Modelling Longitudinal Binary Disease Outcome Data including the effect of covariates and extra variability

By

SIYABONGA NGCOBO

Submitted in partial fulfillment of the academic requirements for the degree of
Master of Science in Statistics
in the
School of Statistics and Actuarial Science,
University of Kwazulu-Natal
2011

Abstract

The current work deals with modelling longitudinal or repeated non-Gaussian measurements for a respiratory disease. The analysis of longitudinal data for non-Gaussian binary disease outcome data can broadly be modeled using three different approaches; the marginal, random effects and transition models. The marginal type model is used if one is interested in estimating population averaged effects such as whether a treatment works or not on an average individual. On the other hand random effects models are important if apart from measuring population averaged effects a researcher is also interested in subject specific effects. In this case to get marginal effects from the subject-specific model we integrate out the random effects. Transition models are also called conditional models as a general term. Thus all the three types of models are important in understanding the effects of covariates and disease progression and distribution of outcomes in a population. In the current work the three models have been researched on and fitted to data. The random effects or subject-specific model is further modified to relax the assumption that the random effects should be strictly normal. This leads to the so called hierarchical generalized linear model (HGLM) based on the h-likelihood formulation suggested by Lee and Nelder (1996). The marginal model was fitted using generalized estimating equations (GEE) using PROC GENMOD in SAS. The random effects model was fitted using PROC GLIMMIX and PROC NLMIXED in SAS (generalized linear mixed model). The latter approach was found to be more flexible except for the need of specifying initial parameter values. The transition model was used to capture the dependence between outcomes in particular the dependence of the current response or outcome on the previous response and fitted using PROC GENMOD. The HGLM was fitted using the GENSTAT software. Longitudinal disease outcome data can provide real and reliable data to model disease progression in the sense that it can be used to estimate important disease parameters such as prevalence, incidence and others such as the force of infection. Problem associated with longitudinal data include loss of information due to loss to follow up such as dropout and missing data in general. In some cases cross-sectional data can be used to find the required estimates but longitudinal data is more efficient but may require more time, effort and cost to collect. However the successful estimation of a given parameter or function depends on the availability of the relevant data for it. It is sometimes impossible to estimate a parameter of interest if the data cannot its estimation.

Declaration

This dissertation represents the original work of the author. The work done by others or by myself previously has been acknowledged and referenced accordingly. This research has not been previously submitted, in any form, to any institution.

....

Siyabonga Ngcobo

....

Prof. H.G. Mwambi

MSc Supervisor

....

Dr. S. Ramroop

MSc Co-Supervisor

Acknowledgement

I would like to thank my supervisor, Prof. Henry G. Mwambi, for his tremendous support, help and guidance throughout my masters.

I would also like to thank my co-supervisor Dr. S. Ramroop for his comments.

I also like to thank the DST/NRF Centre of Excellence in Epidemiological Modelling and Analysis (SACEMA) for funding my studies.

I want to thank my family and friends for their support.

My special thanks go to my parents Ncamisile and Jabulani Ngcobo for their unfailing moral support and they always encouraged me to do my best in all matters of life.

Finally I want to thank God Almighty who made it possible for me by opening the opportunities and gave guidance and support that made me to complete the study.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Respiratory infection is considered as one of the major public health problems in the developing countries (Amir et al., 2009). Baqui et al (2007) recognized that the respiratory infection is the leading cause of morbidity and mortality in many countries. Respiratory infection is an infection that occurs or is associated with the respiratory system. There are several forms of respiratory infections such the lower and upper respiratory infections. These are viral diseases that can also be of different strains or subtypes that can change between years. Some subtypes are contagious and spread easily between people, while other subtypes are dependent on other factors such as suitable weather conditions to aid their spread. Rahman and Rahman (1997) indicate that in developing counties 30% of all patients consultation and 25% of all admission are of acute respiratory tract infections.

In a longitudinal study individuals outcomes of interest are measured repeatedly through time (Diggle et al., 2002). This is in contrast to cross-sectional studies, in which a single observation is measured for each individual at a given time. More generally longitudinal data can be collected either prospectively, following subjects forward in time or retrospectively, by extracting multiple measurements overtime on each person from historical records (Diggle et al., 2002). However the latter might not be reliable because they are not collected in real time. This may introduce inaccuracy due to recall bias or invalidated records. Longitudinal studies can involve a large number of repeated measures such as in drug abuse follow up studies (Yang et al, 2007). These are population based follow up studies. Clinical trials are also by nature prospective studies which often have time to a clinical outcome as the principal response. Analysis of such data needs a particular analysis called survival analysis which is not the focus in the current work. Longitudinal studies are designed to investigate change over time in a characteristic which is measured repeatedly for each study participant. Longitudinal data require special statistical methods because the set of observations on one subject tends to be inter-correlated. This correlation must be taken into account in the analysis in order to draw valid scientific conclusions or inference. The subjects in a longitudinal study can usually be assumed to be independent but the observations within a subject are bound to be correlated. Longitudinal studies are also referred to as repeated measure designs are designs in which the measurement process consists of repeated measurements on the same experimental unit.

The experimental unit can also include plots in the context of agricultural experiments, a single patient in health research or a household in survey studies among many. Repeated measures analysis employs statistical methods to deal with data or outcomes measured on the same experimental unit at different times or under different experimental or observational conditions. In a repeated measurements analysis, one is usually interested in both between-subject and within-subject effects. Between-subject effects are those whose values change only from subject to subject and remain the same for all observations on a single subject, for example, the treatment and gender effects. Within-subject effects are those whose values may differ from measurement to measurement, for example, Level of CD4 count from one time to the next. Repeated measures analysis can be applied to both continuous and categorical outcomes. Examples of the latter include outcomes such as disease status measured or observed over time as in the current analysis and also in disease surveillance studies. The former type of responses include transformed responses or outcomes in order to satisfy normality such as log viral loads, square root CD4 lymphocytes for HIV infected individual and untransformed variables such as blood pressure, weight and many more.

Most often repeated measures analyses for continuous data applies methods based on the mixed model with special parametric forms on the covariance structure for observations from the same subject (Laird and Ware, 1982). The current project will however focus on the application of different approaches of modelling non-Gaussian data namely the generalized linear model (GLM), generalized estimating equations (GEE), the generalized linear mixed model (GLMM) and hierarchical generalized linear model (HGLM) for analyzing longitudinal non-Gaussian data (McCullagh and Nelder, 1989; Liang and Zeger, 1986; Molenberghs and Verbeke, 2005; Lee and Nelder, 1996). The thesis will also consider transition models which allow for the dependence on already observed outcomes or the history of outcomes. The software of analysis in the current study is SAS except under the HGLM where GENSTAT will be used. The data to be modelled in this project (reported in Lee, Nelder and Pawitan, 2006) is in the form of a binary response repeated measurements for individuals exposed or susceptible to infection by a respiratory disease. At a given measurement occasion the observation gives the infection status of an individual in the sample. This means the response Y either takes the value 1 (success) or 0 (failure) and in the case of a disease process as in the current application Y=1 means the respiratory disease status of an individual is 'infected' and Y=0 means the respiratory disease status of an individual is 'not infected'. In Lee et al (2006) the data were analysed using Genstat. Thus the successful implementation of the methods supported in SAS is one of the achievements of the current work. Binary disease outcomes over time can be conveniently modeled as Markov chain process with transition probabilities denoting change from a given disease status to another (in the future) similar to transition from no use to use of a substance as in drug abuse research studies (Yang et al., 2007).

This project will focus on different ways or approaches for modelling such inherently correlated data in the context of a disease process. Thus it is appropriate to assume the type of data used in the project is from a parent distribution that is non-Gaussian contrary to most applications which are based on data that satisfy normality. For such correlated data over time there are several correlation structures that can be assumed. For example the compound symmetry structure assumes the covariance between any two measurements at two different time points on the same unit is the same and further the same variance for all measurements. However, sometimes the correlation between observations that are close together in time is likely to be higher than the correlation for measurements further apart. In this case, the first-order autoregressive correlation structure also denoted as AR(1) may be more appropriate. This can be extended to a general p-order AR(p) correlation structure depending on the complexity of the dependence structure. A more flexible correlation structure is that based on the unstructured model, in which we allow for different parameters for the variance of each repeated measurement as well as different covariance parameters for each pair of repeated measurements. Clearly such a structure is highly parametric and this may explain why analysts and modellers aim at more parsimonious covariance structures as much as possible. The major difference between modelling longitudinal or clustered correlated data compared to independent observations is that the independence assumption under the classical maximum likelihood estimation is now not attainable.

1.1 Exploratory and Description of the data

The current data was a published data which was reported in Lee et al (2006). In the current data, 111 patients were followed prospectively for 4 months or time points and in each month an individual test was conducted to determine if his/her respiratory status as whether infected or not infected. In each of two centres included in the study the patients were randomly assigned to one of the two treatment groups receiving either the active treatment or a placebo. Thus the data is an example of a multi-centre clinical trial where the outcome of interest is binary. Respiratory infection status indicated as 0 for not infected, 1 for being infected was determined for each of four visits. For analysis purpose the two treatments are coded as 1 for placebo and 2 for active treatment. The aim of the current work is to accurately model the resulting respiratory infection disease data accounting for multiple visits per patient who were randomized into placebo or active treatment at baseline.

The model includes both categorical and continuous covariates. The variables centre, treatment, sex, and baseline (baseline respiratory status) are factor variables with two levels each. The variable age (age at time of entry into the study) is a continuous variable. There were 23 females and 88 males in the study. The effect of measured covariates was also an important component in the analysis. Particularly this was a multi-centre study it is important to control for the centre effect. The incidence of a disease in the context of longitudinal follow up data is defined as the number or proportions of new cases within a given time interval while the prevalence of the disease is the proportion of infected individuals at a given time which includes both newly infected and existing cases. In the current analysis, an incident event for a given individual is defined as the change from state Y=0 to state Y=1 in a given time interval.

A summary of the number of patients in the study at each time point and the number of infections at each time point are given in Table 1.1. The marginal sample probabilities showing infected and uninfected with respiratory disease at each time point are given in Table 1.2. It should be noted however the proportion in Table 1.2 gives probabilities of being infected and uninfected estimates calculated at each given time point over all the individuals sampled from the target population and randomized to that given arm. Thus the estimates in Table 1.1 and 1.2 are in fact cross-sectional estimates. The probabilities in Table 1.2 are calculated assuming independence between time and individuals. Table 1.2 indicate that the probability of being infected with respiratory disease is the highest at baseline. The result in Table 1.3 was fitted using SPSS statistical software. The result indicate that there was no statistically significant association between time and the status of respiratory disease. It is clear from Figure 1.1 that the prevalence is not constant over time and is different from zero in the two arms. Figure 1.2 show that 49.1% of individuals who are on placebo and 33.3% of individuals who are on active treatment are infected with the respiratory disease at visit 1. Figure 1.2 also show that 57.9% of individuals who are on placebo and 38.9% of individuals who are on active treatment are infected with the respiratory disease at visit 4. The prevalence is consistently lower in the active treatment arm at all visits. This is also evident in Table 1.1. Figure 1.3 show that 39.8% males individuals

Table 1.1: Treatment arm specific number of infected and prevalence of the respiratory infection at each visit

	Placebo arm			Active arm		
	Number infected	N	Prevalence	Number infected	N	Prevalence
baseline	31	57	54.4	30	54	55.6
visit1	28	57	49.1	18	54	33.3
visit2	35	57	61.4	16	54	29.6
visit3	31	57	54.4	15	54	27.8
visit4	33	57	57.9	21	54	38.9

Table 1.2: The number of being infected and uninfected with respiratory disease at each visits

	Infected	Probability of being infected	uninfected	Probability of being uninfected
Baseline	61	0.55	50	0.45
Visit1	46	0.41	65	0.59
Visit2	51	0.46	60	0.54
Visit3	46	0.41	65	0.59
Visit4	54	0.49	57	0.51

and 47.8% females individuals are infected with the respiratory disease at visit 1. Figure 1.3 also show that 48.9% males individuals and 47.8% females individuals are infected with the respiratory disease at visit 4. Figure 1.4 show that 53.6% of the individuals in centre 1 and 29.1% of the individuals in centre 2 are infected with the respiratory disease at visit 1. Figure 1.4 also shows that 62.5% of the individuals in centre 1 and 34.5% of individuals in centre 2 are infected with the respiratory disease at visit 4. The observed number of disease cases at a given time are generally assumed to have a binomial distribution and the time specific estimate of prevalence is the sample proportion of cases. However as indicated earlier the individual binary observations are most likely dependent leading to correlated data that must be accounted for.

Analysis using the generalized linear model methodology assuming the outcomes within a patient are independent is clearly not appropriate because of the inherent correlation of outcomes within an individual. There are three types of models that will be used to model the current data.

Table 1.3: Association between time and status of respiratory disease

	Value	DF	Pr>Chisq
Pearson Chi-square	1.706	3	0.636
Likelihood ratio	1.706	3	0.636
Linear-by-linear Association	0.657	1	0.418
N of valid cases	444		

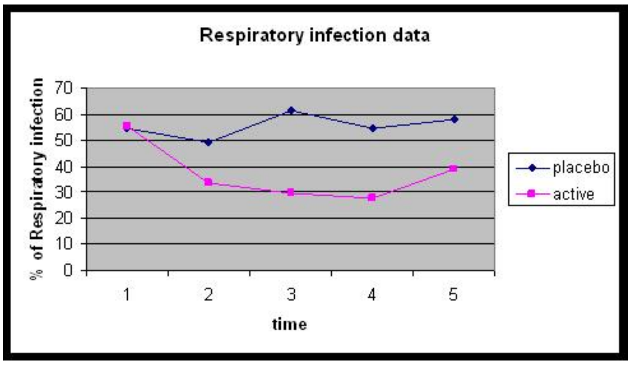

Figure 1.1: Proportion of respiratory infection for placebo and active treatments at baseline and each visits

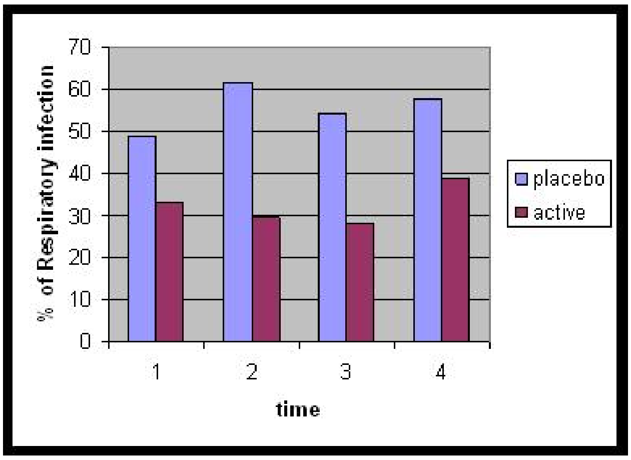

Figure 1.2: The percentage of infection for placebo and active treatment at four visits

These are marginal, random-effects and transition models. The models will be fitted using SAS software except under the hierarchical modelling approach where GENSTAT will be used. Marginal models are appropriate when inference is about the average response in the subpopulation sharing

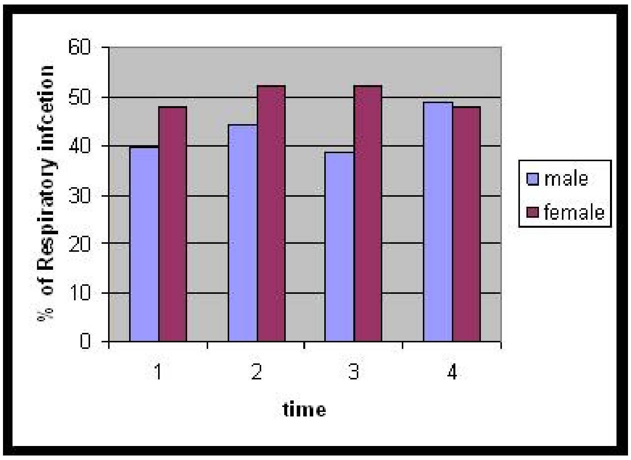

Figure 1.3: The percentage of infection for males and females individuals at the four visits

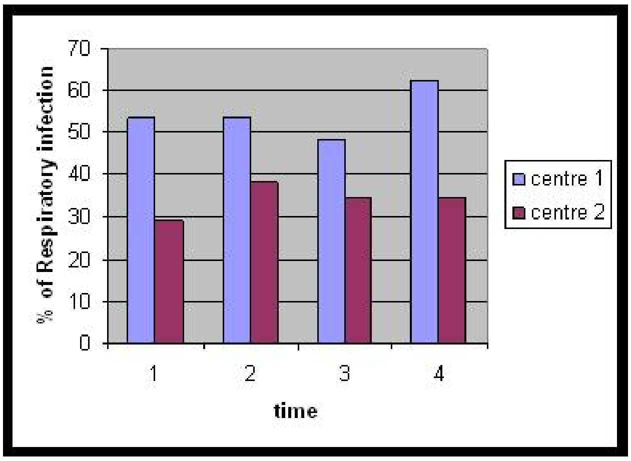

Figure 1.4: The percentage of infections in the two centres at the four visits

or having a common set of covariate values. Marginal models are important in the context of a clinical trial where estimating the average difference between treatments is generally the most important goal. Because the response variable is binary the logit link function under the binomial probability model will be used. More precisely individual observations are best modelled as a Bernoulli type of responses because it is assumed that the outcome variable of interest is binary. Thus the data are coded such that a successful outcome (infection) in the experiment is coded as 1 and a failure (not infected) is coded as 0. The Bernoulli distribution is a special case of the binomial distribution where number of trials is n=1. The analysis of binary and counts data is often faced with the problem of over and under dispersion. The work will also briefly discuss the problem of dispersion in relation to the modelling approaches being applied in the current work.

1.2 Objective of the Research

The question of interest was whether the active treatment is effective in controlling the incidence of the disease or not compared to placebo. An additional question of interest was also whether the evolution of disease status was time dependent and whether this dependence was different for the two treatments given other measured covariates. The primary objective of this study was to review and apply advanced statistical methods to analyze correlated non-Gaussian longitudinal data using both marginal and subject-specific random effects models. The work also includes the study of transition models to model dependent outcomes. This will enable efficient and correct estimation of covariate effects on the outcome which in our case is the respiratory disease status. The work were also considered approaches to deal with non-normal random effects under the hierarchical generalized linear modelling approach. The application were based on multiple outcomes of a respiratory disease infection status of individuals monitored over time. The data was an example of multi-centre clinical trial because the individuals were sample from two centres. The specific objectives of the study were:

- To review statistical methods used in the analysis of correlated non-Gaussian data.
- To investigate approaches that can be used to construct a dispersion model for data that may suffer from overdispersion.
- To determines the effect of the active treatment compared to placebo in controlling the incidence of disease using the marginal effects, subject effects and transition models.
- To review the theory and application of hierarchical generalized linear models not necessary assuming normal subject specific effects.
- Demonstrate the understanding of the above through an application to a disease outcome data.
- Compare the performance of different approaches.

Chapter 2

Marginal models for correlated longitudinal data

The observed or measured status of a disease (infected or not infected) such as a respiratory disease in an individual is a binary response outcome. However, clinically such an outcome is based on a disease threshold or cut off using a biomarker that could well be continuous. In this work we focus on data derived from repeated observations from the same individual. Thus the assumption of independent observations cannot be used because observations within an individual are correlated therefore not independent. The chapter briefly discusses the theory of generalized linear models (GLMs) for non-Gaussian data by first introducing the exponential family. The problem of correlated non-Gaussian longitudinal data is then discussed and addressed using the generalized estimating equations (GEEs) due to Liang and Zeger (1986). The application of this type of marginal model under different correlation structures to respiratory infection data is done using the SAS software and the results discussed.

2.1 Generalized linear models

2.1.1 The exponential family of distributions

Generalized linear models (GLMs) are an extension of the classical linear models, so that the latter form a suitable starting point for our discussion.

Let Y be a random variable from a parent distribution which possibly depends on the parameters θ and ϕ. A detailed discussion of generalized linear models can be found in McCullagh and Nelder (1989). The random variable Y is said to have a distribution from the exponential family if its probability density function can generally be written in the form:

$$f(y \mid \theta, \phi) = \exp\Big[\frac{y\theta - \psi(\theta)}{\phi} + c(y, \phi)\Big]. \tag{2.1}$$

The parameter θ is called the natural parameter and ϕ is called the dispersion or the scale parameter. The function $\psi(.)$ is called the cumulant function which is helpful in generating the mean and variance as will be shown below. The mean and variance of Y can be derived by exploiting the equation

$$\int f(y \mid \theta, \phi) dy = 1. \tag{2.2}$$

Taking the first and second derivative with respect to θ from both sides of Eq.(2.2) leads to the two equations of the form

$$\int (y - \psi^{'}(\theta)) f(y \mid \theta, \phi) dy = 0, \tag{2.3}$$

and

$$\int [\phi^{-1}(y - \psi^{'}(\theta))^2 - \psi^{''}(\theta)] f(y \mid \theta, \phi) dy = 0. \tag{2.4}$$

Solving for $\mu = E(y)$ and $Var(y) = E[(y - \mu)^2]$ in Eq.(2.3) and (2.4) respectively we get the solutions $E(y) = \psi^{'}(\theta)$ and $Var(y) = \phi v(\mu)$ where $v(\mu) = \psi^{''}(\theta)$.

Note that in general the mean and variance are dependent, since

$$Var(y) = \phi\psi^{''}[\psi^{'-1}(\mu)] = \phi v(\mu) \tag{2.5}$$

The function $v(\mu)$ is called the variance function. The function $\psi^{'-1}(.)$ which express θ as a function of μ is the link function and $\psi^{'}$ is the inverse link function. There are several distributions which conform to this structure and for clarification purposes we briefly relate the above formulation to the Normal, Poisson and Bernoulli distributions which all fall under the exponential family of distributions.

In the case when $Y \sim N(\mu, \sigma^2)$ then its density function is given by

$$\begin{aligned} f(y \mid \theta, \phi) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{1}{-2\sigma^2}(y-\mu)^2\right] \\ &= \exp\left[\frac{1}{\sigma^2}\Big(y\mu - \frac{\mu^2}{2}\Big) - \Big(\frac{\log(2\pi\sigma^2)}{2}\Big) - \frac{y^2}{2\sigma^2}\right] \end{aligned}$$

Which is in the form of an exponential family where the natural parameter is $\theta = \mu$, the scale parameter is $\phi = \sigma^2$, $c(y, \phi) = \frac{-\log(2\pi\sigma^2)}{2} - \frac{y^2}{2\phi}$ and $\psi(\theta) = \frac{\theta^2}{2} = \frac{\mu^2}{2}$. A special property about the normal distribution is that the variance function $v(\mu) = 1$, therefore $Var(y) = \sigma^2$ independent of μ.

Let Y be Poisson distributed with mean μ. Then the density function of the Poisson distribution is

$$f(y) = \frac{\mu^y \exp(-\mu)}{y!}$$

The above density function is also part of the exponential family since it can be written as

$$f(y) = \exp(y \log \mu - \mu - \log y!)$$

Thus the natural parameter is $\theta = \log \mu$, $\psi(\theta) = \exp(\theta)$ the scale parameter is $\phi = 1$ and $c(y, \phi) = \log y!$. In this case the variance function $v(\mu) = \mu$, therefore $Var(y) = \mu$ which depend on μ. Thus a Poisson random variable is naturally modelled using a log-link function.

In the case of the Bernoulli distribution $Y \sim Bernoulli(\pi)$ then the density function is given by:

$$\begin{aligned} f(y,\theta,\phi) &= \pi^y(1-\pi)^{1-y} \\ &= \exp[y\log\pi + (1-y)\log(1-\pi)] \\ &= \exp[y\log\Big(\frac{\pi}{1-\pi}\Big) + \log(1-\pi)] \end{aligned}$$

Which is in the form of an exponential family were the natural parameter is $\theta = \log\Big(\frac{\pi}{1-\pi}\Big)$, the scale parameter is $\phi = 1$ and $\psi(\theta) = -\log(1-\pi) = \log(1+\exp(\theta))$ since $\pi = \frac{\exp(\theta)}{1+\exp(\theta)}$

Note that under this model the mean and the variance are given by

$$E(y) = \psi^{'}(\theta) = \pi$$

and

$$Var(y) = \phi v(\mu) = \mu(1-\mu)$$

Which depend on μ as with the case of the Poisson distribution. The canonical link function is here given by the logit link where

$$\theta = \log\Big(\frac{\mu}{1-\mu}\Big)$$

The above comparison shows that the normal distribution possess the unique property that the variance is independent of the mean contrary to almost all other members of exponential family of distributions.

2.1.2 The structural component and link function

Let $Y_1, Y_2, ..., Y_N$ denote N independent observations from the same exponential family distribution such that

$$f(y_i \mid \theta_i, \phi) = \exp\Big[\frac{y_i\theta_i - \psi(\theta_i)}{\phi} + c(y_i,\phi)\Big]. \qquad (2.6)$$

In addition, let x_i denote a vector of p known independent variables, covariate or predictor variable. It is reasonable and practical to assume that the differences between individual mean can be captured through general relation

$$g(E(Y_i)) = g(\mu_i) = x_i^{'}\beta$$

where $E(Y_i) = \mu_i$ and g is a link function that relate the mean of the response to the linear predictor $x_i^{'}\beta$, x_i is a vector of the independent variables for the i^{th} observations and β is a vector of regression parameters to be estimated from data using the maximum likelihood estimation. In the classical GLM analysis (McCullagh and Nelder, 1989) the outcomes Y_i, $i = 1, ..., N$, are assumed to be independent and have the probability distribution from the exponential family. The most commonly used link function is the canonical link function where $g(\mu_i) = \theta_i$ so that the relation $\theta_i = x_{ij}^{'}\beta$ holds.

2.2 Maximum likelihood estimation

The likelihood L(β, ϕ) is given by

$$L(\beta, \phi) = \prod_{i=1}^{N} f(y_i \mid \beta, \phi) = \prod_{i=1}^{N} \exp\Big[\frac{y_i\theta_i - \psi(\theta_i)}{\phi} + c(y_i, \phi)\Big]. \tag{2.7}$$

Estimation of the regression parameters in β is usually done using maximum likelihood estimation. Assuming that the observations are independent and the log-likelihood is defined by

$$l(\beta, \phi) = \frac{1}{\phi}\sum_i [y_i\theta_i - \psi(\theta_i)] + \sum_i c(y_i, \phi). \tag{2.8}$$

The maximum likelihood estimator of the parameter vector β is obtained by solving the estimating equations based on the score equation given below

$$S(\beta) = \sum \frac{d\mu_i}{d\beta} v_i^{-1}(y_i - \mu_i(\beta)) = 0. \tag{2.9}$$

Note that the estimation of β depends on the density function only through the mean and the variance functions $v_i = v(\mu_i)$. The score equations can be solved numerically using the iterative algorithm such as the Newton-Raphson, Fisher scoring and re-weighted least squares (RWLS). The inference on β is based on a classical maximum likelihood theory such as the asymptotic properties of the estimate $\hat{\beta}$ leading to inference based on the wald- test, Likelihood ratio test or the score test.

To solve the score equations by the Newton-Raphson method, the iterative equation is given by

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} - (H^{(t)})^{-1}\mu^{(t)} \tag{2.10}$$

where $H^{(t)}$ is the Hessian matrix while $\hat{\beta}^{(t)}$ and $\mu^{(t)}$ are the vector of regression parameter and the vector of mean at step t respectively . The Hessian matrix is the square matrix of second-order partial derivatives of a function. Fisher scoring iterative equation is given by

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + (\vartheta^{(t)})^{-1}\mu^{(t)}. \tag{2.11}$$

where $\vartheta = X^{'}WX$ is the expected information matrix and W is the diagonal matrix with the main diagonal elements

$$w_i = \frac{(d\mu_i/d\eta_i)^2}{var(Y_i)}.$$

Reweighted least squares iterative equation is given by

$$\hat{\beta}^{(t+1)} = (X^{'}W^{(t)}X)^{-1}X^{'}W^{t}z^{(t)}. \tag{2.12}$$

where z is a linearized form of link function g, evaluated at y. Fisher scoring method is similar to the Newton-Raphson method but the only difference being how the Hessian matrix is used. Fisher scoring method uses the expected value of this matrix called expected information, whereas the Newton-Raphson method uses the matrix itself or the observed information. Once parameters are estimated there is a need to make inference on them. The Wald test is an asymptotic parametric statistical test with a great variety of uses. Whenever a relationship within or between data items can be expressed as a statistical model with parameters to be estimated from a sample, the Wald test can be used to test hypotheses about the true value of the parameter based on the sample estimate. In most statistical tests including the Wald test the maximum likelihood estimate $\hat{\beta}$ of the parameter(s) of interest β is compared with the proposed value β_0, with the assumption that the difference between the two will be approximately normal. Typically the square of the difference is compared to a chi-squared distribution quantile at a given significance level. In the univariate case, the square of the Wald statistic is defined as

$$\omega = \frac{(\hat{\beta} - \beta_0)^2}{var(\hat{\beta})}$$

Alternatively, the standardized difference can be compared to a normal distribution quantile at a given level of significance. In this case the test statistic is

$$z = \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})}$$

Strictly speaking the latter form is what is commonly called the Wald statistic. The likelihood-ratio test can also be used to test whether an effect on the mean response on the link function scale exists or not. Usually the Wald test and the likelihood ratio test give very similar conclusions (as they are asymptotically equivalent). Another alternative is the score test, which have the advantage that it can be formulated in situations where the variability in the estimate is difficult to estimate. Under the non-normal data the likelihood ratio test is generalized to the idea of deviance. The deviance of a model is defined as the likelihood ratio of the saturated model versus the particular model of interest.

Example

Suppose that $y_i \sim B(1, p_i)$ where $\mu_i = p_i$

$$\begin{aligned}
f(y_i) &= p_i^{y_i}(1-p_i)^{1-y_i} \\
L(\mu_i) &= \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \\
l(\mu_i) &= \sum_{i=1}^{n} [y_i \log p_i + (1-y_i)\log(1-p_i)] \\
&= \sum_{i=1}^{n} \Big[y_i \log \mu_i + (1-y_i)\log(1-\mu_i)\Big]
\end{aligned}$$

The saturated log-likelihood is defined as

$$l(y_i) = \sum_{i=1}^{n} [y_i \log y_i + (1-y_i)\log(1-y_i)]$$

The deviance is

$$
\begin{aligned}
D(y_i, \mu_i) &= 2\log\frac{L(y_i)}{L(\mu_i)} \\
&= 2[l(y_i) - l(\mu_i)] \\
&= 2\sum_i \left[y_i \log\left(\frac{y_i}{\mu_i}\right) + (1 - y_i)\log\left(\frac{1 - y_i}{1 - \mu_i}\right)\right]
\end{aligned}
$$

Consider two models M_1 and M_2 where M_1 is a subset of M_2 with corresponding deviance $D(M_1)$ and $D(M_2)$ respectively. Then the change in deviance between M_1 and M_2 is

$$\triangle D = D(M_1) - D(M_2) \geqslant 0$$

which is approximately $\chi^2(k)$ where k is extra number of parameters in M_2. Therefore the deviance can be used to measure the significance of the additional parameters in M_2.

2.3 Logistic regression models for binary response

The aim of this section is to review and apply logistic regression to the binary respiratory infection data. Let Y_{ij} denote the respiratory status of subject i at visit j for $i = 1, ..., 111$ and $j = 1, ..., 4$, This

$$Y_{ij} = \begin{cases} 1 & \text{if respiratory status is 'infected'} \\ 0 & \text{if respiratory status is 'not infected'} \end{cases}$$

Let $x_{ij} = (x_{ij1}, ..., x_{ijp})'$ denote a $p \times 1$ vector of covariates for subject i at visit j. The five explanatory variables are treatment (x_{ij1}), sex (x_{ij2}), baseline disease status (x_{ij3}), centre (x_{ij4}) and age (x_{ij5}) where

$$x_{ij1} = \begin{cases} 1 & \text{if placebo,} \\ 2 & \text{if active,} \end{cases}$$

$$x_{ij2} = \begin{cases} 1 & \text{if male,} \\ 2 & \text{if female,} \end{cases}$$

$$x_{ij3} = \begin{cases} 1 & \text{if infected at baseline,} \\ 0 & \text{if not infected at baseline,} \end{cases}$$

and

$$x_{ij4} = \begin{cases} 1 & \text{if centre1} \\ 2 & \text{if centre2} \end{cases}$$

It is noted that all the five covariates treatment, sex, baseline status, centre and age all fall in the category of baseline covariates. Thus in all the variables we can drop the index j and reduce the notation to x_{i1}, x_{i2}, x_{i3} and x_{i4} respectively. The fifth covariates x_{i5} is age of an individual at baseline which is a continuous covariate. The observed disease status of an individual i denoted by Y_i takes value 1 if the individual is infected and zero otherwise. The logistic regression model is part of a special case of statistical models called generalized linear models (GLMs) under the assumption that observation Y_i are independent and $\phi = 1$. Logistic regression allows one to model the probability of an outcome of an event as a linear function of a set of variables that may be continuous or discrete including dichotomous variables. The response variable is binary therefore a Bernoulli distribution is naturally assumed. The marginal probability of the outcome is denoted by $p_i = Pr(Y_i = 1)$. For simplicity, let us assume that the marginal probability depends on a set of covariates $x_i = (1, x_{i1}, ..., x_{ip})'$ through regression coefficients $\beta_0, \beta_1, ..., \beta_p$. Then a logistic regression model is a function of the form

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} = \eta_i$$

The expression to the right denoted by η_i as earlier stated is called the linear predictor. This means that for an individual with covariate vector x_i.

$$p_i = P(Y_i = 1) = \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)} = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}. \tag{2.13}$$

The interpretation of the parameter β_j, $j = 1, 2, ..., p$ is that for every unit increase in x_j $logit(p_i)$ increases by β_j units conditionally on holding other variables constant where

$$logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = x_{ij}'\beta. \tag{2.14}$$

The exponential family simplifies the construction of the likelihood and log-likelihood. The log-likelihood based on a sample of N individuals is

$$l(\beta) = \sum_{i=1}^{N} \Big[y_i \ln \Big(\frac{p_i}{1-p_i} \Big] + \ln(1-p_i) \Big], \tag{2.15}$$

where the natural parameter $\theta_i = \ln\left(\frac{p_i}{1-p_i}\right)$ which is implicitly a function of β through Eq.(2.13). The mean and variance of the response variable are given by:

$$E(Y_i) = \mu_i = \frac{e^{\theta_i}}{1+e^{\theta_i}} = p_i,$$

and

$$\sigma_i^2 = var(Y_i) = \frac{e^{\theta_i}}{1+e^{\theta_i}} \times \frac{1}{1+e^{\theta_i}} = p_i(1-p_i).$$

In general, we can write the model for the combined data as

$$logit(p) = X\beta \tag{2.16}$$

where X the design or model matrix given by

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{ip} \\ 1 & x_{21} & x_{22} & . & . & . & x_{2_p} \\ 1 & x_{31} & x_{32} & . & . & . & x_{3p} \\ . & . & . & & & & . \\ . & . & . & & & & . \\ . & . & . & & & & . \\ 1 & x_{n1} & x_{n2} & . & . & . & x_{np} \end{bmatrix}$$

and the vector valued parameter β is given by

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{bmatrix}$$

The expression $logit(\mathbf{p})$ now is a vector of the form

$$\mathbf{logit(p)} = \begin{bmatrix} logit(p_1) \\ logit(p_2) \\ . \\ . \\ . \\ logit(p_N) \end{bmatrix}$$

corresponding to the N observations $Y_i, i = 1, 2, ..., N.$

2.3.1 Application of Logistic regression to the respiratory infection data

First, a model was fitted assuming the observations within an individual are independent. In the analysis the reference levels for categorical variables or factor variables were centre two, active treatment, female category, infected status and time period four for the categorical variables centre, treatment, sex, baseline status of infection and time. The results from this analysis based on maximum likelihood estimation and Type 3 analysis effects are shown in Tables 2.1 and 2.2 respectively. The result in Table 2.1 indicates that the covariate effects centre, treatment, age and baseline disease status are significant while sex and time are not significant at 5% significance level. The type 3 table of results also give the same conclusion that it is the centre, treatment, age and baseline disease status effects that are significant at 5% significance level. The result for the adjusted odds ratio estimates are shown in Table 2.3.

The Table 2.3 results show that the odds of infection for individual in centre 1 was 1.362 times that of individuals in centre 2. The odds of infection for an individual who received the placebo treatment was 1.888 times that for individual who received the active treatment. The odds of infection for a male individual was 1.088 times that for a female individual. The odds of infection for an individual who is not infected at baseline is 0.399 times that of an individual who is infected at baseline. The odds of infection at times 1, 2 and 3 are respectively 0.863, 1.085 and 0.851 times that at time 4. It is however clear that the above approach is inferior because it ignores dependence between observations within an individual.

Table 2.1: Analysis of maximum likelihood estimates

parameter	Estimate	Standard Error	Pr>Chisq
intercept	-1.0904	0.3446	0.0016
Centre	0.3086	0.1189	0.0095
treatment	0.6357	0.1176	$< .0001$
Sex	0.0839	0.1468	0.5675
Age	0.0189	0.0088	0.0316
Baseline	-0.9188	0.1198	$< .0001$
time1	-0.1470	0.1918	0.4434
time2	0.0820	0.1908	0.6672
time3	-0.1619	0.1921	0.3993

Table 2.2: Type 3 analysis of effects

parameter	DF	Wald chi-squared	Pr>Chisq
Centre	1	6.7315	0.0095
treatment	1	29.2065	$< .0001$
Sex	1	0.3269	0.5675
Age	1	4.6193	0.0316
Baseline	1	58.8591	$< .0001$
time	3	2.1688	0.5381

The purpose of assuming independence in the analysis is to show the effect of not properly accounting for the correlation structure in the analysis of correlated non-Gaussian data such as the current one. These results will be discussed in relation to estimates obtained when an appropriate correlation structure is assumed. Lack of not correctly accounting for the inherent correlation in the data increases type I error for testing the significance of the various effects.

Table 2.3: The Adjusted odds ratio estimates for logistic regression model

parameter	odds ratios	95% Confident Interval	
Centre	1.362	1.078	1.719
treatment	1.888	1.499	2.378
Sex	1.088	0.816	1.450
Age	1.019	1.002	1.037
Baseline	0.399	0.315	0.505
time1	0.863	0.593	1.257
time2	1.085	0.747	1.578
time3	0.851	0.584	1.239

2.4 Marginal models for correlated data

Now consider an extension of the cross-sectional scenario defined in section 2.3 to a longitudinal or clustered data structure where an experimental or study unit is observed repeatedly for a number of occasions.

Let the observations for individual i be Y_{ij} for $j = 1, 2, ..., n$ and $i = 1, 2, ..., N$. Let $\mathbf{Y}_i$ denote the n dimensional vector of observations from individual i assuming each individual contributes n outcomes. More generally an individual i contributes n_i outcomes where the $n_i's$ are not necessarily equal $\forall i$, $i = 1, 2, ..., N$.

In a marginal model the regression of the response on explanatory variables is modelled separately from within person correlation (Diggle et al., 2002). In the regression model the marginal expectation denoted as, $E(Y_{ij})$ is modelled as a function of explanatory variables $x_{ij1}, x_{ij2}, ..., x_{ijp}$. To state the general marginal model formely, let $\mu_{ij} = E(Y_{ij} \mid x_{ij})$ for $i = 1, 2, ..., N$ and $j = 1, 2, ..., n$. Diggle et al.(1994) and later in Diggle et al.(2002) point out that modelling association among binary responses with a correlated relation has a disadvantage and they recommend using the odds ratio instead. The data influences the range of the correlation since the estimate of the correlation between the j^{th} and k^{th} response for binary data are constrained by the means, $\mu_{ij} = Pr(Y_{ij} = 1)$. Consider

$$Corr(Y_{ij}, Y_{ik}) = \frac{Pr(Y_{ij} = 1, Y_{ik} = 1) - \mu_{ij}\mu_{ik}}{\sqrt{\mu_{ij}(1 - \mu_{ij})\mu_{ik}(1 - \mu_{ik})}} \tag{2.17}$$

This mean $Pr(Y_{ij} = 1, Y_{ik} = 1)$ is constrained to satisfy (Prentice,1988)

$$max(0, \mu_{ij} + \mu_{ik} - 1) < Pr(Y_{ij} = 1, Y_{ik} = 1) < min(\mu_{ij}, \mu_{ik}) \tag{2.18}$$

The odds ratio defined below appears to be a more natural choice for modeling the association in binary data as they are not constrained by the means (Liang et al., 1992).

$$OR(Y_{ij}, Y_{ik}) = \frac{Pr(Y_{ij} = 1, Y_{ik} = 1)Pr(Y_{ij} = 0, Y_{ik} = 0)}{Pr(Y_{ij} = 1, Y_{ik} = 0)Pr(Y_{ij} = 0, Y_{ik} = 1)}$$

whose values in the range $(0, \infty)$, where a value greater than one indicates an association. The measure is unconstrained on the log-scale which is the most commonly used in logistic regression. Under GEE the correlations are treated as nuisance parameters and the use of correlations versus odds ratios usually has little influence on the inference on β, the regression parameters for the marginal mean model. Liang and Zeger (1986) parameterize association in terms of correlation and use moment estimation for the unknown correlation. Let $g(\mu_{ij})$ denote the appropriate link function corresponding to the mean of the random variables Y_{ij} whose realized value is y_{ij}. Often the canonical link function is often a good candidate. The dependence on covariates is achieved through the model.

$$g[E(Y_{ij} \mid x_{ij})] = x_{ij}^{'}\beta \tag{2.19}$$

where x_{ij} are covariates for individual i measured at occasion j and β is a vector of fixed effect parameters. By marginal expectation we mean the average response over the sub-population that share common covariate values.

Fitting marginal models can be quite involving because the marginal association parameters are highly constrained. A marginal model has the following assumptions:

Firstly the marginal expectation $E(Y_{ij}) = \mu_{ij}$ depends on explanatory variables x_{ij} through the relation $g(\mu_{ij}) = x_{ij}^{'}\beta$ where g is the known link function such as the logit for binary responses, log-link for count data, identity link for Gaussian data and so on.

Secondly the marginal variance depends on the marginal mean according to the relation $var(Y_{ij}) = \phi v(\mu_{ij})$ where $v(.)$ is known as the variance function and ϕ is a scale or dispersion parameter which may need to be estimated if its unknown.

Thirdly the correlation between Y_{ij} and Y_{ik} for $j \neq k$ is a function of the marginal means and perhaps of additional vector of parameters α that is $Corr(Y_{ij}, Y_{ik}) = \rho(\mu_{ij}, \mu_{ik}, \alpha)$ where $\rho(.)$ is a known correlation function. The marginal regression coefficient β have similar interpretation as coefficients from a cross-sectional analysis. Marginal models can be viewed as an extension of the generalized linear model (McCullagh and Nelder, 1989) for independent data to correlated non-Gaussian data. To illustrate its application consider the problem of assessing the dependence on vitamin A of a respiratory infection in children. Let x_{ij} indicate whether or not a child is vitamin A deficient (1=yes; 0=no) and let $\mu_{ij} = E(Y_{ij})$. Then the marginal model specification corresponding to the above scenario is given by

$$logit(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1 x_{ij}$$

and

$$Var(Y_{ij}) = \mu_{ij}(1-\mu_{ij})$$

and

$$Corr(Y_{ij}, Y_{ik}) = \alpha, j \neq k.$$

The interpretation of the transformed regression coefficient $\theta_0 = \exp(\beta_0)$ is that it gives the ratio of the frequency of infected to uninfected children among the sub population that is not vitamin A deficient. The parameter $\theta_1 = \exp(\beta_1)$ is the odds of infection among vitamin A deficient children divided by the odds of infection among children replete with vitamin A. That is $\exp(\beta_1)$ gives the odds ratio of infection comparing vitamin A deficient to vitamin A replete children. Thus the advantage of the above extension to the GLM regression model is that we are able to estimate our parameters of interest with more reliable standard errors for interval estimation because now dependence between observations within an individual is correctly accounted for. In addition‘ several covariates can be included in the model and regression coefficients interpreted conditional on other covariates being fixed as in a Gaussian multiple regression but now on the logit link transformed scale. The key modification is that we take into account the correlation between observations. The most popular method of dealing with correlation in marginal models is the generalized estimating equation approach (Liang and Zeger, 1986; Zeger and Liang, 1986).

2.5 Generalized Estimating Equations Theory

Generalized estimating equations (GEE) were introduced by Liang and Zeger (1986) as a method of dealing with correlated data in the case of non-Gaussian clustered responses or outcomes such as in longitudinal studies. The GEE methodology for the analysis of repeated measurements is a marginal model. The approach is an extension of the quasi-likelihood formulation (Wedderburn, 1974) to longitudinal data analysis. The GEE method is also viewed as semi-parametric in nature because the estimating equations are derived without full specification of the joint distribution of within subject or unit observations. However the most important development in the GEE methodology is that it allows the specification of a working correlation structure among the observations, $Y_{i1}, ..., Y_{in}$ within an individual or cluster which is not necessarily the true underlying correlation structure. As in the generalized linear model the GEE method relates the marginal response

$$\mu_{ij} = E(y_{ij})$$

to a linear combination of the covariates as

$$g(\mu_{ij}) = x_{ij}^{'}\beta, \qquad (2.20)$$

where y_{ij} is the response for subject i at time j, x_{ij} is the corresponding $p \times 1$ vector of covariates, β is a $p \times 1$ vector of unknown parameters and g(.) is the link function. The GEE approach also describes the variance of y_{ij} as a function of the mean:

$$Var(y_{ij}) = v(\mu_{ij})\phi,$$

where $v(.)$ is variance function and ϕ is the unknown scale parameter. In terms of the vector y_i of outcomes, $v(.)$ becomes an $n_i \times n_i$ matrix with $v(\mu_{ij})\phi$ in the diagonal but the GEE method also allows for correlation between pairs of observations. If the response variable is binary as in the case of the current analysis the link and variance function are given by

$$g(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right)$$

and

$$v(\mu_{ij}) = \mu_{ij}(1-\mu_{ij})$$

for $\phi = 1$.

In fitting the above model the GEE methodology allows the inclusion of a working correlation among the observations $Y_{i1}, ..., Y_{in}$ to ensure this structural requirement for the data. Over time GEEs have become an important method in the analysis of correlated data. These data sets can arise from longitudinal studies, in which subjects are measured at different points in time, or from clustered data, in which measurements are taken on subjects who share a common cluster characteristic such as belonging to the same litter, household, region and so on. In this case the cluster denotes an experimental unit which is repeatedly observed for example a class using a certain teaching method. The marks scored by students in that class comprise repeated measurements. The method of generalized estimating equations provides consistent estimates of the regression parameters, even when the working correlation structure is mis-specified (Liang and Zeger, 1986). However, the efficiency of a GEE estimate can be affected by the choice of the working correlation model. The GEE approach has been extended to include methods that combine multiple GEEs based on different working correlation models, using the empirical likelihood method (Qin and Lawless, 1994). Analyses show that this hybrid method is more efficient than a GEE using a mis-specified working correlation model. Such an extension to the exponential family has been slow because of the inherent difficultly in modelling the dependence structure in a natural way. In contrast to the classical GLM which assumes independent observations, full likelihood modelling of repeated measures can still be used but the computations quickly become unmanageable. For less than full likelihood approach (Liang and Zeger; 1986) the generalized estimating equation technique became necessary and it is briefly discussed below.

Liang and Zeger(1986) recognized that the standard GLM estimating equation

$$\sum \frac{d\mu_i}{d\beta} v_i^{-1}(y_i - \mu_i) = 0$$

for independent data is immediately applicable for multivariate outcome data. In their extension, y_i and μ_i are now vector valued and v_i is a variance matrix. Thus the estimating equation approach involves a specification of the mean vector and the variance matrix. The Generalized estimating equation approach has some desirable statistical properties that makes it an attractive method for dealing with correlated data. Note that generalized estimating equations reduce to GLM estimating equations for $n_i = 1$. Secondly generalized estimating equations are the maximum likelihood score equation for multivariate Gaussian data or generally if the underlying distribution can fully be specified.

To formally state the estimating equation, let the vector of measurements on the i^{th} subject be $Y_i = (Y_{i1}, ..., Y_{in_i})$ with the corresponding mean vector be $\mu_i = (\mu_{i1}, ..., \mu_{in_i})$ and V_i its covariance matrix as a function of β. Then an extension of the independence estimating equation to the correlated data is given by

$$\sum \frac{d\mu_i}{d\beta} V_i^{-1}(y_i - \mu_i(\beta)) = 0 \tag{2.21}$$

The aim is to be able to estimate β appropriately taking into account of the correlation between measurements from the same unit through the covariance matrix V_i.

2.5.1 Working correlation

Let R_i be an $n_i \times n_i$ working correlation matrix that is fully specified by the vector of parameters α. The covariance matrix of Y_i is modeled as

$$V_i = A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}} / \phi \tag{2.22}$$

The variance function matrix A is an $n_i \times n_i$ diagonal matrix with the j^{th} diagonal element given by $v(\mu_{ij})$. The working correlation $R_i(\alpha)$ is assumed to depend on a set of parameter α. The over dispersion parameter ϕ is assumed to be known or if unknown it is included as one of the parameters to be estimated from the data using known methods such as the method of moments. The working correlation structure $R_i(\alpha)$ is not usually known because it depends on unknown quantities α thus it must be estimated or specified. The unknown quantities α are derived or estimated in terms of the Pearson residuals given by

$$\hat{e}_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{v(\hat{\mu}_{ij})}} \tag{2.23}$$

subject to the structure of the assumed correlation matrix. Thus as a bi-product of GEE estimation $\hat{R}_i(\hat{\alpha})$ is also provided. However it should be noted that $R_i(\alpha)$ is not necessarily the correct correlation and hence the GEE method also allows for V_i to be estimated empirically which gives rise to the so called robust or empirical based standard errors.

2.5.2 Covariance and correlation structures

The covariance structure is not the primary interest of analysis but essential for valid inference. Therefore a lot of effort is usually needed at the beginning of the statistical analysis to assess and determine the best covariance structure to assume for the cluster. Note that in Eq.(2.22) the form of the variance function in the diagonal matrix A is obtained based on the quasi-distribution assumed for the data therefore effort is spared to determine what structure to use for $R_i(\alpha)$. There are several specific choices of the form of the working correlation structures to use for the clustered responses. The four commonly used correlation structures are the compound symmetry (CS), first order autoregressive (AR(1)), Toeplitz (Toep) and the unstructured (UN).

Compound Symmetry (CS) :

The compound symmetry covariance structure assumes non-zero, yet uniform correlations for all pairs of within-subject variables. A similar terminology used for this type of correlation structure is the exchangeable correlation structure. Every observation within an individual or cluster is equally correlated with every other observation from that individual. This choice of covariance structure may not be reasonable with multiple measurements collected over time, since the correlation between two observations Y_{ij} and Y_{ik} for $j \neq k$ most likely exhibit diminishing correlation as the time lag between observations Y_{ij} and Y_{ik} increases. For example if we consider four repeated measures, the compound symmetry covariance structure is given by

$$\mathbf{CS} = \begin{bmatrix} \sigma^2 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix} = \sigma^2 R(\alpha)$$

where σ_1 is the covariance, σ^2 is the variance and $\rho = \frac{\sigma_1}{\sqrt{\sigma^2\sigma^2}}$ gives the correlation between any two observations thus here $\alpha = \rho$.

First order autoregressive (AR(1)) :

The second plausible covariance structure that is relevant for repeated measures in time is the autoregressive structure a term derived from times series analysis that assumes observations are related to their own past values or history through one, two, or a higher order autoregressive (AR) process. A first order autoregressive covariance structure assumes that two observations taken close to each other in time (or space) within an individual tend to be more correlated than two observations taken far apart in time from the same individual or observational unit. The correlation between two measurement that are m time units apart is given by ρ^m. Since $-1 < \rho < 1$ but it is structurally more realistic to assume $0 < \rho < 1$ therefore the greater the power $m \geq 1$ the smaller the correlation. In the case of four repeated measurements covariance structure is

$$\mathbf{AR(1)} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix} = \sigma^2 R(\alpha)$$

Once again here $\alpha = \rho$ meaning only one correlation parameter is needed just as in the case of the CS structure. Note that such a correlation structure also applies to observations repeatedly measured in space. In this case a measure of separation between two observations is the distance between the observations, the correlation d distance units apart is ρ^d. Note that the analogy can be extended to data correlated in space and time leading to a 2-dim correlation structure. A 2-dim correlation structure also arises in agricultural field trials using the so called row-column designs with observation correlated both along rows and columns. However this extension is not the focus of the current project. The decaying correlation structure can be generalized to the case observations Y_{ij} and Y_{ik} measured at time t_{ij} and t_{ik} respectively where now
$Corr(Y_{ij}, Y_{ik}) = \rho^{|t_{ij}-t_{ik}|}$.

Toeplitz (Toep) :

In contrast to the $AR(1)$ structure no assumption of exponential decay is made. The $AR(1)$ structure depends on the single parameter ρ for a complete specification of the correlation but the Toeplitz model has as many parameters as there are distances. In the case of four repeated measurements the Toeplitz covariance structure is

$$\mathbf{TOEP} = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{bmatrix} = \sigma^2 R(\alpha)$$

In this case the correlation parameter vector is $\alpha = (\rho_1, \rho_2, \rho_3)^{'}$

Unstructured (UN) :

The most flexible covariance structure leads to the unstructured pattern of correlations which assumes unconstrained pair-wise correlations where each correlation is estimated from the data (the most complex model). No assumption is made about the relative magnitude of the correlation between any two pairs of observations. All the variances and covariances are different. In the case of four repeated measurements the covariance structure is given by

$$\mathbf{UN} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{bmatrix}$$

An analysis that uses an UN covariance matrix will be less powerful than an analysis that uses a less parametric but more realistic structure. The problem though is knowing aprior what the realistic or parsimonious structure is. The unstructured type of correlation has an immediate disadvantage because it increases the number of parameters to estimate in the overall model hence causing possible non-convergence problems, particularly those associated with boundary values. A strategy used to reduce the number of parameters is to assume that $\sigma_i^2 = \sigma^2 \ \forall i$ along the diagonal. So that α is the set of all k pairwise correlation where $k = \frac{n(n-1)}{2}$. The advantage of the UN structure is that the user does not need to labour too much deciding on the structure to use or not use but with a cost to pay in terms of the size of parameters to estimate. The UN structure can provide a more preferred fit if the dimension of $R_i(\alpha)$ is not large.

2.5.3 Model Fit Analysis under the QIC statistics

The Quasilikelihood for the Independence model Criterion (QIC) statistic was proposed by Pan (2001) and further discussed by Hardin and Hilbe (2003) as the analogous to the familiar AIC (Akaike's Information Criterion) statistic which was used for the comparison of the models fit with likelihood-based methods. Since the generalized estimating equations (GEE) method is not a likelihood-based method, the AIC statistic is not available. QIC statistics can be used for the selection of an appropriate working correlation structure for a GEE model. The QICu is another statistic used to select an appropriate model fit using GEE models. The QIC and the related QICu statistic can be used to compare GEE models (Hoi-Jeong Lim, 2011). Note that the QICu is an approximation to the QIC that can only be used for variable selection in the model. In the current application we use the QIC to select the appropriate correlation structure. Because we were comparing different correlation structures using the same GEE model. Therefore QICu statistic can not be used in the current application for the selection of a working correlation structure. When using QIC or QICu to respectively compare two structures or two models, the model with the smaller statistic is preferred (Hoi-Jeong Lim, 2011).

2.6 Application of GEE to the respiratory infection data

The aim of this section is to apply and discuss the results from the GEE model to the respiratory infection data using the different correlation structures. In this application a series of models relating the logit of the probability of disease over time and measured covariates are fitted. The model that was first fitted contained all the main effects. Only those terms that were found to be significant were retained with the suitable correlation structure. The main effect terms that we considered are age, sex, treatment, baseline outcome, time and centre. The application was carried out using SAS proc GENMOD which is an inbuilt procedure in SAS capable of fitting both generalized linear models and their extensions to GEEs allowing for the specification of the quasi-distribution and the correlation structure for clustered data. The simplest model assumes the independence correlation and the results are given in Table 2.5. Such an assumption is not realistic because repeated measurements from the same individual are bound to be correlated. Note that a model that assumes independence can also be fitted using proc LOGISTIC in SAS. However, results under the independence structure assumption are internally used to provide initial estimates in the iterative GEE estimation algorithm. Note these results are similar to Table 2.1. More realistic correlation structures considered are the unstructured (UN), AR(1) and the compound symmetry (CS) structures. Parameter estimates derived from the generalized estimating equations are based on two different types of the standard errors namely the empirical standard errors and the model based standard errors. Model based standard errors are those calculated using the assumed correlation (hence covariance matrix) structure while empirical standard errors are those based on an empirical covariance structure estimated out directly from the data itself. This is also called the sandwich structure. The results in Table 2.4 shows that the best correlation structure for the GEE model is unstructured correlation structure because it has the smallest QIC compared to the AR(1) and compound symmetry correlation structures. Although QIC suggest the use of unstructured correlation model the difference between the UN and CS correlation structures is very small. Under these circumstances one would probably prefer to take the simplest model that is the CS correlation structure to simplify the fitting process and to avoid the risk of over fitting. The GEE model results for the three different correlation structures (UN, CS, AR(1)) showing empirical and model based standard errors are shown in Tables 2.6 and 2.7 respectively. Model based standard errors are calculated assuming the suggested correlation structure is the correct one for the data thus are different from the empirical standard errors which are data based. The empirical standard errors attempt to adjust for mis-specification of the correlation structure and rather let the standard errors be estimated empirically. In some cases there is a fairly large difference between model based and empirical standard errors. In a way model based standard errors tend to be too conservative leading to an overestimate of treatment effects. However the impact is not sifted through into the interpretation of the results. It should be noted that in comparison to any other assumed correlation structure standard errors under the independence assumption in Table 2.5 are different from those in Tables 2.6 and 2.7 and are infact generally much smaller which is why the GEE approach is important in accounting for any inherent correlation that may exist between observations from the same individual or cluster. This helps to correct for the inflated type I error when making inference about effect parameters. The score statistics for Type 3 GEE analysis are given in Table 2.8. The result for the adjusted odds ratio estimates for the GEE model that are given in Table 2.9 were obtained under the unstructured correlation structure. Overall there are significant treatment and baseline effects. Centre and age are both not significant and also there is no significant time effect in the occurrence of the respiratory infection. The result in Table 2.9 show that the centre estimate is 0.63. This means the log odds ratio is 0.63 and the odds ratio is 1.88. The reference centre is centre 2 which means that the odds of infection for an individual from centre 1 is 1.88 ((95%CI=0.95, 3.72), p-value=0.07) times that of an individual from centre 2. The log odds ratio corresponding to the treatment estimate is 1.19 which means the odds ratio is 3.29. The reference treatment is the active therefore this means that the risk of disease for an individual who receives the placebo treatment is 3.29 ((95%CI=1.70, 6.37), p-value=0.0004) times that of an individual who receives the active treatment. The log odds ratio corresponding to the sex estimate is 0.13 and odds ratio is 1.14. The reference sex is female which means that the male individuals were 1.14 ((95%CI=0.48, 2.67), p-value=0.77) more likely to be infected with respiratory disease than female individuals. For a unit increase in baseline age log(OR) increases by 0.02. This means that the odds of infection for an individual aged a is 1.02 ((95$\%$CI=0.99, 1.04), p-value=0.19) that of an individual aged $(a+1)$. This is an indication of a non-significant age effect on the probability infection. The baseline estimate is -1.85 and the odds ratio is 0.16. The reference level is the infected state this mean that the odds of infection for an individual who is not infected at baseline is 0.16 ((95$\%$CI=0.08, 0.31), p-value$< .0001$) times that of an individual who is infected at baseline. This means that an individual who is initially not infected at baseline is at lower risk of infection. Alternatively one can think of those initially infected at baseline to be more frail than those not infected. The inclusion of the baseline term in the model helps to correct for this disparity when interpreting the treatment effect. The odds of infection at time 1 is 0.69 ((95$\%$CI=0.42, 1.13), p-value=0.14) times the odds of infection at time 4. The odds of infection at time 2 is 0.87 ((95$\%$CI=0.53, 1.41), p-value=0.56) times the odds of infection at time 4 and the odds of infection at time 3 is 0.68 ((95$\%$CI=0.44, 1.06), p-value=0.09) times the odds of infection at time 4. However there is no evidence of any significant time effect. The results show that baseline and treatment effect under all the correlation structures are significant regardless of whether empirical or model based standard errors are used and centre tending to be significant at 5% significant level. The unstructured and compound symmetry correlation structures have their empirical standard errors slightly closer to the model based standard errors than AR(1) correlation structure. An important observation is that comparing models that account for correlation (UN, CS, AR(1)) and those based on the independence assumption (Tables 2.1 and 2.5) we see that age and centre are significant under the independence assumption but not so under the more realistic models. This emphasizes the fact and point made earlier that models that do not correctly account for correlation can lead to very misleading inference and interpretation or conclusion. However it should be noted that the standard errors in Table 2.5 are at least as large as those in Table 2.1 for all effects. Thus a model with some correlation gives more realistic standard errors than a model assuming no correlation at all.

Table 2.4: QIC goodness of fit statistic for GEE

	UN	CS	AR(1)
QIC	517.0019	517.1191	517.5707
QICu	506.5477	506.3774	506.6944

Table 2.5: Initial parameter estimate assuming the independence structure

parameter	DF	Estimate	Standard Error	Pr>Chisq
intercept	1	-0.9730	0.5237	0.0632
Centre	1	0.6171	0.2379	0.0095
treatment	1	1.2714	0.2352	$< .0001$
Sex	1	0.1679	0.2936	0.5675
Age	1	0.0189	0.0088	0.0316
Baseline	1	-1.8375	0.2395	$< .0001$
time1	1	-0.3738	0.3124	0.2314
time2	1	-0.1448	0.3108	0.6413
time3	1	-0.3887	0.3127	0.2139

Table 2.6: GEE parameter estimates and empirical standard errors

	UN			CS			AR(1)		
parameter	Est	SE	Pr> $[z]$	Est	SE	Pr> $[z]$	Est	SE	Pr> $[z]$
intercept	-0.8159	0.7145	0.2535	-0.9476	0.7216	0.1891	-0.8823	0.7219	0.2216
Centre	0.6308	0.3489	0.0706	0.6263	0.3519	0.0751	0.6813	0.3511	0.0523
treatment	1.1900	0.3379	0.0004	1.2237	0.3386	0.0003	1.1717	0.3429	0.0006
Sex	0.1290	0.4354	0.7671	0.1652	0.4376	0.7059	0.1555	0.4419	0.7249
Age	0.0167	0.0128	0.1925	0.0187	0.0130	0.1496	0.0170	0.0129	0.1850
Baseline	-1.8485	0.3421	$< .0001$	-1.8069	0.3450	$< .0001$	-1.8230	0.3457	$< .0001$
time1	-0.3780	0.2537	0.1363	-0.3739	0.2522	0.1382	-0.3736	0.2530	0.1397
time2	-0.1450	0.2504	0.5625	-0.1443	0.2493	0.5628	-0.1404	0.2498	0.5740
time3	-0.3878	0.2261	0.0863	-0.3873	0.2254	0.0857	-0.3839	0.2252	0.0883

Table 2.7: GEE parameter estimates and model based standard errors

	UN			CS			AR(1)		
parameter	Est	SE	Pr> $[z]$	Est	SE	Pr> $[z]$	Est	SE	Pr> $[z]$
intercept	-0.8159	0.7153	0.2541	-0.9476	0.7066	0.1799	-0.8823	0.6657	0.1851
Centre	0.6308	0.3404	0.0639	0.6263	0.3357	0.0621	0.6813	0.3137	0.0299
treatment	1.1900	0.3338	0.0004	1.2237	0.3283	0.0002	1.1717	0.3083	0.0001
Sex	0.1290	0.4211	0.7594	01652	0.4148	0.6905	0.1555	0.3891	0.6894
Age	0.0167	0.0126	0.1857	0.0187	0.0124	0.1332	0.0170	0.0116	0.1427
Baseline	-1.8485	0.3427	$<.0001$	-1.8230	0.3147	$<.0001$	-1.8135	0.2952	$<.0001$
time1	-0.3780	0.2579	0.1427	-0.3739	0.2532	0.1398	-0.3736	0.3009	0.2144
time2	-0.1450	0.2488	0.5599	-0.1443	0.2519	0.5667	-0.1404	0.2834	0.6202
time3	-0.3878	0.2284	0.0895	-0.3873	0.2536	0.1266	-0.3839	0.2406	0.1106

Table 2.8: Score statistics for Type 3 GEE analysis

	UN			CS			AR(1)		
Source	DF	Chi-sq	P-value	DF	Chi-sq	P-value	DF	Chi-sq	P-value
Centre	1	2.87	0.0904	1	2.90	0.0884	1	3.42	0.0642
treatment	1	11.22	0.0008	1	12.19	0.0005	1	10.85	0.0010
Sex	1	0.08	0.7717	1	0.14	0.7085	1	0.12	0.7260
Age	1	1.78	0.1826	1	2.27	0.1317	1	1.92	0.1662
Baseline	1	22.59	$<.0001$	1	22.17	$<.0001$	1	22.29	$<.0001$
time	3	4.21	0.2393	3	4.21	0.2398	3	4.21	0.2401

Table 2.9: The adjusted odds ratio estimates for the GEE model

parameter	Estimate	odds ratios	95% confident interval	
Centre	0.6308	1.8792	0.9484	3.7236
treatment	1.1900	3.2870	1.6952	6.3738
Sex	0.1290	1.1377	0.4846	2.6706
Age	0.0167	1.0168	0.9916	1.0427
Baseline	-1.8485	0.1575	0.0805	0.3079
time1	-0.3780	0.6852	0.4167	1.1268
time2	-0.1450	0.8650	0.5295	1.4131
time3	-0.3878	0.6785	0.4356	1.0570

Chapter 3

Transition models

Let Y_{ij} denote a repeated or longitudinal observations measured at time t_{ij} for $i = 1, 2, ..., N$ and $j = 1, 2, ..., n_i$ in a typical longitudinal study involving N individuals. The transition models are considered as an extension of linear or generalized linear models to describing the distribution of each response Y_{ij} conditional on past responses $Y_{ij-1}, ..., Y_{i1}$ and covariates x_{ij} (Diggle et al., 2002). The vector of previous outcomes is here called the history at time t_{ij} which is denoted by $H_{ij} = (Y_{ij-1}, ..., Y_{i1})$. Under transition models the past outcomes are also treated as predictor variables. The transition models are a very specific class of conditional models. In the transition models, a measurement Y_{ij} in a longitudinal sequence is described as a function of previous outcomes. Diggle et al (2002) describe the case where the longitudinal or repeated observations are equally spaced as in the current example. Although the current application is also based on equal spacing and common observation times ($t_{ij} = t_j \forall i$). Other practical applications may involve unequal spacing our different observation times ($t_{ij} \neq t_j$ for some i). Thus transition models are designed to account for dependence between the repeated outcomes by including the previous outcomes as predictors of the current response. Thus in the case of a continuous response the general model can be written as

$$E[Y_{ij} \mid H_{ij}, x_{ij}] = x_{ij}^{'}\beta + h_{ij}^{'}\alpha \tag{3.1}$$

where β is the vector of fixed effect parameters corresponding to x_{ij} the vector of covariates for individual i measured at occasion j. The parameter vector α is generally a q dimensional vector of fixed effects measuring dependence on $q \leq j-1$ previous outcomes and $h_{ij} = (y_{ij-1}, y_{ij-2}, ..., y_{ij-q})^{'}$. Although dependence on the previous observation(s) is here defined for a continuous response such a model is naturally more plausible when dealing with a binary or categorical responses that exhibit a Markov state dependence property (Feller, 1968) through time. The order of a transition model is the number of previous measurements at previous time occasions that are still considered to influence the current outcome Y_{ij}. A model is called pure stationary if the functional form of the dependence is the same regardless of the actual time and length of time at which it occurs. The model is called stationary if instead dependence is only the length of time interval separating Y_{ij} and Y_{iq}, $q \leq j-1$ given by $\tau_q = t_{ij} - t_{iq}$. Note that in $Eq.(3.1)$ the vector H_{ij} contains the actual terms hence the order of the transition model. A transition model of order one means $h_{ij} = Y_{ij-1}$, and a model of order q means $h_{ij} = (Y_{ij-1}, ..., Y_{ij-q})$ where $q = 1, 2, ..., j-1$. The above model formulation can be extended to non-Gaussian categorical responses but the linear predictor is linked to the marginal or conditional mean via an appropriate link function appropriate to the distribution assumed. In the case of a binary outcome denoting say the infection status of a child with a childhood disease, the response Y_{ij} takes the value 1 if a child is infected and $Y_{ij} = 0$ if not infected at time t_{ij}. Such a transition model was also used by Mwambi et al.(2011) to model the respiratory syncytial virus (RSV) infection in a birth cohort of infants in Kilifi, Kenya for unequally spaced outcomes. Ware et al (1988) focuses on the conditional expectation of Y_{ij} given past outcomes, $Y_{i,j-1}, ..., Y_{i1}$. In general the modeler specifies a regression model for the conditional expectation $E[Y_{ij} \mid Y_{ij-1}, ..., Y_{i1}, x_{ij}]$ or a function of it as an explicit function of x_{ij} and the past responses. As an example consider equally spaced binary outcomes with a logit link regression model as in the current analysis. Assuming first order dependence implies

$$\log \frac{Pr(Y_{ij} = 1 \mid H_{ij}, x_{ij})}{1 - Pr(Y_{ij} = 1 \mid H_{ij}, x_{ij})} = x_{ij}^{'}\beta + \alpha Y_{ij-1} \tag{3.2}$$

where α is now just a real constant. The appealing feature of the transition model in $Eq.(3.2)$ is that it combines the assumption about the dependence of Y_{ij} on x_{ij} and the correlation among repeated Y's into a single equation. The correlation between Y_{ij} and Y_{ij-1} is accountable for, through the statistical significant of the regression parameter α. Diggle, Liang and Zeger (1994) define a transition model as one in which correlation among the discrete responses arise because past response explicitly influence the current outcomes. The general form of the transition model given a link function $g(.)$ can be written as

$$g(E(y_{ij} \mid y_{i1}, ..., y_{i,j-1})) = x_{ij}^{'}\beta + \sum_{r=1}^{s} f_r(y_{i1}, ..., y_{i,j-1}; \alpha_1, ..., \alpha_s) \tag{3.3}$$

where $f_1, ..., f_s$ are functions of previous observations (linear, nonlinear and combination of both) and possibly of an unknown parameter vector $\alpha = (\alpha_1, ..., \alpha_s)^{'}$. The conditional variance of y_{ij} given the past outcomes and known covariates is proportional to a known conditional mean function:

$$var(y_{ij} \mid y_{i1}, ..., y_{i,j-1}) = \phi v(E(y_{ij} \mid y_{i1}, ..., y_{i,j-1}, x_{ij})) \tag{3.4}$$

where v is a known variance function and ϕ is an unknown scale parameter. In the current thesis we consider the case of linear dependence as stated in Eq.(3.2). There is a similarity between the assumptions underlying transition models and those behind the ante-dependence covariance structure. However the ante-dependence covariance structure are not the focus in the current work.

3.1 Transition model for the respiratory infection Data

Consider the generalized linear transition model to the respiratory infection data. We model the conditional distribution of Y_{ij} given the past history in order to assess the dependence structure on previous outcomes. In the case of the current analysis we first assume that the probability of current respiratory infection for individual i at the j^{th} visit has a direct dependence on weather or not an individual was infected at visit $j-1$ as well as on explanatory variable x_{ij}. First consider the first order transition model with a logit link given by

$$logit[Pr(Y_{ij} = 1 \mid H_{ij})] = x_{ij}^{'}\beta + \alpha y_{ij-1} \tag{3.5}$$

where α is a regression measuring the effect of the previous disease status on the current one. Therefore the chance of respiratory infection at time t_{ij} not only depends on explanatory variables but also on whether or not an individual had infection at previous visit. The parameter $\theta = \exp(\alpha)$ is the ratio of odds of infection at time t_{ij} among individuals who did to those who did not have infection at the prior visit given x_{ij} is held constant. It is assumed that the two average individuals to compare share a common covariate vector x_{ij}. Since $Y_{ij-1} = 1$ if infected and $Y_{ij-1} = 0$ if not infected at time t_{ij}, the coefficient β is the change in the log odds of infection among individuals who where free of infection at the previous visit. With a binary response Y_{ij} observed at equally spaced time intervals the 2×2 transition matrix whose element are $Pr(Y_{ij} = y_{ij} \mid Y_{ij-1} = y_{ij-1})$ where each of Y_{ij} and Y_{ij-1} may take values 0 or 1 can be used to display the four possible Markovian types of transitions. The logistic regression Eq.(3.2) above can be used to estimate the transition probabilities as a function of covariates and previous status. For detailed information on the theory of first order transition model including their properties we refer the interested reader to Feller (1968, Vol 1, Pg.372). In the current analysis the significance of second order dependence was also assessed. Such a model is precisely the so called autoregressive model of order two or AR(2), which can be generalized to AR(q) where $q = 1, 2, ..., j-1$. The only additional structure is that the current models also include covariate/dependence.

3.2 Application to respiratory infection data

There are two distinct approaches of how to implement the transition model to disease outcome data over time. These are the marginal or population averaged and conditional or individual specific models. This section is aimed at investigating whether the current infection status for an average individual is dependent on a previous time status using the methods of generalized estimating equations (GEE). The purpose of using the GEE method was to explain the dependence of the response on the measured or observed covariate for clustered data so that we can also account for the correlation between the responses. Under transition models the previous outcomes or responses are also included as covariates in the GEE model. These type of models as described in Chapter 2 fall under the marginal a population averaged models. To achieve this modelling strategy lagged variables were created to identify the previous status of an individual in relation to the current disease status over time. Only first and second order lagged dependence were investigated in the current analysis. This necessitated the creation of two lagged variables based on the order (one and two) dependence. The lagged variables were then included as predictor variables in addition to the set of covariates in the previous model in Chapter 2. The models were fitted using SAS proc GENMOD. The result of the first, second and first jointly with second order dependence models are given in Tables 3.1-3.5. The result for fitting first and second order models separately shows that treatment and baseline status are still significant while centre and age are not significant at 5% significance level. Separate conditional models show that Y_{ij} is significantly depends on Y_{ij-1} and not on Y_{ij-2}. An interesting observation is that when both first and second order lag variables were included in the model age is now significant. The first order lag variable was significant while the second order lag variable was not. From the score statistics for Type 3 GEE analysis (Table 3.4) we see that the treatment and baseline status are significant at 5% significance level using first order lag dependence. The first order model shows that Y_{ij-1} is significant at 5% significance level which indicates that an individual's current status is significantly dependent on the immediate past status as it is expected in many respiratory disease processes (Mwambi et al., 2011), while the second order model shows that Y_{ij-2} is not significant at 5% level of significance. The result from fitting both first and second order lagged variables in the same model indicate that Y_{ij-1} is still significant while Y_{ij-2} is not significant at 5% significance level. Thus we can conclude that only the immediate past information on infection status is important in explaining the current respiratory status of an individual. It is important to state that Y_{ij-2} is not significant at 5% level of significance implying that Y_{ij-1} is more informative about the current status than lags of order 2 and higher. However while age is not significant when lag variables are fitted separately age becomes significant when both Y_{ij-1} and Y_{ij-2} variables are both included in the same model. It may appear Y_{ij} depend on age indirectly through the presence of Y_{ij-1} and Y_{ij-2} in the model. The results in Table 3.5 for fitting first order model shows that the odds of infection among individual who was free of infection at previous visit is 2.412 times that of an individual who was infected at previous visit. The results in Table 3.5 for fitting second order model shows that the odds of infection among individual who was free of infection at prior two step visit is 0.978 times that of an individual who was infected at prior two step visit. In the first order model the odds ratio of comparing placebo and active treatment is 3.114 implying that the odds of infection among individual who was on placebo is 3.114 times that of an individual who was on active treatment. Since the reference level for baseline status was the infected state this means that the odds of infection for an individual who was not infected and disease free at baseline and previous visit is 0.239 that of an individual that was infected at baseline and previous visit. The odds ratios are higher for first order than second order dependence, but higher if both first and second order term are allowed. The disadvantage associated with transition models is that for an order q transition model we loose information for the first q time occasions or visits. This mean either we condition on the first q time points or assume some distribution for $(y_{i1}, y_{i2}, ..., y_{iq})$ which introduces some uncertainty.

Table 3.1: GEE parameter estimates and empirical standard errors including both first and second order transition effects

First order model				Second order model			
parameter	Est	SE	Pr> $[z]$	parameter	Est	SE	Pr> $[z]$
intercept	-2.0716	0.7347	0.0048	intercept	-1.7767	0.9240	0.0545
Centre	0.5999	0.3108	0.0536	Centre	0.6238	0.3788	0.0996
treatment	1.1360	0.3039	0.0002	treatment	1.3501	0.3849	0.0005
Sex	0.1975	0.3883	0.6110	Sex	0.2709	0.4627	0.5582
Age	0.0185	0.0116	0.1102	Age	0.0261	0.0144	0.0697
Baseline	-1.4331	0.3207	$<$.0001	baseline	-1.6822	0.4295	$<$.0001
time	0.1381	0.0837	0.0988	time	0.0706	0.1274	0.5795
Y_{ij-1}	0.8804	0.2463	0.0004	Y_{ij-2}	-0.0227	0.3345	0.9459

Table 3.2: GEE parameter estimates and model based standard errors including both first and second order transition effects

	First order model				Second order model		
parameter	Est	SE	Pr> $[z]$	parameter	Est	SE	Pr> $[z]$
intercept	-2.0716	0.7076	0.0034	intercept	-1.7767	0.8644	0.0398
Centre	0.5999	0.3111	0.0538	Centre	0.6238	0.3629	0.0856
treatment	1.1360	0.3082	0.0002	treatment	1.3501	0.3589	0.0002
Sex	0.1975	0.3848	0.6077	Sex	0.2709	0.4452	0.5428
Age	0.0185	0.0116	0.1096	Age	0.0261	0.0134	0.0518
Baseline	-1.4331	0.3330	$<$.0001	baseline	-1.6822	0.3659	$<$.0001
time	0.1381	0.0910	0.1289	time	0.0706	0.1229	0.5658
Y_{ij-1}	0.8804	0.2531	0.0005	Y_{ij-2}	-0.0227	0.2841	0.9364

Table 3.3: GEE parameter estimates and empirical and model based standard errors for first and second order transition model

	Empirical SEs			Model based SEs		
parameter	Est	SE	Pr> $[z]$	Est	SE	Pr> $[z]$
intercept	-3.0951	0.8483	0.0003	-3.0951	0.8746	0.0004
Centre	0.4891	0.3022	0.1056	0.4891	0.3037	0.1073
treatment	0.9161	0.3058	0.0027	0.9161	0.2995	0.0022
Sex	0.3014	0.3532	0.3935	0.3014	0.3740	0.4203
Age	0.0239	0.0113	0.0341	0.0239	0.0112	0.0326
Baseline	-0.7553	0.3568	0.0343	-0.7553	0.3489	0.0304
time	0.1198	0.1787	0.5027	0.1198	0.1736	0.4903
Y_{ij-1}	1.8814	0.2898	$<$.0001	1.8814	0.3044	$<$.0001
Y_{ij-2}	0.5737	0.3727	0.1237	0.5737	0.3474	0.0986

Table 3.4: Score statistics for Type 3 GEE analysis for first, second and first and second order models

1^{st}order			2^{nd} order			1^{st} and 2^{nd} order		
source	χ^2	Pr> χ^2	source	χ^2	Pr> χ^2	source	χ^2	Pr> χ^2
Centre	3.20	0.0735	Centre	2.58	0.1084	centre	2.60	0.1071
treatment	13.60	0.0002	treatment	12.39	0.0004	treatment	8.29	0.0040
Sex	0.24	0.6255	Sex	0.34	0.5608	Sex	0.69	0.4048
Age	2.27	0.0999	Age	3.79	0.0515	Age	4.90	0.0269
Baseline	18.33	$< .0001$	Baseline	18.20	$<.0001$	Baseline	2.37	0.1233
time	2.51	0.1132	time	0.32	0.5737	time	0.46	0.4988
Y_{ij-1}	6.25	0.0124	Y_{ij-2}	0.00	0.9547	Y_{ij-1}	5.18	0.0228
						Y_{ij-2}	1.39	0.2381

Table 3.5: Odds ratio estimates for different type of transition models

	1^{st}order		2^{nd} order		1^{st} and 2^{nd} order	
parameter	estimate	Odds ratio	estimate	odds ratio	estimate	odds ratio
Centre	0.5999	1.822	0.6238	1.866	0.4891	1.631
treatment	1.136	3.114	1.3501	3.858	0.9161	2.500
Sex	0.1975	1.218	0.2709	1.311	0.3014	1.352
Age	0.0185	1.019	0.0261	1.027	0.0239	1.024
Baseline	-1.4331	0.239	-1.6822	0.186	-0.7553	0.470
time	0.1381	1.148	0.0706	1.073	0.1198	1.127
Y_{ij-1}	0.8804	2.412			1.8814	6.562
Y_{ij-2}			-0.0227	0.978	0.5737	1.775

Chapter 4

The subject-specific approach

4.1 Introduction

In this chapter we consider the problem of modelling extra variability for non-normal data. In the current chapter we also aim to discuss why practical problems may need extra modelling effort to account for extra-variability. There are two possible approaches that can be used. The first is by means of incorporating a dispersion parameter to relax the rigid mean to variance function relationship and secondly by means of random effects terms in the linear predictor. However the first approach makes the GLM not to conform to a likelihood formulation because the extra parameter modified distribution no longer has the classical exponential family structure. This problem was first addressed by Nelder and Pregibon (1987) using the quasi-likelihood formulation to allow the estimation of the extra dispersion parameter. The second approach is more robust particularly if faced with correlated data. In the next sub-section the concepts of under and over-dispersion are demonstrated with reference to binomial model then the link with GLMs dealt with in sub-section 4.1.2

4.1.1 Under-dispersion and Over-dispersion in the binomial model

In this section first we aim to distinguish between overdispersion and under dispersion. Over-dispersion means the presence of greater variability (statistical dispersion) in the data than would be expected based on a given standard probability model for the type of response Y under consideration. Overdispersion is a very common feature in applied data analysis because in practice, populations are frequently heterogeneous contrary to the assumptions implicit within widely used simple parametric models. Over-dispersion occurs when the observed variance is higher than the variance of a theoretical model such as the variance equal to the mean (μ) in the case of the Poisson probability model or the variance equal to $\mu(1-\mu)$ in the case of a Bernoulli probability model where $\mu = P(Y=1)$.

Under-dispersion occurs when there is less variation in the data than expected. To explain this concept suppose $Y_1, ..., Y_n$ are independent Bernoulli outcomes with probabilities $p_1, ..., p_n$. Let $X = \Sigma Y_i$, which gives the total number of successes out of the n trials. It follows that

$$E[X] = \sum p_i = n\theta$$

where $\theta = \frac{\sum P_i}{n}$ but we can easily show that

$$Var(X) = n\theta(1-\theta) - n\sigma_p^2$$

where

$$\sigma_p^2 = \frac{1}{n}\{\sum_{i=1}^{n} p_i^2 - \frac{(\sum_i^n p_i)^2}{n}\}.$$

Thus the important message here is that allowing for heterogeneity in the individual $Bernoulli$ probabilities produces less than standard binomial variance also known as under-dispersion.

To demonstrate over-dispersion suppose $Y_1, ..., Y_m$ are independent binomial (n, p_i) random variables and let I be a random variable taking values in (1,2,...,m). Let $W = Y_i$ be a random choice from the $Y_i's$ with equal probability $P(I=i) = \frac{1}{m}$. That is the $Y_i's$ are equiprobable. Such a process produces a mixture of binomial distributions with marginal probability given by

$$\begin{aligned} P(W = y) &= E[P(Y_i = y) \mid I] \\ &= \frac{1}{m}\sum_{i=1}^{m} P(Y_i = y \mid I = i) \\ &= \frac{1}{m}\sum_{i=1}^{m} \binom{n}{y} p_i^y (1-p_i)^{n-y} = \frac{1}{m}\sum_{i=1}^{m} B(n, p_i) \end{aligned}$$

which does not simplify further. However

$$E(W) = E[E(Y_i \mid I)] = \frac{1}{m}\sum_{i=1}^{m} E(Y_i) = \frac{n}{m}\sum p_i = n\theta$$

where $\theta = \frac{\sum p_i}{m}$ and

$$\begin{aligned} Var(W) &= E[Var(Y_I \mid I)] + Var[E(Y_I \mid I)] \\ &= \frac{1}{m}\sum_i Var(Y_i) + \frac{1}{m}\sum_i E^2(Y_i) - \Big(\sum_i \frac{E(Y_i)}{m}\Big)^2 \\ &= \frac{1}{m}\sum_i np_i(1-p_i) + \frac{1}{m}\sum_i (np_i)^2 - (n\theta)^2 \\ &= n\theta(1-\theta) + n(n-1)\sigma_p^2 \end{aligned}$$

where σ_p^2 is the variance among the p_i 's as defined in the case of sums of Bernoulli trials above. Here we have greater than standard binomial variation which implies over-dispersion. Extra-binomial variance is always implied when we sample from a mixture distribution or population. For example if families in the population have different probabilities p_i 's for an outcome of interest Y_i, then a random sample of families will exhibit values with extra-binomial variance. A wrong binomial assumption can impact the intended inference negatively by producing wrong standard errors. In this example a family in the population can also be regarded as a cluster. Thus a mixture model can be more generally regarded as a model to account for extra cluster to cluster heterogeneity. The beta-binomial distribution (Skellam, 1948) was one of the earliest approach to include this extra variability in a likelihood setting.

4.1.2 The link between over-dispersion and generalized linear models for binary data

Apart from modelling dispersion and other unobserved effect we need models that can also account for observed or measured covariates. The random effects model assumes that variability among clustered binary responses exceeds what would be expected due to the binomial variation alone. The random effects models were developed to account for this so called extra -binomial variation. Let $Y_{i1}, ..., Y_{in_i}$ represent the n_i binomial responses or outcomes from cluster or subject i. The beta-binomial distribution assumes that

• Conditional on μ_i the responses $Y_{i1}, ..., Y_{in_i}$ are independently distributed as binomial with common probability of success μ_i where $0 < \mu_i < 1$

• The μ_i follow a beta distribution with mean μ and variance $\delta\mu(1-\mu)$. It follows that unconditionally, the sum of successes $Y_i = \sum_{j=1}^{n_i} Y_{ij}$ from a cluster $Y_{i1}, ..., Y_{in}$ has a beta-binomial distribution with

$$E(Y_i) = n_i\mu_i$$

and

$$Var(Y_i) = n_i\mu_i(1-\mu_i)[1+(n_i-1)\delta]$$

The over-dispersion parameter δ is the correlation for each pair of binary response from the same cluster. The beta distribution has been used to model the incidence of a non-infectious diseases in households (Griffiths, 1973). The beta distribution can be extended so that an additional structure including covariate dependence may be imposed on the cluster specific mean μ_i.

For example μ_i might be assumed to depend on cluster-level explanatory variables x_i though a logistic function $logit(\mu_i) = x_i^{'}\beta$. Originally, it was assumed that the beta-binomial distribution required each response from the same cluster to have a common probability μ_i. In the regression set up this required the covariates to be the same for all observations within a cluster such that $x_{i1} = ... = x_{in} = x_i$. However the beta-binomial distribution has been extended to allow the covariate to vary within a cluster (Rosner,1984). His model for clustered data is formally equivalent to the following n_i logistic regression:

$$logitPr(Y_{ij} = 1 \mid y_{i1}, ..., y_{ij-1}, y_{ij+1}, ..., y_{in}, x_{ij}) = log\Big(\frac{\theta_{i1} + w_{ij}\theta_{i2}}{1 - \theta_{i1} + (n_i - 1 - w_{ij})\theta_{i2}}\Big) + x_{ij}^{'}\beta^*$$

where $w_{ij} = y_{i.} - y_{ij}$, $j = 1, ..., n$ and where, θ_{i1} is the intercept parameter and θ_{i2} characterizes the association between pairs of responses from the same cluster. β^* is a regression coefficient that measures the effect of x_{ij} on y_{ij} which cannot first be explained by the other responses in the cluster. The effect of cluster level covariates may often be attributed to the other observations within the cluster rather than covariates themselves. This class of models are collectively called conditional models. A special case of such a model are disease state transition models where Y_{ij}, $j = 1, ..., n$ denotes the disease state of an individual i observed at time t_{ij} or accassion j (Diggle et al., 2002). It should be noted that observations from the same cluster are by design correlated not independent. Thus analysis and modelling of such processes should account for such inherent correlation otherwise the results from a model assuming independence may be hard to justify. Infact under the marginal model we noted in the application in chapter 2 that ignoring the structural correlation tend to worsen type I error leading to inappropriate conclusions.

4.2 Modelling correlated data using random effects

The aim of this section is to discuss the different approaches of modelling correlated data by means of random effects using the generalized linear mixed model (Breslow and Clayton, 1993). The current data structure on which application will be based can be described as balanced because each individual in this data set contains the same number of repeated observations measured at equal time intervals in all the individuals. The random effects model is most useful when the objective includes the need to make individual specific inferences rather than population averaged effects only. Models with subject or cluster specific parameter are different from population average models because they include terms that are specific to the subject or cluster. The random effects and population averaged models for correlated binary data describe different type of effects of the covariate on the response probabilities (Neuhous,1992). In the subject-specific modelling approach the response or outcome is modelled as a function of covariates and random effects specific to a subject. As defined in classical design of experimental texts (Milliken and

Johnson, 2009) a factor in a model is random if its levels consist of a random sample from a population of all possible levels. Therefore a model is said to be a random effects model if all the factors in the treatment structure are random effects. A factor in a model is fixed if its levels are selected by means of a non-random process or consists of the entire population of all possible levels. Therefore a model is a fixed effects model if all terms in the model are fixed effects. A model containing both random and fixed effects is therefore called a mixed effects model. Strictly speaking in our case we are dealing with a mixed effects model. For example consider a clinical trial to generate data that can be used to compare drugs intended to reduce blood pressure and the two drugs are administered in nine countries. Suppose in each country a random number of patients are chosen to receive one of the two drugs. Blood pressure is then measured before and after the drug in administered. Finally the difference in blood pressure before and after treatment is obtained for each patient in each country. An appropriate model to consider in this case is

$$Y_{ijk} = \mu + \gamma_i + \rho_j + \epsilon_{ijk}$$

for $i = 1, 2,$ $j = 1, 2, ..., 9$ and $k = 1, 2, ..., n_{ij}$ where Y_{ijk} is blood pressure reduction, μ is the overall mean, γ_i is the drug effect, ρ_j is the country j effect and ϵ_{ijk} is error term for individual k in country j receiving drug i and n_{ij} is the number of patients in country j allocated to drug i. The factor drug is a fixed effect because we are interested to compare only two drugs. However the country effect is considered as a random effect because we are not interested to compare only nine countries. The nine countries are here considered as a random sample from a population of countries. The estimation of ρ_j is attainable because we have multiple observation per country. Note that if the difference within individual was not the outcome of analysis and instead we modelled the actual blood pressure Y_{ijkl} where $l = 1$ if the measurement is before the drug and $l = 2$ after the drug is administered we would need an extra individual specific effect U_{jk} for individual k in country j. Such an effect would also be more realistically assumed to be a random effect because the n_{ij} individuals are a random sample from the population in country j. The estimation of such an effect would rely on the two observations within an individual. Under the random and mixed effects model the subject, unit or cluster specific effects are accounted for by parameters which are themselves considered to be random variables. Thus in the case of a continuous response that is assumed to be Gaussian a linear model including random effects can generally be written as

$$E[Y_{ij} \mid b_i, x_{ij}] = x_{ij}^{'}\beta + z_{ij}^{'}b_i \tag{4.1}$$

where Y_{ij} is the j^{th} observation or response observed or measured from the i^{th} unit or individual at time t_{ij} or occasion j, β is the vector of fixed effects parameters, x_{ij} is the vector of covariates for individual i measured at accasion j, b_i denotes a vector of subject specific random effects and z_{ij} is a vector of covariates for the random effects. A commonly used distributional assumption on the random vector b_i is that $b_i \sim iid.N(0, D)$ where D is variance-covariance matrix for the random effects containing the variance components along the diagonal and covariance parameters in the off-diagonal. The analysis of such a model involves the estimation of fixed parameters including both regression and variance components and the inferences on such parameters. This model has been the subject of analysis by many workers. The model was first analyzed in detail by Laird and Ware (1982). Other key references include the two books by Verbeke and Molenberghs (2000) and Diggle et al (2002). In this work the focus is more on non-Gaussian data in particular binomial distributed outcomes. An example of a random effects linear model is a model which naturally introduces a correlation between observations from the same individual or cluster through a shared random parameters for observations from the same cluster. These include the random intercept and slope models as well as the split-plot design in time models (Milliken and Johnson, 2009). As stated random effects are usually specified to follow a Gaussian distribution. However the most important practical criterion of interest in handling any type of the random effects model chosen is whether the joint likelihood of the outcomes Y and the random effects can be solved analytically. Three random effects models that are referred to frequently in linear or generalized linear models are the random intercept, random slope or both random intercept and slope models. The random intercept model is suitable when there is significant individual to individual variability at baseline and common slope for all subjects. The difference between the random intercept and random slope models is that the former allows the level of the response, Y_{ij} to vary over the clusters or groups only after controlling for predictors. Random slope coefficients are added to random intercepts allowing for the effect of the predictors over time to vary over clusters. The random intercept model allows the intercept to vary by cluster, each of which is distributed according to a common specified probability density function. Closed form solutions for parameter estimates is not possible under the generalized linear mixed model. Methods that are commonly used for estimation include the various types of quadrature methods, including adaptive and non-adaptive Gaussian Hermite quadrature. Quasi-likelihood methods including penalized quasi-likelihood (PQL) and marginal quasi-likelihood (MQL) are also available. These methods will be used in the current chapter as estimation methods. As already stated interest in such methods is on subject specific effects. These helps to identify and control for individuals or units that are outliers compared to average population effects. If that is the case then the analysis also includes the estimation of the random effects via for example empirical Bayes estimation and best linear unbiased prediction (BLUP). Both maximum likelihood (ML) and restricted maximum likelihood estimation (REML) can be used. The more important point about REML is that, by maximizing the likelihood only of the residual contrasts, it gives unbiased estimates for the variance parameters, as compared to the downwardly-biased estimates of ordinary maximum likelihood. The latter is frequently preferred because it corrects for degrees of freedom used in the estimation of fixed effects when estimating the variance components which can be as simple as just the measurement variance and as complex as in multi-level and hierarchical sources of variation. In the current analysis the emphasis is restricted to the estimation of fixed parameters only including variance components. As already stated the thesis is methodologically concerned with the analysis based on the inclusion of random effects in non-Gaussian outcomes to account for individual to individual heterogeneity as a way of coping with over-dispersion. Models resulting to generalized linear mixed models (Breslow and Clayton, 1993) including both fixed and random effects accounting for individual to individual variability will be used. The variance components are the parameters that describe the variance of the distribution of random effects. Here the individual effects are regarded as random effects if the individuals included in the analysis constitute a simple random sample from a bigger population of individuals. In general one can consider a model having q sources of variability each associated with a variance component. We start by adopting the standard assumption that the random effects are from a Gaussian distribution with zero mean and variances σ_s^2 $(s = 1, 2, ..., q)$ also referred to as a variance components. Here one is interested in testing whether $\sigma_s^2 = 0$ or not, where $s = 1, 2, ..., q$. Using the GLM framework (Nelder and wedderburn, 1972; McCullagh and Nelder, 1989) we assume that conditional on individual specific effects the outcomes or responses from an individual i are independent and come from a distribution in the exponential family such that Eq.(4.1) now is more generally modified to

$$g(E(Y_{ij} \mid b_i)) = x_{ij}^{'}\beta + z_{ij}^{'}b_i. \tag{4.2}$$

where g is the link function associated with the conditional distribution of Y_{ij} given b_i while x_{ij} and z_{ij} are respectively p and q dimensional vectors of covariates associated with fixed and random effects respectively. Note that Eq.(4.1) is now a special case of Eq.(4.2) with g equal to identity link function. There are two distinct approaches to follow when making inference about the random effects model. The first approach is appropriate when we are interested in a particular subset of regression coefficients, none of which are assumed to vary across subjects. For example suppose we want to estimate the drug effect in a crossover trial and believe that only the intercept and not the drug effect does vary across individuals. We can treat the b_i as a nuisance parameter and average them out of the joint distribution of response and the random effects. This approach is aimed at finding the marginal distribution of the outcome of interest. This then leads to the marginal likelihood that we use for estimation. The second approach is appropriate when subject specific coefficients are themselves of interest or if averaging out the information about random effects discards too much information about important regression coefficients. One approach here is to regard the b_i as if they are an independent sample from the same distribution and aim to estimate both the fixed effects β and random effects b_i under some specific assumptions. Under this conditional likelihood approach we consider only the longitudinal information and use information within subjects to estimate β. In general we can assume a model that allows the analysis to combine both longitudinal and cross-sectional information. The relative weight given to each source is determined by the variability among b_i. When there is a large variability between subjects the analysis should weight the longitudinal information more heavily since comparisons within a subject are likely to be less variable and more precise than comparison among or between subjects. Note that a fundamental assumption of random effects model is that the variance parameters associated with the b_i are independent of the explanatory variables. If this assumption is incorrect the conditional analysis will still give consistent estimates of β but the inference on dispersion or variance parameters may be inconsistent. If this is the case an extension to include the effect of covariates on dispersion may be necessary. This extension means or requires a model for the dispersion parameter. However this additional complexity is not of immediate focus in the current work. However the problem of considering non-normal distribution for the random effects b_i is addressed in Chapter 5. This is an important practical and valid extension due to Lee and Nelder (1996) because the assumption of normal random effects can be to restrictive and structurally invalid. In the current work the interest was also to investigate if the probability of being infected was dependent on time and also whether there was individual to individual variability on this time dependence.

4.3 Formulation of the generalized linear mixed model

These models are a generalization and extension of the well-known generalized linear model (GLM) (McCullagh and Nelder,1989) by introducing random effects to the linear predictor. Generalized linear models, or GLMs, are ubiquitous tools which extend normal based linear regression models to non-normal data and transformable additive covariate effects (McCullagh and Nelder, 1989, Nelder and Wedderburn,1972) into the linear predictor. To extend GLMs to the case of longitudinal or clustered data one needs to account for the inherent between subject variability and within subject correlation of observations. We already discussed an alternative extension to GLMs to accommodate correlated data. This was through the generalized estimating equations (GEE) (Liang and Zeger ,1986) approach which leads to the so called marginal models. The GEE approach was discussed in detail in chapter 2. An alternative way of accounting for correlated data is to define a random effect term which is constant for observations from the same subject, repeatedly measured experimental unit or cluster but variable between experimental or observational units. A standard generalized linear model assumes that the expectation of the response variable Y_{ij} can be written as a function of a linear predictor $\eta = x_{ij}^{'}\beta$ where $x_{ij}^{'}$ and β are vectors of covariates and fixed effect parameters respectively. Now let b_i denote a random effect parameter to account for individual to individual variability or equivalently interpreted as an individual effect parameter. Assuming observations are conditionally independent given the x_{ij}'s, β and b_i the likelihood of the n_i observations $Y_{i1}, Y_{i2}, ..., Y_{in_i}$ from the i^{th} individual (or cluster) is given by:

$$Pr(Y_{i1}, Y_{i2}, ..., Y_{in} \mid x, \beta, b_i) = \prod_{j=1}^{n_i} Pr(Y_{ij} \mid x_{ij}, \beta, b_i) \tag{4.3}$$

The parameter b_i can also be seen as an unobserved random effect due to the i^{th} individual (or cluster). The structure of dependence among the elements of b_i also adds complexity to the dependence among the $Y_{ij}'s$. That is conditionally on the random effects b_i the repeated measurements Y_{ij} could be independent but marginally they are not. Generalized linear mixed models (GLMMs) are an extension of the generalized linear model (GLM) to longitudinal (or clustered data in general) that accommodates correlated and over-dispersed data by adding random effects to the linear predictor. Non-Gaussian models that are encountered frequently are models for repeated or longitudinal measured outcomes that nominally have a binomial or poisson distributions. The generalized linear mixed model conditionally satisfies the exponential family of distribution structure. For consistency assume the covariate set in z_{ij} is a subset of those in x_{ij}. Given b_i and x_{ij} we postulate that the observations $(y_{i1}, ..., y_{in})$ are independent with density function given by

$$f(y_{ij} \mid b_i, \beta) = \exp\left[\frac{y_{ij}\theta_{ij} - \alpha(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right] \tag{4.4}$$

where $\mu_{ij} = E(Y_{ij} \mid b_i)$ is modeled through a linear predictor containing a fixed regression parameter vector β and subject specific parameter vector b_i or more precisely

$$g(\mu_{ij}) = x_{ij}^{'}\beta + z_{ij}^{'}b_i \tag{4.5}$$

We assume a known link function g and known x_{ij} and z_{ij} which are the vector of covariate values associated with fixed and random effects respectively. With a natural or canonical link function the relation $g(\mu_{ij}) = \theta_{ij}$ holds therefore the model becomes

$$\theta_{ij} = x_{ij}^{'}\beta + z_{ij}^{'}b_i$$

The model is completed by assuming that conditionally on the subject-specific effect b_i, the responses Y_{ij} are independent and b_i are normally distributed with a mean of zero and a variance matrix D among the random effects. The function $c(y_{ij}, \phi)$ may or may not depend on ϕ and/or Y_{ij}. By inverting the link function it implies the conditional mean and variance of Y_{ij} are given by

$$\mu_{ij} = E[Y_{ij} \mid b_i] = g^{-1}(x_{ij}^{'}\beta + z_{ij}^{'}b_i) \tag{4.6}$$

and

$$Var(Y_{ij} \mid b_i) = v(\mu_{ij})\phi \tag{4.7}$$

where g and v are respectively, the link and the variance functions. The random effects $b_i, ..., b_N$ are assumed to be independent (but not necessarily) with a common underlying distribution which depends on some unknown parameter vector α which contains the variance and covariance components. The following are some of the most commonly encountered link functions. In all cases assume η_{ij} is the corresponding linear predictor of the form $\eta_{ij} = x_{ij}^{'}\beta + z_{ij}^{'}b_i$.

- The linear mixed model for a Gaussian continuous response with identity link is specified as:

  $$\mu_{ij} = \eta_{ij}$$

  $$Y_{ij} \mid \mu_{ij} \sim N(\mu_{ij}, \sigma_e^2)$$

- The binomial model with logit or probit link for dichotomous or binary responses are specified as

  $$\text{l}ogit(\mu_{ij}) \equiv \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \eta_{ij}$$

  or

  $$\Phi^{-1}(\mu_{ij}) = \eta_{ij}$$

  and

  $$Y_{ij} \mid \mu_{ij} \sim \text{B}ernoulli(\mu_{ij})$$

  where $\Phi^{-1}(.)$ is the inverse standard normal cumulative distribution function or Probit link and $\mu_{ij} = P(Y_{ij} = 1)$

- The poisson model for counts with log link is specified as

  $$\log(\mu_{ij}) = \eta_{ij}$$

  $$Y_{ij} \mid \mu_{ij} \sim \text{p}oisson(\mu_{ij})$$

By introducing random effects in the above models we end up with the corresponding specific GLMM. The dispersion parameter ϕ stated above accounts for extra variability in a model where random effects are not included. In such a case ϕ is known a prior or may require that it be estimated using methods such as moment estimation. For the Poisson and binomial models, the model implied variance function may not be consistent with the actual distribution. In this case quasi-likelihood methods can be used to estimate the dispersion parameter. The problem of likelihood based estimation of ϕ has also been addressed by Nelder and pregibon (1987) in the context of generalized linear models with independent errors. Given the random effects and covariates the variance of the response is a function of the conditional expectation only. In this case we can assume $\phi = 1$. Generalized linear mixed model are often written as non-linear models with an error term (Goldstein, 2003). This is perhaps encouraged by the availability of the popular powerful penalized quasi-likelihood algorithm to be discuss in section $4.4.3$.

4.3.1 Illustration of the generalized linear mixed model with binary and count response

Consider the case where Y_{ij} is a binary response taking value of 0 or 1. A logistic mixed effects model for Y_{ij} is given by the following three-part specification.

$(a.)$ Conditional on a single random effect b_i, the Y_{ij} are independent and have a bernoulli distribution with $\mu_{ij}^c = E[Y_{ij} \mid b_i]$ where μ_{ij}^c notation is used to emphasize the conditional expectation. The variance is given by

$$Var(Y_{ij} \mid b_i) = E[Y_{ij} \mid b_i](1 - E[Y_{ij} \mid b_i]) = \mu_{ij}^c(1 - \mu_{ij}^c) \tag{4.8}$$

$(b.)$ The conditional mean of Y_{ij} depends upon fixed and random effects via the following linear predictor.

$$\begin{aligned}\eta_{ij} &= x_{ij}^{'}\beta + z_{ij}^{'}b_i \\ &= x_{ij}^{'}\beta + b_i\end{aligned}$$

assuming $z_{ij} = 1$ for all $i = 1, 2, ..., N$ and $j = 1, 2, ..., n_i$ with

$$\log\left[\frac{Pr(Y_{ij} = 1 \mid b_i)}{Pr(Y_{ij} = 0 \mid b_i)}\right] = x_{ij}^{'}\beta + b_i = \eta_{ij} \tag{4.9}$$

The conditional mean of Y_{ij} is related to the linear predictor by a logit link function.

$(c.)$ Suppose a single random effect b_i is assumed to have zero mean and variance g_0^2. Then from $Eq.(4.9)$

$$\mu_{ij}^c = Pr(Y_{ij} = 1 \mid b_i) = \frac{\exp(x_{ij}^{'}\beta + b_i)}{1 + \exp(x_{ij}^{'}\beta + b_i)} \tag{4.10}$$

That is μ_{ij}^c is a non-linear function of both the fixed and random effects parameters (Goldstein, 2003).

Let $\mu_{ij} = E(Y_{ij})$ then in the case of the Gaussian linear mixed model we observed that

$$\mu_{ij}^c = x_{ij}^{'}\beta + z_{ij}^{'}b_i$$

and

$$\mu_{ij} = x_{ij}^{'}\beta$$

While for the binomial model μ_{ij}^c is non-linear and hence

$$\mu_{ij} = E(\mu_{ij}^c) \neq \frac{\exp(x_{ij}^{'}\beta)}{1 + \exp(x_{ij}^{'}\beta)} \tag{4.11}$$

In otherwords switching from the condition to the marginal mean model is not direct as it is in the case of the Gaussian model.

Now suppose that Y_{ij} is a count. A log-linear mixed effect model for Y_{ij} is given by the following three-part specification.

(a) Conditional on a vector of random effects b_i. The Y_{ij} are independent and have a Poisson distribution with

$$Var(Y_{ij} \mid b_i) = E(Y_{ij} \mid b_i) = \mu_{ij}^c$$

$(b.)$ The conditional mean of Y_{ij} depend upon fixed and random effects via the following linear predictor:

$$\eta_{ij} = x_{ij}^{'}\beta + z_{ij}^{'}b_i$$

where for example we can assume

$$x_{ij}^{'} = z_{ij}^{'} = (1, t_{ij})$$

as in the case of a random intercept and slope model. A log-link implies that

$$\log[E(Y_{ij} \mid b_i)] = x_{ij}^{'}\beta + z_{ij}^{'}b_i = \eta_{ij}$$

The conditional mean of Y_{ij} is related to the linear predictor by a log link function, that is

$$\mu_{ij}^{c} = \exp(x_{ij}^{'}\beta + z_{ij}^{'}b_i)$$

$(c.)$ The random effects are assumed to have a bivariate normal distribution with a zero mean and 2×2 covariance matrix D. As shown in $Eq.(4.11)$ in the case of a Binomial (Bernoulli) linear mixed

$$\mu_{ij} = E(Y_{ij}) = E(\mu_{ij}^{c}) \neq \exp(x_{ij}^{'}\beta) \tag{4.12}$$

Equation $Eq.(4.11)$ and $Eq.(4.12)$ demonstrate that in dealing with random effects in non-Gaussian probability models some of the nice properties inherent in the normal case do not directly hold in the non-Gaussian case. Therefore a certain degree of caution is necessary in interpretation. We need to integrate inverse link function over the random effects space in order to get the marginal mean.

4.4 Estimation in the generalized linear mixed model

We now briefly describe some possible estimation methods applicable to the GLMM. Most of the methods rely on maximum or conditional maximum likelihood estimation.

4.4.1 Conditional likelihood Estimation

As an example consider the extension of the logistic regression for binary responses to include individual variability using the conditional likelihood approach. In particular consider the special case of the random intercept logistic model for binary data given by

$$logit[Pr(Y_{ij} = 1 \mid b_i)] = \beta_0 + b_{i0} + x_{ij}^{'}\beta$$

and

$$f(y_{ij}) = p_{ij}^{Y_{ij}}(1 - p_{ij})^{1-Y_{ij}}, j = 1, 2, ..., n_i$$

where $p_{ij} = Pr(Y_{ij} = 1 \mid b_{i0}, x_{ij})$ is found by inverting the logit link function so that

$$p_{ij} = \frac{\exp[\beta_0 + b_{i0} + x_{ij}^{'}\beta]}{1 + \exp[\beta_0 + b_{i0} + x_{ij}^{'}\beta]}$$

and

$$1 - p_{ij} = \frac{1}{1 + \exp[\beta_0 + b_{i0} + x_{ij}^{'}\beta]}$$

Here p_{ij} is an individual or cluster specific probability of experiencing the event of interest. It cannot be viewed as an average value.

To simplify the discussion we will write $\gamma_i = \beta_0 + b_{i0}$ and $\eta_i = \gamma_i + x_{ij}^{'}\beta$ and assume that x_{ij} does not include an intercept term. The density function becomes

$$f(y_{ij}) = \Big(\frac{\exp(\eta_i)}{1 + \exp(\eta_i)}\Big)^{Y_{ij}} \Big(\frac{1}{1 + \exp(\eta_i)}\Big)^{1-Y_{ij}}$$

The conditional likelihood contribution from individual i is

$$\begin{aligned} L_i(\beta) &= \prod_{i=1}^{n_i} f(y_{ij}) \\ &= \Big(\frac{\exp(\eta_i)}{1 + \exp(\eta_i)}\Big)^{\sum Y_{ij}} \Big(\frac{1}{1 + \exp(\eta_i)}\Big)^{n_i - \sum Y_{ij}} \\ &= \exp(\eta_i \sum Y_{ij}) \exp(-\log(1 + \exp(\eta_i))^{n_i}) \\ &= \exp[\eta_i \sum Y_{ij} - n_i \log(1 + \exp(\eta_i))] \end{aligned}$$

The joint likelihood function for β and γ_i is proportional to

$$\prod_{i=1}^{N} \exp[\gamma_i \sum_{j=1}^{n_i} Y_{ij} + (\sum_{j=1}^{n_i} Y_{ij} x_{ij}^{'})\beta - \sum_{j=1}^{n_i} \log(1 + \exp(\gamma_i + x_{ij}^{'}\beta))]$$

where $Y_i = \sum_{j=1}^{n_i} Y_{ij}$. This approach is just but an approximation to the likelihood and not an exact result. The approach is similar to the semi-parametric partial likelihood as developed by Cox (1972) in survival analysis.

4.4.2 Maximum likelihood estimation for estimating fixed parameters

Let

$$f_{ij}(y_{ij} \mid \beta, \alpha, \psi) = f_{ij}(y_{ij} \mid b_i, \beta, \phi) f(b_i \mid \alpha)$$

denote the joint probability density for y_{ij} and b_i expressed as the product of the conditional density of y_{ij} given b_i and the marginal density of b_i . The b_i will be treated as a sample of independent observable variables from a probability distribution. The estimates of β can be obtained by the maximization of the marginal likelihood, obtained by integrating over the random effects. The likelihood contribution to the marginal likelihood from the i^{th} subject becomes

$$f_i(y_i \mid \beta, \alpha, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij} \mid b_i, \beta, \phi) f(b_i \mid \alpha) db_i \tag{4.13}$$

from which the overall marginal likelihood from all the N individuals is derived as

$$\begin{aligned} L(\beta, \alpha, \phi) &= \prod_{i=1}^{N} f_i(y_i \mid \beta, \alpha, \phi) \\ &= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij} \mid b_i, \beta, \phi) f(b_i \mid \alpha) db_i \end{aligned} \tag{4.14}$$

This is just the same as in the case of marginal distribution of Y obtained by integrating the joint distribution of Y and b with respect to b. In the special case such of the Gaussian linear model the integral simplifies to a closed form but in the case of non-Gaussian models closed form integrals are rarely possible therefore often high dimensional numerical integration methods are required for its evaluation. The main difficulty in maximizing $Eq.(4.14)$ is due to the presence of N integrals over the $q-$dimensional random effects b_i. However few special case non-Gaussian problems yield closed form integrals similar to the case of linear mixed models with continuous outcomes for example the case of the probit link binomial linear mixed model. Otherwise in general, no analytic expressions are available for integrals in $Eq.(4.14)$ for non-Gaussian outcomes. Numerical approximation methods that can be used include those that are based on the approximation of the integrand, those based on the approximation of the data and those that are based on the approximation of the integral itself. An extensive overview of these approximations are found in Tuerlinckx et al (2004), Pinheiro and Bates (2000), and Skrondal and Rabe-Hesketh (2004). In this chapter we briefly discuss and review each of the methods.

4.4.3 Estimation based on the approximation of the data

This approach is based on the decomposition of the data into the mean and appropriate error terms using the Taylor series expansion. Note that by inverting the link function the mean of the response is a non-linear function of the linear predictor. All methods or versions based on this approach differ in the order of expansion of the Taylor approximation and the point around which the non-linear mean is approximated. First consider the decomposition of Y_{ij} as

$$Y_{ij} = \mu_{ij} + \epsilon_{ij} = h(x_{ij}^{'}\beta + z_{ij}^{'}b_i) + \epsilon_{ij} \tag{4.15}$$

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

$$v(\mu_{ij}) = h^{'}(x_{ij}\beta + z_{ij}b_i) \tag{4.16}$$

where the derivative is with respect to μ_{ij}. For illustration consider binary outcomes with the logit canonical link function. We then have

$$\mu_{ij} = P(Y_{ij} = 1 \mid b_i) = \pi_{ij} = \frac{\exp[x_{ij}^{'}\beta + z_{ij}^{'}b_i]}{1 + \exp[x_{ij}^{'}\beta + z_{ij}^{'}b_i]} = h(x_{ij}^{'}\beta + z_{ij}^{'}b_i) \tag{4.17}$$

Note that from $Eq.(4.15)$ $\epsilon_{ij} = Y_{ij} - \mu_{ij}$. Since $\mu_{ij} = \pi_{ij}$ and $Y_{ij} = 1$ or 0 this implies that $\epsilon_{ij} = 1 - \pi_{ij}$ with probability π_{ij} and $\epsilon_{ij} = -\pi_{ij}$with probability $1 - \pi_{ij}$ hence $E(\epsilon_{ij}) = 0$ and $Var(\epsilon_{ij}) = \pi_{ij}(1 - \pi_{ij})$. Note that π_{ij} is the conditional mean of Y_{ij} given b_i. Estimation then proceeds by using a Taylor linear

approximation to $h(x_{ij}^{'}\beta + z_{ij}^{'}b_i)$ about $\hat{\theta} = (\hat{\beta}, \hat{b_i})^{'}$. We consider two of the several versions based on the approximation of the data.(Molenberghs and Verbeke, 2005, Pg. 269-270).

Penalized Quasi-likelihood (PQL)

We first discuss a linear Taylor expansion of $Eq.(4.15)$ around current estimates $\hat{\beta}$ of the fixed effect and $\hat{b_i}$ of random effect assuming canonical or natural link. This gives

$$\begin{aligned} Y_{ij} &\approx h(x_{ij}^{'}\hat{\beta} + z_{ij}^{'}\hat{b_i}) \\ &\quad + h^{'}(x_{ij}^{'}\hat{\beta} + z_{ij}^{'}\hat{b_i})x_{ij}^{'}(\beta - \hat{\beta}) \\ &\quad + h^{'}(x_{ij}^{'}\hat{\beta} + z_{ij}^{'}\hat{b_i})z_{ij}^{'}(b_i - \hat{b_i}) + \epsilon_{ij} \\ &= \hat{\mu}_{ij} + v(\hat{\mu}_{ij})x_{ij}^{'}(\beta - \hat{\beta}) + v(\hat{\mu}_{ij})z_{ij}^{'}(b_i - \hat{b_i}) + \epsilon_{ij} \end{aligned} \tag{4.18}$$

where $\hat{\mu}_{ij}$ equals its current predictor $h(x_{ij}^{'}\hat{\beta}+z_{ij}^{'}\hat{b}_i)$ of the conditional mean $E[Y_{ij} \mid b_i]$. In vector form this becomes

$$Y_i \approx \hat{\mu}_i + \hat{V}_i X_i(\beta - \hat{\beta}) + \hat{V}_i Z_i(b_i - \hat{b}_i) + \epsilon_i \tag{4.19}$$

for appropriate design matrices X_i and Z_i and with V_i equal to the diagonal matrix with diagonal entries equal to $v(\hat{\mu_{ij}})$. Re-ordering terms in the above equation yields

$$Y_i^* \equiv \widehat{V_i^{-1}}(Y_i - \hat{\mu}_i) + X_i\hat{\beta} + Z_i\hat{b_i} \approx X_i\beta + Z_i b_i + \epsilon_i^* \tag{4.20}$$

for $\epsilon_i^* = \widehat{V_i^{-1}}\epsilon_i$, which still have the mean of zero. The modified response Y_i^* allows us to approximate the problem as a linear mixed model. The approximate linear mixed model in $Eq.(4.20)$ is used to obtain updated estimates for β, D, ϕ using readily available procedures fitting linear mixed models. The resulting estimates are called the penalized quasi-likelihood estimates (PQL) because they are obtained by optimizing a quasi-likelihood function which only involves first and second order conditional moments, augmented with a penalty term on the random effects. We refer to Breslow and Clayton(1993) and Wolfinger and O'Connell (1993) for its implementation and programming in SAS for more detail. (Molenberghs and Verbeke, 2005, Pg. 270).

Marginal Quasi-likelihood (MQL)

This is an alternative approximation method very similar to the PQL method because it is also based on a linear Taylor expansion of the mean μ_{ij} of $Eq.(4.15)$ around the current estimates $\hat{\beta}$ for the fixed effects but around $b_i = 0$ for random effects. This gives a similar expression as that derived under PQL method with $b_i = 0$. The current predictor $\hat{\mu_{ij}}$ is now of the form $h(x_{ij}^{'}\hat{\beta})$ instead of $h(x_{ij}^{'}\hat{\beta} + z_{ij}^{'}\hat{b}_i)$ as was the case under the PQL method. Re-ordering the terms gives $Y_i^* \equiv \widehat{V_i^{-1}}(Y_i - \hat{\mu}_i) + X_i\hat{\beta}$ which also satisfies the approximate linear mixed model.

$$Y_i^* \approx X_i\hat{\beta} + \widehat{V_i^{-1}}(Y_i - \hat{\mu}_i) \approx X_i\beta + Z_i b_i + \epsilon_i^* \tag{4.21}$$

Similar to $Eq.(4.20)$. The resulting estimates are called marginal quasi-likelihood (MQL). They are obtained by optimizing a quasi-likelihood function which only involve first and second order moments but now evaluated at the marginal linear predictor $x_i^{'}\beta$ rather than the conditional linear predictor $x_i^{'}\beta + z_i^{'}b_i$. For more information refer to Breslow and Clayton(1993) and Goldstein (1991). (Molenberghs and Verbeke, 2005, Pg. 270-271).

Comparison of (PQL) and (MQL) methods

The penalized quasi-likelihood (PQL) approach is the most common estimation procedure for the generalized linear mixed model (GLMM). The essential difference between PQL and MQL is that the MQL does not incorporate the random effects in the linearization process but both methods have the same key idea and will ideally have similar properties. The MQL estimation performs well if the random effects variance is very small. The MQL approach has also commonly been used for inference in the GLMMs. Both approaches require the computations of the joint moments of the clustered observations, up to order four. But the derivations of these moments are not easy. MQL completely ignores the random effect variability in the linearization of the mean. Both MQL and PQL perform poorly for binary outcomes with few repeated measurements per cluster while with an increasing number of measurements per subject MQL remains biased and PQL becomes consistent. Within the PQL and MQL methods, the linear mixed model approximation can be based on maximum likelihood estimation (ML) or Restricted Maximum likelihood estimation (REML) resulting in slightly different results. Rodriguez and Goldman (1995) show that both PQL and MQL may be seriously biased when applied to binary response data. Their simulations reveal that the fixed effects and variance components suffer from substantial, if not severe, attenuated bias in certain situations. Breslow and Lin (1995) and Lin and Breslow (1996) suggest the inclusion of bias correction terms while Kuk (1995) suggested the use of iterative bootstrap. Goldstein and Rasbash (1996) show that one of the ways to improve the accuracy of the approximations is to include a second order term in the Taylor series expansion. They call these methods PQL2 and MQL2. They state that MQL2 performs slightly better than MQL but PQL2 is substantially better than PQL.

4.4.4 Estimation based on the approximation of the integrand

When integrands are approximated, the closed form integral expression is obtain so that numerical maximization of approximated likelihood is feasible. All methods that have been proposed lead to Laplace-type approximations methods of the function to be integrated. Tierny and Kadane (1986) use the laplace method that is designed to approximate integrals of the form.

$$I = \int e^{Q(b)} db \tag{4.22}$$

where $Q(b)$ is a known, unimodal and bounded function of a $q-$dimensional variable b. Let $\hat{b}$ be the value of b that maximizes Q. The second-order Taylor expansion of $Q(b)$ is of the form.

$$Q(b) \approx Q(\hat{b}) + \frac{1}{2}(b-\hat{b})' Q^{''}(\hat{b})(b-\hat{b}) \tag{4.23}$$

where $Q^{''}(\hat{b})$ equal to the approximated Hessian of Q, that is the matrix of second-order derivative of Q, evaluated at $\hat{b}$. When we replace $Q(b)$ in $Eq.(4.22)$ by its approximation in $Eq.(4.23)$ we get

$$I \approx (2\pi)^{\frac{q}{2}} \mid -Q^{''}(\hat{b}) \mid^{\frac{-1}{2}} e^{Q(\hat{b})}$$

Note that the integral in $Eq.(4.14)$ is proportional to an integral of the form given by $Eq.(4.22)$ for functions of $Q(b)$ given by

$$Q(b) = \phi^{-1} \sum_{j=1}^{n_i} [y_{ij}(x_{ij}^{'}\beta + z_{ij}^{'}b) - \psi(x_{ij}^{'}\beta + z_{ij}^{'}b)] - \frac{1}{2} b^{'} D^{-1} b$$

such that the Laplace method can be readily applied. It is important to note the mode $Q(\hat{b})$ depends on the unknown parameter β, ϕ, D such that in each iteration of the numerical maximization of the likelihood, $\hat{b}$ will be re-calculated conditionally on the current values for the estimates of parameters. The approximation will be exact when $Q(b)$ is a quadratic function of b. Raudenbush et al (2000) extended the above Laplace method by including higher-order terms in the Taylor expansion of $Eq.(4.23)$ for Q, up to order six.(Molenberghs and Verbeke, 2005, Pg. 268-269).

4.4.5 Estimation based on the approximation of the integral

As an alternative to the linearization methods, approximations to the integral or numerical integration are possible. Several of these have been implemented in various software tools for generalized linear mixed models. Pinheiro and Bates (1995, 2000) suggest the use of adaptive quadrature rules for the random effects models where the numerical integration is centered around the empirical bayes (EB) estimates of the random effects and the number of the quadrature points is then selected in terms of desired accuracy. We consider Gaussian and adaptive Gaussian quadrature designed for the approximation of the integral of the form:

$$\int f(z)\phi(z)dz \tag{4.24}$$

for a known function $f(z)$ and for $\phi(z)$ the density of a univariate or multivariate standard normal distribution. The first step is to standardize the random effects such that they get the identity covariance matrix. Precisely let δ_i be equal to $\delta_i = D^{-\frac{1}{2}}b_i$ where b_i is the vector of random effects of the model. This transformation implies that δ_i is normally distributed with mean of 0 and covariance I. The linear predictor becomes $\theta_{ij} = x_{ij}^{'}\beta + z_{ij}^{'}D^{\frac{1}{2}}\delta_i$. Hence the variance components in D are now in the linear predictor. The likelihood contribution for subject i is now given by:

$$\begin{aligned} f_i(y_i \mid \beta,\alpha,\phi) &= \int \prod_{j=1}^{n_i} f_{ij}(y_{ij} \mid b_i,\beta,\phi)f(b_i,\alpha)db_i \\ &= \int \prod_{j=1}^{n_i} f_{ij}(y_{ij} \mid \delta_i,\beta,\alpha,\phi)f(\delta_i)d\delta_i \end{aligned} \tag{4.25}$$

Note that $Eq.(4.25)$ is of the form $Eq.(4.24)$ which means that Gaussian or adaptive Gaussian quadrature approximation can be applied to estimate parameters for a generalized linear mixed model.(Molenberghs and Verbeke, 2005, Pg. 273).
The difference between Gaussian and Adaptive Gaussian quadrature methods are shown in the sketch shown in Figure 4.1 below (Molenberghs and Verbeke, 2005, Pg. 274).

The black triangles indicate the position of the quadrature points, and the rectangles indicate the contribution of each point to the integral.

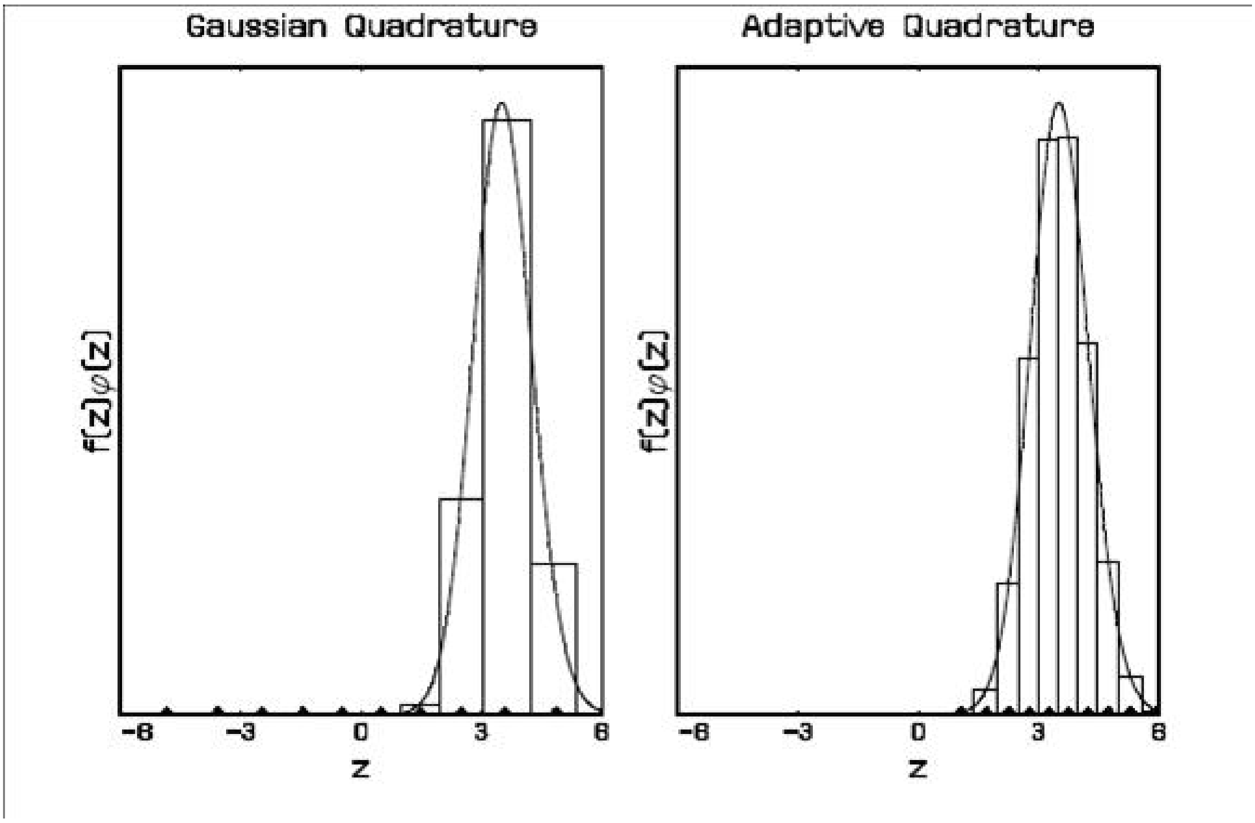

Figure 4.1: Graphical illustration of Gaussian and adaptive Gaussian quadrature of order Q=10

Gaussian Quadrature

In the Gaussian quadrature the integral of $\int f(z)\phi(z)dz$ is approximated by the weighted sum

$$\int f(z)\phi(z)dz \approx \sum_{q=1}^{Q} w_q f(z_q)$$

where Q is the order of approximation. The higher Q is the more accurate the approximation will become. The so called node or quadrature points z_q are solutions to the Qth order Hermite polynomial, while the w_q are called the weights. The nodes z_q and weights w_q are found in the tabulated form. However Press et al (1992) gives an algorithm to calculate these weights less restrictively. The main disadvantage of Gaussian quadrature indicated in literature in the case of univariate integration is that the quadrature points z_q are chosen based on $\phi(z)$, independent of function $f(z)$ in the integrand (Pinheiro and Bates, 1995, 2000). Depending on the support of $f(z)$, the z_q will or will not lie in the region of interest.

Adaptive Gaussian Quadrature

In the Adaptive Gaussian Quadrature, the quadrature points are centered and scaled as if $f(z)\phi(z)$ were a normal distribution with the mean of this distribution being the mode z of $\ln[f(z)\phi(z)]$, and variance is equal to.

$$\Big[-\frac{d^2}{dz^2}\ln[f(z)\phi(z)]\mid_{z=\hat{z}}\Big]^{-1}$$

Here the new adaptive quadrature points are given by

$$z_q^+ = \hat{z} + \Big[-\frac{d^2}{dz^2}\ln[f(z)\phi(z)]\mid_{z=\hat{z}}\Big]^{-\frac{1}{2}} z_q$$

with corresponding weights given by

$$w_q^+ = \Big[-\frac{d^2}{dz^2}\ln[f(z)\phi(z)]\mid_{z=\hat{z}}\Big]^{-\frac{1}{2}} \frac{\phi(z_q^+)}{\phi(z_q)} w_q$$

The integral is then approximated by

$$\int f(z)\phi(z) \approx \sum_{q=1}^{Q} w_q^+ f(z_q^+).$$

In this case the quadrature points are chosen adaptively taking account of the function $f(z)$. Note that when fitting generalized linear mixed models, an approximation is applied to the likelihood contribution of each of the N subject or units in the dataset.

In general the higher the order of Q, the better the approximation will be of the N integrals in the likelihood. The adaptive Gaussian quadrature requires (much) less quadrature points than classical Gaussian quadrature. The adaptive Gaussian quadrature requires calculation of $\hat{z}$ for each unit in the dataset. The function in $Eq.(4.24)$ as well as the quadrature points and weights depends on the unknown parameters β, α, ϕ and hence needs to be updated in every step of the iterative estimation procedure. Note that α is a vector valued parameter containing the unknown variance and covariance parameters in D. (Molenberghs and Verbeke, 2005, Pg. 275-276).

4.5 Software for fitting generalized linear mixed models

The models can be fitted using a number of available statistical software such as SAS, GenStat, Stata and a few more. In the current work we focus on SAS applications. SAS PROC GLIMMIX is capable of fitting statistical models to data with both random and fixed effects and where the response is not necessarily normally distributed. As already stated such models are called generalized linear mixed models (GLMM) in contrasts to linear mixed models (LMM) which strictly assume normal (Gaussian) outcomes where SAS PROC MIXED can readily be used. PROC GLIMMIX performs both estimations and statistical inference for generalized linear mixed models. The model fitting by the GLIMMIX procedure in particular extends the GLM model for longitudinal data by incorporating the cluster or subject to subject variability through the use of random effects. This can be accomplished by including random effects in the linear predictor which in effect means modelling the correlation among the data directly. The GLIMMIX procedure can also fit models for non-normal data with hierarchical random effects, provided that the random effects have a normal distribution. GLIMMIX procedure fits generalized linear mixed models based on linearization. The default estimation method in PROC GLIMMIX for models containing random effects is a technique known as restricted Pseudo-likelihood (RPL) estimation (Wolfinger and O'Connell, 1993). To fit GLMMs via Gaussian and adaptive Gaussian quadrature methods in SAS, PROC NLMIXED is used. SAS PROC NLMIXED allows the user to specify or build the likelihood for the data directly. PROC NLMIXED selects the number of Quadrature points adaptively by evaluating the log likelihood function at the starting values of parameters until two successive evaluations have relative difference less than the value of the $QTOL = option$. The noad option in PROC NLMIXED statement request non-adaptive Gaussian quadrature. PROC NLMIXED computes derivatives of the adaptive Gaussian quadrature approximation and the default method used is the dual quasi-Newton optimization. The main advantage of NLMIXED is that the user is given a high degree of flexibility in the way the model is specified and parameterized. This therefore makes it more appealing to advanced users than other procedures. In the current application both PROC GLIMMIX and PROC NLMIXED are used.

4.6 Application of GLMM to the respiratory infection data

To understand how to fit GLMMs and compare between models three types of GLMM namely the random intercept model, random slope model and random intercept and slope model are fitted. In the case of the random intercept model the estimated variance component and residual variance estimate are shown in Table 4.1. The result for fitting the random intercept model using both the PQL and MQL methods indicates (choosing the unstructured covariance matrix) the two variance components namely the random intercept and the measurement error variances are estimated as 3.6627 and 0.5620 with standard errors given by 0.7648 and 0.0438 respectively in the PQL while under the MQL approximation the estimates are 1.9587 and 0.6593 with standard errors given by 0.4043 and 0.0511 respectively. We note that the random intercept variance component is underestimated under MQL approximation than under the PQL approximation. The parameter estimates measuring the covariate effects are in Table 4.2. Table 4.3 gives the type III analysis for fixed effects. The fixed effects parameter estimates and the type III analysis for the random intercept model for both MQL and PQL methods result indicate that only the treatment and baseline status were significant and other effects were not significant at 5% significance level. The result also indicate that the interaction between treatment and time was not significant at 5% significance level. Similar analysis was done for the random slope model and results showing estimation of variance components are tabulated in Table 4.4. The two variance components namely the variance of the random slope and the measurement error variance are estimated as 0.5251 and 0.5825 with standard errors given by 0.1173 and 0.0461 respectively under the PQL approximation. The MQL estimates are 0.2333 and 0.7057 with corresponding standard errors given by 0.0515 and 0.0549. Again we note that the MQL underestimate the variance component for the random slope compared to the PQL method. The measurement error variance estimate under MQL approach is larger compared to the PQL value. The results for the analysis of fixed effects using the unstructured covariance structure under the random slope model are shown in Table 4.5. Table 4.6 shows a summary analysis based on a type III analysis in SAS. The results for fixed effects parameter estimates and the type III analysis for fitting the random slope model in both MQL and PQL methods indicates that only the treatment and baseline status were significant and other effects were not significant at 5% significance level for unstructured covariance structure. The result for also indicate that the interaction between treatment and time was not significant at 5% significance level. An attempt was made to model both a random intercept and slope to the data. This model was fitted without including the interaction between time and treatment because both PQL and MQL methods were not converged when this interaction was included in the model. Convergence was only possible under the UN covariance structure for the variance components and the MQL approximation when the interaction between time and treatment was not included in the model. The result are shown in Table 4.7. We note that while separately the random intercept (Table 4.1) and slope Table 4.4 variance components were estimated with large values compared to their standard errors the estimate of the variance component for the slope in Table 4.7 is almost the same size as its standard error. The results in Table 4.7 also estimate a negative correlation between the random intercept and slope effects. Nonetheless the estimated magnitude of the covariance is much smaller compared to its standard error. We also note that much of the individual to individual variability is captured more under the random intercept than slope. The results for the analysis of fixed effects for the random slope and intercept model are shown in Table 4.8 and Table 4.9. The result indicate that when we fit both the random slope and intercept in the same model we find that the MQL method does converge better than the PQL which does experience non-convergence under a number of covariance structures. From the unstructured covariance structure specification we note that the elements of the matrix $D = Var(b_i)$ is estimated as

$$\hat{\mathbf{D}} = \begin{bmatrix} 1.9397 & -0.1148 \\ -0.1148 & 0.1044 \end{bmatrix}$$

Which indicates a negative correlation of -0.255 between the random intercept and slope.

Table 4.1: Variance component estimates for the random intercept model

	PQL			MQL	
Covariance Structure		Estimate	Standard error	Estimate	Standard error
Unstructured	UN(1,1)	3.6627	0.7648	1.9587	0.4043
	Residual(VC)	0.5620	0.04383	0.6593	0.0511

The results for the fixed effects analysis in Tables 4.8 and 4.9 allowing both a random intercept and slope also indicate that only the treatment and baseline status were significant and other effects were not significant at 5% significance level. The summary results in Table 4.10 shows that the individuals in centre 1 were 1.840 as likely to be infected with respiratory disease compared to individuals in centre 2 (in terms of odds of infection) but the centre effects was not significant.

Table 4.2: Parameter estimates, standard errors and p-values for fixed effects under the random intercept model

	PQL			MQL		
parameter	Est	SE	P-value	Est	SE	P-value
intercept	-1.9128	1.0612	0.0743	-1.3743	0.7791	0.0806
Centre	0.8135	0.4779	0.0917	0.6097	0.3498	0.0842
treatment	1.8552	0.6502	0.0046	1.3760	0.5221	0.0088
Sex	0.2416	0.5985	0.6867	0.1394	0.4339	0.7482
Age	0.0233	0.0178	0.1904	0.0165	0.0128	0.1986
Baseline	-2.5302	0.4674	<.0001	-1.8280	0.3415	<.0001
time	0.1694	0.1354	0.2117	0.1089	0.1155	0.3464
time*treatment	-0.0573	0.1843	0.7561	-0.0372	0.1605	0.8170

Table 4.3: Type III effects for the random intercept model

	PQL		MQL	
Effect	F-value	P-vale	F-value	P-value
Centre	2.90	0.0917	3.04	0.0842
treatment	8.14	0.0046	6.94	0.0088
Sex	0.16	0.6867	0.10	0.7482
Age	1.72	0.1904	1.66	0.1986
Baseline	29.31	<.0001	28.66	<.0001
time	2.34	0.1274	1.27	0.2608
time*treatment	0.10	0.7561	0.05	0.8170

Table 4.4: Variance component estimates for the random slope model

		PQL		MQL	
Covariance Structure		Estimate	Standard error	Estimate	Standard error
Unstructured	UN(1,1)	0.5251	0.1173	0.2333	0.0515
	Residual(VC)	0.5825	0.04611	0.7057	0.0549

Table 4.5: Parameter estimate, standard errors and p-values for fixed effects under the random slope model

	PQL			MQL		
parameter	Est	SE	P-value	Est	SE	P-value
intercept	-0.4779	0.7825	0.5418	-0.9169	0.6624	0.1673
Centre	0.4635	0.3571	0.1952	0.5164	0.2977	0.0837
treatment	1.3102	0.4986	0.0090	1.3245	0.4692	0.0051
Sex	-0.0571	0.4459	0.8982	0.0253	0.3684	0.9454
Age	0.0025	0.0135	0.8543	0.0093	0.0109	0.3981
Baseline	-2.3666	0.3609	<.0001	-1.9568	0.2971	<.0001
time	0.0337	0.1792	0.8511	0.1074	0.1380	0.4380
time*treatment	0.1252	0.2484	0.6145	-0.0274	0.1916	0.8865

Table 4.6: Type III effects for the random slope model

	PQL		MQL	
Effect	F-value	P-vale	F-value	P-value
Centre	1.68	0.1952	3.01	0.0837
treatment	6.90	0.0090	7.97	0.0051
Sex	0.02	0.8982	0.00	0.9454
Age	0.03	0.8543	0.72	0.3981
Baseline	43.00	<.0001	43.38	<.0001
time	0.60	0.4387	0.96	0.3289
time*treatment	0.25	0.6145	0.02	0.8865

The individuals who received placebo treatment were 3.959 as likely to be infected with the respiratory disease compared to individuals who received the active treatment. Male individuals were 1.150 as likely to be infected with respiratory disease than female individuals but sex was not a significant effect.

The odds of infection for an individual who is not infected at baseline is 0.161 times that of an individual who is infected at baseline. Individual to individual variability was also investigated

Table 4.7: Variance components estimates for the random intercept and slope model using MQL approach

Covariance Structure		Estimate	Standard error
Unstructured	UN(1,1)	1.9397	1.0847
	UN(2,1)	-0.1148	0.3322
	UN(2,2)	0.1044	0.1262
	Residual(VC)	0.6295	0.0595

Table 4.8: Parameter estimates, standard errors and p-values for the fixed effects under the random intercept and slope model using the MQL method

parameter	Est	SE	P-value
intercept	-1.1697	0.7474	0.1206
Centre	0.5729	0.3496	0.1042
treatment	1.2884	0.3331	0.0001
Sex	0.0985	0.4337	0.8206
Age	0.0137	0.0128	0.2858
Baseline	-1.8731	0.3419	<.0001
time	0.0917	0.0846	0.2807

Table 4.9: Type III effects for random intercept and slope model using MQL

Effect	F-value	P-vale
Centre	2.69	0.1042
treatment	14.96	0.0001
Sex	0.05	0.8206
Age	1.15	0.2858
Baseline	30.02	<.0001
time	1.17	0.2807

using the random intercept and slope. The variance components associated with these effect were estimated as 1.9397 and 0.1044 (Table 4.7). The corresponding standard errors were 1.0847

Table 4.10: Odds ratio estimates for the random intercept model

parameters	estimate	odds ratio
Centre	0.6097	1.840
treatment	1.3760	3.959
Sex	0.1394	1.150
Age	0.0165	1.017
Baseline	-1.8280	0.161
time	0.1089	1.115
time*treatment	-0.0372	0.963

and 0.1262 indicating that the inherent individual heterogeneity was more at baseline than over time. In summary the application of PQL and MQL methods were demonstrated using binary longitudinal outcome data and the application was implemented using the SAS software. The methods worked well when only one random effect (intercept or slope) was assumed in the model. When both the intercept and slope were specified the MQL was able to fit both effects but the PQL does not achieve convergence easily. All the models that converged show that only the baseline outcome and treatment effects were significant at 5% significance level.

4.6.1 Direct likelihood estimation via NLMIXED

The result for fitting the model using proc NLMIXED Gaussian quadrature and adaptive Gaussian quadrature assuming a random intercept model are given in Tables 4.11 and 4.12 respectively. In this particular case the result indicates not much difference in the parameter estimates of the adaptive Gaussian quadrature and Gaussian quadrature methods. It is important to stress that each log-likelihood corresponds to the maximum of the approximation to the model likelihood implying that log-likelihoods corresponding to different estimation procedure or different number of quadrature points are not necessarily comparable.

This means that difference in log-likelihood values reflect the difference in the quality of numerical approximation and thus higher log-likelihood values do not necessarily correspond to better

Table 4.11: Solution for the fixed effect model using the Gaussian quadrature method

effect	Q=5	Q=10	Q=20	Q=50
Beta0(intercept)	-0.5464 (1.5477)	-1.5367 (1.4828)	-1.4310 (1.4650)	-1.4310 (1.4582)
Beta1(Centre)	1.3012 (0.5172)	0.8282 (0.5568)	0.9386 (0.5509)	0.9387 (0.5491)
Beta2(treatment)	1.8685 (0.4934)	2.0904 (0.5713)	2.0105 (0.544)	2.0149 (0.5443)
Beta3(Sex)	0.0604 (0.6629)	0.3154 (0.6902)	0.2844 (0.6779)	0.2819 (0.6790)
Beta4(age)	0.0371 (0.020)	0.0253 (0.0207)	0.0273 (0.0205)	0.0274 (0.0203)
Beta5(Baseline)	-3.4251 (0.6947)	-2.9060 (0.6008)	-2.9242 (0.5897)	-2.9237 (0.5866)
Beta6(time)	-0.1440 (0.1258)	-0.1412 (0.1246)	-0.1412 (0.1246)	-0.1413 (0.1246)
tau	4.4273 (1.5587)	3.9775 (1.3371)	4.0017 (1.2983)	4.0314 (1.3303)
-2l	432.6	433.4	433.3	433.3

approximation. The standard errors are approximately the same as those obtained using GLIMMIX procedure in the random intercept model in Table 4.2. The adaptive Gaussian quadrature method gives estimates much closer to the quasi-likelihood approximation under GLIMMIX than the pure Gaussian quadrature approximation.

Table 4.12: Solution for the fixed effect model using the adaptive Gaussian quadrature method

effect	Q=5	Q=10	Q=20	Q=50
Beta0(intercept)	-1.4174 (1.4323)	-1.4301 (1.4563)	-1.4309 (1.4581)	-1.4310 (1.4582)
Beta1(Centre)	0.9291(0.5385)	0.9381 (0.5483)	0.9387 (0.5491)	0.9387 (0.5491)
Beta2(treatment)	1.9957 (0.5331)	2.0135 (0.5434)	2.0148 (0.5443)	2.0148 (0.5443)
Beta3(Sex)	0.2790(0.6663)	0.2817 (0.6781)	0.2819 (0.6790)	0.2819 (0.6790)
Beta4(Age)	0.0271 (0.020)	0.0274 (0.0203)	0.0274 (0.0203)	0.0274 (0.0203)
Beta5(Baseline)	-2.8904 (0.5715)	-2.9212 (0.5851)	-2.9237 (0.5865)	-2.9237 (0.5866)
Beta6(time)	-0.1405 (0.1242)	-0.1412 (0.1246)	-0.1413 (0.1246)	-0.1413 (0.1246)
tau	3.8309 (1.2096)	4.0168 (1.3168)	4.0313 (1.3303)	4.0314 (1.3304)
-2l	433.7	433.3	433.3	433.3

The quadrature methods using NLMIXED present a degree of flexibility because the user is able to specify the likelihood directly. The standard errors under adaptive Gaussian quadrature increase as the number of quadrature point increase.

Chapter 5

Hierarchical generalized linear model

When modelling random effects under the linear mixed model (LMM) (Laid and Ware, 1982) and the generalized linear mixed model (GLMM) (Breslow and clayton, 1993) methodologies the random effects are commonly assumed to be normal which in a way is a very restrictive condition. Under the Hierarchical generalized linear models (HGLMs) the random effects are allowed to come from an arbitrary distribution, in particular the distribution conjugate to that of the response Y. These models were introduced by Lee and Nelder (1996, 1998). In the HGLM the response variable in these models is distributed according to a one parameter exponential family such as binomial, Poisson, exponential and the Gaussian distributions among others. The HGLM as an extension of LMM and GLMM provides a unified modeling frame work for the estimation of cluster-specific quantities of interest, covariate effects and components of variance. This likelihood based approach makes it possible to pool information across clusters to derive more precise estimates of case specific and cluster specific parameters. The application of HGLMs also makes it possible to account for correlations in the data and to derive standard errors of estimates which are more realistic than those obtained by methods ignoring such a correlation. Hierarchical generalized linear models (HGLMs) are similar to GLMMs apart from two distinctions:
(a.) The random effects can have any distribution conjugate to that of the response, whereas current GLMMs nearly always have normal random effects.
(b.) They are not as computationally intensive, instead of integrating out the random effects they are based on a modified form of likelihood known as the hierarchical likelihood or h-likelihood

(Lee and Nelder, 1996).

To introduce the formulation of such a model let Y be the response and u be the (unobserved) random component. The following distribution assumption are stated .

(a). The conditional (log) likelihood for y given u has the GLM form

$$l(\theta^{'}, \phi, y \mid u) = \frac{[y\theta^{'} - b(\theta^{'})]}{a(\phi)} + c(y, \phi)$$

where $\theta^{'}$ denotes the canonical parameter and ϕ is the dispersion parameter. We write $\mu^{'}$ for the conditional mean of y given u, where $\eta^{'} = g(\mu^{'})$ and $g(.)$ is the link function for the GLM describing the conditional distribution of y given u. Thus we use the notation $\theta^{'}$ to reflect this conditionality. The linear predictor $\eta^{'}$ takes the form

$$\eta^{'} = \eta + v \tag{5.1}$$

where $\eta = X\beta$ as for the GLM and $v = v(u)$ for some strictly monotonic function of u.

(b) The distribution of u is assumed appropriately.

The modelling of $\eta^{'}$ in $Eq.(5.1)$ involves not only fixed effects modelling for η but also dispersion modelling for v, which describes the overdispersion.

5.1 Hierarchical Likelihood

The log of the h-likelihood is given by

$$h = l(\theta^{'}, \phi, y \mid v) + l(\alpha, v). \tag{5.2}$$

where $l(\alpha, v)$ is the logarithm of the density function for v with parameter α, and $l(\theta^{'}, \phi, y \mid v)$ is the logarithm of the density function for $y \mid v$. The random component v is the scale on which the random effect u occurs linearly in the linear predictor. The h-likelihood can be derived from the density functions of u and $y \mid v$. $l(\alpha, v)$ can be derived from the density function of u with differential element $dv(u)$ and $l(\theta^{'}, \phi, y \mid v(u)) = l(\theta^{'}, \phi, y \mid v)$ can be derived from the logarithm of the density function for $y \mid u$, since v is the strictly monotonic function of u. The log h-likelihood is the logarithm of the joint density function for v and y. We abbreviate the estimates derived from maximizing the h-likelihood or the maximum h-likelihood estimates as MHLEs and are obtained by jointly solving

$$\frac{dh}{d\beta} = 0.$$

and

$$\frac{dh}{dv} = 0.$$

The MHLEs for random effects are invariant with respect to the transformation of random components u. For example estimating equations $\frac{dh}{dv} = 0$ and $\frac{dh}{du} = 0$ result in the same random effect estimate. In the current work we focus more on HGLMs which assume the random effects follow a conjugate distribution to that of Y. To explain the meaning of a conjugate distribution, suppose conditionally $y \mid u$ is Poisson with mean

$$\mu^c = E(y \mid u) = \exp(x^{'}\beta)u$$

Using the link we have

$$\eta^c = \log(\mu^c) = x^{'}\beta + v$$

where $v = \log(u)$. If u is gamma distributed v is log-normally distributed and the model becomes known as the Poisson-Gamma HGLM. If v is normally distributed then u is log-gamma. This is the Poisson generalized linear model which is now called the Poisson-log-normal HGLM. According to Lee, Nelder and Pawitan (2006) a gamma distribution for u is a conjugate for $y \mid u$ and the resulting model is called a conjugate HGLM. Both Poisson-gamma and Poisson GLMM models belong to the class of HGLM; where the former is a conjugate HGLM while the latter is not.

5.2 Properties of Maximum hierarchical likelihood estimate

Consider the hierarchical model

$$y \mid v \sim f_1(y \mid v, \beta, \phi)$$

$$v \sim f_2(\alpha, \phi)$$

where f_1 and f_2 are arbitrary density function of $y \mid v$ and v respectively. Assume that ϕ and α are given and β are parameters of interest. The log-likelihood of the h-likelihood $h(\beta, \phi, \alpha, y, v)$ has components $l(\beta, \phi, y \mid v) = \log f_1(y \mid v, \beta, \phi)$ and $l(\alpha, v) = \log f_2(v, \alpha)$. It can also be written in the form

$$h = l(\beta, \phi, y \mid v) + l(\alpha, v) = L + l(\beta, \phi, \alpha, v \mid y).$$

where L is the marginal likelihood of y and

$$l(\beta, \phi, \alpha, v \mid y) = \log[f_1(y \mid v, \beta, \phi) f_2(v, \alpha) / \int f_1(y \mid v, \beta, \phi) f_2(v, \alpha) dv].$$

is the logarithm of the density function of $v \mid y$.

5.3 Conjugate Hierarchical generalized linear models

Let the response be y_{ij} for $i = 1, 2, ..., N$ and $j = 1, 2, ..., n_i$ with $n = \Sigma n_i$ and u_i be the unobserved random components from unit i. Assuming $\mu_{ij} = E(Y_{ij})$ and $\mu'_{ij} = E(Y_{ij} \mid u_i)$ we consider the canonical link model such that
$\theta'_{ij} = \theta_{ij} + v_i$, where $\theta'_{ij} = \theta(\mu'_{ij})$, $\theta_{ij} = \theta(\mu_{ij})$ and $v_i = \theta(u_i)$. Then given $\theta(\mu_{ij}) = x_{ij}^T \beta$ were $\beta = [\beta_0, ..., \beta_k]^T$ we have

$$\frac{dh}{d\beta_k} = \sum_{ij} \frac{(y_{ij} - \mu'_{ij}) x_{kij}}{\phi} \tag{5.3}$$

Assume that the kernel of $l(\alpha, v)$ has the form

$$\sum_i (a_1(\alpha) v_i - a_2(\alpha) b(v_i)) \tag{5.4}$$

where $a_1(.)$ and $a_2(.)$ are some function of dispersion parameter α. Then the kernel of the log of the h-likelihood becomes

$$\sum_{ij} \frac{\theta^{'} y - b(\theta^{'})}{\phi} + \sum_{i} (a_1(\alpha) v - a_2(\alpha) b(v)). \tag{5.5}$$

Since $\frac{dh(\theta(\mu))}{d\theta} = \mu$ so that $\frac{db(v)}{dv} = u$, we have

$$\frac{dh}{dv_i} = \frac{\sum_j (y_{ij} - \mu^{'}_{ij}) + \phi a_1(\alpha)}{\phi} - a_2(\alpha). \tag{5.6}$$

Thus equating $\frac{dh}{dv_i}$ to 0 gives an estimate of the random effect as

$$\hat{u_i} = \frac{y_{i+} - \mu^{'}_{i+} + \phi a_1(\alpha)}{\phi a_2(\alpha)}. \tag{5.7}$$

where $y_{i+} = \sum_j y_{ij}$ and $\mu^{'}_{i+} = \sum_j \mu^{'}_{ij}$. This shows that in the conjugate HGLMs the MHLE for the random effects has a simple form on the u-scale.

5.4 Binomial-beta conjugate model

Suppose that the conditional distribution of y given u is the binomial distribution with $\mu^{'} = m\pi^{'}$, then the conjugate HGLM leads to

$$\theta^{'}_{ij} = \theta_{ij} + v_i$$

where $\theta^{'}_{ij} = \log[\frac{\pi^{'}_{ij}}{1-\pi^{'}_{ij}}]$, $v_i = \log[\frac{u_i}{1-u_i}]$ and $\theta_{ij} = \log[\frac{\pi_{ij}}{1-\pi_{ij}}] = X\beta$ where X and β are the design matrix and a vector of fixed regression parameter respectively. Now u is assumed to have the conjugate beta distribution with parameters α_1 and α_2, so that $E(u) = \frac{\alpha_1}{\alpha_1+\alpha_2}$. We have

$$l(\alpha, v) = \sum [\alpha_1 v_i - (\alpha_1 + \alpha_2) \log(\frac{1}{1-u_i}) - \log B(\alpha_1, \alpha_2)]. \tag{5.8}$$

Here $a_1(\alpha) = \alpha_1$ and $a_2(\alpha) = \alpha_1 + \alpha_2$. By $Eq.(5.3)$ the MHL equation for β becomes

$$\frac{dh}{d\beta_k} = \sum_{ij} (y_{ij} - m_{ij}\pi^{'}_{ij}) x_{kij} = 0. \tag{5.9}$$

and by $Eq.(5.7)$ the MHLE for u becomes

$$\hat{u}_i = \frac{(y_{i+} - \mu'_{i+} + \alpha_1)}{(\alpha_1 + \alpha_2)}. \tag{5.10}$$

When $\frac{\alpha_1}{\alpha_2} \rightarrow 1$ and $\alpha_1 \rightarrow \infty$, $\hat{u}_i$ converges to $\frac{1}{2}$ and π' converges to π so that $Eq.(5.9)$ gives the ordinary logit regression equations. When both α_1 and α_2 goes to 0, it implies that $y_{i+} = \mu'_{i+} = \sum_j m_{ij}\pi'_{ij}$ for all i so that $Eq.(5.10)$ gives a type of intrablock estimating equations. Since $\pi' = \frac{\pi u}{[\pi u + (1-\pi)(1-u)]}$, $E(y) = E(m\pi') \neq \mu = m\pi$, so the inference on the marginal mean may not be easy. However within group or cluster comparisons are unaffected because

$$\frac{\pi'_{ij}/(1-\pi'_{ij})}{\pi'_{ik}/(1-\pi'_{ik})} = \frac{\pi_{ij}/(1-\pi_{ij})}{\pi_{ik}/(1-\pi_{ik})}.$$

In general an explicit form of marginal likelihood for binomial-beta models is not available.

5.5 Application of HGLM to the respiratory infection data using GENSTAT

There are very few statistical software that readily can implement HGLMs. The statistical package that we know has made advances in the implementation of HGLMs is the GENSTAT software. We demonstrate how it works with the example of data on respiratory infection. The aim of this section is to apply and discuss the result from hierarchical generalized linear model (HGLM) using GENSTAT. In the analysis the reference levels for categorical variables were centre 1, placebo, male category, non infected status and time one for the categorical variables centre, treatment, sex, baseline status of infection and time. The parameter estimates results from the mean model and the odds ratios are shown in Tables 5.1 and 5.2 respectively. The parameter estimates results from the dispersion model are shown in Table 5.3. The results in Table 5.1 indicate that the covariate effects treatment and baseline disease status are significant and other covariate effects are not significant at 5% significance level. Then Table 5.2 shows that the odds of infection for an individual who receives active treatment is 0.2127 times that of an individual who receives placebo treatment. The odds of infection for an individual who is infected at baseline is 9.089

times that of an individual who is not infected at baseline. The dispersion parameters on the log scale are given in Table 5.3. The estimates on the original scale are $\hat{\phi} = \exp(-0.2277) = 0.7964$ and $\hat{\lambda} = 0.3376$.

Table 5.1: Parameter estimates and standard errors for the mean HGLM model

parameter	Estimate	Standard Error	t-value
intercept	-1.300	0.575	-2.26
Centre	-0.659	0.380	-1.74
treatment	-1.548	0.365	-4.24
Sex	-0.262	0.474	-0.55
Age	0.0217	0.0141	1.54
Baseline	2.207	0.373	5.91
time2	0.289	0.317	0.91
time3	-0.015	0.319	-0.05
time4	0.469	0.317	1.48

Table 5.2: The odds ratio estimates for HGLM

parameters	estimate	odds ratio
Centre	-0.659	0.5174
treatment	-1.548	0.2127
Sex	-0.262	0.7693
Age	0.0217	1.022
Baseline	2.207	9.089
time2	0.289	1.335
time3	-0.015	0.9855
time4	0.469	1.598

Table 5.3: Parameter estimates on the log scale and standard errors from the dispersion model for individuals

parameter	Estimate	Standard Error	t-value
phi(ϕ)	-0.2277	0.0726	-3.14
lambda individual	-1.086	0.165	-6.56

Chapter 6

Discussion and Conclusion

In the current work a series of models to deal with correlated data namely marginal, transition and random effects were fitted to correlated longitudinal respiratory disease outcome data. The data used was a multi-centre trial to compare two types of treatment including the effect of other covariate. The different types of models investigated can broadly be classified as marginal models based on the GEE method, transition models which include outcome history into the linear predictor, conditional or random effects models and lastly the recent hierarchical generalized linear models proposed by Lee and Nelder (1996). The analysis of the first three methods were accomplished in SAS but HGLM was fitted in GENSTAT. The results from fitting all the four different models show a significant dependence of the outcome Y (disease status) only on the treatment and baseline disease status. No significant effects on the other covariates (sex, age, centre, time) at 5% significance level were found. The transition model indicate that only the immediate past history is important in explaining the current respiratory disease status of an individual. Thus the data exhibit a strong first order markovian structure. However an interesting finding noted is that when both first and second order dependence on past outcome effects are included in the transition model age became significant. The results from random effects models are similar to the results from marginal model. However the interpretation of parameter estimates cannot be interchangeable. Proc GLIMMIX and NLMIXED in SAS gave similar results although Proc NLMIXED took much long to converge. This is not surprising because the method is computationally more intensive. The result show that the status of respiratory disease was not time dependence. The result from the GLMM show that the treatment does not affect the linear coefficient of the regression with time. The need to analyse large correlated data using generalized estimating equations and generalized linear mixed model is becoming necessary and in high demand in many areas of applications particularly in medical research, environmental pollution studies, ecology and many more. The results from the GLM and logistic regression assuming independence are important to consider and discuss. The results from the logistic regression model indicate that centre and age effects are significant at 5% significance level while GEE and GLMMs shows non-significance in these effects. The results assuming independence yield smaller standard errors compared to GEE, GLMM and HGLM models. This is because the simpler models assuming independence do not take into account the correlation between observations within an individual which seriously lead to inflated type I error. GEE, GLMM and HGLM models do take into account the correlation between observations. This is precisely the gain in correctly modelling the data by accounting for the inherent correlation in the data. The three model types (GEE, GLMM and HGLM) have capability to hand longitudinal or clustered correlated data. However the GEEs should be used if interest lies in estimating population averaged effects while GLMMs and HGLMs are used when interest lies in cluster specific effects and interpretation. GEEs model the marginal expectation of dependent variable as a function of covariates (Liang and Zeger, 1986; Zorn et al, 2001). Marginal models assume that the relationship between the outcome and covariates is the same for all subjects while the random effects model allows this relationship to differ between subjects (Diggle et al. 2002; Carrier et al, 2002). The random effects models were relevant to modelling the data set as it took into account of individual to individual heterogeneity. Significant variability at baseline was detected via a random intercept model. Under GLMMs and HGLMs one is also interested in differentiating between clusters while under GEEs this effect is averaged out. The generalized linear mixed model provided a flexible method for modelling the data. Clearly generalized linear models without modification cannot be able to handle the longitudinal/repeated measurements or generally clustered data that is why we used the generalized estimating equation and the generalized linear mixed model as methods of dealing with correlated data in the case of non-Gaussian responses. The GEE is a method of extending the generalized linear model to the case of longitudinal correlated data in general. Its derivation is an extension of the quasi-likelihood approach (Nelder and Pregibon, 1987). The extension allows one to model the error distribution through the specification of only the mean and variance structure of the data. GEEs have the capacity to produce both model and empirical based standard errors. The latter is better because it allows the standard errors to be data driven. Under GEEs inference about the mean structure is asymptotically valid even under a mis-specified correlation structure. The generalized linear mixed model is an alternative way of accounting for the correlated data by defining a random effect term which is constant for observations from the same subject or cluster but variable between experimental or observational units. The formulation of these models are likelihood based therefore better distributional properties. The standard errors for the generalized linear mixed model are lager than that of the generalized estimating equation. However the assumptions behind the two modelling approaches are fundamentally different hence one aught to handle this comparison with caution particularly in the case of non-Gaussian data. In the case of Gaussian data the switch between the marginal and conditional random effects models is pretty direct as opposed to the case of the non-Gaussian case. Generalized estimating equations are most naturally adapted for marginal models, not conditional random effect models as in the generalized linear mixed model. The GEE may be inefficient when the goal is estimation of the variance covariance structure. Generalized estimating equations by themselves do not help to separate out different sources of variation when it is often an advantage to be able to attribute variation as being associated with difference factors. Thus in summary GEEs are best suited if one is strictly interested in marginal or population averaged effects while GLMMs are stronger if one wishes to account for extra variability as a result of unit to unit heterogeneity or in more complex multi-level structure models such as interventions targeting girls residing in different households who attend different schools located in different locations within a province. One clearly sees that this type of data has more than one sources of variability which are best modeled using the GLMM approach and its extension HGLMs here (advantage over GLMMs). These are the individual to individual variability and measurement error variance. Finally the study also considered the problem of modelling respiratory infectious disease and the key lesson here is that such processes are complex and one first needs to understand the nature of the underlying disease process which generates the observations. In this particular study we considered data that is best defined as a respiratory process which denotes a disease process which cannot confer permanent immunity thus an infected individual immediately becomes susceptible upon recovery.

To model such disease processes the methods described in the project are adequate subject to some assumptions about the process being satisfied. In cases where the underlying latent disease process is not directly observed advanced techniques such hidden Markov models (HMM) may become necessary in order to adequately link the latent and the observation processes in parameter estimation. Transition models can however be used to capture and model the actual disease processes such as infection and recovery depending on how much information about the process that is contained in the data (Mwambi et al., 2011). However a future area of a research problem is to consider disease outcome data where some of the disease sub-process are partially observed.

Bibliography

[1] Amir, R., Hoque, A.M.W., Khan, R.F. and Rahman, M. (2009). Considering respiratory tract infection and antimicrobial sensitivity: An exploratory analysis. Malaysian Journal of microbiology, 5(2), 109-112.

[2] Baqiu, A.H, Rahman, M., Zaman, K., El Arifeen, S., Chowdhury, H.R., Begum, N., Bhattacharya, G., Chotani, R.A., Yunus, M., Santoshoms, M. and Black, R.E. (2007). A population based study of hospital admission incidence rate and bacteria aetiology of acute lower respiratory infections in children aged less than five years in Bangladesh. Journal of health, population and Nutrition 25, 179-188.

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

[4] Beslow, N.E and Lin, X. (1995).). Bias correction in generalized linear mixed models with single component of dispersion. Biometrka, 82, 81-91.

[5] Carrier, I. nad Bouyer, J. (2002). Choosing marginal or random effects models for longitudinal binary responses application to the self-reported disability among older person. BMC medical research methodology, 2, 15

[6] Chaganty, N.R. and Joe, H. (2004). Efficiency of generalized estimating equations for binary responses. Journal of Royal Statistical Society, B, 66, 851-860.

[7] Cox, D.R. (1972). Regression models and life tables. Journal of Statistical Society, 74, 187-200

[8] Cox, D. R. and Snell, E. J. (1989). Analysis of Binary Data. New York: Chapman and Hall.

[9] Davidian, M. and Giltinan, D. M. (1995).Nonlinear Models for Repeated Measurement Data, 1^{st} edn. Chapman and Hall.

[10] Davis, C. S. (2002). Statistical methods for the Analysis of Repeated Measurements. New York: Springer Verlag.

[11] Demidenko, E. (2004). Mixed Models Theory and Applications. John Wiley and Sons, Inc, Hoboken, New Jersey

[12] Diggle, P.J., Liang, K.-Y. and Zeger, S.L (1994). Analysis of Longitudinal Data. Oxford Science.

[13] Diggle, P.J., Heagert, P., Liang, K.-Y. and Zeger, S.L (2002). Analysis of Longitudinal Data, 2^{nd} edn. Oxford New York

[14] Diggle, P. J.(1998). An approach to the analysis of repeated measurements. *Biometrics* **44**, 959-971.

[15] Diggle, P. J., Heagarty, P., Liang, K-Y., and Zeger, S. (2002). Analysis of longitudinal data. *The Journal of Applied Statistics in the Pharmaceutical industry*. **3,2**, 147-148.

[16] Feller, W. (1968). An introduction to probability theory and its applications, 3^{rd} edn. Vol. 1. New York: Wiley.

[17] Fitzmaurice, G.M, Laird, N.M and Ware, J.H. (2004). Applied Longitudinal Analysis. John Wiley and Sons

[18] Goldstein, H. (1991). Multilevel models with an application to discrete response data. Biometrika, 78, 45-51.

[19] Goldstein, H. and Rasbash, J. (1996). Improving approximations for multilevel models with binary responses. Journal of the Royal Statistical Society, 159, 505-513.

[20] Goldstein, H. (2003). Multilevel statistical models. New York: John Wiley.

[21] Griffiths, D.A. (1973). Maximum likelihood estimation for the beta-distribution and an application to the household distribution of the total number of cases of a disease. Biometrics, 29, 637-648.

[22] Hardin, J.W.and Hilbe, J.M. (2003). Generalized Estimating Equations. London: Chapman and Hall.

[23] Hand, D. and Crowder, M. (1996). Practical Longitudinal Data Analysis, 1^{st} edn. London: Chapman and Hall

[24] Heagerty, P.J. (2002). Marginalized Transition Model and Likelihood Inference for Longitudinal categorical data. Biometrics 58, 342-351.

[25] Hedeker, D. and Gibbons, R.D. (2006). Longitudinal Data Analysis. John Wiley and Sons, Inc, Hoboken, New Jersey.

[26] Hilbe, J.M. (2009). Logistic Regression Models. USA: Chapman and Hall

[27] Hoi-Jeong Lim.(2011).QIC for GEE goodness-of-fit,http://blog.daum.net/dentalstat/7044538

[28] Kuk, A.Y.C. (1995). Asymptotically unbiased estimation in generalize linear models with random effects. Journal of Statistical Society, 57, 395-407.

[29] Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data. Biometrics 38, 963 - 974

[30] Le Cessie, S and Van Houeelingen, J.C. (1994). Logistic Regression for correlated Binary Data. Journal of Royal Statistical Society, 43, 95-108.

[31] Lee, Y. and Nelder, J.A. (1996). Hierarchical Generalized Linear Models. Journal of Royal Statistical Society, 58, 619-678.

[32] Lee, Y. Nelder, J. and Pawitan, Y. (2006). Generalized linear models with random effects. Chapman and hall, New York.

[33] Lee, Y. and Nelder, J. A. (1998). Generalized linear models for the analysis of quality improvement experiments. Canada. J. Statist. 26, 95-105

[34] Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalised linear models. *Biometrika* **73**, 13-22.

[35] Liang, K. Y., Zeger, S. L., Qaqish, B. (1992). Multivariate regression analyisis for categorical data (with Discusion). Journal of Royal Statistical Society, B. 54, 3-40.

[36] Lin, X. and Breslow, N.E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion.

[37] Lindsey, J.K. (1993). Models for Repeated Measurements. New York

[38] Mcullagh, P. (1983). Quasi-likelihood functions. *The Annals of Statistics* **11**, 59-67.

[39] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Second Edition, London: Chapman and Hall.

[40] Milliken, G. and Johnson, D. (2009). Analysis of messy data: Designed Experiments. 2^{nd} Edition, Chapman & Hall/CRC Press, Florida , USA.

[41] Molenberghs, G. and Verbeke, G. (2005). *Models for discrete longitudinal data*. Springer Series in Statistics, Springer-Verlag, New-York.

[42] Mwambi, H., Ramroop, S., White, L. J., Nokes, D. J., Okiro, E. A., Shkedy, Z. and Molenberghs, G. (2011). A frequentist approach to estimating the force of infection for a respiratory disease using repeated measurement data from a birth cohort. *Statistical Methods in Medical Research*, 20, 551-570.

[43] Nelder, J .A. and Wedderburn, R. W (1972). Generalized linear models. Journal of the Royal Statistical Society.135, 370-384

[44] Nelder, J. A. and Pregibon, D. (1987). An extended quasi-likelihood function. Biometrika, 74, 221-232.

[45] Neuahous, J.M. (1992). Statistical methods for longitudinal and clustered designs with binary response. Medicine, 1, 249-273.

[46] Pan, W. (2001), Akaike's information criterion in generalized estimating equations, Biometrics, 57, 120-125.

[47] Pinheiro, J.C. and Bates, D.M. (1995) Approximations to the log-likelihood function in the non-linear mixed- effects model. Journal of computational and graphical statistics, 4, 12-35.

[48] Pinheiro, J.C and Bates, D.M. (2000). Mixed effects models in S and S-plus. Springer-Verlag: New York.

[49] Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary obseravtion. Biometrics, 44, 1033-1048.

[50] Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1992). Numerical Recipes in C: The art of scientific computing. Cambridge university press, Cambridge Uk.

[51] Qin, J. and Lawless, J. (1994). Generalized Estimating Equations. The annals of Statistics, 22, 300-325.

[52] Raghunathan, T.E and Yoichi II. (1993). Analysis of binary data from a multicentre clinical trial. Biometrika 80, 27-39.

[53] Rahman, M.M and Rahman, A.M. (1997). Prevalence of acute respiratory tract infection and its risk factors in under five children. Bangladesh Medical Research Council Bulletin 23, 47-50.

[54] Raudenbush, S.W., Yang, M.L., and Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximations, Journal of computational and graphical statistics, 9, 141-157.

[55] Rodriguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary response. Journal of Statistical Society, 158, 73-89.

[56] Rosner, B. (1984). Multivariate methods in ophthalmology with application to other paired-data situations. Biometrics, 40, 1025-1035.

[57] Skellam, J.G. (1948). A Probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. Journal of Statistical Society, 10, 257,261.

[58] Skrondal, A. and Rabe-Hesketh, s. (2004). Generalized Latent variable modelling. London: Chapman and hall.

[59] Sutradhar, B.C. and Rao, R.P. (2001). On Marginal Quasi-likelihood Inference in Generalized Linear Mixed Models. Journal of Multivariate Analysis, 76, 1-34.

[60] Smyth, G.K. and Verbyla, A.P. (1999). Adjusted Likelihood methods for Modelling dispersion in generalized linear models. Environmetrics, 10, 695-709.

[61] Tierny, L. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association, 81, 82-86.

[62] Tuerlinckx, F., Rijmen, F., Molenberghs, G., Verbeke, G., Briggs, D., Noortgate, W.vanden, Meulders, M., and Boeck, P. De (2004) Estimation and software, In P. De Boeck and M. Wilson, editors, Explanatory item response models, Statistics for social science and public policy, chapter 12, 343-373, Springer-Verlang, New York.

[63] Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer Series in Statistics Springer-Verlag, New-York.

[64] Ware, J.H., Lipsitz, S., and Speize, F.E. (1988). Issues in the analysis of repeated categorical outcomes. Statistics in medicine, 7, 95-107.

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

[66] Whitehead, J. (1992).The design and analysis of sequential clinical trials, edn

[67] Wolfinger, R. and O'Connell, M. (1993). Generalized Linear Mixed Models: a pseudo-Likelihood Approach, Journal of Statistical computation and simulation, 48, 233-243.

[68] Yang, X., Shoptaw, S., Nie, K., Liu, J and Belin, T.R. (2007). Markov transition models for binary repeated measures with ignorable and nonignorable missing values. Statistical methods in medical research, 16(4), 347-364.

[69] Zeger, S. L., and Liang, K. Y. (1986). The analysis of discrete and continuous longitudinal data. *Biometrics* **42**, 121-130.

[70] Zorn, C.S.W (2001). Generalized Estimating Equation Models for Correlated Data, American Journal of Political Science, 45, 470-490

APPENDIX

```
/* SAS code for fitting GEE models/
proc genmod data=sya descending;
class id centre trt sex baseline time;
model status=centre trt sex age baseline time / dist=bin type3;
repeated subject=id(centre) / type=cs modelse ;
run;
```

```
/* The SAS code using MQL and PQL/
```

SAS code (PQL) with random intercept

```
proc glimmix data=sya methods=RSPL;
class id centre trt sex baseline;
model status (event='1') =centre trt sex age baseline time /dist=bin solution;
random intercept /subject=id(centre) type=un;
random residual/ subject=id(centre);
run; MQL is obtained with option 'method=RMPL'
```

Inclusion of random slope:

```
random intercept time /subject=id(centre) type=un;
```

```
/* GENSTAT code for fitting HGLM/
HGFIXEDMODEL [DISTRIBUTION=binomial; LINK=logit; DISPERSION=*; CONSTANT=estimate;
FACT=3]
centre,trt,sex,age,baseline,time
HGRANDOMMODEL [DISTRIBUTION=beta; LINK=logit] id
HGANALYSE [PRINT=model,fixed,random,dispersionest,monitoring,likelihoodstat; MLAPLACE=0;
```

DLAPLACE=0] status; NBINOMIAL=1