Modelling Longitudinal Counts Data with Application to recurrent Epileptic seizure events.

Phathisani Ngulube

Submitted in partial fulfillment of
the requirements for the Degree of

Master of Science

in

Statistics

in the
School of Statistics and Actuarial Science
Pietemaritzburg
University of KwaZulu-Natal

October 2010

Declaration

The research work is the original work done by the author (Phathisani Ngulube, 204518333) and it is not a duplicate of some of the research work done by other authors. All the references that were used to refer to are duly acknowledged.

October 2010.

Student:
_____
Ms Phathisani Ngulube
Supervisor:
_____
Prof Henry Mwambi
Co-supervisor:
_____
Dr Shaun Ramroop

Acknowledgement

I would like thank Dr. Henry Mwambi, my supervisor, for all his guidance, encouragement and long hours he spent helping me. Without him this project would not have been possible. I would also like to extend my gratitude to my co-supervisor Dr. Shaun Ramroop, for his input and encouragement.

I would also like to thank my colleagues for their words of encouragement and input.

To my friends and family, thank you all your moral support and for putting up with me during this project. Now I can catch up on all the socializing I missed while my nose was "buried" in books.

Abstract

The objectives of this thesis is to explore different approaches of modelling clustered correlated data in the form of repeated or longitudinal counts data leading to a replicated Poisson process. The specific application is from repeated epileptic seizure time to events data. Two main classes of models will be considered in this thesis. These are the marginal and subject or cluster specific effects models. Under the marginal class of models the generalized estimating equations approach due to Liang and Zeger (1986) is first considered. These models are concerned with population averaged effects as opposed to subject-specific effects which include random subject-specific effects such that multiple or repeated outcomes within a subject or cluster are assumed to be independent conditional on the subject−specific effects. Finally we consider a distinct class of marginal models which include three common variants namely the approach due to Anderson and Gill (1982), Wei et al (1989) and Prentice et al. (1981)

Contents

Chapter 1

Introduction

There is increasing interest, and need, to extend and apply the theory of the analysis of time to event data to data sets with multiple events per subject. The types of multiplicity can either be multiple events per subject or events of different type. Examples of the former include recurrent infections in AIDS patients (such as recurrent bacterial pneumonia and), recurrent epileptic seizures for individuals who are epileptic. Examples of the latter include recurrent side effects such as toxicity and worsening symptoms in the management of chronic diseases. Such type of processes give rise to a replicated Poisson process since each subject can be considered as generating its own single Poisson process. With the need to evaluate the success of treatment strategies such as HAART (highly active antiretroviral treatment) in the management of HIV/AIDS patients and other dilapidating diseases such analyses are becoming increasingly necessary.

A major issue in the extension of time to event technique such as the Cox proportional hazards regression models is the intra-subject correlation for events from the same subject. Other complexities include multiple time scales, discontinuous intervals of risk, stratum by covariate interactions, and the structure of risks sets.

The counting process modelling approach in this thesis was motivated by the need to more accurately model the disease process of interest which is epilepsy in this case. There is a growing need to understand this condition more and more in order to design better strategies of care for persons having such a dilapidating condition (Madekurozwa, 2009).

This thesis investigates two main categories of modelling approaches to such data. The first category is the marginal models which include three common

variants namely the approach due to Anderson and Gill (1982), Wei et al. (1989) and Prentice et al. (1981). Within the class of marginal models we also consider the class of marginal models for general longitudinal data is the generalized estimating equations approach by Zeger et al. (2002).

A second category of models that the project will be concerned with is the random effects or subject-specific models which include a random subject-specific effect, where multiple outcomes are assumed to be independent conditional on the subject-specific effect. An example of random specific effects model in time to event modelling is the frailty model described by Oakes (1992).

In this thesis the aim is to

- model multiple events per subject as a replicated Poisson process
- Generalize the intensity function generating the events to include the effect of treatment and other subject and/or population specific covariates
- Extend the models to capture the correlation structure of events per subject
- Apply the models to real data in medical research
- Compare the two categories of modelling approaches
- Recommend future extensions

To attain the above aims we need to first look at some preliminary concepts and these are discussed in Chapter two.

Chapter 2

Preliminary concepts

2.1 Longitudinal data

Longitudinal data occurs when a response Y is observed repeatedly on the same individual over time (Verbeke and Molenberghs, 2000; Diggle, Heagerty, Liang and Zeger, 2002). Often in clinical trials we need to examine the effect of a new treatment compared to an existing or standard treatment in curing or alleviating pain on study individuals. In this a set of individuals are randomly allocated to the two treatments then the response of interest measured over time on the two groups of patients with an aim of comparing the benefit accrued from the two treatments. In addition to measuring the response variable other covariates are also measured alongside the response.

The main advantage of longitudinal studies over cross sectional studies is that longitudinal studies can separate the marginal or population averages and individual specific effects in population studies for example cohort and age effects. The changes within individuals over time is what is known as age or time effect. The cohort effect is the differences among people in their baseline values or covariates. Longitudinal studies can distinguish these time and cohort effects while cross-sectional studies cannot. In cross-sectional data, only a single response is available for each of the experimental units (for example human subjects or plants). The following are some of the specific advantages of longitudinal studies:

- They economise on subjects thus they cut on costs.
- Subjects can serve as own control.
- The between-subject variation can be modelled by including a random effect to account for subject to subject variability or heterogeneity.

- They can provide more efficient estimators than cross-sectional designs with the same number and pattern of observations.
- They can provide information about individual change.

Just like any other study, longitudinal studies have challenges of which some of them are listed below:

- Observations are not, by definition, independent implying they are usually correlated, hence it is important to take account of possible association among repeated measurements.
- In most cases analysis methods are not well developed, especially for more sophisticated models, such as models for non-Gaussian data.
- Lack of readily available computational procedures and software.
- The cases of unbalanced designs, missing data and attrition which make analysis of data slightly difficult.
- Time varying covariates.

2.2 Types of correlation structures

In the previous section we mentioned that in longitudinal studies the observations are correlated. Of which they are different types of correlation structures longitudinal data exhibit, some of the structures are :

- Independence: which means observations within an individual or cluster are assumed independent which is somewhat unrealistic in practical terms.
- Exchangeable: that is all measurements on the same unit are equally correlated. This type of correlation structure is sometimes referred to as the compound symmetry or spherical structure. It is used when one is dealing with clustered data. Examples include the case of observations accrued from individuals from the same family or geographic area.
- Completely unstructured: here no assumptions are made about the correlations. Thus there are many parameters to estimate. One problem with using this correlation structure is that the correlation estimates may not converge.

- Autoregressive structure: this kind of correlation structure is used in the cases when one knows that the correlation declines over time.

The above list is not exhaustive because much more structures exist depending on the nature of the data.

2.3 Notation

In the text that follows we will denote the response variable by Y and the explanatory variables, or covariates by **X**. To specify the longitudinal or repeated measurements data setting we let Y_{ij} denote the j^{th} observation for the ith individual for i=1,...,n. The actual observation times are t_{ij} where $j = 1, \ldots, n_i$ assuming individual i is observed or measured n_i times. If we have two treatment groups then $N = N_1 + N_2$ where N_1 is the number of individuals allocated to treatment 1 (for example a placebo) and N_2 the number of those allocated to the second treatment (for example a new or active treatment).

2.4 Incomplete data

Longitudinal data may involve a time to event response. In this context of time to event analysis we distinguish between uncensored (complete) and censored (incomplete) data or observations for both life and non-life studies. In many cases, life data contains uncertainty as to when exactly an event happened (i.e. when the unit failed). Data containing such uncertainty as to exactly when the event happened is termed as censored data or incomplete data depending on the context of application. Under survival analysis it is referred to as censored data and under longitudinal data settings incomplete data (Verbeke and Molenberghs, 2000). Analysis of time to event data or survival analysis is a very popular methodology in health research (Therneau and Grambosch, 2000).

Complete Data

Complete data means that the value of each sample unit is observed or known. For example, if we had to compute the average test score for a sample of ten students, complete data would consist of the known score for each student. Likewise in the case of life data analysis, our data set (if complete) would be composed of the time-to-failure of all units in our sample. Consider a situation in which we are testing 5 (non repairable) units taken randomly from

a population. We are investigating the population to determine if its failure rate is acceptable. In the typical test scenario, we have a fixed time interval T to run the units to see if they survive or fail. If the tested five units all failed and their times-to-failure were recorded, we would then have complete information as to the time of each failure in the sample. Missing or incomplete longitudinal data theory is still an active area of research particularly with reference to clinical trials (Molenberghs and Kenward, 2007).

Censored Data

In many cases when life data are analyzed, all of the units in the sample may not have failed (i.e. the event of interest was not observed) or the exact times-to-failure of all the units are not known. This type of data is commonly called censored data . There are three types of possible censoring schemes, namely right censored (also called suspended data), interval censored and left censored.

Right Censored (Suspended)

The most common case of censoring is what is referred to as right censored data, or suspended data. In the case of life data, these data sets are composed of units that did not fail. For example, if we tested five units and only three had failed by the end of the test, we would have suspended data (or right censored data) for the two units which did not fail. The term "right censored" implies that the event of interest (i.e. the time-to-failure) is to the right of our data point. In other words, if the units were to keep on operating, the failure would occur at some time after our data point (or to the right on the time scale).

Interval Censored

The second type of censoring is commonly called interval censored data. Interval censored data reflects uncertainty as to the exact times the units failed within an interval. This type of data frequently comes from tests or situations where the objects of interest are not constantly monitored. If we are running a test on five units and inspecting them every 100 hours, we only know that a unit failed or did not fail between inspections. More specifically, if we inspect a certain unit at 100 hours and find it is operating and then perform another inspection at 200 hours to find that the unit is no longer operating, we know that a failure occurred in the interval between 100 and 200 hours. In other words, the only information we have is that it failed in a certain interval of time. This is also called inspection data by some authors.

For infectious diseases such as HIV, an individual may test negative when tested at time t_1 then later test positive at time t_2 then the actual time of sero-conversion lies between t_1 and t_2. Then the time to event is interval censored.

Left Censored

The third type of censoring is similar to the interval censoring and is called left censored data. In left censored data, a failure time is only known to have occurred before a certain time. For instance, we may know that a certain unit failed sometime before 100 hours but not exactly when. In other words, it could have failed any time between 0 and 100 hours. This is identical to interval censored data in which the starting time for the interval is zero. Alternatively we may be interested in age at infection with a disease such as HIV/AIDS. But because individuals are not followed continuously if an individual tests positive at age a then this means the age at infection is $a' < a$. The age at infection is left censored. If not the age at infection is $a > a'$ which in this case it will be right censored. This type of data is generally known as current status data.

Ignoring the different forms of censoring can lead to biased results particularly if the censored observations are very different from those remaining. (Therneau and Grambosch,2000 ; Pawitan, 2006)

2.5 Types of longitudinal studies

Some types of longitudinal studies are:

- panel studies
- time series analysis where a single variable(s) is measured at different time points for example monthly for several years.
- cohort data sets, this is where individuals are followed over time and a disease outcome or certain event of interest occurs or until the end of study.
- event history data sets. These are also known as survival data analysis.
- repeated cross sections. This is the most common type of study in longitudinal survey studies. It involves whole surveys with the same variable measured repeatedly at different time points.

At times observation gaps occur in longitudinal studies. This happens when study subjects can be out of the study for a period of time for various reasons and then come back again and this may occur more than once. This is also called intermittent missingness. In the following section other types of missingness will be discussed.

2.6 Missingness and Incompleteness

Longitudinal studies usually suffer from incompleteness even though they are designed to collect information for every subject on each planned occasion. Incompleteness of longitudinal studies results in missing observations. These missing observations maybe due to, for example causes that are not related to the responses (observed or not observed) or the nature of the study procedure. This type of missingness is usually referred to as missing completely at random (MCAR). In most clinical trials the reason for missingness depends on the response variables themselves therefore MCAR assumption rarely holds in clinical trials.

Rubin (1976) classified the missing data mechanism into three types based on how the missing-data processes depend on the responses: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). In the context of longitudinal data, the term MAR refers to the case in which the probability that a response measurement is missing may depend on other observed parts of the response profile, but does not depend on the unobserved response(s). When the missingness depends on the unobserved response(s) then the missing data is said to be MNAR. The type of incompleteness addressed in the current work is from a Poisson process for counts. Individuals are observed over different lengths of time T_i, $i = 1, \ldots, n$ and the number of events experienced by the individual over this duration of time counted. The type of incompleteness occurring here is that the number of observations per individual is not the same for all individuals n and further the time between events is not constant between and within individuals. This leads to unbalanced incomplete data that complicates its analysis within the longitudinal data framework.

Chapter 3

Data Description

In this project we consider modelling a special type of a disease process where the underlying generation process is in the form of a replicated Poisson process. The disease of interest is epilepsy where individuals are randomized into a treatment arm and a placebo arm and then observed repeatedly overtime. This generates a form of repeated or longitudinal non-Gaussian data. Generally each individual generates his/her own short time series as opposed to the classical time series analysis where we have one long time series data (Talke, 2003; Diggle et al, 2002).

3.1 Epilepsy

Epilepsy is a neurological condition that from time to time produces brief disturbances in the normal electrical functions of the brain [2]. Epilepsy is also known as seizure disorder. Normal brain function is made possible by millions of tiny electrical charges passing between nerve cells in the brain and to all parts of the body. When someone is epileptic, the normal pattern may be interrupted by intermittent bursts of electrical energy that are much more intense than usual. This may lead to the person's consciousness, bodily movements or sensations being affected for a short time.

These physical changes are called epileptic seizures, hence epilepsy is sometimes called a seizure disorder. The unusual bursts of energy may occur in just one area of the brain (partial seizures), or may affect nerve cells throughout the brain (generalized seizures)[2]. Normal brain function cannot return until the electrical bursts subside. Conditions in the brain that produce these episodes may have been present since birth, or they may develop later in life due to injury, infections, structural abnormalities in the brain, expo-

sure to toxic agents, or for reasons that are still not well understood. Many illnesses or severe injuries can affect the brain enough to produce a single seizure. When seizures continue to occur for unknown reasons or because of an underlying problem that cannot be corrected, the condition is known as epilepsy. Epilepsy affects people of all ages, all nations, and all races. Epilepsy can also occur in animals, including dogs, cats, rabbits, and mice. In this project two epilepsy data sets will be considered. The first data set which is printed in Table 3.1 is about a clinical trial where a placebo and an active treatment are compared. The data is a typical example of real replicated Poisson process (Pawitan,2006). Individual specific information included in the first data set is the number of epileptic seizures n_i over a period T_i the individual was followed up. The treatment group (placebo or active treatment) was the only covariate in the model. The different time points where the epileptic seizures occurred $(t_{i1}, t_{i2}, \ldots, t_{in_i})$ are also available where $0 \leq t_{ij} \leq T_i$.

The second data set is similar to the first one but now the patients were followed for the same period of time and all patients have the same number of observations. The patients were randomized to receive a control (placebo) and a test drug (progabide) in a two period crossover trial. The second data set was first analyzed by Thall and Vail (1990). These data are reprinted in Table 3.2. From the box plots in Figure 3.1 one can see that there seems to be a strong variation in the count of seizures among the people at the baseline level. The medians for the seizures at the baseline level are higher than the medians at subsequent visits. Figure 3.2 shows the subject profiles of the individuals and from the figure we can see that there is great variability across time among individuals.

The aim in both cases is to model the intensity of epileptic events for individuals under placebo and treatment and to test whether the intensity of occurrence are significantly different between the two treatment groups.

Table 3.1: The epilepsy data example 1: Patient i was followed for T_i weeks, and there were n_i events during the follow-up period.

Subject i	x_i	T_i	n_i	Time of events
1	active	12	3	2.6 3.3 7.2
2	active	5	2	3.5 4.4
3	active	7	4	1.5 1.6 2.2 6.1
4	active	14	3	12.1 12.4 13.4
5	active	10	5	0.7 2.6 3.9 6.9 7.8
6	active	10	2	5.3 6.3
7	active	12	1	10.2
8	active	8	3	0.2 3.2 7.7
9	active	11	3	0.1 2 3.2
10	active	8	3	0.1 3.2 3.7
11	placebo	11	4	2.3 7.9 8 8.8
12	placebo	11	7	5.1 5.2 6.1 6.5 7.9 9.9 10.9
13	placebo	8	6	0.5 0.8 1.9 2.7 5.4 7.2
14	placebo	16	8	1.4 4.3 5 6 7.8 8.4 9.2 11.2
15	placebo	11	11	0.3 0.3 1.9 1.9 2.7 3.1 3.9 5.3 7 8.8 10.1
16	placebo	7	8	1.2 2.6 3.5 4.7 5.3 5.7 5.9 6.1
17	placebo	15	7	0.8 1.5 4.3 4.4 5.1 12.1 14
18	placebo	9	7	0.1 0.1 1 3.6 5.4 6.3 8.7
19	placebo	7	4	0.9 2.2 5.2 6.6
20	placebo	4	2	2.2 3.2
21	placebo	6	6	0.5 1.3 1.3 1.7 2.9 5.6
22	placebo	4	1	1.4

Table 3.2: The epilepsy data example 2: Patient i was followed for T_i weeks, and there were n_i events during the follow-up period.

Subject i	Trt	Age	Base	Y_1	Y_2	Y_3	Y_4
1	0	31	11	5	3	3	3
2	0	30	11	3	5	3	3
3	0	25	6	2	4	0	5
4	0	36	8	4	4	1	4
5	0	22	66	7	18	9	21
6	0	29	27	5	2	8	7
7	0	31	12	6	4	0	2
8	0	36	52	40	20	23	12
9	0	37	23	5	6	6	5
10	0	28	10	14	13	6	0
11	0	36	52	26	12	6	22
12	0	24	33	12	6	8	5
45	1	35	38	19	7	6	7
46	1	25	7	1	1	2	4
47	1	26	36	6	10	8	8
48	1	25	11	2	1	0	0
49	1	22	151	102	65	72	63
50	1	32	22	4	3	2	4
51	1	25	42	8	6	5	7
52	1	35	32	1	3	1	5
53	1	21	56	18	11	28	13
54	1	41	24	6	3	4	0
55	1	32	16	3	5	4	3
56	1	26	22	1	23	19	8
57	1	21	25	2	3	0	1
58	1	36	13	0	0	0	0
59	1	37	12	1	4	3	2

Figure 3.1: Box plots os square-root transformed seizure rates for epileptics at the baseline (0) and for four subsequent two-week periods: (a)control (placebo); (b) test drug (progabide).

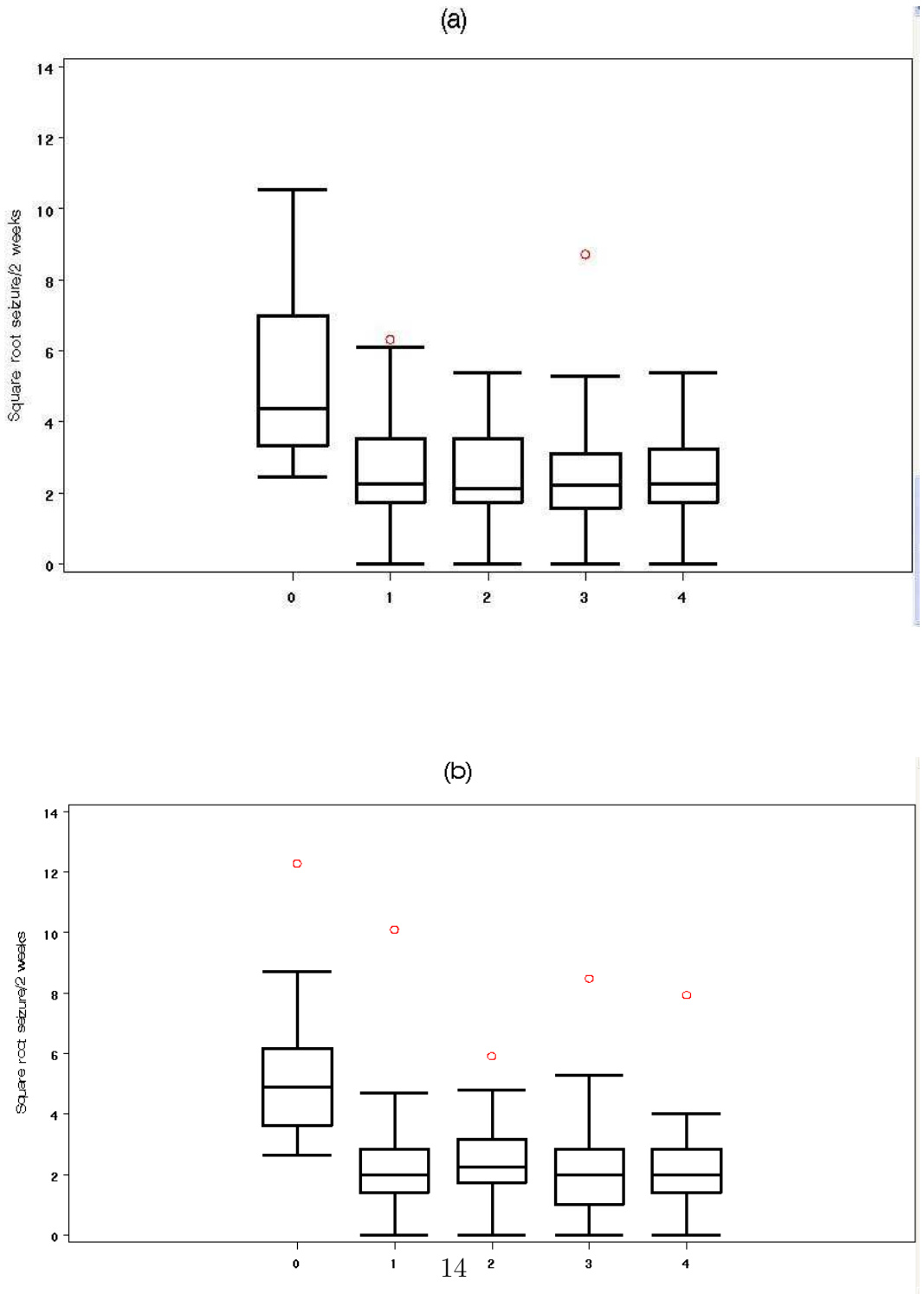
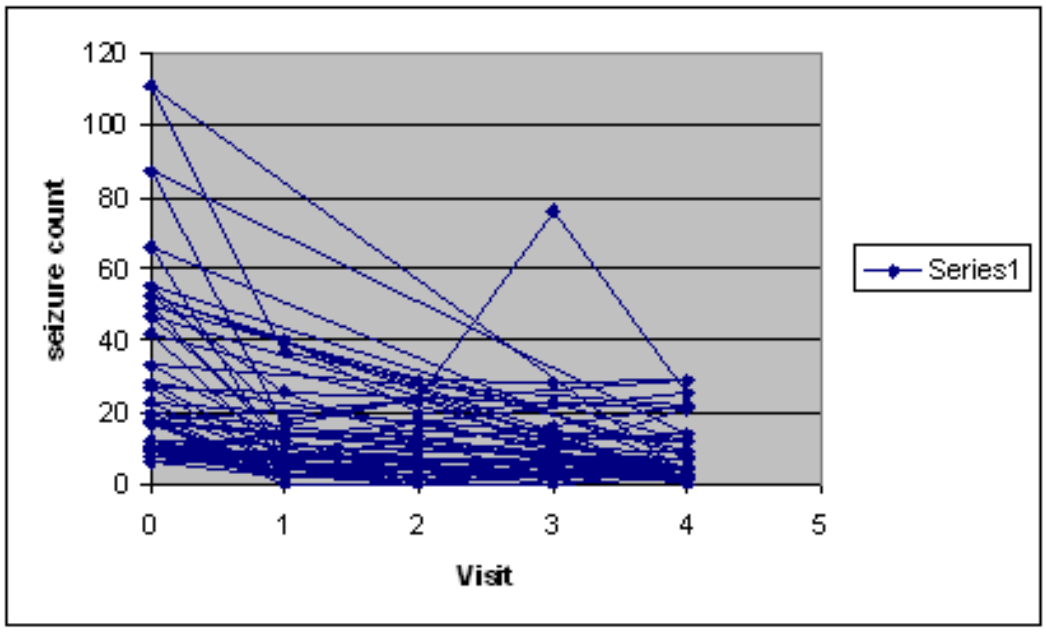
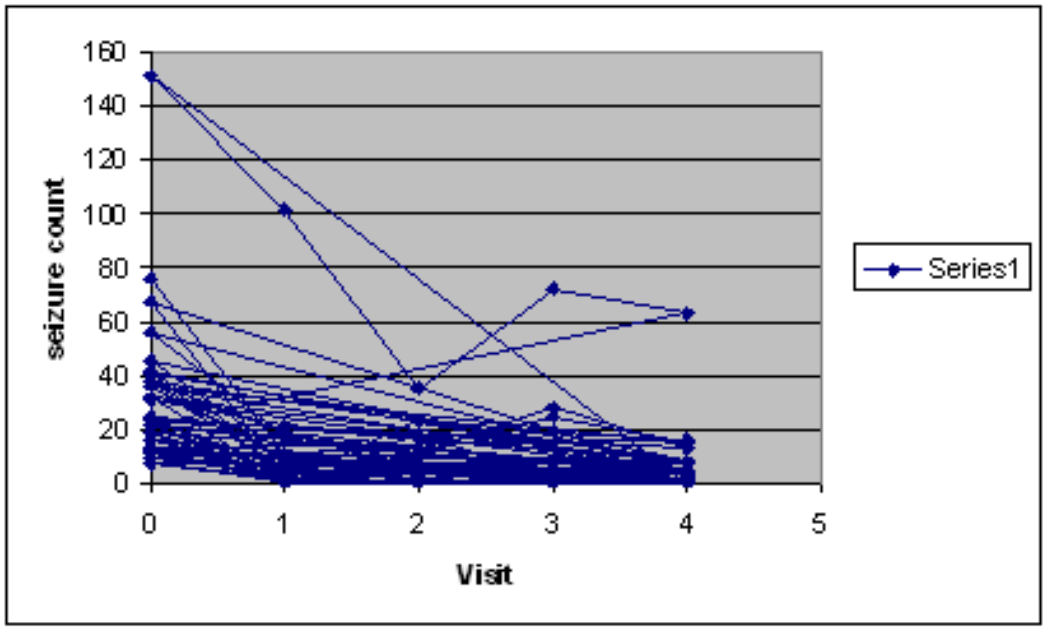


Figure 3.2: Subject profiles from (a) placebo (b) progabide.





Chapter 4

The General Linear Mixed Model for Longitudinal Data

4.1 Introduction

The following chapter summarizes the key ideas about linear mixed models for longitudinal data following the approach in Verbeke and Molenberghs (2000). This is a more general approach to fitting linear models where some covariance structure is expected between observations from the same unit or cluster (e.g. a patient observed repeatedly over time in a clinical trial) of measurement. For example the monotonically declining correlation structure between two observations as the distance between them in time increases. In particular if Y_{ij} and Y_{ik} are two observations from individual i measured at times t_{ij} and t_{ik} respectively we could hypothetically assume that the covariance between the two observations is $\sigma^2\rho^{|t_{ij}-t_{ik}|}$ where σ^2 is the constant variance of any individual observation hence $\rho^{|t_{ij}-t_{ik}|}$ is the correlation between them. If the observations are equally spaced such a covariance model is also known as the auto-regressive model of order one or AR(1. The book by Diggle et al. (2002) among others is also a good references for both balanced and unbalanced incomplete longitudinal data for continuous Gaussian outcomes. There are a number of statistical computing software available to handle the analysis of longitudinal data, from a continuous response. In SAS such models are fitted using proc MIXED. Other correlation structures also exist among them the compund symmetry (CS) where the covariance between any two observations within a cluster is constant. The unstructured (UN) covariance structure is where the covariance between any two observations for example Y_{ij} and Y_{ik} is left completely free. The disadvantage here is that the number of parameters to estimate may be too large particularly

if the number of observations per individual is large.

4.2 The Multivariate Regression Model

To specify the longitudinal or repeated measurements data we let Y_{ij} denote the jth observation for individual $i = 1, \ldots, N$. The actual observation times are t_{ij} where $j = 1, \ldots, n_i$ assuming the response of interest is observed or measured n_i times from individual i. Furthermore we let Y_i be the n_i-dimensional vector of all repeated measurements for the ith subject, that is, $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$. Assuming an average linear trend for Y as a function of time, a multivariate regression model can be obtained by assuming that the elements Y_{ij} in Y_i satisfy the model

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij},$$

with the assumption that the error components ε_{ij} are normally distributed with mean zero. Putting this in vector notation we have

$$Y_i = X_i\beta + \varepsilon_i \tag{4.1}$$

for a design matrix X_i of dimension $n_i \times p$, with β being a p- dimensional vector of regression parameters including the intercept β_0 and ε_i is an n-dimensional vector of error terms ε_{ij}, $j = 1, 2, \ldots, n_i$. The complete multivariate model is then obtained by assuming that $\varepsilon_i \sim N(0, \Sigma_i)$ and that $n_i = n$ for all i so that for the multivariate model the Y_i are independent $\sim N(X_i\beta, \Sigma_i)$ with a common variance-covariance matrix that is Σ equal to $\sigma^2 I$ where I is the identity matrix of dimension n. However this assumption is not necessarily true in the context of repeated or longitudial data settings. Making this assumption means that we ignore the fact that measurements made on the same individual may be (highly) correlated. To take into account the fact the repeated measurements are correlated "one" could assume special forms of Σ depending on the nature of the data and the most appropriate covariance structure that best describes the data. Some of the covariance structures one could assume include the compound symmetry structure (CS) where the covariance between any two observations within a cluster is constant say equal to σ_b^2, the autoregressive (AR) covariance structure similar to that described in the introduction above and many others. The CS structure arises assuming the model

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij} + s_i,$$

where now we assume s_i are iid $N(0, \sigma_b^2)$. Thus

$$Var(Y_{ij}) = \sigma^2 + \sigma_b^2$$

and

$$Cov(Y_{ij}, Y_{ik}) = E(s_i^2) = \sigma_b^2.$$

Thus

$$Corr(Y_{ij}, Y_{ik}) = \frac{\sigma_b^2}{\sigma^2 + \sigma_b^2}$$

where ρ is also known as the intra cluster correlation (ICC) in the context of clustered data. Assuming independence across individuals, β and the parameters in Σ_i can be estimated by maximizing the likelihood given by

$$L_{ML} = \prod_{i=1}^{N} (2\pi)^{\frac{n_i}{2}} |\Sigma_i|^{-\frac{1}{2}} exp(-\frac{1}{2})(Y_i - X_i\beta)'\Sigma_i^{-1}(Y_i - X_i\beta)) \tag{4.2}$$

where y_i is the observed vector of responses from the ith individual or cluster. Inference on the regression parameters β and covariance parameters in Σ are based on classical maximum likelihood theory such as the likelihood ratio (LR) test and asymptotic Wald test. More details on inference will follow in the next section. For now it suffices to say that the multivariate regression model is primarily suitable when measurements are taken at relatively small number of fixed time points and the data is balanced in the sense that each individual contributes equal number of observations n. The model can still be applied even if some measurements are missing provided the software allows for unequal number of measurements per individual or cluster. In SAS procedure MIXED, repeated and unequal number of observations per subject is handled by the REPEATED statement. For example if the repeated time of measurements are held in a categorial variable 'timef' then the statement 'repeated timef/...' is used from which outcomes that have been repeatedly observed and which are missing can be identified. Other software such as GenStat follow the same reasoning but the syntax is software specific.

However in the case of large number of repeated measurements, multivariate regression models can only be applied under very restrictive and very specific covariance structures, even in the case of complete data. The reason for this is obvious since in the case of the unstructured mean and/or unstructured covariance models there will be very many parameters requiring estimation. On the other hand in the case of highly unbalanced unequally spaced observation data, multivariate regression models can again only be applied under very specific mean and covariance structures. For example the AR(1) structure is not directly meaningful since the time points are not equally spaced. Likewise the CS covariance structure is meaningful subject to very strong assumptions. This therefore requires that the above model be modified and extended to address some of these deficiencies and data complexities. In the

following the problem of formulating the general linear model for longitudinal data is addressed.

4.3 A model for Continuous Longitudinal Data

From the Multivariate Regression Model section above one can see that this kind of model can only be used when one is working with a balanced data set but then in most practical cases one usually ends up with unbalanced data, that is either the data set will have unequal number of measurements per subject or measurements in the data set will not have been taken at fixed time points or equally spaced. One can however estimate subject-specific longitudinal profiles using linear regression functions. This leads to the concept of the 2-stage model formulation approach which is discussed in the next subsection. This approach will immediately lead to the general linear mixed effect(s) model. Inference on fixed effects parameters, variance components and random effects will be discussed. A model for residual covariance will also be presented. A motivation for the above model arises from the fact that in most cases longitudinal data is unbalanced due to (i) unequal number of measurements per person and (ii) measurements not taken at same fixed time points and equally spaced (because a situation may arise where individuals contribute equal number of observations but not equally spaced) for all individuals or both. Thus balanced data under such a model is just a special case. Because of (i) and (ii) classical multivariate techniques are often not applicable. The two-stage model formulation or reasoning arises as follows. Often, subject specific longitudinal profiles are well approximated by linear regression functions. This is done under stage 1 part of the formulation. In stage 2 one then builds a mean model to explain variability in the subject specific regression coefficients using covariates. Thus stage 2 formulation is in a way linking individual specific information to population level information.

The linear mixed model for longitudinal data was first described in Laird and Ware (1982). In order to formally derive a model relevant in the analysis of longitudinal data we adopt the two-stage formulation approach also described in Verbeke and Molenbergh (2000) and Diggle et al. (2002).

4.3.1 The two State formulation

Recall that to model longitudinal or repeated measurements data we let Y_{ij} denote the jth observation for individual $i = 1, \ldots, N$. The actual observa-

tion times are t_{ij} where $j = 1, \ldots, n_i$ assuming individual i is observed or measured n_i times. Furthermore we let Y_i be the n_i dimensional vector of all repeated measurements for the ith subject, that is, $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$. In the first stage of the two-stage formulation we assume a linear regression model for each subject separately is given by:

$$Y_i = Z_i\beta_i + \varepsilon_i \tag{4.3}$$

where Z_i is a $(n_i \times q)$ matrix of known individual specific covariates, β_i is a q-dimensional vector of unknown subject specific regression coefficients and ε_i is a vector of residual components ε_{ij}, $j = 1, \ldots, n_i$. Without loss of generality we assume that all ε_i are independent and normally distributed with mean vector zero, and covariance matrix $\sigma^2 I_{n_i}$, where I_{n_i} is the n_i-dimensional identity matrix. Note that the model in equation (4.3) above is describing within subject variability. In the second stage of the two stage formulation, between-subject variability is now modelled by relating β_i to known population level covariates such that

$$\beta_i = K_i\beta + b_i \tag{4.4}$$

K_i is a $(q * p)$ matrix of known covariates, β is a p-dimensional vector of unknown regression parameters and the b_i are assumed to be independent following a q-dimensional multivariate normal distribution with mean vector zero and general covariance matrix G. Next we substitute equation (4.4) into equation (4.3) which then leads to the general linear mixed model general linear mixed model (gLMM) given by

$$Y_i = X_i\beta_i + Z_ib_i + \varepsilon_i \tag{4.5}$$

where $X_i = Z_iK_i$ is a $n_i \times p$ matrix of known covariates while the rest of the terms remain as defined in model (4.3) and (4.4). Thus we finally get the gLMM model which can be specified as

$$\begin{aligned} &Y_i = X_i\beta + Z_ib_i + \varepsilon_i \\ &b_i \sim N(0, G), \\ &\varepsilon_i \sim N(0, \Sigma_i), \\ &i = 1, 2, \ldots, N \end{aligned} \tag{4.6}$$

where β denotes the fixed effects, b_i are the subject specific random effects and the elements in G and Σ_i are known as variance components. There are two ways of specifying model (4.4). Under the conditional model we specify the model for Y_i given the random effects b_i that is

$$Y_i|b_i \sim N(X_i\beta + Z_ib_i, \Sigma_i)$$

Thus

$$E(Y_i|b_i) = X_i\beta + Z_ib_i$$

and

$$Var(Y_i) = \Sigma_i.$$

In contrast under the marginal model specification

$$Y_i \sim N(X_i\beta, Z_iGZ_i' + \Sigma_i).$$

Thus here

$$E(Y_i) = X_i\beta \quad \text{and} \quad Var(Y_i) = Z_iGZ_i' + \Sigma_i = V_i.$$

It should immediately be noted that intrinsically the marginal model allows negative variance components provided V_i is positive semi-definite while in the conditional model negative variance components do not make sense. As stated earlier variance components generally refer to elements in G and Σ_i.

4.4 Estimation under the General Linear Mixed Effects Model

In this section the estimation problem for fixed effects, variance componets and random effects in the general linear model is addressed. First we discuss estimation in the marginal model where the relative merits of two likelihood estimation procedures namely the maximum likelihood (ML) and the restricted maximum likelihood (REML) will be discussed. Inference on fixed effects and variance components will be given attention distinguishing between non-boundary and boundary values in testing hypotheses about variance components. An outline on how to fit linear mixed models using statistical software with reference to SAS will briefly be outlined. Statistical tests about hypotheses concerning both fixed and random effects will be also be discussed. Although the current chapter is based on the normal distribution assumption on Y some of the ideas carry over to the non-Gaussian case (the focus in this work) in subsequent chapters. In particular Chapter 7 where the linera mixed model for non-Gaussian data is described.

4.4.1 Estimation under the Marginal Model

As stated above we consider the estimation for fixed parameters of the marginal model stated again below. The marginal model implied by equation (4.6) translates to

$$Y_i \sim N(X_i\beta, Z_iGZ_i' + \Sigma_i) \tag{4.7}$$

where we assume the fixed parameters are contained in the vector β and the variance component in G and Σ_i are contained in the vector α. One should note that inferences based on the marginal model does not explicitly assume the presence of random effects representing the natural heterogeneity between subjects. The interest here is more on the mean model $E(Y_i) = X_i\beta$ and inference on the parameters in β. However correct inference about the covariance parameters is necessary to ensure efficient inference about the mean model.

Let α denote the vector of all variance and covariance parameters (usually called variance components which are elements in the matrices G and Σ). Thus α consists of the $q(q+1)/2$ different elements in G and of all parameters in Σ. Let $\theta = (\beta', \alpha')'$ be the the s-dimensional vector of all parameters in the marginal model for Y_i. Suppose $\Theta = \Theta_\beta \times \Theta_\alpha$ denote the parameter space for θ, with Θ_β and Θ_α the parameter spaces for fixed effects and for the variance components respectively.

The estiamtes of β and α are obtained from maximising the marginal likelihood function given by

$$L_{ML}(\theta) = \prod_{i=1}^{N}\{(2\pi)^{n_i/2}|V_i(\alpha)|^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(Y_i - X_i\beta)'V_i^{-1}(\alpha)(Y_i - X_i\beta))\} \tag{4.8}$$

with respect to $\theta' = (\beta', \alpha')'$ over Θ.

Maximum Likelihood Estimation

The maximum likelihood estimate of β (MLE) obtained by maximising equation (4.8) is given by

$$\hat{\beta} = (\sum_{i=1}^{N} X_i'W_iX_i)^{-1}\sum_{i=1}^{N} X_i'W_iy_i \tag{4.9}$$

where $W_i = V_i^{-1}$, a result which was first derived by Laird and Ware (1982) when they first proposed the linear mixed model for longitudinal data. Note that the expression for $\hat{\beta}$ implicitly assumes α is known otherwise we need to replace α by its ML or REML estimate.

Restricted Maximum Likelihood Estimation (REML)

To develop the REML concept consider the simple case of a sample of N observations $Y_1, \ldots, Y_N$, from $N(\mu, \sigma^2)$. Given μ is known the MLE of σ^2 is

given by

$$\hat{\sigma}^2_{ML} = \sum_{i=1}^{N}(Y_i - \mu)^2/N \tag{4.10}$$

In this case $\hat{\sigma}^2_{ML}$ is an unbiased for σ^2. However if μ is unknown, the MLE of σ^2 is now given by

$$\hat{\sigma}^2_{ML} = \sum_{i=1}^{N}(Y_i - \overline{Y})^2/N \tag{4.11}$$

and now $\hat{\sigma}^2_{ML}$ is biased downwards for σ^2 because,

$$E(\hat{\sigma}^2_{ML} - \sigma^2) = (-N^{-1})\sigma^2, \text{ where } N,\ \sigma^2 > 0.$$

Thus $\hat{\sigma}^2_{ML}$ is an underestimate of σ^2. However we can note that the bias shrinks as the sample size increases that is as $N \to \infty$ this bias goes to zero asymptotically. The biased expectation leads to the conclusion that an unbiased estimate for σ^2 when μ is unknown should be

$$s^2 = \sum_{i=1}^{N}(Y_i - \overline{Y})^2/(N-1) \tag{4.12}$$

The above discussion shows that having to estimate μ introduces bias in the maximum likelihood estimation of σ^2. Thus one way to circumvent this problem is to find a way of estimating σ^2 without having to estimate μ first. This idea can be generated as follows. Note that all the data can be combined into one vector Y such that

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix} \sim \begin{pmatrix} \begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}, \sigma^2 I_N \end{pmatrix} \tag{4.13}$$

which means $Y \sim N(\mu 1_N, \sigma^2 I_N)$ where 1_N is a N-dimensional vector full of $1's$ and I_N the N-dim identity matrix. To avoid the estimation of μ we transform the vector of observations Y such that μ vanishes from the likelihood. Define

$$U = \begin{pmatrix} Y_1 - Y_2 \\ Y_2 - Y_3 \\ \vdots \\ Y_{N-1} - Y_N \end{pmatrix} = A'Y \sim N(0, \sigma^2 A'A) \tag{4.14}$$

Based on this transformation the MLE of σ^2 is exactly as given by s^2 thus unbiased for σ^2. The transformation operator A which is independent of s^2

defines a set of $N-1$ linearly independent error contrasts and s^2 is called the REML estimate of σ^2. The above formulation can be extended to the case of the linear regression model. To this extension consider a set of N observations $Y_1, Y_2, \ldots, Y_N$ from a normal linear regression model such that $Y \sim N(X\beta, \sigma^2 I_N)$ where Y is a N-dim vector of observations given $Y = (Y_1, Y_2, \ldots, Y_N)'$, X is the design matrix for the fixed vector β (vector of regression parameters) and σ^2 the residual variance. Following the above arguments the MLE of σ^2 for the linear regression model is

$$\hat{\sigma}^2_{ML} = (Y - X\hat{\beta})'(Y - X\hat{\beta})/N \tag{4.15}$$

while the REML estimate is given by

$$\hat{\sigma}^2_{REML} = (Y - X\hat{\beta})'(Y - X\hat{\beta})/(N-p) \tag{4.16}$$

where p is the number of parameters in β including the intercept if any. The REML estimate can also be obtained by transforming the data orthogonal to X to yield the vector U given by

$$U = A'Y \sim N(0, \sigma^2 A'A)$$

such that the estimate proceeds in terms of U and not Y. Note that X is an $N \times (p+1)$ matrix whose ith row is $(1, X_{i1}, \ldots, X_{ip})$ assuming the model involves p explanatory or predictor variables.

4.4.2 REML Estimation for longitudinal data model

We now show how the REML estimation works for the longitudinal data model. Let Y_i denote the individual n_i-dim vector of repeated observations from individual i that is $Y_i = (Y_{i1}, \ldots, Y_{in_i})'$ where it is assumed that $Y_i \sim N(X_i\beta, V_i)$. The strategy is first to combine the N individual specific information into one augmented vector Y such that $Y \sim N(X\beta, V)$ where

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}, X = \begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}, V(\alpha) = \begin{pmatrix} V_{11} & 0 & \ldots & 0 \\ 0 & V_{22} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & V_{NN} \end{pmatrix}. \tag{4.17}$$

Next the data are transformed orthogonal to X to $U = A'Y \sim N(, \sigma^2 A'A)$ where U is a vector of error contrasts defined earlier. The MLE of α, based on U is called the REML estimate and is denoted by $\hat{\alpha}_{REML}$. The resulting

estimate for β will be denoted by $\hat{\beta}_{REML}$. Alternatively $\hat{\alpha}_{REML}$ and $\hat{\beta}_{REML}$ can also be obtained from maximizing

$$L_{REML}(\theta) = |\sum_{i=1}^{N} X_i' W_i(\alpha) X_i|^{-\frac{1}{2}} L_{ML}(\theta) \tag{4.18}$$

or equivalently the log-likelihood

$$\ell_{REML}(\theta) = -\tfrac{1}{2}\ln|\textstyle\sum_{i=1}^{N} X_i' W_i(\alpha) X_i| + \ell_{ML}(\theta)$$

w.r.t. θ (i.e α and β simultaneously). Note that the expression above for $L_{REML}(\theta)$ is $L_{ML}(\theta)$ but subjected to a penalty. Importantly note that $L_{REMl}(\alpha, \hat{\beta}(\alpha))$ is the likelihood of the error contrasts, U, often called the $REML$ likelihood function. Thus $L_{REML}(\theta)$ is not a likelihood of the original data Y, but it is rather based on u. This has an implication when comparing the likelihood of two nested marginal models, because it means under REML they are not comparable as we will see later.

Fitting Linear Mixed Models Using a Statistical Software

A number of statistical software now have capability to fit linear mixed models with ease. These include SAS, GenStat, S-Plus and SPSS among others. In this thesis most of the analysis is carried out in SAS. Thus in this section we will discuss how one would fit the linear mixed models using SAS. For estimation of fixed effects and variance components one uses SAS Proc MIXED statement to primarily specify the data set and method of estimation. PROC MIXED has three options for the method of estimation. They are: ML (Maximum Likelihood), REML (Restricted or Residual maximum likelihood, which is the default method) and MIVQUE0 (Minimum Variance Quadratic Unbiased Estimation). ML and REML are based on a maximum likelihood estimation approach as discussed in subsection 4.4.1. The CLASS statement is used to declare categorical variables or factor variables in the data. The MODEL statement is used to specify the regression models or to state any other model relating the response to the fixed effects variables. This statement also has an option of whether to call or not to call for solutions and whether to fit a model with an intercept or not. To define the random effects (the intercepts and slopes) and their distributions one uses the RANDOM statement. This statement also has options used to specify which variable identifies the subjects, assuming independence across subjects and the form of the matrix G (random effects matrix). The option 'g' and 'gcorr' in the random statement are used when one wants to print the matrix G and the corresponding correlation matrix, option 'v' and 'vcorr is used to print the

matrix V_i and the corresponding correlation matrix. The REPEATED statement is used to first identify the factor variable used for ordering the repeated measurements within a subject e.g. 'time', 'age', 'birth order' in a family and so on. There is also an option within the REPEATED statement to specify which variable identifies the individual or subject, the type of residual covariance matrix Σ_i, option r and rcorr to print Σ_i and the corresponding correlation matrix. Most frequently used covariance structures available to the RANDOM and REPEATED statement are the unstructured (UN), simple independence (SIMPLE), compound symmetry (CS), first-order autoregressive (AR(1)), an so on. A more exhaustive list of possible covariance structures can be found in the books by Verbeke and Molenberghs (2000), Diggle et al.(2002), Molenberghs and Verbeke (2005) among others.

The following is a general form of PROC MIXED statement:
PROC MIXED options;
CLASS variable-list;
MODEL dependent=fixed effects/ options;
REPEATED repeated effects / options;
RANDOM random effects / options;
RUN;

The CONTRAST, ESTIMATE, LSMEANS and RANDOM statements can appear multiple times, all other statements can appear only once. The contrast statement is used if one is interested in testing the significance of a treatment, while the estimate statement is for a linear combination of effects. The random statement contains the list of random effects. The LSMEANS statement is used to request for least square means.

The MODEL statement must appear after the CLASS statement if CLASS statement is used.

4.4.3 Inference for the Marginal Model

In this section we briefly discuss inference for the estimated parameters in the marginal model both for the mean model (fixed effects) and the variance components. In particular the Wald tets, the t-test, the F-test, Robust inference and the LR test for fixed effects will be discussed. For variance components the methods that will be given attention are the Wald test and the LR test. The information criteria (IC) for making inference on non-nested models will also be discussed. Recall that the estimate for β is given by

$$\hat{\beta}(\alpha) = (\sum_{i=1}^{N} X_i' W_i X_i)^{-1} \sum_{i=1}^{N} X_i' W_i Y_i$$

with α replaced by its ML or REML estimate and Y_i is the subject-specific vector of observations. It follows that conditional on α, $\hat{\beta}(\alpha)$ is multivariate normal with mean β and covariance

$$\text{var}(\hat{\beta}) = (\sum_{i=1}^{N} X_i' W_i X_i)^{-1} (\sum_{i=1}^{N} X_i' W_i \text{var}(Y_i) W_i X_i) (\sum_{i=1}^{N} X_i' W_i X_i)^{-1} \qquad (4.19)$$

Using the standard result that if G is a matrix of constants and Y is a random vector valued random variable

$$var(GY) = GVar(Y)G'$$

where G' is the transpose of the matrix G of compatible dimension to the vector Y.

We note that given $\text{var}(Y_i) = V_i = W_i^{-1}$ holds then the expression for $\text{var}(\hat{\beta})$ reduces to

$$\text{var}(\hat{\beta}) = (\sum_{i=1}^{N} X_i' W_i X_i)^{-1} \qquad (4.20)$$

which is what gives rise to model based standard errors in contrast to empirical or robust standandard errors which will be discussed in section 4.4.5.

4.4.4 Approximate wald, t- and F-Tests

The Wald statistic is an alternative test which is commonly used to test the significance of individual regression coefficients for each independent variable (that is, to test the null hypothesis in a regression model that a particular coefficient is zero). Consider any known matrix of constants L and associated hypothesis $H_0 : L\beta = 0$ versus $H_a : L\beta \neq 0$. Then the Wald test statistic for testing such a hypothesis is given by

$$W_s = \hat{\beta} L' [L (\sum_{i=1}^{N} X_i' V_i^{-1} X_i)^{-1} L']^{-1} L \hat{\beta} \qquad (4.21)$$

which is asymptotically distributed as χ^2 with d.f. equal to rank(L) under H_0. Note that the Wald test is based on

$$\text{var}(\hat{\beta}) = (\sum_{i=1}^{N} X_i' W_i X_i)^{-1}.$$

However it should be noted that the variability introduced by replacing α by some estimate (ML or REML) is not taken into account in Wald tests. Robust and empirical standard errors are obtained by finding var($\hat{\beta}$) based on expression (4.19) with var(Y_i) replaced by the estimate from the data. This therefore implies that the Wald test will only provide valid inferences in sufficiently large samples. In practice this is often resolved by replacing the χ^2 distribution by an approximate F-distribution in a similar manner the standardized normal statistic z is replaced by the central t statistic when the variance is unknown and instead estimated by the sample variance s^2 with $n-1$ degrees of freedom. Thus in the above to test H_0 versus H_a the F-test statistic is given by

$$F_s = \frac{\hat{\beta}L'[L(\sum_{i=1}^{N} X_i'V_i^{-1}X_i)^{-1}L']^{-1}L\hat{\beta}}{\text{rank}(L)} \tag{4.22}$$

where approximate null-distribution of F_S is the F distribution with numerator degrees of freedom equal to rank (L) but the denominator degrees of freedom are to be estimated. There are several methods to calculate the denominator d.f. but the most frequently used ones are: the containment, Satterthwaite and the Kenward- Roger methods of approximation. In the context of longitudinal data nearly all methods lead to a large number of denominator degrees of freedom therefore leading to very similar p-values. For a univariate hypothesis rank(L)=1 and the F-test reduces to a t-test.

4.4.5 Robust Inference

Since

$$\hat{\beta}(\alpha) = (\sum_{i=1}^{N} X_i'W_iX_i)^{-1} \sum_{i=1}^{N} X_i'W_iY_i$$

with α replaced by its ML or REML estimate, it implies that $E[\hat{\beta}(\alpha)] = \beta$ provided $E(Y_i) = X_i\beta$. In other words in order for $\hat{\beta}$ to be unbiased it is sufficient that the mean of the response is correctly specified regardless of the assumed structure of V_i. Further still conditional on $\alpha, \hat{\beta}$ has covariance matrix given by

$$\text{var}(\hat{\beta}) = (\sum_{i=1}^{N} X_i'W_iX_i)^{-1} = C_N \tag{4.23}$$

provided var(Y_i) is correctly modelled as $V_i = Z_iGZ_i' + \Sigma_i$. The covariance estimate C_N is called the naive estimate. The so-called robust or sandwich estimate for var($\hat{\beta}$) which we denote as C_R does not require a correct specification of var(Y_i) rather it is obtained by replacing var(Y_i) by

$(Y_i - X_i\hat{\beta})(Y_i - X_i\hat{\beta})' = \tilde{V}_i \neq V_i$ in (4.19). It follows that

$$\text{var}(\hat{\beta}) = (\sum_{i=1}^{N} X_i' W_i X_i)^{-1} \sum_{i=1}^{N} X_i' W_i \tilde{V}_i W_i X_i (\sum_{i=1}^{N} X_i' W_i X_i)^{-1} = C_R$$

The robust variance estimate of $\text{Var}(\hat{\beta})$ is also called the sandwich estimator. Based on the sandwich estimate of $\text{Var}(\hat{\beta})$, robust versions of the Wald, t- and F- tests can be derived.

Note that the above analysis suggests that as long as interest is only on inference of the mean structure, little effort may be spent in modelling the actual covariance structure of Y, provided the data is sufficiently large. However this is not to say an appropriate covariance modelling is not of interest. An appropriate covariance structure may still be of interest for gaining efficiency in parameter estimation. In addition, in the presence of missing data, robust inference is only valid under very restrictive assumptions about the underlying missingnes process such as data be missing completely at random (MCAR).

4.4.6 The Likelihood Ratio Test

The likelihood ratio (LR) test is the most appropriate test to compare nested models with different mean structures, but with equal covariance structure. The general null hypothesis in this case is

$$H_0 : \beta \in \Theta_{\beta,0}; V_i = V_{oi} \quad \text{versus} \quad H_a : \beta \in \Theta_\beta; V_i = V_{oi} \tag{4.24}$$

The second part of the statement of each hypothesis is to emphasize that the covariance structure of the data is the same in both cases. As before we let L_{ML} denote the ML function and let the ML estimate under H_0 and H_a of θ be $\hat{\theta}_{ML,0}$ and $\hat{\theta}_{ML}$ respectively. Then the likelihood ratio test statistic is given by

$$T_{\beta,LR} = -2ln\lambda_N = -2ln[\frac{L_{ML}(\hat{\theta}_{ML,0})}{L_{ML}(\hat{\theta}_{ML})}] \tag{4.25}$$

where

$$\lambda_N = \frac{L_{ML}(\hat{\theta}_{ML,0})}{L_{ML}(\hat{\theta}_{ML})} \tag{4.26}$$

is the ratio of the likelihoods under H_0 and H_a. Values of λ_N close to 1 indicate that H_0 is true while values near 0 are in support of the alternative hypothesis rather than H_0. Thus when H_0 is not true $T_{\beta,LR}$ will be large and postive indicating evidence against H_0.

Note that the LR tests are not valid under REML because here the response vector Y is first transformed into error contrasts $U = A'Y$, for some matrix of constants A such that $A'X = 0$. Then ML estimation is done on U as the data. Thus the likelihood value $L_{REML}(\hat{\theta})$ which is the likelihood at the maximum based on the error contrasts U which are different for different mean models under the null and alternative hypothesis. Thus because two models with different mean structures lead to different REML error contrasts it follows that the subsequent likelihoods are not comparable leading to a breakdown of the apllicabilty of the LR test. Thus the LR test is valid only under ML estimation because both the numerator and denominator are based on the same data Y.

4.5 Inference for the variance Components

Most often it is the inference for the mean structure that is usually of primary interest. However, inferences for the covariance structure could be of interest as well for obvious reasons among them the interpretation of the random variation in the data. A test for a variance component also helps in establishing whether we really do need the inclusion of the random effects or not. It is also important to note that an over-parameterized covariance structure (e.g. the UN structure) may lead to inefficient inferences for the mean model (due to overspending of degrees of freedom in estimating the variance-covariance components). On the other hand a too restrictive covariance model will invalidate inferences for the mean structure. The best covariance model is therefore a balance between a fully unstructured model and the independence assumption.

4.5.1 Approximate Wald Test

Asymptotically, ML and REML estimates of α are normally distributed with correct mean and inverse Fisher information matrix as covariance. Therefore approximate s.e.'s and Wald tests can easily be obtained. However there is need for caution in the context of the hierarchical model in relation to the marginal model interpretation. A null hypothesis of a zero variance component is meaningful only under the marginal model that is when no underlying random effects structure is believed to describe the data. The quality of the normal approximation for $\hat{\alpha}_{ML}$ and $\hat{\alpha}_{REML}$ estimates strongly depends on the true value of α. The approximation is poor once α is relatively close to the boundary of the parameter space. If α is a boundary value, the

normality approximation fails completely. Under the hierarchical normal interpretation a null hypothesis of a zero variance component implies the p-value is based on an incorrect null distribution for the Wald test statistic. The test is only correct, when the null hypothesis is not a boundary value. However even under the hierarchical model interpretation a Wald test is valid for testing a zero covariance parameter such as $g_{12} = 0$ versus $g_{12} > 0$ testing dependence between two random effects since a zero covariance is admissible between any two random variables.

4.5.2 The Likelihood Ratio Test

The LR test is here meant to compare two models with the same mean structure but different variance and covariance parameter structures. The null hypothesis of interest is similar to that of the mean structure, namely

$$H_0 : \alpha = \Theta_{\alpha,0} \quad \text{versus} \quad H_a : \alpha \in \Theta_\alpha \tag{4.27}$$

where $\Theta_{\alpha,0} \subset \Theta_\alpha$. Let $\hat{\alpha}_{ML,0}$ and $\hat{\alpha}_{ML,0}$ be the MLEs under H_0 and H_a. Then the LR test statistic is given by

$$T_\alpha = -2ln\lambda_N = -2ln\left[\frac{L_{ML}(\hat{\alpha}_{ML,0})}{L_{ML}(\hat{\alpha}_{ML})}\right] \tag{4.28}$$

The asymptotic null distribution of T_α is χ^2 with d.f. equal to the difference in dimension of Θ_α and $\Theta_{\alpha,0}$. Not that now as long as the comparison is under the same mean structure, a valid LR test can still be obtained under REML since the error contrasts U are the same in both cases, namely under H_0 and H_a.

4.6 Inference for the Random Effects

In this section the problem of making inference on the random effects b_i is addressed briefly. In particular the idea of empirical Bayes (EB) and best linear unbiased predictors will be given attention (BLUP). The concept of shrinkage estimators will be derived and the normality assumption for random effects discussed. The random intercepts and slopes model will be used as a special case.

4.6.1 Empirical Bayes (EB) Inference

Consider the linear mixed model

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i$$

where $b_i \sim N(0, G)$, $\varepsilon_i \sim N(0, \Sigma_i)$ and that b_i and ε_i are independent stated in equation (4.5). The random effects b_i reflect how the evolution for the ith subject deviates from the expected evolution $X_i\beta$. Estimation of the random effects b_i is helpful for detecting outlying profiles from the expected profile. Thus inference on random effects is only meaningful under the hierarchical model assumptions where

$$Y_i|b_i \sim N(X_i\beta + Z_i b_i, \Sigma_i) \text{ and } b_i \sim N(0, G)$$

implying that

$$E(Y_i|b_i) = X_i\beta + Z_i b_i.$$

Since the b_i here behave like random 'parameters' it is most natural to consider Bayesian approaches where the prior distribution of the random parameters (here random effects) is $b_i \sim N(0, G)$. Thus using the Bayes rule we can express the posterior distribution of the b_i given the data $Y_i = y_i$ as

$$f(b_i|y_i) = \frac{f(y_i|b_i)f(b_i)}{\int f(y_i|b_i)f(b_i)db_i} \tag{4.29}$$

Since we know the marginal distribution of b_i and the conditional distribution $Y_i|b_i$ we can after some algebraic manipulation show that the posterior distribution of b_i is given by

$$b_i|y_i \sim N(GZ_i'W_i(y_i - X_i\beta), \Lambda_i) \tag{4.30}$$

for some positive definite matrix Λ_i. Thus we can use the posterior mean of b_i as an estimate of b_i that is

$$\hat{b}_i(\theta) = E(b_i|Y_i = y_i) = \int f_i(b_i|y_i)db_i = GZ_i'W_i(\alpha)(y_i - X_i\hat{\beta} \tag{4.31}$$

and the variance of this estimate is given by

$$\text{var}(\hat{b}_i(\theta)) = GZ_i'W_i - W_iX_i(\sum X_i'W_iX_i)^{-1}X_i'W_iZ_iG \tag{4.32}$$

However inference on b_i aught to take into account the variability in b_i therefore inference for b_i is usually based on

$$\text{var}(\hat{b}_i(\theta) - b_i) = G - \text{var}(\hat{b}_i(\theta)) \tag{4.33}$$

Thus for inference purposes once the corrected variance in equation (4.33) is found Wald tests can be constructed to test hypotheses about $b_i(\theta)$. Parameters in θ are replaced by their ML or $REML$ estimates, obtained after fitting the marginal model. The estimate $\hat{b}_i = \hat{b}_i(\theta)$ is called the empirical Bayes estimate of b_i. Approximate t and F tests to account for the variability introduced by replacing θ by $\hat{\theta}$ can be constructed similar to tests for fixed effects.

4.6.2 Best Linear Unbiased Prediction (BLUP)

Often, parameters of interest are linear combinations of fixed effects in β and random effects in b_i. For example a subject specific slope is the sum of the average slope for subjects with same covariate values, and the subject-specific random slope for that subject. In general such a linear combination will be of the form

$$\gamma = \ell'_\beta \beta + \ell'_b b_i \tag{4.34}$$

is the quantity of interest. Then conditionally on α,

$$\hat{\gamma} = \hat{\ell}'_\beta \hat{\beta} + \hat{\ell}'_b \hat{b}_i \tag{4.35}$$

is the best linear unbiased predictor of γ. Note that $\hat{\gamma}$ is linear in the observations Y_i, unbiased and minimizes the variance among all unbiased linear estimators. In SAS estimates of random effects are obtained by adding the 'solution' option to the RANDOM statement. However in practice one is also interested in properties of these estimates and as a starting point histograms and scatter plots of certain components of $\hat{b}_i$ can be used to visually detect subjects with exceptional or extreme evolutions over time. Note that the predicted evolution of the ith subject is given by

$$\begin{aligned}
\hat{Y}_i &= X_i\hat{\beta} + Z_i\hat{b}_i \\
&= X_i\hat{\beta} + Z_i G Z_i' V_i^{-1}(y_i - X_i\hat{\beta}) \\
&= (I_{n_i} - Z_i G Z_i' V_i^{-1}) X_i\hat{\beta} + Z_i G Z_i' V_i^{-1} y_i \\
&= \Sigma_i V_i^{-1} X_i\hat{\beta} + (I_{n_i} - \Sigma_i V_i^{-1}) y_i
\end{aligned} \tag{4.36}$$

which is a weighetd average of the population-averaged profile $X_i\hat{\beta}$ and the observed individual data y_i, with weights $\hat{\Sigma}_i V_i^{-1}$ and $I_{n_i} - \hat{\Sigma}_i V_i^{-1}$ respectively. Note that $X_i\hat{\beta}$ gets more weight if the residual variability is large compared to the total variability given by $V_i = Z_i G Z_i^{'} + \Sigma_i$. This phenomenon is the so called shrinkage effect meaning that the observed data are shrunk towards the prior average profile $X_i\beta$ depending on the degree of how much within individual variability there is. This is also reflected in the fact that for any linear combination $\ell' b_i$ of random effects

$$\text{var}(\ell'\hat{b}_i) \leq \text{var}(\ell' b_i) \tag{4.37}$$

We now consider a simple example to demonstrate some of the concepts raised above for purposes of clarity. Consider the random intercepts model given by

$$y_{ij} = \beta_0 + b_{0i} + \beta_1 t_{ij} + \varepsilon_{ij} \tag{4.38}$$

where y_{ij} is the jth observation from the ith individual in the study for $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$, β_0 is the average intercept, b_{0i} is the subject specific intercept which is a random effect assumed to be distributed as $N(0, g_0^2)$, β_1 is the common average slope for all individuals which assumed not to be affected by individual to individual variability, t_{ij} is the actual measurement occasion time and ε_{ij} is the measurement error or residual. Following the above model derivations it follows that the EB estimate for the random intercept b_{0i} is given by

$$\begin{aligned}
\hat{b}_{0i} &= GZ_i' W_i(\alpha)(y_i - X_i\beta) \\
&= g_0^2 1_{n_i}'(\sigma^2 1_{n_i} 1_{n_i}' + \sigma^2 I_{n_i})^{-1}(y_i - X_i\beta) \\
&= \tfrac{g_0^2}{\sigma^2} 1_{n_i}'(I_{n_i} - \tfrac{g_0^2}{\sigma^2 + n_i g_0^2} 1_{n_i} 1_{n_i}')(y_i - X_i\beta) \\
&= \tfrac{n_i g_0^2}{\sigma^2 + n_i g_0^2} \tfrac{1}{n_i} \textstyle\sum_{j=1}^{n_i} (y_{ij} - X_i^{[j]}\beta)
\end{aligned} \tag{4.39}$$

Then we note that $\hat{b}_{0i}$ is a weighted average of 0 (the prior mean) and the average residual for individual i. Thus the larger n_i is the less the shrinkage effect. Likewise the smaller σ^2 relative to g_0^2 the lesser the shrinkage.

4.6.3 Modelling fitting and choice

A more inclusive model for longitudinal data from a continuous response assumed to be from a Gaussian distribution is given by

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_{i(1)} + \varepsilon_{i(2)} \tag{4.40}$$

where now the residual term is split into two components $\varepsilon_{i(1)}$ and $\varepsilon_{i(2)}$ to account for measurement error such that $\varepsilon_{i(1)} \sim \mathrm{N}(0, \sigma^2 I_{n_i})$ and serial correlation for example AR(1) such that $\varepsilon_{i(2)} \sim \mathrm{N}(0, \tau^2 H_i)$. Thus the probability marginal model of the general linear mixed model for Y_i can be written as

$$Y_i \sim N(X_i\beta, Z_i G Z_i' + \sigma^2 I_{n_i} + \tau^2 H_i) \tag{4.41}$$

while conditionally

$$Y_i \sim N(X_i\beta, Z_i G Z_i' + \sigma^2 I_{n_i} + \tau^2 H_i). \tag{4.42}$$

The structure of the matrix H_i can take several forms depending on the serial correlation structure envisaged in the data. A detailed account about the selection of an appropriate structure for H_i can be found in Diggle et al. (2002) with examples.

Chapter 5

Generalized Linear Models for Cross sectional data

5.1 Introduction

The gaussian regression model discussed in section (4.2) assumed that the error terms where normally distributed of which this is not always the case. Generalized linear models (GLMs) were formulated as a way of unifying various other statistical models, including linear regression (gaussian regression model), logistic regression and Poisson regression, under one framework (Nelder and Wedderburn, 1972). A more intensive treatment of GLMs is given by McCullagh and Nelder (1989). This allowed them to develop a general algorithm for maximum likelihood estimation in all these models. It extends naturally to encompass many other models as well. To make use of the algorithm they had to assume that all of the models have distributions in the exponential family. Specifically the algorithm for fitting GLMs used in most statistical packages such as SAS and Genstat is the iterative weighted least squares (IWLS). Applications of GLMs and their extension include the work by Diggle et al. (2002) and Molenberghs and Verbeke (2005) in the context of longitudinal data analysis.

5.2 The Exponential Family

Consider n independent observations each from a distribution in the exponential family with probability density function

$$f(y_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}$$

for some specific functions a(.), b(.) and c(.,.). The parameters θ_i and ϕ are essentially location and scale parameters, respectively. It will be shown in the text that follows that if Y_i has a distribution in the exponential family then it has mean and variance $\mathrm{E}(Y_i) = \mu_i = b'(\theta_i)$and $\mathrm{var}(Y_i) = \sigma_i^2 = b''(\theta_i)a(\phi)$, where $b'(\theta_i)$ and $b''(\theta_i)$ are the first and second derivatives of $b(\theta_i)$. Since the mean depends only on θ_i, in a standard GLM the term $c(y_i, \phi)$ can be left unspecified without affecting the likelihood-based estimation of regression parameters. The exponential family just defined includes as special cases the normal, binomial, Poisson, exponential, gamma and inverse Gaussian distributions. From this common class of distributions McCullagh and Nelder (1989) were able to generalize the Gaussian Linear model to the generalized linear model (GLM). The reason for this is that the family includes several important distributions, and also has 'good' properties which are discussed below. Without loss of generality let $a(\phi) = \phi$ where ϕ is known as the dispersion parameter.

Using the property that $\int f(y|\theta, \phi)dy = 1$ the mean and variance can easily be derived by taking the first and second-order derivatives of the integral with respect to θ. Thus one can easily find the mean and variance to be

$$E(Y) = \mu_{ij} = b^{'}(\theta)$$
$$Var(Y) = b^{''}(\theta)\phi$$

where $b^{'}(\theta)$ and $b^{''}(\theta)$ are the first and second order derivatives of $b(\theta)$ with respect to θ. Thus under the exponential family of distributions the mean and variance are both determined by the function $b(.)$.The reason for restricting GLMs to the exponential family of distributions for Y is that the algorithm applies to the entire family, for any choice of link function. For example suppose that $Y \sim Bin(n, p)$ then

$$f(y|p) = \left(\begin{array}{c} n \\ y \end{array} \right) p^y (1-p)^{n-y} = \exp\left\{ y \ln\left(\frac{p}{1-p}\right) + n\ln(1-p) + \ln\left(\begin{array}{c} n \\ y \end{array} \right) \right\}$$

In this case

$$\theta = \ln \frac{p}{1-p}$$

which means

$$p = \frac{e^\theta}{1+e^\theta}$$

thus the link function is

$$\ln\left(\frac{p}{1-p}\right) = \mathrm{logit}(p) \quad \text{and} \quad b(\theta) = n(1+e^\theta).$$

Therefore

$$b'(\theta) = \frac{ne^{\theta}}{(1+e^{\theta})}$$

and

$$b''(\theta) = \frac{ne^{\theta}}{(1+e^{\theta})^2} = np(1-p)$$

thus in this case $\phi = 1$ and $c(y, \phi) = \ln \begin{pmatrix} n \\ y \end{pmatrix}$.

5.3 The Generalized Linear model

The generalized linear model (GLM) is a flexible generalization of ordinary least squares regression. It relates the random distribution of the measured variable of the experiment (the distribution function) to the systematic (non-random) portion of the experiment (the linear predictor) through a function called the link function.

Suppose we have $Y_1, Y_2, \ldots, Y_N$ independent response observations with mean $\mu_1, \mu_2, \ldots, \mu_N$, respectively. Further suppose the observation Y_i has a distribution that is a member of the exponential family. The basic idea of a GLM is to develop a linear model for the appropriate function of the expected value of the response variable. In order to specify the GLM let η_i denote the linear predictor relating $\mathrm{E}(Y_i)$ to the predictor variable or covariates. Then the relationship can be written as

$$\eta_i = g(E(Y_i)) = g(\mu_i)$$

where we allow the linear predictor to be a monotone function of the mean. The function g is called the link function. The term 'link' is derived from the fact that the function is the link between the mean and the linear predictor. We assume that the link function is a monotonic differentiable function. In the sense that

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

where the x_{ik}'s are the values of the p predictor variables. The parameter $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ is the vector of the unknown regression parameters. The goal is to estimate these unknown parameters, including making inferences about them. Note that β_j measures the change in Y (on the link function scale) for every unit increase in X_j given all other predictor variables are held constant. If $\eta_i = \theta_i$, the canonical parameter, we have the canonical link.

The three fundamental components of a GLM are (i)the response distribution, (ii)the link function and (iii) a set of parameters β and covariates X. We can view the selection of the link function as being similar to the choice of a transformation on the response. However it is the population mean, not the data that is transformed hence the significance of the pioneering work by Nelder and Wedderburn (1972). The link function takes advantage of the natural distribution of the response. If $g(\mu_i) = \theta_i(\mu_i)$ then we have what is known as a canonical link function. The canonical link function simplifies computations with GLMs greatly which otherwise can be complex. For example in the case of the Binomial GLM the canonical link function is

$$g(\mu_i) = \ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{\mu_i}{n_i - \mu_i}\right).$$

Thus in this case $\theta_i = \ln(\frac{p}{1-p})$ and in general the Binomial GLM is

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

5.4 Estimation - Maximum Likelihood

The log-likelihood function of N independent observations from a distribution under the exponential family is

$$\begin{aligned} \ell &= \sum_{i=1}^{N} lnf(y_i, \theta_i, \phi_i) \\ &= \sum_{i=1}^{N} \{\tfrac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\} \end{aligned} \tag{5.1}$$

where $\theta_i = \theta_i(\mu_i)$ a function of the model mean and $\mu_i = \mu_i(\beta)$.

Our main aim is to estimate β, and the first step as is in any maximum

likelihood estimation is to differentiate ℓ with respect to β so

$$\begin{aligned}
\frac{\partial \ell}{\partial \beta} &= \frac{1}{a(\phi)} \sum_{i=1}^{N} [y_i \frac{\partial \theta_i}{\partial \beta} - \frac{\partial b(\theta_i)}{\partial \beta}] \\
&= \frac{1}{a(\phi)} \sum_{i=1}^{N} [y_i \frac{\partial \theta_i}{\partial \beta} - b'(\theta_i) \frac{\partial b(\theta_i)}{\partial \beta}] \\
&= \frac{1}{a(\phi)} \sum_{i=1}^{N} \frac{\partial \theta_i}{\partial \beta} (y_i - \mu_i) \\
&= \frac{1}{a(\phi)} \sum_{i=1}^{N} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta} (y_i - \mu_i) \\
&= \frac{1}{a(\phi)} \sum_{i=1}^{N} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial g(\mu_i)} \frac{\partial g(\mu_i)}{\partial \beta} (y_i - \mu_i) \\
&= \frac{1}{a(\phi)} \sum_{i=1}^{N} \frac{\partial \theta_i}{\partial \mu_i} [g'(\mu_i)]^{-1} \frac{\partial g(\mu_i)}{\partial \beta} (y_i - \mu_i)
\end{aligned} \tag{5.2}$$

But we know that

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{Var(y_i)}{a(\phi)} = v(\mu_i)$$

which implies that

$$\frac{\partial \theta_i}{\partial \mu_i} = [v(\mu_i)]^{-1}$$

where $v(\mu_i)$ is called the variance function. Let $w_i = [v(\mu_i)[g'(\mu_i)]^2]^{-1}$. From this we deduce that:

$$\begin{aligned}
&\frac{\partial \ell}{\partial \beta} = \frac{1}{a(\phi)} \sum_{i=1}^{N} w_i g'(\mu_i) \frac{\partial g(\mu_i)}{\partial \beta}(y_i - \mu_i) \\
&= \frac{1}{a(\phi)} \sum_{i=1}^{N} \frac{\partial g(\mu_i)}{\partial \beta} w_i g'(\mu_i)(y_i - \mu_i) \\
&= \frac{1}{a(\phi)} \sum_{i=1}^{N} X_i w_i g'(\mu_i)(y_i - \mu_i) \\
&\text{(since} \quad \frac{\partial g(\mu_i)}{\partial \beta} = \frac{\partial X_i' \beta}{\partial \beta} = X_i) \\
&= \frac{1}{a(\phi)}[X_1 w_1 g'(\mu_1)(y_1 - \mu_1) + \ldots, X_N w_N g'(\mu_N)(y_N - \mu_N)] \\
&\Rightarrow \frac{\partial l}{\partial \beta} = \frac{1}{a(\phi)} \begin{pmatrix} 1 & 1 & \ldots & 1 \\ X_{11} & X_{21} & \ldots & X_{N1} \\ \vdots & \vdots & \vdots & \vdots \\ X_{1p} & X_{2p} & \ldots & X_{Np} \end{pmatrix} \begin{pmatrix} w_1 g'(\mu_1)(y_1 - \mu_1) \\ w_1 g'(\mu_2)(y_2 - \mu_2) \\ \vdots \\ w_N g'(\mu_N)(y_N - \mu_N)] \end{pmatrix} \\
&= \frac{1}{a(\phi)} X' W \triangle (y - \mu)
\end{aligned} \tag{5.3}$$

where,
W= diag (w_i)
$\triangle$= diag$(g'(\mu_i))$
$y = [y_1, \ldots, y_N]'$
$\mu = [\mu_1, \ldots, \mu_N]'$

Therefore the ML estimating equations are given by equating the score equation $U = \frac{\partial \ell}{\partial \beta}$ to zero that is

$$U = \frac{\partial \ell}{\partial \beta} = \frac{1}{a(\phi)} X' W \triangle (y - \mu) = 0 \tag{5.4}$$

and by further rearrangement of the terms the following identity is obtained namely

$$X' W \triangle y = X' W \triangle \mu \tag{5.5}$$

where $W, \triangle$ and μ are functions of β. Typically this is a non-linear equation which needs to be solved numerically. To estimate $\hat{\beta}$ the score statistic need to be solved using iterative methods such as the Newton Raphson's method, Fisher scoring method or Iterative Reweighted least squares (IRWLS). These methods are briefly discussed in the next section.

5.4.1 Numerical techniques

Before we discuss the numerical techniques we first define the information matrix. The likelihood function for a GLM also determines the asymptotic covariance matrix of the ML estimator $\hat{\beta}$. This matrix is the inverse of the information matrix J. The information matrix J is given by

$$J = -E\Big(\frac{\partial^2 \ell}{\partial \beta^2}\Big) \tag{5.6}$$

From equation (5.4) we know that

$$\tfrac{\partial \ell}{\partial \beta} = \tfrac{1}{a(\phi)} X'W \triangle (y - \mu)$$

and by differentiating the above equation with respect to β once more yields

$$\frac{\partial^2 \ell}{\partial \beta^2} = -\frac{1}{a(\phi)} X'WX + \frac{1}{a(\phi)} \frac{\partial (W\triangle)}{\partial \beta}(y - \mu) \tag{5.7}$$

Substituting equation (5.7) into equation (5.6) we have

$$\begin{aligned} J &= -E(\tfrac{\partial^2 \ell}{\partial \beta^2}) \\ &= \tfrac{1}{a(\phi)} X'WX - \tfrac{1}{a(\phi)} \tfrac{\partial (W\triangle)}{\partial \beta} E(y - \mu) \qquad (5.8) \\ &= \tfrac{1}{a(\phi)} X'WX \end{aligned}$$

Thus we have

$$J = \frac{1}{a(\phi)} X'WX$$

since $\text{var}(\hat{\beta}) = J^{-1}$ it implies that

$$\text{Var}(\hat{\beta}) = J^{-1} = a(\phi)(X'WX)^{-1}.$$

Three most commonly used methods to estimate β are briefly presented below.

Newton Raphson method

In general the Newton-Raphson method of finding the root of the equation $f(x) = 0$ is given by

$$x^{t+1} = x^t - [f'(x^t)]^{-1} f(x^t) \tag{5.9}$$

where, x^t is the current known x-value, $f(x^t)$ represents the value of the function at x^t, and $f'(x^t)$ is the derivative (slope) at x^t. x^{t+1} represents the

updated value of x and the process proceeds in this manner until convergence. Now in our case the objective function of interest is the score statistic $U(\beta)$ thus on the (t+1)th iteration, the algorithm updates the parameter vector β by

$$\beta^{(t+1)} = \beta^{(t)} - [H^{(t)}]^{-1}U^{(t)} \tag{5.10}$$

where $\beta^{(t+1)}$ is the updated solution given $\beta^{(t)}$. H is the hessian matrix and is given by,

$$H = \frac{\partial^2 \ell}{\partial \beta^2}$$

The iteration will be stopped when the parameter estimate does not change significantly anymore, that is when $|\beta^{(t+1)} - \beta^{(t)}| < \varepsilon$ where ε is called the tolerance or convergence limit. We denote the final parameter estimate by $\hat{\beta}$. Thus the Newton-Raphson method is a purely mathematical procedure using the required score statistic as the objective function.

Fisher scoring method

The method of Fisher scoring is a variant of the Newton-Raphson that replaces the Hessian by its expectation with respect to the observations Y_i:

$$\begin{aligned} \beta^{(t+1)} &= \beta^{(t)} - [E(H^{(t)})]^{-1}U^{(t)} \\ &= \beta^{(t)} + J^{(t)}U^{(t)} \end{aligned} \tag{5.11}$$

Since $E(H) = -J$ where J is the Fisher information matrix.

Iterative Reweighted least squares (IRWLS)

According to [1] the method of iteratively re-weighted least squares (IRLS) is a numerical algorithm for minimizing any specified objective function using a standard weighted least squares method such as Gaussian elimination.

Suppose we set

$$Z = X\beta + \varepsilon \tag{5.12}$$

where $\varepsilon = \triangle(y - \mu)$ is the vector of residuals or error terms. It can be shown that

$$E(\varepsilon) = 0, \quad \text{Var}(\varepsilon) = \triangle a(\phi) v(\mu) \triangle,$$

where v(μ)=diag(v(μ_i)) and $\triangle = g'(\mu_i)$ as earlier defined. Therefore Var(ε)=$a(\phi)W^{-1}$ and if we apply the general least squares method to equation (5.12) we get:

$$\begin{aligned}\hat{\beta} &= (X'WX)^{-1}X'WZ \\ &= (X'WX)^{-1}X'W(X\beta + \triangle(y-\mu)) \\ &= \beta + (X'WX)^{-1}X'W \triangle (y-\mu))\end{aligned} \tag{5.13}$$

The second step in equation (5.13) is obtained by substituting equation (5.12). Thus the IRWLS is given by

$$\beta^{(t+1)} = \beta^{(t)} + (X'W^{(t)}X)^{-1}X'W^{(t)} \triangle (y-\mu^{(t)}) \tag{5.14}$$

With these new weights W, the weighted least squares equation is re-solved and the residuals are re-calculated. The process is then iterated many times until convergence is achieved.

5.5 Inference

Our primary interest is to test the following general hypothesis about the vector of parameters β:

$$H_0 : L\beta = 0 \text{ vs } H_a : L\beta \neq 0$$

Since $\hat{\beta}$ is the MLE of β it follows that $\hat{L\beta}$ is the MLE of $L\beta$. Furthermore,

$$\hat{L\beta} \sim \text{N(L}\beta\text{, L Var}(\hat{\beta})\text{ L}')$$

where L is a matrix of known constants of dimension say $r \times p$. Three commonly used statistics for inference are the Wald test, score test and the likelihood ratio test.

5.5.1 Wald Test

The Wald statistic is a test statistic which is commonly used to test the significance about the regression coefficients for each independent variable or a linear combination involving a subject of them. To test the general hypothesis

$$H_0 : L\beta = 0 \text{ versus } H_a : L\beta \neq 0$$

the Wald test statistic for testing such a hypothesis is given by

$$W_s = (L\hat{\beta} - L\beta)'[LVar(\hat{\beta}L']^{-1}(L\hat{\beta} - L\beta) \tag{5.15}$$

which under H_0 is asymptotically distributed as χ^2 with d.f. equal to rank(L). Note that the Wald test is based on

$$\text{Var}(\hat{\beta}) = (\sum_{i=1}^{N} X_i' W_i X_i)^{-1}$$

where W_i=diag(w_i), $w_i = [v(\mu_i)[g'(\mu_i)]^2]^{-1}$ and $X_i^{'} = (1, X_{i1}, X_{i2}, \ldots, X_{ip})$. Thus to test the hypothesis that

$$H_0 : \beta_j = 0 \text{versus} H_a : beta_j \neq 0$$

we let L take the form of a row vector with all components equal to zero except the $(j+1)$ position which takes value 1 assuming $\beta = (1, \beta_1, \ldots, \beta_2)$.

5.5.2 Likelihood Ratio Test

The likelihood ratio test is a widely used procedure for testing hypotheses involving nested models. It rejects the null hypothesis when the maximum likelihood under the null hypothesis is significantly smaller than the maximum likelihood under the alternative hypothesis. In some situations, its p-value can be calculated exactly, but, in general, the p-value must be approximated, usually by using a chi-square approximation.

Suppose that we are interested in comparing two nested models, a full model (which will be denoted as L_1) and a reduced model (which will be denoted as L_0). Suppose that one model (the reduced model) is a special case of the other (the full model). That is, the reduced model is simpler than the full model, so that when the reduced model holds the full model must necessarily hold. The reduced model is then said to be nested within the full model. We can compare the two nested models by comparing their maximized log-likelihoods, say ℓ_0 and ℓ_1. The former is at least as large as the latter i.e $\ell_0 \leq \ell_1$. The larger the difference between L_0 and L_1 the stronger the evidence that the reduced model is inadequate or inappropriate.

In general $L_1 \geq L_0$ since L_0 results from maximising over the restricted set of values thus we have $\bigwedge = \frac{L_0}{L_1} \leq 1$. As the sample size increases -2ln$\bigwedge$ becomes approximately χ^2 with degrees of freedom equal to the difference in the parameters under H_0 and $H_0 \cup H_1$. Formally, this test is called the likelihood ratio test.

The Score statistic

The score U of a likelihood function is the first derivative of ℓ with respect to the model parameters, given by

$$U = \frac{1}{a(\phi)} X'W \triangle (Y - \mu).$$

The expected value and variance of U can be calculated as follows:

$$\begin{aligned} E[U] &= \tfrac{1}{a(\phi)} X'W \triangle E(Y - \mu) = 0 \\ Var(U) &= [\tfrac{1}{a(\phi)}]^2 X'W\triangle \ \mathrm{Var}(Y - \mu)\triangle WX \\ &= [\tfrac{1}{a(\phi)}]^2 X'W\triangle \ \mathrm{Var}(Y)\triangle WX \\ &= [\tfrac{1}{a(\phi)}]^2 X'W \triangle (a(\phi)W^{-1}) \triangle WX \\ &= \tfrac{1}{a(\phi)} X'WX \\ &= J \end{aligned} \tag{5.16}$$

where J is the information matrix. Since $U \sim \mathrm{N}(0, J)$ it follows that $U'J^{-1}U \sim \chi^2_{rank(J)}$.

The main interest is to test the hypothesis:

$$H_0 : \beta = \beta_0 \text{ vs } H_1 : \beta \neq \beta_0$$

thus under H_0 we have

$$U'(\beta_0)[J(\beta_0)]^{-1}U(\beta_0) \sim \chi^2_{rank(J)}.$$

Thus if $T > \chi^{2(\alpha)}_{rank(J)}$, where α is a chosen significance level then $H_0 is rejected.$

When the interest is

$$H_0 : \beta_j = \beta_{0j} \text{ vs } H_1 : \beta_j \neq \beta_{0j}$$

under H_0 we have

$$U'_j(\beta_{0j})[J_{jj}]^{-1}U_j(\beta_{0j}) \sim \chi^2_{rank(1)},$$

that is

$$\frac{[U_j(\beta_{0j})]^2}{J_{jj}} \sim \chi^2_{rank(1)},$$

where $U_j(\beta_{0j})$ is the jth element of U and J_{jj} is the j^{th} diagonal element of the information matrix (J). Thus this implies that

$$\frac{U_j(\beta_{0j})}{\sqrt{J_{jj}}} \sim N(0,1).$$

However it can easily be shown that as N $\longrightarrow \infty$ the Wald test, likelihood ratio test and the score test, have certain asymptotic equivalences. For small to moderate sample sizes, the likelihood ratio test is usually more reliable than the Wald test.

5.6 Adequacy of the GLM Model

In this section we are going to discuss how one can check the adequacy of a model. Usually one wants to know how well a particular GLM describes a set of data. To answer this question let $\ell(\mu, y)$ denote the log likelihood function expressed in terms of the mean $\mu = [\mu_1, \mu_2, \ldots, \mu_N]$ and $\ell(\hat{\mu}, y)$ denote the maximized log-likelihood for the model $g(\mu) = X\beta$. For all possible models, the maximum achievable log-likelihood is $\ell(y, y)$. This occurs when we fit a separate parameter for each observation and the perfect fit is $\hat{\mu} = y$. Such a model is called the saturated model. This model is not useful, since it does not provide any parameter reduction. However, it serves as a baseline for comparison with other model fits.

5.6.1 The deviance

The deviance function is very useful for comparing two models when one model has parameters that are a subset of the second model. The deviance denoted by D^* is the log-likelihood statistic for testing new models against the saturated model and it is expressed as

$$D^* = 2(\ell(y, y) - \ell(\hat{\mu}, y))$$

This has an asymptotic χ^2 distribution with degrees of freedom $N-k$, where k is number of parameters in the reduced model. We use the deviance for model checking and for inferential comparison of models. To show how the deviance is used to compare two nested models let us suppose that D_0 is the

deviance resulting from fitting a GLM and D_1 is the deviance from fitting a submodel. Then the asymptotic distribution of $(D_1 - D_0)$ is χ^2_r where r is the difference in the number of parameters between the two models. Note that when μ is replaced by $\hat{\mu}$ from the estimated model we get what is called observed deviance in some references.

5.6.2 Estimation of the scale parameter

When the scale parameter is unknown, an estimate is obtained using one of the following methods:

- The Deviance method: $\hat{\phi} = \frac{D}{N-(p)}$.
- Pearson χ^2: $\hat{\phi} = \frac{\chi^2}{N-(p)}$.
- Maximum likelihood estimation (also in agreement with the method of moments)where the estimate obtained is

$$\hat{\phi} = \frac{Var(y_i)}{V(\hat{\mu})} = \frac{\sum(y_i - \hat{\mu})^2}{(n-p)V(\hat{\mu})} \tag{5.17}$$

  and p is the number of parameters estimated.

5.7 Quasi-Likelihood Estimation

In some statistical investigations we are uncertain about the distribution of the data and further it might not necessarily be a member of the exponential family. Uncertainty about the distribution makes it impossible to directly use the techniques discussed earlier. Thus it is not possible to directly exploit the nice properties associated with GLMs.

It would therefore be useful to have inferential methods which work as well (or almost as well) as maximum likelihood but without having to make specific distributional assumptions. This is the basic idea behind the Quasi-likelihood. That is to derive the likelihood like quantity whose construction requires few or less restrictive assumptions. Let us restate the score equation as

$$\begin{aligned} U = \frac{\partial l}{\partial \beta} &= \frac{1}{a(\phi)} \sum_{i=1}^{N} (y_i - \mu_i) w_i g(\mu_i) X_i \\ &= \frac{1}{a(\phi)} \sum_{i=1}^{N} \frac{(y_i - \mu_i)}{v(\mu_i) g'(\mu_i)} X_i \end{aligned}$$

It is clear that the likelihood on the assumed distribution for y_i is constructed through μ_i and $v(\mu_i)$. The choice of the distribution determines the mean-variance relationship. The idea of quasi-likelihood estimation is to use the relationship of the mean and variance in a similar manner as above. In contrast to full likelihood method of estimation we do not specify a probability distribution, but only the mean and variance function.

The quasi-likelihood is defined as

$$Q_i(\mu_i; y_i) = \int_{y_i}^{\mu_i} \frac{y_i - \mu_i}{\phi v(\mu_i)} ds$$

which by definition has a derivative with respect to μ_i equal to

$$q_i = \frac{y_i - \mu_i}{\phi v(\mu_i)}$$

The q_i satisfies the same conditions satisfied by $\frac{\partial \ell_i}{\partial \mu_i}$ where

$$\frac{\partial l_i}{\partial \mu_i} = \frac{\partial ln f(y_i, \theta_i, \phi)}{\partial \mu_i}$$

for the exponential family distributions.

Since the components of Y are independent by assumption, the quasi-likelihood for the complete data is the sum of the individual contributions:

$$Q(\mu; y) = \sum Q_i(\mu_i; y_i).$$

By analogy, the quasi-deviance function for a single observation is

$$Q(y_i; \mu_i) = -2\sigma^2 Q(\mu_i; y_i) = 2 \int_{y_i}^{\mu_i} \frac{y_i - \mu_i}{\phi v(\mu_i)} ds$$

(One should note the reversal of the order of integration). The total deviance, $D(y; \mu)$, is the sum of the individual components, and only depends on y and μ, but not σ^2. It should also be noted that the complete quasi-likelihood only depends multiplicatively on σ^2, so that it does not affect the MLEs of β.

Let us restate the log likelihood of an exponential family as

$$\ell_i = \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)$$

It can be shown that

$$\mathrm{E}[q_i]=0,\ \mathrm{E}[\tfrac{\partial l_i}{\partial \mu_i}]=0$$

and that

$$\text{var}(\tfrac{\partial \ell_i}{\partial \mu_i}) = \tfrac{1}{\phi v(\mu_i)}.$$

Thus the roles played by the log-likelihood and $\frac{\partial l_i}{\partial \mu_i}$ can be taken up by Q_i and q_i respectively.

The ϕ occurring in q_i is mainly the constant of proportionality relating $\text{var}(y_i)$ to $v(\mu_i)$. The Maximum Quasi-likelihood method assumes that $\text{var}(y_i)$ is proportional to $v(\mu_i)$. That is $\text{var}(y_i) = \phi v(\mu_i)$ where

$$v(\mu_i) = \tfrac{\partial \ell_i}{\partial \mu_i}.$$

Note that $v(\mu_i)$ is not exactly the same as the earlier definition. The variance function $v(\mu_i)$ is specified using information about how the variance changes with the mean and nothing more.

To find the maximum Quasi-likelihood (MQL) estimator of β, we solve the MQL equation, given by

$$\frac{\partial}{\partial \beta}(\sum Q_i) = 0 \tag{5.18}$$

Evaluating the derivative we have

$$\begin{aligned}
\tfrac{\partial}{\partial \beta}(\sum_{i=1}^{N} Q_i) &= \textstyle\sum_{i=1}^{N} \tfrac{\partial Q_i}{\partial \beta} \\
&= \sum_{i=1}^{N} \tfrac{\partial Q_i}{\partial \mu_i} \tfrac{\partial \mu_i}{\partial \beta} \\
&= \sum_{i=1}^{N} \left[\tfrac{y_i - \mu_i}{\phi v(\mu_i)}\right] \tfrac{\partial \mu_i}{\partial q} \tfrac{\partial q}{\partial \beta} \\
&= \sum_{i=1}^{N} \left[\tfrac{y_i - \mu_i}{\phi v(\mu_i)}\right] \tfrac{1}{g'(\mu_i)} X_i = 0
\end{aligned} \tag{5.19}$$

In matrix notation we have

$$\frac{1}{\phi} X' W \triangle (Y - \mu) = 0 \tag{5.20}$$

This is the same as the score statistic U under the GLM, but here $v(\mu_i)$ is determined from the mean-variance relationship, not from distributional assumptions. Thus the QL model can be fitted using exactly the same method as for fitting a GLM to obtain $\hat{\beta}$. AS in the case of GLMs the estimate of β is not affected by ϕ.

5.8 Extended quasi-likelihood function

The quasi-likelihood can be extended to include terms for the variance. This will allow us to compare different variance functions, and opens up the possibility of modelling the dispersion as a function of covariates.

For a single observation, y, we want to construct a function $Q^+(\mu;\sigma^2;y)$ that, for known σ^2, is the same as $Q(\mu;y)$, but which also has the properties of a log likelihood with respect to derivatives of σ^2. Thus we have to have

$$\begin{aligned} Q^+(\mu;\sigma^2;y) &= Q(\mu;y) + h(\sigma^2;y) \\ &= -\tfrac{D(y;\mu)}{2\sigma^2} + h(\sigma^2;y) \end{aligned} \tag{5.21}$$

for some function $h(\sigma^2;y)$, which we assume to be in the form

$$h(\sigma^2;y) = -\frac{1}{2}h_1(\sigma^2) + h_2(y). \tag{5.22}$$

If Q^+ is to behave like a log likelihood with respect to σ^2, we must have $E(\partial Q^+/\partial\sigma^2) = 0$. Thus

$$0 = \frac{1}{2\sigma^4}E(D(y;\mu)) - \frac{1}{2}h_1'(\sigma^2). \tag{5.23}$$

If we make $E(D(y;\mu))$ the subject of the formula in equation (5.23) we have

$$E(D(y;\mu)) = \sigma^4 h_1'(\sigma^2) \tag{5.24}$$

To a rough first order approximation we have $E(D(y;\mu)) = \sigma^2$, giving $h_1'(\sigma^2) = log(\sigma^2)+$ const. Thus the extended quasi-likelihood is given by

$$Q^+(\mu;\sigma^2;y) = -\frac{1}{2}D(y;\mu)/\sigma^2 - \frac{1}{2}log(\sigma^2). \tag{5.25}$$

If we have information about the higher order moments, we can improve the approximation. It can be shown that

$$E(D(y;\mu)) \simeq \sigma^2 + \frac{1}{12V^2}6\sigma^4 VV'^2 - 3\sigma^4 V^2 V'' - 4V\kappa_3. \tag{5.26}$$

where V is the variance function, and κ_3 is the third order cumulant. Members of the exponential family of distributions (and averages from these), have the property

$$\kappa_{r+1} = \kappa_r' \kappa_2, \text{ for r} \geq 2.$$

Cumulants

Cumulants are constants that, like moments, can be used to describe a probability distribution. Formally,

$$\exp(\textstyle\sum_{r=1}^{\infty} \kappa_r \frac{t^r}{r!}) = \sum_{r=1}^{\infty} \mu_r^{'} \frac{t^r}{r!}$$

where μ_r' is the r^{th} moment (about the origin). The first four moments and their cumulants are related like this:

$\mu_1' = \kappa_1$

$\mu_2' = \kappa_2 + \kappa_1^2$

$\mu_3' = \kappa_3 + 3\kappa_2\kappa_1 + \kappa_1^3$

$\mu_4' = \kappa_4 + 4\kappa_3\kappa_1 + 6\kappa_2\kappa_1^2 + \kappa_1^4$

where $\kappa_2 = \sigma^2 V(\mu)$, and the differentiation is with respect to μ. From this we get

$$\begin{aligned} E(D(y;\mu)) &\simeq \sigma^2(1 + \tfrac{5(\kappa_3^2/\kappa_2^3)^2 - 3(\kappa_4/\kappa_2^2)}{12}) \\ &= \sigma^2(1 + \tfrac{\sigma^2(2V'^2/V - 3V'')}{12}) \end{aligned} \tag{5.27}$$

as well as

$$\begin{aligned} Var(D) &\simeq 2\kappa_2^2/V^2 = 2\sigma^4 \\ Cov(D,Y) &\simeq (\kappa_3 - \kappa_2\kappa_2')/V \end{aligned} \tag{5.28}$$

The covariance obviously reduces to 0 under the property of exponential family cumulants above. If we use the simpler assumption that σ^2 is sufficiently small that $E(D(y;\mu)) \simeq \sigma^2$, we find that the derivatives

$$\frac{\partial Q^+}{\partial \mu} = \frac{Y - \mu}{\sigma^2 V(\mu)} \quad \text{and} \quad \frac{\partial Q^+}{\partial \sigma^2} = \frac{D(y;\mu)}{\sigma^4} - \frac{1}{2\sigma^2} \tag{5.29}$$

have zero mean and approximate covariance matrix

$$\begin{pmatrix} \frac{1}{\sigma^2 V(\mu)} & \frac{\kappa_3 - \kappa_2 \kappa_2'}{2\sigma^6 V^2} \\ \\ \frac{\kappa_3 - \kappa_2 \kappa_2'}{2\sigma^6 V^2} & \frac{1}{2\sigma^4} \\ \\ \end{pmatrix}$$

The off-diagonal terms are zero under the property above, and even if it does not hold, they are often negligible. The expected value of the second derivative matrix is the same as above, except that the off-diagonal terms are zero. From this we can see that, Q^+ has the properties of a quasi-likelihood with respect to both mean and dispersion parameter.

The idea of quasi-likelihood was extended to deal with correlated data in the context of longitudinal data by Liang and Zeger (1986) and Zeger and Liang (1986). This led to the generalized estimating equations (GEEs) which are discussed in the next chapter. A more recent extensive treatment of GEEs is given by Diggle et al. (2002).

‘

Chapter 6

Generalized Linear Models for Longitudinal data

6.1 General Estimating Equations (GEEs)

The basic ideas of GEEs was first introduced by Liang and Zeger in 1986. GEEs are used to model correlated data from longitudinal or repeated measure studies and from clustered or multilevel studies. GEEs can be regarded as an extension of quasilikelihood models for independent measurements. The emphasis is on modelling the expectation of the dependent variable in relation to the covariates (just like with GLMs). The correlation structure is considered to be a nuisance (not of interest in itself), which is accounted for by the method. One should recall that for quasilikelihood models we specify how the mean of the responses depends on the explanatory variables (the link function) and how the variance depends on the mean (the variance function). The setting is as follows, on each of $i = 1, \ldots, N$ subjects or clusters, there are n_i measurements $y_i = (y_{i1}, \ldots, y_{in_i})$. Measurements on different subjects are assumed to be independent and measurements on the same subject or cluster are allowed to be correlated. The model specification of a GEE involves three elements:

- **Systematic part**: This relates the expectation $E(y_{ij}) = \mu_{ij}$ to the linear predictor through the link function

$$g(\mu_{ij}) = \eta_{ij} = x_{ij}^{'}\beta$$

- **Random part**: which specifies how the variance $\mathrm{Var}(y_{ij})$ is related to the mean $E(y_{ij})$ by specifying a variance function $V(\mu_{ij})$ such that

$$Var(Y_{ij}) = \phi V(\mu_{ij}).$$

- **The correlation part**: This is the part which differentiates the GEE model from the GLM. One needs to allow for a correlation structure for observations on the same subject or cluster. This is done by specifying a working correlation matrix.

Hence the specification of a GEE model involves the same steps as specification of a GLM but with the additional specification of a working correlation structure. In the text that follows we will discuss how GEEs are derived and solved.

6.1.1 Deriving GEEs

The score equations for GLM's have been derived in the univariate independent observation case as

$$U = \sum_{i=1} \frac{\partial \mu_i}{\partial \beta} v_i^{-1}(y_i - \mu_i) = 0$$

with $v_i = \text{Var}(Y_i)$. In the case where the outcome Y_i is multivariate that is $Y_i = (Y_{i1}, \ldots, Y_{in_i})'$ with independent components the score function becomes

$$\begin{aligned} U &= \sum_{i=1} \sum_{j=1} \frac{\partial \mu_{ij}}{\partial \beta} v_{ij}^{-1}(y_{ij} - \mu_{ij}) \\ &= \sum_{i=1} \frac{\partial \mu_i'}{\partial \beta} V_i^{-1}(y_i - \mu_i) \end{aligned} \tag{6.1}$$

where μ_i=E(Y_i) and V_i=Var(Y_i)=diag(Var(Y_{ij})). One should note that, when fitting the GLM the same equations had to be solved. GEEs are obtained by allowing a non-diagonal V_i in equation (6.1). V_i is now a $n_i \times n_i$ covariance matrix with diagonal elements given by v_{jj} and will be of the form

$$V_i(\beta, \alpha) = \phi A_i^{\frac{1}{2}}(\beta) R_i(\alpha) A_i^{\frac{1}{2}} \tag{6.2}$$

in which $A_i^{\frac{1}{2}}$ is a diagonal matrix with diagonal elements given by $\sqrt{v_i(\mu_{ij}(\beta))}$. $R_i(\alpha)$ is the correlation matrix of the vector Y_i which depends on a vector α of unknown parameters. The correlation matrix is usually unknown, so therefore one specifies a "working correlation matrix". In GEE models, if the mean is correctly specified, but say the variance and correlation structure are incorrectly specified, then GEE models will still provide consistent estimates of the parameters and thus the mean function as well, while consistent estimates of the standard errors can be obtained via a robust sandwich

estimator. Similarly, if the mean and variance are correctly specified but the correlation structure is incorrectly specified, the parameters can also be estimated consistently and the standard errors can be estimated consistently via the sandwich estimator. If all three are specified correctly, then the estimates of the parameters are more efficient.

GEEs are obtained by solving the score equations in (6.1) allowing for non-zero off diagonal elements for V_i. In order to solve these equations we need to use numerical methods and some of the numerical methods that can be used are explained briefly in the text that follows.

Iteratively Reweighted Least Squares Algorithm

The IRWLS algorithm used to fit models with GEE is an extension of the algorithm to fit generalized linear models. In the score equation (6.1)we let

$$D_i = \frac{\partial \mu_i}{\partial \beta_i} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_i} = \frac{\partial \mu_i}{\partial \eta_i} X_i = D_i^{*-1} X_i \tag{6.3}$$

So that estimation of β is done with iteratively reweighted least squares by regressing the working response vector

$$Z = X\hat{\beta} + D^*(Y - \hat{\mu}) \tag{6.4}$$

on X with block diagonal weight matrix W, whose i^{th} block, corresponding to the i^{th} cluster is

$$W_i = (D_i^*)^{-1} A_i^{-\frac{1}{2}} R_i^{-1}(\alpha) A_i^{-\frac{1}{2}} (D_i^*)^{-1} \tag{6.5}$$

such that $D_i^* = \text{diag}(\partial \eta_{i1}/\partial \mu_{i1}, \ldots, \partial \eta \mu_{in_i}/\partial \mu_{in_i})$, and D^* is the block diagonal matrix with diagonal matrices D_i^*. To update β the equation below is used

$$\begin{aligned} \hat{\beta}^{(t+1)} &= (\sum_{i=1}^{N} X_i' W_i^{(t)} X_i)^{-1} \sum_{i=1}^{N} X_i' W_i^{(t)} Z_i^{(t)} \\ \hat{\beta}^{(t+1)} &= (\sum_{i=1}^{N} D_i^{'(t)} V_i^{-1(t)} D_i^{(t)})^{-1} \sum_{i=1}^{N} D_i^{'(t)} V_i^{-1(t)} D_i^{*(t)} Z_i^{(t)} \end{aligned} \tag{6.6}$$

According to Liang and Zeger (1986) the solution is obtained by alternating between estimation of ϕ, α, and β using method of moments (MoM) estimators for ϕ and α. Thus in summary the IRWLS proceeds as follows:

- **Step 1:** Assuming R = I and $\phi = 1$, provide initial estimate of β with GLM algorithm.

- **Step 2:**Estimate ϕ and α.
- **Step 3:**Use updated ϕ and α to estimate β.
- **Step 4:**Return to step 2. Repeat steps 2 and 3 until convergence.

A MoM estimator of ϕ

$$\hat{\phi} = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{\hat{r_{ij}}}{K-p} \tag{6.7}$$

where $K = \sum_{i=1}^{N} n_i$ and $r_{ij} = \frac{y_{ij}-\mu_{ij}}{\sqrt{v(\mu_{ij})}}$. Given $\hat{\phi}$, a MoM estimator for the exchangeable parameter α is

$$\hat{\alpha} = \hat{\phi} \sum_{i=1}^{N} \sum_{j>j'}^{N} \hat{r}_{ij}\hat{r}_{ij'} / [\sum_{i=1}^{N} \frac{1}{2} n_i(n_i-1) - p] \tag{6.8}$$

Newton Iteration

To solve the system of equations using the Newton Iteration method the following steps must be followed:

- Compute an initial estimate of β from a GLM (i.e. by assuming independence)
- Compute an estimate R(α) of the working correlation on the basis of the current Pearson residuals and the current estimate of β
- Compute an estimate of the variance as

$$V_i = \phi A_1^{\frac{1}{2}} \hat{R}(\alpha) A_1^{\frac{1}{2}}$$

- Compute an updated estimate of β based on the Newton-step

$$\beta = \beta + [\sum_i \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} \frac{\partial \mu_i}{\partial \beta'}]^{-1} [\sum_i \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} (y_i - \mu_i)] \tag{6.9}$$

One should iterate through steps 2-4 until convergence. Note that ϕ needs not to be estimated until the last iteration.

6.1.2 The estimate of β

The GEE estimate $\hat{\beta}$ is often very similar to the estimate obtained if observations were treated as being independent. In other words, the estimate $\hat{\beta}$ for GEEs is often very similar to the estimate obtained by fitting a quasi-likelihood model to the data. However the GEE estimate $\hat{\beta}$ is generally a good estimate in the sense that

- The estimator is asymptotically consistent, that is, when the sample size n increases then $\hat{\beta}$ becomes (almost) indistinguishable from the true value β.
- The estimator is asymptotically normal,

  $$\hat{\beta} \sim N(\beta, \Sigma)$$

  In practice, an estimate of Σ is obtained as a by product of the estimation procedure.

The asymptotic normality holds even if

- the variance function $V(\mu)$ is incorrectly specified
- the working correlation matrix R is NOT the true correlation matrix.

The variance Σ depends on the correlation structure in data and this is the whole point in incorporating a working correlation matrix. One should note that if we have a good estimate for Σ then we can test hypotheses about model parameters, construct and test hypotheses about linear contrasts, and make confidence intervals for the estimated parameters on their functions. The following section explains the estimation of Σ.

6.1.3 Estimating Σ

The covariance matrix Σ is generally unknown and must be estimated from data. In GEEs one can choose between two different forms of estimates. Both depend on the form of the working correlation matrix. The two estimates are:

- The empirical, robust or sandwich estimator which is given by

  $$V(\hat{\beta}) = M_0^{-1} M_1 M_0^{-1}$$

where

$$M_0 = \sum_{i=1}^{N} D_i' \hat{V}_i^{-1} D_i \quad \text{and} \quad M_1 = \sum_{i=1}^{N} D_i' \hat{V}_i^{-1} (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)' V_i^{-1} D_i$$

This estimate tends asymptotically to the true Σ, even if the working correlation is misspecified. Let

$$(y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i) = \tilde{V}_i$$

- If the working correlation is correct, a better estimate is the so-called model-based estimator which is given by $\mathrm{V}(\hat{\beta}) = [\sum_{i=1}^{N} D_i' \tilde{V}_i^{-1} D_i]^{-1}$

One should notice that if $\hat{V}_i = (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)'$ then the two are equal (this occurs only if the true correlation structure is correctly modelled). In practice the empirical estimate is often quite good. If one has no idea about the structure of the correlation, the independence working correlation is often a good choice.

6.1.4 Working correlations discussed

As already mentioned the correlation matrix is usually unknown and must be estimated. It is estimated in the iterative fitting process using the current value of the parameter vector β to compute appropriate functions of the Pearson residual

$$r_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}}$$

There are several specific choices of the form of the working matrix $R_i(\alpha)$ to model the correlations of the individual responses. For additional appropriate choices we refer to Liang and Zeger (1986). Other recent equally good references include Verbeke and Molenberghs (2000), Diggle et al. (2002), and Molenberghs and Verbeke (2005).

The following descriptions present a few of the common choices of working correlations in SAS supported by PROC GENMOD and the resulting covariances.

Autoregressive AR(1) working correlation

Autoregressive is a term derived from times series analysis that assumes observations are related to their own past values through one, two, or a

higher order dependencies. An autoregressive correlation structure indicates that two observations taken close in time (or space) within an individual tend to be more highly correlated than two observations taken far apart in time from the same individual. Formally, $\text{Corr}(y_{ij}, y_{ij}) = \rho_{jj} = 1$ and $\text{Corr}(y_{ij}, y_{ik}) = \rho_{jk}(j \neq k)$ decreases in value as the absolute difference between j and k gets larger. A first-order autoregressive correlation structure (AR (1)) specifies that $\rho_{jk} = \rho^{|j-k|} =$ where ρ is the correlation when $|j - k| = 1$. One should note that the AR(1) working correlation matrix is not helpful when the time points at which measurements are recorded are not equidistant and unequal observations per individual. Thus the AR (1) structure is mainly used when one has balanced data. For illustration purposes consider the case n=4 observations per person and equidistance between two consecutive measurements. Then the AR(1) working correlation for n=4 is given by

$$R = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

The exchangeable working correlation

The exchangeable working correlation structure assumes non-zero, yet uniform correlations for all pairs of within-subject variables, that is every observation within an individual is equally correlated with every other observation from that individual. In other words it means that one can swap the order of any two measurements without changing the correlation structure. This choice of correlation structure may not be reasonable with multiple measurements collected over time, since the correlations most likely will diminish as the time lag between observations increases. This type of correlation structure is sometimes referred to as the compound symmetry or spherical structure. The exchangeable working correlation assumes that $\rho_{12} = \rho_{13} = \ldots = \rho_{jj'} = \alpha$ say which is analogous to applying the compound symmetry assumption of repeated measures with PROC GENMOD or PROC MIXED. It is used when one is dealing with clustered data where particular ordering of observation is not presented. For illustration purposes suppose we consider the case of n=4, then the exchangeable working correlation is given by

$$R = \begin{bmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{bmatrix}$$

Note that in both the AR(1) and the compound symmetry structures only a single correlation parameter needs to be estimated.

The unstructured working correlation

Unstructured working correlation structure assumes unconstrained pair-wise correlations where each correlation is estimated from the data (the most complex model) and is applied to balanced data sets. No assumption is made about the relative magnitude of the correlation between any two pairs of observations. Formally, $\rho_{jj} = 1$ and ρ_{ij} is free to take any value between -1 and +1. The unstructured working correlation matrix estimates $n*(n-1)/2$ correlations from the data. Thus care should be taken when the unstructured working correlation is being used since the number of parameters to estimate becomes large even for moderate n. In practice this means that the correlation parameters can be poorly estimated or that the statistical program may fail to produce a result(converge). For illustration purposes consider the case n=4, then the unstructured working correlation is given by

$$R = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{bmatrix}$$

The independence working correlation

GEE provides a general approach for analyzing discrete and continuous responses with marginal models. Given that a data set consists of repeated measurements within individuals, the simplest possible correlation structure is to (usually incorrectly) assume independence. This assumption is equivalent to observations collected from the same individual being completely uncorrelated with every other observation measured from that individual; correlations are assumed to be 0 for all pair-wise combinations of the within-subject variables. If ρ_{jk} is the correlation between observations j and k, then $\rho_{jj} = 1$ and $\rho_{jk} = 0$, $j \neq k$. Since all off-diagonal correlations are zero, a working correlation matrix is not estimated for this situation. Thus the GEE reduces to the independence (GLM) estimating equation. For the case where n=4 the independence matrix is given by:

$$R(\alpha) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Other working correlations are available depending on the data pattern and problem under consideration. The references by Verbeke and Molenberghs (2000), Diggle et al. (2002) and Molenberghs and Verbeke (2005) among others can be consulted for detailed discussion of the different working correlations and supporting arguments for them.

6.2 Inference

Because the fitting of a model by GEE is not based on a likelihood there is no likelihood ratio test available. Thus the Wald test statistic is often used when making inferences. The hypothesis of primary interest is to test the following general statement about the vector of parameters β namely

$$H_0 : L\beta = 0 \text{ versus } H_a : L\beta \neq 0$$

where L is a known matrix of known constants of dimension say $r \times p$. To use the Wald test we need the variance of $\hat{L\beta}$ which is given by $\mathrm{L}(\mathrm{Var}(\hat{\beta}))\mathrm{L}'$. The Wald test statistic is then given by

$$W_s = (L\hat{\beta} - L\beta)'[LVar(\hat{\beta}L']^{-1}(L\hat{\beta} - L\beta) \tag{6.10}$$

which under H_0 is asymptotically distributed as χ^2 with d.f. equal to rank(L).

One compares two nested models using the distributional result $\hat{\beta} \sim \mathrm{N}(\beta, \Sigma)$. Let M_0 be a sub-model of M_1, where M_1 has the parameter vector β. M_0 is derived from M_1 by setting some parameters in β equal to 0. This can generally be written as a matrix equation with a matrix L such that one has model M_0 if the equation $L\beta = 0$ is fulfilled. The test of model M_0 against the larger model M_1 is obtained by the Wald statistic

$$\beta' L'(L\hat{\Sigma}L')^{-1}L\hat{\beta} \sim \chi^2_{p_2}$$

where $p_2 = p_1 - p_0$ and p_1 and p_0 are the number of parameters of model M_1 and M_0. We note that this statistic is not exact because under GEE the distributional properties of the correlated data is not fully discernible.

6.3 Application to epileptic data

6.3.1 Application to the Thall and Vail data

We are going to apply the method of GEE to the Thall and Vail data which is displayed in Table 3.1 on page 12.

We consider the following model:

$$log(E(Y_{ij})) = \beta_0 + \beta_1 Age + \beta_2 T_i + \beta_3 t_{ij} + \beta_4 T_i t_{ij}$$

where Y_{ij}= number of epileptic seizures in interval j, t_{ij}= length of interval j and T_i is the treatment group.

Below is a SAS code for the standard GEE model:

proc genmod data=elisha;
CLASS ID Trt Tclass;
*MODEL y = Age Trt Time Trt*Time /*
dist = poisson link = log;
REPEATED subject= ID / WITHINSUBJECT=Tclass type=un corrw modelse;
run;

It should be noted that here the model is fitted in exactly the same way as in a standard GLM but the difference is that the code accounts for the correlation structure in the data via the repeated statement of SAS. The statements in the code first produce the usual output from fitting a generalized linear model (GLM) to these data. The estimates are used as initial values for the GEE solution. The REPEATED statement defines the GEE character of the model. In the REPEATED statement the option subject option gives the name of the variable that contains a unique identification of code for each cluster. Since this variable is a categorical variable it must first be named in the CLASS option. The WITHINSUBJECT option names the variable that distinguish different items within a cluster. In this application they are differentiated by different time points. This variables must also be named in the CLASS option. The TYPE option is used to specify the correlation structure. In the SAS code above the unstructured correlation type is specified. The default working correlation type is the independent. Some of the possibilities of type= include autoregressive (AR), exchangeable (EXCH or CS), independent (IND), and user specified correlation matrix (USER or FIXED). The option MODELSE in the REPEATED option tells SAS to print out model based estimates of the standard errors as well as the default empirically corrected standard errors. Corrw displays the estimated working correlation matrix. The COVB option prints out the estimated covariance matrix of the estimate of beta, both the usual estimate and the "robust" version are printed.

Results

Table 6.1: Parameter estimates and standard errors from GEE models for the Thall and Vail (1990) data.

Correlation Structure	Parameter	Estimate	Error	p-value
UN	β_0	4.0648	0.4003(0.4091)	<.0001
	β_1	-0.0308	0.0134(0.0133)	0.0207
	β_2	-0.0073	0.1940(0.2024)	0.9711
	β_3	0.066	0.0286(0.0246)	0.2008
	β_4	-0.0289	0.0397(0.0302)	0.3373
EXCH	β_0	3.8063	0.5774(0.4924)	<.0001
	β_1	-0.0241	0.0194(0.0175)	0.1683
	β_2	-0.0297	0.2446(0.2183)	0.8917
	β_3	-0.3872	0.0678(0.0593)	<.0001
	β_4	-0.0784	0.1012(0.0929)	0.3984
AR(1)	β_0	4.0862	0.5045(0.4831)	<.0001
	β_1	-0.0240	0.0170(0.0169)	0.1570
	β_2	0.0061	0.2129(0.2089)	0.9768
	β_3	-0.4623	0.0857(0.0450)	<.0001
	β_4	-0.0877	0.1239(0.0863)	0.3095
IND	β_0	3.6958	0.3701(0.6413)	<.0001
	β_1	-0.0180	0.0120(0.0199)	0.3668
	β_2	0.0092	0.2038(0.2145)	0.9660
	β_3	-0.3702	0.0802(0.0593)	<.0001
	β_4	-0.0642	0.1136(0.0793)	0.4178

Interpretation of results

Table 6.1 summarizes the GEE results of the four working correlation specifications. The interpretation of the parameters in the marginal (population averaged) and random (mixed) effects model is analogous to the standard logistic regression model, however there are differences (as noted above) in how we adjust for the correlations. In this section we will discuss the results of the AR(1) model of which similar conclusions are reached for the other

models. The p-value for the interaction effect ($T_i t_{ij}$) is 0.3095 which is not significant at the 5% significance level. Time (t_{ij}) is the only significant effect in the AR(1) model at the 5% significance level. The p-value under the UN correlation structure are not in the same structure as those under other correlation structure therefore one is bound to have less reliance in them.

6.3.2 Application to the unbalanced data

We are going to apply the method of GEE to the data set which is displayed in Table 3.2 on page 13.

For a brief discussion of missing values in longitudinal data see section 2.6 . Suppose that you intend to take measurements $Y_{i1}, \ldots, Y_{in_i}$ for the ith unit. Missing values for which Y_{ij} are missing whenever Y_{ik} is missing for all $j > k$ are called dropouts. Otherwise, missing values that occur intermixed with non missing values are intermittent missing values. The GENMOD procedure can estimate the working correlation from data containing both types of missing values by using all available pairs method, in which all non missing pairs of data are used in the moment estimators of the working correlation parameters defined previously. The resulting covariances and standard errors are valid under the missing completely at random (MCAR) assumption.

Thus, because these data are not balanced, we use the SUBJECT option of the REPEATED statement to give SAS the subject variable ID as a class or factor variable so that it can figure out where the missing values are and use this information in estimating the correlation matrix. This time the model of interest is:

$$log(E(Y_{ij})) = \beta_0 + \beta_1 T_i.$$

Below is SAS code for the standard GEE model:

```
proc genmod data=epilepsy;
CLASS ID Trt ;
MODEL y = Trt /
dist = poisson link = log;
REPEATED subject= ID / type=un corrw modelse;
run;
```

Table 6.2: Parameter estimates and standard errors from GEE models for the unbalanced data.

Correlation Structure	Parameter	Estimate	Error	p-value
UN	β_0	-2.0625	0.0533(0.0863)	<.0001
	β_1	-0.0308	0.0134(0.0133)	<.0001
EXCH	β_0	1.7778	0.1101(0.1298)	<.0001
	β_1	-0.7131	0.2045(0.1726)	<.0001
AR(1)	β_0	1.7778	0.1101(0.1298)	<.0001
	β_1	-0.7131	0.2045(0.1726)	<.0001
IND	β_0	1.9480	0.0356(0.0928)	<.0001
	β_1	-0.7615	0.0868(0.1410)	<.0001

Results

From the results in table 6.2 we can see that choosing the AR(1) or exchangeable correlation yields the same results. In all cases the empirical standard errors are close to the model based standard errors. Regardless of the chosen correlation the treatment effect is significant.

6.4 Summary

The GEE works best if the number of observations per subject is small and the number of subjects is large or if in longitudinal studies (e.g. growth curves), the measurements are taken at the same time for all subjects. The main advantage of GEE models is that, if the mean is correctly specified, but the variance and correlation structure are incorrectly specified, then GEE models still provide consistent estimates of the parameters and hence the mean function as well, further consistent estimates of the standard errors can be obtained via a robust sandwich estimator. If the mean and variance are additionally correctly specified but the correlation structure is the only incorrectly specified, then both regression parameters and the standard errors can be estimated consistently with the sandwich estimator. If all three are specified correctly, then the estimates of the parameters are even more efficient.

Chapter 7

Modelling Non-Normal Longitudinal Data with random effects

7.1 Introduction

The generalized linear mixed models (GLMM) extend GLMs by allowing for random, or subject-specific, effects in the linear predictor. These models are useful when the interest of the analyst lies in the individual response profiles rather than the marginal mean $E(Y_{ij})$. The random effects not only determine the structure of correlation between observations on the same subject, they also take account of heterogeneity among subjects, due to unobserved characteristics. Thus proper use of random effects can account for extra variability which cannot be fully accounted for through measured covariates and the dispersion parameter ϕ.

7.2 Generalized linear mixed models (GLMM)

Given a vector b_i of random effects for a unit or cluster i, it is assumed that all responses Y_{ij} are independent, with density

$$f(y_{ij}|\theta_{ij},\phi) = \exp\Big\{\phi^{-1}[y_{ij}\theta_{ij} - \psi(\theta_{ij})] + c(y_{ij},\phi)\Big\}$$

in which θ_{ij}, the natural parameter is now modelled as

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

where x_{ij} is a vector of covariates for fixed effects and Z_{ij} is a vector of covariates for random effects. Similar to GLMs, the following (conditional) relations hold

$$\mu_{ij} = E[Y_{ij}|b_i] = \psi'(\theta_{ij}) \text{ and } \text{Var}[Y_{ij}|b_i] = \phi\psi''(\theta_{ij}) = \phi V(\mu_{ij})$$

where $\theta_{ij} = g(\mu_{ij}) = x_{ij'}\beta + z_{ij'}b_i$. As before, $g(.)$ is called the link function and $V(.)$ the variance function. The p-dimensional vector β denotes the fixed effects parameter vector while the q-dimensional vector b_i denotes the subject specific random effects parameter vector. The p and q dimensional vectors x_{ij} and z_{ij} contain subject i's covariate information for the fixed and random effects, respectively. The specification of the GLMM is completed by assuming that the random effects, $b_i (i = 1, \ldots, N)$, are mutually independent and identically distributed with density function $f(b_i|\alpha)$. Hereby α denotes the unknown parameters in the density. Following the notation used in Chapter 4, it is assumed that

$$\mathbf{b_i} \sim N(\mathbf{0}, G)$$

Let $f_{ij}(y_{ij}|b_i, \beta, \phi)$ denote the conditional density of Y_{ij} given b_i, we then have that the marginal distribution of Y_i is given by

$$f_i(y_i|\beta, \beta, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|b_i, \beta, \phi) f(b_i|G) db_i \tag{7.1}$$

where $f(b_i|G)$ is the density of the $N(0, G)$ distribution. The likelihood function for β, G, and ϕ now equals

$$L(\beta, G, \phi) = \prod_{i=1}^{N} f_i(y_i|\beta, G, \phi) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|b_i, \beta, \phi) f(b_i|G) db_i \tag{7.2}$$

Under the non-Gaussian linear mixed model, the integral cannot be worked out analytically because the random effects enter into the integrand non-linearly. In general, approximations are required and three possible approaches can be followed. These are either an approximation of the integrand, approximation of data or approximation of the integral. Roughly speaking, under the first approach, $\prod_{j=1} f_i(y_i|\beta, G, \phi)$ is approximated by a normal density such that the integral can be calculated analytically, as in the normal linear mixed model. In practice, this approximation will be accurate whenever the response y_{ij} is 'sufficiently continuous' and/or if all the n_i are sufficiently large. The likelihood of the observed data is a marginal likelihood where the random effects have been "integrated out". Unfortunately, this marginal likelihood does not generally have a closed-form expression and approximate

methods of estimation must be used. The most commonly used methods include marginal quasi-likelihood (MQL), penalized quasi-likelihood (PQL), Markov Chain Monte Carlo (MCMC), and Gaussian quadrature (GQ). An added improvement of Gaussian quadrature is adaptive Gaussian quadrature (AGQ).

7.3 Approximation of the Integrand

Since the integral in equation (7.2) cannot be solved analytically one can consider an approximation of the function to be integrated. One of the method that can be used to estimate the integrand is the Laplace approximation of integrand. Laplace's method is an elementary technique for approximating an integral of the form

$$I = \int e^{(Q(b))} db, \tag{7.3}$$

where $Q(b)$ is a smooth real-valued function and has a single maximum in the interior of the domain of integration. One should note that the integrals in $L(\beta, G, \phi)$ can be written in the form $I = \int \exp(Q(b))db$. Letting $\hat{b}$ denote the value that maximizes $Q(b)$, the formula produced by Laplace's method is

$$I \approx \quad (2\pi)^{\frac{q}{2}} \quad | - Q^{''}(\hat{b})|^{-\frac{1}{2}} e^{(Q(\hat{b}))}. \tag{7.4}$$

where q is the dimension of b and $Q^{''}(\hat{b})$ is the Hessian matrix of $Q(b)$ at $\hat{b}$. Taking, for simplicity, the one-dimensional case (q = 1) as an example with the domain of integration being the whole real line, a second-order Taylor series expansion of $Q(b)$ about $\hat{b}$ produces the factor

$$Q(b) \approx Q(\hat{b}) + \frac{1}{2}(b - \hat{b})' Q^{''}(\hat{b})(b - \hat{b}) \tag{7.5}$$

in the integrand. One should note that, as $n \longrightarrow \infty$, the integrand becomes increasingly concentrated near $\hat{b}$. Now equation (7.3) upon substituting equation (7.5) becomes approximated by

$$\int e^{(Q(b))} db \approx e^{(Q(\hat{b}))} \int \exp(\frac{1}{2}(b - \hat{b})' Q^{''}(\hat{b})(b - \hat{b}))db \tag{7.6}$$

of which the integral of the right hand side of equation (7.6) can be found analytically. This gives the one-dimensional version of equation (7.3). Examination of the remainder terms in the Taylor series expansions shows that the order of accuracy of the approximation is as given in equation (7.4). This method is a good approximation in the case of many repeated measures per subject.

7.4 Approximation of the model parameters

Using a Taylor series the pseudo-likelihood technique approximates the original GLMM by a linear mixed model for pseudo-data. In this linearized model the maximum likelihood estimators for the fixed effects and BLUPs for the random effects are obtained using the well-known theory for linear mixed models (as outlined in Chapter 4). The advantage of this approach is that a large number of random effects but also crossed and nested random effects can be handled. A disadvantage is that no true log-likelihood is used. Therefore likelihood-based statistics should be interpreted with great caution. Moreover the estimation process is doubly iterative; a linear mixed model is fit, which is an iterative process, and this procedure is repeated until the difference between subsequent estimates is sufficiently small. In SAS the procedure PROC GLIMMIX enables pseudo-likelihood estimation.

According to Molenberghs and Verbeke (2005), this method of approximation is based on a decomposition of the data into the mean and appropriate error term, with a Taylor expansion of the mean that is a non-linear function of the linear predictor. The methods briefly discussed in this section differ in either the order of the Taylor approximation or the point around which the approximation is expanded.

We will consider the following decomposition of the data (Y_{ij})

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij} = h(x_{ij}^{'} + z_{ij}^{'} b_i) + \varepsilon_{ij} \tag{7.7}$$

where $h(.)$ is the inverse link function and the error terms have the appropriate distribution with

$$Var(Y_{ij}|b_i) = \phi v(\mu_{ij})$$

for $v(.)$ the usual variance in the exponential family. Several methods of implementing the approximation of the data technique have been proposed. Here we discuss two of the commonly referred ones namely the penalized and the marginal quasi-likelihood formulations.

7.4.1 Penalized Quasi-Likelihood (PQL)

The penalized quasi-likelihood (PQL), is based on first-order Taylor expansions around the maximum of current estimates $\hat{\beta}$ and $\hat{b}_i$ of the fixed and random effects via the first-order Laplace approximations to the integrals. This approach produces biased estimates for both the regression and variance components parameters. Breslow and Lin (1995) provided a correction

factor for the estimates of the univariate variance components derived from the second-order Laplace approximations. Breslow and Lin (1996) extend this bias correction to the GLMM with multivariate random effects. This method is implemented in the SAS macro GLIMMIX. To see how the PQL method works consider a linear Taylor expansion of Y_{ij} around $\hat{\beta}$ and $\hat{b}_i$ then

$$\begin{aligned} Y_{ij} &\approx h(x_{ij}^{'}\hat{\beta} + z_{ij}^{'}\hat{b}_i) + h^{'}(x_{ij}^{'}\hat{\beta} + z_{ij}^{'}\hat{b}_i)x_{ij}^{'}(\beta - \hat{\beta}) + h^{'}(x_{ij}^{'}\hat{\beta} + z_{ij}^{'}\hat{b}_i)z_{ij}^{'}(b_i - \hat{b}_i) + \varepsilon_{ij} \\ &\approx \hat{\mu}_{ij} + \upsilon(\hat{\mu}_{ij})x_{ij}^{'}(\beta - \hat{\beta}) + \upsilon(\hat{\mu}_{ij})z_{ij}^{'}(b_i - \hat{b}_i) + \varepsilon_{ij} \end{aligned} \tag{7.8}$$

Thus in vector notation equation (7.8) can be written as

$$Y_i \approx \hat{\mu}_i + \hat{V}_i X_i(\beta - \hat{\beta}) + \hat{V}_i Z_i(b_i - \hat{b}_i) + \varepsilon_i \tag{7.9}$$

Finally re-ordering terms in equation (7.9) yields

$$Y^* = \hat{V}_i^{-1}(Y_i - \hat{\mu}_i) + X_i\hat{\beta} + Z_i\hat{b}_i \simeq X_i\beta + Z_i b_i + \varepsilon_i^* \tag{7.10}$$

where $\varepsilon_i^* = \hat{V}_i^{-1}\varepsilon_i$. Model fitting proceeds by iterating between updating the pseudo responses (Y_i^*) and fitting the above approximate model in equation (7.10) similar to a linear mixed model (covered in chapter 4) until convergence.

7.4.2 Marginal Quasi-Likelihood (MQL)

The MQL is similar to the the PQL, the difference is that the MQL is based on a linear Taylor expansion of the mean μ_{ij} in equation (7.7) around the maximum of current estimate of $\hat{\beta}$ for the fixed effects and around $b_i = 0$ for the random effects. Thus under the MQL

$$Y^* = \hat{V}_i^{-1}(Y_i - \hat{\mu}_i) + X_i\hat{\beta} \simeq X_i\beta + Z_i b_i + \varepsilon_i^* \tag{7.11}$$

otherwise the fitting iteration is exactly the same as in equation (7.10).

The MQL performs reasonably well if a priori the random-effects variance is very small. Both the PQL and MQL perform badly for binary outcomes with few repeated measurements per cluster. With large n_i the PQL provides consistent estimate values while the MQL remains biased. Improvements are possible with higher order Taylor expansions.

7.5 Approximation of the Integral

The likelihood contribution of every subject is generally of the form

$$\int f(z)\phi(z)dz$$

where $\phi(z)$ is the density of the (multivariate) normal distribution. Gaussian quadrature methods replace the integral by a weighted sum:

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

Q is the order of the approximation. The higher Q is the more accurate the approximation will be. The nodes (or quadrature points) z_q are solutions to the Q^{th} order Hermite polynomial. The w_q are well-chosen weights. The nodes z_q and weights w_q are available in tabulated form Alternatively, an algorithm is available for calculating all z_q and w_q for any value Q. With Gaussian quadrature, the nodes and weights are fixed, independent of $f(z)\phi(z)$. With adaptive Gaussian quadrature, the nodes and weights are adapted to the 'support' of $f(z)\phi(z)$. Typically, adaptive Gaussian quadrature needs (much) less quadrature points than the classical Gaussian quadrature. On the other hand, adaptive Gaussian quadrature is much more time consuming. It should also be noted that adaptive Gaussian quadrature of order one is equivalent to Laplace transformation (Molenberghs and Verbeke, 2005).

It is clear in summary that all three methods available to handle random effects in correlated non-Gaussian linear models are highly computer intensive in comparison to the linear mixed model (LMM) which was presented in chapter 4. Thus the fact that the random effects affect the mean response in a non-linear manner makes the non-Gaussian problem much more computationally demanding than in the Gaussian case.

7.6 Inference

In section 7.4 we showed that generalized linear mixed models can be estimated by fitting linear mixed models to the pseudo-data. Since we will be fitting linear mixed models then we can use the same estimation methods used for linear mixed models. Thus we can use the same inference techniques discussed in sections 4.4.3 and 4.6.

7.7 Application to count data

7.7.1 Application to the Thall and Vail data

Background for this data set is given in Chapter 3 and displayed in Table 3.2 on page 13 . We analyse the data using generalized linear mixed models. Since the data are frequency counts interest lies in testing whether the intensities of occurrences are significantly different between the two treatment groups. The covariates used in the analysis are treatment and week. First we assume that conditionally on random effects Y_{ij} is Poisson distributed that is

$$Y_{ij}|b_i \sim \text{Poisson}(\mu_{ij}).$$

Then the following to models are considered namely

$$\log(\mu_{ij}) = \beta_0 + \beta_1 T_i + \beta_2 t_{ij} + \beta_3 T_i t_{ij} + b_{i0} \tag{7.12}$$

$$\log(\mu_{ij}) = \beta_0 + \beta_1 T_i + \beta_2 t_{ij} + \beta_3 T_i t_{ij} + b_{i0} + b_{i1} t_{ij}. \tag{7.13}$$

Hereby Y_{ij} represents the number of epileptic counts measured on subject i in interval j, t_{ij} is the time point at which the jth count is taken for the ith subject and T_i is the treatment indicator for subject i. β_0, β_1, β_2 and β_3 are fixed effects and b_{i0} versus b_{i1}, is the subject specific (random) intercept and slope.

The commands to fit the model in SAS code for penalized quasi-likelihood for GLMM with no slope are

```
Proc glimmix data=Titanic method=RSPL;
Class Subject Treatment;
model Count= Treatment week Treatment*week
/dist= Poisson link=log solution ;
random intercept / subject=subject;
run;
```

The MODEL statement is required and it specifies the dependent variable versus the fixed effects. The fixed-effects determine the X matrix of the model. The option method in the proc glimmix statement specifies the estimation method. The PQL is obtained with the option method=RSPL and the MQL is obtained with the option method=RMPL.

Table 7.1 shows the GLMM results when we do not consider the random slope. Results of GLMM-Repeated Measures Analysis for epileptic patients

Table 7.1: GLMM with intercept only

Parameter	DF	Estimate	Std Error	t value	p-value
β_0	57	2.8750	0.1532	18.77	$< .0001$
β_1	234	-0.00103	0.2114	-0.0	0.9961
β_2	234	-0.3702	0.01797	-20.60	$< .0001$
β_3	234	-0.06424	0.2545	-2.52	0.0123

Table 7.2: Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
b_{i0}	Subject	0.6159	0.1201

(Thall and Vail, 1990) indicate that epileptic seizures did not significantly differ between treatment and control patients since the p-value for treatment is 0.4186. The interaction between treatment and period of visit was also significant, meaning that epileptic seizures did change over time or between control versus treatment patients. The time effect is significant meaning there is a significant variability in the rate of seizure occurrence with time independent of the treatment arm. The estimate of the time parameter is negative which means that an increase in time by one unit implies that the log of mean seizure rate declined by -0.3702 or the mean seizure rate is 0.6906 times that of the time before.

Since $\hat{\beta}_1 = -0.00103$ it follows that the hazard of epilepsy for those on treatment declined by 1.0010 and its 95% confidence interval is (0.5867;1.4154) and since 1 is included in the confidence interval the result that treatment has no effect is further confirmed.

The variance component estimate for model (7.12) is displayed in table 7.2. The variance component was estimated to be 0.6159 with its z Wald statistic of 5.12 which is significant at the 5% level. Thus we conclude that there is a significant individual to individual variability in slope.

The following SAS code was used to fit the GLMM with a random intercept and slope.

SAS code for penalized quasi likelihood (PQL) for GLMM with slope

```
Proc glimmix data=Titanic method=RSPL;
Class Subject Treatment;
model Count= Treatment week Treatment*week
/dist= Poisson link=log solution ;
random intercept week / subject=subject;
run;
```

Table 7.3: GLMM with intercept and slope

Parameter	Estimate	Std Error	t value	p-value
β_0	2.9253	0.1430	20.46	$< .0001$
β_1	0.06110	0.1971	0.31	0.7569
β_2	-0.4098	0.04159	-9.85	$< .0001$
β_3	-0.1356	0.5815	-2.33	0.0209

Table 7.4: Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
b_{i0}	Subject	0.5293	0.1059
b_{i1}	Subject	0.03589	0.009022

Table 7.3 shows the GLMM results when we consider the random slope. Just like with the GLMM model with the random intercept only results of GLMM with an intercept and slope indicate that epileptic seizures did not significantly differ between treatment and control patients since the p-value for treatment is 0.7569. The interaction between treatment and period of visit was significant, meaning that epileptic seizures evolution over time did significantly differ between those on treatment and the control treatment. The time effect is significant in this model and it was also significant under the random intercept only model.

Table 7.4 gives the results for a GLMM with both the random intercept and slope effects. Compared to the results with only the intercept as random (Table 7.1) we see that we come to the same conclusion. Thus accounting for variability in time via random slope renders the same conclusions. Variance component for random slope is estimated as 0.03589 with a standard error of 0.009072 which gives an approximate z-value (Wald) of 3.96 which is clearly significant at 5% level. Thus we conclude that there is a significant

individual to individual variability in slope as was depicted by the individual profile plots in chapter 3. The two variance components under model (7.13) are shown in Table 7.3 which are both significant based on the Wald statistic, z=5.00 and z=3.96.

The following SAS code was used to fit the GLMM with a random intercept and slope.

SAS code for marginal quasi likelihood (MQL)for GLMM with slope
Proc glimmix data=Titanic method=RMPL;
Class Subject Treatment;
*model Count= Treatment week Treatment*week*
/dist= Poisson link=log solution ;
random intercept week / subject=subject;
run;

Table 7.5: GLMM with intercept and slope

Parameter	Estimate	Std Error	t value	p-value
β_0	3.1993	0.2016	15.87	$< .0001$
β_1	0.02275	0.2781	0.08	0.9349
β_2	-0.3691	0.08621	-4.28	$< .0001$
β_3	-0.04568	0.1193	-0.38	0.7022

Table 7.6: Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
b_{i0}	Subject	1.1013	0.2130
b_{i1}	Subject	0.1970	0.03929

Table 7.5 shows the GLMM results when we consider the random slope using the marginal quasi-likelihood (MQL) estimation method. Just like with the GLMM model with the random intercept and slope using the PQL method of estimation the results of GLMM with a intercept and slope using MQL method indicate that epileptic seizures did not significantly differ between treatment and control patients since the p-value for treatment is 0.9349. The interaction between treatment and period of visit was not significant, meaning that epileptic seizures evolution over time did not signif-

icantly differ between those on treatment and the control treatment. The time effect is significant in this model and it was also significant under the PQL approach.

Table 7.5 gives the results for a GLMM with both the random intercept and slope effects. Compared to the results with the PQL approach (Table 7.3) we see that now the interaction effect is no longer significant. Thus accounting for variability in time via random slope renders the same conclusions. Variance components estimates are interpreted in the same way the where interpreted under the PQL approach.

This example shows that the significance of model terms can depend on the structure of the random effects. Thus, one must decide upon a reasonable model for the random effects as well as for the fixed effects. A commonly recommended approach for this is to perform a sequential procedure for model selection. First, one includes all possible covariates of interest into the model and selects between the possible models of random effects using likelihood ratio tests and model fit criteria. Then, once a reasonable random effects structure is selected, one trims model covariates in the usual way.

Chapter 8

Multiple Events per subject

8.1 Stochastic Process

8.1.1 Introduction

According to Ross (1989) and also in Feller (1957) a stochastic process $X(t), t \in T$ is a collection of random variables. That is, for each $t \in T$, X(t) is a random variable. The index t is often interpreted as time. For example, X(t) might equal the total number of customers that have entered a bookshop by time t; or the number of disease episodes by time t; or the total amount of sales that have been recorded in the market by time t, etc. Sometimes a stochastic process is also called a random process. In this regard X(t) is a counting variable.
The following are some of the most important stochastic processes:

- Poisson process
- Markov process in continuous time or discrete time (Markov Chain)
- Renewal process.

Every stochastic process is associated with an index set denoted T in most literature. A discrete-time process is a stochastic process with a set T, of which T is a countable set. On the other hand if T is an interval on the real line, the stochastic process is called a continuous-time process.

The state space of a stochastic process is defined as the set of all possible values that the random variables X(t) can assume.

In a nutshell one could say a stochastic process is a family of random variables that describe the evolution through time of some physical process or phenomenon.

8.1.2 Counting Process

A stochastic process $N(t), t \geq 0$ is said to be a counting process if N(t) represents the total "events" that have occurred up to and inlcuding time t.
From the above definition one can infer that a counting process N(t) must satisfy:

- $N(t) \geq 0$.
- N(t) is integer valued.
- If s<t, then N(s)≤ N(t), that is N(t) is a monotonically increasing function or conforms to a natural ordering in the non-negative integer space.
- For s<t, N(t) - N(s) equals the number of events that have occurred in the interval (s, t].

A counting process is said to possess independent increments if the numbers of events which occur in disjoint time intervals are independent (Ross, 1989). For example suppose N(0)= 0 then independent increment means:

$$P(N(t_1) - N(t_0) = u | N(t_0) = v) = P(N(t_1) - N(t_0) = u).$$

since the intervals $(0, t_0]$ and $(t_0, t_1]$ are disjoint. A counting process is said to possess stationary increments if the distribution of the number of events which occur in any interval of time depends only on the length of the interval (Ross, 1989). This is also known as the time homogeneity property.

8.1.3 Poisson Process

The Poisson process was named after the French mathematician Siméon-Denis Poisson (1781 - 1840).

A counting process $N(t), t \geq 0$ is said to be a Poisson process having rate $\lambda > 0$, if:

- $N(0) = 0$. (This means that the counting of events starts at t=0.)
- The process has independent increments.

- The number of events in any interval of length t is Poisson distributed with mean λt. That is, for all s, $t \geq 0$

$$P[N(t+s) - N(s) = n] = e^{-\lambda t} \frac{(\lambda t)^n}{n!},\ n = 0, 1, \ldots$$

From the third condition one can conclude that the Poisson process has stationary increments and that

$$E[N(t)] = \lambda t$$

of which this explains why λ is called the event rate of the process.

General characteristics of the Poisson process

In its most general form, we require these two conditions for a stochastic process to be a Poisson process are [4]:

- Orderliness: which roughly means for small Δt

$$P[N(t+\Delta t) - N(t) > 1 | N(t+\Delta t) - N(t) \geq 1] = 0$$

  which implies that arrivals don't occur simultaneously (but this is actually a mathematically-stronger statement).

- Memorylessness (also called evolution without after-effects): the number of arrivals occurring in any bounded interval of time after time t is independent of the number of arrivals occurring before time t.

Some real examples of Poisson processes include the number of telephone calls arriving at a switchboard per hour, the number of epileptic attacks in a person in a given interval of time and customer arrival to a queue in a service facility.

8.1.4 Generalizations of the Poisson Process

Further generalizations of the Poisson process include the non-homogeneous Poisson process and the Spatial Poisson process. These are briefly described below.

Non-homogeneous Poisson process

This counting process is sometimes referred to as the non-stationary Poisson process if the event rate (parameter) λ is a function of time t. This is because in general, the rate parameter may not be constant over time. Hence in this case the generalized rate function is given as $\lambda(t)$. Thus the expected number of events between time a and time b is

$$\lambda_{a,b} = \int_a^b \lambda(t)dt.$$

Thus, the number of arrivals in the time interval (a,b], given as N(b)- N (a), follows a Poisson distribution with associated parameter $\lambda_{a,b}$

$$P[N(b) - N(a) = n] = e^{-\lambda_{a,b}} \frac{(\lambda_{a,b})^n}{n!}, \quad \text{n=0,1, } \ldots$$

This generalized form is clearly more realistic because very rarely do real processes in real time such as epileptic attacks behave homogeneously due to individual specific changing characteristics or changing levels of care in general.

Spatial Poisson process

The spatial Poisson process is a type of non-homogeneous process that introduces a spatial dependence on the rate function and is given as $\lambda(x, t)$ where $x \in D$ for some vector space V (e.g. $\mathbf{R}^2 or \mathbf{R}^3$). For any set $S \subset V$(e.g. a spatial region) with finite measure, the number of events occurring inside this region can be modelled as a Poisson process with associated rate function $\lambda_s(t)$ such that

$$\lambda_s(t) = \int_S \lambda(x, t)dx.$$

In the special case where this generalized rate function is a separable function of time and space, we have:

$$\lambda(x, t) = f(x)\lambda(t)$$

for some function $f(x)$. Without loss of generality, let

$$\int_V f(x)dx = 1$$

otherwise we may scale $f(x)$ and $\lambda(t)$ appropriately. Now $f(x)$, represents the spatial probability density function of these random events in the following sense. The act of sampling this spatial Poisson process is equivalent to sampling a Poisson process with rate function $\lambda(t)$, and associating with

each event a random vector x sampled from the probability density function $f(x)$. A similar result can be shown for the general (non-separable) case. The above two dimensional generalization is a reasonable model to capture the occurrence of disease events in space and time. Of particular applications are problems involving diseases which depend on a point source of exposure which is spatially distributed. The epilepsy problem in the context of the current study may not deserve such an analysis. The only discernable heterogeneity is where the epileptic attack rate in an individual is a function due to for example changing treatment modes.

8.2 Poisson Point Process

Poisson point processes have two aspects: the counting process, which follows the number of events in a fixed time interval, and the interval process, which deals with the time intervals between subsequent events.

In most stochastic processes the events being studied happen at particular points in time for example times of epilepsy attacks in a person. In this case counting or Poisson point processes form a rich class of models to deal with such data.

Let N(t) be the number of events up to time *t, dt* a small time interval, and *o(dt)* a quantity of much smaller magnitude than *dt* in the sense $\frac{o(dt)}{dt} \rightarrow 0$ as $dt \rightarrow 0$. The function N(t) captures the point process and we say N(t) is a Poisson point process with *intensity* $\lambda(t)$ if

$$N(t+dt) - N(dt) = \begin{cases} 1 & \text{with probability } \lambda(\text{t})\text{dt} \\ 0 & \text{with probability } 1 - \lambda(\text{t})\text{dt} \\ > 1 & \text{with probability } 0(dt) \end{cases}$$

and $N(t+dt) - N(t)$ is independent of N(u) for $u < t$; the latter is called the independent increment property, which is a continuous version of the concept of independent trials.

The intensity function $\lambda(t)$ in a Poisson point process is really a generalization of the hazard function for time to event or survival analysis data. For this reason we briefly review how to estimate the hazard function assuming a constant hazard model that is assuming the survival times are exponentially distributed. The Cox proportional hazard and the Cox partial likelihood models to deal with measured covariates are also discussed in the subsequent

sections (Cox, 1972, 1975). A further advantage of the Poisson process models is that they are more general than survival models because they allow multiple end points per subject. For this reason the current chapter is also dedicated to dealing with multiple events per subject using the epilepsy data, the focus in this project.

8.3 Censored Data in Survival Analysis

According to Lee (1992) many researchers consider survival analysis to be merely the application of two conventional statistical methods to a special type of problem: parametric if the distribution is known and nonparametric if the distribution is unknown. This assumption would be true if the survival times of all the subjects were exact and known. However in most reliability studies or clinical trials some survival times are not known. One of the reasons being that at times it is not possible to wait for all experimental units to reach their end points simply because every study has a specific study time determined a priori. Thus there is a need for statistical techniques which handle such incompleteness. One of such techniques is discussed in this section. We are going to use the following example to illustrate the technique.

Example 8.1: Two groups of rats were exposed to carcinogenic DBMA, and the number of days to death due to cancer was recorded (Kalbfleisch and Prentice, 1980)

Group 1 : 143, 164, 188, 190, 192, 206, 209, 213, 216, 220, 227, 230,234, 246, 265, 304, 216+, 244+
Group 2 : 142, 156, 163, 198, 205, 232, 232, 233, 233, 233, 233,239, 240, 261, 280, 296, 296, 323, 204+, 344+

The censoring times in the two groups of rats carcinogen example (Kalbfleisch and Prentice, 1980) are said to be right censored. In other words the individual is known to be event free up to time y_i but the actual event time is $t_i > y_i$. There is also a possibility of left censored times. A good example is age at infection for a certain disease. If an individual is tested positive for HIV at age y_i then clearly age at infection is $a_i < y_i$ hence we say that the infection age is left censored. Interval censored data occurs when the actual event time lies between two known time points. For example in follow up studies an individual may be disease free at time t_0 and disease positive at time t_1. Then the actual time of infection is t' where $t_0 < t' \leq t_1$. If an individual did not test positive at known age y_i then clearly the age at

infection is age $a_i > y_i$ thus the age at infection in this case is right censored.

In this example four rats were 'censored' at times 216, 244, 204 and 344; i.e. known not to have died of cancer by those times. Possible reasons for censoring are

- Deaths due to other causes,
- Being alive when the study ends

There are three possible ways (assumptions) of modelling such data:

- ignore the censoring information, i.e. treat all the data as if they are genuine deaths
- drop the censored cases, so we are dealing with genuine deaths
- model the censored data properly

The first two methods can be very biased and misleading if the censoring patterns in the groups differ. The second method is inefficient even if the censoring patterns in the two groups are not similar. With a correct model, the last method is potentially the best as it would take into account whatever information is available in the censored data. However results from the third approach can be very misleading under a mis-specified model. Thus model identification is a key step in analyzing time to event data. On the other hand models which are less dependent on distributional assumptions may become handy at times leading to semi-parametric or fully non-parametric approaches.

In general censored data can be denoted as $(y_1, \delta_1), \ldots, (y_n, \delta_n)$, where δ_i is the last-known status or event indicator: $\delta_i = 1$ if y_i is a true event time, and zero otherwise. If t_i is the true lifetime of subject i, then $\delta_i = 0$ if and only if $t_i > y_i$ where t_i here denotes the unknown true event time for individual i. Our concern would be modelling the true lifetime t_i rather than the observed y_i, since censoring is usually a nuisance occurrence that does not have any substantive meaning, for example it can be determined by the the study design.

Suppose $t_1, \ldots, t_n$ are and iid sample from $f_\lambda(t)$. Let the cumulative density function (cdf) of T be $F(T)$ and $S(T) = 1 - F(T)$ then the function $S(T)$ is called the survivor or simply the survival function. The likelihood contribution of the observation (y_i, δ_i) is

$$L_i(\lambda) = P_\lambda(T_i > y_i) = s(y_i) \text{ if } \delta_i = 0$$

or

$$L_i(\lambda) = f_\lambda(y_i) \text{ if } \delta_i = 1.$$

The contribution $P_\lambda(T_i > y_i)$ is therefore given by $S(y_i)$ where $S(.)$ is the survival function. Thus the overall likelihood as a function of λ involving both contributions can be written as

$$\begin{aligned} L(\lambda) &= \prod_{1=1}^{n} L_i(\lambda) \\ &= \prod_{1=1}^{n} S(y_i)^{1-\delta_i} f_\lambda(y_i)^{\delta_i}. \end{aligned} \tag{8.1}$$

To illustrate this let us consider an exponential model, which despite its simplicity is used regularly in survival studies, defined by

$$\begin{aligned} f_\lambda(t) &= \lambda e^{-\lambda t} \\ P_\lambda(T > t) &= e^{-\lambda t}. \end{aligned}$$

Hence we have the likelihood function as a result of applying equation (8.1)

$$L(\lambda) = (\lambda)^{\sum \delta_i} \exp(-\lambda \sum y_i).$$

Upon taking the derivative of the log-likelihood we get the score function to be

$$U(\lambda) = \frac{\partial}{\partial \lambda} \ln L(\lambda) = \frac{\sum \delta_i}{\lambda} - \sum y_i$$

and setting it to zero, we get

$$\hat{\lambda} = \frac{\sum \delta_i}{\sum y_i}$$

In the above equations one can note that $\sum y_i$ is the total observation time (person days) including both the censored and uncensored cases, while $\sum \delta_i$ is the number of events. In this case $\hat{\lambda}$ is the sample hazard rate given by the rate of event occurrence per unit time per person. In epidemiology such a quantity would estimate the incidence rate or equivalently the force of infection (Mwambi, et al., 2009) of a disease.

With some algebra the observed Fisher information of λ is

$$I(\hat{\lambda}) = \frac{\sum \delta_i}{\hat{\lambda}^2},$$

so that the standard error of $\hat{\lambda}$ is

$$se(\hat{\lambda}) = \frac{\hat{\lambda}}{\left(\sum \delta_i\right)^{\frac{1}{2}}}.$$

Example 8.1: continued. Assuming an exponential model for excess life over 100 days (in principle this cutoff can be set to any unknown parameter), so for Group 1 we get n=19, $\sum y_i = 2195$, $\sum \delta_i = 17$ and

$$L(\lambda_1) = (\lambda)^{17} exp(-2195\lambda).$$

which yields $\hat{\lambda}_1 = \frac{17}{2195} = 0.00774$ (se=0.00188). Equivalently $\hat{\lambda}_2 = \frac{19}{2923} =$ 00.00650 (se=0.00149).The approximate 95% intervals are (0.004063,0.011427) and (0.003577,0.009423) respectively. Since the two intervals overlap we conclude that the death rates in the two groups of rats are not significantly different. Alternatively a Wald statistic can be calculated to test the same hypothesis. A similar approach was used in Mwambi et al (2009) using longitudinal binary disease outcomes for a respiratory disease affecting children within the age of one year to estimate the force of infection and recovery rates of the disease. Thus the above analysis although fairly basic is applicable to relatively complex problems in practice. In Mwambi et al. (2009) piecewise constant monthly rates were estimated to capture the time varying occurrence of the respiratory events in the cohort of children.

8.4 Cox Proportional Hazards Model

The model was first introduced by Cox (1972) and has come to be known as the Cox regression model. It is also sometimes described simply as proportional hazards regression. It is probably the most widespread model used in survival analysis.

According to the Cox proportional hazard model, the failure rate of a system is affected not only by its operation time, but also by the covariates under which it operates. For example, a unit may have been tested under a combination of different accelerated stresses such as humidity, temperature, voltage, etc. It is clear then that such factors affect the failure rate of a unit.

The instantaneous failure rate (or hazard rate) of a unit as a function of time only is given by:

$$h(t) = \frac{f(t)}{S(t)} \tag{8.2}$$

where:

- f(t) is the probability density function, of the event time T.
- S(t) is the survival function, i.e P(T > t) the probability of failure occurring after time t or the probability of surviving up to time t.

To accommodate the fact that the failure rate of a unit being dependent not only on time but also on other covariates, the above equation must be modified in order to be a function of time and of the covariates.

The proportional hazards model assumes that the failure rate (hazard rate) of a unit is the product of two components stated below

- an arbitrary and unspecified baseline failure rate, $h_0(t)$, which is a function of time only.
- a positive function $\psi(x, \beta)$, independent of time, which incorporates the effects of a number of covariates such as humidity, temperature, pressure, voltage, etc.

Thus the hazard (failure) rate of a unit is then finally given by:

$$h(t, \mathbf{x}, \beta) = h_0(t)\psi(\mathbf{x}, \beta) \tag{8.3}$$

where:

- $\mathbf{x}$ is a column vector consisting of the measurable covariates
- β is a column vector consisting of the unknown parameters (also called regression parameters) of the model.

It can be assumed that the form of $\psi(\mathbf{x}, \beta)$ is known and $h_0(t)$ is unspecified. Different forms of $\psi(\mathbf{x}, \beta)$ can be used. However, the exponential form due to Cox (1972,1975) is mostly used which ensures the hazard as a function of the measured covariates maintains positivity which is given by:

$$\psi(\mathbf{x}, \beta) = e^{\mathbf{x}'\beta} \tag{8.4}$$

The failure rate can then be written as:

$$h(t, \mathbf{x}, \beta) = h_0(t)e^{\mathbf{x}'\beta}$$

The form of dependence on covariates also ensures as stated above that the hazard is always positive in order to make practical sense.

A remarkable property of the model that avoids the need to specify $h_0(t)$ can be shown as follows. If lifetimes T_1 and T_2 have proportional hazards say

$$h_i(t) = h_0(t)\eta_i$$

for $i = 1, 2$, respectively, then

$$P(T_1 < T_2) = \frac{\eta_1}{\eta_1 + \eta 2} \tag{8.5}$$

regardless of the shape of the baseline hazard function. Here η_i can be defined as the risk score for unit $i = 1, 2$ entering as a function of individual specific covariates. The result in equation (8.5) is also similar to the case of two independent exponential random variables with parameters λ_1 and λ_2 stating that

$$P(X_1 < X_2) = \frac{\lambda_1}{\lambda_1 + \lambda 2}$$

The intervals between successive death times convey no information about the effect of the explanatory variables on the hazard function of death, since the baseline hazard function has an arbitrary form.

So we could have $h_0(t)$ and $h(t)$ equal to zero in those time intervals in which there are no deaths. In other words only the ordered statistic of times of deaths is required.

It is sufficient to consider the probability that that the j^{th} individual who dies at t_j, conditional on t_j being one of the observed set of death time $t_1, t_2, \ldots, t_r$. Where we assume that the data available has n individuals amongst whom there are r distinct death times and $n - r$ right censored survival times. We further assume that there are no tied death times and that the r deaths are ordered, that is $t_1 < t_2 < \ldots < t_r$, so that t_j is the j^{th} ordered death time.

If the vector of explanatory variables for the individual who dies at t_j is $\mathbf{X}_j$ then extending the result stated in equation (8.5) we deduce that

$$\begin{aligned}
&P[\text{individual with } \mathbf{X}_\text{j} \text{ dies at } \text{t}_\text{j} | \text{One death at } \text{t}_\text{j}] \\
&= \frac{P[\text{individual with } \mathbf{X}_\text{j} \text{ dies at } \text{t}_\text{j}]}{P[\text{One death at } \text{t}_\text{j}]} \\
&= \frac{h_i(t_j)}{\sum\limits_{l \in R_{t_j}} h_i(t_j)} \qquad (8.6) \\
&= \frac{e^{X'_j \beta}}{\sum\limits_{l \in R_{t_j}} e^{X'_j \beta}}
\end{aligned}$$

Taking the product of the conditional probabilities over the r death times we have:

$$P(T_{i_1} < T_{i_2} < \ldots < T_{i_r}) = \prod_{j=1}^{r} \frac{e^{X_j'\beta}}{\sum\limits_{l \in R_{t_j}} e^{X_j'\beta}} \tag{8.7}$$

where R_{t_j} is the number at risk just before the death at time t_j also known as the risk set at time t_j. Expression (8.7) is what is popularly known as the Cox partial likelihood (CPL) because (i) the baseline hazard need not be explicitly specified and (ii) no distributional assumption is made about event times (Cox, 1975). The only problem with CPL is that analytical maximization of the log-likelihood to estimate the vector of parameters β is complex. However most statistical computing system such as SAS and Genstat have numerical procedures for estimating the parameters such as proc PHREG in SAS and RPHFIT in Genstat.

8.5 Replicated Poisson Process

In this section methods suitable for modelling disease processes such as epilepsy data will be developed in increasing order of complexity, pointing out the benefits and deficiencies of each method in a progressive manner. The occurrence of epilepsy events is modelled as a replicated Poisson process. Table 3.1 shows a data set from a study of treatment of epilepsy, where the patients were randomized to either active or placebo groups. Because of staggered entry to the study, patients have different follow up periods. The patients' or their care givers were asked to record the time of epileptic attacks during follow up. Note that each individual sequence of events is a Poisson process. The parameter of interest in these methods is the event rate or hazard rate in the context of survival analysis.

8.5.1 Method One

This method will only make use of the different follow-up periods among patients, but will make no use of the times of attacks hence it can not be generalized if one needs to consider more covariates.

The following assumptions are going to be made:

- Event occurrence within each patient follow a Poisson point process.
- The rate or intensity is constant over time.

- λ_a and λ_b will be the rate of attacks in the active and placebo groups respectively.
- The aim is to compare λ_a and λ_b from the repeated measurement or longitudinal data.

Let n_i denote the total number of attacks for patient i while y_a and y_p denote the total number of events in the active and placebo treatment groups. Thus $y_a = \sum_a n_i$ and $y_p = \sum_p n_i$ where $\sum_a$ and $\sum_p$ means summations over the active and placebo groups respectively. Thus following the assumptions above we can infer the following distributional assumptions.

$$\begin{aligned} n_i &\sim Poisson(T_i\lambda) \quad \text{(general)} && (8.8)\\ y_a &\sim Poisson(\sum_a T_i\lambda_a) \quad \text{(active)} && (8.9)\\ y_p &\sim Poisson(\sum_p T_i\lambda_p) \quad \text{(placebo)}, && (8.10) \end{aligned}$$

using the known result that the sum of Poisson random variables is itself Poisson distributed. The parameter λ in the first equation is either λ_a or λ_p depending on whether individual i is from the active or placebo groups. The parameter of interest here is

$$\theta = \frac{\lambda_a}{\lambda_p} \tag{8.11}$$

Conditional on $y_a + y_b$ it can be shown that y_a is Binomially distributed

Table 8.1: Table of Summaries

	Active	Placebo
y_a or y_p	29	71
$\sum T_i$	97	109

with probability of success given by

$$p(\theta) = \frac{\sum_a T_i\lambda_a}{\sum_a T_i\lambda_a + \sum_p T_i\lambda_p} = \frac{\theta\sum_a T_i}{\theta\sum_a T_i + \sum_p T_i}$$

thus the conditional likelihood for θ can be written as:

$$L(\theta) = (\tfrac{97\theta}{97\theta+109})^{29}(1 - \tfrac{97\theta}{97\theta+109})^{71},$$

since $\sum_a T_i = 97$ and $\sum_p T_i = 109$. Note that the constant terms which are independent of θ are omitted in the likelihood expression.

The maximum likelihood estimate (MLE) of θ is calculated by maximising the function

$$\ln L(\theta) = 29\ln(97\theta) - 29ln(97\theta + 109) + 71ln\Big(1 - \frac{97\theta}{97\theta + 109}\Big) \qquad (8.12)$$

and since

$$\frac{\partial \ln L(\theta)}{\partial \theta} = \frac{29}{\theta} - 100\Big(\frac{97}{97\theta + 109}\Big)$$

equating $\frac{\partial ln L(\theta)}{\partial \theta}$ to zero the equation that yields the maximum likelihood estimate of θ is:

$$\frac{29}{\hat{\theta}} = 100(\frac{97}{97\hat{\theta} + 109}) \qquad (8.13)$$

leading to

$$\hat{\theta} = \frac{109}{97(\frac{100}{29} - 1)} \quad \hat{\theta}$$

that is

$$= \frac{(\frac{29}{97})}{(\frac{71}{109})} = 0.459$$

The standard error is $se(\hat{\theta}) = 0.10$. The likelihood of the null hypothesis $H_0 : \theta = 1$ is very small, hence we conclude that the active treatment leads to fewer attacks of epilepsy. In other words the hazard of epilepsy for the treatment group is less than that for the placebo group. The 95% interval for $\hat{\theta}$ is (0.263,0.655). Clearly the interval does not include one which further confirms the result.

8.5.2 Method Two

This time we will use the Poisson regression which can easily accommodate some covariates. Just like method one this method will make no use of the information about event times. We will use the same assumptions we used in method one. Let $x_i = 1$ if the patient i belongs to the active treatment group and zero if patient i belongs to the placebo group. We will further assume that the number of attacks n_i is Poisson with mean event rate

$$\mu_i = T_i \exp(\beta_0 + \beta_i x_i)$$

of which the above equation is equivalent to the linear model

$$\ln \mu_i = \ln T_i + \beta_0 + \beta_i x_i.$$

with a log link function. This model can be fitted as a GLM for a Poisson distributed response where the term $logT_i$ is declared as an OFFSET. The parameter estimates obtained by using proc GENMOD in SAS are shown in Table 8.2 below. One can note that $e^{\hat{\beta}_1} = e^{-0.7787} = 0.46 = \hat{\theta}$ as computed by method one. The 95 % confidence for θ is

$$(\exp(\hat{\beta}_1 - 1.96 * s.e(\hat{\beta}_1)), \exp(\hat{\beta}_1 + 1.96 * s.e(\hat{\beta}_1))) = (0.280, 0.707)$$

which is slightly shifted to the right compared to the direct likelihood one. It is important to note the similarity between the Cox proportional hazard

Table 8.2: Epilepsy data.

Effect	Parameter	Estimate	se	Chi-Square	p-value
Intercept	β_o	-0.4287	0.1187	13.05	0.0003
Treatment	β_1	-0.7787	0.2204	12.49	0.0004

model and the Poisson regression. In fact we can state the following result relating the two.

Theorem
The poisson regression is a special case of the Cox proportional hazards model.

Proof
Let the common event baseline rate be $\mu_0(t)$ in both groups. The

$$\log \mu_i(x = 1) = \log T_i + \log \mu_0(t) + \beta_0 + \beta_1$$

$$\log \mu_i(x = 0) = \log T_i + \log \mu_0(t) + \beta_0$$

Then log hazard ratio

$$\log \frac{\mu_i(x = 1)}{\mu_i(x = 0)} = \beta$$

which is independently baseline rate $\mu_0(t)$.

8.5.3 Method Three

Unlike the other two methods mentioned in the previous sections this method makes use of the times of attacks and it does not assume constant intensity for the Poisson processes for each individual, hence it can be generalized.

We assume that the attacks for a patient follow a Poisson point process with intensity $\lambda_x(t)$, where x is the covariate vector. We will use the following proportional intensity model

$$\lambda_x(t) = \lambda_0(t, \alpha)g(x, \beta), \tag{8.14}$$

where $\lambda_0(t, \alpha)$ is the baseline intensity function dependent on an unknown parameter α. The effect of covariate x is to modify the baseline intensity proportionally on a time independent function $g(x, \beta)$. The function $g(x, \beta)$ depends on β an unknown parameter which expresses the effect of the covariate on the intensity function, for example using the log linear or equivalently the multiplicatively Cox proportional like model gives rise to the following expression for $\lambda_x(t)$;

$$\lambda_x(t) = \lambda_0(t, \alpha)e^{x'\beta}.$$

The baseline intensity $\lambda_0(t, \alpha)$ requires a parameter α, which is like a nuisance parameter since it is not the parameter of interest.

Denoting the n_i recurrent event times of subject i by t_{i1}, ..., t_{in_i}, implies that the contribution of this subject to the likelihood is given by

$$L_i(\alpha, \beta) = e^{-\Lambda_{x_i}(T_i)} \prod_{j=1}^{n_i} \lambda_{x_i}(t_{ij}), \tag{8.15}$$

where

$$\begin{aligned} \Lambda_{x_i}(T_i) &= \int_0^{T_i} \lambda_{x_i}(t)dt \\ &= g(x_i, \beta) \int_0^{T_i} \lambda_0(t, \alpha)dt \\ &= g(x_i, \beta)\Lambda_0(T_i, \alpha). \end{aligned} \tag{8.16}$$

So

$$\begin{aligned} L_i(\alpha, \beta) &= e^{-g(x_i,\beta)\Lambda_0(T_i,\alpha)}\{g(x_i, \beta)\Lambda_0(T_i, \alpha)\}^{n_i} \prod_{j=1}^{n_i} \frac{\lambda_0(t_{ij}, \alpha)}{\Lambda_0(T_i, \alpha)} \\ &\equiv L_{1i}(\alpha, \beta)L_{2i}(\alpha) \end{aligned} \tag{8.17}$$

where

$$L_{1i}(\alpha,\beta) \;=\; e^{-g(x_i,\beta)\Lambda_0(T_i,\alpha)}\{g(x_i,\beta)\Lambda_0(T_i,\alpha)\}^{n_i}. \tag{8.18}$$

Expression (8.17) is obtained by using result in equation (8.16) in equation (8.15) and writing $\lambda_{x_i}(t_{ij})$ in the general form given in equation (8.14). The total likelihood from all, say m, individuals is

$$\begin{aligned} L(\alpha,\beta) &= \prod_{i=1}^{m} L_i(\alpha,\beta) \\ &= \prod_{i=1}^{m} L_{1i}(\alpha,\beta)\prod_{i=1}^{m} L_{2i}(\alpha) \\ &= L_1(\alpha,\beta)L_2(\alpha). \end{aligned} \tag{8.19}$$

Hence one can see that the information about β is contained only in the first term $L_1(\alpha,\beta)$; of which this is the likelihood based on the number of events N_i from each individual following the model

$$N_i \sim Poisson(\Lambda_0(T_i,\alpha)g(x_i,\beta)). \tag{8.20}$$

Note that method one in section (8.5.1) is obtained by assuming a constant intensity.

Since α is a nuisance parameter having to model $\Lambda_0(T_i,\alpha)$ is a nuisance, since it is not directly relevant to the question of interest namely to assess the treatment effects and comparisons. The data structure makes it difficult to specify an appropriate model for $\Lambda_0(T_i,\alpha)$; for example, we cannot simply plot the histogram of the event times because some of the event times are truncated times and not actual event times. The decomposition of the likelihood suggest the following possible method of estimation.

- Estimate α from $L_2(\alpha)$, which is a conditional likelihood given $n_i's$.This component is fully determined by the set of event times $t_{ij}'s$.
- Use $\hat{\alpha}$ to compute $\Lambda_0(T_i,\hat{\alpha})$.
- Estimate β in the Poisson regression based on the data (n_i,x_i) with $\Lambda_0(T_i,\hat{\alpha})$ as an offset term.

To avoid dealing with the nuisance parameter let us further assume that $T_i \equiv T$. We first need the result that if X_i for $i=1,\ldots,m$, are independent Poisson(λ_i), then the conditional distribution of $(X_1,\ldots,X_m)$ given $\sum X_i$ is multinomial with parameters $(\pi_1,\ldots,\pi_m)$, where $\pi_i = \frac{\lambda_i}{\sum_{j=1}^{m}\lambda_i}$. Applying this result to

$$N_i \sim Poisson(\Lambda_0(T_i, \alpha)g(x_i, \beta)).$$

$i = 1, 2, \ldots, m$ and letting $n = \sum_i n_i$, we now have

$$\begin{aligned} L_1(\alpha, \beta) &= P(N_1 = n_1, \ldots, N_m = n_m) \\ &= P(N_1 = n_1, \ldots, N_m = n_m | \sum N_i = n) P(\sum N_i = n) \\ &= \prod_{i=1}^{m} \Big(\frac{g(x_i, \beta)}{\sum_{j=1}^{m} g(x_i, \beta)} \Big)^{n_i} P(\sum N_i = n) \\ &\equiv L_{10}(\beta) L_{11}(\alpha, \beta), \end{aligned} \tag{8.21}$$

where

$$L_{10}(\beta) = \prod_{i=1}^{m} \left(\frac{g(x_i,\beta)}{\sum\limits_{j=1}^{m} g(x_j,\beta)} \right)^{n_i}.$$

If we use the common log linear model

$$\lambda_x(t) = \lambda_0(t, \alpha) e^{x'\beta},$$

then

$$L_{10}(\beta) = \prod_{i=1}^{m} \left(\frac{e^{x'_i\beta}}{\sum\limits_{j=1}^{m} e^{x'_j\beta}} \right)^{n_i},$$

which is exactly the Cox partial likelihood for this particular setup.

In general when $T_i \neq T$, and assuming a Poisson process with proportional intensity, the Cox partial likelihood is defined as follows. One should note that this formulation allows the covariate to change over time. Let t_{ij} denote the j^{th} event of subject i, x_{ij} the associated covariate vector and R_{ij} the set of subjects still at risk or the risk set at time t_{ij}.
Then

$$L_{10}(\beta) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} \Big(\frac{e^{x'_{ij}\beta}}{\sum\limits_{k \in R_{ij}} e^{x'_{kj}\beta}} \Big).$$

To apply this approach to the epilepsy data we first need to consider the approximate Cox partial likelihood, first assuming the the follow-up period T_i's are the same for all subjects. Let the covariate $x_i = 1$ if i belongs to the active therapy group and $x_i = 0$ otherwise. Then

$$L_{10}(\beta) = \prod_{i=1}^{m} \left(\frac{e^{x'_{ij}\beta}}{\sum\limits_{j=1}^{m} e^{x'_{kj}\beta}} \right)^{n_i},$$

where $e^{x'_i\beta} = \theta$ if $x_i = 1$ and $e^{x'_i\beta} = 1$ otherwise; then $\sum_{j=1}^{m} e^{x'_{kj}\beta} = 10\theta + 12$, so

$$L_{10}(\beta) = (\tfrac{\theta}{10\theta+12})^{y_a}(\tfrac{1}{10\theta+12})^{y_p},$$

The maximum likelihood estimate of θ is calculated as follows:

$$\begin{aligned} lnL_{10}(\beta) &= y_a ln\theta - y_a ln(10\theta+12) - y_p ln(10\theta+12) \\ \tfrac{\partial lnL_{10}(\beta)}{\partial\theta} &= \frac{y_a}{\theta} - \frac{10y_p + 10y_a}{10\theta+12} \\ \tfrac{y_a}{\hat{\theta}} &= \frac{10y_p + 10y_a}{10\hat{\theta}+12} \\ \hat{\theta}(10y_p + 10y_a) &= (10\hat{\theta}+12)y_a \\ \hat{\theta} &= \frac{12y_a}{10y_p} \\ \hat{\theta} &= 0.490. \end{aligned} \tag{8.22}$$

The standard error of $\hat{\theta}$ is 0.103 and the 95% confidence interval of $\hat{\theta}$ is (0.288;0.692). Since 1 is not in the interval we can conclude that the active treatment has led to fewer epilepsy attacks.

Comparing the three methods

Table 8.3 is a summary of the estimates of θ from the three different methods. Method one and two give the same estimate of $\hat{\theta}$, whereas the third method gives a slightly different estimate, this is because in method three we assumed that the patients were followed for the same number of weeks ($T_i = T$) of which this was not the case. The confidence intervals are closer for method two and three. Methods one and three seem to be less conservative than method two which explicitly controlled for variable exposure times by using $logT_i$ as an offset in the regression model. Thus the smaller standard errors in methods one and three may arise due to the assumption $T_i = T$ of which this was not the case.

8.6 Recurrent Events

Recurrent events occur frequently in longitudinal studies involving multiple subjects. Some examples are seizures in epileptic patients , successive tumors in cancer studies, attacks by opportunistic infections for HIV/AIDS patients

Table 8.3: The following table compares the estimates of $\hat{\theta}$ obtained from the three methods mentioned earlier.

Method	$\hat{\theta}$	se	95% Confidence interval
Method One	0.459	0.10	(0.263,0.655)
Method Two	0.46	0.1187	(0.280;0.707)
Method Three	0.490	0.103	(0.288,0.692)

and many more. Treatment strategies such as HAART (Highly Active Anti-Retroviral Therapy) are being implemented in order to reduce the intensity of such attacks. However models for recurrent events or multiple events in a subject are necessary in order to better compare different treatment or intervention strategies in such circumstances.

Models for recurrent events are usually discussed through one of the following three related functions namely,

- their intensity functions
- their hazard functions
- or their cumulative mean function.

For example if one chooses to use the intensity function approach then the kind of approach that can be used is the Andersen and Gill (1982) model which is discussed in section 8.7.

If we are dealing with multiple events in the presence of death one might want to model both events and death following appropriate suggestions by Ghosh (2000). Ghosh (2000) gives an overview on how the models for recurrent events can be modified so that they can incorporate death into the analysis. Here as expected death is treated as an absorbing state as in the Markov chain formulation (Feller, 1957).

For the analysis of multiple events with observation gaps Zhao and Sun (2006) discuss the cumulative mean function and regression analysis approaches for analysis.

Sturmer (2000) shows how the logistic, Poisson, and two different Cox proportional hazards regression approaches for the analysis of multiple events

can be applied in the field of epidemiology.

The following sections focus on discussing methods for the analysis of recurrent events where the main focus is on the three widely used variance-correction models: the "independent increments" model developed by Anderson and Gill (1982), the marginal risk set model of Wei et al.(1989), and the conditional risk-set model (Prentice et al.,1981), which may be estimated in either total or gap time. All are related in that, compared to the standard Cox model, they use the non-independence caused by repeated events to empirically correct the standard error estimates. The key distinction among these models is "the way that the risk sets are defined at each failure" (Cleves, 1999). The risk set defines which observation may fail at a particular time; as a result of the different risk-set definitions, very different processes are modelled by the three alternatives. Thus the estimated coefficients will vary among these three variance corrected models because of the different underlying assumptions.

8.7 Andersen-Gill model

The simplest variance-corrected model is that of Andersen and and Gill (1982) (hereafter AG). The key characteristic of the AG model is "the assumption that the risk of an event for a given subject is unaffected by any earlier events that occurred to the same subject, unless terms that capture such dependence are included explicitly in the model as covariates" (Oakes, 1992). That is, multiple events for any particular observation are assumed to be conditionally independent; for this reason, the AG model is often referred to as the "independent increment" model. When events are not independent, robust variance estimates allowing for clustering within units may be used (for this see for example, White, 1980).

When the time scale is duration since entry (exposure), the intensity process for the ith subject is

$$Y_i(t)\lambda_0(t)exp(X_i(t)\beta) \tag{8.23}$$

In practical terms, the Cox and AG models are essentially indistinguishable, and in fact the former can be shown to be a special case of the latter. That is in the definition of the indicator variable $Y_i(t)$ for survival data is that the individual ceases to be at risk when an event occurs and Y_i goes to zero, but for the AG model for recurrent events, $Y_i(t)$ remains one as events occur. Thus, while the AG approach is straightforward to estimate, the assumption

of independent increments is strong, particularly if the ordering of events may be important. Also, unlike the other models considered here, the AG model restricts the baseline hazard rate for all events to be the same. For many applied problems, the assumption of independent increments will not be acceptable, at least not without empirical testing.

8.8 Wei, Lin and Weissfeld model

The marginal risk-set approach of Wei, Lin, and Weissfeld (1989) (hereafter WLW) applies the traditional competing risks set up for multiple events to repeated events. Ordered events data are treated as if they presented a typical competing risks problem: each individual is "at risk" for the first, second, third, etc., event from the beginning of the study period. The data are then stratified by event number, and separate baseline hazards are estimated at the first occurrence of the event under study, the second, etc. The approach is thus referred to as the "marginal risk set" model because, within these event defined strata, marginal data are used. As a result, at any point in time, all individuals that have not yet experienced k events are assumed to be "at risk" for the kth event.

The intensity or hazard function for the jth event for the ith subject is:

$$Y_{ij}(t)\lambda_{0j}(t)exp(X_i(t)\beta_j) \tag{8.24}$$

Unlike the AG model, stratification by event allows baseline hazards for each event to differ; as in the AG model, however, covariate effects are assumed to be constant across event ranks. The signature characteristic of the WLW approach is that all observations or individuals are at risk for all events at all times prior to experiencing that event. That is, in the case of repeated events of the same type, the "fifth" event can (in theory) occur at any time, even prior to the "first," "second," etc. events.

8.9 Prentice, Williams and Peterson model

In the conditional model of Prentice, Williams, and Peterson (1981) (hereafter PWP), an observation is not at risk for a later event until all prior events have already occurred. That is gap times between recurrent events are modelled equivalently and the data are stratified by the number of previous events.

Accordingly, the "risk set" at time t for the kth occurrence of an event is limited to those observations under study at time t who have already experienced $k-1$ events of that type. As in the WLW model, estimates are then stratified by event rank, so that the different events have varying baseline hazards. As in the previous models, however, covariate effects are again assumed to be constant across strata, though as in the WLW model strata-by-covariate interactions may be estimated. An additional feature of the conditional risks model is that the model may be estimated in either total time (i.e., time from each unit's entry into the observation set) or in gap-time (also referred to as "interevent time "), defined as the duration since the previous event.

8.10 Comparing the three methods

The similarities and differences of the variance-corrected models are summarized in table 8.4 below. All three models use robust variance estimates (Wei et al; 1989) to address the potential for interdependence due to repeated events. Robust standard errors assume that observations are independent across units (or "clusters") but not necessarily within those units. (For more information on the robust variance computation see Therneau and Grambsch, 2000.)

8.11 Fitting the three models

For us to be able to fit the models we need to create appropriate data sets. Creation of appropriate data sets is done using the counting process pioneered by Andersen and Gill (1982). We will illustrate this using the following extract of the data set being modelled as shown in table 8.5, where we assume that subject 19 experiences the maximum number of attacks.

AG model data setup

As an example, let us use the response values for subject 9 (with time-independent indicator covariates treatment = 1 for the active treatment group and 0 for the placebo group) who experiences an event on week 0.1, 2 and 3.2 and has now been followed to week 11. This subject would be coded as having contributed four observations or "lines" of data whose intervals are (0,0.1], (0.1,2], (2,3.2], (3.2,11] with corresponding exit status codes of 1, 1, 1 and 0. The data file for this subject is shown below in Table 8.6 where the pair of variables Tstart, Tstop define the time exposure interval or the risk interval.

Table 8.4: Comparison of Variance-Correction Models for Repeated Events

	Model Properties			
	Risk set for Event k at Time t	Time Scale	Robust Standard Errors	Stratification by Event
AG	Independent Events	Duration since starting observation	Yes	No
WLW	All subjects that haven't experienced event k at Time t	Duration since starting observation	Yes	Yes
PWP Total time	All subjects that have experienced event $k-1$ and haven't experienced event k, at Time t	Duration since starting observation	Yes	Yes
PWP Gap time	All subjects that have experienced event $k-1$ and haven't experienced event k, at Time t	Duration since previous event	Yes	Yes

Table 8.5: Data being used to illustrate how to create appropriate data sets

Subject i	Treatment	T_i	n_i	Time of events
8	active	8	3	0.2 3.2 7.7
9	active	11	3	0.1 2 3.2
10	active	8	3	0.1 3.2 3.7
11	placebo	11	4	2.3 7.9 8 8.8
19	placebo	7	4	0.9 2.2 5.2 6.6
20	placebo	4	2	2.2 3.2

The following gives the SAS code that uses the above data in the PHREG procedure, where Status(0) indicates that an event of interest has not occurred at that exit time, and that the subject is still at risk for the event(s) of interest at that time. SAS assumes that the other exit status values provided in the data set are the event(s) of interest.

Table 8.6: AG model data

Subject	Tstart	Tstop	Status	Treatment
9	0	0.1	1	1
9	0.1	2	1	1
9	2	3.2	1	1
9	3.2	11	0	1

proc phreg;
*model (Tstart, Tstop) * Status(0) = Treatment;*
run;

WLW model

The WLW data set contains 5 lines of data for each subject since the maximum number of events recorded for any one of the participants is 4. As an example, let us use the response values for subject 9 (with time-independent indicator covariates treatment = 1 for the active treatment group and 0 for the placebo group) who experiences an event on week 0.1, 2 and 3.2 and has now been followed to week 11. This subject would be coded as four observations or "lines" of data whose intervals are (0,0.1], (0.1,2], (2,3.2], (3.2,11] with corresponding exit status codes of 1, 1, 1 and 0. The data file for this subject is shown below in Table 8.7 where the variable CT defines the cumulative time since exposure (origin) and enum is the variable to index the multiple events :

The following gives the SAS code that uses the above data in the PHREG

Table 8.7: WLW model data

Subject	CT	Status	Treatment	Enum
9	0.1	1	1	1
9	2	1	1	2
9	3.2	1	1	3
9	11	0	1	4
9	11	0	1	5

procedure to fit the the WLW model:

```
proc phreg covs(aggregate);
model CT*Status(0)=Trt1 Trt2 Trt3 Trt4 ;

strata Enum;
id ID;
Trt1= Trt*(enum=1);
Trt2= Trt*(enum=2);
Trt3= Trt*(enum=3);
Trt4= Trt*(enum=4);
Trt5= Trt*(enum=5);
run;
```

The variable Enum is specified in the STRATA statement so that there is one marginal Cox model for each distinct value of Enum. The variables Trt1, Trt2, Trt3, Trt4, and Trt5 in the MODEL statement are event-specific variables derived from the independent variable treatment by the given programming statements. One can avoid using the programming statements in PROC PHREG if you create these event-specific variables in the input data set by using the same programming statements in a DATA step.

PWP model

To illustrate how the data for the PWP model is setup we will use subject 9 again. The data structure for this subject is shown below in Table 8.8 where the variable gaptime defines the time interval between consecutive events:

The following gives the SAS code that uses the above data in the PHREG

Table 8.8: PWP model data

Subject	Tstart	Tstop	Gaptime	Status	Treatment	Enum
9	0	0.1	0.1	1	1	1
9	0.1	2	1.9	1	1	2
9	2	3.2	1.2	1	1	3
9	3.2	11	7.8	0	1	4
9	3.2	11	7.8	0	1	5

procedure to fit the the PWP model for total time:

proc phreg covs(aggregate);
*model (Tstart,Tstop)*Status(0)=Trt1 Trt2 Trt3 Trt4 ;*

strata Enum;
Trt1= Trt(enum=1);*
Trt2= Trt(enum=2);*
Trt3= Trt(enum=3);*
Trt4= Trt(enum=4);*
Trt5= Trt(enum=5);*
run;

8.12 Application: Fitting the three models to the Epileptic data

In this section we will fit the AG, WLW, and PWP models to the epilepsy data set with treatment as the only covariate given in table 3.1 using proc phreg in SAS. We can not fit these models to the Thall and Vail data because in the Thall and Vail data set does not record the time at which a single event occurs. That is the counting process form is necessary; one observation per time interval or event.

AG model

Table 8.9 shows the results for the treatment effect in the AG model. The coefficient of the AG model is -0.71990, thus the hazard ratio is $e^{-0.71990} = 0.487 = \hat{\theta}$. A 95% confidence interval for $\hat{\theta}$ is (0.258; 0.702). Since 1 is not in the interval we can conclude that the active treatment has led to fewer epilepsy attacks.

Table 8.9: AG model

Parameter	DF	Estimate	Std Error	χ^2	p-value	θ
Treatment	1	-0.71990	0.22296	10.4253	0.0012	0.487

Marginal risk set or WLW

Out of the 22 patients, all patients have at least one epileptic seizure episode, 20 patients have two recurrences, 17 patients have three recurrences, 12 patients have four recurrences, 9 patients have 5 recurrences and so on. These figures are shown in table 8.10. Parameter estimates for the eleven marginal

Table 8.10: Summary of the Number of Event and Censored Values

Stratum	Visit	Total	Event	Censored	Percent Censored
1	1	22	22	0	0.00
2	2	22	20	2	9.09
3	3	22	17	5	22.73
4	4	22	12	10	45.45
5	5	22	9	13	59.09
6	6	22	8	14	63.64
7	7	22	6	16	72.73
8	8	22	3	19	86.36
9	9	22	1	21	95.45
10	10	22	1	21	95.45
11	11	22	1	21	95.45
Total		242	100	142	58.68

models are shown in table 8.11. The p-values in table 8.11 indicate a lack of evidence of a treatment effect in the first three recurrences at the 5% significance level. But then the treatment effect is significant in the rest of the strata. The optimal weights for estimating the parameter of the common treatment effect are 0.51546, 0.49266, -0.53580, 0.25013, -0.04812, 0.55946, -0.25160, 0.12184, 0.55716, -0.33300 and -0.32830 for Trt1, Trt2, Trt3, , Trt4, Trt5, Trt6, Trt7, Trt8, Trt9, Trt10, and Trt11, respectively, which gives a parameter estimate of -6.2062 with a standard error estimate of 0.2651. A more sensitive test for a treatment effect is the 1 degree of freedom test based on this common parameter; of which there is sufficient evidence for such effect at the 5% level (< 0.0001). This 1 degree of freedom estimate is obtained by specifying the following option in the model statement
TREATMENT: test trt1,trt2,trt3,trt4,trt5,trt6,
trt7,trt8,trt9,trt10,trt11/average e;

Table 8.11 shows the treatment effects within strata for the WLW model.

Table 8.11: WLW model

Parameter	DF	Estimate	Std Error	p-value	θ
Trt1	1	-0.78174	0.48508	0.1071	0.458
Trt2	1	-0.62714	0.41647	0.1321	0.534
Trt3	1	-0.92881	0.49663	0.0615	0.395
Trt4	1	-2.51403	0.70428	0.0004	0.081
Trt5	1	-2.62655	0.98041	0.0074	0.072
Trt6	1	-16.49074	0.47705	$< .0001$	0.000
Trt7	1	-16.34941	0.53072	$< .0001$	0.000
Trt8	1	-16.37891	0.62425	$< .0001$	0.000
Trt9	1	-16.20289	1.00000	$< .0001$	0.000
Trt10	1	-16.20289	1.00000	$< .0001$	0.000
Trt11	1	-15.99848	1.02470	$< .0001$	0.000

For strata one, the estimated coefficient is close to the AG model. Thus the hazard ratios are close that is 0.487 for the AG model and 0.458 for the WLW model. The relative risk of having a second epilepsy attack (0.534) is lower than that of experiencing the first epileptic attack (0.458). The data seems to suggest that after a person in the treatment group has had 5 epilepsy attacks they become risk free compared to the placebo group since the hazard ratio becomes 0. However this needs to be interpreted with caution because not all individuals have equal number of events.

PWP model for gap time

Results of the analysis of the PWP gap-time model are shown in table 8.12. The p-values in table 8.12 indicate a lack of evidence of a treatment effect in all the recurrences at the 5% significance level. For strata one the PWP for gap time and the WLW have the same hazard ratio thus the estimated coefficient of the PWP model for gap time is also close to the AG model, the relative risk of having a second epilepsy attack (0.860) is higher than that of having the first epileptic attack (0.458). The data seems to suggest that a person is not at risk of having a fourth epilepsy attack but is at risk of having a fifth epilepsy attack. This does not make sense. Maybe such discrepancies are due to the fact that not all individuals have equal number of events.

Table 8.12: PWP model for gap time

Parameter	DF	Estimate	Std Error	p-value	θ
Trt1	1	-0.78174	0.49834	0.1167	0.458
Trt2	1	-0.62714	0.46156	0.7433	0.860
Trt3	1	-0.81502	0.55636	0.1288	0.430
Trt4	1	-2.22264	1062	0.9875	0.000
Trt5	1	-0.69137	1.15719	0.8586	1.229
Trt6	1	-14.40644	1924	0.9936	0.000
Trt7	0	0	.	.	.
Trt8	0	0	.	.	.
Trt9	0	0	.	.	.
Trt10	0	0	.	.	.
Trt11	0	0	.	.	.

Table 8.13: PWP model for total time

Parameter	DF	Estimate	Std Error	p-value	θ
Trt1	1	-0.78174	0.49834	0.1167	0.458
Trt2	1	-0.62714	0.46920	0.1321	0.534
Trt3	1	-0.81502	0.52869	0.1232	0.443
Trt4	1	-2.22264	0.79841	0.0054	0.108
Trt5	1	-0.69137	1.07104	0.5186	0.501
Trt6	1	-14.40644	1205	0.9888	0.000
Trt7	0	0	.	.	.
Trt8	0	0	.	.	.
Trt9	0	0	.	.	.
Trt10	0	0	.	.	.
Trt11	0	0	.	.	.

PWP model for total time

Table 8.13 shows the results of the PWP model for total time. One should note that the regression coefficients for the first epileptic seizure are the same as those of the gap time model, since the total time and the gap time are the same for the first recurrence. There is no significant treatment effect on the total times for any of the recurrences except the fourth recurrent event.

Chapter 9

Conclusion

The modelling approaches discussed in this project have shown that there are many options available when one is modelling multiple events of the same type. When choosing the appropriate model "one" should bear in mind the objectives of the study and the type of data to be used. In this project we mainly focused on methods which used the counting process theory because the data we were dealing with was in the form of counts.

The GEE works best if the number of observations per subject is small and the number of subjects is large or if in longitudinal studies (e.g. growth curves), the measurements are taken at the same time for all subjects. In chapter 6 we mentioned that main advantage of GEE models is that, suppose the mean is correctly specified, but the variance and correlation structure are incorrectly specified, then GEE models still provide consistent estimates of the parameters and hence the mean function is correctly estimated, further consistent estimates of the standard errors can be obtained via a robust sandwich estimator. Hence when one's main interest is the population averaged estimates one can safely choose the GEE approach. One might also choose to use the Andersen and Gill(1982) model method in this situation.

On the other hand if one's interest is in modelling the different hazard functions for different strata then the WLW or PWP models maybe the most appropriate choices.

Generalized linear mixed models (GLMMs)are variable when the interest is in the lies in the individual response profiles. When one is looking at random effects which encompass variation among individuals one should choose generalized linear mixed models. Generalized linear mixed models (GLMMs) combine the properties of two statistical frameworks that is, linear mixed

models (which incorporate random effects) and generalized linear models (which handle non-normal clustered/repeated data. GLMMs are the best tool for analyzing non normal data that involve random effects: all one has to do, is to specify a distribution, link function and structure of the random effects.

In a nutshell the methods discussed are very useful when it comes to modelling epidemics and relevant parameters such as forces of infection. In particular the analysis of multiple events of the same type per subject can be applied when it comes to the issue of HIV/AIDS. Since HIV/AIDS patients do not usually die because of HIV/AIDS but due to opportunistic diseases such as TB(Tuberculosis) and PCP (Pneumocystis carinii pneumonia). Of which these diseases are recurrent thus the effect of treatment on these diseases can be modelled using the methods discussed in this dissertation. The project has addressed a very important problem in Biostatistics namely the analysis of repeated data.

Future work will include modelling the missingness pattern and their effect to parameter estimation and inferences and including a frailty term in the Cox proportional hazards model.

Bibliography

[1] http://en.wikipedia.org/wiki/Iteratively_re-weighted_least_squares

[2] http://www.epilepsyfoundation.org/about/faq/

[3] http://en.wikipedia.org/wiki/Generalized_linear_model

[4] http://en.wikipedia.org/wiki/Poisson-process

[5] http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/statug_phreg_sect041.htm

[6] Ake C.F., Carpenter A. L. *Extending the Use of PROC PHREG in Survival Analysis.* SAS Conference Proceedings: WUSS 2003, San Francisco, California

[7] Allison P.D., *Logistic regression using the SAS system : theory and application* SAS Institute.

[8] Andersen P. K. and Gill R. D. (1982) Cox's Regression Model for Counting Processes: A Large Sample Study. *Annals of Statistics*, **10**, 4:1100-1120.

[9] Balshaw R. F. (2002)A Semiparametric Model for the Analysis of Recurrent - Event Panel data. *Biometrics*, **58**, 324-331.

[10] Box G. E. P.,Jenkins G. M., and Reinsel G. C.3rd edition, (1994) *Time Series Analysis: Fore-casting and Control.* Englewood Cliffs, NJ: Prentice-Hall.

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

[12] Breslow N.E. and Lin X.(1995) Bias correction in generalized linear mixed models with a single components of dispersion, *Biometrika* **82**, 81-91.

[13] Breslow N.E. and Lin X.(1996) Bias correction in generalized linear mixed models with a multiple components of dispersion, *Journal of the American Statistical Association* **91**, 1007-1016.

[14] Cleves M. (1999) Analysis of Multiple Failure-Time Data with Stata. *Stata Technical Bulletin* **49**, 30-39.

[15] Cook R.J. and Lawless J.F. (2002). Analysis of repeated events. *Statistical Methods in Medical Research,***11**, 141-166.

[16] Cox D.R. (1972) Regression models and life-tables (with discussion). *Journal of Royal Statistical Society, Series B,* **74**, 187-220.

[17] Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269-276.

[18] Cox D.R. and Miller H.D. (1965) *The Theory or Stochastic Processes.* London : Chapman & Hall.

[19] Cowling B.J., Hutton J.L. and Shaw J.E.H. (2006). Joint modelling of event counts and survival times. *Journal of the Royal Statistical Society, Series C,* **55**, 1-39.

[20] Diggle P.J.(1983) *Statistical analysis of spatial point patterns.* London: Academic Press

[21] Diggle P.J, Liang K-Y and Zeger S.L. (1994) *Analysis of Longitudinal data.* Oxford: Oxford University Press.

[22] Diggle, P.J., Heargerty, P., Liang, K-Y. and Zeger, S.L. 2nd Edition, (2002) *Analysis of longitudinal data.* Oxford University Press, Oxford.

[23] Feller W. 2nd Edition, (1957). *An introduction to probability theory and its applications* Wiley, (New York).

[24] Ghosh D. (2000). Methods for Analysis of Multiple Events in the Presence of Death. *Controlled Clinical Trials***21**, 2: 115-126.

[25] Henderson R., Diggle P. and Dobson A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics,* **1**, 465-480.

[26] Kalbfleisch, J. D. and Prentice R.L (1980) *Statistical analysis of failure time data.* New York: Wiley.

[27] Laird M.N. and Ware J.H. (1982) Random effects models for longitudinal data. *Biometics,* **38**, 963-974. In [52]

[28] Lee E.T. 2nd Edition, (1992) *Statistical Methods for Survival Data Analysis* New York : Wiley.

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

[30] Little R.J.A. and Rubin, D.B. (2002) *Statistical with Missing Data.* New York: Wiley Interscience.

[31] Madekurozwa M.N.(2008) *A Quality of Care Audit of Children Referred with Suspected Epilepsy to Two Hospitals in Pietermaritzburg, KwaZulu-Natal.* University of the Witwatersrand.

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

[33] Molenberghs G. and Kenward M.G. (2007) *Missing DAta in Clinical Trials* Wiley: England.

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

[35] Mwambi H., Ramroop S., Shkedy Z. and Molenberghs G. (2009) A frequentist approach to estimating the force of infection and recovery rate for a respiratory disease among infants in coastal Kenya. *Statistical Methods in Medical Research.*

[36] Nelder J.A. and Pregibon D. (1987) An extended quasi-likelihood function. *Biometrika* **74**,2: 221-232.

[37] Nelder J.A. and Wedderburn R.W.M (1972) Generalized Linear Models. *Journal of Royal Statistical Society, Series A (General),* **135**,3:370-384.

[38] Oakes D.A. (1992) *Frailty Models for Multiple Event Times.* In Duration Models for Repeated Events. Box-Steffensmeierm J.M. and Zorn C. (2002) *Journal of Politics,***64**, 4:1069-1094.

[39] Pawitan Y. (2006) *In all likelihood: Staistical Modelling and Inference Using Likelihood* Oxford: Oxford University Press.

[40] Pedroso de Limab A.C, Paes A.T. (2004) A SAS macro for estimating transition probabilities in semiparametric models for recurrent events. *Computer Methods and Programs in Biomedicine* **75**, 1:59-65.

[41] Phipson B. (2006) *Analysis of Time-To-Event Data Including Frailty Modelling.* University of KwaZulu-Natal.

[42] Prentice R. L.,Williams B. J. and Peterson A. V . (1981) On the Regression Analysis of Multivariate Failure Time Data. *Biometrika* **68**, 2: 373-79.

[43] Ross S. M.(1989) *Introduction to Probability Models* London: Academic Press, Inc.

[44] Ross S. M. 2nd Edition, (1996) *Stochastic Processes* New York: John Wiley & Sons, Inc.

[45] Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581-592.

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

[47] Seber, G.A.F (1984) *Multivariate Observations* New York: John Wiley & Sons.

[48] Sturmer T., Glynn R. J., Kliebsch U., Brenner H. (2000) Analytic strategies for recurrent events in epidemiologic studies: background and application to hospitalization risk in the elderly. *Journal of Clinical Epidemiology* **53**, 1: 57 - 64.

[49] Talke, I.S. (2003) *Modelling Votality In Time Series Data.*University of KwaZulu-Natal.

[50] Thall, P.F. and Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, **46**, 657-671.

[51] Therneau T.M. and Grambosch P.M.(2000) *Modeling Survival Data: Extending the Cox Model.* New York: Springer-Verlag.

[52] Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal data.*Springer

[53] Wedderburn, R.W.M (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**,3: 439-447.

[54] Wei L. J., Lin D. Y. and Weissfeld L. (1989) Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *Journal of the American Statistical Association* **84**, 408: 1065-73.

[55] White, Halbert. (1980) A Heteroskedasticity-Consistent Covariance Matrix and a Direct Test for Heteroskedasticity. *Econometrica* **48**, 4: 817-38.

[56] Zhao Q., Sun J. (2006) Semiparametric and nonparametric analysis of recurrent events with observation gaps. *Computational Statistics and Data Analysis* **51**, 1924-1933.