distribution of observations. It only assume a functional form or the marginal distribution at each time and a correlation structure, called working correlation matrix. The specification of the working correlation matrix accounts for the form of within-subject correlation of observations of the response variable. It also provides various choice of correlation structures to reflect the random intercepts.

The objective of this dissertation is to investigate the different statistical models which can be used for binary data, in particular binary data that take values either 0 or 1 and have longitudinal nature. This investigation is planned in order to enable the researchers, for example ecologists, to improve the understanding of the statistical approaches they follow to analyse the binary longitudinal data.

1.2 Data used in the study

The data used in this study arise from an ecological study and consisted of the presence or absence of adults of the chrysomelid beetle *Mesoplatys ochroptera Stàl* in western Kenya. The species was monitored in two experiments established during 1999-2000 and 2000-2001, each consisting of three agroforestry treatments namely, pure *Sesbania sesban* (L.) Merrill, a mixture of *S. sesban* and *Tephrosia vogelii* Hook, and *S. sesban* and *Crotalaria grahamiana* Whight & Arn. The study was conducted at four sites in western Kenya namely, Dudi and Khumusalaba in Butere district, Mutumbu in Siaya district, and Lela in Kisumu district (Sileshi, Girma, and Mafongoya, 2006). However, in this study only the data collected in 1999-2000 were used for analysis. In each treatment, the abundance of *M. ochroptera* was

monitored on 15 randomly selected trees that were tagged using coloured plastic strings. The number of adults were recorded on seven dates (i.e. monthly from July to January) of sampling for each site and treatment in 1999 and 2000 on each tree. On each date, samples were taken from the same tree, and this constituted a repeated measures data set, i.e. longitudinal data. The total sample size was 1260. In the analysis the counts were recoded as 1 if the count is greater than or equal to 1 and 0 if the count is zero. The dates (time), sites and treatments were used as predictors of the presence and absence of adults of the chrysomelid beetle. Note that the data could be modelled using a GLM (Sileshi, Girma and Nyadzi, 2009) / GLMM/ GEE approaches that suits for counts data. Because repeated observations were recorded on the same tree within a site, one might expect observations from the same tree within a site to be more similar than observations from trees in different sites.

1.3 Overview of the Dissertation

In Chapter 2 the concepts of generalized linear model are introduced. This is then followed by more detail description of logistic regression modeling. The chapter also contains methods for model selection and diagnostics of logistic regression modeling. In the GLM approach, the model diagnostics are traditionally based on residual analysis, such as plots of residuals against the predicted values of the response variable and normal plots of residuals. We discuss in Chapter 2 that this method is appropriate but insufficient to detect potential outlying or influential cases in the binary data where the response takes values either 0 or 1. At the end of the chapter the GLM approach is used to analyze the study data.

The generalized linear mixed models (GLMMs) and their relevance for longitudinal binary data are illustrated in Chapter 3. The GLMM with subject-specific random intercepts and GLMM with subject-specific random effects are used to assess the effects of time, sites, treatments and their interactions on the presence and absence of adults of the chrysomelid beetle.

A formal introduction of the generalized estimating equations (GEE) approach, together with discussions on estimation of the model parameters are given in Chapter 4. Similar to Chapters 2 and 3, the GEE and its application to analyze the study data is done at the end of this chapter.

Finally, in Chapter 5, some concluding remarks regarding the three chapters or GLM, GLMMs and GEE approaches are presented.

Chapter 2

Generalized Linear Models

2.1 Exponential Family of Distributions

Consider a random variable y whose probability distributions $f(y;\theta,\phi)$ depends on parameters θ and ϕ. The distribution $f(y;\theta,\phi)$ belongs to the exponential family if $f(y;\theta,\phi)$ can be written as

$$f(y;\theta,\phi) = \exp\left\{\frac{[y\,\theta - b(\theta)]}{a(\phi)} + c(y;\phi)\right\} \tag{2.1}$$

where θ and ϕ are parameters. The parameter θ is called the canonical parameter and ϕ is the dispersion parameter (McCullagh and Nelder, 1989). The choice of the functions $a(.)$, $b(.)$ and $c(.)$ determine the actual probability distribution. For example the Bernoulli distribution $a(\phi) = 1$, $b(\theta) = \log[1+\exp(\theta)]$ and $c(y;\phi) = 1$; and for binomial distribution with n number of trials $a(\phi) = 1/n$, $b(\theta) = \log[1+\exp(\theta)]$ and $c(y;\phi) = \log\left[n!/\{(ny)!\,(n-n\,y)!\}\right]$ where y is the number of successes in n trials. Therefore, for the binary response variable the exponential family can be defined in terms of the location parameter θ only as

$$f(y;\theta) = \exp\left\{\frac{[y\,\theta - b(\theta)]}{a} + c(y)\right\} \tag{2.2}$$

where a is a constant.

The mean and variance of the exponential family in (2.2) are given by

$$E(y) = \mu = b'(\theta) \tag{2.3}$$

and

$$Var(y) = \mu = a\, b''(\theta), \tag{2.4}$$

respectively, where $b'(\theta)$ and $b''(\theta)$ are the first and second derivatives of $b(\theta)$ with respect to θ, respectively (McCullagh and Nelder, 1989). Observe from (2.3) and (2.4) that

$$\theta = b'^{-1}(\mu) \qquad \text{and} \qquad var(y) = a\, b''\left(b'^{-1}(\mu)\right). \tag{2.5}$$

Hence for the exponential family distribution the variance of y will vary with mean.

2.2 Generalized Linear Models

The generalized linear model (GLM) is defined in terms of a set of independent random variables $y_1, \ldots, y_n$ each with a distribution from the exponential family (McCullagh and Nelder, 1989; Dobson, 2002). Its definition needs three components. First is the response variable y (i.e. the random component) and its probability distribution (which is an exponential family). Second is the systematic component which represents the explanatory (predictor) variables $x_1, \ldots, x_p$ in the model. Third is the link function, a function that links the expected value of y and the systematic component by the function

$$g(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = \mathbf{x}'\boldsymbol{\beta} \tag{2.6}$$

where $g(\mu)$ is the link function, $\mathbf{x} = (1, x_1, \ldots, x_p)'$, x_j's are explanatory variables, $j = 1, \ldots, p$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ is vector of parameters. The link function $g(.)$ is a monotone and differentiable function. The explanatory variables might be continuous, covariates, dummy variables for levels of factors. The parameters $\boldsymbol{\beta}$ are estimated by the maximum likelihood estimation method.

2.2.1 The variance matrix

Let $y_1, \ldots, y_n$ be the response variables sharing the same distribution from the exponential family. Note that the means of these variables $\mu_i = E(y_i),\ i = 1, \ldots, n$ are not necessarily the same, such that

$$g(\mu_i) = \mathbf{x}_i' \boldsymbol{\beta}.$$

Since the response variables in GLM are independent, the variance matrix of $\mathbf{y} = (y_1, \ldots, y_n)$, $var(\mathbf{y}) = \mathbf{V}$ is a diagonal matrix with the diagonal terms which are equal to the variances of each variable given the underlying observations. Hence for the binary data with Bernoulli distributions, the mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ the variance matrix is given by

$$\mathbf{V} = \begin{pmatrix} \mu_1(1-\mu_1) & 0 & \ldots & 0 & \ldots & 0 \\ 0 & \mu_2(1-\mu_2) & \ldots & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \mu_i(1-\mu_i) & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & \ldots & \mu_n(1-\mu_n) \end{pmatrix}.$$

2.3 Model validation in GLM

The general aim of model validation is to detect evidence against model assumptions. Unlike the linear models assumptions in the GLM the interest in not in checking for evidence of non-normality, and/or heteroskedasticity of residuals. The assumptions of the GLM are different from those of the normal linear model, leading to different diagnostic tests. The residuals, which are calculated as $e_i = y_i - \hat{y}_i$ where $\hat{y}_i$ is fitted values of the response variable, and their studentized version are central to model checking for the normal linear model. For the GLM, these residuals are neither normally distributed, nor do they have constant variance, for any response distribution other than the normal. The definition of residual in GLM is broadened from the notion of the difference between observed and fitted values, to

a more general quantification of the conformity of a case to the model specification. There are two forms of residual that are commonly used in GLM for model validation, Pearson and deviance residuals. The Pearson residuals are defined by

$$e_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{var(y_i)}}.$$

The deviance residuals are defined by

$$e_i^D = sign(y_i - \hat{\mu}) \times \sqrt{d_i}$$

where *sign* stands for sign and has the value 1 if y_i is larger than $\hat{\mu}$ and -1 if y_i is smaller than $\hat{\mu}$ and d_i is the contribution of the ith observation to the deviance. McCullagh and Nelder (1989, page 398) recommend to use deviance residuals for model checking as these have distributional properties that are closer to the residuals from a normal linear regression model. However, in the GLM analysis for binary response attention is not focused on looking for normality from the Pearson or deviance residuals rather at looking for patterns in the Pearson or deviance residuals.

For model validation, the residuals of choice is plotted, for example deviance residual, against the fitted values, each explanatory variable in the model, each explanatory variable not in the model, i.e. those not used in the model or those dropped during the model selection procedure and against time. As it is discussed in Zuur et al (2009)

- If there are patterns in the graph with residuals against fitted values, it may indicate overdispersion or use of the wrong mean-variance relationship, i.e. wrong choice of distribution. However, since the response variable in our data set is a vector with zeros and ones, overdispersion is not expected (Zuur et al, 2009, page 253).

- If there are patterns in the graph with residuals against omitted explanatory variables, then the possible solution is to include them in the model.

- If there are patterns in the graph showing variability against each explanatory variable used in the model, then either include quadratic terms or conclude that there is violation of independence.

- If there are patterns in the graph with residuals against time, this may indicate violation of independence assumption. As a possible solution consider either a generalized linear mixed model or generalized estimation equation.

2.3.1 Residual deviance or Deviance

The residual deviance, sometimes called deviance, is defined as twice the difference between the log-likelihood of the saturated model, a model that provided a perfect fit, $log(L(\mathbf{y}, \mathbf{y}))$ and the model under study $log(L(\mathbf{y}, \mu))$, that is

$$Deviance = 2 \times [log(L(\mathbf{y}, \mathbf{y})) - log(L(\mathbf{y}, \mu))]$$

where $\mathbf{y}$ and μ refer to a vector of observations and mean, respectively. The residual deviance statistic is approximately Chi-square distributed with $N - p$ degrees of freedom, where p is the number of regression parameters in the model and N the number of observations. The smaller the residual deviance, the better is the model. However, McCullagh and Nelder (1989, page 118-119) argue that for the binomial GLM, a large value of the residual deviance cannot always be seen as evidence of a poor fit.

To compare two models, say M_1 and M_2, we can use Chi-square approximation for deviances. Let D_1 and D_2 be the deviances of M_1 and M_2, respectively and p_1 and p_2 be the number of parameters in M_1 and M_2, respectively. Then the difference between D_1 and D_2 is asymptotically Chi-square distributed with $p_1 - p_2$ degrees of freedom. Therefore, if $D_2 - D_1 > \chi^2_{p_1-p_2,\alpha}$, then M_1 is a better model.

2.4 Logistic Regression Model

In this section we review methodological approaches for logistic regression. The logistic regression also called a binary regression model is a special case of the generalized linear model (GLM). The model is intoduced first and the techniques used in obtaining parameters

and precision estimates are explored. The software used in implementing the model is also highlighted.

Let y_i be a binary dependent variable and $\mathbf{x}_i$ be the corresponding $p\times 1$ vector of explanatory variables or covariates, $i = 1, \ldots, n$. It is assumed that the number of observations n is equal to or greater than the number of explanatory variables, i.e. $n \geq p$ and that the n vectors $\mathbf{x}_i$ have full rank, i.e. $rank(\mathbf{x}_i) = p$, $i = 1, \ldots, n$. In general, a binary response variable is represented as a variable taking either 0 or 1 modelled by the Bernoulli or binomial distribution with number of trials is equal to 1.

Consider a binary response variable y with $y = 0$ or $y = 1$. For example y indicates whether adult of the Chrysomelid beetle is present or not. The aim is to explain y in terms of explanatory variables contained in $\mathbf{x}$. If π is the probability that $y = 1$ then y has a Bernoulli distribution. The density function of Bernoulli distribution is given by

$$f(y;\pi) = \pi^y \, (1-\pi)^{(1-y)}. \tag{2.7}$$

By taking the logarithm and then exponentiating (2.7) the density function can be expressed as

$$f(y;\pi) = \exp\left[y \, \frac{\pi}{1-\pi} + \log(1-\pi)\right].$$

Let $\theta = \pi/(1-\pi)$, hence the mean of the Bernoulli distribution π can be expressed as the inverse of the logit function

$$\mu = \pi = P(y=1) = \frac{\exp(\theta)}{1+\exp(\theta)}.$$

Therefore, using the expression of θ the Bernoulli density function can be expressed in the exponential family form, given in expression (2.2), as

$$f(y;\theta) = \exp\{y\,\theta - \log[1+\exp(\theta)]\}$$

with $b(\theta) = \log[1+\exp(\theta)]$, $a = 1$ and $c(y) = 1$. Further, using expression (2.6), the logistic regression with p-explanatory variables $\mathbf{x}_i$ can be given by

$$g(\mu) = g(\pi) = \mathbf{x}_i' \, \boldsymbol{\beta}.$$

where $g(\pi)$ is a link-function. There are various link functions that can be used to model binary outcomes, such as logit, probit, complementary log-log and user defined link function. The logit link is popular because the coefficients have a clear interpretation via the odds ratio. Whereas probit link is used in situation where only thresholds of a normally distributed latent variable are observable, for example in econometrics and dose-response studies and the complementary log-log link function is used in cases where the original count Poisson data are reduced to binary (Demidenko, 2013).

2.4.1 Estimation of $\boldsymbol{\beta}$

Recall the above notations that y_i be a binary outcome variable that measures the occurrence of an event with $y_i = 1$ means the event took place and $y_i = 0$ means that it did not occur, the $p \times 1$ vector $\mathbf{x}_i$ denotes the vector of explanatory variables (covariates). In logistic regression, the occurrence of the event y_i given $\mathbf{x}_i$ is modelled via probability as

$$P(y_i = 1|\mathbf{x}_i) = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}}, \qquad i = 1, \ldots, n \tag{2.8}$$

where the $p \times 1$ vector $\boldsymbol{\beta}$ is the parameter of interest. The likelihood function for the response vector $\mathbf{y} = (y_1, \ldots, y_n)$ can thus be expressed as

$$L(\boldsymbol{\beta}, \mathbf{y}) = \prod_{i=1}^{n} \left(\frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}} \right) \tag{2.9}$$

and hence the log-likelihood function for the response vector $\mathbf{y}$ is given by

$$\ell = \ln(L(\boldsymbol{\beta}, \mathbf{y})) = \boldsymbol{\beta}' \sum_{i=1}^{n} \mathbf{x}_i - \sum_{i=1}^{n} \ln(1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}). \tag{2.10}$$

The maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ is obtained by maximizing this function with respect to $\boldsymbol{\beta}$. In particular, differentiating ℓ with respect to $\boldsymbol{\beta}$ yields the score or estimating equations

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{1}{1 + e^{\mathbf{x}_i' \boldsymbol{\beta}}} \mathbf{x}_i - \sum_{i=1}^{n} \mathbf{x}_i = \mathbf{0}. \tag{2.11}$$

In practice this equation is solved numerically by iteration, e.g. using the Newton Raphson or Fisher scoring method.

The information matrix for the parameters $\boldsymbol{\beta}$ is given by

$$\mathbf{I}_\beta = -E\left(\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right) = \sum_{i=1}^{n} \frac{e^{\mathbf{x}_i' \boldsymbol{\beta}}}{(1 + e^{\mathbf{x}_i' \boldsymbol{\beta}})^2} \mathbf{x}_i \mathbf{x}_i'. \tag{2.12}$$

The information matrix plays a crucial role in inference. For example, it is well known that the variance of the maximum likelihood estimator of the parameters can be approximated asymptotically by the inverse of the information matrix (Azzalini, 1996, page 83).

Recall that either the Newton Raphson or Fisher scoring method can be used to maximize the log-likelihood function in expression (2.11). Both methods have the following generic form for the MLE of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_{k+1} = \hat{\boldsymbol{\beta}}_k + \lambda_k \mathbf{I}_\beta^{-1} \left(\frac{\partial \ell}{\partial \boldsymbol{\beta}}\right)\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_k}, \tag{2.13}$$

where $0 < \lambda_k \leq 1$ is a step length to provide a decrease of ℓ (Demidenko, 2013). For the Newton Raphson method $\mathbf{I}_\beta$ is the negative Hessian matrix whereas for the Fisher scoring method $\mathbf{I}_\beta$ is the expected negative Hessian matrix. At the final iteration $\mathbf{I}_\beta^{-1}$ gives the asymptotic variance-covariance matrix of the MLE of $\boldsymbol{\beta}$.

2.5 Goodness-of-fit for GLM: Logistic Regression

Unlike other GLMs, e.g. Poisson regression for count response variable, the deviance is not a useful measure of goodness of fit of the logistic regression. To illustrate this, consider the predicted probability of observing the event

$$\hat{\mu}_i = \frac{e^{\mathbf{x}_{ij}' \hat{\boldsymbol{\beta}}}}{1 + e^{\mathbf{x}_{ij}' \hat{\boldsymbol{\beta}}}}$$

where $\hat{\boldsymbol{\beta}}$ is a vector of the ML estimate. Collett (2003) has shown that the deviance expressed as

$$Deviance = -2 \sum_{i=1}^{n} \left[\hat{\mu}_i \ln \frac{\hat{\mu}_i}{1-\hat{\mu}_i} + \ln(1-\hat{\mu}_i) \right].$$

This deviance depends on the observed y_i only through the fitted values $\hat{\mu}_i$. Therefore, the deviance is not informative about the goodness of fit of $\hat{\mu}_i$ to y_i. Further, according to Collett (2003, pages 69-70) under some conditions the deviance can not be approximately Chi-square distributed. However, the Pearson Chi-square statistic, which is defined as

$$\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i(1-\hat{\mu}_i)}$$

(Dobson, 2002) depends on the actual y_i values. It has approximately a Chi-square distribution with $n-p$ degrees of freedom, where n is the number of observations and p is the number of parameters in the logistic model, under the hypothesis that the model is correct. Dobson (2002, page 126) has shown that the Pearson Chi-square statistic is asymptotically equivalent to the deviance. However, if the expected frequencies are too small, the approximation can be poor and the the Pearson Chi-square statistics is not considered a reliable measure of fit.

Therefore, other goodness of fit, such as a 2 by 2 classification table, Receiver Operating Characteristic (ROC) and Hosmer Lemeshow statistic (1980) can be used, for example to assess the predictive usefulness of a logistic model (De Jong and Heller, 2008; page 108).

2.5.1 Model checking for binary data

Consider now a logistic regression for a binary data which takes values 0 and 1. The validity of the fitted model includes, for example, checking the linear systematic component of the model is correctly specified, correct link function used and whether the data contain outliers and/or influential observations. The methods used to investigate these on fitted model are known as model diagnostics (Collett, 2003). Some of these methods are data or model

dependent, so the discussion and consideration is only on those methods which are suitable to binary data.

Checking the linear systematic component of the model

The properties of the deviance and Pearson residuals are going to be discussed first. Following Collett (2003) discussion, the deviance residuals for binary data defined as

$$r_i^D = \frac{sign(y_i - \hat{\pi}_i)}{\sqrt{-2\left[y_i \log(\hat{\pi}_i) + (1-y_i)\log(1-\hat{\pi}_i)\right]}},$$

where $\hat{\pi}_i = P(y_i = 1)$ and the Pearson residuals as

$$r_i^P = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i\,(1-\hat{\pi}_i)}}.$$

Both residuals and their corresponding standardized versions take one of the two values depending on the observed value of y_i, $i = 1, \ldots, n$. For example, for $y_i = 1$

$$r_i^D = \frac{sign(1-\hat{\pi}_i)}{\sqrt{-2\log(\hat{\pi}_i)}} > 0, \quad \text{because} \quad 0 < \hat{\pi}_i < 1$$

and

$$r_i^P = \frac{1-\hat{\pi}_i}{\sqrt{\hat{\pi}_i\,(1-\hat{\pi}_i)}} > 0, \quad \text{because} \quad 0 < \hat{\pi}_i < 1.$$

Similarly, for $y_i = 0$

$$r_i^D = \frac{sign(-\hat{\pi}_i)}{\sqrt{-2\log(1-\hat{\pi}_i)}} < 0, \quad \text{because} \quad 0 < \hat{\pi}_i < 1$$

and

$$r_i^P = -\frac{\hat{\pi}_i}{\sqrt{\hat{\pi}_i\,(1-\hat{\pi}_i)}} < 0, \quad \text{because} \quad 0 < \hat{\pi}_i < 1.$$

The third type of residuals that can be used for model checking are the likelihood residuals, which estimate components of a likelihood ratio test of deleting an individual observation. They are a weighted combination of the standardized Pearson and deviance residuals, that is

$$r_i^L = \frac{sign(y_i - \hat{\pi}_i)}{\sqrt{h_i\,(r_i^P)^2 + (1-h_i)(r_i^D)^2}},$$

where h_i is the ith diagonal element of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ where $\mathbf{X}$ is the $n \times p$ design matrix. Since h_i measures the effect that y_i has on the determination of fitted value, it is also known as the leverage. These diagonal elements are usually small and hence the values of r_i^L are approximately equal to r_i^D and have similar property.

Hence the distribution of residuals obtained from modelling binary data will not be approximately normally distributed and plots of these residuals against predicted probability or explanatory variables in the fitted model will show patterns whether the fitted model is valid or not. The use of these plots for model checking is therefore invalid. However, plots of the deviance or likelihood residuals or leverage (h_i) against the case numbers can provide information about the possible influential or outlier observations. Furthermore, there are statistical measures that help to identify the influential observations and for details refer to Collett (2003, pages 158 - 160). The leverage measures the extent to which the ith observation is separated from the others in terms of the values of the explanatory variables.

Selection of a link function

The assessment or selection of a link function can be affected by outliers or an incorrect linear predictors. Collett (2003, page 149) suggests that selection of the link function should only be carried out after model checking procedures have been used to validate other aspects, such as checking outliers or influential observations, of the fitted model. The AIC can be used for a link function selection.

2.5.2 Software

The standard logistic regression (i.e. GLM) analysis of the study data is done by using PROCEDURE LOGISTIC in SAS (SAS Institute Inc. 2013. SAS/STAT® 13.1 Users Guide). The analysis can also be done, for example, using PROCEDURE GLM of SAS or using the basic function *glm* in R.

2.6 Analysis of adults of the Chrysomelid beetle data using GLM

We assume that the response variable y_i, which is coded as 1 if adults of the Chrysomelid beetle is present on a tree within a site and 0 otherwise, is Bernoulli distributed with probability π_i. Hence the expected mean and variance of y_i are given by $E(y_i) = \pi_i$ and $var(y_i) = \pi_i(1-\pi_i)$. The presence/absence of adults of the Chrysomelid beetle recorded at seven time points (months). There were 15 trees from each of four sites for data collections. An exploratory analysis was done using plot of observed probabilities for presence of adults of the Chrysomelid beetle against time and it is displayed in Figure 2.1.

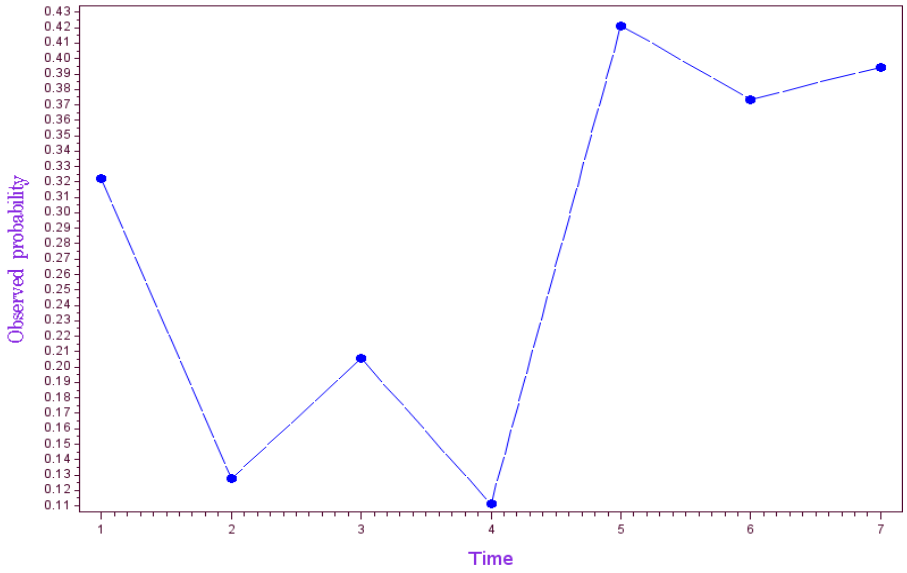

Figure 2.1: Plots of observed probabilities against time.

The plot shows that the probability of the presence of adults of the Chrysomelid beetle various through time and hence $var(y_i)$. A rapid decrease in variance is observed between the first and the second months and the highest increase in the fifth month. Since there is no general pattern in the level of variation, it can be implied that a quadratic term in time

($Time2 = Time \times Time$) should be introduced in the model. Even though it seems that the standard logistic regression model is not suitable the model fitting process is started with the standard logistic regression model for comparison purposes. This analysis ignores the possible correlation structure that might occur due to the repeated measurements from a tree within a site.

The logistic regression model for the data using logit link is therefore given by

$$\begin{aligned} logit(\pi_i) = \beta_0 + \beta_1 \times Site + \beta_2 \times Time + \beta_3 \times Site \times Time + \beta_4 \times Time2 \\ + \beta_5 \times Time2 \times Site + \beta_6 \times Treatment + \beta_7 \times Site \times Treatment + \beta_8 \times Site \times Time2 \\ + \beta_9 \times Time \times Treatment + \beta_{10} \times Time2 \times Treatment \end{aligned} \tag{2.14}$$

This model can also be written using a probability notation as

$$\pi_i = \frac{e^{\eta(\mathbf{x}_i)}}{1 + e^{\eta(\mathbf{x}_i)}}$$

where $\eta(\mathbf{x}_i) = \beta_0 + \beta_1 \times Site + \beta_2 \times Time + \beta_3 \times Site \times Time + \beta_4 \times Time2 + \beta_5 \times Time2 \times Site + \beta_6 \times Treatment + \beta_7 \times Site \times Treatment + \beta_8 \times Site \times Time2 + \beta_9 \times Time \times Treatment + \beta_{10} \times Time2 \times Treatment$.

Note that we have written the model in a simplified notation but in the analysis Site and Treatment are fitted as factors with four and three factor levels, respectively, and Time so $Time2$ are continuous variables.

As it is discussed in Section 2.2, the analysis of a GLM consists of three steps, the distribution of the response variable, the specification of the expected value in terms of explanatory variables and the link function. For the GLM, i.e. the standard logistic regression, analysis the PROCEDURE LOGISTIC in SAS is used. The Type 3 analysis of effects results based on the Wald test are shown in Table 2.1.

Table 2.1: Effects tests in Logistic regression analysis using the Wald (Wald Chi-square) test, DF (degrees of freedom) and p-value (Wald Chi-square p-value)

Effect	DF	Wald	p-value
Site	3	60.8374	< 0.0001
Time	1	7.8573	0.0051
Time×Site	3	69.8835	< 0.0001
Time2	1	16.6312	< 0.0001
Time2×Site	3	64.5677	< 0.0001
Treatment	2	14.4076	0.0007
Site×Treatment	6	5.6654	0.4617
Time×Treatment	2	8.7451	0.0126
Time2×Treatment	2	7.9766	0.0185

The results show that the $Site \times Treatment$ interaction is not statistically significant ($p - value = 0.4617$). This indicates that there is no evidence that the treatments affect the presence of adults of the Chrysomelid beetle in four sites differently.

2.6.1 Model selection

The analysis that is done using PROCEDURE LOGISTIC, utilises either the selection criterion like $-2\log(L)$, or the likelihood ratio test if two models are nested, i.e. one is obtained from the other by putting some of the parameters equal to zero or using AIC if the models are not nested. Let M_1 be the model containing all explanatory variables including the first-order interactions (model 2.14) and M_2 be a model without Site by Treatment interaction. The analyses on models M_1 and M_2 resulted in 1123.459 and 1129.309, respectively for $-2\log(L)$. A difference of these value $\chi^2 = 5.85$ approximately follows a Chi-square

Table 2.2: Effects tests in Logistic regression analysis using the Wald (Wald Chi-square) test, DF (degrees of freedom) and p-value (Wald Chi-square p-value)

Effect	DF	Wald	p-value
Site	3	62.5040	< 0.0001
Time	1	8.3638	0.0038
Time×Site	3	70.8382	< 0.0001
Time2	1	17.3902	< 0.0001
Time2×Site	3	65.1124	< 0.0001
Treatment	2	16.2211	0.0003
Time×Treatment	2	12.5414	0.0019
Time2×Treatment	2	11.3570	0.0034

distribution with 6 ($= DF(M_1) - DF(M_2) = 23 - 17$) degrees of freedom, which gives a p-value of 0.4401. Refitting of model M_2 resulted in all effects being significant at the 5% level (see Table 2.2). Even though the difference is non-significant we prefer model M_2 to model M_1 as it is a parsimonious model.

Except the regression coefficients for interaction levels $Dudi \times Time$ and $Dudi \times Time2$, treatment level SSTV, $SSTV \times Time$ and $SSTV \times Time$ the rest are significant at 5% level (see Table 2.3).

The Hosmer Lemeshow approach classifies the data in categories (for our data it formed 10 categories, results not shown) defined by grouping the predicted values of $\hat{\pi}_i$ so that the total numbers of observations per category, i.e. the observed frequencies (O_l), are approximately equal. The expected frequencies (E_l)are obtained using $\sum n_i \hat{\pi}_i$ for observations with adults

Table 2.3: Parameter estimates, standard error, Wald χ^2 p-value for model M_2.

Parameter	Estimate	Standard error	Wald χ^2	p-value
Intercept	-1.0426	0.4702	4.9177	0.0266
Site (Ref: Ochinga)				
Dudi	0.7224	0.5865	1.5170	0.2181
Khumus	5.0400	0.6921	53.0276	$< .0001$
Marenyo	-4.0762	1.1910	11.7142	0.0006
Time	-0.7368	0.2548	8.3638	0.0038
Time×Site				
Dudi	0.4193	0.3239	1.6761	0.1954
Khumus	-3.7215	0.4941	56.7407	$< .0001$
Marenyo	1.4199	0.5612	6.4030	0.0114
Time2	0.1222	0.0293	17.3902	$< .0001$
Time2×Site				
Dudi	-0.0625	0.0381	2.6876	0.1011
Khumus	0.4359	0.0601	52.5585	$< .0001$
Marenyo	-0.1476	0.0609	5.8839	0.0153
Treatment (Ref: WSS)				
CRSS	1.5889	0.4229	14.1134	0.0002
SSTV	-0.1452	0.4297	0.1141	0.7355
Time×Treatment				
CRSS	-0.7842	0.2414	10.5499	0.0012
SSTV	0.0699	0.2444	0.0818	0.7749
Time2×Treatment				
CRSS	0.0933	0.0290	10.3270	0.0013
SSTV	-0.0198	0.0293	0.4572	0.4980

of the Chrysomelid beetle and $\sum n_i(1-\hat{\pi}_i)$ without adults of the Chrysomelid beetle for each category. The Hosmer-Lemeshow statistic is then calculated as

$$\chi^2_{HL} = \sum_{l=1}^{10} \frac{(O_i - E_i)^2}{E_i}$$

and it has a Chi-square distribution with 8 degrees of freedom. For the current data $\chi^2_{HL} = 10.916$ with $p - value = 0.2065$ and it is non-significant implying good fit.

Receiver Operating Characteristic (ROC) Curve: Once the 2 by 2 classification table is obtained the predictive usefulness of a fitted model can be summarized using the sensitivity (the relative frequency of predicting the presence of adults of the Chrysomelid beetle when it is present) and specificity (the relative frequency of predicting the absence of adults of the Chrysomelid beetle when it is not present). The ROC curve plots sensitivity against specificity for each threshold. Traditionally one minus the specificity is plotted on the horizontal axis, and sensitivity on the vertical axis. With this orientation of the axes, a value near zero on the X-axis, shows high specificity and it generally implies a low value on the Y-axis, i.e. low sensitivity, and vice versa.

All ROC curves start at (0,0) and end at (1,1) as these points correspond to threshold probabilities 0 and 1, respectively. A model with perfect predictive ability has sensitivity and specificity both equal to one, giving a ROC curve consisting of the single point in the top left hand corner. A ROC curve for a fitted model with good predictive ability rises quickly to 1 and will have a curve at a far top left hand corner. In practice the model predictive ability is quantified by computing the area under the ROC curve (AUC). The AUC has a maximum of 1. A model with ROC equal to the 45^o line (AUC = 0.5) has a predictive ability no better than chance. For the current data, model M_2 yields AUC = 0.819 indicating very good predictive ability (see Figure 2.2).

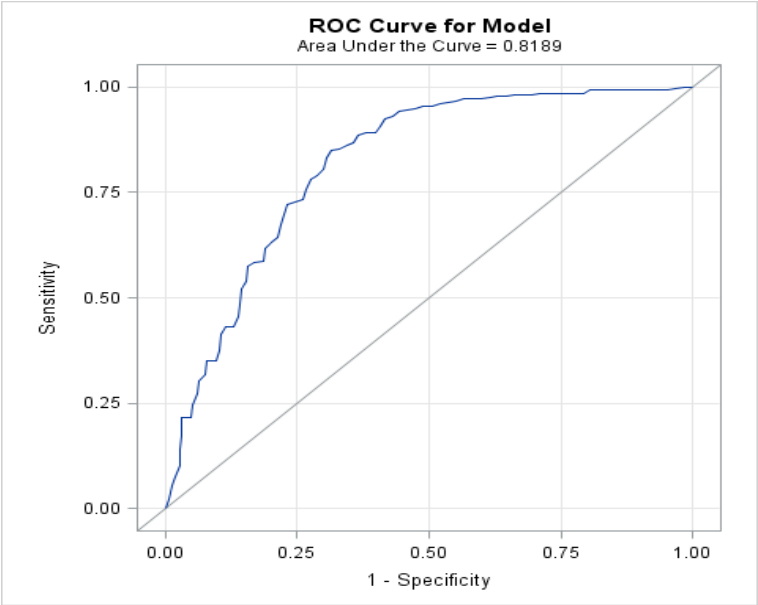

Figure 2.2: ROC curve for model M_2.

2.6.2 Model diagnostic

The Pearson Chi-squared or Deviance residuals can be used for checking the adequacy of a GLM, for example, they should be plotted against Time, continuous explanatory variable in the model, to check if the assumption of linearity is appropriate. However, if there are few distinct values of the residuals, the residual plots of binary response may be relatively uninformative. Therefore, plots of standardized deviance residuals, likelihood residuals and leverages against case numbers are produced (Figures 2.3 - (2.5)).

In the three plots observations 457 and 476 stand out as having relatively large residuals (Figures 2.3 and 2.4) and large leverage values (Figure 2.5). These two observations are not well fitted by the model, M_2. In addition, cases 457 and 476 have large values for change in deviance 10.302 and 10.566, respectively. This indicates that the two observations might be seriously affecting the fit of the model (Collett, 2003, page 166). To assess the effect of these observations on the values of the regression coefficients we have run the analysis of case deletion, i.e. leave-one-out analysis. The effects of deletion of these cases on the

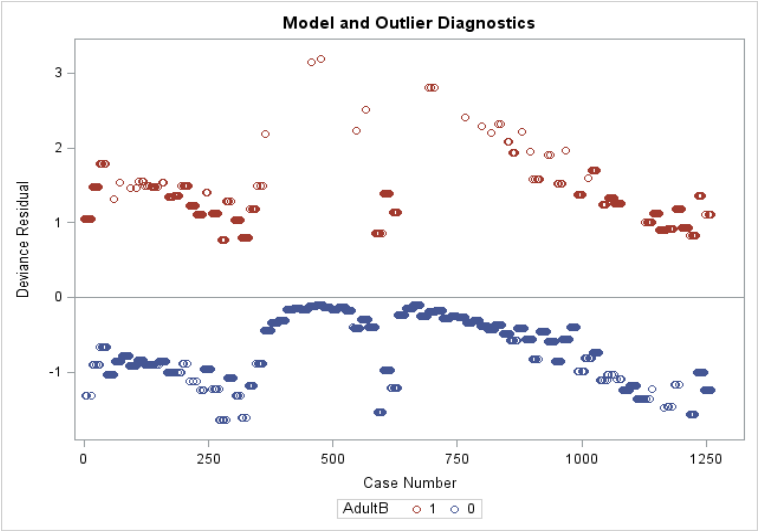

Figure 2.3: Plot of standardized deviance residuals against case numbers for M_2 model.

parameter estimates in the fitted logistic regression model, cases 457 and 476, are presented in Table 2.4 and Table 2.5, respectively as change of regression coefficients values. As shown in the tables, these cases are influential because their deletion have a huge impact in most of the regression coefficients estimates such as, coefficients of Khumus, Time, Time×Khumus, Time2, and Time2×Khumus.

Model M_2 has refitted by deleting cases 457 and 476. The plots in Figures 2.6, 2.7 and 2.8 do not suggest that there are any unusual observations. Furthermore, the Hosmer-Lemeshow statistic has a value 3.6748 with 8 degrees of freedom and p-value = 0.8852 and the area under ROC curve is 0.8233 (slight improvement compared to M_2 with cases 457 and 476); both showing a very good fit. Note that there is inconsistency in sign for the regression coefficient of $Time2 \times SSTV$ interaction level.

Table 2.4: Change in regression coefficients estimates in model M_2 when case 457 is deleted.

Effect	$\Delta\hat{\beta}_i$	Effect	$\Delta\hat{\beta}_i$	Effect	$\Delta\hat{\beta}_i$
Intercept	-0.1617	Time×Khumus	0.5005	CRSS	-0.0852
Dudi	0.1307	Time×Marenyo	-0.1485	SSTV	0.0356
Khumus	-0.3406	Time2	-0.3469	Time×CRSS	0.1250
Marenyo	0.0673	Time2×Dudi	0.2671	Time×SSTV	-0.0555
Time	0.3186	Time2×Khumus	-0.5139	Time2×CRSS	-0.1283
Time×Dudi	-0.2515	Time2×Marenyo	0.1711	Time2×SSTV	0.0591

Table 2.5: Change in regression coefficients estimates in model M_2 when case 476 is deleted.

Effect	$\Delta\hat{\beta}_i$	Effect	$\Delta\hat{\beta}_i$	Effect	$\Delta\hat{\beta}_i$
Intercept	-0.1678	Time×Khumus	0.5019	CRSS	0.0286
Dudi	0.1347	Time×Marenyo	-0.1431	SSTV	-0.0780
Khumus	-0.3393	Time2	-0.3568	Time×CRSS	-0.0581
Marenyo	0.0631	Time2×Dudi	0.2731	Time×SSTV	0.1285
Time	0.3280	Time2×Khumus	-0.5157	Time2×CRSS	0.0621
Time×Dudi	-0.2571	Time2×Marenyo	0.1652	Time2×SSTV	-0.1327

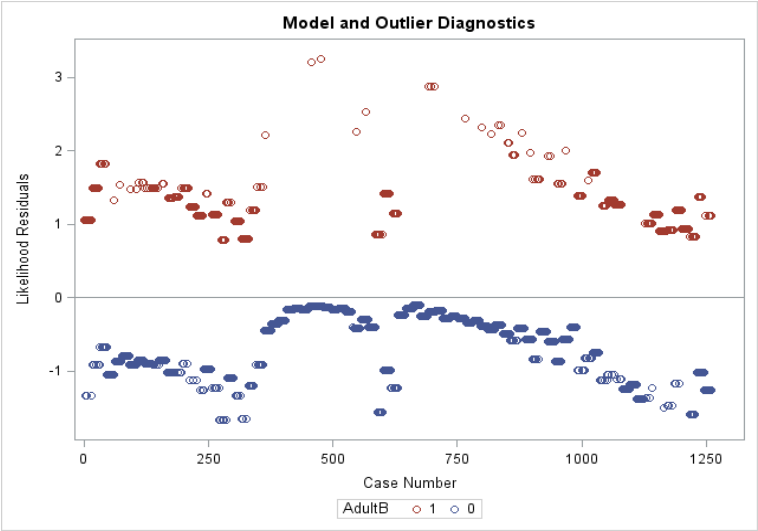

Figure 2.4: Plot of likelihood residuals against case numbers for M_2 model.

To investigate the sensitivity of the analysis with respect to varying link functions, we replaced the logit link function by the probit and complementary log(-log) links and these models were compared using the Akaike information criterion (AIC). There is no noticeable difference among the models with a probit link (AIC = 1142.191), complementary log(-log) link (AIC = 1146.217) and the model with a logit link (AIC = 1143.522). Hence the logit link function is retained throughout the dissertation. The deletion of cases 457 and 476 reduced the AICs for logit, probit and complementary log(-log) links by 27.937, 30.047 and 26.843, respectively.

The common features of the generalized linear models (GLMs) and hence the logistic regression models are that their linear, i.e. systematic components contain terms know as fixed effects (Collett, 2003) and the observations are assumed to be independent. However,when the binary data, for example, are from studies which have hierarchic or multilevel structure or from longitudinal studies, where repeated observations on the same subjects collected through time, the logistic regression in the GLM sense is no longer a valid method. This is due to the fact that a hierarchical design structure may induce correlation between

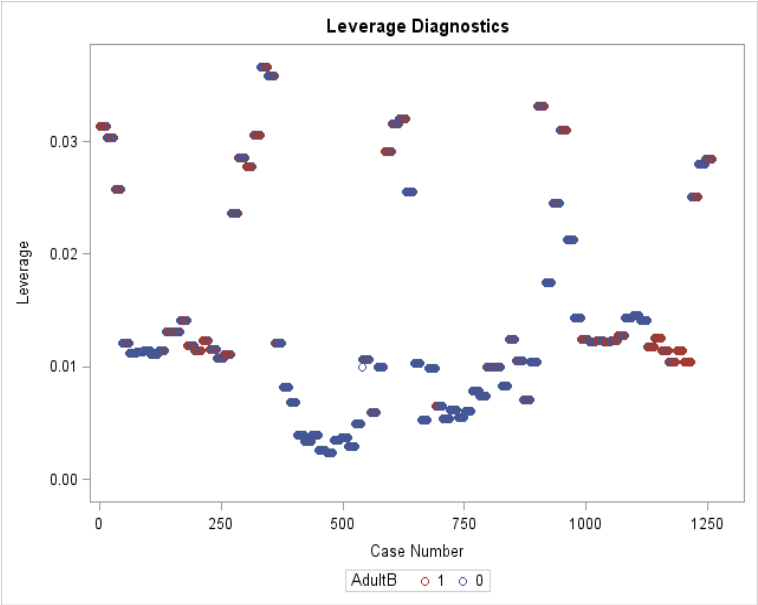

Figure 2.5: Plot of the values of leverage against case numbers for M_2 model.

or within cluster and likewise the repeated measurements in a longitudinal study may also induce within subject correlation.

Two of the statistical methodologies that can be used to account for this correlation for the longitudinal data are the generalized linear mixed models (GLMMs) explained in chapter 3 and generalized estimating equation (GEE) explained in chapter 4. Even though the methods are different from each other they both solve the violation of the independence assumption in GLM. They are equally important as they ensure that the regression parameters are correctly estimated. Both methods are efficient and produce unbiased estimates when correlation are taken into account.

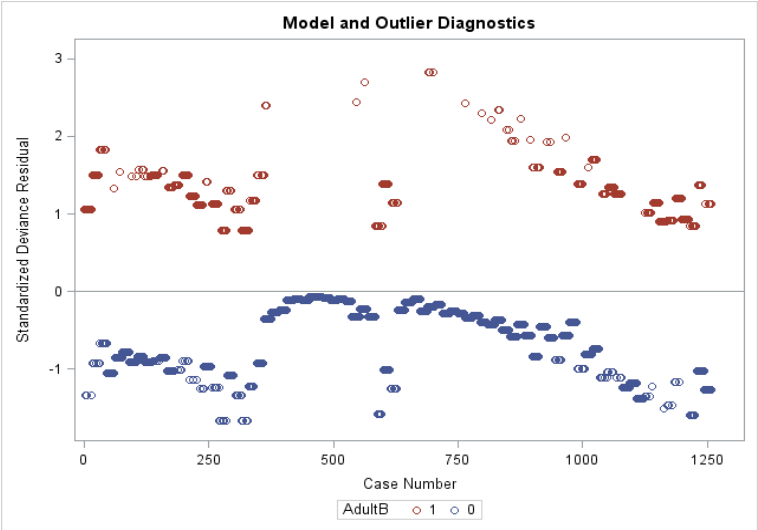

Figure 2.6: Plot of standardized deviance residuals against case numbers for model M_2 after deleting cases 457 and 476.

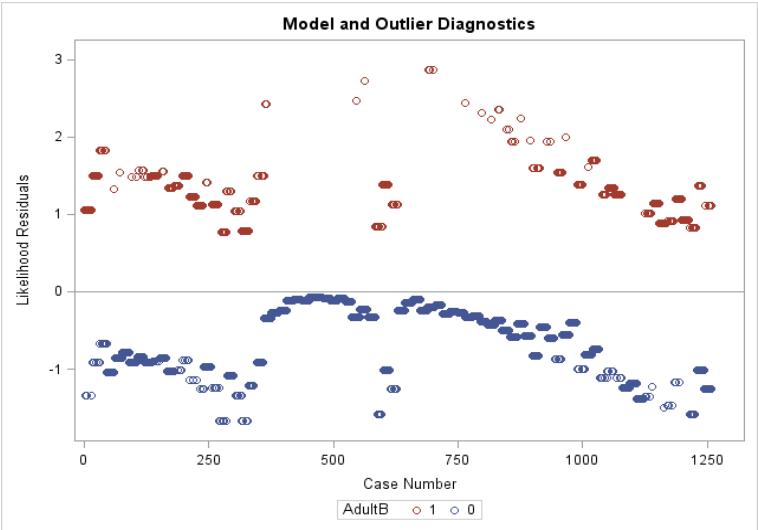

Figure 2.7: Plot of likelihood residuals against case numbers for model M_2 after deleting cases 457 and 476.

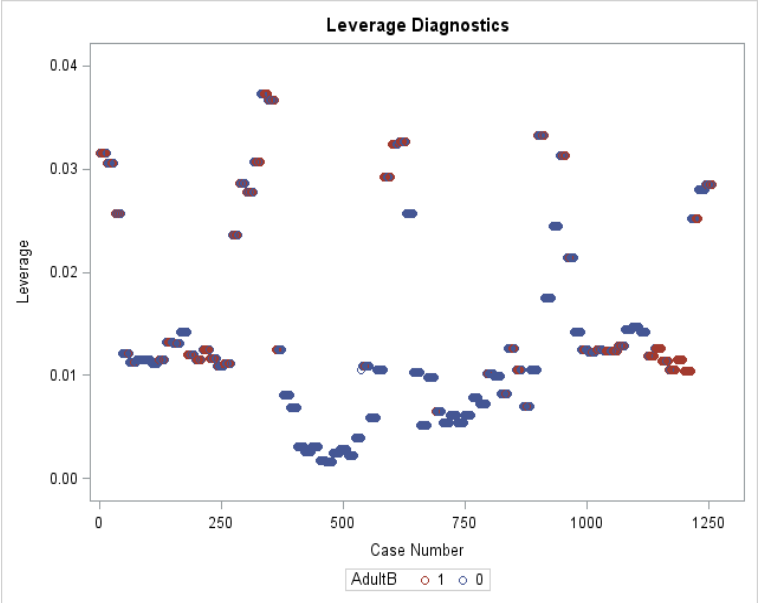

Figure 2.8: Plot of the values of leverage against case numbers for model M_2 after deleting cases 457 and 476.

Table 2.6: Effects tests in model M_2 after deleting cases 457 and 476.

Effect	DF	Wald	p-vale
Site	3	64.0659	< 0.0001
Time	1	20.8782	< 0.0001
Time×Site	3	65.2228	< 0.0001
Time2	1	14.0514	0.0002
Time2×Site	3	60.2016	< 0.0001
Treatment	2	16.7210	0.0002
Time×Treatment	2	13.2256	0.0013
Time2×Treatment	2	11.9576	0.0025

Table 2.7: Parameter estimates, standard error, Wald χ^2 p-value for model M_2 after deleting cases 457 and 476.

Parameter	Estimate	Standard error	Wald χ^2	p-value
Intercept	-4.2208	0.7819	29.1382	< 0.0001
Site (Ref: Ochinga)				
Dudi	2.4127	0.7661	9.9189	0.0016
Khumus	7.5269	1.0143	55.0721	$< .0001$
Marenyo	-2.3873	1.6503	2.0927	0.1480
Time	1.8967	0.4151	20.8782	$< .0001$
Time×Site				
Dudi	-1.4655	0.4204	12.1537	0.0005
Khumus	-6.4572	0.8163	62.5703	$< .0001$
Marenyo	-0.4653	0.7710	0.3643	0.5461
Time2	-0.1814	0.0484	14.0514	0.0002
Time2×Site				
Dudi	0.1635	0.0503	10.5732	0.0011
Khumus	0.7684	0.1008	58.1523	$< .0001$
Marenyo	0.0785	0.0837	0.8792	0.3484
Treatment (Ref: WSS)				
CRSS	3.1012	0.7668	16.3573	$< .0001$
SSTV	-0.6969	0.4325	2.5957	0.1072
Time×Treatment				
CRSS	-1.5496	0.4281	13.1037	0.0003
SSTV	0.0699	0.2444	0.0818	0.7749
Time2×Treatment				
CRSS	0.1729	0.0510	11.5125	0.0007
SSTV	0.0600	0.0514	1.3609	0.2434

Chapter 3

Generalized Linear Mixed models for binary response variables

3.1 Introduction

The common features of the generalized linear models (GLMs) and hence the logistic regression models are that their linear, i.e. systematic components contain terms know as fixed effects (Collett, 2003) and the observations are assumed to be independent. However, when the binary data, for example, are from studies which have hierarchic or multilevel structure or from longitudinal studies, where repeated observations on the same subjects collected through time, the logistic regression in the GLM sense is no longer a valid method. This is due to the fact that a hierarchical design structure may induce correlation between or within cluster and likewise the repeated measurements in a longitudinal study may also induce within subject correlation. Two of the statistical methodologies that can be used to account for this correlation for the longitudinal data are the generalized linear mixed models (GLMMs) and generalized estimating equation (GEE). In this chapter we consider GLMMs and GEE is the topic of Chapter 4.

In the present chapter the generalized linear mixed model and the estimation of its parameters are briefly discussed. Specifically, the GLMM is described in Section 3.2 and the

random effects logistic regression model as a special case of GLMM is discussed in Section 3.3. Finally, possible software for such analysis are briefly discussed in Section 3.4.

3.2 Generalized Linear Mixed models for binary response variables

Generalized linear mixed models are commonly used as an extension of the generalized linear models and they allow for correlated responses through the inclusion of random effect terms in the linear component (or mean structure of GLM), random coefficients and covariance patterns (McCulloch and Searle, 2001). The random effects incorporate correlation between the repeated observations within each cluster and variation between clusters, resulting in GLMMs (Wu, 2010). It is assumed that correlation arises among repeated observations within a given cluster because of the shared random effects, but these repeated observations are assumed to be conditionally independent given the random effects (Wu, 2010).

Let $\mathbf{y} = (y_{i1}, \ldots, y_{in_i})$ be the n_i repeated observations of the response within subject or cluster i, $i = 1, \ldots, K$. We assume that, conditioning on the random effects $\mathbf{b}_i$, the repeated measurements $\mathbf{y} = (y_{i1}, \ldots, y_{in_i})$ are independent and each follows a distribution in the exponential family. Following McCulloch and Searle (2001), a general GLMM can be written as

$$g(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b}_i, \qquad j = 1, \ldots, n_i;, \quad i, \ i = 1, \ldots, K \tag{3.1}$$

where $\mu_{ij} = E(y_{ij}|\boldsymbol{\beta}, \mathbf{b}_i)$ is the conditional mean, $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are vectors of covariates for observation j in subject i, $\boldsymbol{\beta}$ is a p-dimensional vector of regression coefficients, and $\mathbf{b}_i$ is a q-dimensional vector of random effects which are assumed to follow a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{D}$, that is, $\mathbf{b}_i \sim \mathrm{N}_p(\mathbf{0}, \mathbf{D})$, $g(.)$ is a known function which links the mean and the linear form of predictors called the link function (McCullogh and Searle, 2001). The total number of observations is given by

$N = \sum_{i=1}^{K} n_i$. Since the model in (3.1) is specified based on the conditional mean, the model sometimes is called conditional models or subject-specific model (Wu, 2010, page 63).

3.2.1 Statistical inference for a GLMM

Statistically inference on a GLMM is generally based on the likelihood method. In the GLMM given in expression (3.1), the marginal distribution for $\mathbf{y}_i$ is given by

$$f(\mathbf{y}_i|\boldsymbol{\beta},\ \mathbf{D}) = L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{y}_i) = \int \prod_{j=1}^{n_i} [f(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\beta}, \phi, \mathbf{b}_{ij})\ f(\mathbf{b}_i|\mathbf{D})] d\,\mathbf{b}_i. \tag{3.2}$$

which usually does not have an analytic or closed-form expression since the model is nonlinear in the random effects $\mathbf{b}_i$. The likelihood function for $\boldsymbol{\beta}$ and $\mathbf{D}$ is a product of K terms in expression (3.2) and is given by

$$L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{y}) = \prod_{i=1}^{K} \int L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{y}_i) \tag{3.3}$$

where $\mathbf{y}$ is the total response vector over all the subjects. This likelihood function involves an intractable multi-dimensional integral with respect to the random effects $\mathbf{b}_i$ which is dues to the presence of K integrals over a q-dimensional random effects (Molenberghs and Verbeke, 2005; Wu, 2010). The most commonly used inference for GLMMs include methods based on Gauss-Hermite quadrature or Monte Carlo integration techniques, Expectation-Maximization (EM) algorithms, and approximate methods based on Taylor approximations or Laplace approximations (Molenberghs and Verbeke, 2005; Lee, Nelder and Pawitan, 2006).

Approximate methods which avoid integrations are computationally much more efficient have been suggested in the literature (e.g. Molenberghs and Verbeke, 2005; Wu, 2010), however these methods can be computationally intensive when the dimension of the random effects is large. These numerical methods include Gaussian Quadrature and Adaptive Gaussian Quadrature where the former is less precise but less time consuming and the latter is precise but much more time consuming (Molenberghs and Verbeke, 2005).

It is common that interest is on estimating parameters in the marginal distribution of $\mathbf{y}_i$, however it is also necessary to obtain estimates for the random effects. These reflect between-cluster specific variability, which makes them more helpful for detecting special cases, such as outlying observations or a group of individuals evolving differently. These estimates are needed when interest is in the prediction of subject-specific evolutions. Estimation of the random effects will be based on their posterior distributions and obtained estimates are called the Empirical Bayes (EB) estimates. However, in this dissertation we are interested in estimating the fixed effects in the GLMMs. The literature on GLMMs is extensive and the basic results can be found in the Molenberghs and Verbeke (2005), Jiang (2007), McCulloch, Searle, and Neuhaus (2008) books.

3.3 Logistic regression with subject-specific intercept

The logistic regression model with varied subject-specific intercepts is the simplest generalized linear mixed model. Standard logistic regression model applied to $\{y_{ij}, \mathbf{x}_{ij}\}$ in Chapter 2 implicitly assumes that the presence/absence of adults of the Chrysomelid beetle is constant across the *trees within a site*. Since, for example, soil characteristics may vary within a site, if trees are samples from different farms / plots in the same site or treatment effect may be different across the farms / plots within a site, this assumption may be wrong. Therefore, the assumption that the presence/absence of adults of the Chrysomelid beetle is constant across the *trees within a site* would lead to improper conclusions. One of the possible approach is to assume that intercepts in the logistic regression differ from *tree to tree within a site*.

Let y_{ij} be a binary dependent variable that codes the presence/absence of adults of the Chrysomelid beetle in the jth sample from the ith tree, where $y_{ij} = 1$ if present and $y_{ij} = 0$ if absent, and $\mathbf{x}_i$ be the corresponding $p \times 1$ vector of explanatory variables, $i = 1, \ldots, K$. It is assumed that the number of observations $N = \sum_{i=1}^{K} n_i$ is equal to or greater than the number of explanatory variables, i.e. $N \geq p$ and that the N vectors $\mathbf{x}_{ij}$ has full rank, i.e.

$rank(\mathbf{x}_{ij}) = p$, $i = 1, \ldots, K$. Let n_i denote the number of samples taken from the ith tree. The logistic regression with subject-specific intercept is therefore given by

$$P(y_i = 1 \,|\, b_i) = \mathbf{x}'_{ij}\,\boldsymbol{\beta}, \quad i = 1, \ldots, K, j = 1, \ldots, n_i \tag{3.4}$$

where $\mathbf{x}_{ij}$ is the $p \times 1$ vector of explanatory variables. Since b_i is the intercept term, we assume in this section that $\mathbf{x}_{ij}$ does not have a constant component. However, if we are interested in comparing the presence/absence of adults of the Chrysomelid beetle across the trees within the site adjusted by the treatment and time effects, then the intercept term can be considered as a parameter of interest. Observe in model (2.7) that, there is no subject-specific explanatory variables as this may cause non-identifiability problem in the model because of the subject-specific intercept b_i consumes all variation due to the presence of subject-specific explanatory variables (Demidenko, 2013, page 356).

3.3.1 Assumptions on intercept term b_i

The two assumptions on intercept term b_i, namely fixed and random effects, may lead to two different statistical models.

1. *Fixed effects model / fixed intercept*: If $\{b_i\}$ are assumed to be fixed then they will be estimated as unknown parameters with $\boldsymbol{\beta}$. In this case, they are sometimes called nuisance parameters.

2. *Random intercepts model*: If $\{b_i\}$ are assumed to be random, $b_i = \beta_0 + u_i$, $\{u_i\}$ are independently identically distributed (IID) random variables with zero mean and variance σ_b^2 and β_0 is the population-averaged parameter (Demidenko, 2013). Here it is assumed that $\{y_{ij}\}$ are independent conditional on u_i.

According to Demidenko (2013) the disadvantage of the fixed effects model is that it leads to increasing the number of additional parameters, which is equal to the number of subjects. Its advantage is that it does not require special methods of estimation as one can easily

reduce the model to standard logistic regression by introducing dummy variables for each subject. However, if the number of subjects is large, one faces the problem of a large number of nuisance intercepts but still this problem can be handled using conditional logistic regression (Prentice, 1988). This will not be considered in the dissertation because the number of subjects (i.e. trees) is not large, in fact how 'large is large' is not clearly defined in the literature. In contrast, the random intercepts model has only two parameters β_0 and σ_b^2 in addition to $\boldsymbol{\beta}$ (where $\boldsymbol{\beta}$ here is without the intercept parameter),which are estimated by the GLMM approach.

The advantage of the random effects model is that it is more flexible and includes the fixed effects model as an extreme case when the variance of the random effect σ_b^2 goes to infinity. The proof of this is given in (Demidenko, 2013, pages 356-357). However, according to Breslow and Clayton (1993), ignoring randomness of the intercept in the logistic regression model leads to attenuation the regression coefficients. In the following section the random effects logistic regression model is discussed and the random intercept model follows as a special case of the former.

3.4 Random effects logistic regression model

Let y_{ij} be a binary dependent variable that codes the presence/absence of adults of the Chrysomelid beetle in the jth sample from the ith tree, where $y_{ij} = 1$ if present and $y_{ij} = 0$ if absent and $\mathbf{x}_{ij}$ be the corresponding $p \times 1$ vector of explanatory variables or covariates, $i = 1, \ldots, K$. It is assumed that the number of observations N is equal to or greater than the number of explanatory variables, i.e. $N \geq p$ and that the N vectors $\mathbf{x}_{ij}$ have full rank, i.e. $rank(\mathbf{x}_i) = p$, $i = 1, \ldots, K$. Let n_i denote the number of samples taken from the ith tree. The total number of observations is given by $N = \sum_{i=1}^{K} n_i$.

It is assumed that

$$y_{ij}|\mathbf{b}_i \sim Bernoulli(\pi_{ij})$$

where $\pi_{ij} = P(y_{ij} = 1)$ is the probability that at least one adult of the Chrysomelid beetle is present in the jth sample from the ith tree. The general random effects logistic regression model is given by

$$Log\left[\frac{P(y_{ij}=1|\mathbf{b}_i)}{1-P(y_{ij}=1|\mathbf{b}_i)}\right] = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b}_i. \tag{3.5}$$

Using the logit link the model can also be written as

$$P(y_{ij}=1|\mathbf{b}_i) = \pi_{ij} = g^{-1}(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b}_i)$$

where $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are vectors of covariates for observation j in subject i, $\boldsymbol{\beta}$ is a p-dimensional vector of regression coefficients, $\mathbf{b}_i$ is a q-dimensional vector of random effects expressing the the presence/absence of adults of the Chrysomelid beetle in a sample of the ith subject and $g(.)$ is the link function. Usually the random effects are assumed to follow a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{D}$, that is, $\mathbf{b}_i \sim \mathrm{N}_p(\mathbf{0}, \mathbf{D})$.

Conditional on the random effects, the likelihood contribution of samples from subject i are assumed to be independent. To obtain the marginal likelihood contribution for the ith subject, the random effects have to be integrated out. The marginal likelihood for the ith subject is given by:

$$L(\boldsymbol{\beta}, \mathbf{D}|\mathbf{y}_i) = f(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{D}) = \int \prod_{j=1}^{n_i} P(y_{ij}=1|\boldsymbol{\beta}, \mathbf{b}_i) f(\mathbf{b}_i|\mathbf{D}) d\mathbf{b}_i \tag{3.6}$$

with $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})$. The expression in (3.6) is a special case of (3.2). Here as the response variable y_{ij} takes values 0 or 1, the parameter ϕ related to overdispersion is not needed. Hence using expression (3.3) the total marginal likelihood function for $\boldsymbol{\beta}$ and $\mathbf{D}$ is a product of K terms in expression (3.6) and is given by

$$L(\boldsymbol{\beta}, \mathbf{D}|\mathbf{y}) = \prod_{i=1}^{K} L(\boldsymbol{\beta}, \mathbf{D}|\mathbf{y}_i) \tag{3.7}$$

where $\mathbf{y}$ is the total response vector over all the subjects.

3.4.1 Logistic regression model with random intercept

The random intercept model is commonly used to describe longitudinal data. It is a special case of the random effects logistic regression model with the first component of the $p \times 1$ vector $\mathbf{x}_{ij}$ equal to 1, vector $\mathbf{z}_{ij} = 1$ and the random vector $\mathbf{b}_i = b_i$, i.e. subject-specific intercept. The model with random intercept therefore can be rewritten as

$$Log\left[\frac{P(y_{ij}=1\,|\,b_i)}{1-P(y_{ij}=1|b_i)}\right] = \mathbf{x}'_{ij}\boldsymbol{\beta} + b_i \tag{3.8}$$

where the term b_i is the random component of the intercept, and the other terms in the model are the same as those introduced in model (3.6). The random terms b_i, $i = 1, \ldots, K$, are assumed to be independently normally distributed with mean zero and variance σ_b^2. This model implies that observations with in the same subject are correlated.

To obtain the parameter estimates (and their precision) using the maximum likelihood approach, we have to maximize the total marginal likelihood (3.8) which results in the Maximum Likelihood Estimate (MLE). The Standard Error (SE) of the MLE can be obtained from the inverse of the information matrix evaluated at the MLE.

3.4.2 Software

Gaussian and adaptive Gaussian quadrature are implemented in SAS procedures MIXED, GLIMMIX and NLMIXED; in STATA procedure GLAMM; and package lme4 in R. Further, the PQL method is implemented in R package MASS. In this dissertation, the PROCEDURE GLIMMIX (SAS Institute Inc. 2013. SAS/STAT® 13.1 Users Guide) was used in fitting the random effects logistic regression model.

3.5 Analysis of adults of the Chrysomelid beetle data using GLMMs approach

In all GLMMs (logistic regression with subject-specific random effects and logistic regression models with random intercept) the random effects are assumed to be normally distributed. Adaptive Gaussian Quadrature with 100 quadrature points are used to numerically approximate the likelihood and the Newton-Raphson method will be used as an optimization technique. The combination of these methods produce the most reliable results (Molenberghs and Verbeke, 2005). First the logistic regression with subject-specific random intercept is presented, that is

$$y_{ij} \sim Bernoulli(\pi_{ij})$$

$$logit(\pi_{ij}) = \beta_0 + \beta_1 \times Site + \beta_2 \times Time + \beta_3 \times Site \times Time + \beta_4 \times Treatment + \beta_5 \times Site \times Treatment + \beta_6 \times Time \times Treatment + \beta_7 \times Time^2 + \beta_8 \times Time^2 \times Site + \beta_9 \times Time^2 \times Treatment + b_i,$$

$$b_i \sim N(0, \sigma_b^2).$$

If the variance σ_b^2 is small, then the contribution from b_i is also rather small and all the trees in a given site will have a similar logistic curve. Otherwise, if σ_b^2 is relatively large (i.e. the test on $\sigma_b^2 = 0$ is significant), then each tree in a given site will have different intercepts. Note the use of tree within a site as a random intercept has advantage that it introduces the compound symmetrical correlation structure. This implies that the probability of adults of the Chrysomelid beetle present on a tree within a particular site is correlated to other tree on the same site.

3.5.1 Logistic regression model with subject-specific random intercept

Using the GLM approach as was discussed in Chapter 2, explanatory variables that are considered in the analysis are Site, Time, Treatment, $Time2 = Time \times Time$ and the first

order interactions of these four variables. As in the case of GLM,the analysis showed that the interaction effect between site and treatment is non-significant(p-value=0.5117) based on the fixed effect test and the type III test of fixed effects. Hence the interaction term site*treatment was considered for removal. Therefore a model without the non-significant interaction term was fitted to assess the significance of the rest of the model terms. The solutions for fixed effects and the Type III test of fixed effects results for the second model are presented in Tables 3.1 and 3.2, respectively. However, two of the regression coefficients namely, coefficients for treatment $SSTV$ and the interaction level $Time \times SSTV$ are inconsistent in sign compared to those from GLM (see Table 2.7). The estimated random intercept variance, i.e. tree specific variance, is $\hat{\sigma}_b^2 = 0.2621$ (SE = 0.1200) for the model with Site by Treatment interaction and $\hat{\sigma}_b^2 = 0.2583$ (SE = 0.1186) for the model without Site by Treatment interaction. In both models, with and without Site by Treatment interaction, the random intercept was found to be significant with likelihood ratio test (LRT) = 1102.63 and 1107.65, respectively and have equal $p-value = 0.0002$ for the mixture of Chi-square test which are highly significant at 5% level.There are various covariance structures that are applicable to the GLMM procedure (i.e AR (1),CS,UN etc.).The simplest form of these is the unstructured covariance structure which contains no additional constraints on its element. Another common covariance structure is the variance component in which the random effect has its own variance and all covariance in D matrix are zero(West et.al ,2007).From the analysis all covariance structure gave the same variance components. As a result the same model was obtained irrespective of the covariance structure selected.

Table 3.1: Results from subject-specific random intercept model without Site by Treatment interaction

Effect	Estimate	Standard Error	p-value
Intercept	-4.421	0.815	< 0.0001
Site (Ref: Ochinga)			
Dudi	2.530	0.809	0.0028
Khumus	7.853	1.063	< 0.0001
Marenyo	-2.450	1.690	0.1528
Time	1.984	0.428	< 0.0001
Time×Site			
Dudi	-1.536	0.433	0.0004
Khumus	-6.717	0.840	< 0.0001
Marenyo	-0.482	0.785	0.5395
Time2	-0.1895	0.050	0.0002
Time2×Site			
Dudi	0.171	0.052	0.0010
Khumus	0.799	0.104	< 0.0001
Marenyo	0.080	0.085	0.3473
Treatment (Ref: WSS)			
CRSS	3.246	0.787	< 0.0001
SSTV	1.412	0.795	0.0760
Time×Treatment			
CRSS	-1.625	0.439	0.0002
SSTV	-0.725	0.444	0.1025
Time2×Treatment			
CRSS	0.182	0.052	0.0005
SSTV	0.062	0.053	0.2383

Table 3.2: Type III tests of fixed effects for subject-specific random intercept model without Site by Treatment interaction

Effect	p-value for Chi-square	p-value for F
Site	< 0.0001	< 0.0001
Time	0.0006	0.0006
Time×Site	< 0.0001	< 0.0001
Time2	< 0.0001	< 0.0001
Time2×Site	< 0.0001	< 0.0001
Treatment	0.0002	0.0002
Time×Treatment	0.0010	0.0011
Time2×Treatment	0.0019	0.0020

3.5.2 Logistic regression model with subject-specific random effects

We have also considered a random effects model by introducing subject-specific random slopes for the time trend. The analysis obtained using the same SAS code used for the subject-specific random intercept by replacing RANDOM INTERCEPT option in Proc GLIMMIX with RANDOM INTERCEPT TIME. Different structures for the random-effects covariance matrix (which is a 2×2 matrix), such as unstructured or structure which assumes equal variance for intercepts and slopes or that assume independent intercepts and slopes, can be defined using the option "type = " under the RANDOM option of Procedure GLIMMIX. The estimated random intercept and slope variances are equal to $\sigma^2_{b_0} = 0.1467$ (SE = 0.1291) and $\sigma^2_{b_0} = 0.9335$ (SE = 0.2413), respectively. The asymptotic covariances between the two random intercept and slope is -0.0086. The tests of covariance parameters based on the residual pseudo-likelihood is statistically nonsignificant (p-value = 0.3710) for the test on variance-covariance matrix of the random effects equal to diagonal matrix with different diagonal elements (i.e. variances of the random effects are different). In total 420 ($=4 \times 7 \times 15$) random-effects models are generated for 15 trees in each of four sites and

at seven time points. However, only two random intercepts namely, for tree number 4 in Site Ochinga (p-value = 0.0336) and for tree number 11 in Site Ochinga are statistically significant at 5% level. The introduction of a random time effect in the model increased the complexity of the model,as a result there were computational issues. According to West et.al (2007), computational issues may arise depending on the nature of the longitudinal data (i.e. based on the similarity of observations with a subject), and on the optimization technique (i.e. convergence to a point whereby the covariance parameter lies closer or outside the parameter space. The covariance parameter estimate can be zero, and this is an indication of the covariance parameter breaching a lower bound constraint on the method, which in most cases is zero. As a result the singularity issue arises.) Due to these the D matrix might not be positive definite resulting in inadequate results. In their book" Linear Mixed Models: a practical guide using statistical software" a few corrective measure can be taken into account when there exists a singularity problem, and these include amongst other:

1. Providing initial covariance parameter estimate: This can be established from other methods that are similar to the method that is in use.(e.g. If a problem occurs in GLIMMIX ,the mixed and nlmixed can be used to derive the initial values)

2. Remove unnecessary random effect: These include insignificant terms as well as higher order terms. The inclusion of higher order terms introduces a lot of complexity and the computation of the model with these terms might also take longer. Therefore removing them allows one to have a sensible model that is well formulated.

As in any statistical modelling process one of the of the model characteristic that we should consider is the simplicity of a model, i.e. parsimonious, therefore we suggest not to consider this fitted model. Later in Chapter 4, the results that we obtained from GEE supports our suggestion. Further, the results from this analysis did not affect the statistical significance of the effects which we have obtained in subject-specific random intercept model.

Chapter 4

Generalized Estimating Equations

4.1 Introduction

The generalized linear model (GLM) method is based on the maximum likelihood theory of independent observations and an assumption about the distribution of response observations. The generalized linear mixed model (GLMM) method is based on extending the fixed effects GLM to include random effects and covariance patterns, as it was discussed in Chapter 3. Unlike the GLM and GLMM methods, the generalized estimating equation (GEE) method is based on the quasi-likelihood theory and no assumption is made about the distribution of response observations (Liang and Zeger, 1986).

In the present chapter the generalized estimating equation and its feature are discussed in Sections 4.2 and 4.3, respectively. The estimation of working correlation parameters and dispersion parameter are discussed in Section 4.4. The selection of working correlation structure and summary of estimation algorithm for GEE method are briefly discussed in Sections 4.5 and 4.6, respectively. Finally, software for such analysis are briefly discussed in Section 4.4 and this followed by GEE application to adults of the Chrysomelid beetle data.

4.2 Model and estimation of regression parameters

The generalized estimating equations (GEE) was first proposed by Liang and Zeger (1986) for clustered and repeated data, which require only the correct specification of the univariate marginal distribution provided one is willing to adopt working assumption about the association structure. Since their publications several approaches have been developed to improve the technique. The literature on GEE is extensive and the basic results can be found in Ziegler et al. (1996), Greene (1997), Hardin and Hilbe (2003), Fitzmaurice, Laird and Ware (2004), and Molenberghs and Verbeke (2005, Chapter 8).

Here the GEE for longitudinal data is defined. Let y_{ij} be a binary outcome random variable and $\mathbf{x}_{ij}$ be a $p \times 1$ vector of fixed covariate at time j for subject i, where $i = 1, \ldots, K$ and $j = 1, \ldots, n_i$ with the first two moments defined as $E(y_{ij}) = \mu(\mathbf{x}'_{ij}\boldsymbol{\beta})$ and $var(y_{ij}) = \nu(\mathbf{x}'_{ij}\boldsymbol{\beta})$ where μ and ν are known functions. For binary response model $\nu = \mu(1-\mu)$. It is assumed that observations between subjects are independent but observations within subjects are dependent. We may ignore this dependence and still obtain consistent and asymptotically normally distributed estimates of $\boldsymbol{\beta}$ (Liang and Zeger, 1986). However, to improve the efficiency, one needs to account for the correlation of observations within subject and hence to specify the covariance matrix $cov(\mathbf{y}_i)$ where $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})'$ denotes the response vector for subject $i = 1, \ldots, K$. Assume

$$y_{ij}|\,\mathbf{x}_{ij} \sim \; Bernoulli(\mu_{ij}) \quad \text{and} \quad log\left[\frac{\mu_{ij}}{1-\mu_{ij}}\right] = \mathbf{x}'_{ij}\boldsymbol{\beta}$$

and therefore the mean and variance of y_{ij} are equal to μ_{ij} and $\mu_{ij}(1-\mu_{ij})$, respectively. Following Fitzmaurice, Laird and Ware (2004), in the matrix notation the systematic part of the model is given by

$$\eta_i = \mathbf{X}_i\boldsymbol{\beta} \tag{4.1}$$

where $\mathbf{X}_i$ is an $n_i \times p$ matrix with a jth row $\mathbf{x}_{ij}$ and $\boldsymbol{\beta}$ is a vector of regression parameters. The relationship between the conditional mean with respect to the explanatory variables $\mathbf{X}_i$ and the systematic component has the same form as in GLM models and hence

$$E(\mathbf{y}_i|\mathbf{X}_i) = \mu_{ij} \quad \text{and} \quad \eta(\mu_{ij}) = \mathbf{X}_i\boldsymbol{\beta} \tag{4.2}$$

where $\eta(\mu_{ij})$ is the link function.

The next step for GEE analysis is to specify an association structure between two observations y_{ij} and $y_{ij'}$ of the same subject. One can use different working correlation assumptions, for example independence (equivalent to logistic regression), exchangeable (equal correlation among all observations from the same subject, similar to the standard logistic model) and an auto-regressive correlation of order 1, AR(1) structure (i.e. serial dependence).

The estimates for regression parameters in $\boldsymbol{\beta}$ are obtained by solving

$$\sum_{i=1}^{N} \mathbf{D}_i\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\alpha})(\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0} \tag{4.3}$$

where $\mathbf{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}'$ is an $p \times n_i$ matrix of first-order derivatives of the mean vector $\boldsymbol{\mu}_i$ (i.e. the mean response as a function of covariates) with respect to $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\alpha}) = \mathbf{A}_i^{1/2}\mathbf{R}(\boldsymbol{\alpha})\mathbf{A}_i^{1/2}$ is the variance-covariance matrix of the $n_i \times 1$ vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})'$ and the digonal matrix $\mathbf{A}_i$ is a know function of $\boldsymbol{\beta}$. The working correlation matrix $\mathbf{R}(\boldsymbol{\alpha})$ describes within-subject dependencies and it is assumed that its elements are known functions of a p-dimensional parameter $\boldsymbol{\alpha}$, for example in an exchangeable (compound symmetry) correlation structure all elements except the diagonal are α = constant. The $\boldsymbol{\Sigma}(\boldsymbol{\alpha})$ is usually modelled by adopting a common form for the variance based on an appropriate member of the exponential family. Therefore, when the distribution of the response variable is a member of the exponential family we do not need all the details of the probability distribution in order to estimate the regression parameters $\boldsymbol{\beta}$, only its mean and variance. However, we have to determine what form $\mathbf{R}(\boldsymbol{\alpha})$ takes, this means that we choose a correlation structure (e.g. exchangeable correlation or auto-regressive correlation) that closely describes what is observed in the response data.

Liang and Zeger (1986) show that when the marginal mean μ_i has been correctly specified as $f(\mu_i) = \mathbf{X}_i\boldsymbol{\beta}$ and when mild regularity conditions hold, the estimator $\hat{\boldsymbol{\beta}}$, which is obtained

as a solution of (4.3), is consistent and asymptotically normally distributed with mean $\boldsymbol{\beta}$ and asymptotic variance-covariance matrix, called a robust estimate,

$$var(\hat{\boldsymbol{\beta}}) = \mathbf{M}_0^{-1}\mathbf{M}_1\mathbf{M}_0^{-1} \tag{4.4}$$

where

$$\mathbf{M}_0 = \sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \boldsymbol{\beta}} \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\alpha}) \frac{\partial \mu_i}{\partial \boldsymbol{\beta}'} \tag{4.5}$$

and

$$\mathbf{M}_1 = \sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \boldsymbol{\beta}} \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\alpha}) Var(\mathbf{y}_i) \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\alpha}) \frac{\partial \mu_i}{\partial \boldsymbol{\beta}'} \tag{4.6}$$

$Var(\mathbf{y}_i)$ is the true covariance matrix of $\mathbf{y}_i$. Whether or not the working correlation structure is correctly specified, point estimates of $\boldsymbol{\beta}$ and standard errors based on (4.4) are asymptotically correct. These standard errors were called robust by Liang and Zeger (1986), while the variance estimator in (4.4) is sometimes referred to as the sandwich estimator. However, to avoid confusion with methods from robust statistics, the terms *empirically corrected* variance and standard errors are commonly used in literature. The inverse of expression (4.5), i.e. $\mathbf{M}_0^{-1}$, was referred to as the *naive* estimator by Liang and Zeger (1986) but currently the term *model based* estimator is more common.

4.3 Some features of GEE

- *Quasi-likelihood*: Note from the above discussion that in the GEE approach no distribution is specified and the correlation matrix need not coincide with the true one. Therefore, (4.3) sometimes is viewed as a quasi-likelihood approach where a complete data distribution is not specified.

- *Overdispersion parameter*: Apart from a working correlation matrix, it is possible to incorporate an overdispersion parameter as well. Let ϕ be the additional overdispersion parameter. Then this can be introduced in (4.3) replacing $\mathbf{A}_i^{1/2}\mathbf{R}(\boldsymbol{\alpha})\mathbf{A}_i^{1/2}$ by

$$\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\alpha}, \phi) = \phi\, \mathbf{A}_i^{1/2}\mathbf{R}(\boldsymbol{\alpha})\mathbf{A}_i^{1/2}. \tag{4.7}$$

However, overdispersion cannot occur when un-grouped binary data are being modelled (Collett, 2003, page 170) therefore this parameter is not going to be estimated for adults of the Chrysomelid beetle data.

- *Full likelihood*: Observe that when $\mathbf{R}_i(\boldsymbol{\alpha})$ would be correctly specified, $Var(\mathbf{y}_i) = \boldsymbol{\Sigma}_i(\boldsymbol{\alpha})$ in (4.6) and then $\mathbf{M}_0 = \mathbf{M}_1$. Hence the expression in (4.4) reduces to $\mathbf{M}_0^{-1}$ which corresponds to full likelihood. Thus, when the working correlation structure is correctly specified (4.4) reduces to full likelihood.

- According to Molenberghs and Verbeke (2005) the empirically corrected standard error is the one to be used not the model based. When both standard errors are far apart, this can be seen as an indication for a poor choice of the working correlation structure.

4.4 Estimation of working correlation parameters and dispersion parameter

Liang and Zeger (1986) proposed moment based estimates for the working correlation parameters $\boldsymbol{\alpha}$ and the overdispersion parameter ϕ. In the estimation process, we first define residuals as

$$e_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{\nu(\mu_{ij})}}. \tag{4.8}$$

The residual e_{ij} is dependent on $\boldsymbol{\beta}$ through μ_{ij} and $\nu(\mu_{ij})$, the variance at time j and hence the jth diagonal element of $\mathbf{A}_i$. Common choices for the working correlation assumptions in generalized estimating equations and moment-based estimators are (Liang and Zeger, 1986):

i. *Independence correlation structure*: $Corr(y_{ij}, y_{ik}) = 0$ so there is no estimator required. The assumption of independence working correlation structure may scarifies the benefits of using GEE because it does not account for within-subject correlation.

ii. *Exchangeable correlation structure*: $Corr(y_{ij}, y_{ik}) = \alpha$ and its moment-based estimator is given by

$$\hat{\alpha} = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{n_i(n_i-1)}\sum_{j\neq k} e_{ij}e_{ij}.$$

The exchangeable working correlation structure allows for constant correlations between any two observations within a subject for all time points and needs only one parameter α to be estimated.

iii. *AR(1)*: $Corr(y_{ij}, y_{ik}) = \alpha^{|j-k|}$ and its moment-based estimator is given by

$$\hat{\alpha} = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{(n_i-1)}\sum_{j\leq n_i-1} e_{ij}e_{i,j+1}.$$

The AR(1) correlation coefficients diminish as the absolute time difference between two observations, i.e. the exponent $\mid j-k \mid$, within a subject increases.

iv. *Unstructured correlation structure*: $Corr(y_{ij}, y_{ik}) = \alpha_{jk}$ and its moment-based estimator is given by

$$\hat{\alpha}_{jk} = \frac{1}{K}\sum_{i=1}^{K}\sum_{j\leq k} e_{ij}e_{ik}.$$

According to Fitzmaurice, Laird and Rotnitzky (1993), the assumption of unstructured working correlation matrix allows estimation of all possible correlations between within-subject observations and includes then in the estimation of variances.

Similarly, the moment-based estimate of the dispersion parameter is given by

$$\hat{\phi} = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{n_i}\sum_{i=1}^{n_i} e_{ij}^2. \tag{4.9}$$

Note that the independence correlation structure does not introduce additional parameter $\boldsymbol{\alpha}$ and hence there is no overdispersion, parameter estimates $\hat{\boldsymbol{\beta}}$ will not differ from those

obtained from ordinary logistic regression analysis. However, the asymptotic variance-covariance matrix obtained from (4.4) and hence the standard errors will differ from the one obtained with ordinary logistic regression analysis. The latter stemming from the model-based but they are incorrect $\mathbf{M}_1^{-1}$.

4.5 Selection of working correlation structure

According to Molenberghs and Verbeke (2005), independence and exchangeable working assumptions can be used in all applications whether longitudinal, cluster, or otherwise correlated. AR(1) and unstructured are less relevant for clustered data, longitudinal studies with unequally spaced measurements in time, etc. As it is discussed by these authors, there is no price to pay in terms of consistency of asymptotic normality, but there may be efficiency loss when the working correlation structure differs strongly from the true underlying structure.

In general, a correlation matrix $\mathbf{R}(\boldsymbol{\alpha})$ can be selected as follows. Fit a series of models that differ only in the choice of correlation matrix $\mathbf{R}(\boldsymbol{\alpha})$; all of the remaining features of these models are the same. For each model compare the sandwich variance estimates with the model-based variance estimates. The model whose sandwich variance estimates most closely resembles its model-based variance estimates is the one with the best correlation matrix $\mathbf{R}(\boldsymbol{\alpha})$ (Debusho and Sileshi, 2012).

4.6 Estimation Algorithm in GEE approach

The system of equations in (4.3) need $\boldsymbol{\alpha}$ and ϕ to estimate $\boldsymbol{\beta}$, while the moment-based estimates of $\boldsymbol{\alpha}$ and expression (4.7) for ϕ depend on $\boldsymbol{\beta}$. Therefore, the estimation algorithm alternates between $\boldsymbol{\beta}$ estimation from GEE (4.3) and $(\boldsymbol{\alpha}, \phi)$. The estimation algorithm

to fit GEE are based on Liang and Zeger (1986) and solved numerically by iteration that consists of the following steps:

Step 1: For a given $\boldsymbol{\alpha}$ and ϕ (and therefore $\hat{\boldsymbol{\Sigma}_i}(\boldsymbol{\alpha})$, an estimate of $\boldsymbol{\Sigma}_i(\boldsymbol{\alpha})$), obtain an estimate for the regression parameters $\boldsymbol{\beta}$.

Step 2: Given the regression parameters $\boldsymbol{\beta}$, update $\boldsymbol{\alpha}$ and ϕ (and therefore $\hat{\boldsymbol{\Sigma}_i}(\boldsymbol{\alpha})$).

Step 3: Iterate between steps 1 and 2 until convergence.

That is, when $\boldsymbol{\alpha}$ and ϕ are held fixed, we iterate

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k + \left[\sum_{i=1}^{K}\left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}'}\right)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\alpha})\left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}'}\right)'\right]^{-1}\left[\sum_{i=1}^{K}\left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}'}\right)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\alpha})(\mathbf{y}_i - \mu_i)\right].$$

At convergence, the estimated regression parameters are nearly equal to the population parameters, i.e. consistent, and identically normally distributed with mean $\boldsymbol{\beta}$ (Debusho and Sileshi, 2012) and a covariance matrix (4.4). The empirically corrected estimate for $Var(\hat{\boldsymbol{\beta}})$ is obtained from replacing in (4.4) $\boldsymbol{\Sigma}_i(\boldsymbol{\alpha})$ by its estimate $\hat{\boldsymbol{\Sigma}_i}(\boldsymbol{\alpha})$ and $Cov(\mathbf{y}_i)$ by the covariance matrix $(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})'$. The chosen correlation structure $\mathbf{R}(\boldsymbol{\alpha})$ is then used in the covariance matrix, resulting in the sandwich estimate or the robust variance estimate. The diagonal element of this matrix can be used to obtain the standard errors of the regression parameter estimates. For many statistical packages this is the default estimate that is displayed in the standard error column of the summary table of the model.

4.7 Goodness-of-fit for GEE

Recall that generalized estimating equations (GEE) is a non-likelihood based approach. Therefore, the Akaike Information Criterion (AIC) cannot be directly applied since AIC is based on maximum likelihood estimation. Pan (2001) proposed the quasi-likelihood under the independence model criterion (QIC) and its related statistic QIC_u, where QIC_u approximates QIC when the GEE model is correctly specified and they defined as

$$QIC = Q + 2\,trace(\hat{\mathbf{V}}_I\,\hat{\mathbf{V}}_R) \qquad \text{and} \qquad QIC_u = Q + 2\,p$$

where Q is a quasi-likelihood and computed using the working correlation matrix, $\hat{\mathbf{V}}_I = [var(\hat{\boldsymbol{\beta}})]^{-1}$ obtained by fitting an independence model, $\hat{\mathbf{V}}_R = [var(\hat{\boldsymbol{\beta}})]^{-1}$ a modified sandwich estimate of variance (4.4) from the model with the working correlation matrix and p is the number of parameters in $\boldsymbol{\beta}$. Basically QIC_u adds a penalty $2\,p$ to Q. The QIC is further discussed by Hardin and Hilbe (2003) and for detail mathematical derivation and discussion on QIC we suggest to refer Pan (2001) and Hardin and Hilbe (2003). When using QIC or QIC_u to compare two models, the model with the smaller statistic is preferred.

QIC can also be applied to select a working correlation structure in GEE if the goal of selecting a working correlation structure is to estimate $\boldsymbol{\beta}$ more efficiently, although the QIC values may differ slightly. However, in both literatures the authors advise QIC_u should not be used for selecting a working correlation structure.

4.8 Software

We used PROCEDURE GENMOD in SAS (SAS Institute Inc. 2013. SAS/STAT® 13.1 Users Guide) for GEE analysis. There are various packages for GEE analysis in R, *geeglm* function from the *geepack*, *gee* function from the *gee* package and *yags* package.

4.9 Analysis of adults of the Chrysomelid beetle data using GEE method

It is important to realise that the ordinary logistic regression, i.e. GLM, assumes independence of observations including those from the same tree which are separated by a month. However, adults of the Chrysomelid beetle data have longitudinal nature where multiple samples (trees per site) and multiple observations per tree collected at different time points and these observation might be correlated. Ignoring the potential existence of dependence may tend to increase the risk of a Type I error, particularly where within-subject (i.e. tree

Table 4.1: Fit criteria for the four working correlation assumptions

Fit criterion	Independence	AR(1)	Exchangeable
QIC	1143.5709	1143.8956	1143.8432
QIC_u	1143.5219	1143.7024	1143.5526

within a site) correlation is strong. We therefore considered GEE for analysis to include a dependence structure.

Preliminary analyses, GLM and GLMM, have indicated that Site, Treatment, the linear and square effects of Time, and except the Site by Treatment interaction, the other four first order interactions are important for model buildup. Hence for this analysis the logistic regression model is considered, i.e. called the marginal model

$$y_{ij} \sim Bernoulli(\pi_{ij})$$

$$logit(\pi_{ij}) = \beta_0 + \beta_1 \times Site + \beta_2 \times Time + \beta_3 \times Site \times Time + \beta_4 \times Treatment + \beta_5 \times Time \times Treatment + \beta_6 \times Time^2 + \beta_7 \times Time^2 \times Site + \beta_8 \times Time^2 \times Treatment.$$

There are seven follow-up measurements for each and every tree in a site. Assuming the seven follow-up measurements as equally spaced time points, we have considered three working correlation structures namely, independence, exchangeable and autoregressive (AR(1), i.e. serial dependence). The unstructured working correlation is not considered because of convergence problem. Table 4.1 presents the fit criteria, QIC and QIC_u, for the three working correlation assumptions.

The first step in GEE analysis is to choose the best working correlation structure. This helps to estimate the regression coefficients β_i's more efficiently (Pan, 2001, page 122). As it is discussed earlier, GEE models are generally robust to misspecification of the correlation structure (Liang and Zeger, 1986). However, incorrect specification can affect the efficiency of parameter estimates. The independence correlation structure has the smallest QIC and

thus the independence structure is chosen as the preferred correlation matrix (Table 4.1). Table 4.2 displays parameter estimates and standard errors of GEE models.

It is clear from Table 4.2 that all analyses agree closely in terms of parameter estimates, supporting Zeger and Liang (1986) in that misspecification of working correlation would still give consistency in parameter estimates. Further, there is little difference between the empirically corrected and model based standard errors. This may be due to the fact that the correlation between observations is small. However, two of the regression coefficients namely, coefficients for treatment $SSTV$ and the interaction level $Time \times SSTV$ are inconsistent in sign compared to those from GLM (see Table 2.7).

The AR(1) correlation coefficient has been estimated as 0.0999, that is the two measurements from the same tree one month apart have correlation equal to 0.0999, two months apart $(0.0999)^2 = 0.01$, and so on. The correlation matrices for AR(1) and exchangeable working structures are

$$r_{AR(1)} \begin{pmatrix} 1 & 0.0999 & 0.0100 & 0.0010 & 0.0001 & 0.0000 & 0.0000 \\ & 1 & 0.0999 & 0.0100 & 0.0010 & 0.0001 & 0.0000 \\ & & 1 & 0.0999 & 0.0100 & 0.0010 & 0.0001 \\ & & & 1 & 0.0999 & 0.0100 & 0.0010 \\ & & & & 1 & 0.0999 & 0.0100 \\ & & & & & 1 & 0.0999 \\ & & & & & & 1 \end{pmatrix}$$

Table 4.2: Parameter estimates (model-based standard error; empirically corrected standard errors) for GEE under three working assumptions: IND (independence), EXCH (exchangeable) and AR(1) (autoregressive).

Effect	Parameter	IND	EXCH	AR(1)
Intercept	β_0	-4.22(0.78;0.66)	-4.21(0.78;0.66)	-4.13(0.79;0.65)
Site (Ref: Ochinga)				
Dudi	β_1	2.41(0.77;0.74)	2.40(0.77;0.74)	2.43(0.81;0.73)
Khumus	β_2	7.53(1.01;0.85)	7.48(1.01;0.85)	7.42(1.07;0.85)
Marenyo	β_3	-2.39(1.65;1.37)	-2.40(1.68;1.35)	-2.41(1.79;1.32)
Time	β_0	1.89(0.42;0.33)	1.89(0.41;0.33)	1.87(0.42;0.33)
Time×Site				
Dudi	β_4	-1.47(0.42;0.35)	-1.46(0.41;0.35)	-1.51(0.45;0.35)
Khumus	β_5	-6.46(0.82;0.56)	-6.39(0.83;0.56)	-6.39(0.86;0.57)
Marenyo	β_6	-0.46(0.77;0.56)	-0.46(0.77;0.56)	-0.52(0.83;0.54)
Time2	β_7	-0.18(0.05;0.04)	-0.18(0.05;0.04)	-0.18(0.05;0.04)
Time2×Site				
Dudi	β_8	0.16(0.05;0.04)	0.16(0.05;0.04)	0.17(0.05;0.04)
Khumus	β_9	0.77(0.10;0.07)	0.76(0.10;0.07)	0.76(0.11;0.07)
Marenyo	β_{10}	0.08(0.08;0.06)	0.08(0.08;0.06)	0.09(0.09;0.06)
Treatment (Ref: WSS)				
CRSS	β_{11}	3.10(0.77;0.70)	3.09(0.76;0.70)	2.96(0.75;0.70)
SSTV	β_{12}	1.36(0.78;0.67)	1.36(0.77;0.67)	1.29(0.74;0.66)
Time×Treatment				
CRSS	β_{13}	-1.55(0.43;0.38)	-1.55(0.42;0.38)	-1.46(0.42;0.38)
SSTV	β_{15}	-0.70(0.43;0.40)	-0.70(0.43;0.40)	-0.66(0.41;0.39)
Time2×Treatment				
CRSS	β_{16}	0.17(0.05;0.05)	0.17(0.05;0.05)	0.16(0.05;0.05)
SSTV	β_{17}	0.06(0.05;0.05)	0.06(0.05;0.05)	0.06(0.05;0.05)
Correlation	ρ	–	0.0259	0.0999

and

$$r_{EXCH}\begin{pmatrix} 1 & 0.0259 & 0.0259 & 0.0259 & 0.0259 & 0.0259 & 0.0259 \\ & 1 & 0.0259 & 0.0259 & 0.0259 & 0.0259 & 0.0259 \\ & & 1 & 0.0259 & 0.0259 & 0.0259 & 0.0259 \\ & & & 1 & 0.0259 & 0.0259 & 0.0259 \\ & & & & 1 & 0.0259 & 0.0259 \\ & & & & & 1 & 0.0259 \\ & & & & & & 1 \end{pmatrix}.$$

Note that the correlation matrix for independence working assumption is a 7×7 identity matrix. Further, the overall working correlation matrix for each of working assumption is a block matrix of the above respective matrix. The results of Type III tests of fixed effects from Proc GENMOD, i.e. from GEE analysis, for all the working correlation assumptions are consistent with GLM and GLMMs results, that is all effects are statistically significant at 5% level. However, the interpretation of the regression parameters estimates from GLMM and GEE (i.e. marginal models) are different, the theoretical reason for this is discussed in Molenberghs and Verbeke (2005, Chapter 16).

Chapter 5

Conclusion

In Chapter 2, the generalized linear modelling (GLM) and its application to the binary data, i.e. the presence / absence of the Chrysomelid beetle data is discussed. However, one of the underlying assumptions of GLM is that the data are independent, which not always the case, in particular for longitudinal studies that usually produces repeated measurements for the same subject. Then in Chapter 3, the generalized linear mixed models (GLMMs)is considered which take into account the correlation between observations within subject (i.e. tree in a site). Like GLM, GLMM assumes distribution for the response variable. Furthermore, the introduction of a random intercept takes into account for within subject correlation but it assumes a compound symmetry structure, i.e. it allows for equal correlations between two observations in a given subject, and it does not make allowance for other correlation structure.

The generalized estimating equations (GEE), which is discussed in Chapter 4, applies to any regression model including GLM, robust to specification of the correlation structure and always produces consistent and asymptotically normally distributed estimates. Unlike GLM and GLMM, in GEE we do not need to specify the distribution of the response variable. It provides valid inferences for a marginal regression model for longitudinal data without fully specifying the joint distribution of the observations. Furthermore, there are various choice of correlation structures to reflect the random intercepts.

Even though the method of GEE is the most efficient, caution should be exercised when interpreting the results. Ballinger (2004) explain some of the disadvantages of the method of GEE as follows:

1. When the link function or the variance function is misspecified, incorrect statistical conclusion can be made which misrepresent the data. The misspecification of the working correlation matrix potentially results in loss of efficiency in the model, which leads differences in standard errors.

2. The method results in computational issues when the correlation within clusters is high. For example the unstructured working correlation structure will take long to converge, or might cause the programme to crash depending on who sophisticated a machine is.

3. When the numbers of subjects are small, the estimates of the variance could be biased.

Molenberghs and Verbeke (2005) advise that when GEE is deemed unsatisfactory in the sense that there is some scientific interest in the correlation structure then one should consider some of the extensions of GEE, such as Prentice's GEE method (Prentice, 1988, 1991), second-order GEE (Zhao and Prentice, 1990; Liang, Zeger and Qaqish, 1992) or alternating logistic regression (Carey, Zeger and Diggle, 1993).

In this study we have observed that a presence of an influential observation can affect the inference or interpretation of the regression coefficient. It should be noticed that GLM gives a model with inflated p-values, in particular when the correlation between observations within a subject is high. Even for the data used in this dissertation which has small correlation, the p-values reported in Table 2.7 are inflated for most of the effects compared to GEE results (results are not given here). Further, the results also suggest that even with small correlation, ignoring a random effects in a binary model can lead to inconsistent estimation of regression parameters.

Finally, more time was spent on the standard logistic regression model (i.e. GLM) on a selection of predictors and model diagnostics issues to obtain a model with good fit. Then

the same model was used to analyze the data using GLMMs and GEE methods. However, we suggest when one deals with modelling of longitudinal binary data that he or she should apply the model selection procedures of the generalized linear mixed model or generalized estimating equations to obtain optimal model.

Bibliography

Azzalini, A. (1996). *Statistical Inference: Based on the likelihood*. Chapman and Hall: London.

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

Ballinger,G.A.(2004).Using generalized estimating equations for longitudinal data analysis. Organizational research methods,vol 7 no.2,127-150

Collett, D. (2003). *Modelling Binary Data*, 2nd edition. Chapman and Hall, New York.

Debusho, L.K. and Sileshi, G. (2012). Modelling of correlated soil animals count data. *Peer-reviewed Proceedings of the 54th Annual Conference of the South African Statistical Association for 2012 (SASA 2012)*. Nelson Mandela Metropolitan University, Port Elizabeth, South Africa. ISBN: 978-1-86822-621-4.

Demidenko, E. (2013). *Mixed Models: Theory and Applications with R*, 2nd edition. John Wiley & Sons, Inc., New Jersey.

De Jong, P and Heller, G.Z. (2008). *Generalized Linear Models for Insurance Data*. University Press, Cambridge.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*, 2nd edition,

Boca Raton, FL: Chapman & Hall/CRC.

Fieberg, J., Randall H. Rieger, Michael C. Z. and Jonathan S.S. (2009). Regression modelling of correlated data in ecology: subject-specific and population averaged response patterns. *Journal of Applied Ecology*, **46**(**5**), 1018 – 1025.

Fitzmaurice, G.N., Laird N.M. and Ware J. (2004). *Applied longitudinal analysis*. Wiley-IEEE.

Fitzmaurice, G.N., Laird N.M. and Rotnitzky, A. (1993). Regression models for discrete longitudinal response. *Statistical Science*, **8**, 284 – 309.

Greene W.H. (1997). *Econometric analysis*, 3rd. Prentice-Hall, Inc, Upper Saddle River, N.J.

Haberman, S.J. (1974). *The Analysis of Frequency Data*, Vol.1. Chicago: University of Chicago. Hardin, J. W., and J. M. Hilbe. (2003). *Generalized Linear Models and Extension*. Station, TX: Stata Press.

Hardin, J. W., and Hilbe, J. M. (2003). *Generalized estimating equations*. Boca Raton, FL: Chapman and Hall/CRC Press.

Hosmer, D.W. and Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Comm. in Stat. - Theory and Methods*, **A9**, 1043 – 1069.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer, New York.

Lee, Y., Nelder, J. A., and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects*. Chapman and Hall, New York.

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear

models. *Biometrika*, **73**, 13 – 22.

McCulloch, C.E., Searle, S.R., and Neuhaus, J.M. (2008). *Generalized, Linear, and Mixed Models*, 2nd edition. Wiley, New York.

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

McCullogh, C. E. and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. John Wiley and Sons, New York.

Molenberghs, G. and Verbeke, G. (2005). *Models for discrete longitudinal data*. Springer, NewYork,NY.

Ortega, J.M. and Rheinboldt, W.C. (1970). *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York. Pan, W. (2001). Akaikes Information Criterion in Generalized Estimating Equations. *Biometrics*, **57**, 120 – 125.

Prentice, R.L. (1988). Correlated binary regression with covariates speciffic to each binary observation. *Biometrics* **44**, 1033 – 1084.

R Development Core Team. (2012). R: A language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Sileshi, G., Girma, H. and Mafongoya, P.L. (2006). Occupancy-abundance models for predicting densities of three leaf beetles damaging the multipurpose tree Sesbania sesban in eastern and southern Africa. *Bull. Entomol. Res.*, **96**, 61 – 69.

Sileshi, G., Girma, H. and Nydzi, G.I. (2009). Traditional occupancyabundance models are inadequate for zero-inflated ecological count data. *Ecological Modelling*, **220**, 1764 – 1775.

Vens, M. and Ziegler, A. (2012). Generalized estimating equations and regression diagnostic for longitudinal controlled trials: A case study. *Comput. Statist. Data Anal.*, **56**, 1232 – 1242.

Verbeke, G. and Molenberghs, G. (1999). *Linear Mixed Models for Longitudinal Data*. Springer-Verlag: New York.

West,Brady.T,Welch Kathleen B and Galecki Andrzej T(2007).Linear mixed modesl:A practical guide using statistical software.Chapman and Hall/CRC,London.
Wu, L. (2010). *Mixed Effects Models for Complex Data*. Chapman and Hall, New York.

Zeger, S. L., and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121 – 130.

Ziegler, A. Kastner C., Gromping U. and Blettner M. (1996). The generalized estimating equations in the past ten years: An overview and a biomedical application fttp://ftp.stat.uni-muenchen.de/pub/sfb386/paper24.ps.Z

Zuur, A., Ieno, E.N., Walker, N., Saveliev, A.A. and Smith, G.M. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer Science+Media, LLC.