ANALYSIS OF LONGITUDINAL BINARY DATA: AN APPLICATION TO A DISEASE PROCESS

By

SHAUN RAMROOP

Submitted in fulfilment of the academic
requirements for the degree of

DOCTOR OF PHILOSOPHY
in
Statistics

in the
School of Statistics and Actuarial Science
University of KwaZulu-Natal
Pietermaritzburg
2008

Dedication

To my mother, the late 'Sukhrani Ramroop' (may her soul rest in peace), and my wife, 'Shanitha Ramroop'.

Declaration

The research work described in this thesis was carried out in the School of Statistics and Actuarial Science, University of KwaZulu-Natal, Pietermaritzburg Campus, from July 2003 to July 2008, under the supervision of Dr. Henry G Mwambi.

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

July, 2008.

Mr Shaun Ramrooop	Date
Dr Henry Mwambi	Date

Acknowledgement

First and foremost I would like to express my appreciation and gratitude to my Lord and Saviour, Jesus Christ, for giving me life breath and the strength to complete this thesis. I owe you my life Lord!

I thank my supervisor, Dr. H Mwambi. I thank you for putting up with me with your unwavering patience, uplifting encouragement and prudent correction. Your knowledge and insight in modelling infectious diseases combined with your carrot and stick method has made this thesis possible. I have learned rich lessons of statistical research from you that will stay with me for the rest of my life.

I would like to dedicate the work of this thesis to the loving memory of my late mother, Sukhrani Ramroop who passed away in May 2006. I was presenting the work on generalized estimating equations from the paper by Liang and Zeger (1986) to my supervisor Dr. H Mwambi when I received the news of her passing away. I will never for forget this paper for as long as I live.

I must also make mention of my wife Shanitha Ramroop and my son, Elisha Benjamin. I thank my wife for her constant support, patience and encouragement to me as my life partner. I also express my appreciation to her for all the time she availed to me to complete the thesis, by taking care of our son. Thank you Shani!

Finally I would like to thank my HOD, Professor Temesgen Zewotir for the wisdom he has given me in writing this thesis as well as the enriching conversations we have had regarding the finer details of a PhD. I thank you my friend and affirm that you have made a significant contribution in my life!

Abstract

The analysis of longitudinal binary data can be undertaken using any of the three families of models namely, marginal, random effects and conditional models. Each family of models has its own respective merits and demerits. The models are applied in the analysis of binary longitudinal data for childhood disease data namely the Respiratory Syncytial Virus (RSV) data collected from a study in Kilifi, coastal Kenya. The marginal model was fitted using generalized estimating equations (GEE). The random effects models were fitted using 'Proc GLIMMIX' and 'NLMIXED' in SAS and then again in Genstat. Because the data is a state transition type of data with the Markovian property the conditional model was used to capture the dependence of the current response to the previous response which is known as the history. The data set has two main complicating issues. Firstly, there is the question of developing a stochastically based probability model for the disease process. In the current work we use direct likelihood and generalized linear modelling (GLM) approaches to estimate important disease parameters. The force of infection and the recovery rate are the key parameters of interest. The findings of the current work are consistent and in agreement with those in White et al. (2003). The aspect of time dependence on the RSV disease is also highlighted in the thesis by fitting monthly piecewise models for both parameters. Secondly, there is the issue of incomplete data in the analysis of longitudinal data. Commonly used methods to analyze incomplete longitudinal data include the well known available case analysis (AC) and last observation carried forward (LOCF). However, these methods rely on strong assumptions such as missing completely at random (MCAR) for AC analysis and unchanging profile after dropout for LOCF analysis. Such assumptions are too strong to generally hold. In recent years, methods

of analyzing incomplete longitudinal data have become available with weaker assumptions, such as missing at random (MAR). Thus we make use of multiple imputation via chained equations that require the MAR assumption and maximum likelihood methods that result in the missing data mechanism becoming ignorable as soon as it is MAR. Thus we are faced with the problem of incomplete repeated non–normal data suggesting the use of at least the Generalized Linear Mixed Model (GLMM) to account for natural individual heterogeneity. The comparison of the parameter estimates using the different methods to handle the dropout is strongly emphasized in order to evaluate the advantages of the different methods and approaches. The survival analysis approach was also utilized to model the data due to the presence of multiple events per subject and the time between these events.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Crowder and Hand (1990) state that repeated measures arise in many diverse fields and are possibly even more common than single measurements. The term 'repeated' is used to describe measurements which are made of the same characteristic on the same observational unit but on more than one occasion. In longitudinal studies individuals may be monitored over a period of time to record the developing pattern of their observed values. The conditions of the period may be deliberately changed, such as in crossover trials to study the effect of treatment on the individual. Lindsey (1999) state what distinguishes repeated observations from those in the more traditional statistical data modelling as being that:

- the same variable is measured on the same observational unit more than once and as a result the responses are not independent as in the usual regression analysis and where
- more than one observational unit is involved; the responses do not form a simple time series.

Lindsey (1999) gives two factors that are imperative to repeated measures data, namely

1. the two types of stochastic dependence among measurements on the same observational unit,
    - homogeneity of responses on a unit versus heterogeneity across units
    - distance, in time or space, among responses on a unit.
2. the three basic types of responses which may be measured,
    - general continuous data
    - categorical or count data (as in the current study)
    - duration and survival data

Repeated measures may be spatial rather than temporal. Crowder gives the following two examples to validate the above statement i.e. the first concerns individual load bearing cables where the breaking strength may be measured at several points along the length, so 'time' becomes 'distance'. The second is an example in two dimensions, where the intensity of corrosion may be recorded at points over the floor area of a metal tank. Lindsey (1999) give various examples of repeated measures data that range from Agriculture, Biology, Business, Commerce, Engineering, Medicine (successive periods of illness and recovery under different treatment regimes) and Geography, just to mention a few. There are various examples of repeated measures from different fields, but it can be best summarized that in most contexts where a single measurement can be made, repeated measurements can also be made. Lindsey (1999) calls longitudinal studies prospective studies, that is, a sample

of units may be chosen according to the criteria of certain explanatory variables and then followed up in time to see what response is obtained. Diggle et al. (2002) also highlight this point stating that longitudinal data can be collected prospectively, following subjects forward in time or retrospectively, by extracting multiple observations on each person from historical records. They then go on to point out that longitudinal data are more commonly collected prospectively since the quality of repeated measurements collected from past records or from a person's recollection my be inferior due to recall bias (Goldfarb, 1960). Under this broad class, of prospective studies, we find panel, clinical trial and cohort studies. Examples of such prospective (longitudinal) studies include growth studies and longitudinal health studies. Lindsey (1999) further states that time is an explanatory variable within units and the aim is to compare the differences in the way in which the measurements change over time for different units or groups of units.

Diggle and Donnelly (1989) give the following characteristics of longitudinal repeated measurement studies:

(i) The data consist of a relatively large number of time series, structured by some type of more or less complex experimental or sample design

(ii) Usually the time series are relatively short

(iii) The times of measurement may be unequally spaced, may include missing values, and may be different among units. This may be the result of design or of accidental missingness.

(iv) The series will often be nonstationary in mean and/ or in higher order structure

(v) The researcher is usually most interested in the mean response profile, often some sort of location model linking the experimental conditions

> to the observed series of responses. However a reasonable model for the higher order structure, at least the second order covariance structure, is important, if only to provide valid and efficient inferences about the mean response profile. In some cases, it may also be of interest in its own right.

Molenberghs and Verbeke (2005) state that longitudinal data exhibit replication 'in two directions', subjects on the one hand and repeated measurements within subject collected over time on the other hand, with in addition the specific structure imposed by the uni-directional time dimension, makes them rich in structure.

As far as predictions go, the longitudinal data on available units will be used to predict future values of a unit which has a shorter time series, with prediction only up to the end of the time period of the first set of units. Diggle et al. (2002) state that the differences between longitudinal studies and cross-sectional studies is essentially that in a cross-sectional study, a single outcome is measured for each individual as opposed to the several measurements per individual in longitudinal studies. They further state that while it is often possible to address the same scientific questions in a longitudinal or cross-sectional study, the major advantage of longitudinal studies is its capacity to separate what in the context of population studies are called *cohort* (changes over time within subjects from differences within subjects at the baseline level) and *age* (changes over time within subjects) effects.

The data set that we will be analyzing falls under categorical and count data. Lindsey (1999) state that categorical and count data are increasingly becoming more common in statistical inference, and, in a number of

fields, now have far more importance than normal type responses. He then distinguishes the differences between count and categorical data by giving two examples to exemplify this point, one with respect to a study of overweight/obese people and the other example with respect to the number of industrial accidents per week in several factories over time.

Molenberghs and Verbeke (2005) state that there are three families of models that are suitable for modelling longitudinal data. They are *marginal models*, *random effects models* and *conditional models*. They formalize the idea of the longitudinal data setting as follows: Each individual has a vector $\boldsymbol{Y}$ of responses with a natural (time) ordering among the components. This leads to several generally nonequivalent, extensions of univariate models. In a *marginal model*, marginal distributions are used to describe the outcome vector $\boldsymbol{Y}$, given a set X of predictor variables. The correlations among the components of $\boldsymbol{Y}$ can then be captured either by adopting a fully parametric approach or by modelling a limited number of lower order moments only. Alternatively in the *random effects models*, the predictor variables X are supplemented with a vector $\boldsymbol{b}$ (or $\boldsymbol{\beta}$) of subject specific effects, conditional upon which the components of $\boldsymbol{Y}$ are often assumed to be independent. This does not preclude the fact that more elaborate models are possible if residual dependence is detected. Finally a *conditional model* describes the distribution of the components of $\boldsymbol{Y}$ conditional on X but also conditional on (a subset of) the other components of $\boldsymbol{Y}$. In a longitudinal context, a particular relevant class of conditional models that describes a component of $\boldsymbol{Y}$ given the ones recorded earlier in time are the so-called *autoregressive* or *transition models*. These models will be looked at in more detail in subsequent chapters with application to the Respiratory Syncytial Virus (RSV) data set that consti-

tutes the object of analysis in the current study. The current work is based on a repeated measurements or longitudinal data monitoring the infection status from a respiratory disease (RSV) for children within one year of age. The focus in the thesis is a problem with complex data structures particularly involving the dependence and incompleteness. Modelling is crucial for simplifying and clarifying the structure. Compromise with the likelihood, for example using marginal, conditional or approximate likelihoods, become necessary.

The aims and objectives of the research is to:

- Model the RSV data set using different techniques by taking into account the longitudinal structure that is present in the data.
- Estimate the force of infection and per capita loss of infection or recovery rate using likelihood based methods.
- Model the missingness and the dropout process in the data set.
- Review the computational tools that software packages such as SAS and Genstat incorporate to model longitudinal data.

Chapter 2

Exploratory and preliminary data analysis

The Kilifi data set from coastal Kenya is a longitudinal study measuring the prevalence of the Respiratory Syncytial Virus (RSV, a causal agent of pneumonia) in children. By definition, a longitudinal study is one where data are obtained when a response is measured repeatedly on a set of units. The Kilifi data set is part of a study carried out by the Kenyan Medical Research Institute and the Wellcome Trust in Kilifi, Kenya. The data set presents a form of missingness or incompleteness which has to be properly accounted for in order to carry out an appropriate analysis of the data which will lead to correct conclusions. The model that will be built to represent this data will aid in understanding the disease process and in the design of intervention strategies for this disease affecting children mostly under the age of one year because proper inference will be drawn from it. The Kilifi data set had 368 children that were recruited in the study, however only 334 children's data were measured and recorded. For each child the following information was collected with the variable names in brackets and underlined. These are:

- the number of visits (visit);
- the time between visits (dt);
- the type of sampling, active sampling if the field worker went to visit the child and passive if the child was brought to the clinic to be sampled (actipass);
- the age in months of the child at the visit (age);
- whether the child is infected or uninfected (rsv), this is the response variable;
- the time in months since the beginning of the study (timemonth);
- the prevalence of the virus in the blood (prev), a continuous variable;

In all there were 9374 responses that were measured and the number of times each child is measured varied from one child to another, for example, child no. 344 is measured at 12 different occasions with unequal time intervals between measurements and child # 368 was measured at 20 different occasions with equally spaced time intervals. In designed experiments data can be described as *balanced* when each cell in the data set contains the same number of observations and as *unbalanced* when this is not the case. Unbalanced data can also occur when the design of the experiment forces the data to be unbalanced i.e 'planned unbalancedness'. Unbalanced data can also result from unfortunate circumstances or experimental carelessness, for example, if the experimenter loses some of the data points. In this regard the Kilifi data set can be described as a highly *unbalanced* one. The coding and levels of data variables is shown in Table 1.1 below:

Variable	Levels and coding
id	1,. . . ,368
rsv	1=uninfected, 2=infected
dt	0,. . . ,181
visit	1,. . . ,44 (not all the children had 44 visits)
actpass	1=active sampling, 2=passive sampling
age	0,. . . ,12 months
timemonth	1,. . . ,12 months
prev	a continuous variable ranging from min=0 to max=0.047516

Table 2.1: Table of variables

It is clear that the response variable (rsv) in the above data set is a binary non-Gaussian variable. The generalized linear model in a longitudinal setting seems the best option to deal with such a data set. First, in general the linear mixed model makes the following four assumptions:

- First the mean structure or a function of the mean structure in the case of Non-Gaussian data is modelled as a linear function of the measured covariates and other explanatory variables such as time in the above example data set.

- Secondly the model also makes assumptions about the variance function which can be constant, linear, quadratic or take on other forms.

- Thirdly there is an assumption about the correlation structure i.e. a model can assume a constant or serial correlation structure across the individual set of observations or a more complicated correlation structure.

- Finally the model also makes an assumption about the individual-

specific profiles which can be linear, quadratic or general polynomial structure.

In practice the linear mixed model can be built using a two stage formulation as outlined in Verbeke and Molenberghs (2000). This type of formulation will be discussed in detail in Chapter 2. Data exploration is therefore an extremely helpful tool in the selection, identification and inference of an appropriate model. Although the Kilifi data set cannot be appropriately analyzed as Gaussian longitudinal data the similarities and dis-similarities with non-Gaussian data is important in order to develop an appropriate model for it. For this reason Chapter 2 will be devoted to a review of the linear mixed model under the Gaussian assumption.

2.1 Profile plots

Figures 2.1 and 2.2 show plots for disease status against sample visits for individual subjects and a group of individuals respectively. From the profile plots in Figure 2.1 and Figure 2.2, it is clear that many of the children remained uninfected for most of the study time in which they were followed. The zero level in the plots indicate that the child was in an uninfected state. The x-axis is the visitation number of the child whilst the y-axis denotes the disease status of the child.

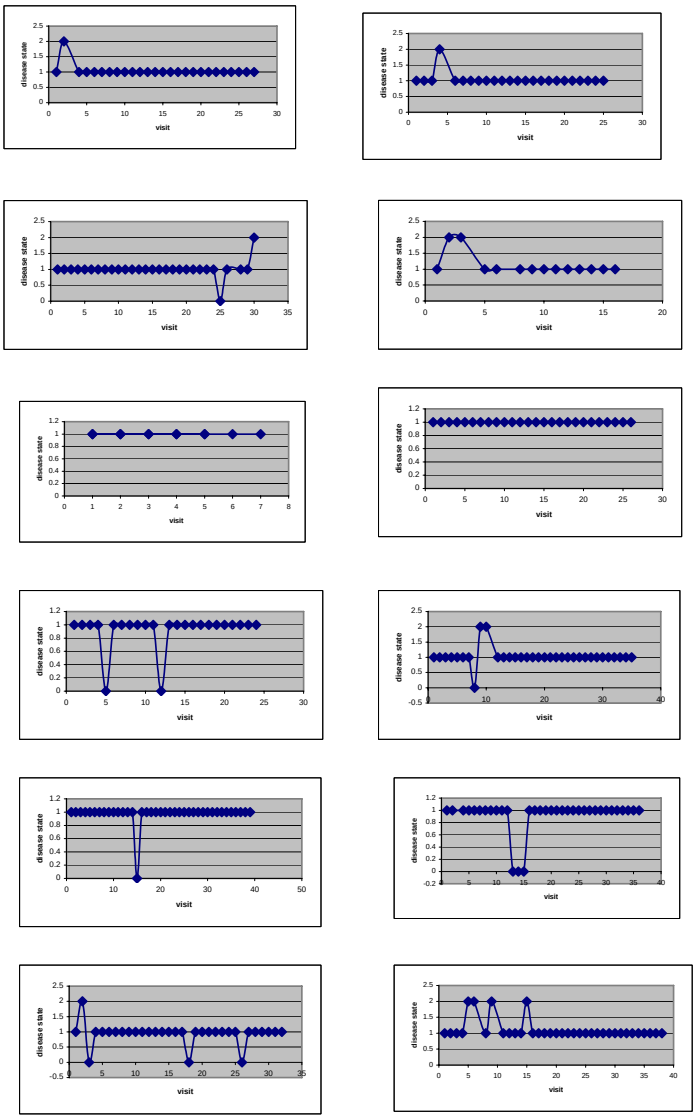

Figure 2.1: A sample of profile plots.

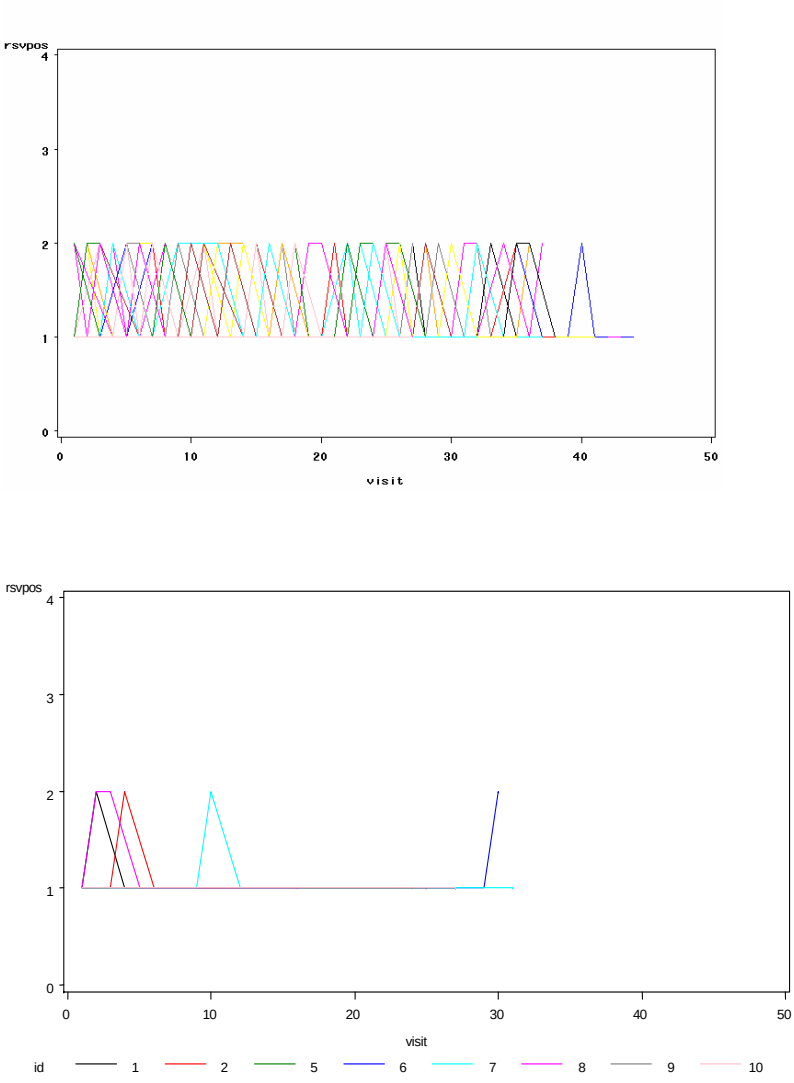

Figure 2.2: Overall profile and a sample of 10 profile plots.

2.2 Overall transitions

A program was written in SAS Proc IML to get the following 2×2 transition matrix:

		Y_{ij}	
		uninfected	infected
Y_{ij-1}	uninfected	8598	132
	infected	131	13

Table 2.2: Matrix of transitions between infected and uninfected states

The 2×2 matrix gives the number of visits to the uninfected and infected states conditional on the state (row) at previous visit. From the above matrix, it is clear that this disease is a rare one because most of the transitions were from uninfected to uninfected states. There are a total of 131 transitions among the children from the uninfected to the infected state and almost a similar number, of 132, transiting from infected to uninfected. It is however important to note that the time interval between transitions was not constant. The time intervals were different within and between the children, which as previously stated, makes the data set highly unbalanced. Therefore standard methods of analysis may not be directly applicable. Appropriate models to deal with the data is one of the aims of this thesis.

2.3 Visits by week

A program was written in SAS using the Proc IML to group the data into weeks and also to count the number of transitions between the two states of uninfected and infected during a given week. The probability of being in the infected and uninfected states was calculated for each week. This information is displayed in Table 2.3.The reason for exploring the data in this way was to possibly isolate temporal trends in the infection process as well as identifying the weeks when the probability of infection had been at its peak. Figure 2.3 which is a plot of the observed weekly probability to be infected against time clearly shows that the respiratory disease is highly seasonal. There are times when the incidence is high and times when it is very low. This is evidence of time dependence in the dynamics of the respiratory disease which is possibly due to the effect of climatic variables which in turn are time dependent.

Week	Uninfected(1's)	Infected(2's)	Probability of being in an infected state	Probability of being in uninfected state
1	248	9	0.03502	0.96498
2	263	7	0.02592	0.97407
3	293	10	0.03300	0.96700
4	304	10	0.031847	0.96815
5	297	7	0.02300	0.97697
6	296	7	0.0231	0.97698
7	307	4	0.0129	0.98713
8	288	5	0.017	0.98294
9	294	9	0.0297	0.9703
10	271	4	0.0145	0.98545
11	266	6	0.0220	0.9779
12	276	2	0.0071	0.9928
13	264	3	0.0112	0.9887
14	258	1	0.0039	0.99613
15	240	1	0.00414	0.99585
16	230	1	0.0043	0.99567
17	199	0	0	1
18	147	0	0	1
19	129	0	0	1
20	131	1	0.0076	0.9924
21	133	0	0	1
22	70	0	0	1
23	79	1	0.0125	0.9875
24	108	0	0	1
25	126	0	0	1
26	86	1	0.01149	0.9885
27	74	0	0	1
28	107	0	0	1
29	115	0	0	1
30	86	0	0	1

Week	Uninfected(1's)	Infected(2's)	Probability of being in an infected state	Probability of being in an uninfected state
31	76	0	0	1
32	107	1	0.0093	0.9907
33	145	0	0	1
34	100	4	0.0385	0.9615
35	115	2	0.0171	0.9829
36	101	2	0.0194	0.9806
37	134	3	0.0219	0.9781
38	117	4	0.0331	0.9669
39	147	2	0.0134	0.9865
40	167	6	0.0457	0.9543
41	184	6	0.032	0.9684
42	192	5	0.0253	0.9746
43	213	7	0.032	0.0.968
44	227	6	0.026	0.974
45	221	3	0.013	0.986
46	215	3	0.0137	0.9863
47	164	3	0.0179	0.9821
48	100	0	0	1
49	51	0	0	1
50	28	0	0	1
51	11	0	0	1
52	1	0	0	1

Table 2.3: Visits by week for being in the infected and uninfected states

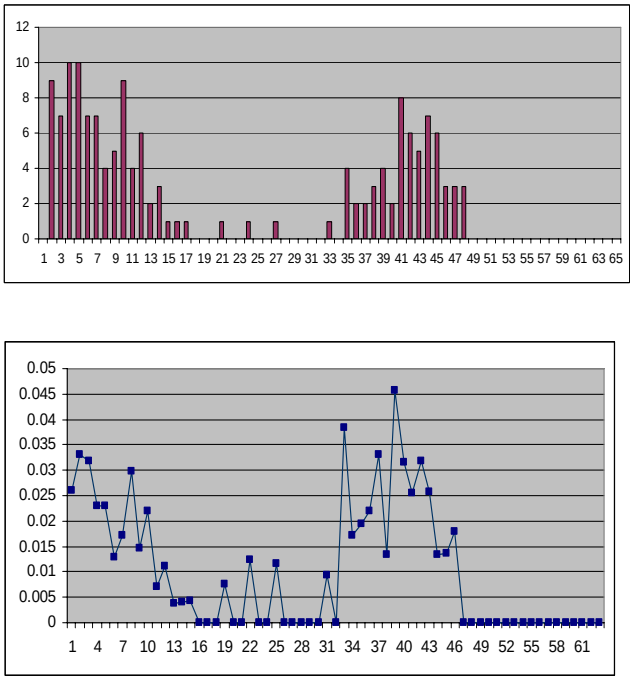

Figure 2.3: The probability of infection over time in weeks .

The graphs indicate that the probability of being in an infected state is highest in weeks 40-43 but noticeably high in weeks 1-4 as well. The probability of infection during the period between week 17-33 is dominated by a sequence of zero probabilities. The bar chart further confirms these facts. In relation to modelling the data, this would imply that choosing and fitting an appropriate model to the data requires a very careful and systematic process in order to deal with certain specific complexities about the data set such as the seasonal effects and missingness.
The exploratory data analysis for the weeks yields the following information:

- The probability of being in an infected state is highest in weeks 1-4 and 40-43 which possibly indicates that the disease, RSV, is affected by some kind of seasonal factors such as temperature, humidity etc. as stated by White et al. (unpublished paper 2003).

- The probability of being in an uninfected state seems to be considerably high throughout all the periods of about 0.98.

- The data for weeks 53-64 have been excluded from the above table because all the entries are zero, implying that none of the children continued to have their data recorded for the full length of 64 weeks. This aspect of missingness will be discussed in Section 2.4 and a full analysis will be done in Chapter 9.

2.3.1 Visits by month

A program was written in SAS 'Proc IML' to group the data into months and also to count the number of times the children were infected or uninfected during those months. The probability of being in the infected state and uninfected state was calculated during each month. The reason for exploring the data in this way was to possibly detect any temporal trends in the infection process as well as identifying the months when the probability of being in the infection state was at its peak. A further reason was to compare the disease infection process when time is discretized weekly and monthly.
The results are tabulated as follows:

Months	Uninfected(1's)	Infected(2's)	Probability of being in the infected state	Probability of being in the uninfected state
1	1136	38	0.0323	0.967
2	1254	24	0.0187	0.9812
3	1252	21	0.0165	0.9835
4	961	3	0.0031	0.9969
5	548	1	0.0018	0.9982
6	417	2	0.0047	0.9953
7	417	0	0	1
8	447	5	0.011	0.989
9	491	11	0.0219	0.978
10	836	27	0.031	0.969
11	855	16	0.0183	0.9817
12	187	0	0	1

Table 2.4: The probability of being in the infected and uninfected states over time in months

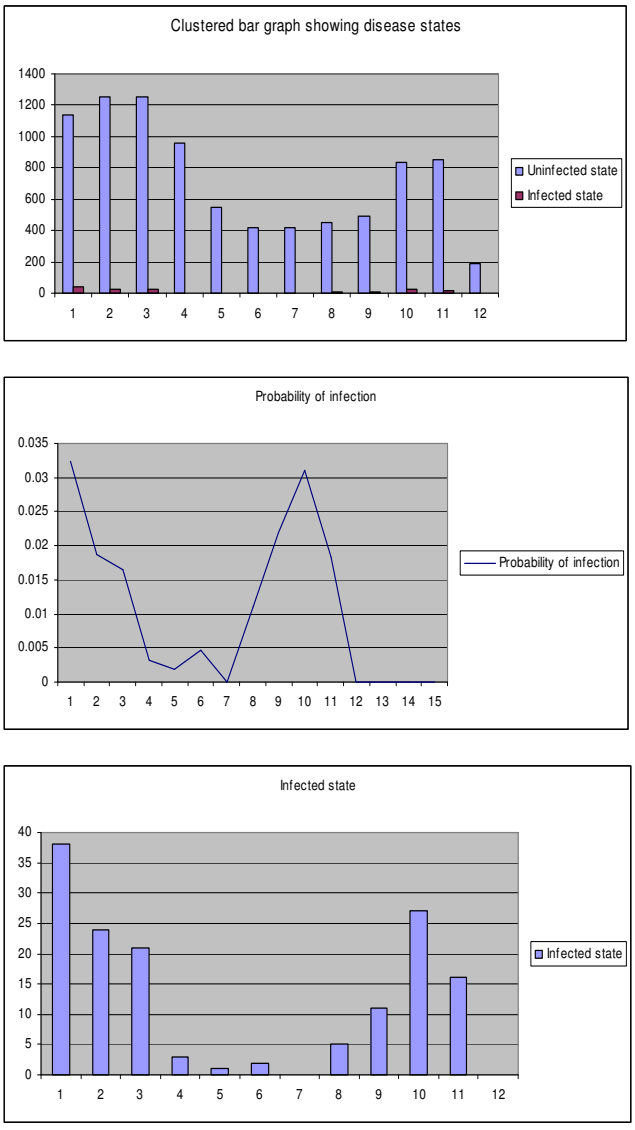

Figure 2.4: The probability of infection in months.

The 13th, 14th and 15th months have been left out of the table deliberately because they have no entries as in the case of the weeks 52-64. From the graphs it is evident that the probability of being in the infected state peaks in months 1-3 and 9-11. The conclusions about the probability of being in an infected state from the months data is similar to that of the exploratory data analysis by weeks. The trends seen in the above graphs are similar to those in the weekly graphs. From the visits by week and the visits by months, it is clear that the chance of a child being in the infected state is related to or affected by some climatic factors in the area, possibly rainfall or temperature, typical of airborne diseases. These trends were also stated in the unpublished paper by White et al. (2003). The probability of infection is highest in months 1-3 and 9-11. The average probability of recurrent infection in the monthly time scales is 0.011427 and the average probability of infection in weekly time steps is 0.010757. The reason for the difference is obvious; the discretization steps are different. Nonetheless the values are of the same order of magnitude as is expected.

2.3.2 Individual transition matrices

In order to fully understand the data, individual transition matrices were constructed using a program written in SAS 'Proc IML'. All the matrices can not be displayed because there are 334 of them. Most of the observed transition matrices indicate that most transitions were from the uninfected to the uninfected states and not from the uninfected to the infected states. In other words most participants remained sero-negative most of the observation time. There were only 5 children in the entire data set whose initial state was the infected state and 229 of them that began the study with the uninfected state. Below are the transition matrices for child 1 and child 2:

Child 1:

		Y_{ij}	
		uninfected	infected
Y_{ij-1}	uninfected	23	1
	infected	1	0

Table 2.5: Matrix of transitions between infected and uninfected states for child 1

Child 2:

		Y_{ij}	
		uninfected	infected
Y_{ij-1}	uninfected	21	1
	infected	1	0

Table 2.6: Matrix of transitions between infected and uninfected states for child 2

2.3.3 Visits

The visits per child are classified as either an actively sampled visit or a passively sampled visit. An actively sampled visit occurred when the field worker visited the child for sampling and a passively sampled visit occurred if the child was brought to the health clinic to be sampled. There were 7506 actively sampled visits and 1868 passively sampled visits. The following graphs depict the total number of visits per child, the number of passively sampled visits per child and the number of actively sampled visits per child. The vertical axis indicates the number of visits and the horizontal axis represents the child identification number (child 1, child 2 etc.) in all three graphs.

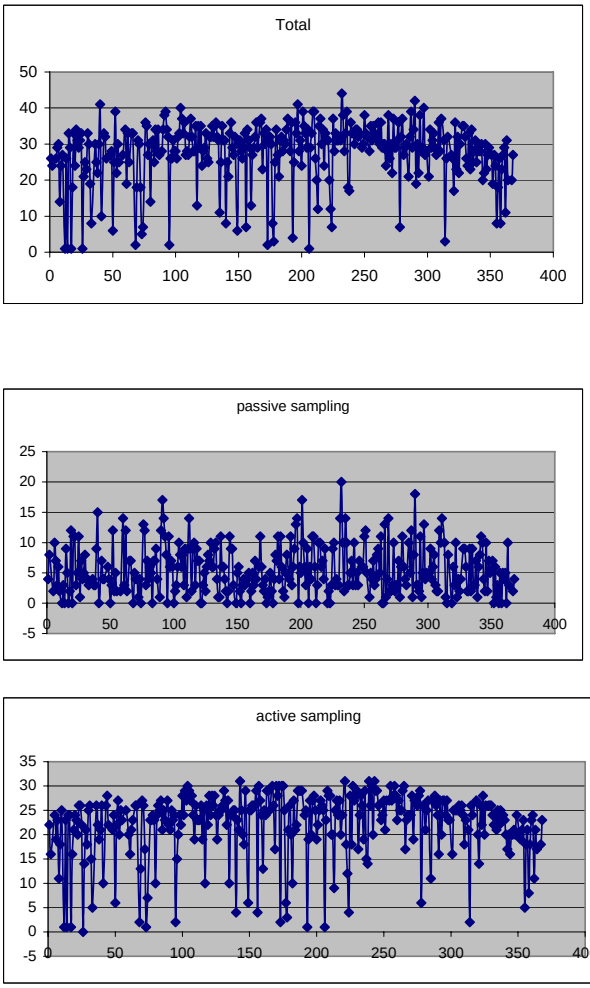

Figure 2.5: The graphs of visits.

2.3.4 Age at the first visit

The following descriptive statistics of the ages of the children (in months) were obtained using SPSS at the first visit:

Descriptive Statistic	Value(months)
Mean	0.6805
Median	1
Mode	1
Sample standard deviation	0.5379
Sample variance	0.2893
Kurtosis	0.002254
Skewness	0.0476
Range	3
Minimum	0
Maximum	3

Table 2.7: Descriptive statistics of the age at first visit

The following frequency distribution table of the children's ages at the first visit was also constructed:

Age(months)	Frequency
0-0.9 months	119
1-1.9 months	209
2-2.9 months	9
3 months	1

Table 2.8: Frequency distribution table of the age at first visit

2.4 Missingness

There is clearly a form of missingness in the data set, that could be one of the following three cases. These are missing at random (MAR), missing

completely at random (MCAR) or missing not at random (MNAR) as defined by Little and Rubin (2002). Little and Rubin (2002) define the complete data $Y = y_{ij}$ and the missing data indicator $M = m_{ij}$ such that $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is present. The missing data mechanism is characterized by the conditional distribution of M given Y say, $f(M|Y, \phi)$ where ϕ denotes some unknown parameters. If the missingness does not depend on the values of the data Y, missing or observed, that is, if

$$f(M|Y, \phi) = f(M|\phi)$$

for all Y and ϕ, the data is said to be missing completely at random (MCAR). The missingness does not depend on the data values. Heitjan (1997) provides the following example of MCAR missing data: Imagine a research associate shuffling raw data sheets and arbitrarily discarding some of the sheets. Then this would constitute data which is MCAR. Another example of data which is MCAR arises when investigators randomly assign research participants to complete two-thirds of a survey instrument. Graham, Hoffer and Mackinnon (1996) illustrate the use of planned missing data patterns of the MCAR type using a survey example where responses are gathered more from survey items from fewer research participants than one would ordinarily obtain from a standard survey completion paradigm in which every research participant receives and answers each survey question. Let Y_{obs} denote the observed components or entries of Y and Y_{miss} the missing components. When the missingness depends only on the Y_{obs}, the observed components or entries, then the missing data mechanism is called missing at random (MAR) and

$$f(M|Y, \phi) = f(M|Y_{obs}, \phi)$$

for all Y_{miss} and ϕ. This is also called ignorability. Ignorability depends of the analysis method used. Strictly this only applies under likelihood analyses.

MAR does not imply ignorability under unweighted GEE, for example. Cases with incomplete data differ from cases with complete data, but the pattern of data missingness is traceable or predictable from other variables in the database rather than being due to the specific variable on which the data are missing. For example, if research participants with low self-esteem are less likely to return for follow-up sessions in a study that examines anxiety level over time as a function of self-esteem, and the researcher measures self-esteem at the initial session, self-esteem can then be used to predict the missingness pattern of the incomplete data. Another example is reading comprehension. Investigators can administer a reading comprehension test at the beginning of a survey administration session; research participants with lower reading comprehension scores may be less likely to complete the entire survey. In both of these examples, the actual variables where data are missing are not the cause of the incomplete data. Instead, the cause of the missing data is due to some other external influence. When the missing data mechanism depends conditionally on the missing response variables, given the observed variables, we have missing not at random (MNAR). This type of missingness is also called non-ignorability. This pattern of data missingness is non-random and it is not predictable from other variables in the database. If a participant in a weight-loss study does not attend a weigh-in due to concerns about his weight loss, his data are missing due to non-ignorable factors. In contrast to the MAR situation outlined above where data missingness is explainable by other measured variables in a study, non-ignorable missing data arise due to the data missingness pattern being explainable by the very variable(s) on which the data are missing. This poses an interesting challenge on how to firstly determine the type of missingness and secondly, on how to model the data under the prevailing missingness. The analysis of missing data will be

carried out in detail in Chapter 9. In Chapter 4 the problem is addressed partly via the use of Weighted Generalized Estimating Equations (WGEE).

2.5 Conclusion

The exploratory data analysis suggests that RSV is a rare disease based on the preceding descriptive statistics and measures. There were not many transitions from the uninfected to the infected states nor from the infected to infected states. The transition pattern could possibly hint at a low force of infection for the disease. The possible low force of infection was also evident when the data was broken down by weeks, then by months and the transition probability plotted over time. The missingness present in the data set presents a challenge in how to estimate the intermittent missing values (85 missing values in the response variable) and more importantly how to model or estimate the missing values in the dropout process where each child is supposed to have had 44 visits but this was not the case. The data presents challenging features to be modelled in a longitudinal data set up in the subsequent chapters of the thesis.

Chapter 3

Modelling Continuous Longitudinal Data

The Linear Mixed Model for Longitudinal Continuous Data

3.1 Introduction

The Kilifi data consists of repeated measurements over time giving the status of infection (infected or uninfected) with the Respiratory Syncytial Virus (RSV) virus among children within the age of one year in Kilifi, Kenya. Thus we have a form of longitudinal data with a binary response as opposed to a continuous type of response. The number of observations per child was not constant, say n. But rather there are n_i observations corresponding to child i in general. The time points at which these observations were made were not equally spaced in days rather in general the j^{th} observation Y_{ij} for child i was in general made at time t_{ij}. Furthermore the t_{ij} were different for different children. Thus we have a highly unbalanced data set which

requires proper modelling procedures in order to come up with meaningful conclusions. The relevant model required for the data falls in the general class of models for repeated discrete (non-Gaussian) data. Nevertheless because of the similarities and dis-similarities between models for those data types and those for repeated continuous data, this chapter will present an overview of methods applicable to linear mixed models for longitudinal continuous (assumed to be Gaussian) data. Methods specific to the analysis of non-Gaussian data will be developed in subsequent chapters taking into account the necessary departure from Gaussian data.

3.2 A 2-stage model formulation of the linear mixed model

Let the vector Y_i of the n_i observations for individual or cluster i be from a continuous variable with a Gaussian distribution. Because of the scenario just described above, that of unequal number of measurements per subject and that the measurements are not taken at fixed time points, multivariate regression techniques are often not applicable (Fahrmeir and Tutz, 1994). However the subject specific longitudinal profiles can be well approximated by linear regression functions. This leads to what is popularly known as a 2-stage model formulation where in stage 1 a linear regression model for each subject is defined and in stage 2, an attempt is made to explain the variability in the subject specific regression coefficients using known covariates (Verbeke and Molenberghs, 2000). In this chapter the approach by Verbeke and Molenberghs (2000) is adopted in formulating the linear mixed model for continuous data. Thus in general let the vector $\boldsymbol{Y_i}$ corresponding to observations from subject i be

$$\boldsymbol{Y_i} = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$$

then the stage 1 models are given by

$$\boldsymbol{Y_i} = Z_i\boldsymbol{\beta_i} + \boldsymbol{\epsilon_i} \tag{3.1}$$

for $i = 1. \ldots, N$. Here Z_i is an $n_i \times q$ matrix of known covariates and $\boldsymbol{\beta}_i$ is a q-dimensional vector of subject specific regression coefficients. It is assumed that

$$\boldsymbol{\epsilon_i} \sim N(\mathbf{0}, \Sigma_i)$$

and often $\Sigma_i = \sigma^2 I_{n_i}$ implying that the observations Y_{ij} are uncorrelated.
It should be noted that the above model describes the observed variability within subjects. In stage 2 the between subject variability can now be incorporated by relating $\boldsymbol{\beta_i}$ to known covariates. That is

$$\boldsymbol{\beta_i} = K_i\boldsymbol{\beta} + \boldsymbol{b_i} \tag{3.2}$$

where K_i is a $q \times p$ matrix of known covariates and $\boldsymbol{\beta}$ is a p-dimensional vector of unknown regression parameters and $\boldsymbol{b_i}$ is a vector that is $q \times 1$. It is assumed that

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

Substituting Eq. (3.2) into Eq. (3.1) leads to the following final expression for Y_i. Namely

$$\begin{aligned} \boldsymbol{Y_i} &= Z_iK_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} + \boldsymbol{\epsilon_i} \\ &= X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} + \boldsymbol{\epsilon_i} \end{aligned}$$

Thus the general linear mixed effects model can generally be represented in the form

$$\boldsymbol{Y_i} = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} + \boldsymbol{\epsilon_i} \tag{3.3}$$

where $\boldsymbol{b_i} \sim N(\mathbf{0}, G)$ denotes the individual random effects assumed to be normally distributed with a mean vector of zero and a variance matrix G. The vector $\boldsymbol{\epsilon_i} \sim N(\mathbf{0}, \Sigma_i)$ of dimension $n_i \times 1$ denotes the measurement error and $\boldsymbol{b_1}, \ldots, \boldsymbol{b_N}$, $\boldsymbol{\epsilon_1}, \ldots, \boldsymbol{\epsilon_N}$ are assumed to be independent. The elements in G and Σ_i are known as variance components. Within this formulation it is therefore crucially important to distinguish between the hierarchical and the marginal model.

3.3 Hierarchical versus Marginal Model

Note that under the above formulation of the general linear mixed model the conditional distribution of Y_i given b_i is given by

$$\boldsymbol{Y_i}|\boldsymbol{b_i} \sim N(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i}, \Sigma_i) \tag{3.4}$$

where

$$\boldsymbol{b_i} \sim N(\mathbf{0}, G). \tag{3.5}$$

This model formulation given jointly by Eq. (3.5) and Eq. (3.4) is therefore called a hierarchical model giving a probability model of $\boldsymbol{Y_i}$ given $\boldsymbol{b_i}$. It is clear that marginally Y_i is distributed as

$$\boldsymbol{Y_i} \sim N(X_i\boldsymbol{\beta}, Z_iGZ_i^{'} + \Sigma_i).$$

Under the above marginal model, very specific assumptions are made about the dependence of the mean and covariance on the covariates X_i and Z_i.

The implied mean is $X_i\boldsymbol{\beta}$ while the implied covariance is $V_i = Z_i G Z_i' + \Sigma_i$. It is also crucially important to note that the hierarchical model implies the marginal one, but not vice versa. As an example, consider the case where individuals are randomized into three possible doses, low, high and control dose respectively denoted by L, H and C. Suppose the interest is to understand the evolution of a continuous response Y_{ij} measured on subject i at time occasions t_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, N$. Then the stage 1 model can be written as

$$Y_{ij} = \beta_{1i} + \beta_{2i} t_{ij} + \epsilon_{ij}$$

for $j = 1, \ldots, n_i$. Let the stage 2 model be given by

$$\begin{cases} \beta_{1i} = & \beta_0 + b_{1i} \\ \beta_{2i} = & \beta_1 L_i + \beta_2 H_i + \beta_3 C_i + b_{2i} \end{cases}$$

where the parameters β_{1i} and β_{2i} are subject specific while β_0, β_1, β_2 and β_3 are common to all subjects. Then the combined model is

$$Y_{ij} = (\beta_0 + b_{1i}) + (\beta_1 L_i + \beta_2 H_i + \beta_{3i} C_i + b_{2i}) t_{ij} + \epsilon_{ij}$$

such that

$$Y_{ij} = \begin{cases} \beta_0 + b_{1i} + (\beta_1 + b_{1i}) t_{ij} + \epsilon_{ij}, & \text{if low dose} \\ \beta_0 + b_{1i} + (\beta_2 + b_{2i}) t_{ij} + \epsilon_{ij}, & \text{if high dose} \\ \beta_0 + b_{1i} + (\beta_3 + b_{3i}) t_{ij} + \epsilon_{ij}, & \text{if control} \end{cases}$$

The implied marginal mean structure is a linear average evolution in each group with equal average intercepts but different average slopes. The unknown fixed effects parameters are $\beta_0, \beta_1, \beta_2$ and β_3. The random effects parameters are the b_{1i}'s denoting the the random deviations from the common

intercept β_0 and the b_{2i} denoting the random deviations from the common group specific slopes. Note that as stated earlier, the assumption under the continuous response is that the vector $(b_{1i}, b_{2i})'$ is distributed as a bivariate normal distribution that is $N(\mathbf{0}, G)$ where G is a 2×2 variance-covariance matrix given generally by

$$G = \begin{pmatrix} g_{11} & g_{12} \\ g_{12} & g_{22} \end{pmatrix}$$

Under this model the implied covariance structure assuming $\Sigma_i = \sigma^2 I_{n_i}$ is given by

$$\begin{aligned} \text{cov}(\boldsymbol{Y_i}(t_1), \boldsymbol{Y_i}(t_2)) &= (1 \;\; t_1) G \begin{pmatrix} 1 \\ t_2 \end{pmatrix} + \sigma^2 \delta_{\{t_1, t_2\}} \\ &= g_{22} t_1 t_2 + g_{12}(t_1 + t_2) + g_{11} + \sigma^2 \delta_{\{t_1, t_2\}} \end{aligned}$$

Note that the model structure implicitly assumes that the variance function is quadratic over time, with positive curvature g_{22}. Suppose instead the following stage 1 model was proposed

$$Y_{ij} = \beta_{1i} + \beta_{2i} t_{ij} + \beta_3 t_{ij}^2 + \epsilon_{ij}$$

and the stage 2 model is given by

$$\begin{cases} \beta_{1i} = & \beta_1 age_i + \beta_2 L_i + \beta_3 H_i + \beta_4 C_i + b_{1i} \\ \beta_{2i} = & \beta_5 age_i + \beta_6 L_i + \beta_6 H_i + \beta_7 C_i + b_{2i} \\ \beta_{3i} = & \beta_8 age_i + \beta_9 L_i + \beta_{10} H_i + \beta_{11} C_i + b_{3i} \end{cases}$$

The implied marginal mean structure would be a quadratic evolution in each group. The average intercept, linear and quadratic slopes are now corrected for age differences. The implied marginal covariance is now:

$$
\begin{aligned}
\mathrm{cov}(\boldsymbol{Y_i}(t_1), \boldsymbol{Y_i}(t_2)) &= (1 \;\; t_1 \;\; t_2^2) G \begin{pmatrix} 1 \\ t_2 \\ t_2^2 \end{pmatrix} + \sigma^2 \delta_{\{t_1, t_2\}} \\
&= g_{33} t_1^2 t_2^2 + g_{23}(t_1^2 t_2 + t_1 t_2^2) + g_{22} t_1 t_2 \\
&+ g_{13}(t_1^2 + t_2^2) + g_{12}(t_1 + t_2) + g_{11} + \sigma^2 \delta_{\{t_1, t_2\}}
\end{aligned}
$$

Note that now the variance-covariance matrix for (b_{1i}, b_{2i}, b_{3i}) is a 3×3 matrix with elements g_{kl} where $l, k = 1, 2, 3$ and the implied variance function is now a fourth order polynomial over time.

3.4 A model for the residual covariance structure

Most often Σ_i is taken as $\sigma^2 I_{n_i}$. This implies the conditional independence assumption, that is conditional on the random effects $\boldsymbol{b_i}$, the elements in the vector of observations $\boldsymbol{Y_i}$ are independent. However in the presence of no, or little, random effects the assumption of conditional independence is often unrealistic. For example, the random intercepts model not only implies constant variance, it also implicitly assumes constant correlation between any two measurements within subjects since in this case $\mathrm{Var}(Y_{ij}) = g + \sigma^2$ and $\mathrm{Cov}(Y_{ij}, Y_{ij'}) = g$ for $i \neq j'$. Hence when there is no evidence for (additional) random effects, or if they would have no substantive meaning, the correlation structure can be almost wholly accounted for in an appropriate model for Σ_i. An example of a frequently used model is:

$$
\boldsymbol{Y_i} = X_i \boldsymbol{\beta} + Z_i \boldsymbol{b_i} + \boldsymbol{\varepsilon_{(1)i}} + \boldsymbol{\varepsilon_{(2)i}} \tag{3.6}
$$

where $\boldsymbol{b_i}$ is a term accounting for between subject variability, $\boldsymbol{\varepsilon_{(1)i}}$ account for measurement error and $\boldsymbol{\varepsilon_{(2)i}}$ is the serial correlation component. This last term represents the belief that part of an individual's observed profile is a response to a time varying stochastic process operating within an individual. This results in a correlation between serial measurements, which is usually a decreasing function of the time separation between these measurements. The correlation matrix D_i of $\boldsymbol{\varepsilon_{(2)i}}$ is assumed to have a general (j, k) element of the form

$$d_{ijk} = h(|t_{ij} - t_{ik}|) \tag{3.7}$$

for some decreasing function of $h(.)$ with $h(0) = 1$. Some frequently used functional forms of $h(.)$ are the exponential decay serial correlation, $h(w) = \exp(-\phi w)$ and the Gaussian serial correlation $h(w) = \exp(-\phi w^2)$. Extreme cases for both types is when $\phi = +\infty$ implying the components in $\varepsilon_{(2)i}$ are independent and the case when $\phi = 0$ meaning that the components in $\varepsilon_{(2)i}$ are perfectly correlated. In general the smaller ϕ is, the stronger the serial correlation is. The resulting linear mixed model is then given by:

$$\boldsymbol{Y_i} = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} + \boldsymbol{\varepsilon_{(1)i}} + \boldsymbol{\varepsilon_{(2)i}} \tag{3.8}$$

where

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

$$\boldsymbol{\varepsilon_{(1)i}} \sim N(\mathbf{0}, \sigma^2 I_{n_i}),$$

$$\boldsymbol{\varepsilon_{(2)i}} \sim N(\mathbf{0}, \tau^2 H_{n_i}).$$

3.4.1 The mean structure

For balanced data, an average can be calculated for each occasion separately, and standard errors for the means calculated. Plots of such summary quantities can help to tell whether there is a linear or non-linear average trend.

Increasing standard errors in this case can be due to dropouts which is a common feature in longitudinal or repeated measurements data. For unbalanced data, the time scale can be discretized and simple averaging within intervals calculated. For example in the case of the Respiratory Syncytial Virus (RSV) the time scale was discretized into monthly and weekly intervals and the average prevalence was calculated for each month. Since the data is non-normal, the mean corresponds to the weekly and monthly average prevalence (See Figure 2.3 and 2.4, pages 35 and 38 in Chapter 1). Smoothing techniques such as the Loess smoothing in SAS or S-Plus and other standard statistical software can be used to estimate the average evolution non-parametrically. If important covariates or factors are known, similar plots can be constructed for subgroups with different values for these covariates or factors. For example, given gender status for the children affected by RSV, weekly and monthly prevalences can be constructed for each sex.

3.5 The variance structure

In general the variance function dependent on time equals

$$\sigma^2(t) = E[Y(t) - \mu(t)]^2 \tag{3.9}$$

Hence an estimate for $\sigma^2(t)$ can be obtained by applying any of the techniques used for exploring the mean structure to squared residuals, r_{ij}^2. For balanced longitudinal data, the correlation structure can be studied through the correlation matrix, or scatter plot matrix. Graphically, pairwise scatter plots can be used for exploring the correlation between any two repeated measurements.

3.5.1 The semi-variogram

For unbalanced data, the same approach can be used, after discretizing the time scale. An alternative method, is the case when the variance function suggest constant variance. We reconsider the general linear mixed model,

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

where

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

$$\boldsymbol{\varepsilon_{(1)i}} \sim N(\mathbf{0}, \sigma^2 I_{n_i}),$$

$$\boldsymbol{\varepsilon_{(2)i}} \sim N(\mathbf{0}, \tau^2 H_{n_i}).$$

are all independent. Based on the knowledge of the mean function, residuals

$$r_{ij} = y_{ij} - \mu(t_{ij})$$

can be obtained. Thus based on the model above the residuals including all terms are assumed to follow the model

$$\boldsymbol{r_i} = Z_i\boldsymbol{b_i} + \boldsymbol{\varepsilon_{(1)i}} + \boldsymbol{\varepsilon_{(2)i}}$$

where $\boldsymbol{r_i} = (r_{i1}, r_{i2}, \ldots, r_{in_i})'$. The semi-variogram assumes constant variance, which implies that the only random effects in the model will at most be intercepts that is, $Z = (1 \ldots 1)'$. If we denote the variance of the random intercepts by ν^2 then the covariance matrix for the n_i observations from subject i takes the form

$$V_i = \mathrm{Var}(\boldsymbol{Y_i}) = \mathrm{Var}(\boldsymbol{r_i}) = \nu^2 Z_i Z_i' + \sigma^2 I_{n_i} + \tau^2 H_i.$$

It therefore follows that the residuals r_{ij} have constant variance $\nu^2 + \sigma^2 + \tau^2$. The correlation between any two residuals r_{il} and r_{ik} from the same subject

indexed i is given by

$$\rho(|t_{il} - t_{ik}|) = \frac{\nu^2 + \tau^2 h(|t_{il} - t_{ik}|)}{\nu^2 + \sigma^2 + \tau^2} \tag{3.10}$$

since the covariance between any two observations is $\nu^2 + \tau^2 h(|t_{il} - t_{ik}|)$. It can easily be shown that for $l \neq k$

$$\frac{1}{2}E(r_{il} - r_{ik})^2 = \sigma^2 + \tau^2(1 - h(|t_{il} - t_{ik}|)) = \gamma(w_{ikl}). \tag{3.11}$$

The function $\gamma(w)$ is called the semi-variogram, and it only depends on the time points t_{ij} through the time lags $w_{ilk} = |t_{il} - t_{ik}|$. Note that a decreasing serial correlation function $h(.)$ yields an increasing semi-variogram $\gamma(w)$ such that $\gamma(0) = \sigma^2$ and converges to $\sigma^2 + \tau^2$ as w tends to infinity.

Obviously an estimate of $\gamma(w)$ can be used to explore the relative importance of the stochastic components b_i, $\varepsilon_{(1)i}$ and $\varepsilon_{(2)i}$ as well as the nature of the serial correlation function $h(.)$. An estimate of $\gamma(w)$ is obtained from smoothing the scatter plot of the $\sum_{i=1}^{N} \frac{n_i(n_i-1)}{2}$ half squared differences $u_{ikl} = \frac{(r_{il}-r_{ik})^2}{2}$ between pairs of residuals within subjects versus the corresponding time lags $w_{ijk} = |t_{il} - t_{ik}|$. It can also be shown that for $i \neq j$ that is for two different individuals and $l \neq k$, that

$$\frac{1}{2}E(r_{il} - r_{jk})^2 = \sigma^2 + \tau^2 + \nu^2$$

because $E(r_{il}r_{jk}) = 0$. This means that the total variability in the data (assumed to be constant) can be estimated by

$$\hat{\sigma^2} + \hat{\tau^2} + \hat{\nu^2} = \frac{1}{2N^*} \sum_{i \neq j} \sum_{l=1}^{n_j} \sum_{k=1}^{n_i} (r_{il} - r_{jk})^2 \tag{3.12}$$

where N^* is the number of terms in the sum. Recall as discussed before, the linear mixed model is often derived using a 2 stage model formulation. This is based on a good approximation of the subject specific profiles by linear

regression models. This therefore emphasizes the need to use exploratory methods for longitudinal data to confirm this assumption. A natural way to explore longitudinal profiles is by plotting them (See randomly selected profiles in Fig 2.1 in Chapter 2 for the RSV data). For large data sets the subjects can be ordered according to subject profile characteristics (mean, variability etc.) then plot the profiles from some subjects.

Some ad hoc statistical procedures for checking the assumption of linear regression models used in the first stage formulation can be used. These include extensions of the classical linear regression techniques such as the coefficient R^2 of multiple determination and a formal test for model extension.

In linear regression

$$R^2 = \frac{SSTOT - SSE}{SSTO}$$

where $SSTO$ and SSE denote respectively the total sum of squares and residual sum of squares. It follows that subject-specific coefficients can be obtained as

$$R_i^2 = \frac{SSTOTi - SSEi}{SSTOTi}.$$

Histograms of R_i^2 or scatter plots of R_i^2 can be used to investigate the visual adherence of the linear model across the subject specific regression models. The overall or pooled estimate of the coefficient of multiple determination is given by

$$R_{meta}^2 = \frac{\sum_{i=1}^{n}(SSTOTi - SSEi)}{\sum_{i=1}^{n} SSTOTi}.$$

A SAS macro is available to work out R_{meta}^2 in the case of the linear mixed model.

3.5.2 Test for model extension

The aim is to test for the need of extending the linear regression model $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with additional covariates in X^*. Let the model containing the full set of covariates $(p + p^*)$ be denoted by MF and the reduced model with the reduced set of covariates p in number be MR. Then the general test statistic for comparing the models is denoted by:

$$F = \frac{(SSE(MR) - SSE(MF))/p^*}{SSE(MF)/(n - p - p^*)}.$$

Thus the overall test for the need to extend the stage 1 model is

$$F_{meta} = \frac{\{\sum_{\{i:n_i \geq p+p^*\}}(SSE_i(MR) - SSE_i(MF))\}/\{\sum_{\{i:n_i \geq p+p^*\}} p^*\}}{\{\sum_{\{i:n_i \geq p+p^*\}} SSE_i(MF)\}/\{\sum_{\{i:n_i \geq p+p^*\}}(n_i - p - p^*)\}}$$

Under the null distribution the test statistic is distributed as F with $\sum_{\{i:n_i \geq p+p^*\}} p^*$ and $\sum_{\{i:n_i \geq p+p^*\}}(n_i - p - p^*)$ degrees of freedom. Note that this statistic requires $n_i \geq p+p^*$. Otherwise subjects with $n_i < p+p^*$ cannot contribute to this statistic. Again it is noted that a SAS macro is available for carrying out this test.

3.6 Estimation of the marginal model

3.6.1 Maximum likelihood estimation (ML) of the variance components

Recall that the general linear mixed model is of the form

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

with the usual assumptions on $\boldsymbol{b_i}$, $\boldsymbol{\varepsilon_{(1)i}}$ and $\boldsymbol{\varepsilon_{(2)i}}$. As stated earlier the implied marginal model is given by

$$\boldsymbol{Y_i} \sim N(X_i\boldsymbol{\beta}, Z_iGZ_i^{'} + \Sigma_i).$$

It is therefore important to note that inferences based on the marginal model do not explicitly assume the presence of random effects representing the natural heterogeneity between subjects. Let $\boldsymbol{\beta}$ denote the vector of fixed effects and let $\boldsymbol{\alpha}$ be a vector of all variance components in G and Σ_i. Then the variance covariance matrix V_i of $\boldsymbol{Y}_i$ is $\boldsymbol{\alpha}$ dependent. Thus we can let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')$ denote the vector of all parameters in the marginal model. It then follows that the marginal likelihood function assuming the above assumptions hold is

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

where $V_i(\boldsymbol{\alpha})$ is the matrix of variance components. If $\boldsymbol{\alpha}$ were known then according to Harville (1974) the MLE of $\boldsymbol{\beta}$ equals

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = (\sum_{i=1}^{n} X_i W_i X_i')^{-1} \sum_{i=1}^{n} X_i' W_i \boldsymbol{y_i}.$$

where $W_i = V_i^{-1}$. Within this framework $\hat{\boldsymbol{\alpha}_{\boldsymbol{ML}}}$ are obtained by maximising

$$L_{ML}(\boldsymbol{\alpha}, \hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}))$$

with respect to $\boldsymbol{\alpha}$. The resulting estimates $\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha_{ML}})$ for $\boldsymbol{\beta}$ are denoted by $\hat{\boldsymbol{\beta}_{\boldsymbol{ML}}}$. Alternatively $\boldsymbol{\alpha_{ML}}$ and $\boldsymbol{\beta_{ML}}$ can be obtained from maximizing $L_{ML}(\theta)$ with respect to $\boldsymbol{\theta}$ that is with respect to both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ simultaneously.

3.6.2 Restricted maximum likelihood estimation (REML)

The maximum likelihood estimation of the variance components does not account for the loss of degrees of freedom used in estimating the fixed parameters, hence the need for an alternative approach such as REML. First for purposes of clarity consider the univariate cross-sectional simple case of a

normal population of n observations $Y_1, \ldots, Y_n$ from $N(\mu, \sigma^2)$. If μ is known, the MLE of σ^2 equals

$$\hat{\sigma}^2 = \sum_i (Y_i - \mu)^2/n$$

and $\hat{\sigma}^2$ is unbiased for σ^2. However when μ is unknown, the MLE of σ^2 becomes

$$\hat{\sigma}^2 = \sum_i (Y_i - \bar{Y})^2/n$$

and $\hat{\sigma}^2$ is biased for σ^2 since

$$E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2.$$

Nonetheless when μ is unknown an unbiased estimate can still be found and is given by

$$S^2 = \sum_i (Y_i - \bar{Y})^2/(n-1)$$

Apparently the simple example above shows that having to estimate μ introduces bias in the estimation of σ^2. Now what follows, is a procedure of how to estimate σ^2 without estimating μ first. Note that the model for the data can be written as

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}, \sigma^2 I_n \right)$$

Now transform the data vector $\boldsymbol{Y}$ such that μ vanishes from the likelihood. Let

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

where $A^{'}$ is an $(n-1)$x n matrix of the form

$$A^{'} = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \\ 0 & 0 & 0 & \dots & 1 & -1 \end{pmatrix}$$

The MLE of σ^2 based on the data transformation $\boldsymbol{U}$, is equal to

$$S^2 = \frac{1}{n-1}\sum_i (Y_i - \bar{Y})^2.$$

The matrix $A^{'}$ defines a set of $n-1$ linearly independent error contrasts. S^2 is called the restricted maximum likelihood estimate (REML) of σ^2 and S^2 is independent of A. Now consider the estimation of residual variance in linear regression. Let $Y_1, \dots, Y_n$ denote a random sample of observations from a linear regression model namely from a population which is $N(X\boldsymbol{\beta}, \sigma^2 I)$. It is well known that the MLE of σ^2 is

$$\hat{\sigma}^2_{ML} = (\boldsymbol{Y} - X\hat{\boldsymbol{\beta}})^{'}(\boldsymbol{Y} - X\hat{\boldsymbol{\beta}})/n.$$

It can be further shown that $\hat{\sigma}^2$ is biased for σ^2 since

$$E(\hat{\sigma}^2) = \frac{n-p}{n}\sigma^2 \neq \sigma^2$$

An unbiased estimator for σ^2 is the mean sum of squares due to error (MSE) given by

$$MSE = (\boldsymbol{Y} - X\hat{\boldsymbol{\beta}})^{'}(\boldsymbol{Y} - X\hat{\boldsymbol{\beta}})/(n-p) = \hat{\sigma}^2_{REML}.$$

The MSE can also be obtained from transforming the data orthogonal to X such that

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

The MLE of σ^2, based on U, now equals the mean squared error, MSE. The MSE is again called the REML estimator of σ^2. Note that $\hat{\sigma}^2_{REML} > \hat{\sigma}^2_{ML}$ since $\hat{\sigma}^2_{REML} = \frac{n}{n-p}\hat{\sigma}^2_{ML}$ and $\frac{n}{n-p} > 1$.

3.6.3 REML estimation for the linear mixed model

For easy manipulation combine all models of the form

$$\boldsymbol{Y_i} \sim N(X_i\boldsymbol{\beta}, V_i)$$

where $V_i = Z_i G Z_i^{'} + \Sigma_i$ into one model given by

$$\boldsymbol{Y} \sim N(X\boldsymbol{\beta}, V)$$

where

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y_1} \\ \vdots \\ \boldsymbol{Y_n} \end{pmatrix},\ X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix},\ V(\boldsymbol{\alpha}) = \begin{pmatrix} V_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & V_n \end{pmatrix}$$

Using a similar approach as in the simple case the data can be transformed to be orthogonal to X so that

$$\boldsymbol{U} = A^{'}\boldsymbol{Y} \sim N(\boldsymbol{0}, A^{'}V(\boldsymbol{\alpha})A)$$

The MLE of $\boldsymbol{\alpha}$, based on $\boldsymbol{U}$ is called the REML estimate and is denoted by $\hat{\boldsymbol{\alpha}}_{REML}$. The resulting estimate $\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}_{REML})$ for $\boldsymbol{\beta}$ will be denoted by $\hat{\boldsymbol{\beta}}_{REML}$. Note that the estimates $\hat{\boldsymbol{\alpha}}_{REML}$ and $\hat{\boldsymbol{\beta}}_{REML}$ can be obtained by maximizing

$$L_{REML}(\boldsymbol{\theta}) = \left| \sum_{i=1}^{n} X_i W_i(\boldsymbol{\alpha}) X_i^{'} \right|^{-\frac{1}{2}} L_{ML}(\boldsymbol{\theta})$$

with respect to $\boldsymbol{\theta} = (\boldsymbol{\beta}^{'}, \boldsymbol{\alpha}^{'})$. Although strictly speaking not a likelihood $L_{REML}(\boldsymbol{\theta})$ is therefore still called the REML likelihood function.

3.6.4 Fitting Linear Mixed Models

Mixed models are fitted using the PROC MIXED in SAS. One can specify the required estimation method as maximum likelihood (ML) or restricted maximum likelihood (REML) which is the default method in SAS. The CLASS statment is for defining factors in the model.

The MODEL statement helps to specify the response variables and fixed effects only. Options are similar to SAS regression procedures. The RANDOM statement is for defining random effects including intercepts. The RANDOM statement also helps to identify 'subjects' under the assumption of independence across subjects. The statement is also used to specify the type of random-effects covariance matrix G and the option "g" and "gcorr" are used to print out G and the corresponding correlation matrix. Further the options "v" and "vcorr" are used to print out V_i and the corresponding correlation matrix.

The purpose of the REPEATED statement is to order the measurements within subjects or clusters. The effects specified must be of the factor-type. The options under this statement also identifies the 'subjects' under the assumption of independence across subjects. The type of residual covariance matrix Σ_i is also specified, for example type=simple. The options "r" and "rcorr" help to print out Σ_i and the corresponding correlation matrix.

Some frequently used covariance structures available in the RANDOM and REPEATED statements include: Unstructured (type=UN),
Simple (type=SIMPLE), Compound Symmetry (type=CS) such as the random intercepts covariance structure and the split plot design covariance structure, Banded (type=UN(2)), First order Autoregressive type (type=AR(1)), Toeplitz (type=TOEP), Toeplitz(1)(type=TOEP(1)), Heterogenous Compound Symmetry(type=CSH), Heterogenous first order Autoregressive
(type=ARH(1)), Heterogenous Toeplitz(type=TOEPH).

When serial correlation is to be fitted, it should be specified in the REPEATED statement and the option "LOCAL" can be added to also include measurement error, if required. Some frequently used serial correlation structures available in RANDOM and REPEATED statements in-

clude Power (type=SP(POW)(list)), Exponential (type=SP(EXP)(list)) and Gaussian (type=SP(GAU)(list)).

Sometimes rescaling the time point t_{ij} may become necessary for efficient convergence in the estimation process. Negative variance components can be allowed in SAS using the option "nobound" to the PROC MIXED statement. A negative variance component may suggest a negative curvature in the variance function (a concave function). Again, this emphasizes the non-equivalence of hierarchical and marginal models. The marginal model allows the negative variance component, as long as the marginal covariance matrices V_i are positive definite. The hierarchical interpretation of the model does not allow negative variance components because $b_i \sim N(0, G)$.

3.7 Inference for the marginal model

Inference for the fixed effects $\boldsymbol{\beta}$ can be based on the Wald test, t-test, F-test, robust inference or the likelihood ratio (LR) test. Inference for the variance components is based on the Wald test and the LR test. The information criteria can generally be useful for making inference about the marginal model.

The estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = (\sum_{i=1}^{n} X_i' W_i X_i)^{-1} \sum_{i=1}^{n} X_i' W_i \boldsymbol{y_i} \tag{3.13}$$

with $\boldsymbol{\alpha}$ being replaced by its ML or REML estimate according Harville (1974) and Laird and Ware (1982). Conditional on $\boldsymbol{\alpha}$, $\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha})$ is multivariate normal

with mean $\boldsymbol{\beta}$ and covariance

$$\begin{aligned}\text{Var}(\hat{\boldsymbol{\beta}}) &= (\sum_{i=1}^{n} X_i' W_i X_i)^{-1} (\sum_{i=1}^{n} X_i' W_i (\text{Var}\boldsymbol{Y_i}) W_i X_i) (\sum_{i=1}^{n} X_i' W_i X_i)^{-1} \\ &= \left(\sum_{i=1}^{n} X_i' W_i X_i\right)^{-1} \end{aligned} \tag{3.14}$$

provided that $W_i = V_i^{-1}$ where $V_i = \text{Var}(Y_i) = Z_i G Z_i + \Sigma_i$.

3.7.1 Approximate Wald test

Let L be a known contrast or transformation matrix and consider testing the hypothesis

$$H_0 : L\boldsymbol{\beta} = \mathbf{0}$$

versus

$$H_A : L\boldsymbol{\beta} \neq \mathbf{0} \tag{3.15}$$

Then the Wald test statistic is given by

$$W_T = \hat{\boldsymbol{\beta}}' L' [L \left(\sum_{i=1}^{n} X_i' V_i^{-1}(\hat{\boldsymbol{\alpha}}) X_i\right) L']^{-1} L\hat{\boldsymbol{\beta}}$$

The asymptotic sum distribution of W_T is chi-square distributed with rank(L) degrees of freedom. Thus using the statistic W_T inference on fixed effects can be made via the transformation $L\boldsymbol{\beta}$.

3.7.2 Approximate t-test and F-test

It should be noted that the Wald test is based on

$$\text{var}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{n} X_i' W_i(\boldsymbol{\alpha}) X_i\right)^{-1}$$

The deficiency with the Wald test statistic is that the variability introduced by replacing α by some estimate (ML or REML) is not taken into account in

the subsequent test. Therefore Wald tests will only provide valid inferences in sufficiently large samples. In practice, this is often resolved by replacing the χ^2 distribution by an appropriate F distribution. Thus to test the hypothesis H_0 versus H_A in Eq. (3.14), the above statistic becomes

$$F_T = \frac{\hat{\boldsymbol{\beta}}' L' [L \left(\sum_{i=1}^{n} X_i' V_i^{-1}(\hat{\boldsymbol{\alpha}}) X_i \right) L']^{-1} L \hat{\boldsymbol{\beta}}}{\text{rank}(L)}$$

The approximate null distribution of F_T is F with numerator degrees of freedom equal to rank(L). The denominator degrees of freedom have to be estimated from the data using common methods such as the containment method, the Sattherwaite approximation and the Kenward and Roger approximation. In the context of longitudinal data, all methods typically lead to large degrees of freedom, and therefore also very similar p-values. For univariate hypotheses, rank(L)=1 and in this case the F-test is equivalent reduces to a t-test. Linear hypotheses of the form given by Eq. (3.14) can be tested in SAS using a CONTRAST statement. The option "chisq" in the CONTRAST statement is needed in order to obtain a Wald test. SAS Proc Mixed also allows the estimation and testing of linear combinations of the elements in β using an ESTIMATE statement. Using similar arguments as for approximate Wald tests, t-tests, and F-tests, approximate confidence intervals can be obtained for such linear combinations, also implemented in the ESTIMATE statement. Specification of L remains the same as for the CONTRAST statement.

3.7.3 Robust Inference

Given the estimate for $\boldsymbol{\beta}$ in Eq. (3.13) with $\boldsymbol{\alpha}$ replaced by its ML or REML estimates then conditional on $\boldsymbol{\alpha}$, $\hat{\boldsymbol{\beta}}$ has the expected value given by,

$$\begin{aligned} \mathrm{E}[\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha})] &= \left(\sum_{i=1}^{n} X_i' W_i X_i\right)^{-1} \sum_{i=1}^{n} X_i W_i \mathrm{E}(\boldsymbol{Y_i}) \\ &= \left(\sum_{i=1}^{n} X_i' W_i X_i\right)^{-1} \sum_{i=1}^{n} X_i W_i X_i \boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

provided that the $E(\boldsymbol{Y_i}) = X_i\boldsymbol{\beta}$. Hence in order for $\hat{\beta}$ to be unbiased, it is only sufficient that the mean of the response is correctly specified. Conditional on α, $\hat{\beta}$ has covariance, $\mathrm{var}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}(X_i' W_i X_i)^{-1}$ as derived in Eq. (3.14)

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\beta}}) &= (\sum_{i=1}^{n} X_i' W_i X_i)^{-1} \sum_{i=1}^{n} (X_i' W_i \mathrm{Var}(\boldsymbol{Y_i}) W_i X_i) (\sum_{i=1}^{n} X_i' W_i X_i)^{-1} \\ &= (\sum_{i=1}^{n} X_i' W_i X_i)^{-1} \end{aligned}$$

Note that this assumes that the covariance matrix is correctly modelled as $\mathrm{Var}(\boldsymbol{Y_i}) = V_i = Z_i G Z_i' + \Sigma_i$ and $W_i = V_i^{-1}$. This form of the covariance estimate is therefore often called the 'naive' estimate. The so-called robust estimate for $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ which does not assume the covariance matrix to be correctly specified is obtained by replacing $\mathrm{Var}(\boldsymbol{Y_i})$ by

$$\widetilde{\mathrm{Var}(\boldsymbol{Y_i})} = [\boldsymbol{Y_i} - X_i\boldsymbol{\beta}][\boldsymbol{Y_i} - X_i\boldsymbol{\beta}]'$$

rather than V_i . The only condition for $\widetilde{\mathrm{Var}(\boldsymbol{Y_i})}$ to be unbiased for $\mathrm{Var}(\boldsymbol{Y_i})$ is that the mean is correctly specified. The 'robust' variance estimate also called the sandwich estimate is now given by

$$\mathrm{Var}(\hat{\beta}) = (\sum_{i=1}^{n} X_i' W_i X_i)^{-1} (\sum_{i=1}^{n} X_i' W_i \widetilde{\mathrm{Var}(\boldsymbol{Y_i})} W_i X_i) (\sum_{i=1}^{n} X_i' W_i X_i)^{-1}$$

Based on this sandwich estimate, robust versions of the Wald test as well as of the t-test and the F-test can be obtained. This signifies the point that as long as interest is only in the inferences in the mean structure, little effort should be spent in modelling the exact covariance structure, provided that the data set is sufficiently large. An extreme point of view involves the use of OLS with robust standard errors. Nevertheless appropriate covariance modelling may still be of interest, firstly for the purpose of interpretation of random variation in the data, secondly for gaining efficiency and thirdly because in the presence of missing data, robust inference is only valid under very severe assumptions about the underlying missingness process. Issues of missingness were discussed briefly in Chapter 2 and will be revisited in more detail in Chapter 9. Robust inference for the fixed effects can be obtained by adding the option 'empirical' in the PROC MIXED statement in SAS namely

```
proc mixed data=data1 method=reml empirical;
```

assuming the data set is 'data 1'. It is quite possible that for some parameters, the robust standard error is smaller than the naive, model based one. For others the opposite can be true. Thus interpretation of both standard errors should be done with caution.

3.7.4 Likelihood ratio test

The likelihood ratio tests are used to compare nested models with different mean structures, but equal covariance structures. The null hypothesis of interest can therefore be stated as

$$H_0 : \boldsymbol{\beta} \in \Theta_{\beta,0}$$

for some subspace $\Theta_{\beta,0}$ of the parameter space Θ_β of the fixed effects $\boldsymbol{\beta}$. Let the notations $L_{ML}, \hat{\boldsymbol{\theta}}_{ML,0}$ and $\hat{\boldsymbol{\theta}}_{ML}$ respectively denote the maximum likelihood (ML) function, the maximum likelihood estimator(MLE) under H_0 and under the general model. Then the test statistic under the LR method is

$$-2\ln\lambda_n = -2\ln\left[\frac{L_{ML}(\hat{\boldsymbol{\theta}}_{ML,0})}{L_{ML}(\hat{\boldsymbol{\theta}}_{ML})}\right].$$

The asymptotic distribution of the statistic under the null distribution is χ^2 with degrees of freedom (df) equal to the difference in dimension of Θ_β and $\Theta_{\beta,0}$ that is

$$\dim\Theta_\beta - \dim\Theta_{\beta,0}.$$

It should be noted that LR tests for the mean structure are not valid under REML. A negative LR test statistic is a very possible outcome under REML. The reason is as follows: under REML the response $\boldsymbol{Y}$ is transformed into error contrasts $\boldsymbol{U} = A'\boldsymbol{Y}$, for some matrix A with $A'X = 0$. Afterwards ML estimation is performed based on error contrasts. Models with different mean structures lead to different sets of error contrasts. Hence the corresponding REML likelihoods are based on different observations, which makes them no longer comparable.

3.8 Inference for variance components

Inference for the mean structure is usually the primary goal in most research problems. However, inference for the covariance structure is of interests as well. This is necessary for the interpretation of the random variation in the data. It is important to note that an overparameterised covariance structure leads to inefficient inference of the mean structure. On the other hand too

restrictive models invalidate inferences for the mean structure. The challenge is to strike a balance between these two extremes.

3.8.1 Approximate Wald test

Asymptotically, ML and REML estimates of α are normally distributed with correct mean and inverse Fisher information matrix as covariance. Hence approximate standard errors and Wald tests can easily be obtained. Standard errors and approximate Wald tests for variance components can be obtained in PROC MIXED by adding the option 'covtest' to the PROC MIXED statement in SAS namely

proc mixed data=data1 method=reml covtest

assuming 'data1' is already loaded in SAS. Caution for Wald tests for variance components arises due to the difference between marginal and hierarchical models. When no underlying random effects structure is believed to represent the observed variation between subjects, then the Wald tests can only be fully interpreted under the marginal model.

3.8.2 The likelihood ratio test for tests on variance components

The quality of the normal approximation for the ML and REML estimates strongly depends on the true value of $\boldsymbol{\alpha}$. The normal approximation fails and performs poorly if $\boldsymbol{\alpha}$ is relatively close to the boundary of the parameter space. If $\boldsymbol{\alpha}$ is a boundary value, the normal approximation fails completely. In this current problem the LR tests are ideal for the comparison of nested models with equal mean structures, but different covariance structure. Let

the hypothesis of interest be

$$H_0 : \boldsymbol{\alpha} \in \Theta_{\alpha,0}$$

for some subspace $\Theta_{\alpha,0}$ of the parameter space Θ_α of the variance components $\boldsymbol{\alpha}$. Let the notations L_{ML}, $\hat{\boldsymbol{\theta}}_{ML,0}$ and $\hat{\boldsymbol{\theta}}_{ML}$ have similar meanings as was the case for the tests for fixed effects. Then the test statistic is given by

$$-2\ln\lambda_n = -2\ln\left[\frac{L_{ML}(\hat{\boldsymbol{\theta}}_{ML,0})}{L_{ML}(\hat{\boldsymbol{\theta}}_{ML})}\right].$$

The asymptotic distribution of the statistic under the null distribution is χ^2 with degrees of freedom equal to the difference in dimension of Θ_α and $\Theta_{\alpha,0}$ given by

$$\dim\Theta_\alpha - \dim\Theta_{\alpha,0}.$$

Note that as long as models being compared are with the same mean structure, a valid LR test can be obtained under REML as well. Both models can be fitted using the same error contrasts, making the likelihood comparable. Note that if, H_0 is a boundary value, the classical χ^2 approximation may not be valid. For some very specific null hypotheses on the boundary, the correct asymptotic null distribution has been derived.

3.8.3 Marginal testing for the need of random effects

Self and Liang (1987) and Stram and Lee (1994, 1995) state that tests for hypotheses such as in Eq.(3.16) below require the use of mixture distributions. They have been able to show that the the asymptotic null distribution for the likelihood ratio test statistic is often a mixture of chi-squared distributions rather than a single chi-squared distribution. This principle applies under a hierarchical model interpretation where the asymptotic null distribution for

the LR test statistic for testing significance of all variance components related to one or more multiple random effects, can be derived. Consider testing

$$H_0 : G = 0$$

versus

$$H_A : G_{11} = g_{11} \tag{3.16}$$

for some non negative scalar g_{11}. The asymptotic null distribution would be

$$-2ln\lambda_n \rightarrow \chi^2_{0,1}$$

which is a mixture of χ^2_0 and χ^2_1 with equal weights 0.5. The intuitive idea follows when one considers the extended parameter space R of g_{11}. Under H_0, g_{11} will be negative in the 50% of the cases leading to $\hat{g}_{11} = 0$. Hence overall $L_{ML}(\hat{\boldsymbol{\theta}}_{ML,0}) = L_{ML}(\hat{\boldsymbol{\theta}}_{ML})$. This idea can be extended generally to the case for the need of q versus $q + k$ random effects requiring a mixture of χ^2_q and χ^2_{q+k} with equal weights of 0.5. However simulations may become necessary to derive the asymptotic null distribution. This means that ignoring the boundary problem too often leads to over-simplified covariance structures. Failing to correct for the boundary problem inflates the p-value. Implying that ignoring the boundary problem may invalidate inferences, even for the mean structure.

3.9 Information Criteria

Definition of IC: LR tests can only be used to compare nested models. However at times there arises a need to compare non nested models. The general idea behind the LR test for comparing model A to a more extensive model B is to select model A if the increase in likelihood under model B

is small compared to increase in complexity. A similar argument is quite possible to compare non-nested models A to B. Here one selects the model with the largest (log-) likelihood provided it is not (too) complex. If ℓ is the log-likelihood under the given model, then the penalized log-likelihood is $\ell - \mathcal{F}(\#\boldsymbol{\theta})$ for some function$\mathcal{F}(.)$ of the number $(\#\boldsymbol{\theta})$ of parameters in the model. The criteria for model selection is to select the model with the highest penalized log-likelihood. Different functions $\mathcal{F}(.)$ lead to different criteria. These are summarised below

Criteria	Definition of $\mathcal{F}(.)^{\star}$
Akaike(AIC)	$\mathcal{F}(\#\boldsymbol{\theta}) = \#\boldsymbol{\theta}$
Schwarz(SBC)	$\mathcal{F}(\#\boldsymbol{\theta}) = (\#\boldsymbol{\theta} \ln n^*)/2$
Hannan and Quinn(HQIC)	$\mathcal{F}(\#\boldsymbol{\theta}) = \#\boldsymbol{\theta} \ln(\ln n^*)$
Bozdogan(CAIC)	$\mathcal{F}(\#\boldsymbol{\theta}) = \#\boldsymbol{\theta}(\ln n^* + 2)/2$

where $\star : n^* = \sum_{i=1}^{N} n_i$ under ML and $\star : n^* = n - p$ under REML. It should however be noted that information criteria are not formal testing procedures. For the comparison of models with different mean structures, information criteria should be based on ML rather than on REML, as otherwise the likelihood values would be based on different sets of error contrasts, and therefore would be no longer comparable. Information criteria can be obtained in SAS by adding the option 'ic' to the PROC MIXED statement, viz

```
proc mixed data=test method=ml ic;
```

It should be noted that different 'ic' may select or lead to different non nested models. Thus care should be taken and other tests such as Wald tests may be used to confirm the results.

3.10 Inference for random effects

Brief discussion on Empirical Bayes (EB) inference and how to carry out best linear unbiased prediction will be outlined.

3.10.1 Empirical Bayes Inference

The purpose of random effects $\boldsymbol{b_i}$ in the model is to reflect how the evolution for the ith subject deviates from the expected evolution $X_i\boldsymbol{\beta}$. The estimation of $\boldsymbol{b_i}$ is helpful for the detection of outlying profiles. This strategy is however, only meaningful under the hierarchical model interpretation. Recall that the hierarchical specification of the model is given as

$$\boldsymbol{Y_i}|\boldsymbol{b_i} \sim N(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i}, \Sigma_i),$$

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

Since the $\boldsymbol{b_i}$ are random, it is natural to use Bayesian methods. Under this setting or approach the prior distribution for $\boldsymbol{b_i}$ will be taken as $N(\boldsymbol{0}, G)$. Its posterior density $f(\boldsymbol{b_i}|\boldsymbol{y_i})$ is then given by

$$\begin{aligned} f(\boldsymbol{b_i}|\boldsymbol{y_i}) &\equiv f(\boldsymbol{b_i}|\boldsymbol{Y_i} = \boldsymbol{y_i}) \\ &= \frac{f(\boldsymbol{y_i}|\boldsymbol{b_i})f(\boldsymbol{b_i})}{\int f(\boldsymbol{y_i}|\boldsymbol{b_i})f(\boldsymbol{b_i})d\boldsymbol{b_i}} \\ &\propto f(\boldsymbol{y_i}|\boldsymbol{b_i})f(\boldsymbol{b_i}) \\ &\propto \ldots \\ &\propto exp\{-\frac{1}{2}(\boldsymbol{b_i} - GZ_i^{'}W_i(\boldsymbol{y_i} - X_i\boldsymbol{\beta}))^{'}\Lambda^{-1}(\boldsymbol{b_i} - GZ_i^{'}W_i(\boldsymbol{y_i} - X_i\boldsymbol{\beta}))\} \end{aligned}$$

for some some positive definite matrix Λ_i. It follows that the posterior distribution of b_i is given by

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

Thus a logical estimate of b_i can be obtained from its posterior mean given by

$$\begin{aligned}\hat{\boldsymbol{b_i}}(\boldsymbol{\theta}) &= E[\boldsymbol{b_i}|\boldsymbol{Y_i} = \boldsymbol{y_i}] \\ &= \int \boldsymbol{b_i} f(\boldsymbol{b_i}|\boldsymbol{y_i}) d\boldsymbol{b_i} \\ &= GZ_i' W_i(\boldsymbol{\alpha})(\boldsymbol{y_i} - X_i\boldsymbol{\beta}) \qquad (3.17)\end{aligned}$$

assume to depend on a parameter $\boldsymbol{\theta}$. It is clear from the above that $\hat{\boldsymbol{b_i}}(\boldsymbol{\theta})$ is normally distributed with covariance

$$\text{Var}(\hat{\boldsymbol{b_i}}(\boldsymbol{\theta})) = GZ_i'\{W_i - W_iX_i(\sum_{i=1}^{N} X_i'W_iX_i)^{-1}X_i'W_i\}Z_iG$$

It follows that the inference about $\boldsymbol{b_i}$ should account for the variability in $\boldsymbol{b_i}$. Because of the this reason, inference for $\boldsymbol{b_i}$ should be based on

$$\text{var}(\hat{\boldsymbol{b_i}}(\boldsymbol{\theta}) - \boldsymbol{b_i}) = G - \text{var}(\hat{\boldsymbol{b_i}}(\boldsymbol{\theta})).$$

It follows that just as for the fixed effects inference discussed in Section 3.7, Wald tests can be derived to test hypotheses about $\boldsymbol{b_i}$. Parameters in $\boldsymbol{\theta}$ are replaced by their ML or REML estimates, obtained from fitting the marginal model. The estimate $\hat{\boldsymbol{b_i}} = \hat{\boldsymbol{b_i}}(\boldsymbol{\theta})$ is called the 'Empirical Bayes' estimate of $\boldsymbol{b_i}$. Approximate t-test and F-tests to account for the variability introduced by replacing $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$ similar to testing for fixed effects can be derived.

3.10.2 Best Linear Unbiased Prediction

Often parameters of interest are linear combinations of fixed effects in $\boldsymbol{\beta}$ and random effects in $\boldsymbol{b_i}$. 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. Thus in general, suppose

$$u = \lambda_\beta'\boldsymbol{\beta} + \lambda_b'\boldsymbol{b_i}$$

is of interest. Conditionally on $\boldsymbol{\alpha}$,

$$\hat{u} = \lambda_{\beta}^{'}\hat{\boldsymbol{\beta}} + \lambda_{b}^{'}\hat{\boldsymbol{b_i}}$$

is a best linear unbiased predictor (BLUP) of u. In fact from the theory of linear models $\hat{u}$ is linear in the observations $\boldsymbol{Y_i}$, unbiased for u and it has minimum variance among all unbiased linear estimators and abbreviated as (UMVUE).

3.10.3 Shrinkage estimators

Consider the the prediction of the evolution of the ith subject. That is

$$\begin{aligned}\hat{\boldsymbol{Y_i}} &\equiv X_i\hat{\boldsymbol{\beta}} + Z_i\hat{\boldsymbol{b_i}} \\ &= X_i\hat{\boldsymbol{\beta}} + Z_iGZ_i^{'}V_i^{-1}(\boldsymbol{y_i} - X_i\hat{\boldsymbol{\beta}})\end{aligned}$$

because

$$\hat{\boldsymbol{b_i}} = GZ_i^{'}V_i^{-1}(\boldsymbol{y_i} - X_i\boldsymbol{\beta}).$$

Now since

$$V_i = Z_iGZ_i^{'} + \Sigma_i$$

it follows that

$$V_i - \Sigma_i = Z_iGZ_i^{'}$$

so that if we make this substitution, we have

$$\begin{aligned}\hat{\boldsymbol{Y_i}} &= X_i\hat{\boldsymbol{\beta}} + (V_i - \Sigma_i)V_i^{-1}(\boldsymbol{y_i} - X_i\hat{\boldsymbol{\beta}}) \\ &= X_i\hat{\boldsymbol{\beta}} - (V_i - \Sigma_i)V_i^{-1}X_i\hat{\boldsymbol{\beta}} + (V_i - \Sigma_i)V_i^{-1}\boldsymbol{y_i} \\ &= X_i\hat{\boldsymbol{\beta}} - X_i\hat{\boldsymbol{\beta}} + \Sigma_iV_i^{-1}X_i\hat{\boldsymbol{\beta}} + (I_{n_i} - \Sigma_iV_i^{-1})\boldsymbol{y_i} \\ &= \Sigma_iV_i^{-1}X_i\hat{\boldsymbol{\beta}} + (I_{n_i} - \Sigma_iV_i^{-1})\boldsymbol{y_i} \qquad (3.18)\end{aligned}$$

Hence, $\hat{\boldsymbol{Y}}_{\boldsymbol{i}}$ is a weighted mean of the population averaged profile $X_i\hat{\boldsymbol{\beta}}$ and the observed data $\boldsymbol{y_i}$, with weights $\hat{\Sigma}_i\hat{V}_i^{-1}$ and $I_{n_i} - \hat{\Sigma}_i\hat{V}_i^{-1}$ respectively. Note that $X_i\hat{\boldsymbol{\beta}}$ gets much higher weight if the residual variability is large in comparison to the total variability contained in V_i. This phenomenon is called 'shrinkage'. The observed data are shrunk towards prior average $X_i\boldsymbol{\beta}$. This is also reflected in the fact that for any linear combination $\boldsymbol{\lambda}'\boldsymbol{b_i}$ of random effects

$$\mathrm{Var}(\boldsymbol{\lambda}'\hat{\boldsymbol{b}}_{\boldsymbol{i}}) \leq \mathrm{Var}(\boldsymbol{\lambda}'\boldsymbol{b_i})$$

3.10.4 The random-intercepts model revisited

Consider the random intercepts model with

$$Z_i = \mathbf{1}_{\boldsymbol{n_i}}$$

a vector of ones and

$$D = \sigma_b^2 I_{n_i},$$

a diagonal $n_i \times n_i$ matrix with only one variance component σ_b^2. Also assume absence of serial correlation such that

$$\Sigma_i = \sigma^2 I_{n_i}$$

so that from Eq. (3.17) Empirical Bayes estimate for the random estimate b_i, equals

$$\begin{aligned}
\hat{b}_i &= \sigma^2 \mathbf{1}'_{\boldsymbol{n_i}}(\sigma_b^2 \mathbf{1}_{\boldsymbol{n_i}}\mathbf{1}'_{\boldsymbol{n_i}} + \sigma^2 I_{n_i})^{-1}(\boldsymbol{y_i} - X_i\boldsymbol{\beta}) \\
&= \frac{\sigma_b^2}{\sigma^2}\mathbf{1}'_{\boldsymbol{n_i}}\left(I_{n_i} - \frac{\sigma_b^2}{\sigma^2 + n_i\sigma_b^2}\mathbf{1}_{\boldsymbol{n_i}}\mathbf{1}'_{\boldsymbol{n_i}}\right)(\boldsymbol{y_i} - X_i\boldsymbol{\beta}) \\
&= \frac{n_i\sigma_b^2}{\sigma^2 + n_i\sigma_b^2}\frac{1}{n_i}\sum_{j=1}^{n_i}(y_{ij} - X_i^{[j]}\boldsymbol{\beta})
\end{aligned}$$

It is important to take note that $\hat{b}_i$ is a weighted average of 0 (prior mean) and the average residual for subject i. The less shrinkage the larger n_i and the smaller σ^2 relative to σ_b^2. The equation above shows that the larger n_i is the smaller σ^2 is relative to σ_b^2 and the less the shrinkage and vice versa.

3.10.5 The normality assumption for Random Effects

In practice, histograms of Empirical Bayes (EB) estimates are often used to check the normality assumption for the random effects. However, since

$$\begin{aligned} \hat{\boldsymbol{b_i}} &= GZ_i'W_i(\boldsymbol{y_i} - X_i\boldsymbol{\beta}) \\ \text{Var}(\hat{\boldsymbol{b_i}}) &= GZ_i\left\{W_i - W_iX_i(\sum_{i=1}^{N} X_i'W_iX_i)^{-1}X_i'W_i\right\}Z_iG \end{aligned}$$

One should at least first standardize the EB estimates. Further due to the shrinkage property, the EB estimates do not fully reflect the heterogeneity in the data. Therefore EB estimates obtained under normality cannot be used to check normality. This suggests that the only possibility to check the normality assumption is to fit a more general model, with a classical linear mixed model as a special case and to compare both models using Likelihood ratio methods.

3.10.6 The heterogeneity model

One possible extension of the linear mixed model is to assume a finite mixture as random effects distribution namely :

$$\boldsymbol{b_i} \sim \sum_{j=1}^{g} p_j N(\boldsymbol{\mu_j}, G)$$

with $\sum_{j=1}^{g} p_j = 1$ and $\sum_{i=1}^{g} p_j\boldsymbol{\mu_j} = \mathbf{0}$.

The interpretation of the above assumption is as follows: The population

consists of g sub-populations. Each sub-population contains a fraction p_j of the total population and in each sub-population, a linear mixed model holds. A very flexible class of parametric models holds for the random effects distribution whilst the classical model is the case where $g = 1$. The fitting of the above model is based on an EM algorithm for which a SAS macro is available and the EB estimates can be calculated under the heterogeneity model.

3.10.7 Power analyses under the linear mixed model

In any statistical test no matter how simple or complex the test is, the statistician is always interested in the power of the test. In this section the F-test for fixed effects is considered. Thus consider the general linear hypothesis:

$$H_0 : L\boldsymbol{\beta} = \mathbf{0}$$

versus

$$H_A : L\boldsymbol{\beta} \neq \mathbf{0}$$

Recall that the F test statistic is given by:

$$F_T = \frac{\hat{\boldsymbol{\beta}}' L' \left[L(\sum_{i=1}^{N} X_i' V_i^{-1}(\hat{\boldsymbol{\alpha}}) X_i) L' \right] L\hat{\boldsymbol{\beta}}}{rank(L)}$$

The approximate null distribution of F_T is F with the numerator degrees of freedom equal to the $rank(L)$. The denominator degrees of freedom need to be estimated from the data. This can be done so using three possible methods namely, the:

1. Containment method

2. Sattherwaite approximation

3. Kenward and Roger approximation

In general, not necessarily under H_0, F_T is approximately F distributed with the same number of the degrees of freedom but with a non-centrality parameter:

$$\phi = \boldsymbol{\beta}' L' \left[L(\sum_{i=1}^{N} X_i' V_i^{-1}(\hat{\boldsymbol{\alpha}}) X_i)^{-1} L' \right] L\boldsymbol{\beta}$$

which equals $\mathbf{0}$ under H_0. This can be used to calculate powers under a variety of models and under a variety of alternative hypotheses. Note that ϕ is equal to $rank(L) \times F_T$ and with $\boldsymbol{\beta}$ replaced by $\hat{\boldsymbol{\beta}}$. The SAS procedure 'MIXED' can therefore be used for the calculation of ϕ and the related numbers of degrees of freedom.

Calculation in SAS

The following is an outline of the steps involved in the calculation of the power of the test.

1. Construct a data set of the same dimension and with the same covariates and factor values as the design for which the power is to be calculated.

2. Use as responses $\boldsymbol{y_i}$ the average values $X_i\beta$ under the alternative model.

3. The fixed effects estimate will then be equal to

$$\begin{aligned} \hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) &= \left(\sum_{i=1}^{N} X_i' W_i(\boldsymbol{\alpha}) X_i \right)^{-1} \sum_{i=1}^{N} X_i' W_i(\boldsymbol{\alpha}) \boldsymbol{y_i} \\ &= \left(\sum_{i=1}^{N} X_i' W_i(\boldsymbol{\alpha}) X_i \right)^{-1} \sum_{i=1}^{N} X_i' W_i(\boldsymbol{\alpha}) X_i \boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

4. Hence the F statistic reported by SAS will be equal to $\frac{\phi}{rank(L)}$

5. This calculated F value and the associated numbers of degrees of freedom can be saved and used afterwards for the calculation of the power

6. Note that this requires keeping the variance components in α fixed, equal to the assumed population values

7. The steps in the calculations are as follows:

   - Use PROC MIXED to calculate ϕ and the degrees of freedom ν_1 and ν_2
   - Calculate the critical value F_c:

   $$P(F_{\nu_1,\nu_2,0} > F_c) = level \;\; of \;\; significance$$

   - Calculate the power

   $$power = P(F_{\nu_1,\nu_2,\phi} > F_c)$$

8. The SAS functions 'finv' and 'probf' are used to calculate F_c and the power

Using the above procedure it is clear that the within subject correlation will increase the power for inferences on within subject effects but decrease the power for inferences on between subject effects.

3.11 Conclusion

There are several advantages to the application of the mixed model to medical and survey data. Duchateau et al. (1998, p.18) highlight the specific features of the mixed model as advantages which include:

- Complex data structures can be described in a natural way by the mixed model;
- The analysis of unbalanced data is a natural extension of the analysis of balanced data in the mixed model framework;
- The mixed model framework allows a flexible choice of the appropriate inference space;
- A mixed model also allows the prediction of random effects of interest by best linear unbiased prediction (BLUP).

As a conclusion to this chapter, much of the theory that is covered here will help us to understand the extension of the linear mixed model for longitudinal continuous data to longitudinal discrete data, as is the case with the Kilifi data set. The normality assumption covered in this chapter is a special case of the generalized linear modelling approach for longitudinal data.(McCullagh and Nelder, 1989; Lee, Nelder and Pawitan, 2006; Verbeke and Molenberghs, 2005 and Molenberghs and Verbeke, 2006). The theoretical results covered in this chapter are helpful in subsequent chapters where the focus will solely be on the analysis of binary longitudinal data for an infectious disease process. More importantly departures from the current classical linear mixed model will be of paramount importance in the current work. Modelling the disease process as a non-Gaussian process is a novel idea in the research work. In particular given the disease process is a reversible type of process it is believed the subsequent analysis will add knowledge on the analysis of infectious disease processes in general.

Chapter 4

The Generalized Linear Model

4.1 Introduction

This chapter will focus on models that are suitable to fit repeated measures data but with discrete responses or outcomes. The generalized linear models, according to McCullagh and Nelder (1989) are one such family of models and are generally suitable for discrete repeated measurements in the context of correlated data. Diggle et al. (2002) state that extensions of the generalized linear models in the context of correlated observations include the following classes of models:

- Marginal models
- Random effects models
- Conditional models

Diggle et al. (2002) and Aerts et al. (2002) distinguish between these three families. In order to be able to study each type of model thoroughly, the exponential family of distributions will briefly be described.

4.1.1 The Exponential Family

A random variable Y is said to have a distribution from the exponential family if its probability density function can generally be written in the form:

$$f(y|\theta,\psi) = \exp\left[\frac{y\theta - \psi(\theta)}{\phi} + c(y,\phi)\right] \tag{4.1}$$

for a specific set of unknown parameters θ and ϕ, and for known functions $\psi(.)$ and $c(.,.)$. The parameter θ is called the natural or canonical parameter and ϕ is called the scale or dispersion parameter. The mean and variance of Y can be derived by making use of the property $\int f(y|\theta,\phi)dy = 1$ and taking the first and second order derivatives with respect to θ from both sides of the equation so that we have:

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

and

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

Thus we have that $E[Y] = \mu = \psi^{'}(\theta)$ and $\text{Var}(Y) = \psi^{''}(\theta)\phi$ where $\psi^{'}$ and $\psi^{''}$ denote the first and second derivatives of $\psi(\theta)$ with respect to θ. The mean and the variance are thus related through the relation

$$\sigma^2 = \phi\psi^{''}[\psi^{'-1}(\mu)] = \phi\upsilon(\mu)$$

where $\upsilon(\mu) = \psi^{''}[\psi^{'-1}(\mu)]$ is known as the variance function.

4.1.2 Some illustrations

Binomial random variable

For a binomial variable Y denoting the number of successes in n independent trials with a probability of success p in each trial, the probability distribution is:

$f(y) = \begin{pmatrix} n \\ y \end{pmatrix} p^y(1-p)^{n-y} = \exp\left[y \log \frac{p}{1-p} + n \log(1-p) + \log \begin{pmatrix} n \\ y \end{pmatrix} \right].$

From Eq (4.1) it follows that the corresponding canonical (natural) parameter is $\theta = \log(\frac{p}{1-p})$ also known as the logit(p). Alternatively $\theta = \log\left(\frac{\mu}{n-\mu}\right)$ where $\mu = np$.

Note that in terms of θ

$$p = \frac{\exp(\theta)}{1+\exp(\theta)}$$

and

$$1-p = \frac{1}{1+\exp(\theta)}$$

In terms of the structure of an exponential family probability density function (p.d.f) $\psi(\theta) = -n\log(1-p) = n\log(1+\exp(\theta))$, $\phi = 1$. and
$c(y,\phi) = \log \begin{pmatrix} n \\ y \end{pmatrix}$

Furthermore,
$E[Y] = \psi'(\theta) = n\frac{\exp(\theta)}{1+\exp(\theta)} = np$
and
$\text{Var}(Y) = \psi''(\theta)\phi = \frac{n\exp(\theta)(1+\exp(\theta)) - n\exp(\theta)\exp(\theta)}{(1+\exp(\theta))^2} = np(1-p)$. Thus in this case $v(\mu) = \mu(1-\frac{\mu}{n})$ since $\mu = np$

The Bernoulli (random variable) model for the binary response is a special

case of the Binomial random variable with $n = 1$ and therefore both share the same canonical or natural parameter (McCullagh and Nelder, 1989; Molenberghs and Verbeke , 2005).

Logistic and Probit link function

Let Y be Bernoulli distributed with the probability of success $P(Y = 1) = \pi$. Since the Bernoulli density is part of the exponential family its p.d.f can be written as

$$f(y) = \exp\{y \ln\left(\frac{\pi}{1-\pi}\right) + \ln(1-\pi)\}$$

where the natural parameter θ is equal to $\ln[\frac{\pi}{1-\pi}]$ or the logit of π, the scale parameter $\phi = 1$, with mean $\mu = \pi$ and variance function $\upsilon(\pi) = \pi(1-\pi)$. Thus the distribution is just a special case of a Binomial distribution with $n = 1$. The function $\ln[\frac{\pi}{1-\pi}]$ is called the link function in the context of generalized linear models. If the function $\Phi^{-1}(\pi)$ is used where Φ is the standard normal distribution function, then we have the probit link function

Poisson model for counts

Let Y be Poisson distributed with mean μ. The density is part of the exponential family and can be written as

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

Thus the natural parameter is $\theta = \ln(\mu)$, the scale parameter is ϕ=1, and the variance function is $\upsilon(\mu) = \mu$. Thus a Poisson response variable is naturally modelled using a log link fucntion

4.2 The Generalized Linear Model

The Generalized Linear Model can be formalized as follows:

1. Assume we have independent response variables $Y_1, Y_2, ..., Y_N$ which are assumed to share the same density $f(y|\theta, \psi)$ from an exponential family with $E[Y_i] = \mu_i$ but different natural parameters θ_i are allowed for all observations

2. We let $\boldsymbol{x_i}= [x_1, ..., x_p]$, be a p-dimensional vector of covariate values, $i = 1, \ldots, N$

3. In generalized linear models, it is believed that the differences between the θ_i can be explained through a linear function of known covariates:

$$\theta_i = \boldsymbol{x_i'\beta}$$

4. The means μ_i need to be modelled in terms of the covariate values and it is assumed that $\eta(\mu_i) = x_i'\beta$ for a known link function $\eta(.)$ and a p-dimensional vector β a vector of fixed unknown regression coefficients. Stacking all the N row vectors x_i into one matrix X gives the well known design matrix for the data of dimension $N \times (p + 1)$

4.3 Extending the examples to Generalized linear models

Logistic and Probit regression for Binary data

As already mentioned, the natural link is the logit link so that if $Y_i \sim Bernoulli(\pi_i)$ then the linear model is

$$\ln\left[\frac{\pi_i}{1-\pi_i}\right] = x_i'\beta$$

where in terms of covariates

$$\pi_i = \frac{\exp(x_i'\beta)}{[1 + \exp(x_i'\beta)]}.$$

Note that the natural parameter is a function of the covariate x_i. Alternatively for the probit link, one uses the model $\Phi^{-1}(\pi_i) = x_i^{'}\beta$ so that $\pi_i = \Phi(x_i\beta)$, where Φ denotes the distribution function of a standard normal random variable. For a Binomial variable the $Y_i \sim B(n_i, p_i)$ and the regression model is of the form $\text{logit}(p_i) = x_i^{'}\beta$.

Poisson Regression for counts

The logarithm is the natural link function, leading to the classical Poisson regression model $Y_i \sim$Poisson(μ_i) with ln $(\mu_i) = x_i^{'}\beta$. where μ_i is the mean occurrence rate. This also implies $\mu_i = \exp(x_i^{'}\beta)$ is a quantity which is always non-negative.

4.4 Maximum Likelihood Estimation and Inference

The following derivation follows that in Molenberghs and Verbeke (2005, p. 30). Estimation of the regression parameters in $\boldsymbol{\beta}$ is usually done using maximum likelihood estimation (ML). We assume independence of observations and therefore the log-likelihood is given by

$$\ell(\beta, \phi) = \frac{1}{\phi}\sum_{i=1}^{N}[y_i\theta_i - \psi(\theta_i)] + \sum_i c(y_i, \phi). \tag{4.2}$$

The score equations $S(\boldsymbol{\beta})$ are obtained by calculating the first order derivatives with respect to $\boldsymbol{\beta}$ of the log-likelihood and equating them to give

$$S(\boldsymbol{\beta}) = \sum_i \frac{\partial\theta_i}{\partial\boldsymbol{\beta}}[y_i - \psi^{'}(\theta_i)] = 0. \tag{4.3}$$

Since $\mu_i = \psi^{'}(\theta_i)$ and $\upsilon_i = \upsilon(\mu_i) = \psi^{''}(\theta_i)$ and under $\phi = 1$, it implies that

$$\frac{\partial\mu_i}{\partial\boldsymbol{\beta}} = \psi^{''}(\theta_i)\frac{\partial\theta_i}{\partial\boldsymbol{\beta}} = \upsilon_i\frac{\partial\theta_i}{\partial\boldsymbol{\beta}}$$

therefore the score equations $S(\boldsymbol{\beta})$ becomes

$$S(\boldsymbol{\beta}) = \sum_i \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \upsilon_i^{-1}(y_i - \mu_i) = 0. \tag{4.4}$$

These score equations can then be solved iteratively using algorithms such as re-weighted least squares, Newton Raphson, or Fisher scoring. Once the ML estimates have been obtained, classical inference based on three asymptotically equivalent methods based on asymptotic likelihood theory, namely the Wald-type tests, likelihood ratio tests and score tests can be used. For example, in the linear normal model, estimation of ϕ may be required to estimate the standard errors of the elements in $\boldsymbol{\beta}$. Since $\text{Var}(Y_i) = \phi\upsilon_i$, an obvious estimate for ϕ is given by

$$\hat{\phi} = \frac{1}{N-p}\sum_i (y_i - \hat{\mu}_i)^2/\upsilon_i(\hat{\mu}_i).$$

Under the normal model, this would yield

$$\hat{\sigma}^2 = \tfrac{1}{N-p}\textstyle\sum_i (y_i - \boldsymbol{x_i}\hat{\boldsymbol{\beta}})^2,$$

which is the mean squared error used in linear regression models to estimate the residual variance. More details on estimation and inference in the generalized linear models can be found in McCullagh and Nelder (1989)and more recent references such as Molenberghs and Verbeke (2005) and Lee, Nelder and Pawitan (2006). McCullagh and Nelder (1989) state that the Generalized linear models is a unifying theory to a wide range of settings:

- For normal outcomes, we could use linear models, multiple regression and Analysis of variance (ANOVA)
- Binary outcomes would involve the use of probit and logit (logistic) regression

- Categorical data makes use of the log-linear modelling
- Outcomes that are counts can be modelled using Poisson regression
- Non-negative continuous time to event data can be modelled using survival analysis

For continuous outcomes, the classes of models that can be used to model the data include:

- Marginal models

$$E(Y_{ij}|\boldsymbol{x_{ij}}) = \boldsymbol{x_{ij}'}\boldsymbol{\beta}$$

  where $\boldsymbol{x_{ij}'}$ denotes the vector of covariates for individual i measured at occassion j and $\boldsymbol{\beta}$ is a vector of fixed parameters.

- Random-Effects models

$$E(Y_{ij}|\boldsymbol{b_i}, \boldsymbol{x_{ij}}) = \boldsymbol{x_{ij}'}\boldsymbol{\beta} + \boldsymbol{z_{ij}'}\boldsymbol{b_i}$$

  where $\boldsymbol{x_{ij}'}$ and $\boldsymbol{\beta}$ carry the same meaning as explained above. The vector $\boldsymbol{b_i}$ denotes a vector of subject specific random effects and $\boldsymbol{z_{ij}'}$ is the corresponding vector of covariates.

- Transition models

$$E(Y_{ij}|Y_{i,j1}, \ldots, Y_{i1}, \boldsymbol{x_{ij}}) = \boldsymbol{x_{ij}'}\boldsymbol{\beta} + \alpha Y_{i,j-1}$$

  where now in addition to the dependence on $\boldsymbol{x_{ij}'}$ the current response depends on the previous response.

These models can then be extended to include the case of repeated discrete outcomes. We will now consider these extensions below. These extensions and their applications will be discussed in the current and subsequent chapters.

4.5 Longitudinal Generalized Linear Models

If we have normal outcomes, then the switch between the cross sectional data to longitudinal data is straightforward. The case of repeated or longitudinal non-normal data is not that straightforward. The lack of key distributions such as the normal distribution results to several modelling options and the introduction of non-linearity. This implies no easy transfer between model families. The following summary best explains the departure from the case of normal to non-normal data:

	Cross-sectional	**Longitudinal**
Normal outcome	Linear model	Linear mixed model(LMM)
Non-normal outcome	Generalized linear model (GLM)	Several options

4.5.1 Marginal Models

Let the observations for individual i be Y_{ij}, for $j = 1, 2, \ldots, n$ and $i = 1, 2, \ldots, N$ and let Y_i denote the n dimensional vector of observations for individual i assuming each individual contributes n outcomes. But more generally an individual i contributes n_i outcome. Marginal models are sometimes also called the population averaged models. The probability of each outcome (or set of outcomes) is directly modelled (integrating or summing the other outcomes away) so that we have to correctly specify $E(Y_{ij}|x_{ij})$. Minimally we have to specify

$$\boldsymbol{\eta}_i(\boldsymbol{\mu}_i) = \{\eta_{i1}(\mu_{i1}), \ldots, \eta_{in}(\mu_{in})\}$$

$$E(\boldsymbol{Y}_i) = \boldsymbol{\mu}_i$$

and

$$\boldsymbol{\eta}_i(\boldsymbol{\mu}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$$

$$\text{var}(\boldsymbol{Y}_i) = \phi \boldsymbol{v}(\boldsymbol{\mu}_i)$$

where $\boldsymbol{v}(.)$ is a known variance function and

$$\text{corr}(\boldsymbol{Y}_i) = R(\boldsymbol{\alpha}),$$

that is the correlation matrix $R(\boldsymbol{\alpha})$ depends on a set of parameters $\boldsymbol{\alpha}$. Fitting marginal models is quite involving because the marginal association parameters are highly constrained. It is also important to note that marginal models are reproducible or upward compatible. For example, Fitzmaurice and Laird (1993) make use of a mixed marginal conditional model, Carey, Zeger and Diggle (1993) use alternating logistic regressions and Molenberghs and Ritter (1996) and Molenberghs and Danielson (1999) suggest the use of 2^{nd} order mixed parameterization and Generalized Estimating Equations type 2 (GEE 2). There are various methods applicable in fitting marginal models for both, non-likelihood and likelihood ones.

In the class of non-likelihood methods Koch et al. (1975) introduced the Empirical generalized least squares (EGLS) method. This method can be fitted in SAS using the " Proc CATMOD" procedure. Generalized Estimating Equations (GEE) have been applied by among others Liang and Zeger (1986), Lipsitz, Laird and Harrington (1991), Liang, Zeger and Qaquish (1992), Zhao and Prentice (1990) and Robins, Rotnitzky and Zhao (1995). The GEE method can be implemented using "Proc GENMOD" procedure in SAS.

In the class of likelihood methods Ashford and Sowden (1970) proposed the use of the Multivariate probit model. Bahadur (1961) used the Bahadur model as a marginal model. Dale (1986) used the odds ratio model for bivariate data based on the Plackett distribution (Plackett, 1965). The odds ratio models for multivariate data have been used in different settings as

marginal models. For example, Lang and Agresti (1994) use the constraint equations approach, Molenberghs and Lesaffre (1994, 1999) extend the Dale model to multivariate ordinal outcomes and Glonek and McCullagh (1995) use the multivariate logit type models. In the next section the GEE non likelihood model is discussed and applied to the RSV data set.

4.6 Generalized Estimating Equations (GEE)

4.6.1 Introduction

The key paper that introduced the Generalized Estimating Equations (GEE) was that by Liang and Zeger (1986). Thereafter reviews have been published by Desmond (1997), Pendergast et al. (1996) and Hall (2001). Recall that the score equations for GLM's were derived in Eq (4.3) as

$$S(\boldsymbol{\beta}) = \sum_i \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \upsilon_i^{-1}(y_i - \mu_i) = 0.$$

In case the outcome Y_i is multivariate, that is, $\boldsymbol{Y_i} = (Y_{i1}, \ldots, Y_{in_i})'$ with independent components Y_{ij}, this would become

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_i \sum_j \frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}} \upsilon_{ij}^{-1}(y_i - \mu_i) \\ &= \sum_i \frac{\partial \boldsymbol{\mu_i'}}{\partial \boldsymbol{\beta}} V_i^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu_i}) \\ &= \sum_i F_i' V_i^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu_i}) \\ &= 0 \end{aligned}$$

where $F_i = \frac{\partial \boldsymbol{\mu_i}'}{\partial \boldsymbol{\beta}}$ and $\boldsymbol{\mu_i} = E(\boldsymbol{Y_i})$ and

$$V_i = \begin{pmatrix} v_{i1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & v_{in_i} \end{pmatrix}.$$

In case of the normal model with $\boldsymbol{\mu_i} = X_i\boldsymbol{\beta}$, this equation becomes

$$S(\boldsymbol{\beta}) = \sum_i X_i' V_i^{-1}(\boldsymbol{y_i} - X_i\boldsymbol{\beta}) = 0. \tag{4.5}$$

It is important to note that when fitting linear mixed models, the same score equation as Eq (4.5) had to be solved. However, V_i was not diagonal but was equal to the modelled covariance matrix of $\mathbf{Y_i}$. GEE's can be obtained by using a non-diagonal V_i in the score equations for GLM's:

$$\sum_i \frac{\partial \boldsymbol{\mu_i}'}{\partial \boldsymbol{\beta}} V_i^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu_i}) = 0$$

where V_i is now a n_i x n_i covariance matrix with diagonal elements given by v_{ij}. In practice then V_i will be of the form

$$V_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \phi A_i^{1/2}(\boldsymbol{\beta}) R_i(\boldsymbol{\alpha}) A_i^{1/2}(\boldsymbol{\beta}) \tag{4.6}$$

in which

$$A_i^{1/2}(\beta) = \begin{pmatrix} \sqrt{v}_{i1}(\mu_{i1}(\boldsymbol{\beta})) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sqrt{v}_{i1}(\mu_{in_i}(\boldsymbol{\beta})) \end{pmatrix}$$

$R_i(\boldsymbol{\alpha})$ is the correlation matrix of $\boldsymbol{Y_i}$ which depends on a vector α of unknown parameters. It is important to note that unlike in the normal case, solving $S(\boldsymbol{\beta}) = 0$ will not yield MLE's. The equations are strictly speaking, not score equations since they are not first-order derivatives of some log-likelihood function for the data under some statistical model. We refer to the above approach as the standard modelling GEE approach.

4.6.2 Large Sample Properties

Let $\hat{\boldsymbol{\beta}}$ be the solution to Eq (4.5) that is $\hat{\boldsymbol{\beta}}$ is the solution to

$$\sum_i \frac{\partial \boldsymbol{\mu_i'}}{\partial \boldsymbol{\beta}} V_i^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu_i}) = 0.$$

Then large sample properties guarantee that conditionally upon $\boldsymbol{\alpha}$, $\hat{\boldsymbol{\beta}}$ is asymptotically ($N \to \infty$) normally distributed with mean $\boldsymbol{\beta}$ and covariance matrix:

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \left(\sum_i \frac{\partial \boldsymbol{\mu_i'}}{\partial \boldsymbol{\beta}} V_i \frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}}\right)^{-1} \text{x} \left(\sum_i \frac{\partial \boldsymbol{\mu_i'}}{\partial \boldsymbol{\beta}} V_i^{-1} \text{Var}(\boldsymbol{Y}_i) V_i^{-1} \frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}}\right) \text{x} \left(\sum_i \frac{\partial \boldsymbol{\mu_i'}}{\partial \boldsymbol{\beta}} V_i \frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}}\right)^{-1}. \tag{4.7}$$

Notationally we can write Eq (4.7) as:

$$\text{Var}(\hat{\boldsymbol{\beta}}) = I_0^{-1} I_1 I_0^{-1}.$$

The above estimator of $\text{Var}(\hat{\boldsymbol{\beta}})$ called the sandwich estimator is also sometimes called the robust estimator. This result holds provided that the mean was correctly specified i.e. provided that $E(\boldsymbol{Y_i}) = \boldsymbol{\mu_i}(\boldsymbol{\beta})$. In practice $\boldsymbol{\alpha}$ is replaced by an estimate. The robust (sandwich) estimator in the linear models case derived earlier on is a special case of the above covariance matrix. In case R_i is indeed the correct correlation model, the covariance matrix $\text{Var}(\hat{\boldsymbol{\beta}})$ reduces to

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \phi \left(\sum_i \frac{\partial \boldsymbol{\mu_i'}}{\partial \boldsymbol{\beta}} V_i^{-1} \frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}}\right)^{-1} = I_0^{-1}.$$

provided $\left(\sum_i \frac{\partial \boldsymbol{\mu_i'}}{\partial \boldsymbol{\beta}} V_i^{-1} \frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}}\right)$ is non-singular.
However, I_0^{-1} is the so called naive estimator or model based estimator. The known variance result is recovered when the guess of the correct correlation model is actually equal to the true model. The estimators $\hat{\boldsymbol{\beta}}$ are consistent even if the working correlation matrix is correct.

In practice, Var($\boldsymbol{Y_i}$) in Var($\hat{\boldsymbol{\beta}}$) is replaced by:

$$[\boldsymbol{Y_i} - \boldsymbol{\mu_i}(\hat{\boldsymbol{\beta}})][\boldsymbol{Y_i} - \boldsymbol{\mu_i}(\hat{\boldsymbol{\beta}})]^{'}$$

which is unbiased for Var($\boldsymbol{Y_i}$), provided that the mean has been correctly specified. This means that there are several implications based on the asymptotic theory, viz.

- Mean structure needs to be correctly specified
- Little effort needs to be spent on specifying the correlation structure because mis-specification does not affect consistency and asymptotic normality
- Considerably large samples may be required
- Efficiency can be affected and this follows from the Cramér-Rao inequality
- Taken to the extreme, one could make the working assumptions of independence between two repeated measures
- It also implies that the correlation structure should not be interpreted
- GEE's validity is limited when there are incomplete data

4.6.3 The Working Correlation Matrix

When fitting marginal models it is possible to specify what is known as the working correlation matrix $R_i(\boldsymbol{\alpha})$ for the n observations from subject i We can therefore write

$$V_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \phi A_i^{1/2}(\boldsymbol{\beta}) R_i(\boldsymbol{\alpha}) A_i^{1/2}(\boldsymbol{\beta}).$$

The variance function A_i is the $n_i \times n_i$ diagonal matrix with elements $v(\mu_{ij})$, the known GLM variance function. The working correlation $R_i(\boldsymbol{\alpha})$, is possibly dependent on a different set of parameters $\boldsymbol{\alpha}$. The over dispersion parameter ϕ, is assumed to be 1 or estimated from the data.

The unknown quantities are expressed in terms of the Pearson residuals

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

Note that the e_{ij} implicitly depends on $\boldsymbol{\beta}$ which is unknown and has to be estimated.

4.6.4 Estimation of the Working Correlation Matrix

Liang and Zeger (1986) proposed the moment based estimates for the working correlation matrix. Some of the more popular estimation assumptions include:

Assumption	$\text{Corr}(Y_{ij}, Y_{ik})$	Estimate
Independence	0	
Exchangeable	α	$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{n_i(n_i-1)} \sum_{j \neq k} e_{ij}e_{ik}$
AR(1)	α^t	$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{n_i-1)} \sum_{j \leq n_i-1} e_{ij}e_{i,j+1}$
Unstructured	α_{jk}	$\hat{\alpha}_{jk} = \frac{1}{N}\sum_{i=1}^{N} e_{ij}e_{ik}$

The dispersion parameter is then estimated by

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

Albert and McShane (1995), Fitzmaurice (1995) and Hall and Severini (1998) all point out that accurate modelling of the correlation structure generally improves statistical inference on means. However the moment based estimates of the working correlation structure has led to doubts about efficiency

of correlation modelling and convergence of GEE solution algorithms in certain situations. Crowder (1995) indicated that the solution to the first order GEE for $\boldsymbol{\beta}$ does not exist under certain types of severe misspecification of the working correlation structure. Further work along these lines was done by Sutradhar and Das (1999) who using unbalanced data, showed that these estimates of $\boldsymbol{\beta}$ obtained under a working independence assumption are sometimes more efficient than those with a misspecified nondiagonal working correlation structure. Chaganty (1997), Segal, Neuhaus and James (1997) and O'Hara-Hines (1998) all point out criticisms with the parameter estimation of GEE. Chaganty (1997) proposes the quasi-least squares (QLS) method for estimating the correlation parameters. This method was further extended and investigated by Shults and Chaganty (1998). Recently, Wang and Carey (2004) propose two ways of constructing unbiased estimating equations from general correlation models for irregularly timed repeated measures to supplement and enhance GEE. The equations are obtained by differentiation of the Cholesky decomposition of the working correlation or as the score equations for decoupled Gaussian pseudolikelihood. These equations can be solved with computational effort equivalent to that required for a first order GEE. Wang and Carey (2004) also state that methods are well defined for highly unbalanced and irregularly timed data sets and are applicable for working correlation patterns outside the first order Markovian time series model. They also state that if convergence in the unbiased estimating equations are not achieved then choosing a different correlation structure or switching to a working independence model may be a solution.

4.6.5 Fitting GEE

The standard fitting procedure for GEE's in SAS is 'PROC GENMOD'. The steps involved in the computational procedure are as follows:

Step 1

Compute initial estimates for $\boldsymbol{\beta}$ using a univariate GLM (i.e. assuming independence among the n_i responses for subject i)

Step 2

Compute Pearson residuals e_{ij} using Eq (4.8)

Step 3

Compute estimates for $\boldsymbol{\alpha}$.

Step 4

Compute $R_i(\boldsymbol{\alpha})$ under a given assumption of a correlation structure

Step 5

Compute an estimate for ϕ using Eq (4.9)

Step 6

Compute $V_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \phi A_i^{1/2}(\boldsymbol{\beta}) R_i(\boldsymbol{\alpha}) A_i^{1/2}(\boldsymbol{\beta})$.

Step 7

Update estimate for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left[\sum_{i=1}^{N} F_i' V_i^{-1} F_i\right]^{-1} \left[\sum_{i=1}^{N} F_i' V_i^{-1} (\boldsymbol{y_i} - \boldsymbol{\mu_i})\right]$$

Step 8

Iterate 2-7 until convergence is reached

Estimates of precision can be achieved by comparing I_0^{-1} and $I_0^{-1} I_1 I_0^{-1}$.

4.7 Some developmental notes on GEE over time

We note the following important notes about GEE's

1. Burton et al. (1998) give a detailed comparison of the several parallels between GEE algorithm and the IGLS algorithm used in multilevel modelling. They note that in the case of normal data and the identity link, the two procedures are in fact equivalent.

2. GEE's have been derived using the marginal distribution of $y_{ij}{}'s$. Liang and Zeger (1986) note that it may not be the appropriate technique when interest centres on "growth" or more general change over time in the repeated measures.

3. There is a related method known as GEE2 which is described by Zhao and Prentice (1990) and Kenward (1994). This method attempts to gain efficiency of $\boldsymbol{\beta}$ by substituting a parametric model for $\boldsymbol{\alpha}$ and then relies on assuming that both models for the mean and dependency parameters are correct i.e. there is a tradeoff on robustness for efficiency. This method is further discussed by Liang, Zeger and Qaquish (1992). They found that there is a non-negligible loss in efficiency with GEE1 (the standard GEE described above) compared to GEE2 in the estimation of dependency parameters $\boldsymbol{\alpha}$. However, there was very little loss in efficiency in the estimation of $\boldsymbol{\beta}$ and thus they recommended that GEE1 be used when the $\boldsymbol{\alpha}$ can be regarded as nuisance parameters, and the number of clusters k is large relative to n_i, the size of each cluster. This is the situation usually encountered in survey analysis. Note that a cluster corresponds to an individual in the context of repeated or

longitudinal data.

4. An alternative to GEE is the *alternating logistic regressions (ALR)* proposed by Carey, Zeger and Diggle (1993), but not of interest in the current work.

5. Le Cessie and Van Houwelingen (1994) suggested an approximation to the true likelihood by means of a pseudo-likelihood (PL) function that is easier to evaluate and to maximize. Both GEE and PL give consistent and asymptotically normal estimators provided an empirically corrected variance estimator which we have called the sandwich estimator is used. GEE is well suited only to marginal models while PL can be used for marginal models (Geys, Molenberghs and Lipsitz, 1998) and conditional models (Geys, Molenberghs and Ryan, 1997, 1999).

6. Wang and Lin (2005) investigate the impacts of misspecifing the variance function which is known to be a function of the mean. They state that in the framework of GEE, the correct specification of the variance function can improve the estimation efficiency even if the correlation structure is misspecified. However misspecification of the variance function impacts much more on the estimators for within cluster covariates than for cluster level covariates and also if the variance function is misspecified, the correct choice of the correlation structure may not necessarily improve estimation efficiency.

7. Mainstream statistical software packages such as SAS (PROC GENMOD), STATA(XTGEE command) and GENSTAT has the methodology of the GEE described above in-built .

4.7.1 Application of fitting GEE models to the RSV data set

A series of various models were fitted using the 'Proc Genmod' procedure in SAS by changing the correlation structure within individual responses and then assessing the main effects. The model that was first fitted included all the main effects terms. Only those terms that were found to be significant were retained with the suitable correlation structure. The main effects terms that we consider are: age, dt, prev, actipass and timemonth. These variables were described in detail in Chapter 1. All the interaction terms were assessed by sequentially adding them to the full model of main effects one at a time and then assessing the p-values of the Wald test of the model but none of the interaction terms were found to be significant. Hence they are not reported here. The results are summarized below.

	Exchangeable			Independent			AR(1)		
Parameter	Est.	Std. Error	Pr> |Z|	Est.	Std. Error	Pr> |Z|	Est.	Std. Error	Pr> |Z|
Intercept	-5.0329	1.4454	0.001	-5.0363	1.4479	0.001	-5.0337	1.458	0.001
age 0	-0.9253	1.2175	0.447	-0.9197	1.2194	0.451	-0.9261	1.2343	0.453
age 1	-0.6499	1.0714	0.544	-0.647	1.0736	0.547	-0.6011	1.0824	0.579
age 2	-0.2792	1.0337	0.787	-0.276	1.0356	0.790	-0.241	1.0437	0.817
age 3	-0.0714	0.9744	0.942	-0.0689	0.9759	0.944	-0.0305	0.9835	0.975
age 4	-0.6709	0.9491	0.480	-0.669	0.9502	0.481	-0.6499	0.9579	0.498
age 5	-2.6057	1.298	0.045	-2.6025	1.2979	0.045	-2.5411	1.2919	0.049
age 6	-1.5989	1.0086	0.113	-1.596	1.0087	0.114	-1.566	1.0105	0.121
age 7	-2.2518	1.1538	0.051	-2.25	1.1538	0.051	-2.2603	1.1692	0.053
age 8	-1	0.5944	0.093	-0.9989	0.5946	0.093	-0.96	0.5969	0.108
age 9	-0.7399	0.5124	0.149	-0.7389	0.5126	0.149	-0.7361	0.5189	0.156
age 10	-0.3234	0.4528	0.475	-0.3221	0.4528	0.477	-0.2992	0.4574	0.513
age 11	-0.5684	0.4612	0.218	-0.5685	0.4612	0.218	-0.5365	0.4637	0.247
age 12	0.000	0.000	.	0.000	0.000	.	0.000	0.000	.
dt	0.0008	0.0084	0.919	0.0009	0.0084	0.919	0.0014	0.0082	0.866
prev	44.6065	8.1063	< .0001	44.5942	8.1055	< .0001	43.8948	8.1214	< .0001
timemonth	-0.0457	0.1044	0.662	-0.0454	0.1046	0.664	-0.0437	0.1053	0.678
actipass 0	2.2345	0.1768	< .0001	2.2341	0.1769	< .0001	2.2049	0.1759	< .0001
actipass 1	0.000	0.000	.	0.000	0.000	.	0.000	0.000	.

Table 4.1: Model based standard errors and estimates GEE

	Exchangeable			Independent			AR(1)		
Parameter	Est.	Std. Error	Pr> |Z|	Est.	Std. Error	Pr> |Z|	Est.	Std. Error	Pr> |Z|
Intercept	-5.033	1.165	< .0001	-5.036	1.165	< .0001	-5.034	1.161	< .0001
age 0	-0.925	1.230	0.452	-0.920	1.229	0.454	-0.926	1.226	0.450
age 1	-0.650	0.906	0.473	-0.647	0.906	0.475	-0.601	0.902	0.505
age 2	-0.279	0.858	0.745	-0.276	0.858	0.748	-0.241	0.857	0.779
age 3	-0.071	0.801	0.929	-0.069	0.801	0.932	-0.031	0.800	0.970
age 4	-0.671	0.746	0.368	-0.669	0.746	0.370	-0.650	0.749	0.385
age 5	-2.606	1.194	0.029	-2.603	1.194	0.029	-2.541	1.172	0.030
age 6	-1.599	0.869	0.066	-1.596	0.869	0.066	-1.566	0.860	0.069
age 7	-2.252	1.040	0.030	-2.250	1.039	0.030	-2.260	1.033	0.029
age 8	-1.000	0.606	0.099	-0.999	0.607	0.100	-0.960	0.606	0.113
age 9	-0.740	0.561	0.187	-0.739	0.561	0.188	-0.736	0.554	0.184
age 10	-0.323	0.473	0.494	-0.322	0.473	0.496	-0.299	0.473	0.527
age 11	-0.568	0.443	0.199	-0.569	0.443	0.199	-0.537	0.447	0.231
age 12	0.000	0.000	.	0.000	0.000	.	0.000	0.000	.
dt	0.001	0.011	0.937	0.001	0.011	0.937	0.001	0.010	0.893
prev	44.607	6.554	< .0001	44.594	6.552	< .0001	43.895	6.527	< .0001
timemonth	-0.046	0.085	0.589	-0.045	0.085	0.592	-0.044	0.084	0.603
actipass 0	2.235	0.181	< .0001	2.234	0.181	< .0001	2.205	0.178	< .0001
actipass 1	0.000	0.000	.	0.000	0.000	.	0.000	0.000	.

Table 4.2: Empirical based standard errors and estimates GEE

The algorithm for the unstructured correlation matrix option did not converge and the results are omitted. The results of the model based estimates and standard errors are not very different between the three correlation structures. The magnitude of the estimates are somewhat similar. Moreover, we see that the model based and the empirical parameter estimates are not very

different in magnitude. This is a feature of GEE because the choice between naive and empirical only affects the estimation of the covariance matrix of the regression parameter $\boldsymbol{\beta}$. The output for the correlation between two repeated measurement for the exchangeable correlation matrix was found to be -0.00035. A possible reason why the unstructured correlation matrix did not produce convergence is because the observations can not be aligned that is the observations were not equally spaced. Table 4.1 and 4.2 shows that for the model and empirical based estimates that at the 5% significance level there were significant differences between age group 5 relative to age group 12 and mildly between age group 7 relative to age group 12 in determining whether a child is infected or not. The variables prevalence (prev) and type of sampling (actipass), whether a child was actively or passively sampled (actipass 0 versus actipass 1) were both significant at the 5% level in influencing whether a child is infected or not. The full results are tabulated in Tables 4.1 and 4.2 for the types of standard errors and the three correlation structures. It is also worthwhile noting that the exchangeable and independent correlation structures have their empirical standard errors slightly closer to the model based standard errors than the AR(1) correlation structure. The estimated GEE correlation matrices are all essentially independent, so we expect to see no appreciable differences among the columns of Table 4.1 and 4.2. It is however interesting that the sandwich estimator appears to be picking up dependence not captured by the working correlation matrices given the estimated correlation parameters. It is necessary to reiterate that the unstructured correlation matrix is found to be unsuitable in this scientific setting and is dropped.

Correlation Type	Source	DF	Chi-Square	Pr > Chi-Sq
Exchangeable	age	12	30.39	0.0024
	dt	1	0.01	0.9379
	prev	1	23.32	< .0001
	timemonth	1	0.3	0.5860
	actipass	1	61.86	<.0001
Independent	age	12	30.39	0.0024
	dt	1	0.01	0.9378
	prev	1	23.32	< .0001
	timemonth	1	0.29	0.5882
	actipass	1	61.81	<.0001
AR(1)	age	12	30.54	0.0023
	dt	1	0.02	0.8974
	prev	1	22.94	< .0001
	timemonth	1	0.27	0.6008
	actipass	1	62.00	<.0001

Table 4.3: Score statistics for Type III GEE

The type III score statistics show that the age, prev and actipass variables to be significant at the 5% level in all three correlation structures. The magnitude of the estimates do not differ by vast amounts from each other in the three correlation structures.

4.8 Weighted Generalized Estimating Equations (WGEE)

Missing data is a problem that is frequently encountered in the analysis of clustered and repeated measurement data. Liang and Zeger (1986) pointed out that, GEE inferences are only valid under MCAR. Robins, Rotnitzky

and Zhao (1995) proposed a class of weighted estimating equations to allow for MAR, as an extension to the GEE.

The idea is to weight each subject's contribution in the GEE's by the inverse probability that a subject drops out at the time he dropped out. This can be calculated, for example, as

$$\begin{aligned}\nu_{id_i} \equiv P[D_i = d_i] = & \prod_{k=2}^{d_i-1} (1 - P[R_{ik} = 0 | R_{i2} = \ldots = R_{i,k-1} = 1]) \\ & \times \quad P[R_{id_i} = 0 | R_{i2} = \ldots = R_{i,d_i-1} = 1]^{I\{d_i \leq T\}}\end{aligned} \tag{4.10}$$

where R_{ij} is a dropout indicator taking a value of 1 if the individual did not drop out and a value of 0 if dropout occurred. Recall that $\boldsymbol{Y}_i$ can be partitioned into unobserved components $\boldsymbol{Y}_i^m$ and the observed components $\boldsymbol{Y}_i^o$. Similarly, we can allow the same partitioning of $\boldsymbol{\mu}_i$ into $\boldsymbol{\mu}_i^m$ and $\boldsymbol{\mu}_i^o$. In the weighted GEE approach, which is intended to reduce possible bias of $\hat{\boldsymbol{\beta}}$, the score equations to be solved, taking into account the correlation structure are:

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^{N} \frac{1}{\nu_{id_i}} \frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}'} \left(A_i^{1/2} R_i A_i^{1/2} \right)^{-1} (\boldsymbol{y_i} - \boldsymbol{\mu_i}) = \mathbf{0} \\ &= \sum_{i=1}^{N} \sum_{d=2}^{n+1} \frac{I(D_i = d)}{\nu_{id_i}} \frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}'}(d) \left(A_i^{1/2} R_i A_i^{1/2} \right)^{-1} (d)(\boldsymbol{y_i}(d) - \boldsymbol{\mu_i}(d)) = \mathbf{0} \end{aligned}$$

where $\boldsymbol{y_i}(d)$ and $\boldsymbol{\mu_i}(d)$ are the first $d-1$ elements of $\boldsymbol{y_i}$ and $\boldsymbol{\mu_i}$ respectively. We define $\frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}'}(d)$ and $\left(A_i^{1/2} R_i A_i^{1/2} \right)^{-1} (d)$ analogously. Each child should have had a maximum of 44 visits and this was not the case in the data set due to dropout. The pattern of the missingness was monotone and WGEE was applied to the data by firstly running the dropout macro and then deriving the weights to run WGEE in SAS. The results for fitting WGEE to the RSV data are summarized below:

Correlation Type	Source	DF	Chi-Square	Pr > Chi-Sq
Exchangeable	age	12	29.30	0.0036
	dt	1	0.54	0.4625
	prev	1	14.91	0.0001
	timemonth	1	1.26	0.2608
	actipass	1	58.93	<.0001
Independent	age	12	27.82	0.0059
	dt	1	0.54	0.4606
	prev	1	14.48	0.0001
	timemonth	1	1.30	0.2551
	actipass	1	58.07	<.0001
AR(1)	age	12	27.83	0.0059
	dt	1	0.56	0.4528
	prev	1	14.31	0.0002
	timemonth	1	1.32	0.2508
	actipass	1	58.45	<.0001

Table 4.4: Score statistics for Type III WGEE

	Exchangeable			Independent			AR(1)		
Parameter	Est.	Std. Error	Pr> |Z|	Est.	Std. Error	Pr> |Z|	Est.	Std. Error	Pr> |Z|
Intercept	-8.526	1.733	< .0001	-8.306	1.632	< .0001	-8.311	1.635	< .0001
age 0	1.373	1.459	0.347	1.339	1.382	0.333	1.303	1.391	0.349
age 1	-2.456	1.354	0.070	-2.517	1.279	0.049	-2.493	1.282	0.052
age 2	0.132	1.279	0.918	-0.005	1.209	0.997	0.005	1.212	0.997
age 3	0.169	1.177	0.886	0.035	1.113	0.975	0.046	1.116	0.967
age 4	0.760	1.090	0.486	0.781	1.028	0.448	0.788	1.030	0.445
age 5	-1.671	1.489	0.262	-1.495	1.338	0.264	-1.478	1.338	0.269
age 6	-0.941	1.105	0.394	-0.888	1.034	0.390	-0.877	1.035	0.397
age 7	-1.899	1.294	0.142	-1.743	1.171	0.137	-1.739	1.174	0.139
age 8	-0.754	0.642	0.240	-0.722	0.617	0.242	-0.709	0.618	0.251
age 9	-0.492	0.554	0.375	-0.455	0.533	0.393	-0.455	0.535	0.395
age 10	-0.095	0.487	0.846	-0.072	0.469	0.878	-0.064	0.470	0.891
age 11	-0.494	0.489	0.313	-0.441	0.469	0.347	-0.434	0.470	0.356
age 12	0.000	0.000	.	0.000	0.000	.	0.000	0.000	.
dt	0.008	0.008	0.304	0.008	0.008	0.311	0.008	0.008	0.303
prev	40.413	8.876	< .0001	38.698	8.461	< .0001	38.465	8.473	< .0001
timemonth	0.154	0.126	0.224	0.151	0.119	0.203	0.152	0.119	0.201
actipass 0	3.415	0.260	< .0001	3.325	0.250	< .0001	3.313	0.250	< .0001
actipass 1	0.000	0.000	.	0.000	0.000	.	0.000	0.000	.

Table 4.5: Model based standard errors and estimates for WGEE

	Exchangeable			Independent			AR(1)		
Parameter	Est.	Std. Error	Pr> |Z|	Est.	Std. Error	Pr> |Z|	Est.	Std. Error	Pr> |Z|
Intercept	-8.526	1.966	< .0001	-8.306	1.841	< .0001	-8.311	1.840	< .0001
age 0	1.373	1.799	0.445	1.339	1.700	0.431	1.303	1.712	0.447
age 1	-2.456	1.681	0.144	-2.517	1.617	0.120	-2.493	1.613	0.122
age 2	0.132	1.420	0.926	-0.005	1.350	0.997	0.005	1.350	0.997
age 3	0.169	1.289	0.896	0.035	1.231	0.977	0.046	1.230	0.970
age 4	0.760	1.049	0.469	0.781	0.987	0.429	0.788	0.987	0.425
age 5	-1.671	1.639	0.308	-1.495	1.336	0.263	-1.478	1.333	0.267
age 6	-0.941	1.051	0.370	-0.888	0.944	0.347	-0.877	0.943	0.352
age 7	-1.899	1.324	0.151	-1.743	1.083	0.108	-1.739	1.082	0.108
age 8	-0.754	0.728	0.300	-0.722	0.675	0.285	-0.709	0.675	0.293
age 9	-0.492	0.672	0.464	-0.455	0.625	0.467	-0.455	0.623	0.466
age 10	-0.095	0.565	0.867	-0.072	0.525	0.891	-0.064	0.526	0.903
age 11	-0.494	0.510	0.333	-0.441	0.468	0.346	-0.434	0.470	0.356
age 12	0.000	0.000	.	0.000	0.000	.	0.000	0.000	.
dt	0.008	0.010	0.396	0.008	0.009	0.385	0.008	0.009	0.374
prev	40.413	8.355	< .0001	38.698	7.784	< .0001	38.465	7.782	< .0001
timemonth	0.154	0.135	0.255	0.151	0.128	0.236	0.152	0.127	0.231
actipass 0	3.415	0.628	< .0001	3.325	0.578	< .0001	3.313	0.575	< .0001
actipass 1	0.000	0.000	.	0.000	0.000	.	0.000	0.000	.

Table 4.6: Empirical based standard errors and estimates for WGEE

The model based results show that at the 5% significance level, the parameter estimates for 'prev' and 'actipass' are significant under three correlation structure assumptions while the parameter estimate for 'age 1' versus 'age 12' is significant only under the independent assumption. Empirical based estimates in Table 4.6 show only 'prev' and 'actipass' to be significant at the

5% level. There seem not to be huge differences between the model based and empirical standard errors for the exchangeable correlation structure, however for the Independent and the AR(1) correlation structures the differences between the standard errors are slightly bigger. This difference is happening due to the presence of the weights in the estimating procedure. Table 4.7 is a table comparing the three correlation structures under GEE and WGEE marginal models.

	GEE		WGEE		GEE		WGEE		GEE		WGEE	
	Exchangeable		Exchangeable		Independent		Independent		AR(1)		AR(1)	
Parameter	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Intercept	-5.03	1.45	-8.53	1.73	-5.04	1.45	-8.31	1.63	-5.03	1.46	-8.31	1.64
age 0	-0.93	1.22	1.37	1.46	-0.92	1.22	1.34	1.38	-0.93	1.23	1.30	1.39
age 1	-0.65	1.07	-2.46	1.35	-0.65	1.07	-2.52	1.28	-0.60	1.08	-2.49	1.28
age 2	-0.28	1.03	0.13	1.28	-0.28	1.04	-0.01	1.21	-0.24	1.04	0.00	1.21
age 3	-0.07	0.97	0.17	1.18	-0.07	0.98	0.03	1.11	-0.03	0.98	0.05	1.12
age 4	-0.67	0.95	0.76	1.09	-0.67	0.95	0.78	1.03	-0.65	0.96	0.79	1.03
age 5	-2.61	1.30	-1.67	1.49	-2.60	1.30	-1.50	1.34	-2.54	1.29	-1.48	1.34
age 6	-1.60	1.01	-0.94	1.11	-1.60	1.01	-0.89	1.03	-1.57	1.01	-0.88	1.03
age 7	-2.25	1.15	-1.90	1.29	-2.25	1.15	-1.74	1.17	-2.26	1.17	-1.74	1.17
age 8	-1.00	0.59	-0.75	0.64	-1.00	0.59	-0.72	0.62	-0.96	0.60	-0.71	0.62
age 9	-0.74	0.51	-0.49	0.55	-0.74	0.51	-0.46	0.53	-0.74	0.52	-0.45	0.53
age 10	-0.32	0.45	-0.09	0.49	-0.32	0.45	-0.07	0.47	-0.30	0.46	-0.06	0.47
age 11	-0.57	0.46	-0.49	0.49	-0.57	0.46	-0.44	0.47	-0.54	0.46	-0.43	0.47
age 12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
dt	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.00	0.01	0.01	0.01
prev	44.61	8.11	40.41	8.88	44.59	8.11	38.70	8.46	43.89	8.12	38.46	8.47
timemonth	-0.05	0.10	0.15	0.13	-0.05	0.10	0.15	0.12	-0.04	0.11	0.15	0.12
actipass 0	2.23	0.18	3.41	0.26	2.23	0.18	3.33	0.25	2.20	0.18	3.31	0.25
actipass 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.7: Model based standard errors and estimates for GEE and WGEE

Table 4.4 shows that the type III score statistics under WGEE for age, prev and actipass are all significant at the 5% same as in the case of GEE except that the magnitude of the estimates are slightly smaller. Tables 4.5 and 4.6 show that the model based parameter estimates as well as the standard errors for the WGEE are not very different from each other under all three correlation structures. Table 4.7 gives the results for both the GEE and WGEE analysis for the three working correlation structures. The WGEE parameter estimates are generally larger than those of GEE parameter es-

timates in all three correlation structures. The results also show that the standard errors of the parameter estimates for WGEE are larger than those for GEE. It should be emphasized that although this has happened it may not always hold in every scientific setting.

4.9 Conclusion

The generalized linear model provided a flexible method for modelling the data. The WGEE provided a means to model the dropout process and the results will be discussed in detail comparatively with the last observation carried forward (LOCF), avaliable data and likelihood based analyses in Chapter 9. It should also be noted that the WGEE analysis carried out in SAS took longer to converge when compared to an ordinary GEE analysis. The obvious extension of the generalized linear model is to include the child effect as the random effect in the model so that we now work in the class of generalized linear mixed models. This leads us to the next family of models namely, random effects models.

Chapter 5

Random Effects Models

5.1 Introduction

Searle (1988) stated that the concept of the mixed model was first described by Jackson (1939) even though the term "mixed model" was not used. In describing a 2-factor-no-interaction situation Jackson states that one factor is a "measure of the trial effect" and the other as a "measure of the individual effect". This seemed to be the first occurrence of the word "effect" in what is now its customary usage of linear models. Jackson further described his model as having one factor random and one non-random which seems as a clear specification of a mixed model, although not called so such, at the time. Associated with the term mixed model is whether or not it has fixed or random effects and whether it can be applied to balanced data or unbalanced data. So we first need to describe and define these concepts.

Balanced and unbalanced data

This concept was highlighted in Chapter 1 but will be defined here again for completeness.

- Data can be described as balanced when each cell in the data set contains the same number of observations and as unbalanced when this is not the case.

- For designed experiments, unbalanced data occur when the design of the experiment force the data to be so, in other words, when there is "planned" unbalancedness. Unbalanced data can also result from unfortunate circumstances or experimental carelessness, for example if the experimenter loses some of the data points. As a result the cell containing these missing observations vary in number with respect to other cells. Note that the current application involves longitudinal or repeated measurement data which is most often unbalanced because of unequal observations per individual. Hence the focus of the current study is on the analysis of unbalanced longitudinal binary response data with random individual effects.

Fixed and random effects models

- **Definition**: A factor in a model is random if its levels consist of a random sample from a population of all possible levels. A model is then a random effects model if all the factors in the treatment structure are random effects.

- **Definition**: A factor in a model is fixed if its levels are selected by a non-random process or consists of the entire population of all possible levels. A model is then a fixed effects model if all the factors in the treatment structure are fixed effects.

A model containing both fixed and random effects is therefore called a mixed effects model.

5.2 The Generalized Linear Mixed Model

One of the classical random effects models is known as the Beta-binomial model. This model was proposed by Skellam (1948) and then Kleinman (1973). As the name suggests this model is made up of a binomial part and a beta part. The Beta-binomial model is outside of the scope of this thesis and will not be considered. However the most frequently used random effects model for discrete outcomes is the generalized linear mixed model (GLMM). Generalized linear mixed models are also a straightforward extension of the generalized linear models for univariate data to the context of clustered measurements. With the advent of computational power, there has been a wide range of software tools available for fitting generalized linear mixed models. Examples of areas where GLMMs can be used include:

- estimating trends in disease rates
- modelling CD4 counts in a clinical trial over time for HIV infected individuals
- modelling the proportion of infected plants on experimental units in a design with randomly selected treatments or randomly selected blocks
- predicting the probability of high ozone levels for randomly selected countries
- modelling skewed data over time
- analyzing customer preference with respect to certain brands of clothing groceries etc.
- joint modelling of multivariate outcomes

Much of the work in this chapter is taken directly from Levin (1999).

5.2.1 Model Formulation

As in the case of the linear mixed model, discussed in Chapter 2, Y_{ij} is the jth outcome measured for subject i, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$ and $\boldsymbol{Y_i}$ is the n_i-dimensional vector of all measurements available for subject i. The random effects $\boldsymbol{b_i}$ is assumed be drawn independently from the $N(\mathbf{0}, D)$ and the outcomes Y_{ij} are conditionally independent given $\boldsymbol{b}_i$ with densities of the form

$$f_i(y_{ij}|\boldsymbol{b_i}, \boldsymbol{\beta}, \phi) = \exp\{\phi^{-1}[y_{ij}\theta_{ij} - \psi(\theta_{ij})] + c(y_{ij}, \theta)\}$$

for a known link function η such that $\eta(\mu_{ij}) = \eta(E(\boldsymbol{Y_{ij}}|\boldsymbol{b_i})) = \boldsymbol{x_{ij}'}\boldsymbol{\beta} + \boldsymbol{z_{ij}'}\boldsymbol{b_i}$where $\boldsymbol{x_{ij}}$ and $\boldsymbol{z_{ij}}$ are respectively, p-dimensional and q-dimensional vectors of known covariate values corresponding to the fixed and random effects $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$. The vector $\boldsymbol{\beta}$ is a p-dimensional vector of unknown fixed regression coefficient and $\boldsymbol{b}_i$ is a q-dimensional vector of random regression effects. The quantity ϕ is a scalar parameter. Finally let $f(\boldsymbol{b_i}|D)$ be the density of the $N(\mathbf{0}, D)$ distribution for the random effects $\boldsymbol{b_i}$.

5.2.2 Maximum Likelihood Estimation

The random effects models can be fitted by maximization of the marginal likelihood, obtained by integration of the random effects. The likelihood contribution of the i^{th} subject becomes

$$f_i(\boldsymbol{y_i}|\boldsymbol{\beta}, D, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{b_i}, \boldsymbol{\beta}, \phi) f(\boldsymbol{b_i}|D) d\boldsymbol{b_i} \tag{5.1}$$

from which the likelihood for $\boldsymbol{\beta}, D$, and ϕ is derived

$$\begin{aligned} L(\boldsymbol{\beta}, D, \phi) &= \prod_{i=1}^{N} f_i(\boldsymbol{y_i}|\boldsymbol{\beta}, D, \phi) \\ &= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{b_i}, \boldsymbol{\beta}, \phi) f(\boldsymbol{b_i}|D) d\boldsymbol{b_i} \end{aligned} \tag{5.2}$$

The problem here is maximizing Eq. (5.2) the product of N integrals over the q-dimensional random effects $\boldsymbol{b_i}$. However in some special cases these integrals can be worked out analytically such as in the case of the linear mixed model with continuous outcomes as well as the case of the Probit-normal model. In general there are no analytic expressions available for for the integrals of equation Eq. (5.2) and as a result numerical approximation methods are needed. Numerical approximations can be divided into the following classes:

1. Those that are based on the approximation of the integrand
2. Those based on an approximation of the data
3. Finally those that are based on the approximation of the integral itself.

Tuerlinckx et al (2004), Pinheiro and Bates (2000) and Skrondàl and Rabe-Hesketh (2004) give an overview of the currently available numerical approximations. We will now consider these methods of approximation briefly.

5.2.3 Estimation based on the approximation of the integrand

Whenever integrands are approximated, closed form expressions must be obtained so that numerical maximization of the approximated likelihood is feasible. All the proposed methods lead to Laplace-type methods. Tierny and Kadane(1986) use the Laplace method which is designed to approximate integrals of the form

$$I = \int e^{Q(\boldsymbol{b})} d\boldsymbol{b} \tag{5.3}$$

where $Q(\boldsymbol{b})$ is a known, unimodal and bounded function of a q-dimensional variable $\boldsymbol{b}$. Let $\hat{\boldsymbol{b}}$ be the value of $\boldsymbol{b}$ for which Q is maximized. The second-

order Taylor series expansion of $Q(\boldsymbol{b})$ is of the form

$$Q(\boldsymbol{b}) \approx Q(\hat{\boldsymbol{b}}) + \frac{1}{2}(\boldsymbol{b}-\hat{\boldsymbol{b}})' Q''(\hat{\boldsymbol{b}})(\boldsymbol{b}-\hat{\boldsymbol{b}}) \tag{5.4}$$

for $Q''(\hat{\boldsymbol{b}})$ equal to the Hessian of Q i.e. the matrix of the second-order derivative of Q evaluated at $\hat{\boldsymbol{b}}$. If we replace $Q(\boldsymbol{b})$ in Eq. (5.3) by its approximation from equation Eq. (5.4), we have

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

Clearly the integral in Eq. (5.2) is proportional to an integral of the form given by equation Eq. (5.3), for functions $Q(\boldsymbol{b})$ given by

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

such that Laplace's method can be applied. It is imperative to note that the mode $\hat{\boldsymbol{b}}$ of Q depends on the unknown parameters $\boldsymbol{\beta}$, ϕ and D such that in each iteration of the numerical maximization of the likelihood, $\hat{\boldsymbol{b}}$ will be recalculated conditionally on the current values for the estimates of these parameters. The Laplace approximation will be exact when $Q(\boldsymbol{b})$ is a quadratic function of $\boldsymbol{b}$ i.e. the integrands in Eq. (5.2) are exactly equal to normal kernels. Raudenbaush, Yang and Yosef (2000) extend the above Laplace method to include higher order terms in the Taylor series expansion in Eq. (5.4) up to the order 6. They note that this improves the overall approximation.

5.2.4 Estimation based on the approximation of the data

We now consider the second approach which is based on the decomposition of the data into the mean and an appropriate error term with a Taylor series

expansion of the mean which is a non-linear function of the linear predictor. Different methods in this approach exist because they use different orders in the Taylor series expansion and the point around which the approximation is expanded. Consider the decomposition

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij} = h(\boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i) + \varepsilon_{ij} \tag{5.5}$$

in which $h(.)$ equals the inverse link function $\eta(.)$ and where the error terms have the distribution with variance equal to $\text{Var}(Y_{ij}|\boldsymbol{b}_i) = \phi\upsilon(\mu_{ij})$ where $\upsilon(.)$ is as the usual variance function in the exponential family of distributions. We consider the decomposition where we have binary outcomes with the logit natural link function. This is directly applicable to our data set. We then have:

$$\mu_{ij} = P(Y_{ij} = 1) = \pi_{ij} = \frac{\exp(\boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{'}\boldsymbol{b}_i)}{1 + \exp(\boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{'}\boldsymbol{b}_i)}$$

with $\varepsilon_{ij} = 1-\pi_{ij}$ with probability π_{ij} and equals $-\pi_{ij}$ with probability $1-\pi_{ij}$. We will now consider two methods of approximation of the mean and hence the parameters. These methods are called *Penalized Quasi-Likelihood(PQL)* and *Marginal Quasi-Likelihood(MQL)*

Penalized Quasi-Likelihood (PQL)

We will firstly look at the Taylor series expansion of Eq. (5.5) around the current estimates $\hat{\beta}$ and $\hat{b}_i$ of the fixed effects and random effects. This gives us:

$$\begin{aligned}
Y_{ij} &\approx h(\boldsymbol{x}_{ij}^{'}\hat{\boldsymbol{\beta}} + \boldsymbol{z}_{ij}^{'}\hat{\boldsymbol{b}}_i) \\
&+ h^{'}(\boldsymbol{x}_{ij}^{'}\hat{\boldsymbol{\beta}} + \boldsymbol{z}_{ij}^{'}\hat{\boldsymbol{b}}_i)\boldsymbol{x}_{ij}^{'}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\
&+ h^{'}(\boldsymbol{x}_{ij}^{'}\hat{\boldsymbol{\beta}} + \boldsymbol{z}_{ij}^{'}\hat{\boldsymbol{b}}_i)\boldsymbol{z}_{ij}^{'}(\boldsymbol{b}_i - \hat{\boldsymbol{b}}_i) + \varepsilon_{ij} \\
&= \hat{\mu}_{ij} + \upsilon(\hat{\mu}_{ij})\boldsymbol{x}_{ij}^{'}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \upsilon(\hat{\mu}_{ij})\boldsymbol{z}_{ij}^{'}(\boldsymbol{b}_i - \hat{\boldsymbol{b}}_i) + \varepsilon_{ij}
\end{aligned}$$

where $\hat{\mu}_{ij}$ equals the current predictor $h(\boldsymbol{x}_{ij}'\hat{\boldsymbol{\beta}} + \boldsymbol{z}_{ij}'\hat{\boldsymbol{b}}_{\boldsymbol{i}})$ for the conditional mean $E(Y_{ij}|\boldsymbol{b_i})$ and h' the derivative of h w.r.t $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$ in the second and third terms respectively . In vector notation, this reduces to

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

for appropriate design matrices X_i and Z_i with $\hat{V}_i$ equal to the diagonal matrix with diagonal entries equal to $v(\hat{\mu}_{ij})$. Now if we reorder the above expression, it becomes

$$\boldsymbol{Y_i^*} \equiv \hat{V}_i^{-1}(\boldsymbol{Y_i} - \hat{\boldsymbol{\mu}}_{\boldsymbol{i}}) + X_i\hat{\boldsymbol{\beta}} + Z_i\hat{\boldsymbol{b}}_{\boldsymbol{i}} \approx X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} + \boldsymbol{\varepsilon_i^*} \tag{5.6}$$

for $\boldsymbol{\varepsilon_i^*}$ equal to $\hat{V}_i^{-1}\boldsymbol{\varepsilon}$ which has a zero mean. Eq. (5.6) can be viewed as a linear mixed model for the pseudo data $\boldsymbol{Y}_i^*$ with fixed and random effects as $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$ and error terms $\boldsymbol{\varepsilon}_i^*$. The theory of the linear mixed model was covered in detail in Chapter 3.

This then gives a useful and familiar algorithm for fitting the generalized linear mixed model. Given the starting values for the parameters $\boldsymbol{\beta}, D$ and ϕ in the marginal likelihood, empirical Bayes estimates are calculated for $\boldsymbol{b_i}$ and the pseudo data $\boldsymbol{Y_i^*}$. The approximate linear mixed model in Eq. (5.6) is fitted yielding estimates for $\boldsymbol{\beta}, D$ and ϕ. These estimates are then used in the pseudo data and and the procedure iterated until convergence is reached. The resulting estimates are called the *p*enalized quasi-likelihood (PQL) estimates. The PQL procedure was studied in detail independently by Breslow and Clayton (1993) and Wolfinger and O' Connell (1993).

Marginal Quasi-Likelihood(MQL)

This method is very similar to PQL except that we consider the Taylor series expansion of Eq. (5.5) of the mean around the current estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}_{\boldsymbol{i}} = \boldsymbol{0}$

of the fixed effects and random effects. This yields similar results as above except that the current predictor of the mean $\hat{\mu}_{ij}$ is of the form $h(\boldsymbol{x_{ij}}\hat{\boldsymbol{\beta}})$ rather than $h(\boldsymbol{x}'_{\boldsymbol{ij}}\hat{\boldsymbol{\beta}} + \boldsymbol{z}'_{\boldsymbol{ij}}\hat{\boldsymbol{b}}_{\boldsymbol{i}})$. The pseudo data is now $\boldsymbol{Y}_{\boldsymbol{i}}^{*} \equiv \hat{V}_i^{-1}(\boldsymbol{Y_i} - \hat{\boldsymbol{\mu}}_{\boldsymbol{i}}) + X_i\hat{\boldsymbol{\beta}}$ and thus satisfies the linear mixed model:

$$\boldsymbol{Y}_{\boldsymbol{i}}^{*} \approx X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} + \boldsymbol{\varepsilon}_{\boldsymbol{i}}^{*} \tag{5.7}$$

Again the model is fitted by iterating between the calculation of the pseudo-data and the approximate linear mixed model for this pseudo-data. The estimates are called *m*arginal quasi-likelihood(MQL) estimates. More of the details of the above method can be found in Breslow and Clayton (1993) and Goldstein (1991)

5.2.5 Some notes about the PQL and MQL methods

1. The essential difference between *PQL* and *MQL* is that *MQL* does not incorprate the random effects b_i in the linear predictor but both methods have the same key idea and will ideally have similar properties

2. The pseudo data Y_i^* determines the accuracy of both approximations.

3. Rodriguez and Goldman (1995) show that both *PQL* and *MQL* may be seriously biased when applied to binary response data, as is our case. Their simulations reveal that the fixed effects and variance components suffer from substantial, if not severe, attenuation bias in certain situations

4. Wolfinger (1998) showed that the Laplace, *PQL* and *MQL* methods perform badly in cases of binary repeated observations, with a relatively small number of observations available for all persons

5. Goldstein and Rasbash (1996) and Rodriguez and Goldman (1995) show that one of the ways to improve the accuracy of the approximations is to include a second order term in the Taylor series expansion. They call these methods *PQL 2* and *MQL 2*. They state that and *MQL2* performs only slightly better than *MQL* but *PQL 2* is substantially better than *PQL*

6. Breslow and Lin (1995) and Lin and Breslow (1996) suggest the inclusion of bias correction terms while Kuk (1995) suggested the use of iterative bootstrap

7. Within the *PQL* and *MQL* methods, the linear mixed model can be based on Maximum likelihood estimation(ML) or Restricted Maximum likelihood estimation(REML) both yielding slightly different results.

8. Quasi-likelihood methods are very similar linearization methods for fitting GEE's covered in Chapter 3.

5.2.6 The methods of Schall and Breslow and Clayton

In this derivation $\boldsymbol{u}$ denotes the vector of random effects. If we assume the random effects are $\boldsymbol{u} \sim N(0, \boldsymbol{G})$, the normal "errors" linear mixed model is given by

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}.$$

the conditional mean of the observations given the random model effects is

$$\mathrm{E}[\boldsymbol{y}|\boldsymbol{u}] = \boldsymbol{X\beta} + \boldsymbol{Zu}$$

and the conditional variance

$$\mathrm{var}[\boldsymbol{y}|\boldsymbol{u}] = \boldsymbol{R} = \mathrm{var}(\boldsymbol{e}).$$

The observations can then be described as

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

where $\boldsymbol{\mu}$ denotes the conditional mean $E[\boldsymbol{y}|\boldsymbol{\mu}]$. The Generalized Linear Mixed Model (GLMM) can also be described in terms of the conditional means. It takes the form:

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

where $\boldsymbol{\eta}$ is the linear predictor through the link function $\mathrm{g}(\boldsymbol{\mu})$**.** Thus we could write the GLMM as

$$\mathrm{g}\{\mathrm{E}[\boldsymbol{y}|\boldsymbol{u}]\} = \boldsymbol{X\beta} + \boldsymbol{Zu}$$

As in the conventional mixed model, the random model effects $\boldsymbol{u}$ are assumed to have a multivariate normal distribution with mean $\mathbf{0}$ and variance covariance matrix $\boldsymbol{G}$ while as in the conventional generalized linear model, the underlying distribution of $\boldsymbol{y}$ is assumed to be a member of the exponential family (for any given $\boldsymbol{u}$)

To see the difficulty that arises with GLMM's, suppose that we have a cluster sampled survey data with a random sample of clusters indexed $i = 1, \ldots, N$ with elementary units $j = 1, \ldots, n_i$ within each cluster. The observations satisfy generalized linear models with common distributions and a link function. We now introduce a random term u_i corresponding to the random cluster effect, and assume that the random effects u_i are normally distributed with a mean 0 and a variance σ_u^2. We put $u_i = \gamma z_i$ where $z_i \sim NID(0, 1)$

As an example we consider a logistic regression model for binary outcomes. Let

$$p_{ij} = \mathrm{Prob}(y_{ij} = 1)$$

so that

$$(p_{ij}|z_i) = \text{Prob}(y_{ij} = 1|z_i)$$

gives the probability of "success" for the j^{th} unit in cluster i, given the value of the random cluster effect Z_i in cluster i. Then the model can be written as

$$\text{logit}(p_{ij}|z_i) = x_{ij}\beta + \gamma z_i. \tag{5.8}$$

Overall therefore regardless of the cluster effect, p_{ij} is found by integrating over the random cluster effects to get

$$p_{ij} = \int_{z_i=-\infty}^{\infty} (p_{ij}|z_i)\phi(z_i)dz_i \tag{5.9}$$

where ϕ is the density function the standard normal random variable $(N(0,1))$. For model Eq. (5.8), the joint likelihood for $y_{ij}, j = 1, \ldots, n_i$ and $i = 1, \ldots, N$ involves the integral in Eq. (5.9) since

$$l(\beta, \gamma) = \sum_{ij} \{y_{ij}\text{log}p_{ij} + (n_{ij} - y_{ij})\text{log}(1 - p_{ij})\} + \text{contsant}. \tag{5.10}$$

The difficulty is that the integral has to be evaluated numerically. This can be done for example using the Gaussian Hermite formula for numerical quadrature, (discussed in detail in Section 5.7.10 of this chapter) but briefly under this method, an integral such as above is approximated by means of sums as follows:

$$\int_{-\infty}^{\infty} f(u)e^{-u^2}du \approx \sum_{j=1}^{m} c_j f(s_j) \tag{5.11}$$

where the values of c_j and s_j are given in standard tables. When we integrate out (numerically) the random effects, we obtain the **marginal** likelihood $l(\beta, \gamma)$ and this **marginal** likelihood can be maximized to find the maximum likelihood estimates (MLEs) of β and γ. This approach was used by Hedeker

and Gibbons (1994) who proposed a random effects ordinal regression model for the analysis of clustered response data. Now a binary response can be considered a special case of an ordinal response with only two(ordered) outcome categories. They developed the model for both the probit and the logistic response functions using the "threshold" concept in which it is assumed that the observed ordered category is determined by the value of a latent unobservable continuous response that follows a linear model incorporating random effects. Hedeker and Gibbons (1996) in addition developed a program called MIXOR (and an extended version MIXORE) to implement this method of marginal maximum likelihood estimation (MMLE). MIXORE is a public domain computer program that can be downloaded from the internet, together with a manual that describes how the data should be prepared for analysis, together with a specification file MIXORE.DEF which describes the setup of the the data in terms of which columns contain the random effects, which columns contain the fixed effects and which column contains the (ordinal) response. In this program, any given set of factors (i.e. categorical explanatory variables) should be first converted into the required number of indicator variables which are then stored as separate columns in the input data set. This could be seen as a drawback for analyses with a large number of factors each having a large number of levels (In the Kilifi data set, factors such as "visit"-44 levels). In addition, a constant term is required in the linear predictor of the model, thus the data file should have a column of 1's as one of the explanatory variables. For these reasons We will not use the MIXORE program in the analysis of the current data set.

There are a number of issues associated with this approach, among them:

- It is in fact relatively easy to extend NLMIXED analyses for correlated random effects such as those found in random coefficient regression

models. Such models are quite easy to fit in WinBUGS.

- It may be computationally demanding
- It is not widely implemented in existing commercial software packages (it is only freely available in MIXORE and also available in Stata versions 6 and 9)

One might wonder why we opt to use the likelihood conditional on the random effects $\boldsymbol{u}$ in the case of the non-normal response. The answer is that in the case of a normal response and the identity link, the random effects do not appear explicitly in the likelihood, but only appear through the variance covariance parameters $\boldsymbol{\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2}$ in the case of k random effects or equivalently through the "ratios or gammas" $\gamma_i = \frac{\sigma_i^2}{\sigma^2}$ where σ^2 is the residual variance. In the above example $\mathrm{E}(z_j) = 0$, thus $\mathrm{E}[\boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{\gamma z}_j] = \boldsymbol{x}_{ij}\boldsymbol{\beta}$. However $\mathrm{E}[g(\mu)] \neq g[\mathrm{E}(\mu)]$ in the case of the non-identity link function g, so taking the expectations will not cause the random terms to vanish. Hence we will consider a number of alternative approaches based on modifications to the mixed model equations.

5.2.7 Estimation approaches by Schall and by Breslow and Clayton

First consider the methods for obtaining parameter estimates for GLMMs. The fitting algorithm for the GLMM models is analogous to that for generalized linear models. The description below follows Waddington et al. (1994). For data $\boldsymbol{y}$ with mean $\boldsymbol{\mu}$, the mean is related to the linear predictor η by the link function: $\mathrm{g}(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zu}$ where $\boldsymbol{\beta}$ represents the fixed effects and $\boldsymbol{u}$ is a vector of random effects with $\mathrm{var}(\boldsymbol{u}) = \boldsymbol{G}$, a function of unknown variance components $\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2$. A working dependent

variate $\boldsymbol{y^*}$ is created by linearizing the link function applied to the data about the mean values, as follows $\boldsymbol{y^* = X\beta + Zu} + \mathrm{D}(\boldsymbol{\mu})(\boldsymbol{y - \mu})$ where $\mathrm{D}(\boldsymbol{\mu}) = \frac{dg}{d\mu} = \mathrm{diag}\{\mathrm{g}'(\mu_1), \mathrm{g}'(\mu_2), \ldots\}$. Thus the working variate has three components:

(i) fixed effects represented by $\boldsymbol{X\beta}$

(ii) random effects represented by $\boldsymbol{Zu}$ and

(iii) an error term represented by $\mathrm{D}(\boldsymbol{\mu})(\boldsymbol{y - \mu})$which depends on the distribution of $\boldsymbol{y}$ and the link function g through $\mathrm{D}(\boldsymbol{\mu})$

The first and third terms numbered above are the same as those of a standard working variate for the standard generalized linear model discussed under the PQL and MQL methods. The second and third terms give:
$\mathrm{var}(\boldsymbol{y^*}) = \boldsymbol{ZGZ'} + \boldsymbol{DVD}$ where $\boldsymbol{V}$is the variance of $\boldsymbol{y}$ conditional on the random effects $\boldsymbol{u}$, and $\boldsymbol{DVD}$ is a diagonal matrix.
Thus the working variate $\boldsymbol{y^*}$ is described by a linear mixed model with fixed effects $\boldsymbol{\beta}$ and random effects $\boldsymbol{u}$ and weights $\boldsymbol{W = (DVD)^{-1}}$. Given an estimate of $\boldsymbol{\mu}$, the standard mixed model equations can be used to estimate $\boldsymbol{\beta}$ and $\boldsymbol{u}$

$$\begin{bmatrix} X'WX & X'WZ \\ Z'WX & Z'WZ + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\mu} \end{bmatrix} = \begin{bmatrix} X'Wy^* \\ Z'Wy^* \end{bmatrix} \tag{5.12}$$

These equations can be solved using the REML algorithm discussed in Chapter 3, which will also give estimates of the variance components $\boldsymbol{G}$. It has been found that restricting the REML algorithm to two iterations to provide an approximate solution rather than allowing it to converge, does not affect the speed of convergence of the GLMM algorithm, and saves computing time (Welham, 1993). Given estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$, further estimates of $\boldsymbol{\mu}$ are

formed and used to update the working variate $\boldsymbol{y}^*$ and weights $\boldsymbol{W}$(which is also true in the case of generalized linear models) and the estimation is repeated until convergence.

The method of Schall (1991) uses an estimate of the conditional mean of $\boldsymbol{y}$ given $\boldsymbol{u}$ i.e. $\hat{\boldsymbol{\mu}} = \mathrm{g}^{-1}(\boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{u}}) = h(\boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{u}})$.

This method takes slightly different forms suggested by a number of authors. Breslow and Clayton (1993) refer to this method as penalized quasi-likelihood or PQL and derive the methods using a quasi-likelihood argument, which following Littell et al. (1996, Chapter 11) can be summarized as follows. In the the case of generalized linear models the estimating equations are determined by

$$Q(\mu_i, y_i) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)}$$

where μ_i is involved in the right hand side of the equation through the mean function θ_i.

In the normal "errors" mixed model of the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$$

the conditional mean of the observations given the random model effects is

$$\mathrm{E}[\boldsymbol{y}|\boldsymbol{u}] = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}$$

and the conditional variance

$$\mathrm{var}[\boldsymbol{y}|\boldsymbol{u}] = \boldsymbol{R} = \mathrm{var}(\boldsymbol{e})$$

The observations can be then described as

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

where $\boldsymbol{\mu}$ denotes the conditional mean $\boldsymbol{E}[\boldsymbol{y}|\boldsymbol{u}]$. In the generalized linear mixed model the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{u}$ plays the same role

as the distribution of $\boldsymbol{y}$ in the fixed effects generalized linear model i.e. the conditional quasilikelihood of an observation y_i given μ_i is

$$Q(\mu_i, y_i|u_i) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)}.$$

The joint quasilikelihood of the observations $\boldsymbol{y}$ is the sum of the quasilikelihood of $\boldsymbol{y}$ given $\boldsymbol{u}$ and the quasilikelihood of $\boldsymbol{u}$. In matrix terms the joint quasilikelihood is

$$\mathrm{Q}(\boldsymbol{\mu}, \boldsymbol{u}; \boldsymbol{y}) = \boldsymbol{y}^{'}\boldsymbol{A}^{-1}\boldsymbol{\theta} - (\boldsymbol{b}_{\boldsymbol{\theta}}^{0.5}\boldsymbol{A}^{-1}\boldsymbol{b}_{\boldsymbol{\theta}}^{0.5}) + \frac{1}{2}\boldsymbol{u}^{'}\boldsymbol{G}^{-1}\boldsymbol{u} \tag{5.13}$$

where $\boldsymbol{A}$ is the matrix of $a(\phi_i)$'s, $\boldsymbol{\theta}$ is the vector of $\theta(\mu_i)$'s, $\boldsymbol{b_\theta}$ is the vector of $b(\theta_i)$'s and $\boldsymbol{G}$ is as defined as var($\boldsymbol{u}$).

Breslow and Clayton (1993) and Wolfinger and O' Connell (1993) show that solutions for $\boldsymbol{\beta}$ and $\boldsymbol{u}$ can be obtained from $\mathrm{Q}(\boldsymbol{\mu}, \boldsymbol{u}; \boldsymbol{y})$ by iteratively solving the modified mixed model equations as described above. Engel and Keen (1994, 1996) discuss the relationship between Schall's method which is an extension of the iteratively reweighted least squares algorithm used in the estimation of generalized linear models, and PQL which is an approximation of ML estimation using Laplace integration and showed that they are equivalent. Schall's method (1991) or the PQL procedure assume that the scale parameter $a(\phi) = 1$. The Wolfinger-O´ Connell procedures which are implemented in the SAS macro GLIMMIX and which they call pseudolikelihood (PL) or restricted pseudolikelihood (REPL), assume that $a(\phi)$ is unknown. PL obtains a maximum likelihood type estimate of $a(\phi)$, while REPL obtains a REML like estimate. PQL is a special case of PL when $a(\phi) = 1$ (Wolfinger and O´ Connell (1993)). Note that an additional dispersion factor is also incorporated as a residual variance in the IRREML approach of Engel and Keen(1994). In the terminology of Zeger et al.(1988), Schall's method is a 'subject specific' (SS) model.

Goldstein (1995) derived the PQL method as an extension of IGLS to non-linear models, using a Taylor's series expansion for the non-linear function. We should note that in his notation, $a(\phi)$ is not assumed to be unity i.e. his model corresponds to the PL and REPL models of Wolfinger and O'Connell. The PQL model of Goldstein can be fitted using the software MLn.

An alternative to Schall's model is what the developers of the statistical software, Genstat refers to as the marginal model of Breslow and Clayton, variantly referred to by other authors (for example Goldstein, 1995) as "marginal quasilikelihood" or MQL. The algorithim for MQL is identical to that for Schall's model, except for the estimation of the mean $\boldsymbol{\mu} = \mathrm{g}^{-\mathbf{1}}(\boldsymbol{X\beta})$ (where the random term $\boldsymbol{Zu}$ is not used). This model is implemented in Genstat's GLMM procedure and in MLn. It can be considered as an approximation to that of Schall when the σ_i^2's are small, and in the terminology of Zeger et al. (1998) as a population averaged or PA model.

Breslow and Clayton (1993) point out that the detailed derivation of the PQL (Schall) model using quasilikelihood involves several ad hoc adjustments and approximations of which no formal justification is given; thus the derivation should be considered as providing heuristic motivation for the estimating algorithm that is used, and that this algorithm should be studied on its own right. They further note that some of the approximations made in the derivation are likely to improve as the normal approximation becomes more possible, for example as the denominators of binomial proportions increase or as the means of Poisson observations increase. This will not be the case for a binary outcome, which is what we observe in many surveys. Breslow and Clayton (1993) feel that even with binary data the PQL estimates of the fixed effects and their standard errors are sufficiently accurate for many practical purposes; however, inference on variance parameters are

less satisfactory and in simulation studies the procedure often converged to a non-positive definite matrix when the binomial denominator is 1 or 2. They point out that this is due to the fact that when the response probabilities are small and the data are highly discrete, only limited information is present for estimating random effects and their associated variances and covariances. While this applies to ML estimates as well, these may be consistent where PQL are not, for example for paired binary data with random pair effects.

Breslow and Clayton (1993) derive the model leading to MQL estimation (the marginal model of Breslow and Clayton as implemented in Genstat) by specifying the generalized linear model in terms of the marginal mean as $\mathrm{E}[y_i] = \mu_i = h(\boldsymbol{x_i'\beta})$ where h is the inverse link function. If the link function is not the identity, the marginal mean so defined does not in general coincide with the marginal mean calculated from the conditional formulation $\mathrm{E}[\boldsymbol{y}|\boldsymbol{u}] = \mathrm{h}(\boldsymbol{X\beta + Zu})$.

Zeger et al. (1988) investigated marginal models of the above form for longitudinal designs. They showed that the true marginal mean of the hierarchical model with normally distributed random effects could (at least approximately) be expressed in the form of the equation above, but with altered values for the regression variables or regression coefficients. With the log link for example they showed that $E[y_i] = \mu_i = \exp(\boldsymbol{x_i'\beta + 0.5 z_i' D z_i})$ so the random effects add an offset to the equations for the marginal mean. With the logit link the $\boldsymbol{\beta}$ coefficients are attenuated by a factor c_i such that

$$E(y_i) = \frac{\exp(c_i x_i'\beta)}{1 + \exp(c_i x_i'\beta)} \tag{5.14}$$

where $c_i = \det(c^2 D z_i z_i' + I)^{-0.5} = (1 + c^2 z_i' D z_i)^{-0.5}$ where $c = \frac{16}{\sqrt{3}}/[15\pi]$ (Breslow and Clayton (1993), equation (18)).

This result could be used to alleviate the bias in the estimation of both mean and variance parameters by treating the c_i as a multiplicative offset whose

values depend on the current model parameters θ. A number of suggested improvements to the estimation algorithms for both PQL and MQL will be considered below.

Breslow and Clayton (1993) compare the PQL and MQL methods; they point out that the regression coefficients of PQL depend strongly on the estimated variance components when we do not have an identity link, even in large samples, whereas this is not the case for MQL. They regard marginal models as being more appropriate when interest is focussed on the marginal relationship between covariates and the response, in which case the random effects model serves mainly to suggest a plausible covariance structure that enables us to get reasonably efficient estimating equations for the mean value parameters, while PQL is the method of choice for estimating parameters in the hierarchical model.

To consider this in more detail, we will now return to the concepts of population average (PA) and cluster specific (CS) models in survey analysis. The term subject specific (SS) and population average (PA) were introduced by Zeger et al. (1988) in an article discussing extensions of GLMs for the analysis of longitudinal data. As an example they considered data for 537 children from Steubenville Ohio, each of whom was examined from age 7 to 10 annually. Whether the child had respiratory infection in the year prior to each examination was reported by the mother. The mother's smoking status [regular smoker (1) or not (0)] was determined at the first interview, and was regarded as a time independent variable. They considered the following models:

(1) In the population average (PA) models the marginal probability of respiratory infection μ_{it} was assumed to satisfy the model

$$\text{logit}(\mu_{it}) = \beta_0^* + \beta_1^* x_1 + \beta_2^* x_2 + \beta_3^* x_3$$

where

$$\begin{aligned} x_1 &= 1 \text{ if the mother smokes and 0 otherwise} \\ x_2 &= \text{age in years since the 9th birthday and} \\ x_3 &= x_1 x_2 = \text{ age if the mother smokes and 0 otherwise} \end{aligned}$$

Zeger et al. (1988) argue that the $\boldsymbol{\beta^*}$ describe how the population averaged response depends on covariates, so in particular $\boldsymbol{\beta_1^*}$ compares the rate of respiratory disease for children whose mothers smoke to the rate for children whose mothers do not smoke.
(2) In the subject specific (SS) model, the probability of respiratory infection for an individual is described as

$$\text{logit}(w_{it}) = \beta_0^* + \beta_1^* x_1 + \beta_2^* x_2 + \beta_3^* x_3 + u_i$$

where in the $\text{logit}(w_{it}), w_{it} = \text{E}[y_{it}|u_i]$ and assumed $u_i \sim N(0, \sigma_u^2)$. Zeger et al. (1988) argue that in this case β_1 indicates how one child's risk would change if his/her mother changed his/her smoking status. They present a number of findings:

1. As has been noted above, with the logit link, the PA parameters are attenuated i.e. the random effects variability shrinks the fixed effects parameters towards 0, with the degree of shrinkage depending on the variance of the random effects

2. Estimates of the SS parameters are correlated with the variance parameters of the random effects, even asymptotically. Thus the precision of estimating $\boldsymbol{\beta}$ depends on that of the variance parameters, and these parameters are more difficult to estimate from the longitudinal data with a nonlinear link and a few observations per subject.

3. In PA models, only the link function needs to be correctly specified to make consistent inferences about PA coefficients. On the other hand the SS model uses not only the information contained in the population averaged response, but also a distributional assumption about the heterogeneity among subjects, thus for the SS models both the link function and the random effects distribution must be correctly specified for consistent inferences.

4. SS models are desirable when the response for an individual rather than the population is the focus, they give growth curves as an example. PA models are most efficiently used in population studies, such as in epidemiology. The difference in the population averaged response between two groups with different risk factors is more the focus than the change in an individual's response

Pendergast et al.(1996) comment in detail on the example given in Zeger et al.(1988) which has been discussed above. They point out that the difference in the interpretation of the maternal smoking effect in the two models is difficult to conceptualize, and insight may be gained from considering a slightly different scenario. Suppose that the covariate measured was whether or not the mother currently smoked, and that this was measured at the same time points as the child's respiratory disease status i.e. "mothers's smoking status" could now be considered as a within cluster covariate, since it could change from timepoint to timepoint (i.e. from subunit to subunit). The SS model would then allow direct observation and estimation of the average effect (in terms of the log odds ratio) of the change in smoking status upon respiratory disease status, assuming that there were subjects in which a change in smoking status was observed. The coefficient for "maternal smoking" represents the common log odds ratio for respiratory disease of mother's

smoking status across children. The PA model on the other hand ignores the fact that the effect of change in smoking status of the mother given a child had actually been measured, and succeeds in estimating only the odds ratio between smoking and nonsmoking mothers. Mothers who had changed smoking status would appear in both groups.

However in fact if no mother changes smoking status during the study, the covariate would again be a between subject covariate and no effect of changing smoking status can be directly observed. The PA model measures the log odds ratio between groups of mothers, whereas the SS model purports to measure the effect of change in the mother's smoking status and the interpretation of that coefficient is entirely model based; in fact it can be considered as a form of extrapolation with no data available to check it's validity. Thus Pendergast et al.(1996) conclude that SS interpretation of covariates which do not change within a subject is difficult. On the other hand with covariates that do change within a subject, marginal models ignore the observable information obtained when subjects serve as their own controls and a change is observed.

Neuhaus, Kalbfleish and Hauck (1991) compare SS and PA model approaches by analyzing clustered binary data, looking in particular at parameter interpretation. The data that is used is from a study of breast disease conducted in San Francisco. One component in the study consisted of obtaining a sample of fluid from both breasts of all study women. The binary outcome was whether a sample of breast fluid could ($Y = 1$) or could not ($Y = 0$) be obtained from each breast and the covariates considered were $X_2 =$ age in years, $X_3 =$ age in years at menarche, a binary indicator $X_4 = 1$ if the woman was parous and $X_4 = 0$ if not and a binary indicator of whether ($X_1 = 1$) or not ($X_1 = 0$) physical examination of each breast found evidence of dys-

plasia. The sample comprised 490 white premenopausal women who had no breast disease. Their comparison is between a model fitted using numerical quadrature and a model fitted using GEE, rather than PQL and MQL. They examine one representative of each approach; for SS models they look at the mixed effects logistic model, while PA models they use the GEE approach of Liang et al.(1986). They compare the two approaches for the case of a single covariate x, pointing out that the results generalize easily to the case of several covariates.

5.2.8 Inference for Generalized Linear Mixed Models

We know that from the section above that inference in the GLMM was carried out as follows:

- From section 5.2.7 the observed information matrix is $\boldsymbol{I}(\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$ could be used to construct confidence intervals and carry out hypothesis tests for individual elements of $\boldsymbol{\beta}$.

- We could test hypotheses about subsets of the regression variables by looking at the change in deviance when these terms are dropped from the model.

Littell et al. (1996, section 11.4) state that little research is done on small sample properties for GLMMs. By extension from the mixed model (with normal response) it should be noted that the ratio of the parameter estimates to their standard errors will not in general follow a Student's t distribution. Thus reference is carried out mainly using the Wald statistics. Littell et al. (1996, section 11.4) point out that these test statistics are basically reasonable as extensions of standard tests for mixed models and generalized linear models, and that more work is needed on these procedures, either to

validate them or develop corrections. Simulations carried out by Engel and Buist (1996) suggests that the procedures for confidence intervals and significance tests as developed for ordinary mixed models appear to perform well enough for practical use when applied to the adjusted dependent variate. The GLIMMIX procedure (macro) in SAS, which can fit both MQL (Breslow and Clayton, 1993) and PQL (Schall, 1991) models using repeated calls to PROC MIXED, provides for type III Wald statistics i.e. these enable us to test the significance of any terms in the model, conditional on the remaining model terms. When the scale parameter $a(\phi)$ is known, these statistics are approximately distributed as Chi-squared. If we have overdispersion (or underdispersion) in the case of the Poisson or Binomial models and $a(\phi)$ is unknown, then the Wald statistic is divided by the rank of $\mathbf{L}$ (the matrix used to formulate the hypothesis test) and this is approximately distributed as F with ν_1 and ν_2 degrees of freedom, where ν_1 =rank(L) and ν_2 in simple cases correspnds to the degrees of freedom required to estimate $a(\phi)$, but in more complex cases must be approximated using a Sattherwaite-type procedure. The GLMM procedure in Genstat can produce Type I Wald statistics, but it should be noted that these depend on the order in which terms are included in the model. So if we are using these statistics for model selection, we should refit the model with terms in a number of different orderings. An alternative likelihood based test statistic analogous to change in deviance in generalized linear models and will be considered later on.

As a comparison to SAS GLMM implementation in Genstat, the parameter estimates are calculated either using the the method of Schall (PQL) or the marginal method of Breslow and Clayton (MQL). Ignoring the random effects $\boldsymbol{u}$, this gives a linear predictor $\boldsymbol{X\beta}$ on the scale of the link function

g(.). Predicted means are calculated on this transformed scale in the way that REML calculates them by ignoring the random effects. Consider the case of how REML calculates the predicted means when the response is normally distributed. The predicted means are based on the estimates of the parameters in the model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$. If the design is balanced and orthogonal then the table of means produced in REML for fixed model terms are the same as the ordinary means. There is no such correspondence with unbalanced data, as with the Kilifi data. With REML the means are calculated from a linear transformation of the estimated parameter values, taking no account of the frequency counts for the different factor combinations. Therefore these predicted means will correspond to averages over the factor combinations only with orthogonal data. In the other cases, tables of means can be thought of as mean effects of factor levels adjusted for the mean values of any covariates and for any lack of balance in other factors; that is, as the means we would have expected if the data had been orthogonal. We should note that these means are not of the same type as those produced by default PREDICT directive in Genstat. In this case the marginal frequencies are used as weights for the averages of the factor combinations. Predicted means are calculated using all the parameter estimates and taking means over the model terms not present in the table. If we require the predicted means for the fixed model terms, these means need only to be taken over the estimates for the fixed model terms, since means over the the random terms will always be zero. These predicted means on the transformed scale are referred to as, for example, "predicted means for age". GLMM will also print a table headed "Back transformed means (on the original scale)", which in the case of binary data with a logit link are simply found by applying the

antilogit function

$$\mu = g^{-1}(\eta) = \exp(\eta)/\{1 + \exp(\eta)\}.$$

Genstat also issues a statement to emphasize that the "means are probabilities not expected values". Thus inference should be carried out on the transformed scale, and the back-transformed means are only to give an intuitive guide in interpreting the results. Note in Genstat there is a choice to use either the Schall (PQL) method or the marginal method of Breslow and Clayton (MQL) method. As stated earlier the only difference in the two methods is in the way the parameter estimates are formed.

5.2.9 Remarks on the problem of Bias in Generalized Linear Models

A number of simulation studies have been carried out to investigate the performance of MQL (Breslow and Clayton's method) and PQL (Schall's method). Breslow and Clayton (1993) investigated the performance of both MQL and PQL, both with respect to the estimation of the fixed effects and the estimation of the variance components/random effects. Their results for MQL confirmed the attenuation in the estimation regression coefficients that would be expected on theoretical grounds; they report that much of the bias in the estimation of both regression coefficients and the variance parameters could be alleviated by treating the terms c_i in Eq. (5.14) as a multiplicative offset whose value depended on the current θ (estimate of the random parameters)

The simulation studies revealed that for the PQL method, they found out that inference on the regression coefficients was approximately correct, even for binary data, with improved accuracy as the binomial denominators in-

creased. Inference on variance parameters was less satisfactory under the PQL method, with a tendency for the procedure to converge to a non-positive definite variance matrix when the binomial denominator was 1 or 2. They point out that this is due, at least partially, to the fact that when the response probabilities are small and the data are highly discrete, only limited information is present for estimating random effects and their associated variances and covariances. They also note that a further limitation of the PQL method is the failure to account for the contribution of the estimated variance components when the link function is not the identity, but this is not the case with the MQL method.

Rodriguez and Goldman (1995) carried out a large simulation study to investigate the performance of the MQL approach.The study was implemented using the statistical program ML3, which was the forerunner of MLn. The simulation study was based on a Guatemalan data set which consisted of a sub-sample of respondents in the 1987 National Survey of Maternal and Child Health. The survey was based on a national multistage clustered sample of 5610 women aged 15-44 years living in 240 communities. The logistic model selected as the basis of their simulations examined the determinants of use of modern (versus traditional) prenatal care during pregnancy among women who reported having obtained some kind of prenatal care, for births during the 5 years before the survey. The sample size of interest here consists of the 2449 births that occurred to mothers who received any prenatal care, with 45% of this sample having received modern prenatal care. The data are clustered on two higher order levels; the 2449 births were to 1558 mothers who live in 161 communities. Among the 1558 families, 52.4% had one child (born during the 5 years before the survey), 38.2% had two children, 9.1% had three children and 0.3% had four children. The sample sizes per com-

munity ranged from 1 to 50 with a mean of 15 children.

In the simulation Rodriguez and Goldman considered three covariates, one at each level. In the first four simulation sets they included random family and cluster effects in addition to the fixed components, with two magnitudes for the random effects, namely small (standard deviation 0.4) and large (standard deviation 1). These random effects were generated from independent normal distributions with mean 0 and designated variance (0.16 or 1) while the individual random component was generated from a standard logistic distribution.

They found that the fixed effects exhibited a clear downward bias, as Breslow and Clayton (1993) suggested would happen on theoretical grounds. Rodriguez and Goldman (1995) found that the bias was moderate when the random effects were small ($\sigma = 0.4$) but was fairly substantial when at least one of the random effects was large ($\sigma = 1$). For both random effects large they obtained fixed effects estimates of the order of 0.75 of the true values. They point out that the bias of 0.25 translates to an odds ratio of 0.78, indicating that the effects of the covariates on the odds of using modern prenatal care services would be underestimated by 22%. The estimates of the variance components had an even more pronounced downward bias, with a large number of simulations leading to estimates of the family effect equal to zero. Rodriguez and Goldman (1995) then carried out further simulations

1. To see whether the problems of bias were related to the use of a three level model. To achieve this, they used a two level model including only the family random effect (set to large or small) or the community random effect(set to large or small).

2. To see whether the problems of bias were related to the fact that in the the Guatemalan data structure (with an average of 1.5 children per

mother) they had very limited information on within family variation. To do this they carried out two additional data sets of simulations using rectangular structure of 20 communities, each of which having 20 families with 20 children each, for a total sample size of 800.

For the fixed effects the downward bias remained, and was of a similar magnitude with the corresponding previous simulation (i.e. large or small variance) for the three level Guatemalan structure. For the two level model including only a family random effect, the estimates of the variance components still had a downward bias, and in the case of the small variance 19% of the simulations led to estimates of the family effect equal to zero. The two level model including only a community random effect and the three level rectangular structure model showed a slight downward bias, but none of the family effect equal to zero. This suggests that the cluster size is important in estimating the variance components. It is noted that in their estimation, Rodriguez and Goldman used Iterative Generalized Least Squares (IGLS) which is equivalent to using maximum likelihood (ML) and this would lead to a downward bias in the variance components anyway; for improved results the Restricted Iterated Generalized Least Squares (RIGLS) or REML should be used in each iteration. Goldstein (1995, Chapter 5) and Goldstein and Rasbash (1996) suggest an improved approximation which largely eliminates the downward biases in the estimates from GLMMs. Goldstein (1995, Chapter 5 and Chapter 7) proposed an alternative approach of the estimation for nonlinear models, including GLMMs. Rodriguez and Goldman (1995) point out that Goldstein's procedure in general will produce the same results as those produce by the GLMM algorithm based on quasilikelihood. This follows since the two methods use exactly the same approximating linear model, based on a result due to Browne (1974) who proves that Generalized

Least Squares (GLS) and maximum likelihood are equivalent in the normal case, and that the Fisher's scoring method and GLS coincide when the variance matrix is linear in the unknown parameters, as in the case in variance component models if the parameterization is in terms of σ_u^2 (the variance component corresponding to the random effect $\boldsymbol{u}$.

Rodriguez and Goldman (1995) considered second order MQL estimation, and found that this improved the estimates, but only slightly. Goldstein and Rasbash (1996) carried out simulations based on those of Rodriguez and Goldman, and found that the PQL procedure considerably improved the model estimates with the only bias being in about 20% underestimation in their level-2 model. The greatest improvement occurred from a move from first to second order PQL. They also report the results of Ayis (1995) who showed that the second order PQL produced almost unbiased estimates for the fixed parameters and estimates that are no greater than 4% for the random parameters. Goldstein and Rasbash (1996) also reported on an iterated form of bootstrapping due to Kuk (1995) for producing asymptotically unbiased estimates.

Breslow and Lin (1995) and Lin and Breslow (1996) consider the bias in estimates of both fixed and random parameters, and give suggestions for bias correction procedures. Breslow and Lin (1995) studied the asymptotic bias of the variance component i.e. a single random component and the regression parameter (fixed parameter) estimates in GLMMs with a canonical link function and a single source of extraneous variation. In addition to PQL, they also considered approximations based on Laplace approximations of the integrated likelihood (to be revisited in later sections); the Laplace approxi-

mation approach has been used by Liu and Pierce (1993), Solomon and Cox (1992) and Wolfinger (1993) among others. Breslow and Lin (1995) provided a correction factor for the variance component estimate derived from Laplace approximation and from PQL, and also a first order correction term for the regression coefficients estimated by PQL. They found that the proposed bias corrected PQL estimates significantly improve the asymptotic performance of the uncorrected quantities.

Lin and Breslow (1996) generalize these results of Breslow and Lin (1995) to GLMMs with multiple sets of random effects. They focus on correcting the bias in PQL estimates, since Breslow and Lin (1995) found that in some circumstances the Laplace approximation methods may be numerically unstable. However they use a generalization of the asymptotic expressions derived by Solomon and Cox (1992) for the Laplace approximations to multiple components of dispersion to derive their bias correction procedure and derive a quadratic expansion of the integrated log-quasilikelihood. These issues then led Lin and Breslow to propose a 4-step algorithm to achieve the bias corrected PQL estimates of the regression coefficients and variance components. Lin and Breslow (1996) evaluated the performance of these correction procedures by reanalyzing the well known example of the salamander mating experiment reported by McCullagh and Nelder (1989, section 14.5) and carrying out simulation studies. For the salamander data they found that the performance of the bias correction procedure was unsatisfactory and they attributed this to the large variability of the random effects in the actual salamander data. Their simulation studies showed more positive results; in particular the simple correction procedure for the variance components effectively reducing the bias in the PQL estimates of $\boldsymbol{\theta}$ and the associated mean

square error when the sample size was reasonably large. Note that $\boldsymbol{\theta}$ is the vector of variance component parameters.

They note that attempts to reduce bias are not always desirable, and that the effectiveness of the correction procedure for a particular problem will depend on both the sample size and the conditional form of responses. The corrections often inflate the variances of the parameter estimates, especially in problems involving very small variance components and small sample sizes. The biases in first order and second order corrected regression coefficients are negligible for small amounts of dispersion. When the variance components are between 0.5 and 1 in problems involving binary outcomes, the second order correction perform better. However they point out that both corrections break down for larger variance components. They also note that, from the results of other simulations studies, caution is required when applying corrected PQL (CPQL) to the regression coefficients when the binomial denominators are small. Further as the binomial denominator increases, the PQL method itself yields satisfactory estimates and the corrections may not be necessary. They suggest that the best procedure for general use may be the correction of the variance components and recalculation of the PQL regression components $\boldsymbol{\beta}$ using the corrected PQL variance components.

Engel and a number of his co-workers have also investigated the problem of bias in the estimates from GLMM, and in particular the estimates of the variance components. Engel and Keen (1994) point out that the obvious estimates of the iterative weights, used in GLMMs are:

$$\hat{w}^{-1} = V(\hat{\mu})[g^{'}(\hat{\mu})]^2$$

where the estimate of μ is obtained in the first step of the estimation procedure. They note that $V(\hat{\mu})[g^{'}(\hat{\mu})]^2$ may often be an accurate prediction for $V(\mu)[g^{'}(\mu)]^2$, particularly in the case of a single random effect, where the variance component σ_u^2 is small and is not necessarily a consistent estimator. They recommend the use of alternative weights which depend on both the link and the assumed variance function. In the case of the logit link with a binomial variance function, the alternative weights suggested by Engel and Keen (1994) are given by

$$w_0 = \{2 + 2\exp\frac{\sigma_u^2}{2}\cosh(x^{'}\beta)\}^{-1}.$$

Engel, Buist and Visscher (1995) carried out a number of simulation studies in animal breeding and found out that both the magnitude and direction of the bias in the estimate of the variance component depend on the number of fixed effects and also on the underlying response probability with over estimation of the variance component σ_u^2 when there are a large number of fixed effects and the overall incidence is above 0.9. They consider models in animal breeding which potentially can have over 100 fixed effects.

Engel and Buist(1998) further investigated bias in GLMMs, also looking at the correction method of Lin and Breslow (1996). They found out that while the correction of Lin and Breslow is useful for small or moderate numbers of fixed effects, it is of little benefit in animal breeding studies which commonly have a large number of fixed effects. They also report that the alternative weights of Engel and Keen may alleviate the bias and reduce the MSE, but not for a large number of observations per random effect (in survey data analysis, say more than 40 observations per cluster). Engel (1998) considers a single example to illustrate the asymptotic bias in GLMM estimates of the variance component σ_u^2 and finds that this can be underestimated by

almost half. He suggests that the best procedure for overcoming bias could be the use of a Markov chain Monte Carlo method such as the Gibbs sampler. However such methods will not be the focus in the current work.

5.2.10 Estimation based on the approximation of the integral

When approximation methods fail, approximations to the integral or numerical integration proves to be very useful. Pinheiro and Bates (1995, 2000) suggest the use of adaptive quadrature rules for random effects models where the numerical integration is centred around the EB estimates of the random effects and the number of quadrature points is then selected in terms of the desired accuracy. We will consider Gaussian and adaptive Gaussian quadrature designed for integrals of the form:

$$\int f(z)\phi(z)dz, \tag{5.15}$$

for a known function $f(z)$ and for $\phi(z)$ the density of the univariate or multivariate standard normal distribution. We will standardize the random effects so that they get the identity covariance matrix. That is, let $\boldsymbol{\delta_i}$ be equal to $\boldsymbol{\delta_i} = D^{-1/2}\boldsymbol{b_i}$, where $\boldsymbol{b}_i$ is the original vector of random effects of the model. Note that $\boldsymbol{\delta_i}$ is normally distributed with a mean of $\mathbf{0}$ and covariance I. The linear predictor then becomes $\theta_{ij} = \boldsymbol{x_{ij}'}\boldsymbol{\beta} + \boldsymbol{z_{ij}'}D^{1/2}\boldsymbol{\delta_i}$. Hence the variance components in D is now in the linear predictor. The likelihood contribution for subject i is now

$$f_i(\boldsymbol{y_i}|\boldsymbol{\beta}, D, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{b_i}, \boldsymbol{\beta}, \phi) f(\boldsymbol{b_i}|D) d\boldsymbol{b_i} \tag{5.16}$$

$$= \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\delta_i}, \boldsymbol{\beta}, D, \phi) f(\boldsymbol{\delta_i}) d\boldsymbol{\delta_i} \tag{5.17}$$

Notice that Eq (5.17) is of the form as Eq (5.15) which means that it can be applied to Gaussian or adaptive Gaussian quadrature approximations can be applied to it.

Gaussian Quadrature

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

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

Here Q is the order of the approximation. The higher Q is the more accurate the approximation will be. The so-called nodes or quadrature points z_q are solutions to the Qth order Hermite polynomial whilst the w_q are called the weights. The weights w_q and the nodes z_q can be found in tables. However Press et al. (1992) give an algorithm to calculate these weights. One of the main disadvantages of Gaussian quadrature is highlighted in the case of univariate integration i.e. the quadrature points z_q are chosen based on $\phi(z)$, independent of the function $f(z)$ in the integrand. Hence, depending on the support of $f(z)$, the z_q will or will not lie in the region of interest. (Molenberghs and Verbeke, 2005, p.273-274)

Adaptive Gaussian Quadrature

In this modification the quadrature points are scaled as if the $f(z)\phi(z)$ were a normal distribution with the mean of this distribution being the mode $\hat{z}$ of $\ln[f(z)\phi(z)]$, while the variance is equal to

$$\left[-\frac{\partial^2}{\partial z^2}\ln[f(z)\phi(z)]|_{z=\hat{z}}\right]^{-1}.$$

The new adaptive quadrature points are given by

$$z_q^+ = \hat{z} + \left[-\frac{\partial^2}{\partial z^2} \ln[f(z)\phi(z)]|_{z=\hat{z}} \right]^{-1/2} z_q$$

and the corresponding weights are given by

$$w_q^+ = \left[-\frac{\partial^2}{\partial z^2} \ln[f(z)\phi(z)]|_{z=\hat{z}} \right]^{-1/2} \frac{\phi(z_q^+)}{\phi(z_q)} w_q.$$

The integral is then approximated by

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

When fitting generalized linear mixed models, an approximation is applied to the likelihood contribution of each of the N subjects in the data set. The higher the order of Q, the better the the approximation will be of the N integrals in the likelihood. Adaptive Gaussian quadrature needs (much) less quadrature points than classical Gaussian quadrature but is certainly more time consuming in its iteration process. This is so because the functions in Eq. (5.15) as well as the quadrature points and the weights depend on the unknown parameters, β, D and ϕ, and hence need to be updated in every step of the iterative estimation procedure. (Molenberghs and Verbeke, 2005, p.275-276)

5.2.11 A note on the inference on the fixed and random effects in GLMMs

The fitting of generalized linear mixed models is based on maximum likelihood principles and the inferences for the parameters are readily obtained from classical maximum likelihood theory. Therefore, if we assume that the fitted model is a suitable one, then the estimators that are obtained are

asymptotically normally distributed with the correct values as the means and the inverse of the Fisher information matrix as the covariance matrix. As a result tests such as the Wald type tests, can be used for comparing standardized estimates to the standard normal distribution. Composite hypothesis tests where the Wald statistic which is a standardized quadratic form can be compared to the Chi-squared distribution or likelihood ratio and score tests can be used as well. The Z, t and F tests can be used to test for the fixed effects since the precision estimates for the fixed and random effects are obtained using the linear mixed model methodology discussed in Chapter 3. It is imperative to realize here that the inference of the estimates depends on the assumed sampling distribution. In linear mixed model methodology, the t or F require the normality assumption of the response vector $\boldsymbol{Y_i}$. Asymptotic normality for the fixed effects will follow as a result of the Central Limit Theorem. However for the EB estimates of the underlying normal distribution approximation may be questionable because the posterior distribution of the estimates may be skewed. Molenberghs and Verbeke (2005, p.277) state that one should not throw caution to the wind in trying to interpret output from the linear mixed model that was fitted to the pseudo data (discussed above), for example likelihood ratio tests must be based on the likelihood in Eq. (5.2) of the observed data and not on the likelihood associated with the linear mixed model for the pseudo data.

The validity of inference for the variance components, D will hold for classical Wald, likelihood ratio and score tests as long as the hypotheses that are being tested are not on the boundary of the parameter space. Stram and Lee (1994,1995), Verbeke and Molenberghs (2000, chapter 6) and Verbeke and Molenberghs (2003) all note this point which can be illustrated by considering the example where one wants to test whether the variance τ^2

of a single random effect in generalized linear mixed modelis equal to zero, meaning that one has got to test the following hypothesis:

$$H_0 : \tau^2 = 0$$

against

$$H_1 : \tau^2 > 0$$

Clearly the null hypothesis is on the boundary of the parameter space that is, $\tau^2 \geq 0$ meaning that none of the classical Wald, likelihood ratio and score tests are still valid. This is notable because the classical Wald test is based on the standard normal approximation to the standardized maximum likelihood estimate $\hat{\tau}^2$. This means that this Z statistic can not be normally distributed with a mean of zero since $\tau^2 > 0$. Thus under H_0, this Z statistic follows the positive normal distribution on 50% of the cases and will equal to zero in the other 50% of the cases. This leads to the well known mixture of the Chi-square distributions as the null distributions. These facts will also hold for the one-sided likelihood ratio and score tests.

5.3 Software for Generalized Linear Mixed Models

Due to the vast computational power that has been developed in modern times, there are many commercially available computing software packages that are available for fitting generalized linear mixed models. SAS is one the more flexible packages that allow do the fitting of generalized linear mixed models. The GLMM procedure in Genstat and the SAS procedure GLIMMIX and NLMIXED in SAS version 9.1 are suitable for fitting the GLMM's. Other packages include HLM by Raudenbaush et al.(2001), EGRET (Cytel Software

Corporation 2000), *gllamm* in Stata (Rabe-Hesketh, Pickles and Skrondal, 2001) and MIXOR and MIXREG (Hedeker and Gibbons, 1994). Details of the GLIMMIX procedure in SAS are briefly outlined below because this will be used to model analyse the RSV data in the current research problem.

5.3.1 SAS GLIMMIX for Quasi-likelohood

As already mentioned, the GLIMMIX procedure is still under experimentation in SAS version 9.1. Some of the aspects of the GLMM procedure in Genstat have been discussed in previous sections and hence those details will not be repeated here. The code for the GLIMMIX procedure is very similar to that of PROC MIXED, and the reason for this is that the GLIMMIX procedure calls the PROC MIXED procedure each time a linear mixed model needs to be fitted to newly updated pseudo data. It is imperative to realize that the most important option is the 'method=' in the GLIMMIX statement. In this statement the type of quasi-likelihood is specified namely PQL or MQL. If we chose the PQL option based on REML for the linear mixed models then we will set 'method=RSPL' (RSPL refers to Residual PL). Other available options include:

GLIMMIX option	Quasi-likelihood type PQL/MQL	Inference pseudo-data ML/REML
'method=RSPL'	PQL	REML
'method=MSPL'	PQL	ML
'method=RMPL'	MQL	REML
'method=MMPL'	MQL	ML

The 'CLASS' statement specifies which variables should be considered as factors and such classification variables can be either set as character or numeric. The 'MODEL' statement names the response variable and all covariate vec-

tors corresponding to the fixed effects. By default the intercept is added, however if one does not require a model to be fitted with an intercept then the option 'noint'is used. The 'solution' option is used to request the printing of all the estimates of the fixed effects included in the model together with standard errors, t-statistics, p-values and confidence intervals. The 'dist=' is used to specify the conditional distribution of the data given the random effects. Various distributions are available and include the normal, Bernoulli, Binomial and Poisson distribution. The link function function is set as the natural link by default but other link functions such as the probit, log-log, or identity link can be requested by adding in the appropriate 'lin=' option The 'RANDOM' statement defines the vectors corresponding to the random effects in the models and when random intercepts are required then they should be explicitly specified as opposed to the 'MODEL' statement where an intercept is included by default. The 'subject=' option is used to identify the subjects in the data set. If one requires random slopes for the time trend to be included, then this can be obtained by replacing the 'RANDOM' statement by

" random intercept time/ subject=... type=un;"

In this statement "type=un" specifies that the random effects covariance matrix D is an unstructured 2x2 matrix. Other special structures are available such as for models which assume equal variance for the intercepts and slopes or models which assume independent intercepts and slopes and others. In earlier versions of the SAS software, the PROC GLIMMIX was only available through the GLIMMIX macro which gave rise to the GLIMMIX procedure in SAS version 9.1.

5.3.2 The NLMIXED Procedure for Numercial Quadrature

Gaussian and adaptive Gaussian quadrature as approximations to the integrals in marginal likelihood have been implemented in the SAS procedure NLMIXED. This procedure is a versatile one with many statements and options. In the current study, focus is on the statements needed to fit a generalized linear mixed model. The NLMIXED procedure requires completely different model specifications than most SAS procedures but allows the user a very high degree of flexibility in the way the model is specified and parameterized. One of the consequences of this flexibility is that the user needs not only to specify the model but also has to specify names for all the parameters in the model. SAS then considers the symbols in the model specification that are not referring to variables in the input data set as unknown parameters to be estimated from the data.

The option'noad' in the NLMIXED statement is to request non-adaptive quadrature since the default that is used is adaptive quadrature. The option 'qpoints=' specifies the number of quadrature points. If this option is omitted then the number of quadrature point is selected adaptively by evaluating the log-likelihood function at the starting values of the parameters until two successive evaluations show sufficiently small relative change. It is also important to note that the model fitting based on Laplace approximation for the marginal integrals can be specified using adaptive Gaussian quadrature with only one quadrature point. The PARMS statement is used to specify the starting values for all parameters in the model. Parameters not listed in the PARMS statement are given an initial value of 1. This is one of the major drawbacks of the current version of the NLMIXED procedure, the fact that the procedure does not generate starting values except for the default

value of 1 which is given to all parameters that do not appear in the PARMS statement. In complex models however, convergence of the numerical optimization algorithms may highly depend on the specified initial values.

The MODEL statement is used to specify the conditional distribution of the data given the random effects. Various distributions such as the Normal, Bernoulli, Binomial and Poisson distributions can be specified. In case models are needed which do not fit within any of the classical distributions, user-defined likelihoods can be specified through the option 'model y $\sim$ general(ll)' in which ll is the user defined log-likelihood.

The RANDOM statement defines the random effects in the model. One has the flexibility of specifying the random effects in the sense that one can estimate the random-intercepts variance rather than the standard deviation. Further to this a mean model can also be specified for the random effects. There are also ways in which one can specify the model to incorporate independence of random intercepts and slopes as well as if one wanted to estimate directly the correlation between random intercepts and slopes rather than their covariance.

The 'subject=' option determines when new realizations of the random effects occur. The procedure assumes the occurrence of a new realization whenever the value of the variable specified in the 'subject=' option changes from the previous observation. The RANDOM statement also allows the inclusion of an output option of the form 'out=data set' which request and output data set containing empirical Bayes estimates for the random effects together with their approximate standard errors. The SAS version 9.1 only allows one RANDOM statement meaning that multilevel models can not be fitted to incorporate random effects at different levels. SAS does have other procedures to fit these multilevel models.

5.3.3 The Random Intercept Model

The random intercept model is given a special attention because of its relevance to the current childhood respiratory disease data, the subject of study in the current research. First we briefly revisit the case of a normal response Y.

Fitzmaurice, Laird and Ware (2004, pp. 188-199) define a random intercept model as a linear model with a randomly varying subject effect. In this model, each subject is assumed to have an underlying level of response that persists over time. This is incorporated into the linear mixed effects model by regarding this subject effect as random, yielding the following model

$$Y_{ij} = X_{ij}^{'}\beta + b_i + e_{ij} \tag{5.18}$$

where b_i is the random subject effect and the e_{ij} are regarded as measurement or sampling errors. From the model above, the response for the i^{th} subject at the j^{th} occasion is assumed to differ from the population mean, $X_{ij}^{'}\beta$, by a subject effect b_i, and a within subject measurement error, e_{ij}. Both the subject effect and the measurement error are assumed to be random, with a mean 0 and variances $\text{Var}(b_i) = \sigma_b^2$ and $\text{Var}(e_i) = \sigma^2$, respectively. Furthermore it is assumed that b_i and e_i are independent. The model describes the mean response trajectory over time for any individual as the conditional mean of Y_{ij} given the subject specific effect,

$$E(Y_i|b_i) = X_{ij}^{'}\beta + b_i$$

and the marginal mean of Y_{ij}, the mean response profile in the population as,

$$E(Y_{ij}) = X_{ij}^{'}\beta.$$

The interpretation of the model,

$$Y_{ij} = X_{ij}^{'}\beta + b_i + e_{ij}$$

is that the regression parameters β describe the patterns of change in the mean response over time (and their relation to covariates) in the population of interest, while b_i describes how the trend over time for the i^{th} individual deviates from the population average where b_i represents an individuals deviation from the population mean intercept, after the effects of covariates have been accounted for and when b_i is combined with the fixed effects it describes the trajectory over time for any individual. This is seen if we expand the above model as:

$$\begin{aligned} Y_{ij} &= X_{ij}^{'}\beta + b_i + e_{ij} \\ &= \beta_1 X_{ij1} + \beta_2 X_{ij2} + \ldots + \beta_p X_{ijp} + b_i + e_{ij} \\ &= \beta_1 + \beta_2 X_{ij2} + \ldots + \beta_p X_{ijp} + b_i + e_{ij} \\ &= (\beta_1 + b_i) + \beta_2 X_{ij2} + \ldots + \beta_p X_{ijp} + e_{ij} \end{aligned}$$

where $X_{ij1} = 1$ for all i and j and β_1 is therefore the fixed effect intercept term in the model. When expressed this way, it can be seen that the intercept for the i^{th} individual is $\beta_1 + b_i$ and varies randomly from one individual to another. Because the mean of the random effect b_i is assumed to be zero, b_i represents deviation of the i^{th} individual's intercept $\beta_1 + b_i$ from the population intercept β_1. Next, the marginal mean of Y_{ij} is given as

$$E(Y_{ij}) = \mu_{ij} = X_{ij}^{'}\beta$$

and the marginal variance of each response is given by,

$$\begin{aligned}
\text{Var}(Y_{ij}) &= \text{Var}(X_{ij}^{'}\beta + b_i + e_{ij}) \\
&= \text{Var}(b_i + e_{ij}) \\
&= \text{Var}(b_i) + \text{Var}(e_{ij}) \\
&= \sigma_b^2 + \sigma^2.
\end{aligned}$$

Similarly the marginal covariance between any pair of responses, Y_{ij} and $Y_{ik}, j \neq k$ is given by

$$\begin{aligned}
\text{Cov}(Y_{ij}, Y_{ik}) &= \text{Cov}(X_{ij}^{'}\beta + b_i + e_{ij}, X_{ik}^{'}\beta + b_i + e_{ik}) \\
&= \text{Cov}(b_i + e_{ij}, b_i + e_{ik}) \\
&= \text{Cov}(b_i, b_i) \\
&= \text{Var}(b_i) \\
&= \sigma_b^2
\end{aligned}$$

The marginal covariance pattern of the repeated measurements is exhibited in the following compound symmetry pattern

$$\text{Cov}(Y_i) = \begin{pmatrix}
\sigma_b^2 + \sigma^2 & \sigma_b^2 & \sigma_b^2 & \dots & \sigma_b^2 \\
\sigma_b^2 & \sigma_b^2 + \sigma^2 & \sigma_b^2 & \dots & \sigma_b^2 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \dots & \sigma_b^2 + \sigma^2
\end{pmatrix}$$

The random intercept model can be extended into the Generalized Linear Mixed model and this is illustrated in Section 5.3.4 and 5.3.5 below.

5.3.4 Generalized Linear Mixed Model for Counts

Suppose that the response Y_{ij} is a count. Then using a three part specification (described in section 5.3.3 above):

1. Conditional on a vector of random effects b_i, the Y_{ij} are independent and have a Poisson distribution with $\text{Var}(Y_{ij}|b_i) = E(Y_{ij}|b_i)$ with $\phi = 1$.

2. The conditional mean of Y_{ij} depends upon the fixed and random effects via the following linear predictor:

$$\eta_{ij} = X_{ij}^{'}\beta + Z_{ij}^{'}b_i$$

   where $X_{ij}^{'} = Z_{ij}^{'} = (1, t_{ij})$, with

$$\log\{E(Y_{ij}|b_i)\} = \eta_{ij} = X_{ij}^{'}\beta + Z_{ij}^{'}b_i$$

   That is, the conditional mean of Y_{ij} is related to the linear predictor by a log link function, this is an example of a log linear mixed effects model.

3. The random effects are assumed to have a bivariate normal distribution, with a zero mean and 2×2 covariance matrix G

This example is one of a log linear regression model with randomly varying intercepts and slopes. The models implies that there is natural heterogeneity among individuals in both their baseline level and changes in their expected counts over time.

5.3.5 Generalized Linear Mixed Model for a Binary Response

Suppose that Y_{ij} is a binary response, taking values 0 or 1. A logistic mixed effects model for Y_{ij} is given below:

1. Conditional on a single random effect b_i, the Y_{ij} are independent and have a Bernoulli distribution with $\text{Var}(Y_{ij}|b_i) = E(Y_{ij}|b_i)\{1-E(Y_{ij}|b_i)\}$ with $\phi = 1$.

2. The conditional mean of Y_{ij} depends upon the fixed and subject specific random effects via the following linear predictor:

$$\eta_{ij} = X_{ij}^{'}\beta + Z_{ij}^{'}b_i = X_{ij}^{'}\beta + b_i$$

   where $Z_{ij} = 1$ for all $i = 1, \ldots, n_i$ with

$$\log\left\{\frac{Pr(Y_{ij}=1|b_i)}{Pr(Y_{ij}=0|b_i)}\right\} = \eta_{ij} = X_{ij}^{'}\beta + b_i,$$

   That is, the conditional mean of Y_{ij} is related to the linear predictor by a logit link function. This is the random intercept model discussed earlier.

3. The single random effect are assumed to have a univariate normal distribution, with a zero mean and variance, say g_{11} because G is now of dimension 1×1

This example is one of a simple logistic regression model with randomly varying intercepts. The model implies that there is natural heterogeneity in individuals propensity to respond positively that persists throughout all binary responses obtained on any individual.

5.4 Analysis and Application to the RSV data

A marginal model was first fitted in Genstat with the RSV infection status (infected and not infected), being the response variable, age, dt, prevalence, actipass and time-month being the fixed effects. Note that dt=time

between events variable, actipass=sampling method, active or passive and time-month=time in months. The child effect was taken as a random effect in the model. The codes for the other variables used in this model have already been discussed in Chapter 1. The Binomial distribution with a logit link was used. More specifically, it is assumed that, conditionally on subject specific, random effects, u_i, the RSV status response variable, Y_{ij} for child i at time j is Bernoulli distributed with mean π_{ij} that is, $Y_{ij}|u_i \sim \text{Bernoulli}(\pi_{ij})$ and can be modelled as

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + u_i$$

where x_1 represents the 'age' effect with levels from $0, \ldots, 12$, x_2 is the 'dt' effect, x_3 represents the 'prev' effect, x_4 represents the 'actipass' effect with levels 0 and 1, x_5 represents the 'timemonth' effect and u_i is the random effect of the child that is normally distributed with a mean of 0 and a variance of τ^2. The results from Genstat after the algorithm had converged using both the PQL and MQL methods are shown in Tables 5.1 to 5.4. The Wald tests for fixed effects under the PQL method of Schall (1991) are summarized in Tables 5.1 and 5.2 while the methods of Breslow and Clayton (1993) are summarized in Tables 5.3 and 5.4.

Fixed term	Wald statistic	d.f	Wald/d.f.	chi pr
age	29.31	12	2.44	0.004
dt	0.02	1	0.02	0.876
prev	55.67	1	55.67	< 0.001
actipass	153.86	1	153.86	< 0.001
timemonth	0.15	1	0.15	0.703

Table 5.1: Wald tests by adding terms sequentially to the model

Fixed term	Wald statistic	d.f	Wald/d.f.	chi pr
timemonth	0.15	1	0.15	0.703
actipass	153.62	1	153.62	< 0.001
prev	28.98	1	28.98	< 0.001
dt	0.07	1	0.02	0.784
age	19.09	12	1.59	0.086

Table 5.2: Wald tests by dropping terms sequentially to the model

Fixed term	Wald statistic	d.f	Wald/d.f.	chi pr
age	29.32	12	2.44	0.004
dt	0.02	1	0.02	0.876
prev	55.68	1	55.68	< 0.001
actipass	153.88	1	153.88	< 0.001
timemonth	0.15	1	0.15	0.703

Table 5.3: Wald tests by adding terms sequentially to the model

Fixed term	Wald statistic	d.f	Wald/d.f.	chi pr
timemonth	0.15	1	0.15	0.703
actipass	153.64	1	153.64	< 0.001
prev	28.99	1	28.99	< 0.001
dt	0.08	1	0.08	0.784
age	19.10	12	1.59	0.086

Table 5.4: Wald tests by dropping terms sequentially to the model

One can see that for both models the actipass and prev terms are significant. The order in which one fits the terms in Genstat is very important since different orders of the terms may cause the algorithm not to converge. The above model was found to be the most viable one amongst others. Different interaction effects were assessed for their significance in the model by adding

interaction terms one at a time into the model, however none of the interaction terms were found to be significant at the 5% level. It is appropriate now to fit a model only with the actipass and prev as the only terms in the model. Only the results reported by the Schall (1991) method are reported in Table 5.5 and 5.6 below since the Breslow and Clayton (1993) model gave similar results.

Fixed term	Wald statistic	d.f	Wald/d.f.	chi pr
prev	89.23	1	89.23	< 0.001
actipass	140.89	1	140.89	< 0.001

Table 5.5: Wald tests by adding terms sequentially to the model

Fixed term	Wald statistic	d.f	Wald/d.f.	chi pr
actipass	140.89	1	153.62	< 0.001
prev	88.42	1	88.42	< 0.001

Table 5.6: Wald tests by dropping terms sequentially to the model

5.4.1 Analysis and Application to the RSV data using Proc GLIMMIX in SAS

Random Effects Models

The optimal model that was fitted included the explanatory variables age, dt, prev, actipass and timemonth. The variable visit was excluded from this model because the model did not converge when this term was included in the optimal model. Different covariance structure models were investigated in Table 5.7: The results for the analysis of fixed effects for all the different covariance structure models are shown in Table 5.8:

Covariance Structure		Estimate	Standard Error
Unstructured	UN(1,1)	0.000	0.000
	Residual(VC)	1.0378	0.01524
Compound symmetry	Var(child)	0.000	0.000
	CS(child)	-2.11E-6	3.393E-6
	Residual(VC)	1.401	0.01524
Power	Var(child)	0.000	0.000
	SP(POW)(child)	0.000	0.000
	Residual(VC)	1.0378	0.01524
Spherical	Var(child)	0.000	0.000
	SP(SPH)(child)	0.000	0.000
	Residual(VC)	1.0378	0.01524
Gaussian	Var(child)	0.000	0.000
	SP(GAU)(child)	0.000	0.000
	Residual(VC)	1.0378	0.01524

Table 5.7: Covariance Parameter Estimates-random effects model

Once again we find the prev and actipass variables are significant at the 5% level and the age variable is tending towards significance. Furthermore there is a difference in the age 5 and age 12 groups with respect to whether a child is infected or not.

Effect	Estimate	Standard Error	Pr> $\|t\|$
Intercept	-5.0363	1.4750	< 0.0001
Age 0	-0.9197	1.2422	0.4591
Age 1	-0.6470	1.0937	0.5541
Age 2	-0.2760	1.0550	0.7936
Age 3	-0.06886	0.9942	0.9448
Age 4	-0.6690	0.9681	0.4895
Age 5	-2.6025	1.3222	0.0491
Age 6	-1.5960	1.0276	0.1204
Age 7	-2.2500	1.1754	0.0556
Age 8	-0.9989	0.6057	0.0992
Age 9	-0.7389	0.5222	0.1571
Age 10	-0.3221	0.4613	0.4851
Age 11	-0.5685	0.4699	0.2264
Age 12	0.0000	0.0000	0.0000
Dt	0.000851	0.008526	0.9205
Prev	44.5942	8.2574	< 0.0001
Actipass 0	2.2341	0.1803	< 0.0001
Actipass 1	0.000	0.000	0.000
Timemonth	-0.04538	0.1065	0.6701

Table 5.8: Parameter estimates and standard errors of the fixed effects-random effects model

Effect	F-Value	P-value
Age	1.62	0.0777
Dt	0.01	0.9205
Prev	29.17	< 0.0001
Actipass	153.61	< 0.0001
Timemonth	0.18	0.6701

Table 5.9: Type III Effects for random effects model

The estimates of the fixed effects for the random effects model with the compound symmetry covariance structure is slightly different from the above estimates:

Effect	Estimate	Standard Error	Pr> $\|t\|$
Intercept	-2.8180	1.4540	0.0526
age 0	-0.9175	1.2309	0.4561
age 1	-0.6314	1.0810	0.5591
age 2	-0.2620	1.0434	0.8017
age 3	-0.0596	0.9844	0.9517
age 4	-0.6572	0.9600	0.4936
age 5	-2.5925	1.3183	0.0493
age 6	-1.5922	1.0248	0.1203
age 7	-2.2360	1.1703	0.0561
age 8	-0.9913	0.6018	0.0995
age 9	-0.7301	0.5203	0.1606
age 10	-0.3231	0.4614	0.4838
age 11	-0.5656	0.4705	0.2293
age 12	0	.	.
dt	0.000855	0.008519	0.9201
prev	44.5376	8.2548	$< .0001$
actipass 0	-2.2344	0.1802	$< .0001$
actipass 1	0.000	0.000	0.000
timemonth	-0.04429	0.1053	0.6739

Table 5.10: Parameter estimates and standard errors of the fixed effects-random effects model using compound symmetry

Effect	F-Value	P-value
Age	1.62	0.0799
Dt	0.01	0.9201
Prev	29.11	< 0.0001
Actipass	153.81	< 0.0001
Timemonth	0.18	0.6739

Table 5.11: Type III Effects for random effects model-compound symmetry

Random Intercept Model

The above optimal model was also fitted as a random intercept model. Different covariance structure models were again investigated (see Table 5.12: It

Covariance Structure		Estimate	Standard Error
Unstructured	UN(1,1)	0.000	0.000
	Residual(VC)	1.0378	0.01524
Compound symmetry	Var(child)	No convergence	No convergence
	CS(child)	No convergence	No convergence
	Residual(VC)	No convergence	No convergence
Power	Var(child)	0.000	0.000
	SP(POW)(child)	0.000	0.000
	Residual(VC)	1.0378	0.01524
Spherical	Var(child)	0.000	0.000
	SP(SPH)(child)	0.000	0.000
	Residual(VC)	1.0378	0.01524
Gaussian	Var(child)	0.000	0.000
	SP(GAU)(child)	0.000	0.000
	Residual(VC)	1.0378	0.01524

Table 5.12: Covariance Parameter Estimates random intercept model

must be said the the CS model led to non-convergence as well as a different residual variance component, evident in Table 5.12 and Table 5.7. This is a purely computational artifact. The solution for the fixed effects for all the different covariance structure models are summarized in Tables 5.13 and 5.14 :

Effect	Estimate	Standard Error	Pr> $\|t\|$
Intercept	-5.0363	1.4750	< 0.0001
Age 0	-0.9197	1.2422	0.4591
Age 1	-0.6470	1.0937	0.5541
Age 2	-0.2760	1.0550	0.7936
Age 3	-0.06886	0.9942	0.9448
Age 4	-0.6690	0.9681	0.4895
Age 5	-2.6025	1.3222	0.0491
Age 6	-1.5960	1.0276	0.1204
Age 7	-2.2500	1.1754	0.0556
Age 8	-0.9989	0.6057	0.0992
Age 9	-0.7389	0.5222	0.1571
Age 10	-0.3221	0.4613	0.4851
Age 11	-0.5685	0.4699	0.2264
Age 12	0.0000	0.0000	0.0000
Dt	0.000851	0.008526	0.9205
Prev	44.5942	8.2574	< 0.0001
Actipass 0	2.2341	0.1803	< 0.0001
Actipass 1	0.000	0.000	0.000
Timemonth	-0.04538	0.1065	0.6701

Table 5.13: Parameter estimates and standard errors of the fixed effects-random intercept model

Effect	F-Value	P-value
Age	1.62	0.0777
Dt	0.01	0.9205
Prev	29.17	< 0.0001
Actipass	153.61	< 0.0001
Timemonth	0.18	0.6701

Table 5.14: Type III Effects for random intercept model

The results for the random intercept model are exactly the same as for the random effects model. Both models show that the 'prev' and 'actipass'

variables are significant at the 5% significance level as well as 'age' tending towards significance and the difference between the age 5 and age 12 groups are significant here as well with respect to the disease process.

5.4.2 Adaptive and Non-adaptive Gaussian Quadrature

The GLMM was also fitted using PROC NLMIXED. The sample program is used to achieve this is given below The fitted model was:

$rsvpos = \beta_{00} + \beta_0 age0 + \beta_1 age1 + beta_2 age2 + \beta_3 age3 + \beta_4 age4 + \beta_5 age5 + \beta_6 age6 + \beta_7 age7 + \beta_8 age8 + \beta_9 age9 + \beta_{10} age10 + \beta_{11} age11 + \beta_{13} dt + \beta_{14} prev + \beta_{15} actipass + \beta_{16} timemonth + childeffect(\tau)$.

The sample program used to specify the model in SAS is given below:

```
proc nlmixed data =lisa   qpoints=20 tech=nmsimp; parms
beta00=-5.06 beta0=-0.9 beta1=-0.65 beta2=-0.27 beta3=-0.067
beta4=-0.66 beta5=-2.5 beta6=-1.6 beta7=-2.2 beta8=-0.99
beta9=-0.74 beta10=-0.32 beta11=-0.55 beta13=-0.0009 beta14=45
tau2=1.02; teta=beta00+beta0*age0+beta1*age1+beta2*age2+beta3*age3
+beta4*age4+beta5*age5+beta6*age6
+beta7*age7+beta8*age8+beta9*age9+
beta10*age10+beta11*age11+beta13*dt+beta14*prev+beta15*actipass
+beta16*timemonth; expteta=exp(teta); p=expteta/(1+expteta); model
rsvpos~binary(p); random b~normal(0,tau2) subject=rsv; run;
```

The results for adaptive Gaussian quadrature and non-adaptive Gaussian quadrature are given in Tables 5.15 and 5.16:

Gaussian Quadrature			
Effect	$Q = 3$	$Q = 5$	$Q = 20$
Intercept	-5.04(1.488)	-5.72(1.584)	-5.466(1.384)
beta0	-0.92(1.244)	-0.94(1.862)	-0.91(1.35)
beta1	-0.65(1.082)	-0.68(1.985)	-0.69(1.857)
beta2	-0.28(1.081)	-0.27(1.056)	-0.28(1.045)
beta3	-0.07(1.001)	-0.08(0.991)	-0.07(0.984)
beta4	-0.67(0.981)	-0.69(0.967)	-0.67(0.991)
beta5	-2.71(1.381)	-2.12(1.354)	-2.74(1.311)
beta6	-1.61(1.021)	-1.59(1.044)	-1.60(1.058)
beta7	-2.23(1.184)	-2.13(1.192)	2.20(1.188)
beta8	0.99(0.624)	-1.05(0.652)	-1.01(0.652)
beta9	-0.76(0.521)	-0.74(0.601)	-0.75(0.504)
beta10	-0.33(0.456)	-0.42(0.498)	-0.32(0.416)
beta11	-.57(0.481)	-0.54(0.453)	-0.57(0.485)
beta13	-0.0008(0.009)	0.0009(0.007)	-0.0008(0.006)
beta14	47.2(7.995)	49.3(8.994)	46.1(8.774)
beta15	2.31(0.168)	2.33(0.183)	2.38(0.199)
beta16	-0.05(0.109)	-0.04(0.137)	-0.05(0.108)
τ	1.03(0.0114)	1.01(0.018)	1.03(0.013)
-2ℓ	2243.7	2242.2	2243.9

Table 5.15: Solution for the fixed effects-Gaussian quadrature

Adaptive Gaussian Quadrature			
Effect	$Q = 3$	$Q = 5$	$Q = 20$
Intercept	-5.02(1.433)	-5.71(1.498)	-5.786(1.354)
beta0	-0.92(1.244)	-0.94(1.862)	-0.91(1.145)
beta1	-0.65(1.012)	-0.68(1.385)	-0.68(1.017)
beta2	-0.28(1.051)	-0.23(1.034)	-0.26(1.026)
beta3	-0.07(0.991)	-0.08(0.988)	-0.07(0.999)
beta4	-0.66(0.989)	-0.69(0.997)	-0.67(0.982)
beta5	-2.61(1.341)	-2.12(1.369)	-2.72(1.311)
beta6	-1.61(1.071)	-1.59(1.022)	-1.63(1.758)
beta7	-2.24(1.191)	-2.13(1.189)	2.20(1.198)
beta8	0.99(0.674)	-1.02(0.652)	-1.01(0.699)
beta9	-0.74(0.501)	-0.74(0.561)	-0.75(0.540)
beta10	-0.33(0.459)	-0.42(0.498)	-0.32(0.446)
beta11	-0.57(0.498)	-0.54(0.422)	-0.57(0.494)
beta13	-0.0008(0.007)	0.0009(0.009)	-0.0008(0.006)
beta14	43.8(8.395)	49.3(8.994)	46.1(8.214)
beta15	2.22(0.178)	2.23(0.183)	2.41(0.189)
beta16	-0.05(0.109)	-0.03(0.158)	-0.05(0.104)
τ	1.03(0.0115)	1.01(0.018)	1.03(0.013)
-2ℓ	2243.8	2242.8	2243.8

Table 5.16: Solution for the fixed effects-adaptive Gaussian quadrature

In this particular case there seems not to be much differences between the adaptive and non-adaptive Gaussian quadrature estimates. The standard errors are also consistent with those from the GLIMMIX procedure. It must also be stated that there is a correct maximum to the likelihood for these models. If different quadrature methods lead to different answers, then at most one can be lading to the correct MLE. Different choices of quadrature starting values, and convergence criteria should be used until one is able to consistently obtain the same correct MLE's. Differences among the estimates merely indicate lack of, or inappropriate, convergence. Next a generalized linear mixed model with only the two variables, 'prev' and 'actipass' was

fitted since they were the only significant variables. The results are given below for different covariance structure models as shown in Tables 5.17, 5.18 and 5.19. The solution for the fixed effects for all the different covariance

Covariance Structure		Estimate	Standard Error
Unstructured	UN(1,1)	No convergence	No convergence
	Residual(VC)	No convergence	No convergence
Compound symmetry	Var(child)	7.187E-6	2.885E-6
	CS(child)	-6.71E-6	.
	Residual(VC)	0.8814	0.01306
Power	Var(child)	4.745E-7	2.885E-6
	SP(POW)(child)	0.000	0.000
	Residual(VC)	0.8814	0.01306
Spherical	Var(child)	4.745E-7	2.885E-6
	SP(SPH)(child)	0.000	0.000
	Residual(VC)	0.8814	0.01306
Gaussian	Var(child)	4.745E-7	2.885E-6
	SP(GAU)(child)	0.000	0.000
	Residual(VC)	0.8814	0.01306

Table 5.17: Covariance parameter estimates in a random effects model-prev and actipass

structure models are:

Effect	Estimate	Standard Error	Pr> $\|t\|$
Intercept	-5.9583	0.1796	< 0.0001
Prev	50.7801	4.8589	< 0.0001
Actipass 0	2.0576	0.1608	< 0.0001
Actipass 1	0.000	0.000	0.000

Table 5.18: Parameter estimates and standard errors of the fixed effects-using prev and actipass in a random effects model

Effect	F-Value	P-value
Prev	109.22	< 0.0001
Actipass	163.74	< 0.0001
Timemonth	0.18	0.6701

Table 5.19: Type III Effects for random effects model-prev and actipass

Fitting the above model as a random intercept model yielded the following results summarized in Tables 5.20, 5.21 and 5.22:: The solution for the fixed effects for all the different covariance structure models are:

Once again the random intercept model estimates are very similar to those of the random effects estimates. The above model was also fitted as a GLMM but using PROC NLMIXED. The fitted model was:

$$rsvpos = \beta_{00} + \beta_0 prev + \beta_1 actipass + childeffect(\tau)$$

The sample program is:

```
proc nlmixed data =lisa   qpoints=20 tech=nmsimp; parms
beta00=-5.9 beta0=50.7 beta1=2.03 tau2=0.85;
teta=beta00+b+beta0*prev+beta1*actipass; expteta=exp(teta);
p=expteta/(1+expteta); model rsvpos~binary(p); random
b~normal(0,tau2) subject=rsv; run;
```

The results for adaptive Gaussian quadrature and non-adaptive Gaussian quadrature are given below in Tables 5.23 and 5.24 :

The results here, not surprisingly are similar to those achieved by using Proc GLIMMIX. It is important to stress that each log-likelihood equals the

Covariance Structure		Estimate	Standard Error
Unstructured	UN(1,1)	0.2086	0.1638
	Residual(VC)	0.8782	0.01308
Compound symmetry	Var(child)	0.1053	0.1638
	CS(child)	0.1034	.
	Residual(VC)	0.8782	0.01308
Power	Var(child)	0.2086	0.1638
	SP(POW)(child)	0.000	0.000
	Residual(VC)	0.8782	0.01308
Spherical	Var(child)	0.2086	0.1638
	SP(SPH)(child)	1.000	0.000
	Residual(VC)	0.8782	0.01308
Gaussian	Var(child)	0.2086	0.1638
	SP(GAU)(child)	1.000	0.000
	Residual(VC)	0.8782	0.01308

Table 5.20: Covariance parameter estimates in a random intercept model-prev and actipass

Effect	Estimate	Standard Error	Pr> $\|t\|$
Intercept	-5.9564	0.1802	< 0.0001
Prev	50.6917	4.8703	< 0.0001
Actipass 0	2.0601	0.1615	< 0.0001
Actipass 1	0.000	0.000	0.000

Table 5.21: Parameter estimates and standard errors of the fixed effects-using prev and actipass in a random intercept model

Effect	F-Value	P-value
Prev	108.33	< 0.0001
Actipass	163.77	< 0.0001
Timemonth	0.18	0.6701

Table 5.22: Type III Effects for random intercept model-prev and actipass

maximum of the approximation to the model likelihood implying that log-likelihoods corresponding to different estimation procedures and/or different

Gaussian Quadrature			
Effect	$Q = 3$	$Q = 5$	$Q = 20$
Intercept	-3.9555(0.1805)	-3.6407(0.1764)	-3.9257(0.1790)
beta0	50.2189(5.7557)	52.7442(6.1598)	50.1865(5.703)
beta1	-1.9171(0.1686)	-2.1389(0.1853)	-1.9715(0.1685)
τ	0.8700(0.0114)	0.7300(0.018)	0.7800(0.0190)
-2ℓ	1494.7	1458.2	1494.6

Table 5.23: Solution for the fixed effects-gaussian quadrature using prev and actipass

Adaptive Gaussian Quadrature			
Effect	$Q = 3$	$Q = 5$	$Q = 20$
Intercept	-3.9515(0.1807)	-2.8854(0.1482)	-3.9243(0.1789)
beta0	50.5363(5.7469)	49.8533(6.1178)	50.1887(5.700)
beta1	-1.9558(0.1696)	-1.9917(0.1673)	-1.9717(0.1684)
τ	0.8700(0.012)	0.8590(0.0114)	0.8700(0.0119)
-2ℓ	1494.7	1420	1494.6

Table 5.24: Solution for the fixed effects-adaptive gaussian quadrature using prev and actipass

number of quadrature points are not necessarily comparable. This means that difference in log-likelihood values reflect the differences in the quality of the numerical approximations and thus higher log-likelihood values do not necessarily correspond to better approximations.

The random intercept and random effects models differed slightly in their covariance structure estimated but not in their parameter estimates, this can be expected. The adaptive and non-adaptive Gaussian results were similar to each other and in the current scientific setting the PROC GLIMMIX and PROC NLMIXED results are similar to each other. In the context of the RSV data, the 'prev' and 'actipass' variables are significant at the 5% level and are influential in contributing to a child's RSV status. The 'age' variable

was tending to significance at the 5% level. There are significant differences in the 'age 5 versus age 12'and 'age 7 versus age 12' groups at the 5% level.

5.5 Conclusion

The analysis in SAS and Genstat gave similar results with the only difference being in that SAS uses the last level of a factor as a baseline whilst Genstat uses the first level of a factor as a baseline and then performs comparisons within those levels in the output of the parameter estimates. The Proc GLIMMIX and Proc NLMIXED gave similar results although the Proc NLMIXED took much longer to converge in SAS. The random effects and random intercept models gave similar results as can be expected. The child effect accounted for very little variation in the random effects models. The random effects models were relevant to modelling the data set as it took into account the child effect albeit that this effect was not substantially accounting for much variation. In many ways, the random effects models are thorough in its approach to modelling any data set.

Chapter 6

The Conditional Model

6.1 Introduction

Conditional models or conditionally specified models are those in which any response within the sequence of repeated measures is modelled conditional upon (subsets of) the other outcomes. This could be the set of all past measurements or a subset thereof, in so-called transition models. Cox (1972) describes a conditional model as one where the parameters describe a feature (expectation, probability, odds, logit,...) of (a set of) responses, given values for the other responses. The best known example of this is undoubtedly the log-linear model. Rosner (1984) give a conditional logistic model. In this thesis the focus is mainly on transition type of conditional models dictated by the type of data studied.

6.2 The Cox Model

The log-linear model proposed by Cox (1972) is given as follows: Let Y_i denote a vector responses of dimension n from an individual or cluster i

whose components are $y_{ij}, j = 1, \ldots, n$. Then according to Cox (1972)

$$\begin{aligned} f(\boldsymbol{y}_i, \boldsymbol{\theta}_i) &= \exp\left(\sum_{j=1}^{n} \theta_{ij} y_{ij} + \sum_{j_1<j_2} \theta_{ij_1 j_2} y_{ij_1} y_{ij_2} + \ldots + \theta_{i1\ldots n} y_{i1} \ldots y_{in} - A(\boldsymbol{\theta}_i)\right) \\ &= \boldsymbol{c}(\theta_i) \exp\left(\sum_{j=1}^{n} \theta_{ij} y_{ij} + \sum_{j_1<j_2} \theta_{ij_1 j_2} y_{ij_1} y_{ij_2} + \ldots + \theta_{i1\ldots n} y_{i1} \ldots y_{in}\right) \end{aligned} \tag{6.1}$$

where $\boldsymbol{A}(\theta_i)$ and $\boldsymbol{c}(\theta_i)$ represent the same normalizing constant, written in additive and multiplicative way respectively. $\boldsymbol{\theta}_i$ is the canonical parameter, consisting of first, and second, up to nth order components.
The interpretation of the parameters are a conditional one

$$\theta_{ij} = \ln\left(\frac{\Pr(Y_{ij} = 1 | Y_{ik} = 0; k \neq j)}{\Pr(Y_{ij} = 0 | Y_{ik} = 0; k \neq j)}\right)$$

Here the first order parameters (main effects) are interpreted as the conditional logits.
Similarly the second order parameter θ_{ij} is defined as

$$\theta_{ij} = \ln\left(\frac{Pr(Y_{ij} = 1, Y_{ik} = 1 | Y_{il} = 0; k, j \neq l)(Pr(Y_{ij} = 0, Y_{ik} = 0 | Y_{il} = 0; k, j \neq l)}{Pr(Y_{ij} = 1, Y_{ik} = 0 | Y_{il} = 0; k, j \neq l)(Pr(Y_{ij} = 0, Y_{ik} = 1 | Y_{il} = 0; k, j \neq l)}\right)$$

These are the conditional log odds ratios. Due to this conditional interpretation, these models are less useful for regression. The dependence of $E(Y_{ij})$ on covariates involves all parameters not only the main effects. The interpretation of the parameters depends on the length n_i of a sequence. Shorter sequences simply imply that one conditions on less outcomes, changing interpretations with the length of a sequence. The advantages of this model is that the parameter vector is not constrained since all the values of $\boldsymbol{\theta} \in \mathbb{R}$ yield positive probabilities and secondly calculation of joint probabilities is fairly straightforward because by evaluating and summing the density over all possible sequences of $\boldsymbol{y}$ will yield $\boldsymbol{c}(\boldsymbol{\theta})^{-1}$.

Due to the popularity of marginal and random-effects models, conditional models have not received widespread attention. Diggle et al (2002, pp. 142-144) criticized the conditional models because the interpretation of a fixed effect parameter, for example, the evolution or treatment effect, of one response is conditional on the responses of other responses for the same subject, outcomes of other subjects and the number of repeated measures. Not only may such parameters make answering the substantive question difficult, they are ill founded when the number of measurements per subject is not constant. On the other hand, conditional models have received a portion of popularity due to their mathematical convenience such as the log-linear model considered above. The advantages of such a conditional model have already been highlighted. Agresti (2002) also highlights these advantages in a classical log-linear model. Molenberghs and Ryan (1999) and Aerts et al.(2002) discuss the problem at great length and detail, in the context of exchangeable binary data, the advantages of conditional models and show with great care how the disadvantages can be overcome for their setting. They constructed the joint distribution for clustered multivariate binary outcomes, based on the multivariate exponential family model. Fitzmaurice, Laird and Tosteson (1996) take a slightly different approach, but based on the exponential model. This approach is a likelihood based one and has great efficiency over other procedures such as GEE's.

6.3 Transition Models

A very specific class of conditional models are the so-called transition models. In a transition model, a measurement $\boldsymbol{Y}_{ij}$ in a longitudinal sequence is described as a function of previous outcomes, or history $\boldsymbol{h_{ij}} = (Y_{i1}, \ldots, Y_{ij-1})$.

One can then write a regression equation for the outcome $\boldsymbol{Y}_{ij}$ in terms of $\boldsymbol{h_{ij}} = (Y_{i1}, \ldots, Y_{ij-1})$ or alternatively the error term ε_{ij} can be written in terms of previous error terms. Feller (1968) also states that specific classes of transition models fall in the class of Markov models. The order of the transition model is the number of previous measurements that is still considered to influence the current one. A model is called stationary if the functional form of the dependence is independent of the actual time at which it occurs. An example of a stationary first-order autoregressive model for **continuous** data is described by the following equation

$$Y_{i1} = \boldsymbol{x}_{\boldsymbol{i1}}^{'}\boldsymbol{\beta} + \varepsilon_{i1} \tag{6.2}$$

$$Y_{ij} = \boldsymbol{x}_{\boldsymbol{ij}}^{'}\boldsymbol{\beta} + \alpha Y_{ij-1} + \varepsilon_{ij}, \quad j = 2, \ldots, n_i \tag{6.3}$$

Assuming $\varepsilon_{i1} \sim N(0, \sigma^2)$ and $\varepsilon_{ij} \sim N(0, \sigma^2(1-\alpha^2))$ for $j > i$ recursively yields $\text{var}(Y_{ij}) = \sigma^2$ and $\text{cov}(Y_{ij}, Y_{ij'}) = \alpha^{|j-j'|}\sigma^2$. This model produces a marginal multivariate normal model with AR(1) variance covariance matrix and is useful for equally spaced outcomes. If we include random effects in Eqns.(6.2) and (6.3) and if we vary the assumptions regarding the autoregressive structures, then we see that the general linear mixed effects model formulation with serial correlation encompasses wide classes of transition models. Diggle et al. (2002, pp.190-207) give a detailed intensive formulation of Markov models for binary, categorical and Poisson data.

6.4 Transition Models for outcomes of a general type

The generalized linear model ideas can be used to formulate transition models for outcomes of a more general type. If we decompose an outcome as $Y_{ij} = \mu_{ij}^c + \varepsilon_{ij}$, then the first and second moment of a GLM can be written in terms of the history, $\boldsymbol{h}_{ij}$ as:

$$\mu_{ij}^c = E(Y_{ij}|\boldsymbol{h_{ij}}) \tag{6.4}$$

$$\phi\upsilon^c(\mu_{ij}^c) = \text{var}(Y_{ij}|\boldsymbol{h_{ij}}). \tag{6.5}$$

We note that $\upsilon^c(\mu_{ij}^c)$ is a function of the variance by writing it in terms of the mean and ϕ is the overdispersion parameter. The only difference from the conventional GLM is that by including the history, $\boldsymbol{h_{ij}}$, an outcome is described in terms of its predecessors. A function of the mean components is now equated to a linear function of the predictors to get:

$$\eta_{ij}(\mu_{ij}^c) = \boldsymbol{x_{ij}^{'}}\boldsymbol{\beta} + \kappa(\boldsymbol{h_{ij}}, \boldsymbol{\beta}, \boldsymbol{\alpha}) \tag{6.6}$$

where κ is a function, often a linear one, of the history. Now the contribution for Y_{ij} given the history $\boldsymbol{h_{ij}}$, lead to independent GLM contributions due to the law of total probability:

$$f(y_{i1}, \ldots, y_{in_i}) = f(y_{i1}).f(y_{i2}|y_{i1}).f(y_{i3}|y_{i1}, y_{i2}).f(y_{in_i}|y_{i1}, \ldots, y_{in_i-1})$$

which can be written as:

$$f(y_{i1}, \ldots, y_{in_i}) = f(y_{i1}).\prod_{j=2}^{n_i} f(y_{ij}|\boldsymbol{h_{ij}}) \tag{6.7}$$

$$= f(y_{i1}, \ldots, y_{iq}).\prod_{j=q+1}^{n_i} f(y_{ij}|\boldsymbol{h_{ij}}) \tag{6.8}$$

The latter decomposition in Eq. (6.8) is relevant when the $\boldsymbol{h_{ij}}$ contains the q immediately preceding measurements. Eq. (6.8) yields $n_i - q$ independent

univariate GLM contributions. This may now mean that a separate model may be needed for the first q measurements, since these are left undescribed by the conditional GLM. Molenberghs and Verbeke (2005, p.237) give the following example of a logistic type of regression model,

$$\text{logit}[P(Y_{ij} = 1|\boldsymbol{x}_{ij}, Y_{i,j-1} = y_{i,j-1}, \boldsymbol{\beta}, \alpha)] = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \alpha y_{i,j-1}. \tag{6.9}$$

By evaluating Eq.(6.9) at $y_{i,j-1} = 0$ and $y_{i,j-1} = 1$, respectively produces the transition probabilities between occasions $j-1$ and j. If there are no covariates in this model then, these transition probabilities would be constant across the population. When there are time independent covariates only, these transition probabilities change in a straightforward way with the level of covariate.

6.5 A Transition Model for the RSV data

Diggle et al (2002) state that transition models are considered as extensions of generalized linear models (GLMs) for describing the conditional distribution of each response y_{ij} as an explicit function of past responses $y_{ij-1}, \ldots, y_{i1}$ and covariates x_{ij}. Hence the past outcomes are treated as predictor variables.
If we consider the generalized linear transition model with respect to the Kilifi data set, we can model the conditional distribution of Y_{ij} given the past as an explicit function of the q preceding responses. We can assume that the probability of RSV for child i at visit j has a direct dependence on whether or not the child had RSV at visit $j-1$ as well as on explanantory variables, x_{ij}. This is the first case of a first order transition model. If we take the logit link then a first order transition model is given by
$\text{logit}[P(Y_{ij} = 1|Y_{ij-1}, \ldots, Y_{i1})] = \boldsymbol{x}_{\boldsymbol{ij}}^{'}\boldsymbol{\beta} + \alpha Y_{ij-1}$.

Therefore the probability of RSV at time t_{ij} depends on the measured covariates or explanatory variables but also on whether or not the child had RSV at the previous visit. The parameter $\exp(\alpha)$ is the ratio of the odds of infection among the children who did and did not have RSV at the prior visit. The $\boldsymbol{\beta}$ coefficient is the change per unit change in x in the log odds of infection among children who were free of RSV at the previous visit. The transition model stated above is a first order Markov chain according to Feller (1968, vol 1, p. 132). At equally spaced time intervals the 2×2 transition matrix whose elements are $P(Y_{ij} = y_{ij}|Y_{ij-1} = y_{ij-1})$ where each of Y_{ij} and Y_{ij-1} may take values of 0 and 1 is given by inverting the logistic regression equation for every pair (y_{ij}, y_{ij-1}) as

		Y_{ij}	
		0	1
Y_{ij-1}	0	$\frac{1}{1+\exp(x_{ij}^{'}\beta}$	$\frac{\exp(x_{ij}^{'}\beta)}{1+\exp(x_{ij}^{'}\beta)}$
	1	$\frac{1}{1+\exp(x_{ij}^{'}\beta+\alpha)}$	$\frac{\exp(x_{ij}^{'}\beta+\alpha)}{1+\exp(x_{ij}^{'}\beta+\alpha)}$

However in the general transition model we let $H_{ij} = \{Y_{i1}, \ldots, Y_{ij-1}\}$ represent the past responses for the i-th subject, $\mu_{ij}^{c} = E(Y_{ij}|H_{ij})$ and let $v_{ij}^{c} = var(Y_{ij}|H_{ij})$ be the conditional mean and variance of Y_{ij} given past responses and the explanatory variables. We can specify the model analogous to the GLM for independent data, where we assume:

$$g(\mu_{ij}^{c}) = \boldsymbol{x_{ij}^{'}\beta} + \sum_{\boldsymbol{r=1}}^{\boldsymbol{s}} \boldsymbol{f_r(H_{ij};\alpha)} = \boldsymbol{x_{ij}^{'}\beta} + \boldsymbol{h_{ij}^{'}\alpha} \tag{6.10}$$

and

$$v_{ij}^{c} = v(\mu_{ij}^{'})\phi$$

We model the transition from the prior state by the functions f_r to the present response. The past outcomes after transformation by the known

functions f_r are treated as explanatory variables. Interactions among the prior responses may be considered. We can then fit the transition model using GLM techniques and treat the repeated transitions for a child/subject as independent events.

General

Diggle et al (2002) focus on the case where the observation times t_{ij} are equally spaced. The history for subject i at visit j is denoted as $H_{ij} = \{y_{ik}, k = 1, \ldots, j-1\}$. The most useful transition models are Markov chain for which the conditional distribution of Y_{ij} given H_{ij} depends only on the q prior responses $Y_{ij-1}, \ldots, Y_{ij-q}$. The integer q represents the order of the model. Writing the conditional p.d.f of Y_{ij} as an exponential family type of distribution gives

$$f(y_{ij}|H_{ij}) = \exp\{[y_{ij}\theta_{ij} - \psi(\theta_{ij})]/\phi + c(y_{ij}, \phi)\} \tag{6.11}$$

for known functions $\psi(\theta_{ij})$ and $c(y_{ij}, \phi)$. The conditional mean and variance:
$\mu_{ij}^c = E(Y_{ij}|H_{ij}) = \psi^{'}(\theta_{ij})$ and
$v_{ij}^{'} = var(Y_{ij}|H_{ij}) = \psi^{''}(\theta_{ij})\phi$
Diggle et al. (2002) consider models where the conditional mean and variance satisfy the equations

$$g(\mu_{ij}^c) = \boldsymbol{x}_{\boldsymbol{ij}}^{'}\boldsymbol{\beta} + \sum_{r=1}^{s} f_r(H_{ij}; \alpha)$$

for suitable functions f_r and

$$v_{ij}^c = v(\mu_{ij}^{'})\phi$$

where h and v are known link and variance functions determined from the

density function. Hence the transition model expresses the conditional mean as a function of both covariates x_{ij} and of the past responses $Y_{ij-1}, \ldots, Y_{ij-q}$ in a much more general setting. We assume that the past affects the present through the sum of s terms each of which may depend on the q prior values. As an example: A logistic regression model for binary responses assuming a first order Markov chain (Cox, 1970, Korn and Whittemore, 1979, Zeger et al., 1985) specified as:

$$g(\mu_{ij}^c) = \boldsymbol{x_{ij}'}\boldsymbol{\beta} + \alpha y_{ij-1} \tag{6.12}$$

where $g(\mu_{ij}^c) = \text{logit}(\mu_{ij}^c), v(\mu_{ij}^c) = \mu_{ij}^c(1-\mu_{ij}^c),\ f_r(H_{ij}, \alpha) = \alpha_r y_{ij-r}$,
$s = q = 1$ and $\mu_{ij}^c = \text{Prob}(Y_{ij} = 1|H_{ij})$.

A first order Markov model can be fitted by making use of the likelihood function. The contribution to the likelihood for the i^{th} subject can be written as:

$L_i(y_{i1}, \ldots, y_{in_i} = f(y_{i1}) \prod_{j=2}^{n_i} f(y_{ij}|H_{ij})$ where H_{ij} is the history measurement at occasion j given by $H_{ij} = \{y_{ij-1}\}$
In a Markov model of order q, the conditional distribution of Y_{ij} is

$f(y_{ij}|H_{ij}) = f(y_{ij}|y_{ij-1}, \ldots, y_{ij-q})$
so that the likelihood is:
$f(y_{i1}, \ldots, y_{iq}) \prod_{j=q+1}^{n_i} f(y_{ij}|y_{ij-1}, \ldots, y_{ij-q})$
The GLM in Eq.(6.8) specifies only the conditional distribution $f(y_{ij}|H_{ij})$ whilst the likelihood of the first q observations $f(y_{i1}, \ldots, y_{iq})$ is not specified directly. In the logistic and log-linear models $f(y_{i1}, \ldots, y_{iq})$ is not determined from the GLM assumption about the conditional model and the full likelihood is unavailable. An alternative is to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by maximizing

the conditional likelihood given by:

$\prod_{i=1}^{N} f(y_{iq+1},\ldots,y_{in_i}|y_{i1},\ldots,y_{iq}) = \prod_{i=1}^{N}\prod_{j=q+1}^{n_i} f(y_{ij}|H_{ij})$ where N is the number of subjects or clusters in the study.

There are 2 distinct cases to consider in the maximization process of the likelihood

CASE 1

$f_r(H_{ij},\boldsymbol{\alpha},\boldsymbol{\beta})=\alpha_r f_r(H_{ij})$

so that

$g(\mu_{ij}^c)=\boldsymbol{x_{ij}'}\boldsymbol{\beta}+\sum_{r=1}^{s}\alpha_r f_r(H_{ij})$

Clearly $g(\mu_{ij}^c)$ is a linear function of both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_s)'$ so that estimation is the same as for GLMs for independent data. We regress Y_{ij} on the $(p+s)$ dimensional vector of extended explanatory variables $(\boldsymbol{x_{ij}'}, f_1(H_{ij}),\ldots,f_s(H_{ij}))'$.

CASE 2

Case 2 occurs when functions of past responses include both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Examples are linear and log-linear models. The Iterative weighted least squares method is used to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. This exposition is given in Diggle et al (2002, pg. 193-194). As a summary; the derivative of the log conditional likelihood or conditional score function has the form

$$S'(\boldsymbol{\delta}) = \sum_{i=1}^{m}\sum_{j=q+1}^{n_i}\frac{\partial\mu_{ij}^c}{\partial\boldsymbol{\delta}}v_{ij}^{c^{-1}}(y_{ij}-\mu_{ij}^c)=0 \qquad (6.13)$$

where $\boldsymbol{\delta} = (\boldsymbol{\beta},\boldsymbol{\alpha})$. The above equation is analogous to the GLM score equation. The derivative $\frac{\partial\mu_{ij}^c}{\partial\boldsymbol{\delta}}$ is analogous to $\boldsymbol{x}_{ij}$ but it can depend on both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The iterative weighted least squares procedure is formulated as

follows. Let $\boldsymbol{Y_i}$ be the $(n_i - q)$ vector of responses for $j = q+1, \ldots, n_i$ and μ_{ij}^c its expectation given by H_{ij}.

Let $\boldsymbol{X_i^*}$ be an $(n_i - q)$x $(p+s)$ matrix with the k^{th} row given by

$\frac{\partial \mu_{iq+k}}{\partial \boldsymbol{\delta}}$ and $\boldsymbol{W_i} = \text{diag}\left(\frac{1}{v'_{ik+q}}\right)$, $k = 1, \ldots, n_{i-q}$ be an $(n_i - q)$x $(n_i - q)$ diagonal weighting matrix.

Finally let $\boldsymbol{Z_i}$=$\boldsymbol{X_i^*}\hat{\boldsymbol{\delta}}$+ $(\boldsymbol{Y_i} - \hat{\boldsymbol{\mu}}_{\boldsymbol{i}}^{\boldsymbol{c}})$ then an updated $\hat{\boldsymbol{\delta}}$ can be obtained by iteratively regressing $\boldsymbol{Z}$ on $\boldsymbol{X^*}$using weights in $\boldsymbol{W}$. When the correct model is assumed for the conditional mean and variance, the solution $\hat{\boldsymbol{\delta}}$ of Eq.(6.13) asymptotically follows a Gaussian distribution, as N goes to infinity, with mean equal to the true value $\boldsymbol{\delta}$ and $(p+s) \times (p+s)$ variance matrix:

$$\boldsymbol{V_\delta} = \left(\sum_{i=1}^{N} X_i'^* W_i X_i^*\right)^{-1}$$

The variance $\boldsymbol{V_\delta}$ depends on both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and a consistent estimate $\hat{\boldsymbol{V}}_{\boldsymbol{\delta}}$ is obtained by replacing $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by their estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$. However when the conditional mean is correctly specified and variance is not, consistent inferences about $\boldsymbol{\delta}$ can be still obtained using the robust variance:

$$V_R = \left(\sum_{i=1}^{m} X_i'^* W_i X_i^*\right)^{-1} \left(\sum_{i=1}^{m} X_i'^* W_i V_i W_i X_i^*\right) \left(\sum_{i=1}^{m} X_i'^* W_i X_i^*\right)^{-1}.$$

A consistent estimate of V_R can be obtained by replacing $V_i = \text{var}(\boldsymbol{Y_{ij}}|H_i)$ by its estimate $(\boldsymbol{Y_i} - \widehat{\boldsymbol{\mu_i}^c})(\boldsymbol{Y_i} - \widehat{\boldsymbol{\mu_i}^c})'$. Interestingly, even when the Markov assumtption is violated, the robust variance will give more consistent confidence intervals for $\hat{\boldsymbol{\delta}}$. This concludes the estimation process for the transition model.

6.6 Software for fitting Conditional Models in SAS

The standard GLM software in SAS can be used to fit the transition models because subsequent measurements and their past history are independent of each other. The SAS procedures such as PROC GENMOD and PROC LOGISTIC can be used to fit these models. However, one must ensure that the previous measurement(s) can be used as a covariate. The longitudinal data set then needs to be rearranged one record per measurement rather than per subject using the DROPOUT macro. Depending on how many history variables needed to be included, the DROPOUT macro needs to be called as many times. If we use the two most recent measurements, then the macro needs to be called twice. The LAG statement can also be used to prepare a data set in SAS.

6.7 Fitting Conditional Models in SAS to the RSV data

In this section results from fitting a series of transition models (first, second and third order history models) to the RSV data are presented. The models were fitted in SAS using 'Proc Genmod' and 'Proc Logistic' with a log link and a binomial distribution. The results of the analyses are presented in Table 6.1. First the previous responses Y_{ij-1}, Y_{ij-2} and Y_{ij-3} were included independently into the model. Then the models with (Y_{ij-1}, Y_{ij-2}) and $(Y_{ij-1}, Y_{ij-2}, Y_{ij-3})$ were fitted.

Effect	DF	Wald Chi-Square	P-value
age	12	14.529	0.268
dt	1	0.409	0.523
prev	1	32.996	$< .0001$
actipass	1	125.754	$< .0001$
time	1	0.592	0.442
Y_{ij-1}	1	9.809	0.0017
age	12	15.381	0.221
dt	1	0.701	0.403
prev	1	34.672	$< .0001$
actipass	1	128.873	$< .0001$
time	1	0.658	0.417
Y_{ij-2}	1	1.342	0.247
age	12	15.197	0.231
dt	1	0.743	0.389
prev	1	34.644	$< .0001$
actipass	1	128.430	$< .0001$
time	1	0.671	0.413
Y_{ij-3}	1	0.477	0.490
age	12	14.571	0.266
dt	1	0.400	0.527
prev	1	33.104	$< .0001$
actipass	1	126.219	$< .0001$
time	1	0.579	0.447
Y_{ij-1}	1	9.383	0.0022
Y_{ij-2}	1	0.959	0.327
age	12	14.407	0.276
dt	1	0.407	0.523
prev	1	33.134	$< .0001$
actipass	1	126.382	$< .0001$
time	1	0.571	0.450
Y_{ij-1}	1	9.486	0.0021
Y_{ij-2}	1	0.895	0.344
Y_{ij-3}	1	0.490	0.484

Table 6.1: Type III Effects for first, second, third order, first and second and full model

The type III statistics show that 'prev' and 'actipass' variables are significant all three models and the full model that includes all three history terms at the 5% level. However from the history terms, the first order model shows that Y_{ij-1} is significant at the 5% level, while the model with Y_{ij-2} and

Y_{ij-3} included independently show that they are not significant at the 5% level. Thus we conclude that only the immediate past history is important in explaining the current disease status. The Wald Chi-square values do not differ numerically by vast amounts in all three models for all the variables. The model fit statistics show that the first order model has the lowest fit

Criterion	First order	Second order	Third order	$1^{st}, 2^{nd}$	$1^{st}, 2^{nd}$ & 3^{rd}
AIC	1063.176	1069.903	1070.591	1064.355	1065.918
SC	1189.102	1195.829	1196.517	1197.277	1205.836
-2log-likelihood	1027.176	1033.903	1034.591	1026.355	1025.918

Table 6.2: Model fit statistics for first, second and third, first and second and the full order models

statistics for the AIC and SC criterion whilst the full model has the lowest fit statistics for the -2log-likelihood. The significant parameter estimates are confirmed by the type III score statistics in Table 6.1. The standard error for 'age 0' are extremely inflated in comparison to the other age group standard errors (Table 6.7). The model including the first order term has the smallest standard error estimates in comparison to the other models. However the results show that 'age' was not significant in the model.

From Table 6.5, with respect to the first order model, the rsvhistory 1 vs 2 (Y_{ij-1}) is the $\frac{P_{2|1}/(1-P_{2|1})}{P_{2|2}/(1-P_{2|2})} = 0.289$ implying that the odds of a child becoming infected if his/her prior state was uninfected is about 0.3 times more than a child whose prior state is infected and remains infected. This odds ratio is small implying that the denominator is larger than the numerator. The reverse of this is $\frac{P_{2|2}/(1-P_{2|2})}{P_{2|1}/(1-P_{2|1})} = 3.46$ implying that the odds of a child becoming infected if his/her prior state was infected is about 3.5 times more than a child whose prior state is uninfected and becoming infected. This confirms the fact that RSV is indeed a rare occurrence disease. Considering the model with

only Y_{ij-2} included as an explanatory variable, the odds ratio of the disease for those with $Y_{ij-2} = 1$ against $Y_{ij-2} = 2$ is $\frac{P_{2|1}/(1-P_{2|1})}{P_{2|2}/(1-P_{2|2})} = 0.493$ implying that the probability of a child becoming infected if his/her prior two step state was uninfected is about 0.5 times more than a child whose prior state is infected and remains infected. The reverse of this is $\frac{P_{2|2}/(1-P_{2|2})}{P_{2|1}/(1-P_{2|1})} = 2.03$ implying that the probability of a child becoming infected if his/her prior two step state was infected is about 2 times more than a child whose prior two step state is uninfected and becoming infected. For a model with only Y_{ij-3} included as a predictor the odds ratio is the $\frac{P_{2|1}/(1-P_{2|1})}{P_{2|2}/(1-P_{2|2})} = 0.655$ implying that the odds of a child becoming infected if his/her prior three step state was uninfected is about 0.66 times more than a child whose prior three step state is infected and remains infected. The reverse of this is $\frac{P_{2|2}/(1-P_{2|2})}{P_{2|1}/(1-P_{2|1})} = 1.53$ implying that the odds of a child becoming infected if his/her prior three step state was infected is about 1.5 times more than a child whose prior three step state is uninfected and becoming infected. The odds ratios are also decreasing as the order of history dependence increases. The patterns of the odds ratios decreasing is seen across all the variables: 'age', 'dt', 'prev', 'actipass' and 'timemonth' in all three models.

In the first order model the odds ratio for comparing 'age group 1 and age group 12' is 0.408 implying that implying that the odds of a child becoming infected if his/her prior state was uninfected is about 0.4 times more than a child whose prior state is infected and remains infected but when we compare 'age group 10 and age group 12', we find the odds ratio is 0.725 implying that the probability of a child becoming infected if his/her prior state was uninfected is about 0.73 times more than a child whose prior state is infected and remains infected. In other words, this comparative probability of transmitting between the infected and uninfected states increases with age. It is

also important to state that the Y_{ij-2} and Y_{ij-3} terms were not significant at the 5% level implying that Y_{ij-1} is more informative about the current states than the states before time visits t_{ij-1}.

6.8 Conclusion

This chapter has investigated the problem of including the history of outcomes as predictor variables in a model for the analysis of repeated measurement non-Gaussian data in addition to other covariates. In particular transition models within a broader class of conditional models were applied to RSV disease data for children within the age of one year. The outcomes were binary responses denoting the infection status of a child (0 =uninfected, 1 =infected) at any measurement and sampling occasion. Three types of transition models namely first, second and third order history models were investigated in addition to assessing the significance of other covariates. The analysis reveal that the first order history model gave a better fit compared to the other two as well the inclusion of models with first and second and first, second and third order history terms. The results imply that the immediate past history is important in explaining the current status of a child's infection state. The further back the status history of a child the less relevant it is in explaining the current disease status of a child. Other predictor variables that were found to be significant were the prevalence (antibody level) in the blood, the type of sampling method (actively or passively sampled) and age of a child in months.

Parameter	DF	Estimate	Standard Error	Wald Chi-square	P-value
Intercept	1	-4.500	28.754	0.025	0.876
age 0	1	-7.271	344.800	0.000	0.983
age 1	1	0.692	28.744	0.001	0.981
age 2	1	1.060	28.742	0.001	0.971
age 3	1	1.197	28.741	0.002	0.967
age 4	1	0.611	28.741	0.001	0.983
age 5	1	-1.075	28.753	0.001	0.970
age 6	1	0.054	28.745	0.000	0.999
age 7	1	-0.581	28.752	0.000	0.984
age 8	1	0.642	28.742	0.001	0.982
age 9	1	0.803	28.743	0.001	0.978
age 10	1	1.267	28.743	0.002	0.965
age 11	1	1.013	28.743	0.001	0.972
dt	1	-0.006	0.010	0.409	0.523
prev	1	50.377	8.770	32.996	$< .0001$
actpass 0	1	1.134	0.101	125.754	$< .0001$
timemonth	1	-0.090	0.117	0.592	0.442
Y_{ij-1}	1	-0.620	0.198	9.809	0.002
Intercept	1	-4.763	40.200	0.014	0.906
age 0	1	-7.938	482.200	0.000	0.987
age 1	1	0.667	40.192	0.000	0.987
age 2	1	1.098	40.191	0.001	0.978
age 3	1	1.255	40.190	0.001	0.975
age 4	1	0.671	40.190	0.000	0.987
age 5	1	-1.046	40.199	0.001	0.979
age 6	1	0.097	40.193	0.000	0.998
age 7	1	-0.532	40.198	0.000	0.989
age 8	1	0.684	40.191	0.000	0.986
age 9	1	0.911	40.191	0.001	0.982
age 10	1	1.323	40.191	0.001	0.974
age 11	1	1.086	40.192	0.001	0.978
dt	1	-0.009	0.010	0.701	0.403
prev	1	51.480	8.743	34.672	$< .0001$
actpass 0	1	1.149	0.101	128.873	$< .0001$
timemonth	1	-0.094	0.116	0.658	0.417
Y_{ij-2}	1	-0.354	0.305	1.342	0.247
Intercept	1	-4.887	39.829	0.015	0.902
age 0	1	-7.937	477.800	0.000	0.987
age 1	1	0.696	39.820	0.000	0.986
age 2	1	1.091	39.819	0.001	0.978
age 3	1	1.246	39.818	0.001	0.975
age 4	1	0.659	39.818	0.000	0.987
age 5	1	-1.055	39.827	0.001	0.979
age 6	1	0.089	39.822	0.000	0.998
age 7	1	-0.529	39.827	0.000	0.989
age 8	1	0.683	39.819	0.000	0.986
age 9	1	0.924	39.820	0.001	0.982
age 10	1	1.325	39.820	0.001	0.974
age 11	1	1.089	39.820	0.001	0.978
dt	1	-0.009	0.010	0.743	0.389
prev	1	51.349	8.724	34.644	$< .0001$
actpass 0	1	1.143	0.101	128.430	$< .0001$
timemonth	1	-0.095	0.116	0.671	0.413
Y_{ij-3}	1	-0.211	0.306	0.477	0.490

Table 6.3: Parameter estimates for first, second and third order models

Parameter	DF	Estimate	Standard Error	Wald Chi-square	P-value
Intercept	1	-4.235	28.399	0.022	0.882
age 0	1	-7.228	340.600	0.001	0.983
age 1	1	0.666	28.387	0.001	0.981
age 2	1	1.058	28.385	0.001	0.970
age 3	1	1.199	28.384	0.002	0.966
age 4	1	0.614	28.384	0.001	0.983
age 5	1	-1.067	28.397	0.001	0.970
age 6	1	0.061	28.389	0.000	0.998
age 7	1	-0.581	28.396	0.000	0.984
age 8	1	0.641	28.386	0.001	0.982
age 9	1	0.790	28.386	0.001	0.978
age 10	1	1.264	28.386	0.002	0.965
age 11	1	1.001	28.387	0.001	0.972
dt	1	-0.006	0.010	0.400	0.527
prev	1	50.609	8.796	33.104	$< .0001$
actpass 0	1	1.141	0.102	126.219	$< .0001$
time	1	-0.089	0.117	0.579	0.447
Y_{ij-1}	1	-0.607	0.198	9.383	0.002
Y_{ij-2}	1	-0.300	0.306	0.959	0.327
Intercept	1	-4.037	28.070	0.021	0.886
age 0	1	-7.196	336.600	0.001	0.983
age 1	1	0.669	28.056	0.001	0.981
age 2	1	1.053	28.055	0.001	0.970
age 3	1	1.194	28.054	0.002	0.966
age 4	1	0.605	28.054	0.001	0.983
age 5	1	-1.066	28.066	0.001	0.970
age 6	1	0.064	28.058	0.000	0.998
age 7	1	-0.576	28.065	0.000	0.984
age 8	1	0.640	28.055	0.001	0.982
age 9	1	0.791	28.055	0.001	0.978
age 10	1	1.264	28.055	0.002	0.964
age 11	1	0.991	28.056	0.001	0.972
dt	1	-0.006	0.010	0.407	0.523
prev	1	50.679	8.804	33.134	$< .0001$
actpass 0	1	1.142	0.102	126.382	$< .0001$
time	1	-0.088	0.117	0.570	0.450
Y_{ij-1}	1	-0.610	0.198	9.486	0.002
Y_{ij-2}	1	-0.290	0.306	0.895	0.344
Y_{ij-3}	1	-0.214	0.306	0.490	0.484

Table 6.4: Parameter estimates for the model including first and second order terms and the full model including first, second and third order terms

		First Order	Second Order	Third Order	$1^{st}, 2^{nd}$	$1^{st}, 2^{nd}, 3^{rd}$
Effect	Comparison	Estimate	Estimate	Estimate	Estimate	Estimate
age	0 vs 12	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
age	1 vs 12	0.408	0.348	0.359	0.450	0.408
age	2 vs 12	0.589	0.535	0.533	0.592	0.599
age	3 vs 12	0.675	0.625	0.622	0.682	0.690
age	4 vs 12	0.376	0.349	0.346	0.380	0.383
age	5 vs 12	0.07	0.063	0.062	0.071	0.072
age	6 vs 12	0.215	0.197	0.196	0.219	0.223
age	7 vs 12	0.114	0.105	0.105	0.115	0.117
age	8 vs 12	0.388	0.354	0.354	0.391	0.396
age	9 vs 12	0.456	0.444	0.451	0.453	0.461
age	10 vs 12	0.725	0.67	0.674	0.728	0.740
age	11 vs 12	0.562	0.529	0.532	0.560	0.563
dt		0.994	0.991	0.991	0.994	0.994
prev		> 999.999	> 999.999	> 999.999	> 999.999	> 999.999
actipass	0 vs 1	9.664	9.944	9.831	9.794	9.821
timemonth		0.914	0.910	0.910	0.915	0.915
Y_{ij-1}	1 vs 2	0.289			0.297	0.295
Y_{ij-2}	1 vs 2		0.493		0.549	0.560
Y_{ij-3}	1 vs 2			0.655		0.652

Table 6.5: Odds ratio estimates for first, second and third order models

Parameter	DF	Estimate	Standard Error	P-value	Estimate	Standard Error	P-value	Estimate	Standard Error	P-value	Estimate	Standard Error	P-value	Estimate	Standard Error	P-value
Intercept	1	-4.500	28.754	0.876	-4.763	40.200	0.906	-4.8865	39.8285	0.9024	-4.235	28.399	0.882	-4.037	28.070	0.886
age 0	1	-7.271	344.800	0.983	-7.938	482.200	0.987	-7.9371	477.8	0.9867	-7.228	340.600	0.983	-7.196	336.600	0.983
age 1	1	0.692	28.744	0.981	0.667	40.192	0.987	0.6963	39.8203	0.986	0.666	28.387	0.981	0.669	28.056	0.981
age 2	1	1.060	28.742	0.971	1.098	40.191	0.978	1.0911	39.819	0.9781	1.058	28.385	0.970	1.053	28.055	0.970
age 3	1	1.197	28.741	0.967	1.255	40.190	0.975	1.246	39.8183	0.975	1.199	28.384	0.966	1.194	28.054	0.966
age 4	1	0.611	28.741	0.983	0.671	40.190	0.987	0.6588	39.8184	0.9868	0.614	28.384	0.983	0.605	28.054	0.983
age 5	1	-1.075	28.753	0.970	-1.046	40.199	0.979	-1.0554	39.8272	0.9789	-1.067	28.397	0.970	-1.066	28.066	0.970
age 6	1	0.054	28.745	0.999	0.097	40.193	0.998	0.0891	39.8215	0.9982	0.061	28.389	0.998	0.064	28.058	0.998
age 7	1	-0.581	28.752	0.984	-0.532	40.198	0.989	-0.5294	39.8267	0.9894	-0.581	28.396	0.984	-0.576	28.065	0.984
age 8	1	0.642	28.742	0.982	0.684	40.191	0.986	0.6826	39.8193	0.9863	0.641	28.386	0.982	0.640	28.055	0.982
age 9	1	0.803	28.743	0.978	0.911	40.191	0.982	0.924	39.8196	0.9815	0.790	28.386	0.978	0.791	28.055	0.978
age 10	1	1.267	28.743	0.965	1.323	40.191	0.974	1.3252	39.8196	0.9735	1.264	28.386	0.965	1.264	28.055	0.964
age 11	1	1.013	28.743	0.972	1.086	40.192	0.978	1.0886	39.82	0.9782	1.001	28.387	0.972	0.991	28.056	0.972
dt	1	-0.006	0.010	0.523	-0.009	0.010	0.403	-0.00888	0.0103	0.3887	-0.006	0.010	0.527	-0.006	0.010	0.523
prev	1	50.377	8.770	$< .0001$	51.480	8.743	$< .0001$	51.3488	8.7241	$< .0001$	50.609	8.796	$< .0001$	50.679	8.804	$< .0001$
actipass 0	1	1.134	0.101	$< .0001$	1.149	0.101	$< .0001$	1.1428	0.1008	$< .0001$	1.141	0.102	$< .0001$	1.142	0.102	$< .0001$
time	1	-0.090	0.117	0.442	-0.094	0.116	0.417	-0.0949	0.1158	0.4129	-0.089	0.117	0.447	-0.088	0.117	0.450
Y_{ij-1}	1	-0.620	0.198	0.002							-0.607	0.198	0.002	-0.610	0.198	0.002
Y_{ij-2}	1				-0.354	0.305	0.247				-0.300	0.306	0.327	-0.290	0.306	0.344
Y_{ij-3}	1							-0.2112	0.3058	0.4899				-0.214	0.306	0.484

Table 6.6: Comparative Parameter estimates for the full model including first, second,third and first and second order terms

Chapter 7

Estimating the force of infection and the rate of recovery for the RSV disease process

7.1 Introduction

In this chapter a disease model is developed and applied to the Kilifi data set for Respiratory Syncytial Virus (RSV) and compared to an independent analysis by White et al. (2003) and Nokes et al. (2004). This chapter also aims to develop statistical methods for the estimation of disease model parameters given observed data from the process. These methods involve the use of direct likelihood and generalized linear modelling (GLM) to estimate important disease parameters. The force of infection and the recovery rate are here the key parameters of interest. The findings of the current chapter are consistent and in agreement to results from a mechanistic model in White

et al. (2003). The aspect of time varying disease parameters for the RSV disease is also briefly discussed and addressed in the chapter. Some of the theory of this chapter is an extension to that introduced in Chapter 3 because we need to show and emphasize the application of the statistical techniques to the exact theory.

7.2 Introduction of Disease Dynamics

In the field of infectious disease modelling, one area that is now attracting a lot of attention, is that of the statistical estimation of the key parameters associated with the disease processes. These key parameter estimates are based on observed data that is generated by the underlying disease process. In this chapter we in particular consider the problem of the force of infection and the recovery rate for estimating a disease process. The disease of interest is a respiratory infection on children mainly under the age of one year. It is a viral disease caused by the Respiratory Syncytial Virus (RSV). Mathematical models to study the disease are not new. Greehalgh et al.(2000) used both theoretical and deterministic models to study the RSV dynamics. In the chapter we address the problem of combining the estimation of model parameters and disease dynamics. Recall that the data used in our case is repeated measurements data representing the status of a child as either infected (1) or not (0) at a particular time point t_{ij} where the index i denotes a child and j denotes the order of observation for $j = 1, \ldots, n_i$; implying that child i contributed n_i observation. A fixed number of children were observed in this study. Thus, the data constitute repeated non–normal data suggesting the use of statistical methods of analysis (Generalized Linear Modelling approach suggested by McCullagh and Nelder (1989) and an ex-

tension to generalized linear mixed models (Molenberghs and Verbeke, 2005; Lee, Nelder and Pawitan, (2006)) to be able to account for the correlation of responses within the same subject or cluster. In the current study we employ direct likelihood estimation and discuss the implementation of the generalized linear modelling approach (McCullagh and Nelder, 1989) for the estimation of the recovery rate and time-dependent force of infection. First the basic dynamics of the infection process will be reviewed after which we shall present how we carried out the estimation of the model parameters. One complicating factor in the process is that of time dependence of the some of the parameters in the underlying process, hence the need to allow time dependence in the estimation of the parameters. The time varying parameters may be due to seasonal dynamics inherent in a disease process or due to other factors such as genetic or other unobserved effects.

7.3 Brief History and Discussion of RSV

Respiratory syncytial virus (RSV) infection, which manifests primarily as bronchiolitis and/or viral pneumonia, is the leading cause of lower respiratory tract (LRT) infection in infants and young children. The clinical entity of bronchiolitis was described at least 100 years ago. In 1956, RSV, as the causative agent of most epidemic bronchiolitis cases, initially was isolated by Morris (1956) and colleagues from chimpanzees with upper respiratory tract (URT) infections. Subsequently, Channock et al.(1996) associated this agent with bronchiolitis and LRT infection in infants. Since then, multiple epidemiologic studies have confirmed the role of this virus as the leading cause of LRT infection in infants and young children. Cane (2001) state that human RSV causes LRT disease in about 40% of primary cases and

is responsible for the hospitalization of $0.1\% - 2\%$ of infants under the age group of 1 year annually. Peak incidence of occurrence is observed at age 2-8 months. Overall, about 4 million children younger than 4 years acquire an RSV infection, and in a country such as the United States more than 100,000 children are hospitalized annually because of this infection. This translates to 9-14 per 1000 children younger than 1 year who are hospitalized annually for this condition. Virtually all children have had at least one RSV infection by their third birthday. Given the prevalence and potential severity of this condition, it is not surprising that the World Health Organization has targeted RSV for vaccine development. The frequency of RSV can be categorized as follows:

- Internationally: RSV infection is prevalent worldwide, with similar clinical manifestations and young age of RSV LRT infection.
- Race: All races appear susceptible to RSV, with similar disease patterns.
- Sex: Although boys and girls are affected equally by milder RSV disease, the frequency of hospitalization for RSV disease is higher in males, with a male-female-ratio of approximately 2:1.
- Age: Severe RSV disease is primarily a disease of young infants and children, with a peak occurrence at age 2-8 months. Reinfection with RSV occurs throughout life, with disease becoming more limited to the URT

7.4 Brief RSV Data Description

In Chapter 1, we noted that the Kilifi RSV data study is a repeated measurement (longitudinal) data set measuring the prevalence of the Respiratory Syncytial Virus (RSV, a causal agent of pneumonia) in children in coastal Kenya. By definition, a longitudinal study is one where data are obtained when a response is measured repeatedly on a set of units. The Kilifi data set is part of a study carried out by the Wellcome Trust of the UK Centre, Kilifi, in collaboration with the Kenyan Medical Research Institute. The model that will be built to here will aid in further understanding the dynamics of the disease and possibly aid in the design of intervention strategies for this disease affecting mostly children. Statistical inference about the disease process can also be drawn from such a model. In the current analysis we adopt the available case (AC) kind of analysis using likelihood and generalized linear modelling approaches. The data set description has already been covered in Chapter 1 and will not be repeated here. It is clear that the response variable (rsv) in the data set is a binary non-Gaussian variable. The generalized linear model for longitudinal data seems possibly the best option to deal with such a data set. This idea will be investigated in this chapter. Although the Kilifi data set cannot be appropriately analyzed as Gaussian longitudinal data the similarities and dis-similarities with non-Gaussian data is important in order to develop an appropriate model for it. The data set exhibits a form of incompleteness which has to be properly accounted for in order to carry out an appropriate analysis of the data which will lead to correct conclusions. This incompleteness refers to the real underlying process of the disease which is not observed directly except only through the outcomes of such a process. Incompleteness is addressed in Chapter 9. The model that will be built to represent this data will aid in understanding and in the

design of intervention strategies for this disease affecting mostly children. Proper inference about the disease process can also be drawn from such a model. Let Y_{ij} denote the outcome at observation time t_{ij} for individual i. Then assuming a first order Markov model (Diggle et al., 2002), the observed transition matrix can be represented as in Table 7.1 below. As part of the exploratory data analysis in Chapter 1, a program was written in SAS Proc IML to get the following 2×2 overall transition matrix:

		Y_{ij}	
		uninfected	infected
Y_{ij-1}	uninfected	8598	132
	infected	131	13

Table 7.1: Matrix of transitions between infected and uninfected states

Table 7.1 gives the overall number of visits to the uninfected and infected states conditional on the previous state indicated by the row label. From the above matrix, it is clear that this disease is a rare one because most of the transitions were from uninfected to uninfected states. There are a total of 131 transitions among the children from the uninfected to the infected state and almost a similar number, of 132, transiting from infected to uninfected. This outcome is in agreement with the fact that RSV is a rare disease process. It is important to note that the time interval between transitions was not constant. The time intervals were different within and between the children, which as previously stated, makes the data set highly unbalanced. Therefore standard methods of analysis may not be directly applicable. Since the data consists of individuals who got infected and also changed state back to the uninfected state it is possible to use the data to estimate both the force of

infection and the per capita loss of infection or the recovery rate.

7.5 The Susceptible–Infected–Susceptible (SIS) Model

In the SIS disease model, each individual in the population is either infected (I) or susceptible to infection (S). When a susceptible individual becomes infected, it is immediately infectious and when an infected individual is cured, it is immediately susceptible again. In other words the disease does not confer permanent immunity. This is a homogenous mixing type of model, in which every infected individual has the same probability to infect each susceptible, each infected individual has the same probability of being cured. Ross (1915) introduced the deterministic SIS model while Weiss and Dishon (1971) introduced the stochastic SIS model which is a continuous time Markov birth-and-death process that is used to model a variety of processes that range from epidemics, transmission of rumors and chemical reactions. It is also important to note that the short and long term behaviour of the deterministic and stochastic versions of the SIS model are quite different and we will not go into the details of this difference. However a deterministic model is like an average process while a stochastic model accounts for the random variability in the model parameters and variables. In the current problem it should be noted that according to the biology of RSV the disease process may not necessarily be an SIS disease process but rather a more appropriate model would be the susceptible-infected-recovered (SIR) process. A SIR disease process is defined as a process where a subject is susceptible, becomes infected and then recovered with an immunity not to be infected again, for example Rubella or Mumps are diseases which can be modelled

as a SIR process. However since the data present currently was collected on children within the age of one year followed over a period of approximately one year, we model the the transition rates from the disease free to the diseased state (λ) and back (ν) using an SIS model because most children were infected soon after recovery implying a negligible duration in the recovered state for this particular cohort of individuals. The problem is to model the observed data which is a repeated (longitudinal) type of data where each child presents a sequence responses of 1's (diseased) and 0's (disease-free). The model we construct assumes that the disease dynamics have attained equilibrium hence a constant population. In effect deaths are balanced by new births, therefore natural mortality and births are not included in the model.

7.5.1 SIS governing differential equation

The SIS basic governing differential equation is given as follows

$$\frac{\partial q(a,t)}{\partial t} + \frac{\partial q(a,t)}{\partial a} = -\lambda(a,t)q(a,t) + \nu(a,t)p(a,t) \tag{7.1}$$

where $q(a,t)$ and $p(a,t)$ are respectively the proportion of susceptible and infected individuals in the population at time t aged a such that

$$p(a,t) + q(a,t) = 1$$

Thus for a purely SIS model it is enough to study the solution for equation (7.1).The quantities $\lambda(a,t)$ and $\nu(a,t)$ are respectively the force of infection and recovery rate here, both expressed as a function of age and time. However as already mentioned above RSV is a viral disease therefore the most appropriate model is the SIR model where R is the class of recovered individuals with a possible loss of immunity to revert back to the S class. Thus in this case the equation for $p(a,t)$ would become

$$\frac{\partial p(a,t)}{\partial t} + \frac{\partial p(a,t)}{\partial a} = \lambda(a,t)q(a,t) - \{\nu(a,t) + r(a,t)\}p(a,t)$$

where $r(a,t)$ is the rate at which individuals move from the infected class to the recovered class of the preocess. But because the data currently in use was based on children within the age of one year the immunity against the disease for such individuals is still not yet developed therefore we assume $r(a,t) = 0$. It therefore suffices to deal with equation (7.1). In addition note that natural mortality is here assumed to be balanced by new births therefore in effect we assuming a constant population model. If the individuals in the study are within the same age bracket, such as in the Kilifi data set where the children were all within one year of age then we can drop age, in the above equation and therefore write

$$\frac{dq(t)}{dt} = -\lambda(t)q(t) + \nu(t)p(t). \tag{7.2}$$

If we assume $\lambda(t)$ and $\nu(t)$ are time-independent then

$$\frac{dq(t)}{dt} = -\lambda q + \nu(1-q) = -(\lambda + \nu)q + \nu \tag{7.3}$$

because $p(t) + q(t) = 1$. This equation can now be solved using the 'variation of coefficients' technique. The steps to the solution of the SIS governing differential equation (7.3) are outlined below. Put the linear equation in the standard form as

$$\frac{dy}{dt} + P(t)y = f(t).$$

The integrating factor of the standard form is given by $e^{\int P(t)dt}$. Next multiply the standard form of the equation by the integrating factor and note that the left hand side of the resulting equation is automatically the derivative of the product of the integrating factor and y that is,

$$\frac{d}{dt}[e^{\int P(t)dx}y] = e^{\int P(t)dt}f(t).$$

Lastly integrate both sides of this last equation and solve for y subject to the initial conditions of the system.

Thus, the solution to the equation

$$\frac{dq(t)}{dt} = -\lambda q + \nu(1-q),$$

can be constructed by first noting that,

$$\begin{aligned} \frac{dq(t)}{dt} &= -\lambda q + \nu - \nu q, \\ \Rightarrow \frac{dq(t)}{dt} &= -(\lambda+\nu)q + \nu, \end{aligned}$$

implying that

$$\frac{dq(t)}{dt} + (\lambda+\nu)q = \nu.$$

Multiplying both sides by the integrating factor yields

$$\begin{aligned} e^{(\lambda+\nu)t}\frac{dq(t)}{dt} + e^{(\lambda+\nu)t}(\lambda+\nu)q &= e^{(\lambda+\nu)t}\nu \\ \frac{d}{dt}[e^{(\lambda+\nu)t}q(t)] &= \nu e^{(\lambda+\nu)t} \\ \int \frac{d}{dt}[e^{(\lambda+\nu)t}q(t)] &= \int \nu e^{(\lambda+\nu)t}dt \\ e^{(\lambda+\nu)t}q(t) &= \frac{\nu}{\lambda+\nu}e^{(\lambda+\nu)t} + c \\ q(t) &= \frac{\nu}{\lambda+\nu} + ce^{-(\lambda+\nu)t} \end{aligned}$$

Imposing the initial condition that at $t = 0$ that the proportion infected is 0 implies that $q(0) = 1$ and $p(0) = 0$, we can solve for c and get $c = 1 - \frac{\nu}{\lambda+\nu} = \frac{\lambda}{\lambda+\nu}$. Hence we can solve for $q(t)$ and the solution obtained as:

$$q(t) = \frac{\nu}{\lambda+\nu} + \frac{\lambda}{\lambda+\nu}e^{-(\lambda+\nu)t}, \tag{7.4}$$

assuming $q(0) = 1$ and $p(0) = 0$ as the initial conditions and since $p(t)+q(t) = 1$ we get

$$p(t) = \frac{\lambda}{\lambda+\nu} - \frac{\lambda}{\lambda+\nu}e^{-(\lambda+\nu)t} \tag{7.5}$$

as the general solutions for $p(t)$.

Note that equations (7.4) and (7.5) imply that $q(\infty) = \frac{\nu}{\lambda+\nu}$ and hence $p(\infty) = \frac{\lambda}{\nu+\lambda}$ which give the equilibrium proportions of susceptible and infected individuals respectively. This means that for a rare disease $\nu >> \lambda$. Now let the indicators 1 and 0 denote respectively the infected and uninfected states of an individual so that we can define the four conditional transition probabilities as follows

$$\begin{aligned}
\pi_{00}(t) &= P(Y_{it} = 0 | Y_{i,0} = 0) \\
\pi_{01}(t) &= P(Y_{it} = 1 | Y_{i,0} = 0) \\
\pi_{10}(t) &= P(Y_{it} = 0 | Y_{i,0} = 1) \\
\pi_{11}(t) &= P(Y_{it} = 1 | Y_{i,0} = 1)
\end{aligned}$$

As before assume initially at $t = 0$ the proportion infected is 0 that is $q(0) = 1$ and $p(0) = 0$. Note that since the disease process is a reversible process, individuals cannot remain infected forever. The solution $q(t)$ implies that given an individual was infected at the beginning of a time duration t then,

$$\pi_{00}(t) = \frac{\nu}{(\lambda+\nu)} + \frac{\lambda}{(\lambda+\nu)} e^{-(\lambda+\nu)t} \tag{7.6}$$

and since $\pi_{00} + \pi_{01} = 1$, then

$$\pi_{01}(t) = \frac{\lambda}{(\lambda+\nu)} - \frac{\lambda}{(\lambda+\nu)} e^{-(\lambda+\nu)t} \tag{7.7}$$

Following similar arguments we can write expressions for $\pi_{11}(t)$ and $\pi_{10}(t)$ as:

$$\pi_{11}(t) = \frac{\lambda}{(\lambda+\nu)} + \frac{\nu}{(\lambda+\nu)} e^{-(\lambda+\nu)t} \tag{7.8}$$

and

$$\pi_{10}(t) = \frac{\nu}{(\lambda+\nu)} - \frac{\nu}{(\lambda+\nu)} e^{-(\lambda+\nu)t} \tag{7.9}$$

We can also use these transition probabilities, $\pi_{00}(t), \pi_{01}(t), \pi_{10}(t)$ and $\pi_{11}(t)$ to form the following transition matrix,

$$P = \begin{pmatrix} \pi_{00}(t) & \pi_{01}(t) \\ \pi_{10}(t) & \pi_{11}(t) \end{pmatrix} = \begin{pmatrix} 1-\pi_{01}(t) & \pi_{01}(t) \\ \pi_{10}(t) & 1-\pi_{10}(t) \end{pmatrix} \tag{7.10}$$

Note that the process satisfies the ergodic property namely, $\pi_{00}(\infty) = \pi_{10}(\infty) = \frac{\nu}{\nu+\lambda}$ and $\pi_{01}(\infty) = \pi_{11}(\infty) = \frac{\lambda}{\nu+\lambda}$ the ultimate equilibrium proportion of susceptible and infected respectively. Estimates of λ and ν can be obtained from these equations via the maximum likelihood estimation since the transitions represent conditionally i.i.d Bernoulli observations with probabilities π_{10} and π_{01}. The general form of the likelihood can be written as:

$$\{\prod_{i=1}^{N} P(Y_{i,0})\} \prod_{i=1}^{N} \prod_{j=1}^{n_i} P(Y_{i,j}|Y_{i,j-1})$$

using the notation that $Y_{i,j}$ denotes the binary observation for child i at time occasion j out of n_i time occasions in total. The second part of the likelihood, which is the partial likelihood obtained by conditioning on the first measurement $Y_{i,0}$ is proportional to a set of two Binomial distributions that is,

$$\prod_{i=1}^{N} \prod_{i=1}^{n_i} P(Y_{i,j}|Y_{i,j-1}) \propto (\pi_{01})^{n_{01}} (1-\pi_{01})^{n_{00}} (\pi_{10})^{n_{10}} (1-\pi_{10})^{n_{11}}$$

where $n_{k,l}$ are the total number of transitions from state $k \in (0,1)$ to state $l \in (0,1)$ and therefore explicit maximization is possible. Thus conditional on the initial state $\{Y_{i,0}\}$, the disease parameters π_{01} and π_{10} are orthogonal. The two Binomial distributions correspond to the recovery and infection process of the disease. There is an inherent assumption here that the time intervals are of equal length and that the transition probabilities are time independent. It is possible to estimate the transition probabilities by maximizing this partial likelihood instead of the full likelihood, since the first

measurement $Y_{i,0}$ contributes a limited amount of information only if some steady state assumptions are made. Alternatively one can assume the initial state of the child is known with probability one. The maximum likelihood estimates of the transition probabilities obtained this way are:

$$\tilde{\pi}_{01} = \frac{n_{01}}{n_{01} + n_{00}}$$

and

$$\tilde{\pi}_{10} = \frac{n_{10}}{n_{10} + n_{11}}$$

By equating these estimates of the transition probabilities in Eq.(7.6) and Eq.(7.8) one can obtain estimators of the transition rate λ and ν. Furthermore if we consider the table of transitions calculated earlier, that is

		Y_{ij}	
		uninfected	infected
Y_{ij-1}	uninfected	8598	132
	infected	131	13

Table 7.2: Matrix of transitions between infected and uninfected states

Then we can work out the probability transition matrix P as :

$$P = \begin{pmatrix} 0.985 & 0.015 \\ 0.91 & 0.09 \end{pmatrix}$$

The problem with this approach is that the estimating equations so obtained are highly non-linear but the method works well for equally spaced observation times, as in Nagelkerke et al. (1990). The added problem with this approach of estimating λ and ν is that it is not straightforward to find standard errors of estimates. It should be noted that under such assumptions the two disease processes namely the infection and recovery processes are completely disentangled.

7.6 Estimation of the model parameters

An alternative estimation procedure is developed by assuming that the residence times in each disease state is exponentially distributed. In the current case we assume that the duration in the disease free state is exponentially distributed with parameter λ while the duration in the disease state is exponentially distributed with parameter ν. Thus we can interpret λ and ν as the force of infection and the recovery rate respectively. Both λ and ν have units of measurement as days^{-1}. The aim is to use available data to obtain estimates for the disease parameters. From a time to event approach the two parameters can also be seen as the hazard of infection and recovery respectively. We therefore define the four transition probabilities as follows:

$$\begin{aligned}
\pi_{00} &= P(Y_{ij} = 0 | Y_{i,j-1} = 0, d_{ij}) = e^{-\lambda d_{ij}} \\
\pi_{01} &= P(Y_{ij} = 1 | Y_{i,j-1} = 0, d_{ij}) = 1 - e^{-\lambda d_{ij}} \\
\pi_{10} &= P(Y_{ij} = 0 | Y_{i,j-1} = 1, d_{ij}) = 1 - e^{-\nu d_{ij}} \\
\pi_{11} &= P(Y_{ij} = 1 | Y_{i,j-1} = 1, d_{ij}) = e^{-\nu d_{ij}}
\end{aligned}$$

where $d_{ij} = t_{ij} - t_{ij-1}$, is the time interval between the visit at time t_{ij} and t_{ij-1}. So far the approach is similar to that developed by White et al. (2003). The point of departure is that we now propose to use the direct likelihood approach instead of the Bayesian approach. The full likelihood can therefore be written as:

$$L(\nu, \lambda, dt) = (\theta_1)^{\sum \delta_i}(1-\theta_1)^{N-\sum \delta_i} \prod_{0\to 0} e^{-\lambda d_{ij}} \prod_{0\to 1}(1-e^{-\lambda d_{ij}}) \prod_{1\to 0}(1-e^{-\nu d_{ij}}) \prod_{1\to 1} e^{-\nu d_{ij}}$$

Now δ_i is an indicator variable denoting the initial state of a child where $\delta_i = 1$ when the child is initially infected and 0 otherwise. Hence θ_1 is the probability that the child is initially in the infected state such that $\theta_0 = 1 - \theta_1$, N is the total number of individuals in the study and $\sum \delta_i$ are individuals who are

initially in the infected state and $N - \sum \delta_i$ are initially not infected. It is thus simpler to consider the conditional likelihood given the initial states $\{Y_{i,0}\}$ in order to find the maximum likelihood estimates (MLEs) of the parameters λ and ν If we take the log-likelihood then we have:

$$\ell = \log L = \log(constant) - \lambda \sum_{0\to 0} d_{ij} + \sum_{0\to 1} \log(1 - e^{-\lambda d_{ij}}) + \sum_{1\to 0} \log(1 - e^{-\nu d_{ij}}) - \nu \sum_{1\to 1} d_{ij}$$

Taking the first and second partial derivative with respect to λ and ν gives us the following set of equations

$$\begin{aligned}
\frac{\partial \ell}{\partial \lambda} &= -\sum_{0\to 0} d_{ij} + \sum_{0\to 1} [1/(1 - e^{-\lambda d_{ij}})](e^{-\lambda d_{ij}})(d_{ij}) \\
\frac{\partial \ell}{\partial \nu} &= -\sum_{1\to 1} d_{ij} + \sum_{1\to 0} [1/(1 - e^{-\nu d_{ij}})](e^{-\nu dij})(d_{ij}) \\
\frac{\partial^2 \ell}{\partial \lambda^2} &= -\sum_{0\to 1} [1/(1 - e^{-\lambda d_{ij}})]^2 [(e^{-\lambda d_{ij}}) d_{ij}]^2 + \sum_{0\to 1} [1/(1 - e^{-\lambda d_{ij}})](e^{-\lambda d_{ij}}) d_{ij}^2 \\
\frac{\partial^2 \ell}{\partial \nu^2} &= -\sum_{1\to 0} [1/(1 - e^{-\nu d_{ij}})]^2 [(e^{-\nu d_{ij}}) d_{ij}]^2 + \sum_{1\to 0} [1/(1 - e^{-\nu d_{ij}})](e^{-\nu d_{ij}}) d_{ij}^2 \\
\frac{\partial^2 \ell}{\partial \nu \partial \lambda} &= \frac{\partial^2 \ell}{\partial \lambda \partial \nu} = 0
\end{aligned}$$

Below the Fisher's scoring method to iteratively solve for λ and ν is briefly described. Searle (1992) and Longford (1993) state that the Fisher's scoring method is preferred to Newton-Raphson method since it avoids the heavy computational burden of finding the Hessian matrix (the matrix of second derivatives of the loglikelihood) by using the inverse of the information matrix I^{-1} (i.e. we replace the Hessian by the negative of its expected value, which is often easier to compute than the Hessian). The inverse of the information matrix will be required to get the estimated variance-covariance matrix of our parameters. For generality purposes let the parameters λ and ν to be contained in a vector θ. The iterative scheme is then given by:

$$\theta^{(m+1)} = \theta^{(m)} + [I(\theta)^{(m)}]^{-1} \left[\frac{\partial \ell}{\partial \theta}\right]_{\theta = \theta^{(m)}}$$

where the superscript (m) denotes the m^{th} iteration and $I(\theta)^{(m)}$ is the estimate of the information matrix given $\theta = \theta^{(m)}$. The parameters λ and ν were then estimated using the Fisher's scoring method to yield the following estimates along with their standard errors:

Estimator	Estimate	Standard Error
$\hat{\lambda}$	0.001169	0.0001114
$\hat{\nu}$	0.45495	0.067

Table 7.3: Parameter Estimates

Hence a 95% confidence interval for λ is $(0.000951, 0.001388)$ and likewise for ν is $(0.32362, 0.58626)$. It is clear indeed based on these parameter estimates that RSV is a very rare disease since $\lambda << \nu$. Alternatively note that $\pi_{11} = \pi_{01} = 0.0026$ and $\pi_{00} = \pi_{10} = 0.9974$. This means that in the long run a child is disease free 99.7% of the times and infected 0.3% of the times. Also the (maximum likelihood, ML) limiting transition probability matrix $\tilde{P}$ is given by:

$$\tilde{P} = \begin{pmatrix} 0.9974 & 0.026 \\ 0.9974 & 0.026 \end{pmatrix}$$

Estimator	Estimate	Standard Error
$\hat{\lambda}$	0.00135	0.0001
$\hat{\nu}$	0.4979	0.067

Table 7.4: White et al. (2003) Parameter Estimates

Table 7.4 gives estimates by White et al. (2003). Although the authors used a Bayesian MCMC approach our estimates give similar results.

7.7 Application of the GLM to RSV data

As earlier defined, let λ and ν denote the force of infection and the recovery rate for the disease transition process. If we apply the generalized linear model to derive the force of infection for RSV, it will be necessary to consider data on the transitions from the uninfected to infected states namely, from state 0 to state 1 or $0 \rightarrow 1$ and the transitions from uninfected to uninfected that is $0 \rightarrow 0$. These transitions would make up 2 binary events for the response variable and once these transitions are coded as 1 for $0 \rightarrow 0$ and a 0 for $0 \rightarrow 1$, the response variable can be seen to conditionally follow a binomial distribution. Likewise, another pair of binary responses can be similarly defined by considering the transitions $1 \rightarrow 1$ and $1 \rightarrow 0$. The residence times in the infected and uninfected states are assumed to follow the exponential distribution with parameters λ and ν, respectively. In survival analysis terminology, λ can also be interpreted as the hazard of infection or per capita risk of infection. The simpler model is where the only explanatory variable is the inter-state time duration that is, the quantity d_{ij}. Using generalized linear model (GLM) with log link function we obtain

$$\log(\pi_{00}) = -\lambda d_{ij}$$

and

$$\log(\pi_{11}) = -\nu d_{ij}$$

Since the data consist of 4 possible transition probabilities in equation (7.11), in order to formulate an appropriate GLM we define an indicator variable

$$Z_{ij} = \begin{cases} 1 & Y_{ij} = 0, Y_{i,j-1} = 0, \\ 0 & Y_{ij} = 0, Y_{i,j-1} = 1, \\ 0 & Y_{ij} = 1, Y_{i,j-1} = 0, \\ 1 & Y_{ij} = 1, Y_{i,j-1} = 1. \end{cases} \tag{7.11}$$

Let $\theta_{ij} = P(Z_{ij} = 1)$ and consider the following linear predictor

$$\log(\theta_{ij}) = -\lambda d_{ij} \times (1 - Y_{i,j-1}) - \nu d_{ij} \times (Y_{i,j-1}), \tag{7.12}$$

it follows that

$$\log(\theta_{ij}) = \begin{cases} -\lambda d_{ij} & \text{if } Y_{ij} = 0, Y_{i,j-1} = 0, \\ -\nu d_{ij} & \text{if } Y_{ij} = 1, Y_{i,j-1} = 1. \end{cases} \tag{7.13}$$

Thus, using this approach we obtained $\hat{\lambda} = 0.0021$ (95% C.I: 0.0018-0.0024) and $\hat{\nu} = 0.503$ (95% C.I: 0.386-0.657) for the force of infection and the recovery rate, respectively. Note that the parameter estimates obtained in our analysis are slightly an overestimate compared to those by White et al. (2003), $0.00135\,(0.00114 - 0.00157)$ and $0.498\,(0.387, 0.648)$ for the force of infection and the recovery rate, respectively. These estimates are slightly higher than those obtained by the direct likelihood method in Section 5 but are very similar. The cause of the difference albeit a very small one could be attributed to the computational process.

7.8 Time dependent force of infection

The above estimation procedures only helped us to estimate a constant force of infection and recovery rate over the time period of the study. However, there is enough evidence that a disease such as RSV does exhibit clear temporal variation in its incidences, which is a function of the force of infection. Thus, we extended the above approach to obtain monthly piecewise estimates of the force of infection. However, the recovery rate is assumed to be constant or time homogeneous. For months 14 and 15, there are no data because none of the children completed the study up to months 14 and 15. A piecewise constant force of infection with log link function was assumed.

Hence, the linear predictor is given by

$$\log(\theta_{ij}) = -\lambda_k d_{ij} \times (1 - Y_{i,j-1}) - \nu_k d_{ij} \times (Y_{i,j-1}). \tag{7.14}$$

Here, λ_k is the monthly force of infection. Note that the model in (7.14) can be expressed also as a model with complementary-log-log link, in which the linear predictor is given by

$$g(\theta_{ij}) = \log(-\lambda_k) d_{ij} \times (1 - Y_{i,j-1}) - \log(\nu_k) d_{ij} \times (Y_{i,j-1}), \tag{7.15}$$

where g is the complementary-log-log link function. In such a model, the monthly parameter estimates for the force of infection and the constant recovery rate are equal to $\log(\lambda_k)$ and $\log(\nu_k)$, respectively. As a result, the parameter estimates for the monthly force of infection and the constant force of infection are constrained to be non- negative, as required. In this chapter, the complementary-log-log link function was used to estimate the model's parameters. 95% confidence intervals were obtained either by exponentiating the model parameters and their confidence intervals or by applying the delta method for the log of the parameters. Tables 7.5 and 7.6 present the parameter estimates for the monthly force of infection and recovery rate respectively. The force of infection peaks with different heights in months 3 ($\hat{\lambda}_3 = 0.007$), then it decreases to zero at month 9 and increase to secondary peaks at months 11 and 12 ($\hat{\lambda}_{11} = 0.0022$ and $\hat{\lambda}_{12} = 0.0022$, respectively). Month 1 had too few transitions recorded in it while months 14 and 15 did not have any data in them since the children did not complete the study for these months. Hence, these months have been omitted in the analysis. Figures 7.1 and Table 7.5 shows a plot of the force infection against time together with 95% confidence intervals from both direct exponentiation and the delta method. Table 7.6 and Figure 7.2 show that the recovery rate remains virtually constant in the entire year of follow up of study cohorts.

Thus while the force of infection portrays a time varying characteristic, the rate of recovery is not much affected by time. It is possible that there would be a biological reason for this observation.

		Exponentiation		Delta Method	
Month	Lambda	95% Confidence Interval		95% Confidence Interval	
$\hat{\lambda}_2$	0.0053	0.0032	0.0086	0.0027	0.0079
$\hat{\lambda}_3$	0.0070	0.0053	0.0092	0.0051	0.0089
$\hat{\lambda}_4$	0.0051	0.0038	0.0070	0.0036	0.0067
$\hat{\lambda}_5$	0.0024	0.0016	0.0037	0.0014	0.0034
$\hat{\lambda}_6$	0.0019	0.0011	0.0033	0.0009	0.0029
$\hat{\lambda}_7$	0.0010	0.0005	0.0020	0.0003	0.0017
$\hat{\lambda}_8$	0.0001	0.0000	0.0008	-0.0001	0.0003
$\hat{\lambda}_9$	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\lambda}_{10}$	0.0001	0.0000	0.0009	-0.0001	0.0004
$\hat{\lambda}_{11}$	0.0022	0.0014	0.0033	0.0013	0.0031
$\hat{\lambda}_{12}$	0.0022	0.0015	0.0032	0.0013	0.0030
$\hat{\lambda}_{13}$	0.0014	0.0007	0.0029	0.0004	0.0024

Table 7.5: Monthly estimates of the force of infection and confidence Intervals

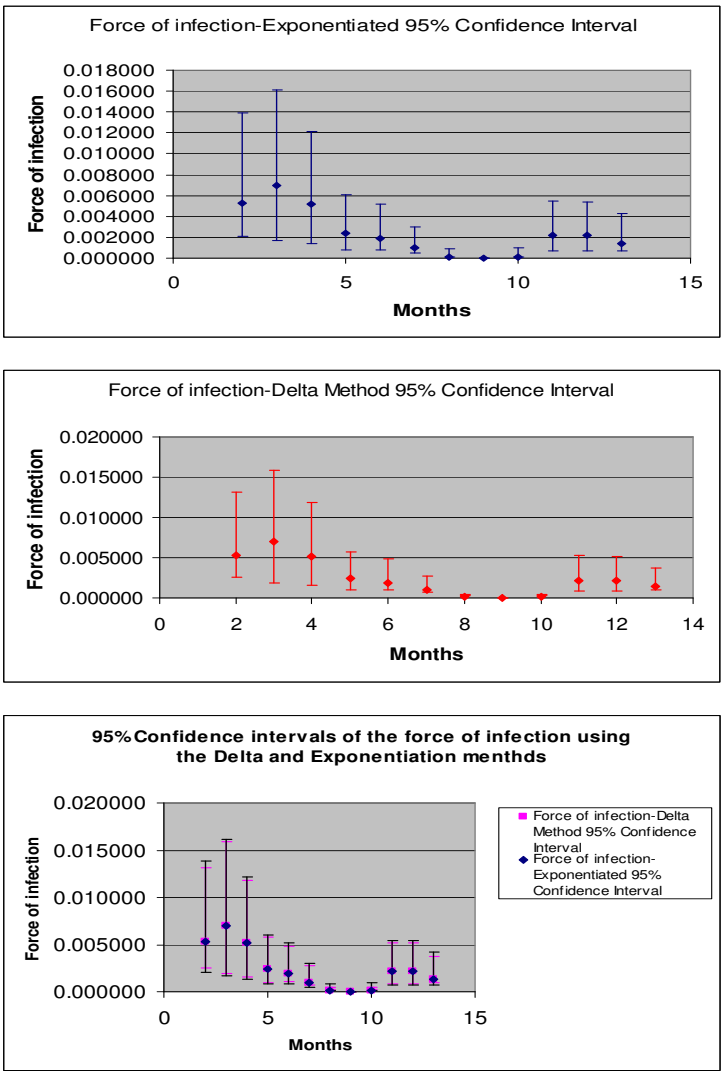

Figure 7.1: The force of infection in months together with 95% confidence intervals using the exponentiated and delta methods.

Monthly estimates of the recovery rate were also similarly obtained and the values are tabulated below for comparison purposes.

Month	Nu	Estimate	Standard Error
2	$\hat{\nu}_2$	0.4990	0.067
3	$\hat{\nu}_3$	0.5000	0.06
4	$\hat{\nu}_4$	0.5036	0.064
5	$\hat{\nu}_5$	0.5021	0.062
6	$\hat{\nu}_6$	0.4990	0.066
7	$\hat{\nu}_7$	0.500	0.076
8	$\hat{\nu}_8$	0.5002	0.072
9	$\hat{\nu}_9$	0.5022	0.065
10	$\hat{\nu}_{10}$	0.5009	0.06
11	$\hat{\nu}_{11}$	0.5006	0.071
12	$\hat{\nu}_{12}$	0.4996	0.061
13	$\hat{\nu}_{13}$	0.5004	0.069

Table 7.6: Monthly estimates of the recovery rate

Months 14 and 15 did not have any data in them because none of the children completed the study up to months 14 and 15. The rate of recovery is fairly constant over all the months with no unusual peaks in the estimates. Graphically the estimates of the recovery rate plotted monthly over the study period is shown in Figure 7.2

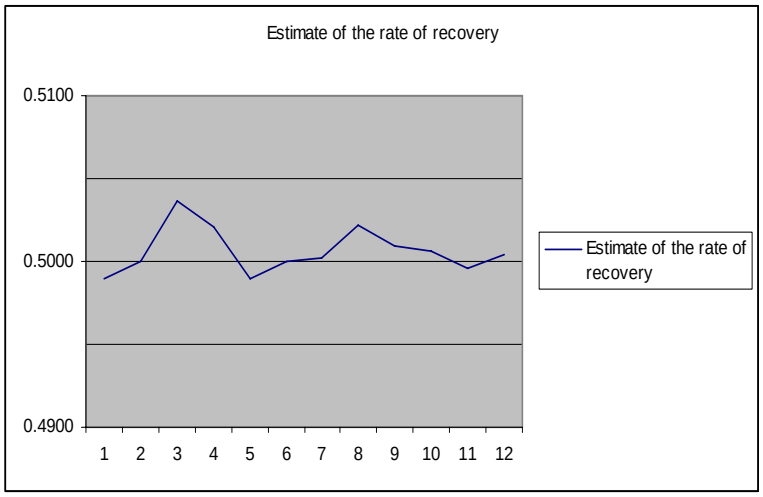

Figure 7.2: The probability of rate of recovery in months.

Table 7.7 gives a comparison of the 15-month piecewise force of infection estimated via the GLM method used in the current analysis and those by White et al. (2003).

	White et al(2003).			GLM-Exponentiation		
Month	Lambda	95% Confidence Int.		Lambda	95% Confidence Int.	
λ_2	0.0045	0.0031	0.0061	0.0053	0.0032	0.0086
λ_3	0.0032	0.0021	0.0044	0.0070	0.0053	0.0092
λ_4	0.0017	0.001	0.0044	0.0051	0.0038	0.0070
λ_5	0.0022	0.0001	0.0027	0.0024	0.0016	0.0037
λ_6	0.0027	0.0001	0.0005	0.0019	0.0011	0.0033
λ_7	0.0022	0.0000	0.0006	0.0010	0.0005	0.0020
λ_8	0.0003	0.0000	0.0008	0.0001	0.0000	0.0008
λ_9	0.0006	0.0001	0.0012	0.0000	0.0000	0.0000
λ_{10}	0.0028	0.0017	0.0042	0.0001	0.0000	0.0009
λ_{11}	0.0028	0.0016	0.0044	0.0022	0.0014	0.0033
λ_{12}	0.0026	0.0014	0.0041	0.0022	0.0015	0.0032
λ_{13}	0.0006	0.0001	0.0019	0.0014	0.0007	0.0029

Table 7.7: Comparison of the monthly estimates of the force of infection

From table 7.7 we can see that the GLM approach gives very similar estimates to those obtained by White et al. (2003). Months 2 (April),3 (May),11 (January) and 13 (March) have the highest forces of infection respectively. A comparison of the overall force of infection and the recovery rate using both the direct likelihood and the GLM approach to that obtained by White et al. (2003) is shown in Table 9 below

	White et al.(2003)		Direct Likelihood(ML)		GLM	
Estimator	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
$\hat{\lambda}$	0.00135	0.0001	0.001169	0.000114	0.0021	0.00015
$\hat{\nu}$	0.4979	0.067	0.45495	0.067	0.5030	0.079

Table 7.8: Comparative Parameter Estimates

7.9 Conclusion

In conclusion, we note that generalised linear modelling combined with likelihood estimation was used to estimate the force of infection and the recovery rate of a childhood respiratory viral disease (RSV). Construction of the full likelihood was not possible therefore a form of conditional likelihood was used to model the data. The generalised modelling approach was modified to estimate monthly specific force of infection for the disease thus allowing the model to capture the temporal trends of disease incidence. One can see from the comparative table above that the ML and GLM estimates are similar to those of White et al. (2003). The force of infection is estimated as $\hat{\lambda} = 0.001169$ and the rate of recovery is estimated as $\hat{\nu} = 0.45495$ using the direct maximum likelihood estimation method. Corresponding estimates using the generalized linear modelling approach are 0.0021 and 0.5030. These two approaches gave quite similar sets of parameter estimates. Thus the ML estimates are closer to the Bayesian MCMC estimates of White et al. (2003) than the GLM estimates. However we prefer the latter because of its flexibility in allowing us to come up with monthly piecewise parameter estimates. It is also seen from the estimation of the monthly parameters that RSV peaks in the $3^{rd}, 11^{th}$ and 12^{th} months that is around January, September and October . This is consistent with the discussions by Cane (2001),

Chew et al. (1998) and Simoes (1999) who all state that RSV has a seasonal signal attributed to meteorological or sociological factors. Furthermore the force of infection is not constant and varies with time. This is characteristic of childhood infections such as the RSV. The parameter estimates also imply that the equilibrium proportion of susceptible and infected children stabilizes at 99.74% and 0.26% which implies that RSV clearly falls in the class of very rare diseases. Nonetheless the disease can be very harmful to children particularly under the age of one year. Thus statistical and mathematical models are a useful tool in understanding it's dynamics and hence assist in designing control and intervention strategies for it.

Chapter 8

Joint Modelling Approach

8.1 Introduction

In various applications, it is common to observe statistical problems with outcomes of a mixed nature. These type of problems, as stated by Molenberghs and Verbeke (2005, pp. 1-2), have been around for almost half a century and are common occurrences in these present times. They state further that the most common situation arises by observing the joint occurrence of a continuous and a binary or ordinal outcome. Areas of application include the fields of psychometry or biometry. Hence the determination of the joint distribution of both outcomes or on specific aspects, such as the association or correlation between the outcomes are usually imperative to the modelling and estimation of parameters. Molenberghs and Verbeke (2005) further state that there are three approaches (briefly discussed below) to modelling a continuous and binary or ordinal outcome.

Approach 1: Postulate a marginal model for the binary outcome and then formulate a conditional model for the continuous outcome, given the cate-

gorical one.

Approach 2: This approach starts with reverse factorization, combining a marginal model for the continuous outcome with a conditional one for the categorical outcome. These conditional models have also been discussed in Cox and Wermuth (1992, 1994), Krzanowski (1988) and Little and Schluchter (1985).

Approach 3: This third model family directly formulates a joint model for the two outcomes. One starts from a bivariate continuous variable, one component of which is explicitly observed and the other one observed in dichotomized, or generally discretized version only. Molenberghs and Verbeke (2005, pp. 441-444) state that a Plackett-Dale approach can be used in this case and that general multivariate exponential family based models have been proposed by Prentice and Zhao (1991), Zhao, Prentice and Self (1992) and Sammel, Ryan and Legler (1997).

Joint modelling is closely related to multivariate modelling and hierarchical modelling and it is inevitable to state that literature in this area for modelling outcomes of various natures is diverse and growing. One can extend the ideas of the three above approaches encompassing the bivariate case above to cases of multivariate continuous outcome and/or a multivariate categorical outcome. In the multivariate outcome setting, it means that for *approaches 1* and *2*, one starts from conditional and marginal multivariate normal and appropriately choose multinomial models, such as one presented in Olkin and Tate (1961). As far as *Approach 3* is concerned, such models were formulated by Hannan and Tate (1965) and Cox (1974) for multivariate

normal with a univariate bivariate or discrete variable. We now look at two examples of joint modelling. One example concerns a bivariate linear mixed model given in Thiébaut et al. (2002) and the other a joint model made up of a binary and continuous outcome described in Molenberghs and Verbeke (2005).

8.2 Examples of joint modelling

8.2.1 Bivariate linear mixed model with normally distributed outcome

Thiébaut et al. (2002, pp 249-251) define a general bivariate linear mixed model including a random component, a first order auto-regressive process and an independent error. They state that in HIV infection, several markers are available to measure the quantity of virus (plasma viral load noted as HIV RNA), the status of the immune system (CD4+ T lymphocytes which are a specific target of the virus, CD8+ T lymphocytes or the inflammation process (β_2 microglobuline). These markers are expected to be associated because as the infection measured by HIV RNA increases it induces inflammation and the destruction of immune cells. Several models have been developed to fit the evolution of CD4 and CD8 cells or CD4 and β_2 microglobuline. Thiébaut et al. (2002) propose the use of multivariate linear mixed models to be fitted to multivariate longitudinal Gaussian data using the SAS MIXED procedure.

The model is formulated as follows: Let $Y_i = \begin{bmatrix} Y_i^1 \\ Y_i^2 \end{bmatrix}$ be the response vector for subject i with Y_i^k denoting an n_i^k-dimensional vector of measurements of marker k $(k = 1, 2)$ with $n_i^1 = n_i^2 = n_i$. If the two markers are independent,

then the two models can be used

$$\left\{ \begin{array}{lcl} Y_i^1 & = & X_i^1\beta^1 + Z_i^1\gamma_i^1 + W_i^1 + \varepsilon_i^1 \\ Y_i^2 & = & X_i^2\beta^2 + Z_i^2\gamma_2^1 + W_i^2 + \varepsilon_i^2 \end{array} \right. \tag{8.1}$$

where $\varepsilon_i^1 \sim N(0, \sigma^2_{\varepsilon^1} I_{n_i})$, $\gamma_i^1 \sim N(0, G^1)$, $W_i^1 \sim N(0, R_i^1)$ and $\varepsilon_i^2 \sim N(0, \sigma^2_{\varepsilon^2} I_{n_i})$, $\gamma_i^2 \sim N(0, G^2)$, $W_i^2 \sim N(0, R_i^2)$ where X_i^k is a $n_i \times p^k$ design matrix, β^k is a p^k vector of fixed effects, Z_i^k is a $n_i \times q^k$ design matrix of individual random effects which is usually a subset of X_i^k, γ_i^k is a q^k vector of individual random effects with $q^k \leq p^k$. W_i^k is a vector of realization of a first order autoregressive process, $w_i^k(t)$ with a covariance structure given by $R_i^k(s,t) = \sigma^2_{w^k} \mathrm{e}^{\lambda^k |t-s|}$ and I_{n_i} is a $n_i \times n_i$ identity matrix.

To take into account correlation between both markers, one could then use the following bivariate linear mixed model:

$$Y_i = X_i\beta + Z_i\gamma_i + W_i + \varepsilon_i \tag{8.2}$$

$$\text{with} \left\{ \begin{array}{ll} \varepsilon_i & \sim N(0, \Sigma_i) \\ W_i & \sim N(0, R_i) \\ \gamma_i & \sim N(0, G_i) \end{array} \right. \text{where}$$

$$X_i = \left[\begin{array}{cc} X_i^1 & 0 \\ 0 & X_i^2 \end{array} \right], \beta = \left[\begin{array}{c} \beta_1 \\ \beta_2 \end{array} \right], Z_i = \left[\begin{array}{cc} Z_i^1 & 0 \\ 0 & Z_i^2 \end{array} \right], \gamma_i = \left[\begin{array}{c} \gamma_i^1 \\ \gamma_i^2 \end{array} \right] \text{ and } W_i = \left[\begin{array}{c} W_i^1 \\ W_i^2 \end{array} \right]$$

is a $2n_i$ vector of realization of a bivariate first order auto-regressive process $w_i(t) = \left[\begin{array}{c} w_i^1(t) \\ w_i^2(t) \end{array} \right]$ and $\varepsilon_i = \left[\begin{array}{c} \varepsilon_i^1 \\ \varepsilon_i^2 \end{array} \right]$ represents independent measurement errors.

The covariance matrix of measurement errors is defined by $\Sigma_i = \Sigma \otimes I_{n_i}$ and $\Sigma = \left[\begin{array}{cc} \sigma^2_{\varepsilon^1} & 0 \\ 0 & \sigma^2_{\varepsilon^2} \end{array} \right]$.

The covariance function of the bivariate auto-regressive process $w_i(t) = \left[\begin{array}{c} w_i^1(t) \\ w_i^2(t) \end{array} \right]$ is given by $R_i(s,t) = C \times \mathrm{e}^{B|t-s|}$ where $C = \left[\begin{array}{cc} \sigma^2_{w^1} & \sigma_{w^1 w^2} \\ \sigma_{w^1 w^2} & \sigma^2_{w^2} \end{array} \right]$ is the process covariance matrix at $t = s$ and B is a 2×2 matrix such that:

1. the eigenvalues of B have negative real parts and

2. C and $D = -(CB + B'C)$ are positive definite symmetric.

The covariance matrix of random effects is the matrix $G = \begin{bmatrix} G^1 & G^{12} \\ G^{12} & G^2 \end{bmatrix}$. With the assumption that γ_i, W_i and ε_i are mutually independent, it is obvious that $\text{var}(Y_i) = V_i = Z_i G_i Z_i' + R_i + \Sigma_i$.

8.2.2 Generalized linear mixed model with continuous and binary endpoint

Molenberghs and Verbeke (2005, p. 442) give an example of modelling a continuous and a binary endpoint. We will consider this example since we have a similar situation to model in the RSV data. They state that there are two modelling strategies available for the modelling of a continuous and a binary endpoint. Since the joint distribution of a mixed continuous and discrete outcome vector can be expressed as the product of the marginal distribution of one of the responses and the conditional distribution of the remaining response given the former response, one can choose either the continuous or the discrete outcome for the marginal model. The problem with such an approach is that no easy expressions for the association between both endpoints are obtained. Molenberghs and Verbeke (2005) opt for a more symmetric treatment of the two outcome variables. They modelled the case where the surrogate is binary and the true endpoint is continuous. The model that is used is specific for a random-effects logistic regression for repeated measures with residual correlation. They assume the following model formulation. Let $\tilde{S}_i$ be a latent variable of which S_i is the dichotomized version. T_i is the true end point which is continuous. The following model

without random effects can then be assumed.

$$\begin{aligned} T_i &= \mu_T + \beta X_i + \varepsilon_{T_i} \\ \tilde{S}_i &= \mu_S + \alpha X_i + \varepsilon_{S_i} \end{aligned}$$

In general we write,

$$\boldsymbol{Y_i} = \boldsymbol{\mu_i} + \boldsymbol{\varepsilon_i} \tag{8.3}$$

where

$$\boldsymbol{\mu_i} = \boldsymbol{\mu_i}(\boldsymbol{\eta_i}) = \boldsymbol{h}(\boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z_i}\boldsymbol{b_i}) \tag{8.4}$$

$\boldsymbol{\mu}_i$ is specified by means of a GLMM, $\boldsymbol{\varepsilon}_i$ is the residual error structure and assume that $\boldsymbol{b_i} \sim N(\mathbf{0}, D)$. The key relaxing assumption is that the components of the inverse link functions $\boldsymbol{h}$ are allowed to differ with the nature of the various outcomes in $\boldsymbol{Y_i}$. The variance of $\boldsymbol{\varepsilon}$ depends on the mean-variance links of the various outcomes, and can contain, in addition a correlation matrix $R_i(\boldsymbol{\alpha})$ and overdispersion parameters ϕ_i. Now using straightforward derivations, a general first order approximate expression for the variance-covariance matrix of $\boldsymbol{Y_i}$ is:

$$V_i = \mathrm{Var}(\boldsymbol{Y_i}) = \Delta_i Z_i D Z_i^{'} \Delta_i^{'} + \Sigma_i$$

Here,

$$\Delta_i = \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \Big|_{b_i = 0}$$

and

$$\Sigma_i = \Phi_i^{1/2} A_i^{1/2} R_i(\boldsymbol{\alpha}) A_i^{1/2} \Phi_i^{1/2}$$

where A_i is a diagonal matrix containing the variances derived from the generalized linear model specification of Y_{ij} given the random effects $\boldsymbol{b_i} = 0$, therefore the diagonal elements are given by $v(\mu_{ij}|\boldsymbol{b_i} = 0)$. Likewise Φ_i is a diagonal matrix with the overdispersion parameters along the diagonal.

When an outcome is normally distributed the, the overdispersion parameter is σ_i^2 and the variance function is 1. For a binary outcome with logit link, we have

$$\upsilon(\mu_{ij}|\boldsymbol{b_i}=0)=\mu_{ij}(\boldsymbol{b_i}=\mathbf{0})[1-\mu_{ij}(\boldsymbol{b_i}=\mathbf{0})]. \tag{8.5}$$

The evaluation under $\boldsymbol{b_i}=\mathbf{0}$ derives from the Taylor series expansion of the mean components around $\boldsymbol{b_i}=\mathbf{0}$. When an exponential family specification is used for all components, with canonical link, $\Delta_i=A_i$, we can then write:

$$V_i=\text{Var}(\boldsymbol{Y_i})=\Delta_i Z_i D Z_i'\Delta_i'+\Phi_i^{1/2}\Delta_i^{1/2}R_i(\boldsymbol{\alpha})\Delta_i^{1/2}\Phi_i^{1/2}$$

Under the conditional independence $R_i=0$ therefore

$$V_i=\text{Var}(\boldsymbol{Y_i})=\Delta_i Z_i D Z_i'\Delta_i'+\Phi_i^{1/2}\Delta_i\Phi_i^{1/2}.$$

The model

$$\boldsymbol{Y_i}=\boldsymbol{\mu_i}+\boldsymbol{\varepsilon_i}$$

can now be written as

$$\begin{pmatrix} S_i \\ T_i \end{pmatrix}=\begin{pmatrix} \mu_s+\lambda b_i+\alpha X_i \\ \frac{\exp(\mu_T+b_i+\beta X_i)}{1+\exp(\mu_T+b_i+\beta X_i)} \end{pmatrix}+\begin{pmatrix} \varepsilon_{S_i} \\ \varepsilon_{T_i} \end{pmatrix}.$$

Note that the inclusion of the scale parameter λ in the continuous component of an otherwise random intercept model. Note also that the continuous and binary outcomes are measured on different scales. Therefore,

$$Z_i=\begin{pmatrix} \lambda \\ 1 \end{pmatrix},\Delta_i=\begin{pmatrix} 1 & 0 \\ 0 & \upsilon_{i2} \end{pmatrix},\Phi=\begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix},$$

with $\upsilon_{i2}=\mu_{i2}(\boldsymbol{b_i}=0)[1-\mu_{i2}(\boldsymbol{b_i}=\mathbf{0})]$ So,

$$V_i=\begin{pmatrix} \lambda^2 & \upsilon_{i2}\lambda \\ \upsilon_{i2}\lambda & \upsilon_{i2}^2 \end{pmatrix}\tau^2+\begin{pmatrix} \sigma^2 & \rho\sigma\sqrt{\upsilon_{i2}} \\ \rho\sigma\sqrt{\upsilon_{i2}} & \upsilon_{i2} \end{pmatrix}$$

Hence,

$$V_i=\begin{pmatrix} \lambda^2\tau^2+\sigma^2 & \upsilon_{i2}\lambda\tau^2+\rho\sigma\sqrt{\upsilon_{i2}} \\ \upsilon_{i2}\lambda\tau^2+\rho\sigma\sqrt{\upsilon_{i2}} & \upsilon_{i2}^2\tau^2+\upsilon_{i2} \end{pmatrix}$$

The correlation derived from the above model specification equals

$$\rho(\boldsymbol{\beta}) = \frac{v_{i2}\lambda\tau^2+\rho\sigma\sqrt{v_{i2}}}{\sqrt{\lambda^2\tau^2+\sigma^2}\sqrt{v_{i2}^2\tau^2+v_{i2}}}$$

If the model does not have random effects then it can simply be written as:

$$\begin{pmatrix} S_i \\ T_i \end{pmatrix} = \begin{pmatrix} \mu s + \alpha X_i \\ \frac{\exp(\mu_T+\beta X_i)}{1+\exp(\mu_T+\beta X_i)} \end{pmatrix} + \begin{pmatrix} \varepsilon_{S_i} \\ \varepsilon_{T_i} \end{pmatrix}.$$

The correlation $\rho(\boldsymbol{\beta})$ can be simplified if for example there are no random effects or if both endpoints are binary for example in Molenberghs and Verbeke (2005). The example finally concludes with the fact, that the above calculations can be performed with ease for general random effects model or design matrices Z_i and for more than two components, of arbitrary nature and not just continuous and binary.

In the general model, no full joint distribution need to be specified, even when we assume the first one to be normally distributed and the second one to be Bernoulli distributed. We can still leave the specification of the joint moments to be of the second order, by way of the marginal correlation. A full joint specification would need full bivariate model specification, conditional upon the random effects.

Under the conditional independence, the specification of the outcome distribution conditional upon the random effects, together with the normality assumptions made about the random effects, fully specifies the joint distribution.

8.3 Application of joint modelling to the RSV data set

Much of the theory of the generalized linear mixed model (GLMM) has already been covered extensively in Chapter 5 and will not be repeated again

in this chapter. We will consider two approaches to the joint modelling of the RSV status (a binary response) of each child together, with the time interval in days between events, d_{ij}. Let the outcomes d_{ij} be observations from a variable D (continuous or discrete). D is discrete if the interval between observations is measured in terms of the whole number of days and continuous otherwise. The RSV status is a binary outcome variable whilst the d_{ij} can be thought to follow:

1) a Poisson distribution, if the we count the number of days, to the next event, defined as the RSV status of the child or,

2) an Exponential distribution, if the we model the d_{ij} as the time in days between two successive RSV events of the child

Hence we will jointly model the RSV status of each child together with the time interval in days between events, d_{ij} as a generalized linear mixed model. In matrix formulation the model is:

$$\begin{bmatrix} RSV\,status \\ D \end{bmatrix} = \begin{bmatrix} age + prev + actipass + timemonth \\ age + prev + actipass + timemonth \end{bmatrix} + \begin{bmatrix} child \\ child \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

that is, any type of response can be modelled as,

$$\text{Response} = \text{fixed effects} + \text{random effect} + \text{error term}$$

According to Littell et al. (2006, p.199), it is recommended that for unequally spaced repeated measures data, to consider some kind of time series covariance structure, where the correlations of the repeated measurements are assumed to be smaller for observations that are further apart in time. The covariance structures that are suitable are the Unstructured (UN), but this structure can be too general, Compound Symmetry (CS) which assumes that the correlations remain constant, Spatial Power Law SP(POW), Gaussian SP(GAU) and Spherical SP(SPH). The latter three covariance structures

Fit Statistics	
-2 Log Pseudo-Likelihood	96903.11
Generalized Chi-Square	16568.08
Gener. Chi-Square / DF	0.89

Table 8.1: Fit statistics-Exponential distribution

are time series type in which the correlations decline as a function of time. Joint modelling of RSV status and D is considered taking each of the five covariance structures into account.

8.3.1 Fitting D the time between events using an Exponential distribution

The fit statistics are as follows: Different covariance structure models were fitted: Only the compound symmetry (CS) and SP(GAU) covariance struc-

Covariance Structure		Estimate	Standard Error
Compound symmetry	Var(child)	0.002029	0.002487
	CS(child)	-0.00670	.
	Residual(VC)	0.8919	0.009374
Gaussian	Var(child)	0.00000	.
	SP(GAU)(child)	1.000	.
	Residual(VC)	0.8833	0.009164

Table 8.2: Covariance Parameter Estimates-Exponential distribution

tures led to convergence. The models using the UN, SP(POW) and SP(SPH) did not converge and hence are not considered further.

The solution for the fixed effects for all the different covariance structure models are tabulated in Tables 8.3 and 8.4. The results for fitting the different covariance structures are shown in Table 8.2. Only the residual variance seems to be the significant source of variation.

Effect	dist	age	Estimate	Standard Error	DF	t Value	Pr > |t|
dist	Binary		-5.0114	1.3729	18209	-3.65	0.0003
dist	Exponential		0.9965	0.1369	18209	7.28	<.0001
age*dist	Binary	0	-0.9373	1.1535	18209	-0.81	0.4165
age*dist	Exponential	0	-0.8113	0.1218	18209	-6.66	<.0001
age*dist	Binary	1	-0.6738	1.0176	18209	-0.66	0.5079
age*dist	Exponential	1	0.2372	0.1038	18209	2.28	0.0224
age*dist	Binary	2	-0.2942	0.9817	18209	-0.30	0.7644
age*dist	Exponential	2	0.5975	0.09937	18209	6.01	<.0001
age*dist	Binary	3	-0.08718	0.9253	18209	-0.09	0.9249
age*dist	Exponential	3	0.5316	0.09604	18209	5.54	<.0001
age*dist	Binary	4	-0.6824	0.9007	18209	-0.76	0.4487
age*dist	Exponential	4	0.5957	0.09270	18209	6.43	<.0001
age*dist	Binary	5	-2.6078	1.2284	18209	-2.12	0.0338
age*dist	Exponential	5	0.7646	0.08975	18209	8.52	<.0001
age*dist	Binary	6	-1.5948	0.9513	18209	-1.68	0.0937
age*dist	Exponential	6	1.1759	0.08480	18209	13.87	<.0001
age*dist	Binary	7	-2.2402	1.0860	18209	-2.06	0.0391
age*dist	Exponential	7	1.2589	0.07978	18209	15.78	<.0001
age*dist	Binary	8	-0.9934	0.5560	18209	-1.79	0.0740
age*dist	Exponential	8	1.1170	0.06899	18209	16.19	<.0001
age*dist	Binary	9	-0.7400	0.4817	18209	-1.54	0.1245
age*dist	Exponential	9	0.7605	0.06222	18209	12.22	<.0001
age*dist	Binary	10	-0.3260	0.4271	18209	-0.76	0.4452
age*dist	Exponential	10	0.4104	0.05887	18209	6.97	<.0001
age*dist	Binary	11	-0.5660	0.4362	18209	-1.30	0.1944
age*dist	Exponential	11	0.1078	0.05672	18209	1.90	0.0574
age*dist	Binary	12	0	.	.	.	.
age*dist	Exponential	12	0	.	.	.	.
prev*dist	Binary		44.8029	7.5487	18209	5.94	<.0001
prev*dist	Exponential		4.9599	0.9961	18209	4.98	<.0001
actipass*dist	Binary	0	2.2277	0.1666	18209	13.37	<.0001
actipass*dist	Exponential	0	-0.2488	0.02515	18209	-9.89	<.0001
actipass*dist	Binary	1	0	.	.	.	.
actipass*dist	Exponential	1	0	.	.	.	.
timemonth*dist	Binary		-0.04742	0.09909	18209	-0.48	0.6323
timemonth*dist	Exponential		0.09877	0.009526	18209	10.37	<.0001

Table 8.3: Solution for the fixed effects-Exponential distribution

The prev, timemonth and actipass variables are significant as well as the majority of the age variable levels i.e.1-4 and 6-10 at the Exponential distribution.

The type III tests for the fixed effects are given as: One can infer from the

Effect	Num. DF	Den. DF	F-value	Pr>F
dist	2	18209	159.49	<.0001
age*dist	24	18209	53.22	<.0001
prev*dist	2	18209	30.02	<.0001
actipass*dist	2	18209	138.05	<.0001
timemonth*dist	2	18209	53.83	<.0001

Table 8.4: Type III tests for the fixed effects-Exponential distribution

type III tests in Table 8.4, that all the fixed effects of age, prev, actipass and timemonth are significant in the joint model.

8.3.2 Fitting D the time between events using a Poisson distribution

In the joint model the variable dt was regarded as counts in days, hence a Poisson distribution is used as the distribution of choice. The fit statistics are shown in Table 8.5:

Fit Statistics	
-2 Log Pseudo-Likelihood	103676.0
Generalized Chi-Square	75419.98
Gener. Chi-Square / DF	4.06

Table 8.5: Fit statistics-Poisson distribution

Different covariance structure models were fitted:

Covariance Structure		Estimate	Standard Error
Unstructured	UN(1,1)	0.06807	0.009169
	Residual(VC)	4.0596	0.04290
Compound symmetry	Var(child)	0.000062	0.009169
	CS(child)	0.06801	.
	Residual(VC)	4.0596	0.04290
Power	Var(child)	0.06807	0.009169
	SP(POW)(child)	0.9999	.
	Residual(VC)	4.0596	0.04290

Table 8.6: Covariance Parameter Estimates-Poisson distribution

Now the estimated variance component for the child random effect is 0.06807. The models using the SP(GAU) and SP(SPH) did not converge and are hence not suitable covariance structures. The solution for the fixed effects for all the different covariance structure models are tabulated in Table 8.7

Effect	dist	age	Estimate	Standard Error	DF	t Value	Pr > |t|
dist	Binary		-5.3831	2.8540	18209	-1.89	0.0585
dist	Poisson		1.3822	0.1488	18209	9.29	<.0001
age*dist	Binary	0	-0.6522	2.4089	18209	-0.27	0.7866
age*dist	Poisson	0	-1.0309	0.1673	18209	-6.16	<.0001
age*dist	Binary	1	-0.3465	2.1082	18209	-0.16	0.8694
age*dist	Poisson	1	0.03291	0.1209	18209	0.27	0.7855
age*dist	Binary	2	-0.01106	2.0328	18209	-0.01	0.9975
age*dist	Poisson	2	0.3709	0.1103	18209	3.36	<.0001
age*dist	Binary	3	0.1883	1.9154	18209	0.10	0.9217
age*dist	Poisson	3	0.3313	0.1010	18209	3.28	<.0001
age*dist	Binary	4	-0.4500	1.8554	18209	-0.24	0.8094
age*dist	Poisson	4	0.4072	0.09160	18209	4.45	<.0001
age*dist	Binary	5	-2.4181	2.5610	18209	-0.94	0.3451
age*dist	Poisson	5	0.5885	0.08242	18209	7.14	<.0001
age*dist	Binary	6	-1.4204	1.9628	18209	-0.72	0.4693
age*dist	Poisson	6	1.0352	0.07180	18209	14.42	<.0001
age*dist	Binary	7	-2.1365	2.2638	18209	-0.94	0.3453
age*dist	Poisson	7	1.1440	0.06185	18209	18.50	<.0001
age*dist	Binary	8	-0.8704	1.1438	18209	-0.76	0.4467
age*dist	Poisson	8	1.0077	0.05176	18209	19.47	<.0001
age*dist	Binary	9	-0.6550	0.9993	18209	-0.66	0.5122
age*dist	Poisson	9	0.6929	0.04498	18209	15.40	<.0001
age*dist	Binary	10	-0.2460	0.8869	18209	-0.28	0.7815
age*dist	Poisson	10	0.4000	0.04038	18209	9.90	<.0001
age*dist	Binary	11	-0.5428	0.9093	18209	-0.60	0.5506
age*dist	Poisson	11	0.09986	0.03791	18209	2.63	0.0084
age*dist	Binary	12	0	.	.	.	.
age*dist	Poisson	12	0	.	.	.	.
prev*dist	Binary		44.3812	15.6472	18209	2.84	<.0001
actipass*dist	Poisson	0	2.2521	0.3486	18209	6.46	<.0001
actipass*dist	Poisson	1	.	.	.	.	.
actipass*dist	Poisson	0	-0.2113	0.01658	18209	-12.75	<.0001
actipass*dist	Poisson	1	.	.	.	.	.
timemonth*dist	Poisson		-0.01431	0.2049	18209	-0.07	0.9443
timemonth*dist	Poisson		0.07461	0.01091	18209	6.84	<.0001

Table 8.7: Solution for the fixed effects-Poisson distribution

The actipass and prev variables are significant as well as the majority of the age variable levels i.e.1-10 at the Poisson distribution. The type III tests for the fixed effects are given as:

Effect	Num. DF	Den. DF	F-value	Pr>F
dist	2	18209	191.92	<.0001
age*dist	24	18209	124.61	<.0001
prev*dist	2	18209	20.23	<.0001
actipass*dist	2	18209	102.32	<.0001
timemonth*dist	2	18209	23.46	<.0001

Table 8.8: Type III tests for the fixed effects-Poisson distribution

One can infer from the above type III tests in Table 8.8, that all the fixed effects of age, prev, actipass and timemonth are significant in the joint model.

8.4 Conclusion

The estimates of the fixed effects for modelling Binary-Poisson and Binary-Exponential models gave very similar results. The variables of age, prev, actipass and timemonth are all significant at the 5% level and are both suitable explanatory variables for the response variable which is the RSV status of a child and the time in days between events, dt, between the visits as a joint model. The joint model has the distinct advantage of modelling two response variables and this allows the experimenter or researcher a degree of flexibility.

Chapter 9

Missing Data

9.1 Introduction

In studies where data is acquired from individuals over time or other study designs, the problem of estimating and handling missing data is bound to surface. Some examples of situations where missing data may arise include survey non-response due for example people moving out of a city, death, missing data in longitudinal studies due to censoring, dropout and missing data by design. There are various methods of dealing with missing data, that range from simple classical methods to model based methods. These methods must be fully understood theoretically before they can be used practically. Furthermore each method is based upon a specific missing data mechanism but one needs to realize that at the heart of the missing value problem it is impossible in practice to identify the missingness mechanisms. Little (1992) gives a detailed account on the methods of handling missing data that had been used thus far. These methods of estimating missing data have progressively advanced over the past years. Up to mid 1970's the methods that were popularly used included complete case analysis, imputation and maximum

likelihood estimation. During the Mid 1970's-1980's, the methods included maximum likelihood estimation to a broad range of problems, the expectation maximization (EM) algorithm and multiple imputation. In the 1990's, maximum likelihood estimation was applied to harder problems. This period also saw extensions of the EM algorithm to the various forms such as the stochastic EM algorithm (SEM), the expectation conditional maximization algorithm (ECM) and the stochastic expectation conditional maximization algorithm (SECM). In the same period Bayes simulation methods (Markov Chain Monte Carlo methods (MCMC) and data augmentation) were also developed. The future for missing data analysis in general, would require the development and refinement of computational tools, diagnostics and the need for non ignorable non-response analysis methods where the missing data mechanism depends on the missing values. As already stated approaches to deal with estimating missing data range from simple classical to complex modern methods that are still under development. For example, Molenberghs and Kenward have recently authored a book dedicated to missing data in clinical studies (Molenberghs and Kenward, 2007).

9.2 The Longitudinal Data Setting

The following is a summary of the concepts of missing data or non-response from Molenberghs and Verbeke (2005, Chapters 26 and 27). They state that in a longitudinal setting, each unit is measured on several occasions and hence it is not unusual in practice for some sequences of measurements to terminate early for reasons outside of the control of the experimenter or investigator, and any unit so affected is called a dropout. Early work on missing values was largely concerned with algorithmic and computational solutions to

the induced lack of balance or deviations from the intended designs. More recently, general algorithms such as the expectation-maximization (EM) by Dempster, Laird and Rubin (1977) and data imputation and augmentation procedures, given in Rubin (1987) combined with powerful computing resources have largely provided a solution to the aspect of this problem. There still remains the very difficult and important question of assessing the impact of missing data on subsequent statistical inference.

Associated with missing data is the missing data patterns, such as, general patterns, monotone patterns and univariate non-response patterns just to name a few and more importantly the missing data mechanisms given in Little and Rubin (2002, p.12 and 1987, Chapter 6). These are:

- missing at random (MAR) if conditional on the observed data the missingness is independent of the unobserved measurements,
- missing completely at random (MCAR) if the missingness is independent of both unobserved and observed data and
- not missing at random or non-random (NMAR) if the process is neither MCAR or MAR. These mechanisms have been classified by Rubin (1976) as well.

In the context of likelihood inference, and when the parameters describing the measurement process are functionally independent of the parameters describing the missingness process, then MAR and MCAR are *ignorable* while a NMAR process is *non-ignorable*.

The mathematical description of the missing data mechanisms is discussed in Section 9.4. The problem of missing data is really about estimating the observations that have gone missing for various reasons. Historically Afifi and Elashoff (1966) and Hartley and Hocking (1971) have given a taxonomy

of methods of estimating missing data. Orchard and Woodbury (1972) and Demptser et al. (1977) have all led to the following categories for analyzing missing data viz. procedures based on completely recorded units, imputation based procedures, weighting procedures and model based procedures.

Many missing data methods are formulated as selection models as those in Little and Rubin (1987) as opposed to the pattern mixture-modelling in Little (1993, 1994). A selection model factors the joint distribution of the the measurement and the response mechanisms into the marginal measurement distribution and the response distribution, conditional on these measurements. This is intuitively appealing since the marginal measurement distribution would be of interest also with the complete data. Within this framework, Little and Rubin (1987) develop their taxonomy in the selection model setting. Hence parameterizing and making inference about the effect of treatment and its evolution over time is straightforward in the selection model context.

In the clinical trial setting, the standard methodology used to analyze longitudinal data subject to missing data or non-response is based on methods such as *last observation carried forward* (LOCF), *complete case analysis* (CC) or simple forms of imputation. These methods are used without questioning the possible influence of these assumptions on the final results. Many authors that have written about these methods and historically, Heyting, Tolboom and Essers (1992) give the earliest account. Mallinckrodt et al. (2003) and Lavori, Dawson and Shera (1995) propose the direct likelihood and multiple imputation methods to deal with incomplete longitudinal data while Siddiqui and Ali (1998) compare direct likelihood and LOCF methods. LOCF and CC methods are based on strong assumptions and in particular even the strong MCAR assumtpion does not suffice to guarantee that a

LOCF analysis is valid. However under MAR the likelihood based analysis can produce a valid analysis without the need for modelling the dropout process and Verbeke and Molenberghs (2000) state that as a result of this linear and generalized linear mixed models can be used. Such an analysis enjoys wider validity than the simpler methods and they are simple to conduct.

Thus, longitudinal data modelling in the presence of missing data should shift away from the *ad hoc* methods and focus on likelihood based ignorable analyses instead. Molenberghs and Verbeke (2005, Chapters 26 and 27) promote the use of direct likelihood methods and demote the use of LOCF and CC approaches while reflecting on the status of MNAR approaches.

9.3 A Taxonomy

The following taxonomy by Rubin (1976) and Little and Rubin (1987) and recently by Molenberghs and Verbeke (2005) is adopted. We assume that for subject i in a study, a sequence of measurements Y_{ij} is designed to be measured at occasions $j = 1, \ldots, n_i$. As before, the outcomes are grouped into a vector $\boldsymbol{Y_i} = (Y_{i1}, \ldots, Y_{in_i})'$. We now define, for each occasion j, an indicator

$$R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

Thus the missing data indicator R_{ij} are grouped into a vector $\boldsymbol{R_i}$ which is the same length as $\boldsymbol{Y_i}$. We then partition $\boldsymbol{Y_i}$ into two sub-vectors such that $\boldsymbol{Y_i^o}$ is the sub-vector containing those Y_{ij} for which $R_{ij} = 1$ and $\boldsymbol{Y_i^m}$ contains the remaining components. The sub-vectors are referred to as the observed and missing components. This partition is allowed to differ with subject and $\boldsymbol{Y_i^o}$ can contain components which are measured later than occasions at which components of $\boldsymbol{Y_i^m}$ ought to have been measured. Complete data refers to

the vector $\boldsymbol{Y_i}$ of scheduled measurements. This is the outcome vector that would have been recorded if there were no missing data. The missing data indicators are assembled into the vector $\boldsymbol{R_i}$ and the process generating $\boldsymbol{R_i}$ is referred to as the missing data process. The full data $(\boldsymbol{Y_i}, \boldsymbol{R_i})$ consists of the complete data together with the missing data indicators. It is obvious that unless all the components of $\boldsymbol{R_i}$ are equal to 1, the full data components are never all observed. Then, the observed data refer to $\boldsymbol{Y_i^o}$ and the missing data to $\boldsymbol{Y_i^m}$. One would then observe the measurements $\boldsymbol{Y_i^o}$ together with the dropout indicators $\boldsymbol{R_i}$. When missingness is restricted to dropout or attrition, we can replace the vector $\boldsymbol{R_i}$ by a scalar variable D_i, the dropout indicator. Then in this case each vector $\boldsymbol{R_i}$ is of the form $(1, \ldots, 1, 0, \ldots, 0)$ and we define the scalar dropout indicator as

$$D_i = 1 + \sum_{j=1}^{n_i} R_{ij} \tag{9.1}$$

For an incomplete sequence, D_i denotes the occasion at which dropout occurs. For a complete sequence, $D_i = n_i + 1$. In both cases D_i indicates one plus the length of the measurement sequence, whether complete or incomplete. Dropout or attrition is a particular monotone pattern of missingness. In order to have monotone missingness there has to exist a permutation of the measurement occasions where a measurement earlier in the permuted sequence is observed for at least those subjects that are observed at later measurements. For this definition to be meaningful, we need to have a balanced design in the sense of a common set of measurement occasions. Other patterns are called non-monotone (Molenberghs and Kenward, 2007).

9.4 Missing data frameworks

As stated earlier, we will take the following derivation and notation of the selection, pattern mixture and shared parameter frameworks from Molenberghs and Verbeke (2005, pp 484-488) and Molenberghs and Kenward (2007). When data are incomplete due to a stochastic mechanism one starts from the full density given as

$$f(\boldsymbol{y_i},\, \boldsymbol{r_i}|X_i, Z_i, W_i, \boldsymbol{\theta},\boldsymbol{\psi}) \tag{9.2}$$

where X_i, Z_i and W_i are design matrices for the fixed effects, random effects and missing data process and where $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are vectors that parameterize the joint distribution. We use $\boldsymbol{\theta}$=$(\boldsymbol{\beta}',\boldsymbol{\alpha}')'$ and $\boldsymbol{\psi}$ to describe the measurement and missingness processes, relatively where $\boldsymbol{\beta}$ is the fixed effects parameter vector and $\boldsymbol{\alpha}$ assembles variance components and/or association parameters. The term "selection model" originates from Heckman (1976) in an econometric literature setting. The selection model factorization then equals

$$f(\boldsymbol{y_i},\, \boldsymbol{r_i}|X_i, Z_i, W_i, \boldsymbol{\theta},\psi) = f(\boldsymbol{y}_i|X_i, Z_i, \boldsymbol{\theta})f(\boldsymbol{r}_i|\boldsymbol{y}_i, W_i, \boldsymbol{\psi}) \tag{9.3}$$

where the first factor is the marginal density of the measurement process and the second factor is the density of the missingness process, conditional on the outcomes. Factor $f(\boldsymbol{r_i}|\boldsymbol{y_i}, W_i,\boldsymbol{\psi})$ describes one's self-selection mechanism to either continue or leave the study.

Little (1993, 1995), Molenberghs, Kenward and Lesaffre (1997) give an alternative family to the selection models called the "pattern mixture models" which are based on the following factorization

$$f(\boldsymbol{y_i},\, \boldsymbol{r_i}|X_i, Z_i, W_i, \boldsymbol{\theta},\boldsymbol{\psi}) = f(\boldsymbol{y_i}|\boldsymbol{r_i}, X_i, Z_i, \boldsymbol{\theta})f(\boldsymbol{r_i}|, W_i, \boldsymbol{\psi}) \tag{9.4}$$

The pattern mixture model allows for a different response model for each pattern of missing values, the observed data being a mixture of these weighted

by the probability of each missing value or dropout pattern and this model has got particular distinct advantages.

The third family of models studied earlier by Wu and Carroll (1988) and Wu and Bailey (1988, 1989) is referred to as the "shared parameter models" given as

$$f(\boldsymbol{y_i},\, \boldsymbol{r_i}|X_i, Z_i, W_i, \boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{b_i}) = f(\boldsymbol{y_i}|\boldsymbol{r_i}, X_i, Z_i, \boldsymbol{\theta},\boldsymbol{b_i})f(\boldsymbol{r_i}|Z_i, W_i, \boldsymbol{\psi},\boldsymbol{b_i}) \quad (9.5)$$

where $\boldsymbol{b_i}$ is a vector of random effects of which one or more components are shared between both factors. sensible assumption of this model is that $\boldsymbol{Y_i}$ and $\boldsymbol{R_i}$ are independent, given the random effects $\boldsymbol{b_i}$. The random effects can be used to define a linear, generalized linear or non-linear mixed effects model. The same vector can be used to define the missing data process. The natural parameters of selection models, pattern mixture models and shared parameter models have a different meaning, and transforming one probability model into one of the other framework is in general not straight-forward, not for normal measurement models but even less so in the general case.

9.5 Missing data mechanisms

Rubin's (1976) taxonomy is based within the selection modelling framework given above, as

$$f(\boldsymbol{r_i}|\boldsymbol{y_i},W_i,\boldsymbol{\psi})=f(\boldsymbol{r_i}|\boldsymbol{y_i^o},\boldsymbol{y_i^m}, W_i,\boldsymbol{\psi})$$

This classification also in Little and Rubin (1987) essentially distinguishes settings in which important simplifications of this process are possible. Thus using this specification on the different types of missingness can then be distinguished.

9.5.1 Missing Completely at Random (MCAR)

Under MCAR the probability of an observation being missing is independent of the responses, thus

$$f(\boldsymbol{r_i}|\boldsymbol{y_i},W_i,\boldsymbol{\psi})=f(\boldsymbol{r_i}|,W_i,\boldsymbol{\psi})$$

The selection model then reduces to

$$f(\boldsymbol{y_i},\ \boldsymbol{r_i}|X_i,Z_i,W_i,\boldsymbol{\theta},\boldsymbol{\psi})=f(\boldsymbol{y_i}|X_i,Z_i,\boldsymbol{\theta})f(\boldsymbol{r_i}|W_i,\boldsymbol{\psi})$$

implying that both the components are independent. The implication is that the joint distribution of $\boldsymbol{y_i^o}$ and $\boldsymbol{r_i}$ becomes

$$f(\boldsymbol{y_i^o},\ \boldsymbol{r_i}|X_i,Z_i,W_i,\boldsymbol{\theta},\boldsymbol{\psi}) = f(\boldsymbol{y_i^o}|X_i,Z_i,\boldsymbol{\theta})f(\boldsymbol{r_i}|W_i,\boldsymbol{\psi}) \tag{9.6}$$

Under MCAR the observed data can be analyzed as though the pattern of missing values were predetermined. In whatever way the data is analyzed, the process(es) generating the missing values can be ignored

9.5.2 Missing at Random (MAR)

Under the MAR mechanism, the probability of an observation being missing is conditionally independent of the unobserved data, given the values of the observed data,

$$f(\boldsymbol{r_i}|\boldsymbol{y_i},W_i,\boldsymbol{\psi})=f(\boldsymbol{r_i}|\boldsymbol{y_i^o},W_i,\boldsymbol{\psi})$$

and the joint distribution of the observed data can now be partioned as

$$f(\boldsymbol{y_i},\ \boldsymbol{r_i}|X_i,Z_i,W_i,\boldsymbol{\theta},\boldsymbol{\psi})=f(\boldsymbol{y_i}|X_i,Z_i,\boldsymbol{\theta})f(\boldsymbol{r_i}|\boldsymbol{y_i^o},W_i,\boldsymbol{\psi})$$

and hence at the observed data level

$$f(\boldsymbol{y_i^o},\ \boldsymbol{r_i}|X_i,Z_i,W_i,\boldsymbol{\theta},\boldsymbol{\psi}) = f(\boldsymbol{y_i^o}|X_i,Z_i,\boldsymbol{\theta})f(\boldsymbol{r_i}|\boldsymbol{y_i^o},W_i,\boldsymbol{\psi}) \tag{9.7}$$

The MAR assumption leads to considerable simplification in the issues surrounding the analysis of incomplete longitudinal data. It is however rare in practice for an investigator to be able to justify its adoption.

9.5.3 Missing not at Random (MNAR)

Under MNAR, neither the MAR and MCAR hold. The probability here, of a measurement being missing depends on the unobserved data. No simplification of the joint distribution is possible and the joint distribution of the observed measurements and the missingness process is written as

$$f(\boldsymbol{y_i^o}, \boldsymbol{r_i}|X_i, Z_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = \int f(\boldsymbol{y_i}|X_i, Z_i, \boldsymbol{\theta}) f(\boldsymbol{r_i}|\boldsymbol{y_i}, W_i, \boldsymbol{\psi}) d\boldsymbol{y_i^m} \qquad (9.8)$$

Inferences can only be made by making further assumptions about which the observed data alone carry no information. Ideally the choice of such assumptions should be guided by external information but the degree to which this is possible in practice varies greatly. Such models can be formualted within the selection models, pattern-mixture models and the shared-parameter models. Molenberghs and Verbeke (2005) state that the differences between these families under the MNAR case is especially important and to different but complimentary views of the missing value problem. Little (1995), Hogan and Laird (1997) and Kenward and Molenberghs (1999) give detailed reviews of these differences. Recently, Molenberghs and Kenward (2007) have added this literature addressing specifically the problem of missing data in clinical studies although their methods are still applicable to other data settings.

9.5.4 Ignorability

The MAR assumption states that once appropriate account is taken of what we have observed, there remains no dependence on unobserved data, at least

in terms of the probability model. We should as a consequence expect much of the missing value problem to disappear under the MAR mechanism and this is in fact the case. This can be shown through the consideration of the likelihood but we will not consider this, here.

9.6 Simple Methods

We will now consider briefly the more simple methods such as LOCF, CC and Buck's method. The LOCF and CC methods for handling missing data are very popular but make strong and unrealistic assumptions that can impact negatively on statistical inference. For the validity of many of these simple methods, MCAR is required. For specific methods such as LOCF, MCAR is necessary but not sufficient. Molenberghs and Verbeke (2005, pp. 490-491) state that serious attention needs to be given to the views for the measurement model on one hand and the philosophy adopted for the missingness model on the other hand. They then state the following:

Model for Measurements.

A choice has to be made regarding the modelling approach and several views are possible:

i) *View 1*: One can choose to analyze the entire longitudinal profile irrespective of whether interest, is to focus on the entire profile or on a specific time point. In the latter case one would make inferences about such an occasion using the posited model.

ii) *View 2*: One states the scientific question in terms of the outcome at a well defined point in time. Several choices are possible.

iii) *View 2a*: The scientific question is defined in terms of the the last planned occasion. In this case, one can either accept the dropout as it is or use one or other strategy (eg. imputation) to incorporate the missing outcomes.

iv) *View 2b*: One can choose to define the question and the corresponding analysis in terms of the last observed measurement.

Method for handling missingness.

A choice has to be made regarding the modelling approach for the missingness process. Under certain assumptions this process can be ignored (eg. a likelihood-based ignorable analysis). Some simple methods such as CC and LOCF, do not explicitly address the missingness process either.

The measurement model will depend on whether or not a full longitudinal analysis is done and when the longitudinal analysis is deemed necessary. As is the case of the RSV data (current study), then the choice depends on the nature of the outcome where options include the linear, generalized linear mixed models and generalized estimating equations.

9.6.1 Complete Case Analysis (CC)

Molenberghs and Verbeke (2005, pp 492-493) state that the complete case analysis includes only those cases for which all measurements were recorded. This method has the obvious advantage that it is simple to describe and almost any software can be used because there are no missing data. This method however suffers from several drawbacks. Firstly, there is nearly always a substantial loss of information and the impact on power and precision may be dramatic. Secondly, severe bias can result when the missingness mechanism is MAR and not MCAR. Furthermore, should an estimator be

consistent in the complete data problem, then the derived complete case analysis is consistent only if the missingness process is MCAR. A CC analysis can be used when views 1 and 2, stated previously, are adopted.

An alternative way to obtain a data set on which complete data methods can be used is to fill in rather than to delete. However concern has been raised regarding imputation strategies and the user of imputation strategies faces several dangers. Molenberghs and Verbeke (2005) refer to Little and Rubin (1987) who show that the application of imputation could be considered acceptable in a linear model with one fixed effect and one error term but that it is generally unacceptable for hierarchical models, split plot designs, repeated measures with a complicated error structure, random effects and mixed effects models.

9.6.2 Last Observation Carried Forward (LOCF)

As the name suggests, this method makes use of the last recorded value under a variable and this value is substituted wherever there are missing values under that variable i.e. whenever a value is missing, the last one is substituted. This method is used extensively in clinical trials and longitudinal studies as stated by Molenberghs and Verbeke (2005, p. 493), Molenberghs et al. (2002) and Heyting et al. (1992) give insight as to why the LOCF method is not suitable to use in estimating missing data in a clinical trial or a longitudinal study (see also Molenberghs and Kenward, 2007, for a more recent analysis). Firstly, since LOCF can be regarded as a form of imputation one has to assume that it is plausible that a subject's measurements do not change from the moment of dropout onwards. In a clinical trial setting, one might believe that the response profile changes as soon as a patient goes off treatment and even it would flatten. The constant profile assumption is

even stronger. Hence the LOCF method fails. Secondly, LOCF artificially increases the amount of information on the data, by treating imputed and actually observed values on an equal footing. This is especially true if the longitudinal view is taken. Molenberghs and Verbeke (2005) show that all the features of a linear mixed model can be disproportionately affected. LOCF will suffice to validate view 2b stated earlier.

9.6.3 Unconditional Mean Imputation

Little and Rubin (1987) state that this technique involves substituting a variable's mean value computed from available cases to fill in missing data values on the remaining cases. This option appears in several SPSS procedures. This method is not well suited for discrete outcome responses, such as binary outcomes but is suitable for continuous data. In addition statistical models such as the linear mixed model are distorted by employing unconditional mean imputation methods. Thus much care need to be exercised when using such a method.

9.6.4 Bucks Method or Conditional Mean Imputation for multivariate data

An alternative quick method is to impute the missing data in one of various ways and then to proceed with standard statistical analysis on the "filled in" data set. The simplest of these is the (unconditinal) mean imputation, where the missing x_{ij} values are replaced by the $\overline{x}_j$, the mean of the observed values for the j^{th} variable. Although this form of imputation preserves the variables means, Little and Rubin (1987) state that the variances and covariances are biased towards zero, as would be expected when imputing values from the

centre of a distribution, but it does lead to a positive semidefinite matrix. The factor of underestimation of the sample variance of the j^{th} variable can be calculated as

$$\frac{n^{(j)} - 1}{n - 1}$$

where $n^{(j)}$ is the number of complete cases for the j^{th} variable.

An improved imputation scheme is to impute the means that are conditional on the observed values in each case. If $\boldsymbol{x}$ is multivariate normally distributed with mean, $\boldsymbol{\mu}$ and variance-covariance matrix, $\boldsymbol{\Sigma}$, then the missing values in each case can be linearly regressed on the observed variables with regression coefficients which can be expressed easily in terms of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Buck (1960) proposed that one first estimates $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from the complete cases, then uses these to calculate the regression coefficients for each case. Finally, substitution of the observed values into the regression line yields predictions for the missing values. This method performs well if the data is MCAR and if the normality assumption of regression is reasonable, but must be used with care if prediction from the prediction line involves extrapolation. If $\boldsymbol{x}$ is not multivariate normally distributed then Buck's method can still be used but with the added assumption that all the regressions between the variables are linear. Although the means are reasonably well estimated the variance-covariance structure is underestimated, although this is less sever than when imputing unconditional means (Little, 1992). One of the limitations of Buck's method arises when at least one of the variables is categorical and not fully observed, as the linear regression may yield predictions which are beyond the scope of the categorical variable. An alternative imputation technique which is often used in survey environments is where imputations are randomly selected from a distribution of plausible values, rather than only from the centre of the distribution. One way to achieve this, as Little

and Rubin (1987) states is to add a suitable random perturbation to the conditional mean. Various other imputation methods also exist including several forms of hot and cold deck imputation. Hot deck imputation lends itself most readily to purely categorical data, since it usually involves imputing values drawn from similar responding cases (Tanner, 1993). Finding these "similar" cases when the data is purely categorical can be complex, but when dealing with continuous variables achieving this is nearly impossible and highly subjective. Cold deck imputation is specific to sample surveys as it involves imputing values obtained from another source of information, most often from a similar previously performed survey.

Little (1992) points out that it is also possible to impute missing values via principle components analysis (PCA), but like the rest of these simple methods, its effectiveness is affected by many aspects. Huisman (2000) compared several simple methods for discrete data and points out that together with the proportion of missing data and the missing data mechanism, when the data set involves categorical variables the scale (length and number of response options) also affects the performance of these simple methods.

9.6.5 Healy-Westmacott procedure

An alternative simple procedure for dealing with missing data, in a single dependent variable, is a simple iterative scheme, proposed by Healy and Westmacott (1956) and consists of the following steps:

*S*tep 1: Impute trial values for missing data

*S*tep 2: Perform standard complete-data analysis to obtain model parameters

*S*tep 3: Use the model parameters to predict missing values

*S*tep 4: Substitute the predicted values for missing values

*S*tep 5: Repeat Steps 2-4 until the missing values do not change considerably

The complete case means of the variables are often used as the initial trial values and if multiple regression is required, the model parameters of interest would be the regression coefficients. An alternative way to monitor convergence would be through the residual sum of squares.

It can be shown that both Buck's method and the Healy-Westmacott procedure are closely related to the EM algorithm which will be explored in the next section. In particular the same idea underlies both the Healy-Westmacott procedure and the EM algorithm, that is, that by exploiting the simplicity of the computations if all the cases were completely observed, the more complex computations involving the incomplete cases can be avoided (McLachlan and Krishnan, 1997).

In general all of the above simple techniques are not necessarily recommended as once the missing values have been imputed, the variability due to those imputations is ignored. This is evident after standard statistical tests are performed, sometimes yielding misleading results, as stated by Schafer (1997) where p-values and standard errors are sometimes misleading. They also tend to be very sensitive to any violation of the MCAR assumption, as assumption which in practice is rarely completely valid.

9.7 The Expectation-Maximization (EM) Algorithm

The Expectation-Maximization algorithm has appeared in scientific literature dating as far back as the 1920's. McKendrick (1926) considers a medical application of a method that has aspects in common with the EM algorithm and other specific applications of it appeared in Hartley (1958) as well as in Beale and Little (1975). However it was in 1976 when a paper by Dempster, Laird

and Rubin was read before the Royal Statistical Society and published in the following year that the EM algorithm was named, formulated in a general context, applied to various applications and its basic statistical properties established. The Dempster, Laird and Rubin (1977) paper is now amongst the six most cited statistical papers in the world (Stigler, 1994). By 1992 over one thousand journal articles had been citing the paper (Meng and Pedlow, 1992). Ryan and Woodall (2005) state that out of the 25 most cited papers, the Dempster, Laird and Rubin (1977) paper is ranked as the 11^{th} with approximately 492 citations per year.

The work by Dempster, Laird and Rubin (1977) also suggested that missing data should be seen as a source of variation that is to be averaged over instead of something that was best removed as quickly as possible from the analysis (Schafer and Olsen 1998). Meng (1997) looks at the link between the EM algorithm and medical studies by linking McKendrick's (1926) work to the EM algorithm. Little and Rubin (2002) make reference to Meng and Pedlow (1992) where they give a wide range of problems that can be solved by the EM algorithm and includes ML for problems not usually considered to involve missing data, such as variance component estimation and factor analysis. McLachlan and Krishnan (1997) state that nowadays the EM algorithm has increasingly found applications in AIDS epidemiology, neural networks, medical imaging, dairy science and genetics. At the same time Meng and Van Dyk (1997) extended the EM algorithm by means of simulation and Monte Carlo methods showing the connection of the extensions of it with stochastic algorithms for missing data that is now at the forefront of research.

Its widespread popularity and use is due mainly to its computational simplicity and stability as well as its conceptual appeal as it solves a complex incomplete-data problem by repeatedly solving easier complete data problems. Let $\boldsymbol{X}$, $\boldsymbol{X_{obs}}$ and $\boldsymbol{X_{mis}}$ denote the complete data, observed data and the missing data of a measurement process. The EM algorithm capitalizes on the interdependence between $\boldsymbol{X_{mis}}$ and $\boldsymbol{\theta}$, since $\boldsymbol{X_{mis}}$ contains information relevant to estimating $\boldsymbol{\theta}$ and $\boldsymbol{\theta}$ in turn helps to find likely values of $\boldsymbol{X_{mis}}$. Thus in a nutshell the EM algorithm "fills in " $\boldsymbol{X_{mis}}$ based on an estimate of $\boldsymbol{\theta}$, re-estimates $\boldsymbol{\theta}$ based on $\boldsymbol{X_{mis}}$ and repeatedly performs these two steps until $\boldsymbol{\theta}$ has met some pre-specified convergence criteria. Note that $\boldsymbol{\theta}$ is a vector that parameterizes the measurement process.

9.7.1 The Theory of the Expectation-Maximization (EM) Algorithm

In this section, we define $\mathbf{X}$ to be a $n \times p$ matrix of data which is not fully observed. Thus, $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{mis})$ where $\mathbf{X}_{obs}$ denotes the observed values and $\mathbf{X}_{mis}$ the missing values, with the n rows corresponding to the observational units or cases and the p columns corresponding to the variables. The probability density function of the complete data, under the assumption that the rows are independently and identically distributed, is

$$P(\mathbf{X}|\theta) = \prod_{i=1}^{n} f(\mathbf{x}_i|\theta)$$

where θ is a vector of unknown parameters of the distribution and $f(\mathbf{x}_i|\theta)$ is the probability density function of the i^{th} row, where $\mathbf{x}_i$ denotes the i^th row written as a column vector. In keeping with section 9.3 and 9.4, more formally both MAR and MCAR can be defined by introducing $\boldsymbol{R}$, a $n \times p$ matrix of indicator variables whose elements are either 0 or 1 depending on

whether the corresponding elements of $\boldsymbol{X}$ are missing or observed. Following Rubin (1976), it is reasonable to assume that the missing data mechanism, and therefore $\boldsymbol{R}$, depends on $\boldsymbol{X}$ as well as some unknown parameters $\boldsymbol{\zeta}$, and thus has probability density function $P(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{\zeta})$. MAR means that this distribution depends on the data $\boldsymbol{X}$ only through observed values, or mathematically

$$P(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{\zeta}) = P(\boldsymbol{R}|\boldsymbol{X}_{\boldsymbol{obs}}, \boldsymbol{\zeta}) \qquad \forall \;\; \boldsymbol{X}_{\boldsymbol{mis}}. \tag{9.9}$$

Similarly MCAR means that the distribution of $\boldsymbol{R}$ does not depend on either the observed or missing values of $\boldsymbol{X}$, that is

$$P(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{\zeta}) = P(\boldsymbol{R}|\boldsymbol{\zeta}) \qquad \forall \;\; \boldsymbol{X}. \tag{9.10}$$

In order to utilize maximum likelihood estimation (ML) the log- likelihood of the observed data is required, $\ell(\boldsymbol{\theta}|\boldsymbol{X}_{obs})$, since under the ignorability assumption one does not need to consider the model of $\boldsymbol{R}$ nor the nuisance parameter, $\boldsymbol{\zeta}$ when making likelihood inferences about $\boldsymbol{\theta}$ (Rubin, 1976). Following the arguments given by Rubin (1976), Schafer (1997) and Little and Rubin (1987, 2002), since the data consists of both $\boldsymbol{X}_{obs}$ and $\boldsymbol{R}$, the probability distribution of the entire data is actually given by

$$\begin{aligned} P(\boldsymbol{X}_{\boldsymbol{obs}}, \boldsymbol{R}|\boldsymbol{\theta}, \boldsymbol{\zeta}) &= \int P(\boldsymbol{X}_{\boldsymbol{obs}}, \boldsymbol{X}_{\boldsymbol{mis}}, \boldsymbol{R}|\boldsymbol{\theta}, \boldsymbol{\zeta}) d\boldsymbol{X}_{\boldsymbol{mis}} \\ &= \int P(\boldsymbol{R}|\boldsymbol{\theta}, \boldsymbol{\zeta}) P(\boldsymbol{X}|\boldsymbol{\theta}) d\boldsymbol{X}_{\boldsymbol{mis}}. \end{aligned} \tag{9.11}$$

Under the MAR assumption of equation (9.9), equation (9.11) becomes

$$\begin{aligned} P(\boldsymbol{X}_{\boldsymbol{obs}}, \boldsymbol{R}|\boldsymbol{\theta}, \boldsymbol{\zeta}) &= \int P(\boldsymbol{R}|\boldsymbol{X}_{\boldsymbol{obs}}, \boldsymbol{\zeta}) P(\boldsymbol{X}|\boldsymbol{\theta}) d\boldsymbol{X}_{\boldsymbol{mis}} \\ &= P(\boldsymbol{R}|\boldsymbol{X}_{\boldsymbol{obs}}, \boldsymbol{\zeta}) \int P(\boldsymbol{X}|\boldsymbol{\theta}) d\boldsymbol{X}_{\boldsymbol{mis}} \\ &= P(\boldsymbol{R}|\boldsymbol{X}_{\boldsymbol{obs}}, \boldsymbol{\zeta}) P(\boldsymbol{X}_{\boldsymbol{obs}}|\boldsymbol{\theta}). \end{aligned} \tag{9.12}$$

This factorization means that the likelihood can be split into two sections, one pertaining to the parameter of interest, $\boldsymbol{\theta}$ and the other to the nuisance parameter, $\boldsymbol{\zeta}$. For any incomplete data problem the distribution of the complete data can be factored from equation (9.12) as

$$\begin{aligned} P(\boldsymbol{X}|\boldsymbol{\theta}) &= P(\boldsymbol{X}_{obs}, \boldsymbol{X}_{mis}|\boldsymbol{\theta}) \\ &= P(\boldsymbol{X}_{obs}|\boldsymbol{\theta})P(\boldsymbol{X}_{mis}|\boldsymbol{X}_{obs}, \boldsymbol{\theta}) \end{aligned}$$

Thus treating

$$L(\boldsymbol{\theta}|\boldsymbol{X}_{obs}) \propto P(\boldsymbol{X}_{obs}|\boldsymbol{\theta}),$$

then the corresponding loglikelihood equation is

$$\ell(\boldsymbol{\theta}|\boldsymbol{X}) = \ell(\boldsymbol{\theta}|\boldsymbol{X}_{obs}) + \ln P(\boldsymbol{X}_{mis}|\boldsymbol{X}_{obs}, \boldsymbol{\theta}) + c \tag{9.13}$$

where c is any arbitrary constant which is due to the proportional relationship between the likelihood function and the distribution of the the data. Equation (9.13) implies that the complete-data likelihood is equal to the observed-data likelihood plus another term, referred to as the predictive distribution of missing data by Schafer (1997), which takes into account the interdependence between $\boldsymbol{X}_{mis}$ and $\boldsymbol{\theta}$ and as such plays a pivotal role in the algorithm.

However because ML estimation requires $\ell(\boldsymbol{\theta}|\boldsymbol{X}_{\boldsymbol{obs}})$, therefore rewriting equation (9.13) in terms of $\ell(\boldsymbol{\theta}|\boldsymbol{X}_{\boldsymbol{obs}})$ yields

$$\ell(\boldsymbol{\theta}|\boldsymbol{X}_{obs}) = \ell(\boldsymbol{\theta}|\boldsymbol{X}) - \ln P(\boldsymbol{X}_{mis}|\boldsymbol{X}_{obs}, \boldsymbol{\theta}) \tag{9.14}$$

Since $\boldsymbol{X}_{\boldsymbol{mis}}$ is not known, equation (9.14) needs to be averaged over $P(\boldsymbol{X}_{\boldsymbol{mis}}|\boldsymbol{X}_{\boldsymbol{obs}}, \boldsymbol{\theta}^{(\boldsymbol{t})})$ where $\boldsymbol{\theta}^{(\boldsymbol{t})}$ is the current estimate of $\boldsymbol{\theta}$, thus dropping the constant c as it does

not affect the maximization that follows. Equation (9.14) therefore becomes

$$\ell(\boldsymbol{\theta}|\boldsymbol{X}_{obs}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \tag{9.15}$$

where

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \int \ell(\boldsymbol{\theta}|\boldsymbol{X}) P(\boldsymbol{X}_{mis}|\boldsymbol{X}_{obs}, \boldsymbol{\theta}^{(t)}) d\boldsymbol{X}_{mis} \tag{9.16}$$

$$H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \int \ln P(\boldsymbol{X}_{mis}|\boldsymbol{X}_{obs}, \boldsymbol{\theta}) P(\boldsymbol{X}_{mis}|\boldsymbol{X}_{obs}, \boldsymbol{\theta}^{(t)}) d\boldsymbol{X}_{mis} \tag{9.17}$$

If $\boldsymbol{\theta}^{(t+1)}$ denotes the value of $\boldsymbol{\theta}$ that maximizes $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, then $\boldsymbol{\theta}^{(t+1)}$ is a better estimate than $\boldsymbol{\theta^{(t)}}$ in the sense that its observed-data likelihood $\ell(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{X})$ is at least as high as that for $\theta^{(t)}$. Therefore after each iteration of the EM algorithm the observed-data likelihood is either increased or remains constant. This is one of the central results in Dempster, Laird and Rubin (1977) and mathematically it can be formalized as

$$\ell(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{X}_{obs}) \geq \ell(\boldsymbol{\theta}^{(t)}|\boldsymbol{X}_{obs})$$

This follows from equation (9.15) since

$$\ell(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{X}_{obs}) - \ell(\boldsymbol{\theta}^{(t)}|\boldsymbol{X}_{obs}) = Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) - [H(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})].$$

Hence $Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$ is always positive due to the fact that $\boldsymbol{\theta^{(t+1)}}$ has been chosen to maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$. Using equation (9.17) the remaining terms can be written as

$$\int -\ln \left[\frac{P(\boldsymbol{X}_{mis}|\boldsymbol{X}_{obs}, \boldsymbol{\theta}^{(t+1)})}{P(\boldsymbol{X}_{mis}|\boldsymbol{X}_{obs}, \boldsymbol{\theta}^{(t)})} \right] P(\boldsymbol{X}_{mis}|\boldsymbol{X}_{obs}, \boldsymbol{\theta}^{(t)}) d\boldsymbol{X}_{mis}$$

Since the function $x \ln x$ is convex, Jensen's inequality may be used, which mathematically states

$$\int \varphi(f(t)) dt \geq \varphi \left[\int f(t) dt \right],$$

or in statistical terms $E[\varphi(f(t))] \geq \varphi(E[f(t)])$, if $f(t)$ is a convex function.
Thus $-H(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) + H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$

$$\begin{aligned}
&\geq \int -\ln\left[\frac{P(\boldsymbol{X}_{mis}|\boldsymbol{X}_{obs},\boldsymbol{\theta}^{(t+1)})}{P(\boldsymbol{X}_{mis}|\boldsymbol{X}_{obs},\boldsymbol{\theta}^{(t)})}\right] P(\boldsymbol{X}_{mis}|\boldsymbol{X}_{obs},\boldsymbol{\theta}^{(t)})d\boldsymbol{X}_{mis} \\
&= -\ln \int P(\boldsymbol{X}_{mis}|\boldsymbol{X}_{obs},\boldsymbol{\theta}^{(t+1)})d\boldsymbol{X}_{mis} \\
&= -\ln 1 \\
&= 0
\end{aligned}$$

Thus the sum of the remaining terms, $-H(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) + H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$ is non-negative which means that at each iteration the observed-data likelihood is either increased or remains constant. It is convenient to split each iteration of the EM algorithm into two distinct stages, namely the expectation and maximization steps.
*T*he Expectation (E) step: In this step the function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is calculated by averaging the complete data likelihood $\ell(\boldsymbol{\theta}|\boldsymbol{X})$ over $P(\boldsymbol{X}_{mis}|\boldsymbol{X}_{obs},\boldsymbol{\theta})$.
*T*he Maximization (M) step: In this step $\boldsymbol{\theta}^{(t+1)}$ is found by maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

Thus overall the EM algorithm starts with an initial estimate of $\boldsymbol{\theta}$, $\boldsymbol{\theta}^{(0)}$, and produces a sequence $\{\boldsymbol{\theta}^{(t)} : t = 1, 2, \dots\}$. This sequence converges to a stationary point of the observed-data likelihood subject to certain conditions which are given in Dempster, Laird and Rubin (1977) and explored further in Wu (1983). In well behaved problems, where the loglikelihood function is unimodal and concave, this stationary point is also the global maximum and hence the EM algorithm yields the unique MLE of $\boldsymbol{\theta}$, the maximizer of both $L(\boldsymbol{\theta}|\boldsymbol{X}_{obs})$ and $\ell(\boldsymbol{\theta}|\boldsymbol{X}_{obs})$ (Dempster, Laird and Rubin ,1977). This convergence to the MLE of $\boldsymbol{\theta}$ will also occur regardless of the initial value of $\boldsymbol{\theta}$, in well behaved problems (Schafer, 1997).

There are several possible ways of choosing the starting values of the parameters $\boldsymbol{\theta}^{(0)}$.Those advocated by Little and Rubin (1987,2002) include

- starting values based on complete case (CC) analysis.
- One of the possible available case analysis.
- One of the single imputation methods.
- The means and variances from all the observed values and set all correlations to zero

However the solutions from available case analysis may result in a non-positive definite variance-covariance matrix which will cause problems at the first iteration, and using solutions from the CC analysis may yield unsatisfactory starting values when the proportion of missing data is large (Little and Rubin, 1987, 2002). But one advantage is that both of these options are most easily implemented and since CC analysis is often run before the EM algorithm it has the advantage that using its solution as starting values involves no additional computation.

As with all iterative procedures convergence criteria must be defined and the convergence carefully monitored. The convergence of the EM algorithm can be defined in various ways. First one can consider the overall convergence, which was shown originally by Dempster, Laird and Rubin (1977) to be linear and also by following the arguments of Schafer (1997).

Since the EM algorithm is an iterative algorithm at each stage a vector function $\boldsymbol{M}$ can be defined that maps the parameter space onto itself, since

$$\begin{aligned} \boldsymbol{\theta}^{(t+1)} &= \boldsymbol{M}(\boldsymbol{\theta}^{(t)}) \\ &= (M_1(\boldsymbol{\theta}^{(t)}), M_2(\boldsymbol{\theta}^{(t)}), \ldots, M_p(\boldsymbol{\theta}^{(t)}))^T \end{aligned}$$

The vector valued mapping of $\boldsymbol{M}$ incorporates both the E and M steps. Using Taylor's expansion about $\hat{\boldsymbol{\theta}}$ yields a first order approximation of $\boldsymbol{M}(\boldsymbol{\theta}^{(t)})$ hence

$$\boldsymbol{M}(\boldsymbol{\theta}^{(t)}) - \boldsymbol{M}(\hat{\boldsymbol{\theta}}) \approx \boldsymbol{M}'(\hat{\theta})(\theta^{(t)} - \hat{\theta}) \tag{9.18}$$

where $\boldsymbol{M}(\hat{\boldsymbol{\theta}})$ is a $p \times p$ matrix with typical element $\frac{\partial M_i(\theta)}{\partial \theta_j}\Big|_{\theta_j=\hat{\theta}_j}$. If $\hat{\boldsymbol{\theta}}$ is a stationary point of the algorithm then $\boldsymbol{M}(\hat{\boldsymbol{\theta}})$=$\hat{\boldsymbol{\theta}}$ and thus equation (9.18) can be rewritten as

$$\boldsymbol{\theta}^{(t+1)} - \hat{\boldsymbol{\theta}}^{(t)} \approx \boldsymbol{M}'(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}}) \tag{9.19}$$

or in terms of an error term at step t with $\epsilon^{(t)} = \boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}}$ as,

$$\epsilon^{(t+1)} \approx \boldsymbol{D}\boldsymbol{\epsilon}^{(t)}$$

where D is defined as the asymptotic rate matrix , $\boldsymbol{M}'(\hat{\boldsymbol{\theta}})$.

Thus the EM algorithm's convergence is said to be linear as $\epsilon^{(t+1)}$ is approximately a linear transformation of $\epsilon^{(t)}$ near $\hat{\boldsymbol{\theta}}$.

For a scalar θ, D is a single number between zero and one, with the small values of D leading to faster convergence. For multivariate data the convergence is governed by the eigenstructure of $\boldsymbol{D}$, and in particular by the largest fraction of missing information which corresponds to the largest eigenvalues of $\boldsymbol{D}$, denoted by λ_{max}.

Convergence can also be defined in terms of the individual parameters, such that element wise convergence rates are given by

$$\lambda_j^{(t)} = \frac{\theta_j^{(t+1)} - \theta_j^{(t)}}{\theta_j^{(t)} - \theta_j^{(t-1)}}, \qquad\qquad j = 1, 2, \ldots, p$$

where the largest eigenvalue of $\boldsymbol{D}$ and hence the largest fraction of missing information can be approximated by $\lambda_j^{(t)}$, $\forall j$, (Schafer, 1997).

Fraley (1999) showed that a better approximation to the largest fraction of

missing information can be obtained, with simple computation of the exact eigenvalues of $\boldsymbol{D}$ by using the iterative power method, which needs only an initial eigenvector estimate, and then computes only the largest eigenvalue and corresponding eigenvector, iterating until the desired accuracy is obtained. Since the entire matrix is not computed, this method is computationally efficient (Fraley, 1999).
Currently when the EM algorithm is implemented via computers convergence is monitored and defined in one of two steps namely:

- Successive parameter values, $\boldsymbol{\theta}^{(t)}$.
- Successive observed-data likelihoods $\ell(\boldsymbol{\theta}^{(t)}|\boldsymbol{X}_{obs})$.

The latter will immediately point out programming problems as it should never decrease, while the former is generally used in computer programmes to state that convergence has occurred for a particular accuracy of ϵ if

$$|\boldsymbol{\theta}_j^{(t)} - \boldsymbol{\theta}_j^{(t-1)}| \leq \epsilon |\boldsymbol{\theta}_j^{(t)}|$$

for a suitably small ϵ, for example 0.0001, and for $j = 1, 2, \ldots, p$.

9.7.2 Comparison of the Expectation-Maximization (EM) Algorithm with other iterative procedures

The EM algorithm has several advantages over other iterative procedures, such as the Newton-Raphson and method of scoring, namely:

- It is numerically stable, in that at each iteration the observed-data likelihood increases, except at a fixed point where it remains constant.
- Under fairly general conditions, it exhibits global convergence.

- Both steps are generally easy to compute as they are complete-data computations.
- No evaluation of the loglikelihood or its derivatives are needed.
- Its convergence is easily monitored.

However its disadvantages include:

- It has not built-in procedure for calculating an estimate of the variance-covariance matrix of the parameters.
- It may exhibit slow convergence, especially when the amount of missing data is considerable.
- It does not guarantee convergence to a global maximum when there are multiple maxima.

Although these disadvantages might seem considerable, the first two have been minimized by several recent methods and the latter is a problem inherent in almost all optimization techniques. It cannot, in general be avoided, but Little and Rubin (1987) and Schafer (1997) both recommended that, when dealing especially with multivariate data, the EM algorithm should be started from several different initial values for the parameters in order to be certain that it is converging to a global maximum and not to a local minimum.

One of the original criticisms of the EM algorithm was that it provided no built-in mechanism for estimating the variance-covariance matrix of the parameters. Potentially one can calculate

$$I(\hat{\theta}) = \left. \frac{\partial^2 \ell(\theta|X_{obs})}{\partial \theta^2} \right|_{\theta=\hat{\theta}}$$

directly where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ and $I(\theta)$ can be used to estimate the variance-covariance matrix, since

$$\theta - \hat{\theta} \simeq N(0, C)$$

where C can be estimated by $I^{-1}(\hat{\theta}) = I^{-1}(\theta)|_{\theta=\hat{\theta}}$. However the computation involved may be complex. Various alternative methods have therefore been developed to simplify the computation.

Louis (1982) gave a generally applicable method for finding the observed-data loglikelihood in terms of the gradient and second derivatives of the complete-data loglikelihood function which is generally easier to calculate. However this method still requires the calculation of the first and second derivatives of the complete-data loglikelihood. Meilijson (1989) avoids such calculations by using numerical approximations and information obtained from both the E and M steps, but is only applicable when the data are independent and identically distributed samples. Meng and Rubin (1991)generalized the approach by formulating the Supplemented EM (SEM) algorithm which does not require the calculation of the loglikelihoods nor their derivatives to obtain an asymptotic estimate of the variance-covariance matrix. While Oakes (1999) give a direct calculation of the information matrix using the function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, namely

$$I(\hat{\theta}) = \left[\frac{\partial^2 Q(\theta|\theta^{(t)})}{\partial\theta^2} + \frac{\partial^2 Q(\theta|\theta^{(t)})}{\partial\theta^{(t)}\partial\theta}\right]_{\theta=\theta^{(t)}=\hat{\theta}} \tag{9.20}$$

An alternative method for calculating the variance-covariance matrix, without calculating the information matrix, is to use the bootstrap method as formulated by Efron (1979) and described by Efron (1994). This method requires that B new samples are independently drawn from the original data set with replacement where each case has got the same probability of being

selected as n^{-1}. The new samples therefore consist of n cases, where the cases from the original data set may occur more than once or not at all. The EM algorithm is then run on each of these data sets, obtaining $\hat{\boldsymbol{\theta}}^{(b)}$, for $b = 1, 2, \ldots, B$. The bootstrap estimate of $\boldsymbol{\theta}$ is

$$\hat{\theta}_{boot} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{(B)}$$

and the bootstrap estimate of the variance-covariance matrix of $\hat{\theta}$ is

$$\hat{V} = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}^{(b)} - \hat{\theta}_{boot})(\hat{\theta}^{(b)} - \hat{\theta}_{boot})^T$$

It can be shown that under quite general conditions $\hat{V}$ is a consistent estimate of the variance of $\hat{\theta}$, as n and B tend to infinity (Little and Rubin 2002). Efron and Tibsharani (1993) showed that 50 to 100 bootstrap replications are generally sufficient for variance estimation. However although confidence intervals can be computed, for the bootstrap distribution that is approximately normal, these require in the order of 200 bootstrap replications (Efron, 1994).

Another criticism of the EM algorithm is of its slow convergence under certain conditions, one of these is that, if the number of variables is nearly equal to the number of cases, even a small amount of missing data may lead to slow convergence. Many solutions to convergence problems have been proposed, each with their own advantages and disadvantages. Due to the increase in computer power, they are not discussed in detail as they only exhibit worthwhile increases in speed when the data sets are large or when the amount of missing data is considerable. The former is a problem in only specific cases while the latter may bring validity of the analysis into question, as a large proportion of missing data may result in invalid inferences even if convergence can be obtained (Little and Rubin 1987).

However for completeness some of these modifications to the EM algorithm which enhance convergence and also require less computing power or enable the solving of complex E and M steps, are:

i). The Generalized EM (GEM) algorithm: it is the most useful when a solution to the M-step does not exist in closed form which, means that it may not be possible to maximize $Q(\theta|\theta^{(t)})$ globally. Thus GEM increases $Q(\theta|\theta^{(t)})$ over its value at $\theta^{(t)}$ rather than maximizing it over the entire parameter space of θ.

ii). Use of a multivariate generalization of Aitken's acceleration (Louis 1982) or of generalized conjugate gradient approach (Jamshidian and Jennrich, 1993) called the Accelerated EM (AEM) algorithm, however in both of these methods there is no longer any guarantee that the observed-data likelihood increases at each iteration.

iii). Expectation/Conditional Maximization (ECM) algorithm, where each M-step is split into several Conditional Maximization (CM) problems which are subject to the various constraints of θ such that the collection of all such constraints ensures that the maximization is over the full parameter space of θ (Meng and Rubin, 1993). Although this algorithm typically requires more iterations than the EM, it requires less computing time (Liu and Rubin, 1994).

iv). Expectation/Conditional Maximization Either (ECME) algorithm, which is an extension of the ECM algorithm where the M-step is again split into several CM-steps but some or all of the CM-steps are replaced by steps that conditionally maximize the incomplete-data loglikelihood function, $\ell(\theta|X_{obs})$ rater than $Q(\theta|\theta^{(t)})$, making it faster than the ECM

algorithm in computing time and more comparable with the EM algorithm in terms of the number of iterations needed (Liu and Rubin, 1994)

v). Alternating Expectation/Conditional Maximization (AECM) algorithm which is obtained by combining aspects of the ECM and Space-Alternating Generalized Expectation-Maximization (SAGE) algorithm (Fessler and Hero, 1994), where the augmentation of the observed data is allowed to vary over the CM-Steps (Meng and Van Dyk, 1997).

vi). Recently a large area of augmenting data into monotone pattern and then applying some variant of the EM algorithm has been widely investigated as it will often converge faster than the EM algorithm, thus the EM algorithm has been extended into the MEM (Monotone Expectation Maximization) algorithm and the ECME has been extended to the MECME (Monotone Expectation/Conditional Maximization Either) algorithm (Liu, 1999). As expected these techniques are most successful when only a small amount of data is needed to be augmented in order to make the missing data pattern monotone.

vii). The parameter-extended EM (PXEM) algorithm (Liu, Rubin and Wu, 1998) is one of the several incredibly fast EM type algorithms and is based on the idea that the inefficiency of the M-step is generally due to the fact that it acts as if the values of the sufficient statistics are correct, whereas they are actually interim values. These discrepancies are revealed by considering the difference between the imputed values and their expectations and the new differences are introduced into the algorithm by associating new parameters with them (Rubin, 1997).

viii). Iterative simualtion based techniques have aso been used to extend or

improve the EM algorithm resulting in Monte Carlo EM algorithm (Wei and Tanner, 1990) and the Stochastic EM (SEM) algoithm (Tanner, 1993).

9.7.3 The Missing Information Principle

The fundamental relationship that the complete information is equal to the observed information plus the missing information, was first shown by Orchard and Woodbury (1972) and was termed the "missing information principle". It can be shown by rewriting equation (9.13), but omitting the constant term that

$$\ell(\theta|X) = \ell(\theta|X_{obs}) + \ln P(X_{mis}|X_{obs}, \theta)$$

Differentiating both sides twice with respect to $\boldsymbol{\theta}$ and multiplying throughout by a negative one yields,

$$-\frac{\partial^2}{\partial\theta^2}\ell(\theta|X) = -\frac{\partial^2}{\partial\theta^2}\ell(\theta|X_{obs}) - \frac{\partial^2}{\partial\theta^2}\ln P(X_{mis}|X_{obs}, \theta)$$

Taking the expectation over $P(X_{mis}|X_{obs}, \theta)$ gives

$$\mathcal{J}_c(\theta) = \mathcal{J}_o(\theta) + \mathcal{J}_m(\theta) \tag{9.21}$$

where

$$\begin{aligned}
\mathcal{J}_c(\theta) &= -\frac{\partial}{\partial\theta^2}Q(\theta|\theta) \qquad \text{from equation (9.16)} \\
\mathcal{J}_o(\theta) &= -\frac{\partial}{\partial\theta^2}\ell(\theta|X_{obs}) \quad \text{from} \quad \theta - \hat{\theta} \simeq N(0, C) \\
\mathcal{J}_m(\theta) &= -\frac{\partial}{\partial\theta^2}H(\theta|\theta) \qquad \text{from equation (9.17)}
\end{aligned}$$

Equation (9.19) is termed the missing information principle and assumes only sufficient regularity to interchange the orders of differentiation and integration (taking expectations).

Dempster, Laird and Rubin (1977) showed that the above information matrices are connected to the asymptotic rate matrix D of the convergence of the EM algorithm by,

$$D = \mathcal{J}_c^{-1}(\theta)\mathcal{J}_m(\theta)$$

which implies that D is simply the ratio of the missing information to complete information. Thus the rate of convergence is also given by the largest eigenvalue of the matrix, $\mathcal{J}_c^{-1}(\theta)\mathcal{J}_m(\theta)$, and therefore the greater the proportion of missing data the slower the convergence, as stated earlier. (Molenberghs and Verbeke, 2005, pp 518-519)

9.7.4 Test for MCAR

Testing whether the data is MAR is impossible, as this would require information about the missing data. However it is possible to test whether data is MCAR or not. If missingness is confined to a single variable, then the simplest method to test MCAR is to compare distributions of the completely observed cases to the incomplete cases, either informally or formally by t-tests for the differences between means. Dixon (1983) using the program BMDP8D extends this test for data where missing values may occur on any of the p variables. Each variable which has missing values is divided into cases where that variable is observed and missing, the means of the other variables are then tested for differences. This is repeated for all variables in which missingness is present, therefore resulting in $p(p-1)$ t-tests if all the variables have some missing values. This may not only yield a large number of statistics, but simultaneous inference may be complicated due to possible correlations between the variables and the test has limited power when the number of incomplete cases is small (Little and Rubin, 1987). It is also possible that this method may produce results that could be regarded

as evidence against MCAR, even when a MCAR mechanism was permissible in the observed missingness pattern (Little, 1988).

Little (1988) thus proposed a single test statistic whose null distribution is asymptotically chi-squared. For data with J distinct missingness patterns then for each missingness pattern an indicator matrix D_j is defined. These matrices have the same number of rows as columns as X and with the number of columns equal to the number of observed variables for that missingness pattern. Each column consists of $p-1$ zeroes and a single one corresponding to the variable identified. Thus for $p = 4$ a misingness pattern with X_2 missing and the other variables observed would have indicator matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Also if the MLEs of $\boldsymbol{\mu}$ and $\sum$ are given by $\hat{\boldsymbol{\mu}}$ and $\hat{\sum}$, with $\tilde{\sum}= \frac{n}{n-1}\hat{\sum}$ being the corrected unbiased variance-covariance matrix of $\hat{\boldsymbol{\mu}}$ then the test statistic, derived under the assumption that $\boldsymbol{x}$ is multivariate normally distributed is

$$d^2 = \sum_{j=1}^{J} (\overline{x}_{obs,j} - \hat{\mu}) \tilde{\sum}_{obs,j}^{-1} (\overline{x}_{obs,j} - \hat{\mu})^T$$

where $\hat{\mu}_{obs,j} = \hat{\mu} D_j$ and $\tilde{\sigma}_{obs,j} = D_j^T \tilde{\sum} D_j$, $x_{obs,j}$ is the $m_j \times p_j$ matrix of observed values for missingness pattern j, with corresponding mean vector $\overline{x}_{obs,j}$ and m_j the number of cases in missingness pattern j with p_j the number of variables observed in that missingness pattern.

The test statistic d^2 is asymptotically chi-squared with degrees of freedom given by $\sum_{j=1}^{J} p_j - p$. This test was formulated assuming that if the

data is not MCAR, then the means of the variables may vary between distinct missingness patterns, but the variances and covariances are assumed to be the same for each pattern. This assumption can be relaxed, although the resulting test is more likely to be sensitive to departures from the normality assumption and less reliable for small samples (Little, 1988). The more restrictive test, as given above remains valid even under non-normality, although it tends to be conservative when the sample sizes are small (Little, 1988). Testing for MCAR will not only reveal whether simple methods will yield valid results but also provides guidance on whether to use the expected or observed information matrix. This is crucial because standard errors for parameters based on expected information matrix are only valid if the data is MCAR (Little, 1988). While the use of the observed information matrix only requires the assumption of MAR, but generally involves more computations.

9.7.5 Results for LOCF, CC and EM algorithm for the intermittent missingness-the 85 missing values in the response variable

There were 85 intermittent missing values in the response variable, the RSV status of a child. This constitutes $\frac{85}{9374} = 0.0091$ i.e. about 1% of missingness. Thus when the EM algorithm will be used to estimate these values using SPSS version 15, the estimated values will be rounded off, due to the fact that the intermittent missingness is extremely small and will not affect the precision of estimators significantly. The following results were obtained after using the above methods to estimate and "fill in" the missing values. The generalized linear mixed model and the random intercepts slopes model was fitted with the following results. Co-incidentally the EM estimated values were exactly

the same as those for the LOCF method. So the the results for those two methods will be the same. The CC method yielded exactly the same results as those in Chapter 4 and will not be reproduced here. We will only look at the optimal model.

LOCF-Random Effects Models

Different covariance structure models were fitted but in most of them only the residual term was estimable. The child random component was negligible.

	LOCF and EM Algorithm		
Covariance Structure		Estimate	Standard Error
Unstructured	UN(1,1)	0.002	0.001
	Residual(VC)	1.0428	0.01694
Compound symmetry	Var(child)	0.000	0.000
	CS(child)	-2.32E-6	3.513E-6
	Residual(VC)	1.0428	0.01694
Power	Var(child)	0.000	0.000
	SP(POW)(child)	0.000	0.000
	Residual(VC)	1.0428	0.01694
Spherical	Var(child)	0.000	0.000
	SP(SPH)(child)	0.000	0.000
	Residual(VC)	1.0428	0.01694
Gaussian	Var(child)	0.000	0.000
	SP(GAU)(child)	0.000	0.000
	Residual(VC)	1.0428	0.01694

Table 9.1: Covariance Parameter Estimates in a random effects model- LOCF and EM Algorithm methods

The solution for the fixed effects for all the different covariance structure models are:

	LOCF and EM Algorithm		
Effect	Estimate	Standard Error	Pr> $\|t\|$
Intercept	-5.0492	1.6250	< 0.0001
Age 0	-0.9213	1.295	0.4822
Age 1	-0.6594	1.0999	0.5991
Age 2	-0.2420	1.0645	0.8231
Age 3	-0.06896	0.9995	0.9898
Age 4	-0.6712	0.9786	0.4895
Age 5	-2.705	1.4210	0.0521
Age 6	-1.413	1.0416	0.1314
Age 7	-2.2900	1.1816	0.0646
Age 8	-1.002	0.6109	0.10012
Age 9	-0.7569	0.5948	0.1691
Age 10	-0.3861	0.4895	0.4879
Age 11	-0.5899	0.4812	0.2289
Age 12	0.0000	0.0000	0.0000
Dt	0.000892	0.009003	0.9205
Prev	46.652	8.9724	< 0.0001
Actipass 0	2.7231	0.1993	< 0.0001
Actipass 1	0.000	0.000	0.000
Timemonth	-0.09712	0.1097	0.7205

Table 9.2: Solution for the fixed effects using a random effects model -LOCF and EM algorithm

Effect	F-Value	P-value
Age	1.68	0.0812
Dt	0.02	0.9434
Prev	30.07	< 0.0001
Actipass	154.42	< 0.0001
Timemonth	0.18	0.6704

Table 9.3: Type III Effects in a random effects model-LOCF and EM algorithm

LOCF- Random Intercept Models

Different covariance structure models were fitted under this model. Again as earlier noted only the residual term was estimable under all covariance

structures. The non-convergence under the CS model is because the random intercept model is ideally the same as the CS model.

	LOCF and EM Algorithm		
Covariance Structure		Estimate	Standard Error
Unstructured	UN(1,1)	0.000	0.000
	Residual(VC)	1.0392	0.01588
Compound symmetry	Var(child)	No convergence	No convergence
	CS(child)	No convergence	No convergence
	Residual(VC)	No convergence	No convergence
Power	Var(child)	0.000	0.000
	SP(POW)(child)	0.000	0.000
	Residual(VC)	1.0392	0.01588
Spherical	Var(child)	0.000	0.000
	SP(SPH)(child)	0.000	0.000
	Residual(VC)	1.0392	0.01588
Gaussian	Var(child)	0.000	0.000
	SP(GAU)(child)	0.000	0.000
	Residual(VC)	1.0392	0.01588

Table 9.4: Covariance Parameter Estimates in a random intercept model-LOCF and EM Algorithm methods

The solution for the fixed effects for all the different covariance structure models shown in Table 9.5.

	LOCF and EM Algorithm		
Effect	Estimate	Standard Error	Pr> $\|t\|$
Intercept	-5.0379	1.4785	< 0.0001
Age 0	-0.9952	1.2492	0.4654
Age 1	-0.6789	1.1005	0.5649
Age 2	-0.2760	1.0550	0.8041
Age 3	-0.069526	0.9989	0.9521
Age 4	-0.6782	0.9791	0.4999
Age 5	-2.6025	1.3222	0.0501
Age 6	-1.5941	1.0105	0.1301
Age 7	-2.2502	1.1789	0.0579
Age 8	-0.9999	0.6060	0.0999
Age 9	-0.7394	0.5286	0.1681
Age 10	-0.3299	0.4689	0.4901
Age 11	-0.5791	0.4804	0.2299
Age 12	0.0000	0.0000	0.0000
Dt	0.000894	0.008662	0.9304
Prev	45.1033	8.3484	< 0.0001
Actipass 0	2.3541	0.1901	< 0.0001
Actipass 1	0.000	0.000	0.000
Timemonth	-0.04599	0.1094	0.6794

Table 9.5: Solution for the fixed effects in a random intercept model-LOCF and EM Algorithm methods

Effect	F-Value	P-value
Age	1.68	0.0832
Dt	0.02	0.9494
Prev	30.08	< 0.0001
Actipass	154.42	< 0.0001
Timemonth	0.18	0.6704

Table 9.6: Type III Effects in a random intercept model-LOCF and EM Algorithm methods

LOCF-Estimating the force of infection and the rate of recovery

The force of infection and the recovery rate of the process were again estimated after the EM algorithm was used to estimate the intermittent miss-

ingness in the response variable. The estimation was carried out via direct likelihood estimation and via a GLM. Both cases gave similar results.

LOCF and EM Algorithm Estimates				
	Direct Likelihood(ML)		GLM	
	Estimate	Standard Error	Estimate	Standard Error
$\hat{\lambda}$	0.001099	0.000117	0.0011	0.0001
$\hat{\nu}$	0.47695	0.068	0.5072	0.068

Table 9.7: Comparative Parameter Estimates

The CC estimates are the same as those given in Chapter 6. The estimates of the LOCF and the EM Algorithm are roughly the same as those of the CC analysis with no unusual observations.

LOCF- Piecewise force of infection estimates

The force of infection and recovery rate were then estimated at different time intervals, namely the 15 months, spanning the study. In a given month the force of infection was assumed to be constant leading to a piecewise constant step function. The results of the process are tabulated in Table 9.8 The force of infection peaks with different heights in months $2, 3, 11$ and 13. The highest peak occurs in month 13 as was the case earlier.

LOCF and EM Algorithm Estimates			
Month	Lambda	Estimate	Standard Error
2	$\hat{\lambda}_2$	0.0055	0.0008
3	$\hat{\lambda}_3$	0.0033	0.0007
4	$\hat{\lambda}_4$	0.0010	0.0004
5	$\hat{\lambda}_5$	0.00024	0.00011
6	$\hat{\lambda}_6$	0.00023	0.00013
7	$\hat{\lambda}_7$	0.00027	0.00016
8	$\hat{\lambda}_8$	0.0004	0.0003
9	$\hat{\lambda}_9$	0.00059	0.0003
10	$\hat{\lambda}_{10}$	0.0024	0.0005
11	$\hat{\lambda}_{11}$	0.0033	0.00072
12	$\hat{\lambda}_{12}$	0.0018	0.00064
13	$\hat{\lambda}_{13}$	0.0255	0.00044
14	$\hat{\lambda}_{14}$	0.0000	0.0000
15	$\hat{\lambda}_{15}$	0.0000	0.0000

Table 9.8: Monthly estimates of the force of infection using LOCF and EM Algorithm

LOCF- Piecewise rate of reovery estimates

As observed with earlier analyses, the rate of recovery results was fairly constant throughout the 15 month period except months 14 and 15 which did not have any data.

LOCF and EM Algorithm Estimates			
Month	Nu	Estimate	Standard Error
1	$\hat{\nu}_1$	0.4878	0.061
2	$\hat{\nu}_2$	0.4990	0.066
3	$\hat{\nu}_3$	0.5000	0.05
4	$\hat{\nu}_4$	0.5016	0.062
5	$\hat{\nu}_5$	0.5011	0.062
6	$\hat{\nu}_6$	0.4980	0.066
7	$\hat{\nu}_7$	0.499	0.076
8	$\hat{\nu}_8$	0.5002	0.070
9	$\hat{\nu}_9$	0.5033	0.065
10	$\hat{\nu}_{10}$	0.5019	0.05
11	$\hat{\nu}_{11}$	0.5026	0.070
12	$\hat{\nu}_{12}$	0.4999	0.060
13	$\hat{\nu}_{13}$	0.5012	0.067
14	$\hat{\nu}_{14}$	0.0000	0.0000
15	$\hat{\nu}_{15}$	0.0000	0.0000

Table 9.9: Monthly estimates of the recovery rate

9.8 Modelling the dropout

We first revisit the vastly quoted taxonomy of the classification of missing data mechanisms by Little and Rubin (1987) on the context of dropout. This taxonomy is relevant in longitudinal data where partially observed sequences, especially due to dropout (eg. a patient leaves the study at some time after which no more measurements are taken), are very common. (Kenward and Molenberghs, 1998). Little and Rubin (1987) define *missing completely at random*(MCAR) to be a process in which the probability of dropout is completely independent of the measurement process. A process is termed *missing at random* (MAR) if the probability of dropout is conditionally independent of the unobserved measurements given the observed measurements. Ignorability depends on the analysis methods used and applies strictly un-

der likelihood analyses. MAR does not imply ignorability under unweighted GEE, for example. Processes that are neither MCAR or MAR are called *non-ignorable*, in which case the probability of dropout depends on unobserved measurements. Dropout is a special case of incompleteness. Since incompleteness usually occurs for reasons outside the control of investigators, and may be related to the outcome measurement of interest, it is generally necessary to address the process that governs incompleteness. Only in special but important cases is it possible to ignore the measurement process (Jansen et al. 2006). This is not the case with the dropout in the current Kilifi data used in this thesis. Possible reasons for patients dropping out of the study (withdrawals) include death, adverse reactions, unpleasant study procedures, lack of improvement, early recovery, and other factors related or unrelated to the study procedure and intervention. In the context of the Kilifi data set every child was initially supposed to have 44 visits, but this was not the case. Tables (9.10) and (9.11) summarize the dropout process in the Kilifi RSV data. Table (9.10) is a summary of the full data set while Table (9.11) is an extract of the first three children in the study.

Kenward (2006) states that any analysis of incomplete longitudinal data must have an assumption with respect to the missing data mechanism (Kenward M, 2006), whether MAR, MCAR or MNAR. If it is MAR then a standard likelihood analysis such as the Generalized Linear Mixed Model can follow. For modelling longitudinal binary data with dropout, there are 2 alternatives (Kenward M, 2006, p. 51)

- Alternative 1: use multiple imputation with an uncongenial imputation distribution. However since the Kilifi data set has a mixture of discrete and continouous variables we can appropriately use multiple imputation using chained equations (MICE) or as it is sometimes called,

Visit	Percent missing	Percent observed
1	0.0	100.0
2	3.0	97.0
3	5.6	94.4
4	6.5	93.5
5	5.9	94.1
6	6.8	93.2
7	5.6	94.4
8	7.1	92.9
9	8.0	92.0
10	8.3	91.7
11	9.8	90.2
12	9.8	90.2
13	9.8	90.2
14	10.1	89.9
15	9.8	90.2
16	10.4	89.6
17	10.7	89.3
18	12.7	87.3
19	11.8	88.2
20	13.6	86.4
21	14.5	85.5
22	16.9	83.1
23	17.5	82.5
24	21.0	79.0
25	22.8	77.2
26	26.0	74.0
27	27.8	72.2
28	33.1	66.9
29	41.4	58.6
30	47.3	52.7
31	53.6	46.4
32	60.4	39.6
33	68.6	31.4
34	76.3	23.7
35	79.3	20.7
36	85.5	14.5
37	89.6	10.4
38	91.7	8.3
39	94.7	5.3
40	97.0	3.0
41	98.2	1.8
42	99.1	0.9
43	99.4	0.6
44	99.7	0.3
Total	37	63.0

Table 9.10: Dropout percentage table

Visit	Child		
	1	2	3
1	O	O	O
2	O	O	O
3	O	O	O
⋮	⋮	⋮	⋮
25	O	M	O
26	O	M	O
27	O	M	O
28	M	M	M
29	M	M	M
30	M	M	M
⋮	⋮	⋮	⋮
44	M	M	M

Table 9.11: Dropout table for first three children

multiple imputation by fully conditional specification.

- Alternative 2: use a subject-specific model with likelihood such as the Generalized Linear Mixed Models(GLMM's)

The view that likelihood methods that ignore the missing value mechanism are valid under an MAR process has evolved out of the work of Rubin and Little and the likelihood in that case must be interpreted in a frequentist sense. Kenward and Molenberghs (1998) give an excellent expository of the qualification of this statement. They first state that Rubin (1976) has showed that MAR is necessary and sufficient to ensure validity of direct likelihood inference when ignoring the process that causes missing data. They further state the essence of this approach is that of identifying and using the appropriate sampling distribution. This is obviously relevant for determining the distributions of test statistics, expected values of the information matrix and measures of precision. Little and Rubin (1987) look into the associated aspects of the above mentioned issues and suggest the use of the observed in-

formation matrix to circumvent problems associated with the determination of the correct, expected information matrix. Several authors such as Meng and Rubin (1991), Baker (1992), just to mention two of them, have also explored this area of research. Louis (1982), Meilijson (1989) and Kenward, Molenberghs and Lesaffre (1994) have all looked at the use of the observed information matrix, without making reference to the problems associated with the expected information matrix. Kenward and Molenberghs (1998) then further explore these aspect using three different illustration and conclude that as long as the observed information matric is used, conventional likelihood based frequentist inference is applicable in the MAR setting. Using the Alternative 2 (Kenward , 2006, p. 51), we will assume that the dropout mechanism for our missingness is MAR and a standard likelihood analysis, such as the Generalized Linear Mixed Model follows. The model was fitted using PROC NLMIXED and the results are given below in Tables refmondo and 9.13 using Gaussian and Adaptive Gaussian Quadrature methods. The results are discussed in relation to those obtained in Chapter 4 using the GLMM approach The fitted model was:

$rsvpos = \beta_{00} + \beta_0 age0 + \beta_1 age1 + beta_2 age2 + \beta_3 age3 + \beta_4 age4 + \beta_5 age5 + \beta_6 age6 + \beta_7 age7 + \beta_8 age8 + \beta_9 age9 + \beta_{10} age10 + \beta_{11} age11 + \beta_{13} dt + \beta_{14} prev + \beta_{15} actipass + \beta_{16} timemonth + childeffect(\tau)$.

The sample program used to fit the above model is:

```
proc nlmixed data =lisa   qpoints=20 tech=nmsimp; parms beta00=-5.06
beta0=-0.9 beta1=-0.65 beta2=-0.27 beta3=-0.067 beta4=-0.66
beta5=-2.5 beta6=-1.6 beta7=-2.2 beta8=-0.99 beta9=-0.74
beta10=-0.32 beta11=-0.55 beta13=-0.0009 beta14=45 tau2=1.02;
teta=beta00+b+beta0*age0+beta1*age1+beta2*age2+beta3*age3
+beta4*age4+beta5*age5+beta6*age6 +beta7*age7+beta8*age8+beta9*age9+
```

```
beta10*age10+beta11*age11+beta13*dt+beta14*prev+beta15actipass
+beta16timemonth; expteta=exp(teta); p=expteta/(1+expteta); model
rsvpos~binary(p); random b~normal(0,tau2) subject=rsv; run;
```

The results for adaptive Gaussian quadrature and non-adaptive Gaussian quadrature are given below:

Gaussian Quadrature			
Effect	$Q = 3$	$Q = 5$	$Q = 20$
Intercept	-5.04(1.488)	-5.72(1.584)	-5.466(1.384)
beta0	-0.92(1.244)	-0.94(1.862)	-0.91(1.35)
beta1	-0.65(1.082)	-0.68(1.985)	-0.69(1.857)
beta2	-0.28(1.081)	-0.27(1.056)	-0.28(1.045)
beta3	-0.07(1.001)	-0.08(0.991)	-0.07(0.984)
beta4	-0.67(0.981)	-0.69(0.967)	-0.67(0.991)
beta5	-2.71(1.381)	-2.12(1.354)	-2.74(1.311)
beta6	-1.61(1.021)	-1.59(1.044)	-1.60(1.058)
beta7	-2.23(1.184)	-2.13(1.192)	2.20(1.188)
beta8	0.99(0.624)	-1.05(0.652)	-1.01(0.652)
beta9	-0.76(0.521)	-0.74(0.601)	-0.75(0.504)
beta10	-0.33(0.456)	-0.42(0.498)	-0.32(0.416)
beta11	-.57(0.481)	-0.54(0.453)	-0.57(0.485)
beta13	-0.0008(0.009)	0.0009(0.007)	-0.0008(0.006)
beta14	47.2(7.995)	49.3(8.994)	46.1(8.774)
beta15	2.31(0.168)	2.33(0.183)	2.38(0.199)
beta16	-0.05(0.109)	-0.04(0.137)	-0.05(0.108)
τ	1.03(0.0114)	1.01(0.018)	1.03(0.013)
-2ℓ	2243.6.7	2242.2	2243.9

Table 9.12: Solution for the fixed effects to model the dropout-gaussian quadrature

Adaptive Gaussian Quadrature			
Effect	$Q = 3$	$Q = 5$	$Q = 20$
Intercept	-5.02(1.433)	-5.71(1.498)	-5.786(1.354)
beta0	-0.92(1.244)	-0.94(1.862)	-0.91(1.145)
beta1	-0.65(1.012)	-0.68(1.385)	-0.68(1.017)
beta2	-0.28(1.051)	-0.23(1.034)	-0.26(1.026)
beta3	-0.07(0.991)	-0.08(0.988)	-0.07(0.999)
beta4	-0.66(0.989)	-0.69(0.997)	-0.67(0.982)
beta5	-2.61(1.341)	-2.12(1.369)	-2.72(1.311)
beta6	-1.61(1.071)	-1.59(1.022)	-1.63(1.758)
beta7	-2.24(1.191)	-2.13(1.189)	2.20(1.198)
beta8	0.99(0.674)	-1.02(0.652)	-1.01(0.699)
beta9	-0.74(0.501)	-0.74(0.561)	-0.75(0.540)
beta10	-0.33(0.459)	-0.42(0.498)	-0.32(0.446)
beta11	-.57(0.498)	-0.54(0.422)	-0.57(0.494)
beta13	-0.0008(0.007)	0.0009(0.009)	-0.0008(0.006)
beta14	43.8(8.395)	49.3(8.994)	46.1(8.214)
beta15	2.22(0.178)	2.23(0.183)	2.41(0.189)
beta16	-0.05(0.109)	-0.03(0.158)	-0.05(0.104)
τ	1.03(0.0115)	1.01(0.018)	1.03(0.013)
-2ℓ	2243.8	2242.8	2243.8

Table 9.13: Solution for the fixed effects to model the dropout-adaptive gaussian quadrature

Both the gaussian quadrature and adpative gaussian quadrature show the type of visit (actipass) and prevalence variable (prev) to be significant. Thus a generalized linear mixed model with only the prev and actipass covariates was fitted. The results are given below. Different covariance structure models were again used for comparison. The solution for the fixed effects for all the different covariance structure models are tabulated in tables that follow.

Fitting the above model as a random intercept model yielded the following results: The solution for the fixed effects for all the different covariance

Covariance Structure		Estimate	Standard Error
Unstructured	UN(1,1)	No convergence	No convergence
	Residual(VC)	No convergence	No convergence
Compound symmetry	Var(child)	7.187E-6	2.885E-6
	CS(child)	-6.71E-6	.
	Residual(VC)	0.8814	0.01306
Power	Var(child)	4.745E-7	2.885E-6
	SP(POW)(child)	0.000	0.000
	Residual(VC)	0.8814	0.01306
Spherical	Var(child)	4.745E-7	2.885E-6
	SP(SPH)(child)	0.000	0.000
	Residual(VC)	0.8814	0.01306
Gaussian	Var(child)	4.745E-7	2.885E-6
	SP(GAU)(child)	0.000	0.000
	Residual(VC)	0.8814	0.01306

Table 9.14: Covariance Parameter Estimates in a random effects model to handle the dropout-using prev and actipass

Effect	Estimate	Standard Error	Pr> $\|t\|$
Intercept	-5.9583	0.1796	< 0.0001
Prev	50.7801	4.8589	< 0.0001
Actipass 0	2.0576	0.1608	< 0.0001
Actipass 1	0.000	0.000	0.000

Table 9.15: Solution for the fixed effects in a random effects model to handle the dropout-using prev and actipass

structure models are:

Effect	F-Value	P-value
Prev	109.22	< 0.0001
Actipass	163.74	< 0.0001

Table 9.16: Type III Effects in a random effects model to handle the dropout-using prev and actipass

Covariance Structure		Estimate	Standard Error
Unstructured	UN(1,1)	0.2086	0.1638
	Residual(VC)	0.8782	0.01308
Compound symmetry	Var(child)	0.1053	0.1638
	CS(child)	0.1034	.
	Residual(VC)	0.8782	0.01308
Power	Var(child)	0.2086	0.1638
	SP(POW)(child)	0.000	0.000
	Residual(VC)	0.8782	0.01308
Spherical	Var(child)	0.2086	0.1638
	SP(SPH)(child)	1.000	0.000
	Residual(VC)	0.8782	0.01308
Gaussian	Var(child)	0.2086	0.1638
	SP(GAU)(child)	1.000	0.000
	Residual(VC)	0.8782	0.01308

Table 9.17: Covariance Parameter Estimates in a random intercept model to handle the dropout

Effect	Estimate	Standard Error	Pr$> \|t\|$
Intercept	-5.9564	0.1802	< 0.0001
Prev	50.6917	4.8703	< 0.0001
Actipass 0	2.0601	0.1615	< 0.0001
Actipass 1	0.000	0.000	0.000

Table 9.18: Solution for the fixed effects in a random effects model to handle the dropout-using prev and actipass

Once again the random intercept model estimates are very similar to those of

Effect	F-Value	P-value
Prev	108.33	< 0.0001
Actipass	163.77	< 0.0001

Table 9.19: Type III Effects in a random effects model to handle the dropout-using prev and actipass

the random effects estimates. The above model was also fitted as a GLMM but using PROC NLMIXED. The fitted model was:

$$rsvpos = \beta_{00} + \beta_0 prev + \beta_1 actipass + childeffect(\tau)$$

The sample program is:

```
proc nlmixed data =lisa   qpoints=20 tech=nmsimp; parms beta00=-5.9
beta0=50.7 beta1=2.03 tau2=0.85;
teta=beta00+b+beta0*prev+beta1*actipass; expteta=exp(teta);
p=expteta/(1+expteta); model rsvpos~binary(p); random
b~normal(0,tau2) subject=rsv; run;
```

The results for adaptive Gaussian quadrature and non-adaptive Gaussian quadrature are given in Tables (9.20) and (9.21):

Gaussian Quadrature			
Effect	$Q = 3$	$Q = 5$	$Q = 20$
Intercept	-3.9555(0.1805)	-3.6407(0.1764)	-3.9257(0.1790)
beta0	50.2189(5.7557)	52.7442(6.1598)	50.1865(5.703)
beta1	-1.9171(0.1686)	-2.1389(0.1853)	-1.9715(0.1685)
τ	0.8700(0.0114)	0.7300(0.018)	0.7800(0.0190)
-2ℓ	1494.7	1458.2	1494.6

Table 9.20: Solution for the fixed effects fitting a GLMM using a NLMIXED gaussian quadrature to handle the dropout-using prev and actipass

Adaptive Gaussian Quadrature			
Effect	$Q = 3$	$Q = 5$	$Q = 20$
Intercept	-3.9515(0.1807)	-2.8854(0.1482)	-3.9243(0.1789)
beta0	50.5363(5.7469)	49.8533(6.1178)	50.1887(5.700)
beta1	-1.9558(0.1696)	-1.9917(0.1673)	-1.9717(0.1684)
τ	0.8700(0.012)	0.8590(0.0114)	0.8700(0.0119)
-2ℓ	1494.7	1420	1494.6

Table 9.21: Solution for the fixed effects fitting a GLMM using a NLMIXED non gaussian quadrature to handle the dropout-using prev and actipass

The results from the non-adaptive and adaptive gaussian quadrature estimation are very similar. Here not much is achieved by adopting the more complex adaptive quadrature method. The smaller model fits much better than the full model judging from the -2 log-likelihood values.

9.8.1 Multiple Imputation

Multiple Imputation (MI) was first comprehensively described in Rubin (1987) in the context of analyzing large sample surveys with non-responses, where these surveys were used to create public-use data sets to be shared by many ultimate users (Rubin, 1996). Although the idea first appears in Rubin (1977) it was not originally widely used due to a lack of computational tools, but the large amount of recent research into Markov Chain Monte Carlo (MCMC) methods as well as current advanced technology has ensured that it is now more easily used. It can be simply described as a technique in which the missing data values X_{mis} are replaced by $X^1_{mis}, X^2_{mis}, \ldots, X^m_{mis}$ thus forming m complete data sets, which are then analyzed by standard statistical methods. The results of the m analyses are then combined into a single inferential statement, ensuring that the uncertainty due to the missing data is incorporated, by rules provided by Rubin (1987). The most critical objection to MI

was that it is a simulation technique. However simulation has now become far more accepted and widely used in statistics. Further still simulations in MI is only used to handle the missing information, the rest of the information is handled by complete case (CC) methods and thus it can only distort part of the inference (Rubin, 1996) and not the entire inference. A summary of Rubin's (1987) terminology on MI entails three distinct stages or phases in the process:

1. The missing values are filled in m times to create m complete data sets.

2. The m complete data sets are analyzed using standard statistical procedures.

3. The results from the m analyses are combined into a single inference.

The first and third tasks can be done using the SAS procedures PROC MI and PROC MIANALYZE. The second task can be done using using any standard statistical procedures.

The theory underlying MI is based on a Markov chain consisting of independent draws from

$$X_{mis} \sim P(X_{mis}|X_{obs}).$$

This distribution is often difficult to draw from in practice and an approximation is used whereby,

$$X_{mis}^{(t)} \sim P(X_{mis}|X_{obs}, \theta^{(t)}).$$

In order that the imputations be independent draws, the Markov chain must be sampled at every k^{th} iterate, where k is large enough so that the dependence between values is negligible (Schafer, 1997). There are no strict rules

for determining k but the graphical representation of Autocorrelation functions (ACFs) can be useful.

Alternatively, multiple sequences can be formed from different starting values and each subsampled once at the k^{th} iteration, where k must be large enough so that stationarity has occurred. Gelman and Rubin (1992) and Schafer (1997) advocate this method of multiple sequences as the simplest and most reliable way to assess the Monte Carlo error of an estimate. Schafer (1997) also states that single chains are recommended when one is certain that reliable convergence to a stationary distribution will occur, otherwise multiple sequences with overdispersed starting values are recommended. Multiple sequences can also be used as a diagnostic tool, since if results from multiple sequences disagree, then the value of k is possibly too small and valid inferences cannot be obtained (Geyer, 1992). Using multiple sequences also means that it is possible to use a statistic, called the potential scale reduction, as defined by Gelman and Rubin (1992), to monitor convergence.

This statistic is based on the idea that convergence can be studied by comparing variation between and within the simulated chains, until the within variation is approximately equal to the between variation. That is, only when the distribution of each sequence is close to the distribution of the sequences mixed together can each be approximating the same target distribution (Little and Rubin, 2002). For each scalar parameter estimate θ, from D parallel chains, the draws, at each iteration t are denoted as $\theta_{d,t}$ where $d = 1, 2, \ldots, D; t = 1, 2, \ldots, T$.

The within and between variances are given respectively as

$$\upsilon = \frac{1}{D}\sum_{d=1}^{D}\left[\frac{1}{T-1}\sum_{t=1}^{T}(\theta_{d,t} - \overline{\theta}_{d.})^2\right]$$

$$b = \frac{T}{D-1}\sum_{d=1}^{D}(\overline{\theta}_{d.} - \overline{\theta}_{..})^2$$

where

$$\overline{\theta}_{d.} = \frac{1}{T}\sum_{t=1}^{T}\theta_{d,t} \quad \text{and} \quad \overline{\theta}_{..} = \frac{1}{D}\sum_{d=1}^{D}\overline{\theta}_{d.}$$

The test statistic is given by

$$\sqrt{\hat{R}} = \sqrt{\frac{\frac{T-1}{T}\upsilon + \frac{1}{T}b}{\upsilon}}.$$

This potential scale reduction statistic decreases to one as $T \to \infty$. Thus when $\sqrt{\hat{R}}$ is high then there is evidence that running the chains for further iterations will be worthwhile. Once $\sqrt{\hat{R}}$ is near one for all the scale parameters, then subsequent draws can be treated as draws from the target distribution (Gelman and Rubin, 1992). Little and Rubin (2002) recommend an upper level 1.2 as "near" enough in general, and also that the scalar parameters are transformed to be approximately normal.

One of the major advantages of MI is that a small m value of typically between three and five will generally yield efficient estimates and therefore unless the fraction of missing data is high, producing and analysing a large number of data sets is simply not advantageous (Schafer, 2002). There are two reasons why only such a small m is necessary, firstly MI only relies on simulation to solve the missing data aspect of the problem and although increasing m would decrease the Monte Carlo error, this error is a relatively small percentage of the overall uncertainty and therefore the gain in efficiency

is small (Rubin, 1987). The second reason is that the rule for combining the m complete data analyses explicitly take into account the Monte Carlo error (Schafer, 1999).

Although only a small m is needed, Meng (1994) points out that in large public-use databases, it is not uncommon for many different analyses to be performed on the same imputed data sets, since the imputer and analyst are generally different people. In order to avoid this Meng (1994) suggests that some 30 imputed data sets are computed by the imputer, then each analyst, in turn randomly chooses the number they require from this pool of imputed data sets. Although this number may seem to be prohibitive, in the current computing environment, storage and computing time are no longer problems.

The fact, that the two stages are computationally separate and thus different people may perform the imputation and analysis independently, is one of MI's advantages especially in the field of large sample surveys. However this advantage may mean that the statistical assumptions or model used for the imputation may somehow be incompatible with the later analysis (Schafer, 1997). If the imputation model is more general, with fewer assumptions, then MI will still yield valid inferences, but with a possible loss of power because the additional generality may add variation among the imputes, $X_{mis}^{(1)}, X_{mis}^{(2)}, \ldots, X_{mis}^{(m)}$. Superefficiency, as termed by Rubin (1996) occurs when the imputer makes more valid assumptions than the analyst, then the MI estimate $\overline{\theta}$ may be more precise than any estimate derived only from the observed data and analyst's model. Thus the resulting confidence intervals are narrower and any hypothesis testing tends to be conservative.

If the extra assumptions made by the imputer are not valid then likewise MI inferences may not be valid, for example, in the case if the imputer assumes no-interaction, then if the analyst later tests for interaction, the conclusion found may not be valid. In the case where, for convenience, non-normal data is modelled as normal for the imputation, inference based on the means for variance-covariance matrix, such as regression or principle components analysis (PCA) will perform reasonably well, but analysis sensitive to tail behaviour should not be performed. Schafer and Olsen (1998) propose that the imputation model should be rich enough to preserve any associations or relationships that might be the focus of later investigations.

Combining the m data sets

The basic rules for pooling the m imputed data sets first appeared in Rubin (1987) and have been extended by several authors. The following summary is taken from Molenberghs and Verbeke (2005, p. 513). After the data is complete, suppose that inference about the parameter $\boldsymbol{\beta}$ is made by assuming

$$(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \sim N(\mathbf{0}, \boldsymbol{U}).$$

The m within-imputation estimates for $\boldsymbol{\beta}$ are pooled to give the multiple imputation estimate

$$\hat{\beta}^* = \frac{\sum_{m=1}^{M} \hat{\beta}^m}{M}$$

Further it follows that one can make normal based inferences for $\boldsymbol{\beta}$ based upon:

$$(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^*) \sim N(\mathbf{0}, \boldsymbol{V}).$$

where

$$V = W + \left(\frac{M+1}{M}\right) B$$

and

$$W = \frac{\sum_{m=1}^{M} U^m}{M}$$

is the average *within* imputation variance, and

$$B = \frac{\sum_{m=1}^{M} (\hat{\beta}^m - \hat{\beta}^*)(\hat{\beta}^m - \hat{\beta}^*)'}{M-1}$$

is the *between* imputation variance.

Hypothesis testing

The asymptotic results as well as the the reference χ^2 distribution depend not only on the the sample size N but also on the number of imputations m. Molenberghs and Verbeke (2005, p. 514) use the proposition of Li, Raghunathan and Rubin (1991), who propose the use of the F distribution. They use the following method to test the hypothesis:

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta \neq \theta_0$$

They use the following method to calculate the p-values:

$$p = P(F_{k,w} > F)$$

where k is the length of the parameter vector θ, $F_{k,w}$ is an F random variable with k numerator and w denominator degrees of freedom, and

$$\begin{aligned} F &= \frac{(\theta^* - \theta_0)' W^{-1} (\theta^* - \theta_0)}{k(1+r)} \\ \text{where} \quad w &= 4 + (\tau - 4)\left[1 + \frac{(1+2\tau^{-1})}{r}\right]^2 \\ r &= \frac{1}{k}\left(1 + \frac{1}{M}\right)\operatorname{tr}(BW^{-1}) \\ \tau &= k(M-1) \end{aligned}$$

Here r is the average relative increase in variance due to nonresponse across the components of θ. The limiting behaviour of this F variable is that if $M \to \infty$, then the reference distribution of F approaches an $F_{k,\infty} = \chi^2/k$ distribution. This procedure is applicable when one component, a subvector, or a set of linear contrasts in θ is the subject of hypothesis testing. In the case of a subvector or one component, the corresponding sub-matrices of B and W are used in the formulae. For a set of linear contrasts, $L\beta$, one should use the appropriately transformed covariance matrices: $\tilde{W} = LWL', \tilde{B} = LBL', \text{and } \tilde{V} = LVL'$.

9.8.2 Methods for creating multiple imputations

The task of creating multiple imputations may be difficult to achieve as they should properly reflect the uncertainty about the missing values given all available information and as such the imputing model should be as objective and general as possible, however the more general the model, the more complex the computation and the more computing power and storage required. One of the most widely used methods, which allows, in theory, the inclusion of complex models is, the Data Augmentation (DA) algorithm which obtains multiple imputations from the Markov chain with draws from

$$X_{mis}^{(t+1)} \sim P(X_{mis}|X_{obs}, \theta^{(t)})$$

where

$$\theta^{(t+1)} \sim P(\theta|X_{obs}, X_{mis}^{(t+1)}).$$

However in multivariate data sets where there are non-linear relationships between the variables, constructing a suitable model, programming the random draws and assessing the convergence of a Markov chain may be time consuming and complex (Little and Rubin, 2002). Simpler methods can thus

be used, which approximate draws from

$$X_{mis}^{(t)} \sim P(X_{mis}|X_{obs}).$$

Although such methods are usually less formally rigorous, they are generally easier to implement and will yield approximately valid inferences when used in conjunction with the MI rules. Little and Rubin (2002) note that inferences obtained from the simpler methods may be more valid than inferences obtained under the DA algorithm with an incorrect imputation model.

Improper multiple imputation

Improper MI as termed by Rubin (1987) does not propagate the uncertainty in estimating θ, since the Markov chain is formed from draws of,

$$X_{mis}^{(t)} \sim P(X_{mis}|X_{obs}, \tilde{\theta}),$$

where $\tilde{\theta}$ is some estimate of θ, often its MLE obtained from the final iteration of the EM algorithm or from the CC analysis. Improper MI works well if the amount of missing data is not too large but can lead to anti-conservative confidence intervals.

Methods that propagate uncertainty

In order that the uncertainty about θ is propagated, X_{mis} can be drawn from

$$X_{mis}^{(t)} \sim P(X_{mis}|X_{obs}, \tilde{\theta}^{(t)}),$$

where

$$\tilde{\theta}^{(t)} \sim P(\theta|X_{new}).$$

The matrix X_{new} can be defined in one of several ways, including

- A subset of the data, if such a subset can be found, having a monotone data pattern.
- The completely observed cases.
- The entire data set with sufficient data values imputed, by one of the simple methods, in order to make a monotone pattern.

The last method is intuitively best suited to situations where only a few data values need to be imputed in order to obtain a monotone pattern.

In large data sets an alternative method can be used which is often preferable to the above methods (Little and Rubin, 2002). If the MLE, $\hat{\theta}$ is available together with some consistent estimate of its large sample variance-covariance matrix, $C(\hat{\theta})$, then $\theta^{(t)}$ can be drawn from

$$\theta^{(t)} \sim N(\hat{\theta}, C(\hat{\theta})).$$

Thus draw t has the form

$$\theta^{(t)} = \hat{\theta} + z^{(t)}$$

with $z^{(t)} \sim N(0, C(\hat{\theta}))$.

Regression methods

Rubin (1987) states that a regression model is fitted for each variable with the missing values, with the previous variables as covariates. Based on the resulting model, a new regression model is then fitted and used to impute the missing values for each variable. Because the data set has a monotone missing data pattern, the process is repeated sequentially for variables with missing values (Molenberghs and Verbeke 2005, p. 515).

Propensity score method

The propensity score, as defined by Rosenbaum and Rubin (1983), is the conditional probability of an assignment to a particular treatment given a vector of observed covariates. In this method the propensity score is generated for each variable with missing values to indicate the probability of observations being missing. The observations are then grouped based on these propensity scores, and an approximate Bayesian bootstrap imputation is applied to each group. This method does not take into account the correlations among variables but only utilizes the covariate information that is associated with whether the imputed variable values are missing. It is advantageous for inferences about distributions about imputed variables but not appropriate for analyses involving relationships between variables (Molenberghs and Verbeke 2005, p.515-516).

Methods that use pragmatic models

In practice it is not unusual that a set of conditional distributions can be formulated relating each variable in turn, in a multiparameter data set, to all other variables. These models are generally only reasonable when taken individually and are rarely able to be described by a single joint distribution (Little and Rubin, 2002). An iterative algorithm can however be developed whereby for each variable's model, suitable parameters are drawn, used to estimate that variables missing values and then imputed. Such algorithms cycle both through all variables in the data set as well as through iterations until convergence has occurred.

There are other methods such as importance sampling, which is a refinement to the above methods that propagate uncertainty, using MLE's from

bootstrapped sampling that involves the EM algorithm and MCMC methods (Molenberghs and Verbeke, 2005). We can not consider all of these methods but it is worthwhile mentioning their existence.

9.8.3 Software used for MI

Due to the explosions and extensions of the methods that have been developed for handling missing data, in the not so distant past, this trend has been mirrored by an increase in the availability of statistical software that is either specifically designed for the analysis of missing data or includes procedures for handling missing data. Together with many freestanding, smaller, freely downloadable programmes, the more mainstream statistical packages have developed procedures for handling missing data. Some of the more popular software include (out of a rather long list):

- NORM (version 2.03, Schafer 2001)
- Norm library versions, CAT, MIX, PAN (Schafer 2001)
- SPSS (version 13)
- SAS (version 9.1.3)

We will focus our attention on SAS (SPSS version 13 was used in earlier sections regarding the EM Algorithm). SAS has got two procedures 'PROC MI' and 'PROC MIANALYZE'. There are three key sequences of tasks for handling multiple imputation in SAS. These are the *I*mputation task followed by the *A*nalysis task and finally the *I*nference task. These are effected using 'PROC MI' followed by any of the standard statistical procedures for example, PROC GENMOD, and then finally using 'PROC MIANALYZE'.

PROC MI

PROC MI is used to generate the imputations. It will create m imputed data sets from an input data set and store the imputed data sets physically in a single data set with indicator variable _IMPUTATION_ to separate the imputed copies. Some of the options available within PROC MI statement include 'simple' which gives the simple descriptive statistics and pairwise correlations based on the complete cases from the input data set. The number of imputations can be specified using the 'nimpute='. The default number is 5. The option 'round=' controls the number of decimal places in the imputed values, with no rounding by default. If more than one number is specified, one should use the 'VAR' statement and the specified number must correspond to the number of variables in the 'VAR' statement. The option 'seed=' is used to specify a positive integer which is used by PROC MI to start the pseudo-random number generator. The default number is generated from the time and date of the computer's clock.

The imputation task is carried out separately for each level of the variables specified in the 'BY' statement. One can also choose between three different methods using the option 'method='. When the missingness is confined to dropout the 'MONOTONE' statement can be used, but does not have to be used since the parametric regression 'method=reg' or the nonparametric propensity method, 'method=propensity' can be used. For general missingness patterns the 'MCMC' statement can be used, which is the default as well. There are numerous options within the MCMC method such as 'ngroups=', 'initial=', 'pmm=' etc. The 'propensity' and 'regression' methods are used for incomplete continuous outcomes, incomplete categorical outcomes can be imputed by including them into the 'CLASS' statement, in addition to the inclusion of the 'VAR' statement. In such a case the 'MONOTONE' option

should be used and one can make use of the logistic regression and discriminant analysis imputation by means of the options 'logistic' and 'discrim' respectively.

With the default 'initial=EM' option the procedure uses the means and standard deviations from available cases as the initial estimates for the EM Algorithm. The final estimates after applying the EM algorithm are then used to start the MCMC process. One can also use a SAS input data set as the initial estimates of the mean and covariance matrix after each imputation using 'initial=input name' option. The option 'niter=' controls the number of iterations between imputations in a single chain with the default being 100.

PROC MIANALYZE

PROC MIANALYZE combines the m inferences into a single one. Parameter and standard error are passed on using the options 'data=', 'parms=', 'covb=', 'xpxi=' to the PROC MIANALYZE statement. The option 'data=' data sets of the types COV, CORR or EST can also be passed on. If one wants to pass on parameter estimates and variance-covariance matrices, it is better to use 'parms=' and 'covb=' or 'parms=' and 'xpxi='. Within the PROC MIANALZE statement one can also get the within-imputation, between-imputation and total covariance matrices using the 'wcov', 'bcov' and 'tcov'. The parameter or effects for which multiple imputation inference is needed can be passed on using the MODELEFFECTS statement (previously the VAR statement). Categorical effects can be handled using the CLASS statment by creating appropriate dummy variables. The 'TEST' statement is used to test hypotheses about linear combinations of the parameters.

9.8.4 Multiple Imputation by fully conditional specification or by chained equations

The goal of multiple imputation, as discussed earlier on in Section 8.12, is to provide valid inference for the statistical analysis of missing data. There are, according to van Buuren (2007) two approaches to imputation of multivariate data namely joint modelling (JM) and fully conditional specification (FCS). The FCS approach is also termed multiple imputation via chained equations (MICE). Both the FCS and JM approaches have attractive features but also certain disadvantages. JM is theoretically sound and is based on parametric statistical theory. However the JM approach may incur bias in the sense that the joint model may lack the flexibility needed to represent typical data features. van Buuren (2007) states that FCS or multiple imputation using chained equations is a semi parametric and flexible alternative that specifies the multivariate model by a series of conditional models, one for each incomplete variable. FCS has tremendous flexibility and is easy to apply but its statistical properties are difficult to establish. We firstly look at some of the univariate imputation methods before describing the theory of FCS and the sofware, MICE, that was used for our imputation to handle the dropout in the Kilifi data set. The following table, taken directly from van Buuren (2007, p. 226), is a summary of the imputation methods in univariate missing data problems.

Rubin and Schafer (1990) first elaborate on the JM approach. The JM approach partitions the observations into groups of identical missing data patterns and imputes the missing entries within each pattern according to a joint model for X, Y and R (X, Y and R will be defined below). We will not describe the JM approach as it was not used to handle the dropout values in the Kilifi data set, the FCS approach was used and we now describe

Type of variable	Method	References
Ignorable methods		
Continuous	Linear regression	Rubin (1987)
		Schenker and Taylor (1986)
	Linear regression + empirical residual	Rubin (1987)
		Schenker and Taylor (1986)
	Predictive mean matching	Rubin (1986)
		Little(1988)
		Schenker and Taylor (1986)
	Nonlinear regression	Harrell (2001)
	Truncated normal model	Schafer (1997, p. 204)
Binary	Logistic regression	Rubin (1987, p.169)
	Probit regression	Albert and Chib (1993)
	Measurement error and reporting model	Yucel and Zaslavsky (2005)
Categorical	Polytomous logistic regression	Brand et al. (2003)
	Discriminant analysis	Brand (1999)
Semi-continuous	Two step: logistic + linear	Rubin (1987, p. 180)
	General location model	
Counts	Poisson regression	Raghunathan et al.(2001)
General	Approximate Bayesian bootstrap	Rubin (1987)
		Parzen et al. (2005)
	Hot deck	Reily and Pepe (1997)
	Machine learning methods	Junninen (2004)
	Polya tree	Paddock (2002)
Nonignorable methods		
Continuous	Normal selection model	Heckman (1976)
	Logit selection model	Greenlees et al. (2002)
Censored data	Data augmentation	Wei and Tanner (1991)
Clustered censored data	GEE	Pan and Connett (2001)
Interval censored	Proportional hazard model	Goetghebeur and Ryan (2000)
		Pan (2000)
Limited dependent variables	DeFries-Fulker model	Bechger et al. (2002)
Below detection limit	Custom model	Hopke et al. (2001)
		Lubin et al. (2004)
Pedigree relations	Custom model	Fridley (2003)
Bracketed responses	Custom model	Heeringa et al. (2002)

Table 9.22: Overview of imputation methods in univariate missing data problems

the method, following van Buuren (2007, p. 227). First Table 9.22 gives an overview of imputation methods in univariate data problems.

The basic idea of FCS is quite old and is known by a variety of names that include: stochastic relaxation, variable-by-variable imputation, regression switching, sequential regressions, ordered pseudo-Gibbs sampler, partially incompatible MCMC, iterated univariate imputation and **chained equations**.

We define our notation, as follows: Let Y_j be one of k incomplete random variables where $j = 1, \ldots, k$ and let $Y = (Y_1, \ldots, Y_k)$. The observed and missing parts of Y_j are denoted, as previously, as Y_j^{obs} and Y_j^{mis} respectively, so that $Y_j^{obs} = (Y_1^{obs}, \ldots, Y_k^{obs})$ and $Y_j^{mis} = (Y_1^{mis}, \ldots, Y_k^{mis})$ stand for the observed and missing data in Y. Let $Y_{-j} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_k)$ denote the collection of the $k-1$ variables in Y except Y_j. Let R_j be the response indicator of Y_j, with $R_j = 1$ if Y_j is observed and $R_j = 0$ if Y_j is missing. Let $R = (R_1, \ldots, R_k)$ and $R_{-j} = (R_1, \ldots, R_{j-1}, R_{j+1}, \ldots, R_k)$. Let $X = (X_1, \ldots, X_l$ be a set of l complete covariates on the same subjects. We also assume that the observations in Y, X and R correspond to a simple random sample from the population of interest.

According to van Buuren (2007, p. 227) the FCS approach is to impute the data on a variable-by-variable basis by specifying an imputation model per variable. FCS is an attempt to define $P(Y, X, R|\theta)$ by specifying a conditional density $P(Y_j|X, Y_{-j}, R, \theta_j)$ for each Y_j. This density is used to impute Y_j^{mis} given X, Y_{-j} and R. Starting from simple guessed values, imputation under FCS is done by iterating over all conditionally specified imputation models. The building blocks of the methods are listed in Table 9.22. One iteration consists of one cycle through all Y_j. If the joint distribution defined by the specified conditional distribution exists, then this process is a Gibbs sampler.

FCS allows great flexibility in creating multivariate models. One can easily specify models that are outside any known standard multivariate density $P(Y, X, R|\theta)$. FCS can use specialized imputation methods that are difficult to formulate as a part of a multivariate density $P(Y, X, R|\theta)$. Imputation methods that preserve unique features in the data, such as bounds and skip patterns can be incorporated. It also must be said that it is easy to maintain

constraints between different variables in order to avoid logical inconsistencies in the imputed data. Such constraints would be difficult to formulate in terms of the multivariate density $P(Y, X, R|\theta)$. Each conditional density has to be specified separately, hence some modelling effort may be required on the part of the user.

The software that is available for creating multiple imputations by FCS include FRITZ, IVEWARE in SAS, HERMES missing data engine, MICE in S-PLUS and R and ICE, an implementation of MICE in Stata.

9.9 Application to the Kilifi RSV data

The dropout values in the Kilifi data set were imputed using MICE in S-Plus. The following table summarizes the type of imputation that was done per variable.

Variable	Type	Method	MICE command
Age	categorical	polytomous regression	polyreg
rsv status	binary	logistic regression	logreg
actipass	binary	logistic regression	logreg
dt	continuous	predictive mean matching or linear regression	pmm
prev	continuous	predictive mean matching or linear regression	pmm
timemonth	continuous	predictive mean matching or linear regression	pmm

Table 9.23: Overview of imputation methods used for the Kilifi data in MICE

After the MICE software was run the imputed data sets were pooled into a final one and this data set was re-analysed using GEEs, GLMM's and the force of infection was re-estimated for the months and overall using the likelihood method and the generalized linear model methods. The results are summarized and presented in tables that follow.

MICE Generalized Estimating Equations-GEEs

Correlation Type	Source	DF	Chi-Square	Pr > Chi-Sq
Exchangeable	age	12	33.93	0.0007
	dt	1	1.65	0.1986
	prev	1	22.42	< .0001
	timemonth	1	4.63	0.0314
	actipass	1	51.27	<.0001
Independent	age	12	33.98	0.0007
	dt	1	1.69	0.1938
	prev	1	22.59	< .0001
	timemonth	1	4.48	0.0343
	actipass	1	51.08	<.0001
AR(1)	age	12	33.85	0.0007
	dt	1	1.70	0.1925
	prev	1	22.79	< .0001
	timemonth	1	4.49	0.0341
	actipass	1	50.50	<.0001

Table 9.24: MICE-Score statistics for Type III GEE

Table 9.24 shows the type III score statistics indicate that the age, prev and actipass variables to be significant at the 5% level. The magnitude of the estimates are quite similar in the the three correlation structures.

	Exchangeable			Independent			AR(1)		
Parameter	Estimate	Standard Error	Pr> |Z|	Estimate	Standard Error	Pr> |Z|	Estimate	Standard Error	Pr> |Z|
Intercept	-3.6386	0.7990	< .0001	-3.6708	0.8084	< .0001	-3.6743	0.8116	< .0001
age 0	-1.2457	0.8495	0.1425	-1.2135	0.8556	0.1561	-1.2173	0.8599	0.1569
age 1	-1.0113	0.6442	0.1164	-0.9919	0.6525	0.1285	-0.987	0.6553	0.132
age 2	-0.8063	0.6107	0.1867	-0.7876	0.6177	0.2023	-0.7855	0.6204	0.2055
age 3	-0.6351	0.5705	0.2656	-0.6195	0.5763	0.2824	-0.6157	0.5787	0.2874
age 4	-0.9	0.5605	0.1083	-0.8851	0.5648	0.1171	-0.8816	0.5672	0.1201
age 5	-2.3722	0.827	0.0041	-2.3732	0.8342	0.0044	-2.3574	0.8342	0.0047
age 6	-0.9628	0.5798	0.0968	-0.9473	0.5812	0.1031	-0.9417	0.5829	0.1062
age 7	-1.7064	0.7636	0.0254	-1.7027	0.7665	0.0263	-1.7197	0.7753	0.0266
age 8	-0.6609	0.4432	0.136	-0.6534	0.444	0.1411	-0.6454	0.4446	0.1466
age 9	-0.3512	0.3627	0.3329	-0.3467	0.3635	0.3401	-0.3511	0.3655	0.3367
age 10	0.4163	0.2562	0.1042	0.4208	0.2567	0.1012	0.4245	0.2576	0.0994
age 11	-0.095	0.2449	0.6979	-0.0955	0.2451	0.6969	-0.0971	0.246	0.6931
age 12	0	0	.	0	0	.	0	0	.
dt	0.0092	0.0053	0.0807	0.0093	0.0052	0.0765	0.0093	0.0052	0.0751
prev	30.9359	5.2662	< .0001	31.0184	5.2754	< .0001	31.07	5.2918	< .0001
timemonth	-0.149	0.0583	0.0106	-0.1467	0.0591	0.013	-0.1466	0.0593	0.0135
actipass 0	1.4236	0.1388	< .0001	1.4233	0.1394	< .0001	1.4188	0.1395	< .0001
actipass 1	0	0	.	0	0	.	0	0	.

Table 9.25: MICE-Model based standard errors and estimates

The algorithm for the unstructured correlation matrix option did not converge and the results are omitted. The results of the model based estimates and standard errors are not very different between the three correlation structures. The magnitude of the estimates are somewhat similar. Moreover, we see that the model based and the empirical parameter estimates in Tables refvelo and 9.26 are not very different in magnitude. This is a feature of GEE because the choice between naive and empirical only affects the estimation of the covariance matrix of the regression parameter $\boldsymbol{\beta}$. The output for the correlation between two repeated measurement for the exchangeable

	Exchangeable			Independent			AR(1)		
Parameter	Estimate	Standard Error	Pr> |Z|	Estimate	Standard Error	Pr> |Z|	Estimate	Standard Error	Pr> |Z|
Intercept	-3.6386	0.8943	< .0001	-3.6708	0.9079	< .0001	-3.6743	0.9127	< .0001
age 0	-1.2457	1.0429	0.2323	-1.2135	1.0468	0.2463	-1.2173	1.0494	0.2461
age 1	-1.0113	0.7019	0.1496	-0.9919	0.7122	0.1637	-0.987	0.7155	0.1678
age 2	-0.8063	0.6766	0.2333	-0.7876	0.6864	0.2512	-0.7855	0.6901	0.2551
age 3	-0.6351	0.6287	0.3124	-0.6195	0.6377	0.3314	-0.6157	0.6413	0.337
age 4	-0.9	0.6227	0.1484	-0.8851	0.6312	0.1609	-0.8816	0.635	0.1651
age 5	-2.3722	0.8217	0.0039	-2.3732	0.8342	0.0044	-2.3574	0.8332	0.0047
age 6	-0.9628	0.6493	0.1382	-0.9473	0.6532	0.147	-0.9417	0.6546	0.1503
age 7	-1.7064	0.7875	0.0302	-1.7027	0.7915	0.0315	-1.7197	0.7965	0.0308
age 8	-0.6609	0.452	0.1437	-0.6534	0.4524	0.1486	-0.6454	0.4526	0.1538
age 9	-0.3512	0.4236	0.407	-0.3467	0.4232	0.4126	-0.3511	0.424	0.4076
age 10	0.4163	0.2784	0.1348	0.4208	0.2789	0.1313	0.4245	0.2793	0.1286
age 11	-0.095	0.2456	0.6987	-0.0955	0.2468	0.6989	-0.0971	0.2479	0.6953
age 12	0	0	.	0	0	.	0	0	.
dt	0.0092	0.0062	0.1362	0.0093	0.0061	0.1301	0.0093	0.0061	0.1284
prev	30.9359	5.3218	< .0001	31.0184	5.3155	< .0001	31.07	5.3278	< .0001
timemonth	-0.149	0.0642	0.0203	-0.1467	0.0652	0.0243	-0.1466	0.0655	0.0252
actipass 0	1.4236	0.1479	< .0001	1.4233	0.1484	< .0001	1.4188	0.1487	< .0001
actipass 1	0	0	.	0	0	.	0	0	.

Table 9.26: MICE-Empirical based standard errors and estimates

correlation matrix was found to be -0.00035. A possible reason why the unstructured correlation matrix did not attain convergence is because the observations can not be aligned. At the 5% significance level there were differences between age group 5 and 12 and age group 7 and 12 with respect to influencing whether a child is infected or not. The prev variable and a difference between whether a child was actively or passively sampled (actipass 0 versus actipass 1) was also significant at the 5% level in influencing whether a child is infected or not. These significant differences was present for all 3 correlation structures and are denoted by p-values less than 0.05. It

is also worthwhile noting that the exchangeable and independent correlation structures have their empirical standard errors slightly closer to the model based standard errors than the AR(1) correlation structure.

MICE-Estimate the force of infection and the prevalence rate

The parameters λ and ν were then estimated using the Fisher's scoring method to yield the following estimates along with their standard errors:

Estimator	Estimate	Standard Error
$\hat{\lambda}$	0.00092	0.000007334
$\hat{\nu}$	0.49634	0.005585

Table 9.27: MICE-Parameter Estimates

Hence a 95% confidence interval for λ is $(0.000905, 0.000933)$ and likewise for ν is $(0.4854, 0.5073)$. Table9.27 gives the estimates of the force of infection and rate of recovery using the maximum likelihood approach and the estimates are consistent to those from the available data.
Using the GLM approach we found the following estimates:

Estimator	Estimate	Standard Error
$\hat{\lambda}$	0.001202	0.0000829
$\hat{\nu}$	0.62983	0.088366

Table 9.28: MICE-GLM Parameter Estimates

Hence a 95% confidence interval for λ is $(0.0010398, 0.001365)$ and likewise for ν is $(0.4566, 0.8030)$. Table 9.28 gives the estimates of λ and ν, but is worthwhile noting that the estimate of ν is slightly higher than the maximum likelihood approach.

The monthly force of infection together with the plots are given in Table 9.29 and Figure 9.1.

		Exponentiation		Delta Method	
Month	Lambda	95% Confidence Interval		95% Confidence Interval	
$\hat{\lambda}_2$	0.0045	0.0027	0.0075	0.0022	0.0068
$\hat{\lambda}_3$	0.0033	0.0023	0.0048	0.0021	0.0045
$\hat{\lambda}_4$	0.0033	0.0023	0.0047	0.0021	0.0045
$\hat{\lambda}_5$	0.0011	0.0006	0.0019	0.0004	0.0017
$\hat{\lambda}_6$	0.0005	0.0002	0.0013	0.0000	0.0010
$\hat{\lambda}_7$	0.0007	0.0003	0.0015	0.0001	0.0012
$\hat{\lambda}_8$	0.0002	0.0001	0.0008	0.0001	0.0005
$\hat{\lambda}_9$	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\lambda}_{10}$	0.0002	0.0001	0.0008	0.0001	0.0005
$\hat{\lambda}_{11}$	0.0016	0.0011	0.0024	0.0010	0.0022
$\hat{\lambda}_{12}$	0.0019	0.0014	0.0025	0.0014	0.0024
$\hat{\lambda}_{13}$	0.0008	0.0006	0.0011	0.0005	0.0011

Table 9.29: MICE-Monthly estimates of the force of infection and confidence Intervals

Once again, we see the time effect of the Respiratory Syncytial Virus, with peaks and troughs in the monthly estimates as depicted in Figure 9.1.

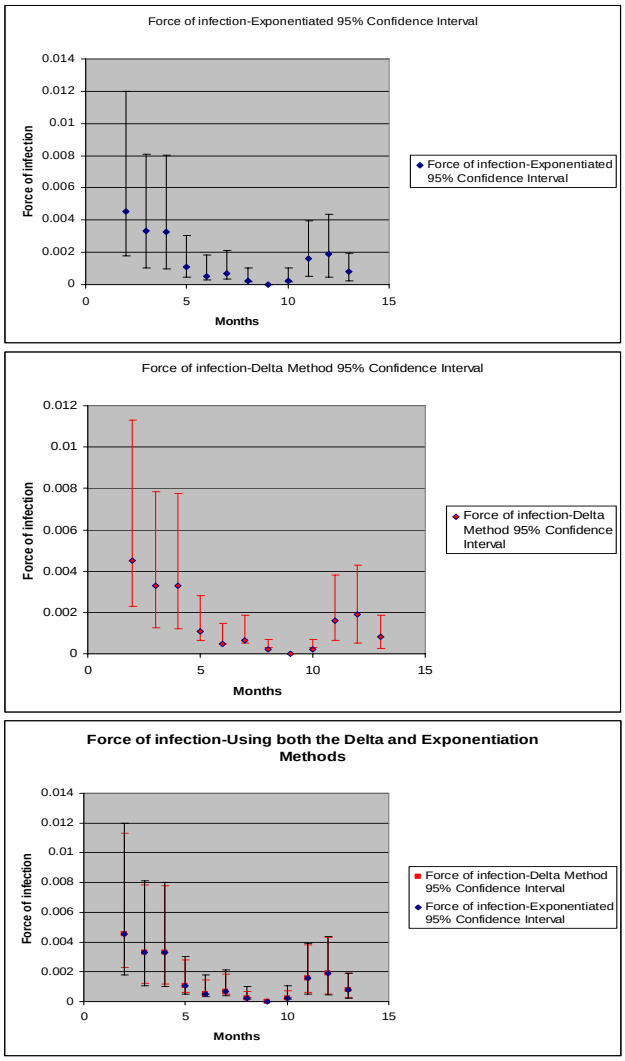

Figure 9.1: MICE-The force of infection in months together with 95% confidence intervals using the exponentiated and delta methods.

MICE-Generalized Linear Mixed Model

In this section MICE was combined with GLMM approach. The generalized linear mixed model was fitted yielding results in the tables below. Different covariance structure models were fitted: The solution for the fixed effects

Covariance Structure		Estimate	Standard Error
Unstructured	UN(1,1)	0.000	0.000
	Residual(VC)	1.1233	0.01303
Compound symmetry	Var(child)	0.000	0.000
	CS(child)	-6.36E-6	2.042E-6
	Residual(VC)	1.1388	0.01330
Power	Var(child)	0.000	0.000
	SP(POW)(child)	0.000	0.000
	Residual(VC)	1.1233	0.01303
Spherical	Var(child)	0.000	0.000
	SP(SPH)(child)	0.000	0.000
	Residual(VC)	1.1233	0.01303
Gaussian	Var(child)	0.000	0.000
	SP(GAU)(child)	0.000	0.000
	Residual(VC)	1.1233	0.01303

Table 9.30: MICE-Covariance Parameter Estimates random effects model

for all the different covariance structure models are given in Tables 9.31 and 9.32.

Here again we find the prev and actipass variables are significant at the 5% level and the age variable is tending towards significance. Furthermore there

Effect	Estimate	Standard Error	Pr> $\|t\|$
Intercept	-3.6708	0.8568	< .0001
age 0	-1.2135	0.9068	0.1808
age 1	-0.9919	0.6916	0.1515
age 2	-0.7876	0.6546	0.2289
age 3	-0.6195	0.6108	0.3105
age 4	-0.8851	0.5986	0.1393
age 5	-2.3732	0.8841	0.0073
age 6	-0.9473	0.616	0.1241
age 7	-1.7027	0.8124	0.0361
age 8	-0.6534	0.4706	0.165
age 9	-0.3467	0.3852	0.3681
age 10	0.4208	0.2721	0.122
age 11	-0.09548	0.2598	0.7133
age 12	0	.	.
dt	0.009293	0.005559	0.0946
prev	31.0184	5.5911	< .0001
actipass 0	1.4233	0.1478	< .0001
actipass 1	0	.	.
timemonth	-0.1467	0.06259	0.0191

Table 9.31: MICE-Solution for the fixed effects random effects model

Effect	F-Value	P-value
Age	1.77	0.0464
Dt	2.79	0.0946
Prev	30.78	< 0.0001
Actipass	92.77	< 0.0001
Timemonth	5.50	0.0191

Table 9.32: MICE-Type III Effects

is a significant difference in the age 5, 7 and age 12 groups with respect to whether a child is infected or not.

The estimates of the fixed effects for the random effects model with the compound symmetry covariance structure is slightly different from the above estimates:

Effect	Estimate	Standard Error	Pr> $\|t\|$
Intercept	-3.611	0.8455	$< .0001$
age 0	-1.2565	0.894	0.1599
age 1	-1.0152	0.6773	0.1339
age 2	-0.8191	0.6435	0.2031
age 3	-0.6506	0.6025	0.2802
age 4	-0.9092	0.5932	0.1254
age 5	-2.3972	0.8858	0.0068
age 6	-0.9794	0.6184	0.1133
age 7	-1.7025	0.8133	0.0363
age 8	-0.6576	0.4702	0.1619
age 9	-0.336	0.3847	0.3825
age 10	0.4209	0.2736	0.1239
age 11	-0.0973	0.2628	0.7112
age 12	0	.	.
dt	0.009008	0.005622	0.1091
prev	30.5458	5.6195	$< .0001$
actipass 0	1.4295	0.1486	$< .0001$
actipass 1	0	.	.
timemonth	-0.1521	0.06122	0.013

Table 9.33: MICE-Compound symmetry solution for the fixed effects

Effect	F-Value	P-value
Age	1.77	0.0464
Dt	2.57	0.1091
Prev	29.55	< 0.0001
Actipass	92.53	< 0.0001
Timemonth	6.17	0.013

Table 9.34: MICE-Compound symmetry Type III Effects

The random intercept model was also fitted with the following results: The

Covariance Structure		Estimate	Standard Error
Unstructured	UN(1,1)	0.000	0.000
	Residual(VC)	1.1233	0.01303
Compound symmetry	Var(child)	6.746E-07	0.1077
	CS(child)	-0.2760	.
	Residual(VC)	1.1303	0.01325
Power	Var(child)	0.000	0.000
	SP(POW)(child)	0.000	0.000
	Residual(VC)	1.1233	0.01303
Spherical	Var(child)	0.000	0.000
	SP(SPH)(child)	0.000	0.000
	Residual(VC)	1.1233	0.01303
Gaussian	Var(child)	0.000	0.000
	SP(GAU)(child)	0.000	0.000
	Residual(VC)	1.1233	0.01303

Table 9.35: MICE-Random Intercept Covariance Parameter Estimates

solution for the fixed effects for all the different covariance structure models are:

Effect	Estimate	Standard Error	Pr> $\|t\|$
Intercept	-3.6708	0.8568	< .0001
age 0	-1.2135	0.9068	0.1808
age 1	-0.9919	0.6916	0.1515
age 2	-0.7876	0.6546	0.2289
age 3	-0.6195	0.6108	0.3105
age 4	-0.8851	0.5986	0.1393
age 5	-2.3732	0.8841	0.0073
age 6	-0.9473	0.616	0.1241
age 7	-1.7027	0.8124	0.0361
age 8	-0.6534	0.4706	0.165
age 9	-0.3467	0.3852	0.3681
age 10	0.4208	0.2721	0.122
age 11	-0.09548	0.2598	0.7133
age 12	0	.	.
dt	0.009293	0.005559	0.0946
prev	31.0184	5.5911	< .0001
actipass 0	1.4233	0.1478	< .0001
actipass 1	0	.	.
timemonth	-0.1467	0.06259	0.0191

Table 9.36: MICE-Random Intercept solution for the fixed effects

Effect	F-Value	P-value
Age	1.77	0.0464
Dt	2.79	0.0946
Prev	30.78	< 0.0001
Actipass	92.77	< 0.0001
Timemonth	5.50	0.0191

Table 9.37: MICE-Type III Effects for random intercept model

Here again we find the prev and actipass variables are significant at the 5% level and the age variable is tending towards significance. Furthermore there is a difference in the age 5, 7 and age 12 groups with respect to whether a

child is infected or not. The estimates of the fixed effects for the random effects model with the compound symmetry covariance structure is slightly different from the above estimates. The results under the CS structure are given in Tables 9.38 and 9.39.

Effect	Estimate	Standard Error	Pr> $\|t\|$
Intercept	-3.611	0.8455	< .0001
age 0	-1.2565	0.894	0.1599
age 1	-1.0152	0.6773	0.1339
age 2	-0.8191	0.6435	0.2031
age 3	-0.6506	0.6025	0.2802
age 4	-0.9092	0.5932	0.1254
age 5	-2.3972	0.8858	0.0068
age 6	-0.9794	0.6184	0.1133
age 7	-1.7025	0.8133	0.0363
age 8	-0.6576	0.4702	0.1619
age 9	-0.336	0.3847	0.3825
age 10	0.4209	0.2736	0.1239
age 11	-0.0973	0.2628	0.7112
age 12	0	.	.
dt	0.009008	0.005622	0.1091
prev	30.5458	5.6195	< .0001
actipass 0	1.4295	0.1486	< .0001
actipass 1	0	.	.
timemonth	-0.1521	0.06122	0.013

Table 9.38: MICE-Random intercept model Compound symmetry solution for the fixed effects of the Optimal Model

Effect	F-Value	P-value
Age	1.77	0.0464
Dt	2.57	0.1091
Prev	29.55	< 0.0001
Actipass	92.53	< 0.0001
Timemonth	6.17	0.013

Table 9.39: MICE-Random intercept model Compound symmetry Type III Effects for Optimal Model

9.9.1 Results for using LOCF to handle the dropout

Using LOCF, we fitted GEE models with the different correlation structure and the results are summarized in Tables 9.40 and 9.41 below:

	Exchangeable			Independent			AR(1)		
Parameter	Est.	Std. Error	Pr> |Z|	Est.	Std. Error	Pr> |Z|	Est.	Std. Error	Pr> |Z|
Intercept	-5.024	1.239	< .0001	-4.917	1.083	< .0001	-5.285	1.252	< .0001
age 0	-0.739	1.013	0.466	-0.887	0.897	0.323	-0.942	1.123	0.401
age 1	-0.679	0.897	0.449	-0.706	0.772	0.360	-0.313	0.926	0.736
age 2	-0.344	0.859	0.689	-0.617	0.748	0.410	-0.127	0.888	0.887
age 3	-0.247	0.802	0.758	-0.407	0.702	0.562	0.145	0.830	0.861
age 4	-0.623	0.766	0.416	-0.643	0.685	0.348	-0.540	0.810	0.505
age 5	-2.533	0.994	0.011	-2.448	0.920	0.008	-1.965	0.986	0.046
age 6	-2.338	0.882	0.008	-2.639	0.863	0.002	-1.447	0.780	0.064
age 7	-2.999	1.164	0.010	-2.857	1.063	0.007	-2.425	1.009	0.016
age 8	-1.514	0.411	0.000	-1.574	0.383	< .0001	-0.878	0.413	0.034
age 9	-1.287	0.314	< .0001	-1.574	0.299	< .0001	-1.200	0.388	0.002
age 10	-0.272	0.235	0.247	-0.213	0.212	0.317	-0.234	0.292	0.424
age 11	-1.268	0.263	< .0001	-1.285	0.241	< .0001	-0.786	0.299	0.009
age 12	0.000	0.000	.	0.000	0.000	.	0.000	0.000	.
dt	0.006	0.004	0.138	0.004	0.004	0.252	0.008	0.004	0.045
prev	48.958	6.275	< .0001	48.263	6.011	< .0001	42.704	6.346	< .0001
timemonth	-0.046	0.091	0.615	-0.041	0.079	0.605	-0.006	0.092	0.951
actipass 0	2.707	0.147	< .0001	2.628	0.125	< .0001	2.188	0.142	< .0001
actipass 1	0.000	0.000	.	0.000	0.000	.	0.000	0.000	.

Table 9.40: LOCF-Model based standard errors and estimates

	Exchangeable			Independent			AR(1)		
Parameter	Est.	Std. Error	Pr> |Z|	Est.	Std. Error	Pr> |Z|	Est.	Std. Error	Pr> |Z|
Intercept	-5.024	1.466	0.001	-4.917	1.358	0.000	-5.285	1.416	0.000
age 0	-0.739	1.428	0.605	-0.887	1.396	0.525	-0.942	1.484	0.526
age 1	-0.679	1.224	0.579	-0.706	1.121	0.529	-0.313	1.197	0.794
age 2	-0.344	1.160	0.767	-0.617	1.070	0.564	-0.127	1.161	0.913
age 3	-0.247	1.109	0.824	-0.407	1.041	0.696	0.145	1.069	0.892
age 4	-0.623	1.010	0.537	-0.643	0.951	0.499	-0.540	1.036	0.602
age 5	-2.533	1.451	0.081	-2.448	1.318	0.063	-1.965	1.106	0.076
age 6	-2.338	1.016	0.021	-2.639	1.035	0.011	-1.447	0.929	0.119
age 7	-2.999	1.333	0.024	-2.857	1.135	0.012	-2.425	0.955	0.011
age 8	-1.514	0.832	0.069	-1.574	0.749	0.036	-0.878	0.733	0.231
age 9	-1.287	0.718	0.073	-1.574	0.748	0.035	-1.200	0.706	0.089
age 10	-0.272	0.575	0.636	-0.213	0.531	0.689	-0.234	0.599	0.696
age 11	-1.268	0.599	0.034	-1.285	0.547	0.019	-0.786	0.593	0.185
age 12	0.000	0.000	.	0.000	0.000	.	0.000	0.000	.
dt	0.006	0.009	0.523	0.004	0.010	0.664	0.008	0.007	0.242
prev	48.958	8.202	< .0001	48.263	8.623	< .0001	42.704	7.194	< .0001
timemonth	-0.046	0.100	0.648	-0.041	0.091	0.652	-0.006	0.096	0.953
actipass 0	2.707	0.244	< .0001	2.628	0.250	< .0001	2.188	0.184	< .0001
actipass 1	0.000	0.000	.	0.000	0.000	.	0.000	0.000	.

Table 9.41: LOCF-Empirical based standard errors and estimates

The results show that at the 5% significance level, 'age 5' versus 'age 12', 'age 6' versus 'age 12', 'age 7' versus 'age 12','age 8' versus 'age 12','age 9' versus 'age 12','age 11' versus 'age 12','prev' and 'actipass 0 versus actipass

1' are all significant with respect to the disease state of the child for both the exchangeable and independent correlation structures. However for the AR(1) structure, all the same variables stated above ('age 5' versus 'age 12', 'age 7' versus 'age 12','age 8' versus 'age 12','age 9' versus 'age 12','age 11' versus 'age 12','prev' and 'actipass 0 versus actipass 1') are significant except for'age 6' versus 'age 12'. There are not huge differences between the model based and empirical standard errors for the exchangeable and independent correlation structures, however for the AR(1) the differences between the standard errors are slightly bigger. The LOCF-GEE give more significant individual effects than the MICE-GEE due to the fact that LOCF artificially inflates the correlation between successive time point.

Correlation Type	Source	DF	Chi-Square	Pr > Chi-Sq
Exchangeable	age	12	25.66	0.0120
	dt	1	0.26	0.6128
	prev	1	18.28	< .0001
	timemonth	1	0.23	0.6284
	actipass	1	36.90	< .0001
Independent	age	12	22.76	0.029
	dt	1	0.15	0.7012
	prev	1	16.01	< .0001
	timemonth	1	0.20	0.6514
	actipass	1	29.67	< .0001
AR(1)	age	12	26.46	0.0092
	dt	1	0.69	0.4055
	prev	1	16.45	< .0001
	timemonth	1	0.00	0.9552
	actipass	1	54.02	< .0001

Table 9.42: LOCF-Score statistics for Type III GEE

The type III score statistics show that the age, prev and actipass variables to be significant at the 5% level. The magnitude of the estimates are similar in the the exchangeable and independent correlation structures but different when compared to the AR(1) correlation structure. A random intercept and random effects model was also fitted using LOCF and yielded the following results shown in Table 9.43.

LOCF-Random intercept model

Different covariance structure models were again fitted: The solution for the

Covariance Structure		Estimate	Standard Error
Unstructured	UN(1,1)	No convergence	No convergence
	Residual(VC)	No convergence	No convergence
Compound symmetry	Var(child)	0.005947	0.2598
	CS(child)	2.2205	.
	Residual(VC)	0.9475	0.0111
Power	Var(child)	2.2263	0.2598
	SP(POW)(child)	1.000	.
	Residual(VC)	0.9475	0.0111
Spherical	Var(child)	2.2263	0.2598
	SP(SPH)(child)	94.6551	.
	Residual(VC)	0.9475	0.0111
Gaussian	Var(child)	2.2263	0.2598
	SP(GAU)(child)	7.6551	.
	Residual(VC)	0.9475	0.0111

Table 9.43: Covariance Parameter Estimates in a random intercept -LOCF

fixed effects for all the different covariance structure models are: The results show that at the 5% significance level, 'age 5' versus 'age 12', 'age 6' versus 'age 12', 'age 7' versus 'age 12','age 8' versus 'age 12','age 9' versus 'age 12'(tending towards significance),'age 11' versus 'age 12','prev' and 'actipass 0 versus actipass 1' are all significant with respect to the disease state of the child for all correlation structures. The results for the random intercept model show that the , 'age','prev' and 'actipass' variables are significant at the 5% level.

Effect	Estimate	Standard Error	Pr> $\|t\|$
Intercept	-5.150	1.375	0.000
age 0	-0.448	1.138	0.694
age 1	-0.546	1.033	0.597
age 2	-0.009	0.969	0.993
age 3	0.133	0.902	0.883
age 4	-0.287	0.849	0.735
age 5	-2.173	1.025	0.034
age 6	-1.938	0.892	0.030
age 7	-2.640	1.173	0.024
age 8	-1.076	0.470	0.022
age 9	-0.714	0.372	0.055
age 10	-0.163	0.284	0.567
age 11	-0.886	0.286	0.002
age 12	0.000	.	.
dt	-0.003	0.005	0.616
prev	51.558	6.479	$< .0001$
actipass 0	2.612	0.138	$< .0001$
actipass 1	0.000	.	.
timemonth	-0.048	0.101	0.632

Table 9.44: Solution for the fixed effects of the random intercept model - LOCF

Effect	F-Value	P-value
Age	3.76	$< .0001$
Dt	0.25	0.6163
Prev	63.33	< 0.0001
Actipass	360.65	< 0.0001
Timemonth	0.23	0.6321

Table 9.45: Type III Effects for random intercept model-LOCF

LOCF-Random effects model

Different covariance structure models were again fitted: The solution for the fixed effects for all the different covariance structure models are: The results

Covariance Structure		Estimate	Standard Error
Unstructured	UN(1,1)	0.000078	0.000012
	Residual(VC)	1.0147	0.01196
Compound symmetry	Var(child)	2.878E-7	0.000012
	CS(child)	0.000077	.
	Residual(VC)	1.0147	0.01196
Power	Var(child)	0.000078	0.000012
	SP(POW)(child)	1.000	.
	Residual(VC)	1.0147	0.01196
Spherical	Var(child)	0.000078	0.000012
	SP(SPH)(child)	84.7272	.
	Residual(VC)	1.0147	0.01196
Gaussian	Var(child)	0.000078	0.000012
	SP(GAU)(child)	5.9208	.
	Residual(VC)	1.0147	0.01196

Table 9.46: Covariance Parameter Estimates for random effects model-LOCF

show that at the 5% significance level, 'age 5' versus 'age 12', 'age 6' versus 'age 12', 'age 7' versus 'age 12','age 8' versus 'age 12','age 9' versus 'age 12'(tending towards significance),'age 11' versus 'age 12','prev' and 'actipass 0 versus actipass 1' are all significant with respect to the disease state of the child for all correlation structures. The results, not surprisingly for the random effects model are the same as the random intercept model and show that, 'age','prev' and 'actipass' variables are significant at the 5% level.

Effect	Estimate	Standard Error	Pr> $\|t\|$
Intercept	-5.622	1.451	0.000
age 0	-0.179	1.217	0.883
age 1	-0.241	1.102	0.827
age 2	0.013	1.043	0.990
age 3	0.219	0.963	0.820
age 4	-0.189	0.911	0.836
age 5	-2.071	1.087	0.057
age 6	-1.990	0.952	0.037
age 7	-2.726	1.250	0.029
age 8	-1.215	0.489	0.013
age 9	-0.872	0.370	0.018
age 10	-0.345	0.269	0.200
age 11	-1.200	0.271	$< .0001$
age 12	0.000	.	.
dt	-0.005	0.005	0.397
prev	52.101	6.744	$< .0001$
actipass 0	2.608	0.137	$< .0001$
actipass 1	0.000	.	.
timemonth	0.011	0.110	0.921

Table 9.47: Solution for the fixed effects for the random effects model -LOCF

Effect	F-Value	P-value
Age	4.12	$< .0001$
Dt	0.72	0.3973
Prev	59.69	< 0.0001
Actipass	360.34	< 0.0001
Timemonth	0.01	0.9209

Table 9.48: Type III Effects for random effects model-LOCF

The random effects models were also fitted using the PROC NLMIXED procedure in SAS. The models were fitted using adaptive and non-adaptive Gaussian quadrature with $3, 5$ and 20 quadrature points. The model fitted was:

$rsvpos = \beta_{00} + \beta_0 age0 + \beta_1 age1 + beta_2 age2 + \beta_3 age3 + \beta_4 age4 + \beta_5 age5 + \beta_6 age6 + \beta_7 age7 + \beta_8 age8 + \beta_9 age9 + \beta_{10} age10 + \beta_{11} age11 + \beta_{13} dt + \beta_{14} prev + \beta_{15} actipass + \beta_{16} timemonth + childeffect(\tau)$. The results are summarized below:

	$Q = 3$			$Q = 5$			$Q = 20$		
Parameter	Est.	Std Err.	Pr> $\|t\|$	Est.	Std Err.	Pr> $\|t\|$	Est.	Std Err.	Pr> $\|t\|$
Intercept	-2.02	1.38	0.14	-4.45	1.43	0.00	-3.53	2.43	0.15
beta 0	-1.34	1.17	0.25	0.59	1.15	0.61	-0.24	2.09	0.91
beta 1	-1.43	1.06	0.18	0.26	1.04	0.81	-0.52	1.99	0.79
beta 2	-0.93	0.99	0.35	0.69	0.99	0.48	-0.01	1.81	1.00
beta 3	-0.68	0.92	0.46	0.77	0.92	0.41	0.14	1.65	0.93
beta 4	-1.10	0.86	0.20	0.30	0.87	0.73	-0.31	1.51	0.84
beta 5	-3.06	1.04	0.00	-1.91	1.05	0.07	-2.42	1.48	0.10
beta 6	-2.77	0.95	0.00	-1.55	0.98	0.12	-2.08	1.35	0.12
beta 7	-3.40	1.11	0.00	-2.55	1.14	0.03	-3.00	1.42	0.04
beta 8	-1.69	0.47	0.00	-0.91	0.50	0.07	-1.28	0.80	0.11
beta 9	-1.23	0.39	0.00	-0.66	0.43	0.12	-0.95	0.63	0.13
beta 10	-0.44	0.30	0.14	-0.06	0.32	0.84	-0.18	0.49	0.72
beta 11	-1.25	0.29	$< .0001$	-0.89	0.30	0.00	-1.00	0.41	0.02
beta 13	0.00	0.01	0.789	0.00	0.02	0.83	0.00	0.01	0.83
beta 14	58.32	7.02	$< .0001$	60.01	7.14	$< .0001$	60.28	7.17	$< .0001$
beta 15	-3.07	0.15	$< .0001$	-2.84	0.17	$< .0001$	-2.97	0.17	$< .0001$
beta 16	-0.12	0.10	0.24	0.02	0.11	0.82	-0.05	0.18	0.778
τ	0.00	0.00	$< .0001$	0.00	0.00	$< .0001$	0.00	0.00	$< .0001$

Table 9.49: Parameter estimate for 3,5 and 20 quadrature points, non adaptive Gaussian quadrature -LOCF

	$Q = 3$			$Q = 5$			$Q = 20$		
Parameter	Est.	Std Err.	Pr> $\|t\|$	Est.	Std Err.	Pr> $\|t\|$	Est.	Std Err.	Pr> $\|t\|$
Intercept	-3.45	1.58	0.03	-3.47	1.61	0.03	-3.53	2.43	0.15
beta 0	-0.27	1.32	0.84	-0.30	1.34	0.83	-0.25	2.09	0.91
beta 1	-0.53	1.20	0.66	-0.57	1.23	0.64	-0.52	1.99	0.79
beta 2	-0.03	1.12	0.98	-0.07	1.15	0.95	-0.01	1.81	1.00
beta 3	0.12	1.04	0.91	0.09	1.06	0.93	0.14	1.65	0.93
beta 4	-0.33	0.97	0.73	-0.36	0.99	0.71	-0.31	1.51	0.84
beta 5	-2.42	1.11	0.03	-2.45	1.12	0.03	-2.42	1.48	0.10
beta 6	-2.11	1.02	0.04	-2.11	1.03	0.04	-2.18	1.35	0.12
beta 7	-2.99	1.17	0.01	-3.01	1.18	0.01	-3.00	1.42	0.04
beta 8	-1.28	0.53	0.02	-1.30	0.54	0.02	-1.28	0.80	0.11
beta 9	-0.95	0.42	0.03	-0.95	0.43	0.03	-0.95	0.63	0.13
beta 10	-0.17	0.31	0.59	-0.18	0.32	0.57	-0.18	0.49	0.72
beta 11	-1.00	0.31	0.00	-1.01	0.31	0.00	-1.00	0.41	0.02
beta 13	0.00	0.01	0.83	0.00	0.01	0.85	0.00	0.01	0.83
beta 14	59.77	7.05	$< .0001$	59.89	7.07	$< .0001$	60.31	7.27	$< .0001$
beta 15	-2.97	0.15	$< .0001$	-2.97	0.15	$< .0001$	-2.97	0.17	$< .0001$
beta 16	-0.05	0.12	0.66	-0.06	0.12	0.64	-0.05	0.18	0.778
tau	0.00	0.00	$< .0001$	0.00	0.00	$< .0001$	0.00	0.00	$< .0001$

Table 9.50: Parameter estimate for 3,5 and 20 quadrature points, adaptive Gaussian quadrature -LOCF

The results of the adaptive and nonadaptive Gaussian quadrature from fitting a GLMM using PROC NLMIXED are not too different from each other but are a bit different from the LOCF random effects and random

intercept model using PROC GLIMMIX.

9.9.2 Comparative Results: LOCF,Original data and MICE

We will firstly compare the force of infection and the rate of recovery for the LOCF, MICE and the Available data, overall and the for individual months.

	LOCF		MICE		Available case	
Parameter	Estimate	S.E	Estimate	S.E	Estimate	S.E
λ	0.00012	0.0001	0.000919	7.334E-9	0.001169	0.00011
ν	0.0313	0.0029	0.49634	0.00559	0.45495	0.067

Table 9.51: Comparative estimates of the force of infection and rate of recovery using maximum likelihood estimation

	LOCF		MICE		Available case	
Parameter	Estimate	S.E	Estimate	S.E	Estimate	S.E
λ	0.00072	0.00006	0.0012	0.000083	0.0021	0.000153
ν	0.0356	0.0031	0.6298	0.0884	0.5030	0.0587

Table 9.52: Comparative estimates of the force of infection and rate of recovery using GLM estimation

Table 9.51 shows that the maximum likelihood estimates of the force of infection are similar with the MICE and available case approach. The MICE

estimate also exhibits the smallest standard error. The LOCF estimates are different and are smaller when compared to the MICE and available case approach. This highlights the potential bias that LOCF introduces when used to handle the dropout in the longitudinal setting. Table 9.52 highlights the potential bias that the LOCF method can introduce as the GLM estimates are again much smaller when compared to the consistent estimates of the MICE and available data. Thus in such a case the LOCF is not a recommended method to estimate such important parameters as the force of infection and the recovery rate for a disease process. Much efficient methods to handle the missingness present in the data are thus more important in the estimation process. This will ensure more valid inference about the process than just to use LOCF because of its simplicity. In Tables 9.53 and 9.54 monthly process parameters are estimated under LOCF, Available data (original data) and MICE with the GLM approach. Confidence intervals were calculated using exponentiation and the delta method respectively.

	LOCF	LOCF-Exponentiation		Original data	Exponentiation		MICE	Exponentiation	
Month	Lambda	95% Confidence Interval		Lambda	95% Confidence Interval		Lambda	95% Confidence Interval	
$\hat{\lambda}_2$	0.0043	0.0026	0.0071	0.0053	0.0032	0.0086	0.0045	0.0027	0.0075
$\hat{\lambda}_3$	0.0026	0.0018	0.0038	0.0070	0.0053	0.0092	0.0033	0.0023	0.0048
$\hat{\lambda}_4$	0.0028	0.0019	0.0041	0.0051	0.0038	0.0070	0.0033	0.0023	0.0047
$\hat{\lambda}_5$	0.0008	0.0004	0.0016	0.0024	0.0016	0.0037	0.0011	0.0006	0.0019
$\hat{\lambda}_6$	0.0004	0.0001	0.0013	0.0019	0.0011	0.0033	0.0005	0.0002	0.0013
$\hat{\lambda}_7$	0.0002	0.0001	0.0009	0.0010	0.0005	0.0020	0.0007	0.0003	0.0015
$\hat{\lambda}_8$	0.0000	0.0000	0.0000	0.0001	0.0000	0.0008	0.0002	0.0001	0.0008
$\hat{\lambda}_9$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\lambda}_{10}$	0.0001	0.0000	0.0008	0.0001	0.0000	0.0009	0.0002	0.0001	0.0008
$\hat{\lambda}_{11}$	0.0012	0.0008	0.0018	0.0022	0.0014	0.0033	0.0016	0.0011	0.0024
$\hat{\lambda}_{12}$	0.0012	0.0008	0.0017	0.0022	0.0015	0.0032	0.0019	0.0014	0.0025
$\hat{\lambda}_{13}$	0.0004	0.0002	0.0009	0.0014	0.0007	0.0029	0.0008	0.0006	0.0011

Table 9.53: Comparative monthly estimates of the force of infection and confidence Intervals-Exponentiation using LOCF, Available data and MICE

	LOCF	LOCF-Delta method		Original data	Delta method		MICE	Delta method	
Month	Lambda	95% Confidence Interval		Lambda	95% Confidence Interval		Lambda	95% Confidence Interval	
$\hat{\lambda}_2$	0.0043	0.0021	0.0065	0.0053	0.0027	0.0079	0.0045	0.0022	0.0068
$\hat{\lambda}_3$	0.0026	0.0016	0.0036	0.0070	0.0051	0.0089	0.0033	0.0021	0.0045
$\hat{\lambda}_4$	0.0028	0.0017	0.0038	0.0051	0.0036	0.0067	0.0033	0.0021	0.0045
$\hat{\lambda}_5$	0.0008	0.0002	0.0014	0.0024	0.0014	0.0034	0.0011	0.0004	0.0017
$\hat{\lambda}_6$	0.0004	-0.0001	0.0009	0.0019	0.0009	0.0029	0.0005	0.0000	0.0010
$\hat{\lambda}_7$	0.0002	-0.0001	0.0005	0.0010	0.0003	0.0017	0.0007	0.0001	0.0012
$\hat{\lambda}_8$	0.0000	0.0000	0.0000	0.0001	-0.0001	0.0003	0.0002	-0.0001	0.0005
$\hat{\lambda}_9$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\lambda}_{10}$	0.0001	-0.0001	0.0003	0.0001	-0.0001	0.0004	0.0002	-0.0001	0.0005
$\hat{\lambda}_{11}$	0.0012	0.0007	0.0017	0.0022	0.0013	0.0031	0.0016	0.0010	0.0022
$\hat{\lambda}_{12}$	0.0012	0.0007	0.0016	0.0022	0.0013	0.0030	0.0019	0.0014	0.0024
$\hat{\lambda}_{13}$	0.0004	0.0001	0.0007	0.0014	0.0004	0.0024	0.0008	0.0005	0.0011

Table 9.54: Comparative monthly estimates of the force of infection and confidence Intervals-Delta Method using LOCF,Available data and MICE

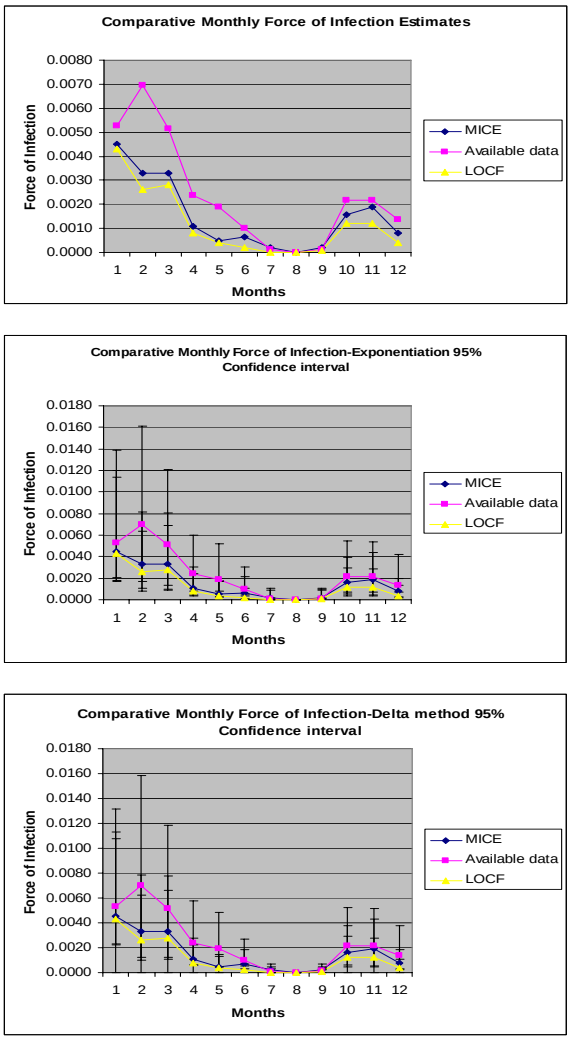

Figure 9.2: MICE-The force of infection in months together with 95% confidence intervals using the exponentiated and delta methods.

Tables 9.53 and 9.54 both reveal differences in the monthly estimates of the force of infection when comparing the LOCF estimates to the MICE and available data estimates. The MICE monthly estimates are somewhat smaller when compared to the available data estimates and also exhibits smaller confidence limits for both the exponentiation and delta methods. The delta method confidence limits for all three approaches are much smaller when compared to the exponentiation confidence limits. Once again the time effects of the disease process are revealed by monthly estimates of the force of infection. Figure 9.2 shows the MICE estimates to be closer to those of the available data estimates. As can be expected, LOCF is a poor performer.

9.10 Comparative GEEs

			LOCF (GEE)		GEE		MICE-GEE		WGEE	
Correlation Type	Source	DF	Chi-Square	Pr> Chi-Sq	Chi-Square	Pr> Chi-Sq	Chi-Square	Pr> Chi-Sq	Chi-Square	Pr> Chi-Sq
Exchangeable	age	12	25.66	0.0120	30.39	0.0024	33.93	0.0007	29.30	0.0036
	dt	1	0.26	0.6128	0.01	0.9379	1.65	0.1986	0.54	0.4625
	prev	1	18.28	<.0001	23.32	<.0001	22.42	< .0001	14.91	0.0001
	timemonth	1	0.23	0.6284	0.30	0.5860	4.63	0.0314	1.26	0.2608
	actipass	1	36.90	<.0001	61.86	<.0001	51.27	<.0001	58.93	<.0001
Independent	age	12	22.76	0.029	30.39	0.0024	33.98	0.0007	27.82	0.0059
	dt	1	0.15	0.7012	0.01	0.9378	1.69	0.1938	0.54	0.4606
	prev	1	16.01	<.0001	23.32	<.0001	22.59	< .0001	14.48	0.0001
	timemonth	1	0.20	0.6514	0.29	0.5882	4.48	0.0343	1.30	0.2551
	actipass	1	29.67	<.0001	61.81	<.0001	51.08	<.0001	58.07	<.0001
AR(1)	age	12	26.46	0.0092	30.54	0.023	33.85	0.0007	27.83	0.0059
	dt	1	0.69	0.4055	0.02	0.8974	1.70	0.1925	0.56	0.4528
	prev	1	16.45	<.0001	22.94	<.0001	22.79	< .0001	14.31	0.0002
	timemonth	1	0.00	0.9552	0.27	0.6008	4.49	0.0341	1.32	0.2508
	actipass	1	54.02	<.0001	62.00	<.0001	50.50	<.0001	58.45	<.0001

Table 9.55: Score statistics for Type III LOCF GEE,GEE,MICE-GEE,WGEE

	Available data-GEE						LOCF-GEE						MICE-GEE						WGEE					
	Exchangeable		Independent		AR1		Exchangeable		Independent		AR1		Exchangeable		Independent		AR1		Exchangeable		Independent		AR1	
Parameter	Estimate	S.E	Estimate	S.E	Estimate	S.E	Estimate	S.E	Estimate	S.E	Estimate	S.E	Estimate	S.E	Estimate	S.E	Estimate	S.E	Estimate	S.E	Estimate	S.E	Estimate	S.E
Intercept	-5.03	1.45	-5.04	1.45	-5.03	1.46	-5.02	1.24	-4.92	1.08	-5.28	1.25	-3.64	0.80	-3.67	0.81	-3.67	0.81	-8.53	1.73	-8.31	1.63	-8.31	1.64
age 0	-0.93	1.22	-0.92	1.22	-0.93	1.23	-0.74	1.01	-0.89	0.90	-0.94	1.12	-1.25	0.85	-1.21	0.86	-1.22	0.86	1.37	1.46	1.34	1.38	1.30	1.39
age 1	-0.65	1.07	-0.65	1.07	-0.60	1.08	-0.68	0.90	-0.71	0.77	-0.31	0.93	-1.01	0.64	-0.99	0.65	-0.99	0.66	-2.46	1.35	-2.52	1.28	-2.49	1.28
age 2	-0.28	1.03	-0.28	1.04	-0.24	1.04	-0.34	0.86	-0.62	0.75	-0.13	0.89	-0.81	0.61	-0.79	0.62	-0.79	0.62	0.13	1.28	-0.01	1.21	0.00	1.21
age 3	-0.07	0.97	-0.07	0.98	-0.03	0.98	-0.25	0.80	-0.41	0.70	0.14	0.83	-0.64	0.57	-0.62	0.58	-0.62	0.58	0.17	1.18	0.03	1.11	0.05	1.12
age 4	-0.67	0.95	-0.67	0.95	-0.65	0.96	-0.62	0.77	-0.64	0.69	-0.54	0.81	-0.90	0.56	-0.89	0.56	-0.88	0.57	0.76	1.09	0.78	1.03	0.79	1.03
age 5	-2.61	1.30	-2.60	1.30	-2.54	1.29	-2.53	0.99	-2.45	0.92	-1.97	0.99	-2.37	0.83	-2.37	0.83	-2.36	0.83	-1.67	1.49	-1.50	1.34	-1.48	1.34
age 6	-1.60	1.01	-1.60	1.01	-1.57	1.01	-2.34	0.88	-2.64	0.86	-1.45	0.78	-0.96	0.58	-0.95	0.58	-0.94	0.58	-0.94	1.11	-0.89	1.03	-0.88	1.03
age 7	-2.25	1.15	-2.25	1.15	-2.26	1.17	-3.00	1.16	-2.86	1.06	-2.43	1.01	-1.71	0.76	-1.70	0.77	-1.72	0.78	-1.90	1.29	-1.74	1.17	-1.74	1.17
age 8	-1.00	0.59	-1.00	0.59	-0.96	0.60	-1.51	0.41	-1.57	0.38	-0.88	0.41	-0.66	0.44	-0.65	0.44	-0.65	0.44	-0.75	0.64	-0.72	0.62	-0.71	0.62
age 9	-0.74	0.51	-0.74	0.51	-0.74	0.52	-1.29	0.31	-1.57	0.30	-1.20	0.39	-0.35	0.36	-0.35	0.36	-0.35	0.37	-0.49	0.55	-0.46	0.53	-0.45	0.53
age 10	-0.32	0.45	-0.32	0.45	-0.30	0.46	-0.27	0.23	-0.21	0.21	-0.23	0.29	0.42	0.26	0.42	0.26	0.42	0.26	-0.09	0.49	-0.07	0.47	-0.06	0.47
age 11	-0.57	0.46	-0.57	0.46	-0.54	0.46	-1.27	0.26	-1.28	0.24	-0.79	0.30	-0.10	0.24	-0.10	0.25	-0.10	0.25	-0.49	0.49	-0.44	0.47	-0.43	0.47
age 12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
dt	0.00	0.01	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
prev	44.61	8.11	44.59	8.11	43.89	8.12	48.96	6.27	48.26	6.01	42.70	6.35	30.94	5.27	31.02	5.28	31.07	5.29	40.41	8.88	38.70	8.46	38.46	8.47
timemonth	-0.05	0.10	-0.05	0.10	-0.04	0.11	-0.05	0.09	-0.04	0.08	-0.01	0.09	-0.15	0.06	-0.15	0.06	-0.15	0.06	0.15	0.13	0.15	0.12	0.15	0.12
actipass 0	2.23	0.18	2.23	0.18	2.20	0.18	2.71	0.15	2.63	0.13	2.19	0.14	1.42	0.14	1.42	0.14	1.42	0.14	3.41	0.26	3.33	0.25	3.31	0.25
actipass 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 9.56: Comparative model based estimates-Available data, LOCF MICE and WGEE

	Available data-GEE						LOCF-GEE						MICE-GEE						WGEE					
	Exchangeable		Independent		AR1		Exchangeable		Independent		AR1		Exchangeable		Independent		AR1		Exchangeable		Independent		AR1	
Parameter	Estimate	S.E	Estimate	S.E	Estimate	S.E	Estimate	S.E	Estimate	S.E	Estimate	S.E	Estimate	S.E	Estimate	S.E	Estimate	S.E	Estimate	S.E	Estimate	S.E	Estimate	S.E
Intercept	-5.03	1.17	-5.04	1.17	-5.03	1.16	-5.02	1.47	-4.92	1.36	-5.28	1.42	-3.64	0.89	-3.67	0.91	-3.67	0.91	-8.53	1.97	-8.31	1.84	-8.31	1.84
age 0	-0.93	1.23	-0.92	1.23	-0.93	1.23	-0.74	1.43	-0.89	1.40	-0.94	1.48	-1.25	1.04	-1.21	1.05	-1.22	1.05	1.37	1.80	1.34	1.70	1.30	1.71
age 1	-0.65	0.91	-0.65	0.91	-0.60	0.90	-0.68	1.22	-0.71	1.12	-0.31	1.20	-1.01	0.70	-0.99	0.71	-0.99	0.72	-2.46	1.68	-2.52	1.62	-2.49	1.61
age 2	-0.28	0.86	-0.28	0.86	-0.24	0.86	-0.34	1.16	-0.62	1.07	-0.13	1.16	-0.81	0.68	-0.79	0.69	-0.79	0.69	0.13	1.42	-0.01	1.35	0.00	1.35
age 3	-0.07	0.80	-0.07	0.80	-0.03	0.80	-0.25	1.11	-0.41	1.04	0.14	1.07	-0.64	0.63	-0.62	0.64	-0.62	0.64	0.17	1.29	0.03	1.23	0.05	1.23
age 4	-0.67	0.75	-0.67	0.75	-0.65	0.75	-0.62	1.01	-0.64	0.95	-0.54	1.04	-0.90	0.62	-0.89	0.63	-0.88	0.64	0.76	1.05	0.78	0.99	0.79	0.99
age 5	-2.61	1.19	-2.60	1.19	-2.54	1.17	-2.53	1.45	-2.45	1.32	-1.97	1.11	-2.37	0.82	-2.37	0.83	-2.36	0.83	-1.67	1.64	-1.50	1.34	-1.48	1.33
age 6	-1.60	0.87	-1.60	0.87	-1.57	0.86	-2.34	1.02	-2.64	1.03	-1.45	0.93	-0.96	0.65	-0.95	0.65	-0.94	0.65	-0.94	1.05	-0.89	0.94	-0.88	0.94
age 7	-2.25	1.04	-2.25	1.04	-2.26	1.03	-3.00	1.33	-2.86	1.14	-2.43	0.96	-1.71	0.79	-1.70	0.79	-1.72	0.80	-1.90	1.32	-1.74	1.08	-1.74	1.08
age 8	-1.00	0.61	-1.00	0.61	-0.96	0.61	-1.51	0.83	-1.57	0.75	-0.88	0.73	-0.66	0.45	-0.65	0.45	-0.65	0.45	-0.75	0.73	-0.72	0.67	-0.71	0.68
age 9	-0.74	0.56	-0.74	0.56	-0.74	0.55	-1.29	0.72	-1.57	0.75	-1.20	0.71	-0.35	0.42	-0.35	0.42	-0.35	0.42	-0.49	0.67	-0.46	0.63	-0.45	0.62
age 10	-0.32	0.47	-0.32	0.47	-0.30	0.47	-0.27	0.57	-0.21	0.53	-0.23	0.60	0.42	0.28	0.42	0.28	0.42	0.28	-0.09	0.56	-0.07	0.53	-0.06	0.53
age 11	-0.57	0.44	-0.57	0.44	-0.54	0.45	-1.27	0.60	-1.28	0.55	-0.79	0.59	-0.10	0.25	-0.10	0.25	-0.10	0.25	-0.49	0.51	-0.44	0.47	-0.43	0.47
age 12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
dt	0.00	0.01	0.00	0.01	0.00	0.010	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
prev	44.61	6.55	44.59	6.55	43.90	6.53	48.96	8.20	48.26	8.62	42.70	7.19	30.94	5.32	31.02	5.32	31.07	5.33	40.41	8.36	38.70	7.78	38.46	7.78
timemonth	-0.05	0.09	-0.05	0.09	-0.04	0.08	-0.05	0.10	-0.04	0.09	-0.01	0.10	-0.15	0.06	-0.15	0.07	-0.15	0.07	0.15	0.14	0.15	0.13	0.15	0.13
actipass 0	2.24	0.18	2.234	0.18	2.21	0.18	2.71	0.24	2.63	0.25	2.19	0.18	1.42	0.15	1.42	0.15	1.42	0.15	3.41	0.63	3.33	0.58	3.31	0.58
actipass 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 9.57: Comparative empirical based estimates-Available data, LOCF,MICE, WGEE

Table 9.55 reveals that the 'age', 'prev' and 'actipass' variables are all significant at the 5% level for all three correlation structures in all cases of LOCF, GEE(available data), MICE and WGEE. However, the 'timemonth' variable is found to be significant at the 5% level in all three correlation structures but only under the MICE approach. The actual values of the scores statistics are the lowest in the LOCF case and highest in the MICE case. The WGEE and GEE(available data) are consistent in their values. Table 9.56and 9.57 summarize the parameter estimates for all three correlation structure in all the approaches. The MICE parameter estimates are the smallest with the smallest standard errors while the WGEE parameter and standard error estimates are the largest. The LOCF and and available data estimates are consistent with each other for both the model and empirical based estimates. The correlation under the exchangeable structure was -0.00035(available case), 0.0303(LOCF), -0.0018(MICE) and 0.0024(WGEE). We see how the potential danger of LOCF is revealed here as well with the artificial inflating of correlation between successive time points. The -2log likelihoods were calculated as -645.19(available case), -1288.22(LOCF), -1078.00(MICE) and -850.13(WGEE). The lowest log likelihood was in the available case followed by WGEE and MICE whilst the highest log likelihood was in the LOCF approach, once again highlighting the danger of this approach.

9.11 Comparative Random Effects Model

	Available Data			LOCF			MICE		
Effect	Estimate	Std.Err	Pr> $\|t\|$	Estimate	Std.Err	Pr> $\|t\|$	Estimate	Std.Err	Pr> $\|t\|$
Intercept	-5.04	1.48	0.00	-5.62	1.45	0.00	-3.67	0.86	< .0001
age 0	-0.92	1.24	0.46	-0.18	1.22	0.88	-1.21	0.91	0.18
age 1	-0.65	1.09	0.55	-0.24	1.10	0.83	-0.99	0.69	0.15
age 2	-0.28	1.06	0.79	0.01	1.04	0.99	-0.79	0.65	0.23
age 3	-0.07	0.99	0.94	0.22	0.96	0.82	-0.62	0.61	0.31
age 4	-0.67	0.97	0.49	-0.19	0.91	0.84	-0.89	0.60	0.14
age 5	-2.60	1.32	0.05	-2.07	1.09	0.06	-2.37	0.88	0.01
age 6	-1.60	1.03	0.12	-1.99	0.95	0.04	-0.95	0.62	0.12
age 7	-2.25	1.18	0.06	-2.73	1.25	0.03	-1.70	0.81	0.04
age 8	-1.00	0.61	0.10	-1.22	0.49	0.01	-0.65	0.47	0.17
age 9	-0.74	0.52	0.16	-0.87	0.37	0.02	-0.35	0.39	0.37
age 10	-0.32	0.46	0.49	-0.35	0.27	0.20	0.42	0.27	0.12
age 11	-0.57	0.47	0.23	-1.20	0.27	< .0001	-0.10	0.26	0.71
age 12	0.00	.	.	0.00	.	.	0.00	.	.
dt	0.00	0.01	0.92	0.00	0.01	0.40	0.01	0.01	0.09
prev	44.59	8.26	< .0001	52.10	6.74	< .0001	31.02	5.59	< .0001
actipass 0	2.23	0.18	< .0001	2.61	0.14	< .0001	1.42	0.15	< .0001
actipass 1	0.00	.	.	0.00	.	.	0.00	.	.
timemonth	-0.05	0.11	0.67	0.01	0.11	0.92	-0.15	0.06	0.02

Table 9.58: Solution for the fixed effects of the random effects model -LOCF ,Original Data and MICE

Table 9.58 show that the smallest parameter estimates and standard errors are given by the MICE approach however similar comparisons of vari-

	Available data		LOCF		MICE	
Effect	F-Value	P-value	F-Value	P-value	F-Value	P-value
Age	1.62	0.0777	4.12	$< .0001$	1.77	0.0464
Dt	0.01	0.9205	0.72	0.3973	2.79	0.0946
Prev	29.17	$< .0001$	59.69	< 0.0001	30.78	< 0.0001
Actipass	153.61	$< .0001$	360.34	< 0.0001	92.77	< 0.0001
Timemonth	0.18	0.6701	0.01	0.9209	5.50	0.0191

Table 9.59: Type III Effects for random effects model-LOCF,Original data and MICE

ables such as 'age 5 vs age 12', 'age 7 vs age 12', 'dt', 'prev', 'timemonth' and 'actipass 0 vs actipass 1' are all significant at the 5% level. Table 9.59 also reveal the type III statistics for the same variables to be similar. The MICE and available type III statistics are dissimilar in magnitude when compared to the LOCF statistics. The -2 log likelihood values were 73005.67 (available data), 114632.7 (LOCF) and 112877.1 (MICE). As can be expected the LOCF log likelihood value is the highest.

9.12 Comparative Random Intercept Model

	Available Data			LOCF			MICE		
Effect	Estimate	Std.Err	Pr> $\|t\|$	Estimate	Std.Err	Pr> $\|t\|$	Estimate	Std.Err	Pr> $\|t\|$
Intercept	-5.04	1.48	0.00	-5.62	1.45	0.00	-3.67	0.86	< .0001
age 0	-0.92	1.24	0.46	-0.18	1.22	0.88	-1.21	0.91	0.18
age 1	-0.65	1.09	0.55	-0.24	1.10	0.83	-0.99	0.69	0.15
age 2	-0.28	1.06	0.79	0.01	1.04	0.99	-0.79	0.65	0.23
age 3	-0.07	0.99	0.94	0.22	0.96	0.82	-0.62	0.61	0.31
age 4	-0.67	0.97	0.49	-0.19	0.91	0.84	-0.89	0.60	0.14
age 5	-2.60	1.32	0.05	-2.07	1.09	0.06	-2.37	0.88	0.01
age 6	-1.60	1.03	0.12	-1.99	0.95	0.04	-0.95	0.62	0.12
age 7	-2.25	1.18	0.06	-2.73	1.25	0.03	-1.70	0.81	0.04
age 8	-1.00	0.61	0.10	-1.22	0.49	0.01	-0.65	0.47	0.17
age 9	-0.74	0.52	0.16	-0.87	0.37	0.02	-0.35	0.39	0.37
age 10	-0.32	0.46	0.49	-0.35	0.27	0.20	0.42	0.27	0.12
age 11	-0.57	0.47	0.23	-1.20	0.27	< .0001	-0.10	0.26	0.71
age 12	0.00	.	.	0.00	.	.	0.00	.	.
dt	0.00	0.01	0.92	0.00	0.01	0.40	0.01	0.01	0.09
prev	44.59	8.26	< .0001	52.10	6.74	< .0001	31.02	5.59	< .0001
actipass 0	2.23	0.18	< .0001	2.61	0.14	< .0001	1.42	0.15	< .0001
actipass 1	0.00	.	.	0.00	.	.	0.00	.	.
timemonth	-0.05	0.11	0.67	0.01	0.11	0.92	-0.15	0.06	0.02

Table 9.60: Random Intercept Model-Solution for the fixed effects -LOCF ,Original Data and MICE

The results for and conclusions for the random intercept model are the same as those for the random effects model. The -2 log likelihood values

	Available data		LOCF		MICE	
Effect	F-Value	P-value	F-Value	P-value	F-Value	P-value
Age	1.62	0.0777	4.12	$< .0001$	1.77	0.0464
Dt	0.01	0.9205	0.72	0.3973	2.79	0.0946
Prev	29.17	$< .0001$	59.69	< 0.0001	30.78	< 0.0001
Actipass	153.61	$< .0001$	360.34	< 0.0001	92.77	< 0.0001
Timemonth	0.18	0.6701	0.01	0.9209	5.50	0.0191

Table 9.61: Random Intercept Model-Type III Effects for Optimal Model-LOCF,Original data and MICE

were 73005.67 (available data), 113657.8 (LOCF) and 112597.2 (MICE). As can be expected the LOCF log likelihood value is the highest. The results of this chapter show that with respect to handling the dropout, LOCF has the advantage of simplicity but clearly can lead to over or under estimation of parameters and must be used with extreme caution or not be used at all.

9.13 Conclusion

This chapter focused on estimating the intermittent missingness, the 85 missing values in the response variable and the dropout in the data set. The methods used to handle the estimation of the intermittent missingness were LOCF and the EM algorithm. The estimated data sets were then analyzed using GEE, GLMM and the direct likelihood and GLM estimation of the force of infection and rate of recovery. Comparatively the results did not differ by much, the reason being that this missingness is only about 1% of the available data. If the proportion of missingness were higher, there would have been distinct comparative differences in the analyses that would have

originated from the EM algorithm and the LOCF estimation techniques. The dropout was handled using LOCF, direct likelihood, WGEE and multiple imputation (MICE). The estimated data sets were then analyzed using GEE, GLMM and the direct likelihood and GLM estimation of the force of infection and rate of recovery. Comparative studies were then done along with the available data. LOCF showed serious weaknesses as a technique to handle the dropout due to the artificial inflation of correlation between successive time points. Thus LOCF should in general not be used to handle the dropout in any longitudinal data set. MICE and WGEE proved extremely useful in handling the dropout. WGEE took a longer to converge in SAS and was computationally more involved. Kenward and Carpenter (2007) state that multiple imputation has at least three distinct advantages. Firstly, it can be applied very generally, to very large data sets with complex patterns of missingness among covariates, and only uses complete data quantities with very simple rules of combination. This is attractive for observational studies. Secondly, multiple imputation provides a relatively flexible and convenient route for investigating sensitivity to postulated NMAR mechanisms. Thirdly, the imputation model may include variables not in the substantive model, which can lead to additional efficiency, or most importantly in the clinical trial setting, allow post randomization covariates in the imputation model if they are predictive of dropout. Kenward and Carpenter (2007) also state that a more rigorous theoretical basis is needed for the chained equation approach and this is an avenue for further research. GEE requires the MCAR assumption but the direct likelihood methods, multiple imputation and WGEE require the fairly general assumption of the MAR mechanism. It must be emphasized that LOCF should no longer be the preferred mode of analysis in order to handle the dropout in any longitudinal setting.

Chapter 10

Survival Analysis Approach: Multiple Events per Subject

10.1 Introduction

The origin of survival analysis can be traced back to early work on mortality tables, which was followed up and expanded by statistical research for engineering applications. Survival analysis is the term that is used to study time-to-event data that correspond to the time from a well-defined time origin until the occurrence of some particular event or *end point* (Collett, 1994). Therefore we are often interested in the waiting time until an event can occurs, or more succinctly put, we are interested in the *end point* of a process. The end point could include events such as death, infection by a disease pathogen, first marriage and many more. The time origin on the other hand could be for example, the diagnosis of a disease or the recruitment of an individual into a study (such as the children that were recruited into the RSV study or the recruitment of an HIV/AIDS patient to start receiving ARVs). The time that an individual spends in a study is the subject time or as in

the case of a patient entering a study, his or her time spent in the study is also called the patient time. The methodology of survival analysis has got a vast number of applications that range from the survival time of animals in a experimental study, the time taken by an individual to complete a task in a psychological experiment, the storage times of seeds being kept in a seed bank and the identification of risk factors for a particular disease. Le (1997, pp 14-16) gives several examples in medical research that can be analyzed as survival or time-to-event data.

Survival data has got special characteristics that are associated with it, for example, most survival data are rarely normally distributed. A histogram of the survival times data will reveal positive skewness. Thus methods to deal with the problem of non-normality are necessary. One approach is that of data transformation typically by taking the logarithm of the data. Another defining characteristic of survival data is that, it is frequently censored or incomplete. Collett (1994, p. 2) states that the survival time of an individual is said to be censored when the end point of interest has not been observed for that individual. This could happen due to various reasons:

i) The study could have reached an end, and an individual may not yet have experienced the event of interest (for example, a child in the Kilifi RSV study could just remain uninfected throughout the study, or in another context, the patient may still be alive at the end of the study when the event is death)

ii) The subject or patient could have been lost due to follow up meaning that the only information that the investigator may have about such an individual is only the last time the individual remained to the study and visited the clinic/hospital or the last time the individual was known to be alive.

The type of censoring described above is known as right censored data, and these patient/subject times are less than the actual survival time that could have been observed. Likewise left censoring occurs when an individual experienced the event before the study commenced. For example, say the interest in a particular study is time to remission from a disease, and the first visit is 3 months after the individual is known to be disease free. If there is evidence of a tumour at this first visit as in the case of cancer, then the event occurred before the study started and this individual is left censored.

Another type of censoring is interval censoring, when an event occurs within an interval of time. Here individuals are known to have experienced a failure within an interval of time. A good example of interval censoring is when the event is defined as the first Tuberculosis (TB) positive test. If an individual is Tuberculosis negative (TB-)on the fifth visit, say, and is Tuberculosis positive(TB+) on the sixth visit; then the exact date of seroconversion is only known to be between the two visits, and the individual is interval censored.

More often than not, investigators are usually interested in right censoring, which is formalized as follows. If an individual enters the study at time t_0 and dies at time $t_0 + t$, then t is the uncensored survival time. However if the individual is last known to be alive at time $t_0 + c$, then c is known as the right censored survival time. (The individual may have been lost due to follow up, or may not have experienced the event by the end of the study). Alternatively, if we define t to be the time to event, and c the time at which censoring occurs, then an individual is right censored if $t > c$ and uncensored if $t \leq c$.

10.2 The Survivor Function and the Hazard Function

Suppose we have a group of patients/subjects with survival times $t_1, t_2, \ldots, t_N$ some of which may be censored. These values can be regarded as realizations of the continuous variable T, which has a probability density function, $f(t)$, and cumulative distribution function $F(t)$, where $F(t)$ is given by,

$$\begin{aligned} F(t) &= P(T \leq t) \\ &= \int_0^t f(u)du \end{aligned}$$

This represents the probability that the survival time is less than some value t (Collett, 1994). The survival function, which represents the probability that an individual will survive beyond time t, is given by

$$S(t) = P(T > t) = 1 - F(t). \tag{10.1}$$

Because survival distributions are usually skewed and there are many censored observations, the mean and the variance are not used to summarize the distribution of T, but rather medians and quantiles are used instead. These can be estimated from the survival function. For example, the median survival time is that value t_m of T satisfying $S(t_m) = 0.5$. In general the p^{th} percentile survival time is t_p such that $S(t_p) = 1 - p$. The hazard function is defined as the probability that an individual experiences an event (eg. death) at time t, given that he or she has survived up until that point. It thus measures the instantaneous death rate for an individual surviving to time t (Collett, 1994). Thus $h(t)$ is essentially a positive quantity and mathematically $h(t)$ is defined by

$$h(t) = \lim_{\delta t \to 0} \frac{P(t \leq T \leq t + \delta t | T \geq t)}{\delta t} \tag{10.2}$$

The above equation can be rewritten as

$$\begin{aligned} h(t) &= \lim_{\delta \to 0} \frac{P(t \leq T \leq t+\delta)}{\delta P(T \geq t)} \\ &= \lim_{\delta \to 0} \left[\frac{F(t+\delta) - F(t)}{\delta} \right] \frac{1}{P(T \geq t)} \\ &= \frac{f(t)}{S(t)} \end{aligned} \tag{10.3}$$

The following relationships can easily be derived from:

$$\begin{aligned} h(t) &= -\frac{d}{dt} \{\log S(t)\} \\ S(t) &= \exp\left\{ -\int_0^t h(u)du \right\} \\ H(t) &= \int_0^t h(u)du. \end{aligned} \tag{10.4}$$

The function $H(t)$ is known as the integrated or cumulative hazard function and is sometimes denoted as $\Lambda(t)$. The mathematical relationship between $H(t)$ and $S(t)$ is given as

$$H(t) = -\log S(t). \tag{10.5}$$

Both the survivor function and the hazard function can be estimated from the given survival data. The methods of estimation can be broadly grouped into parametric and nonparametric methods (Le, 1997). The Cox proportional hazards model, an example of a semi-parametric method, has also been fundamental in survival analysis and will be discussed before we come to the multi-state models. The purpose of this chapter is to apply multi-state survival models to model disease outcome data such as the RSV data currently being studied. In the context of infectious diseases the hazard function corresponds to the force of infection (FOI). The force of infection is the probability that an individual is infected by a disease at time t given the individual was

disease free up to time t. The force of infection can be modelled as time or age dependent. Other simpler models assume that the parameter is time independent.

10.3 Types of Survival Distribution

There are several distributions that are useful and are widely used in the survival context. These include the, Exponential, Weibull and Log-Logistic distributions. We describe the Weibull and Exponential distributions because they are the most relevant to our current study.

10.3.1 Exponential Distribution

The exponential distribution is characterized by the following probability density function (p.d.f)

$$f(t;\lambda) = \lambda e^{-\lambda t}, \quad t > 0$$

The cumulative distribution function (c.d.f) is given by

$$F(t) = 1 - \lambda e^{-\lambda t}$$

and the survivor function is

$$\begin{aligned} S(t) &= 1 - F(t) \\ &= e^{-\lambda t} \end{aligned}$$

The hazard function is

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} \\ &= \lambda \\ H(t) &= \lambda t. \end{aligned}$$

Thus the force of infection resembles the hazard function in its definition. This therefore makes methods in survival analysis very relevant in modelling the force of infection. The exponential distribution for survival time implies that the hazard function is a constant which means that the risk of death or rate of occurrence of events is independent of time. This is practically a very unrealistic assumption, because intuitively the risk of death or an event may increase or decrease as an individual ages or survives, for example. Another unique property associated with the exponential distribution is the lack of memory property. Suppose that the random variable T is the survival time of interest, and is exponentially distributed with parameter λ. Consider the probability that an individual survives for a time greater than t_1, given that he survived up until $t_0(t > t_0)$. Then

$$\begin{aligned} P(T > t_1 | T \geq t_0) &= \frac{P(T > t_1 \text{ and } T \geq t_0)}{P(T \geq t_0)} \\ &= \frac{P(T > t_1)}{P(T \geq t_0)} \\ &= \frac{S(t_1)}{S(t_0)} \\ &= \frac{e^{-\lambda t_1}}{e^{-\lambda t_0}} \\ &= e^{-\lambda(t_1 - t_0)} \end{aligned}$$

This can be interpreted as, given survival to time t_0, the excess life beyond t_0 still has the exponential distribution with parameter λ. This property is

popularly known as the memoryless property of the exponential distribution. Furthermore, this result also explains why the exponential distribution is not a realistic distribution for time-to-event data. The advantages of this distributions are its simplicity in the analyses associated with it.

10.3.2 Weibull Distribution

The two parameter p.d.f. for the Weibull function is given by

$$f(t;\gamma,\delta) = \delta\gamma t^{\gamma-1}e^{-\delta t^{\gamma}}, \quad t > 0$$

The parameter γ is the shape parameter, while δ is the scale parameter. Note that when $\gamma = 1$ the Weibull distribution reduces to the Exponential distribution with parameter δ. The c.d.f. of the Weibull distribution is given by

$$F(t) = 1 - e^{-\delta t^{\gamma}}, \quad t > 0.$$

The survivor function is therefore given by

$$S(t) = 1 - F(t) = e^{-\delta t^{\gamma}} \tag{10.6}$$

and the hazard function is thus

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= \delta\gamma t^{\gamma-1} \\ H(t) &= \delta t^{\gamma}. \end{aligned}$$

The key property here is that the hazard is a function of survival time. Note that for $\gamma = 1$, the hazard function is constant reducing to the case of the exponentially distributed survival time. The hazard function takes a different shape depending on the shape parameter γ as depicted in the following table:

Value of γ	Shape of $h(t)$
$0 < \gamma < 1$	Exponential decay
$\gamma = 1$	Constant, $h(t) = \delta$
$\gamma = 2$	Straight line
$\gamma > 2$	Exponential Growth

Table 10.1: Hazard function for different values of γ

10.4 The Proportional Hazards Model or the Cox Regression Model

The Cox regression model is very popular in the analysis of multi-state models and time-to-event data. Collett (1994, p.54) state that there are essentially two main reasons for modelling survival data. One reason is to find the right combination of the potential explanatory variables that affect the form of the hazard function and the other reason is to actually obtain an estimate for the hazard function through modelling it. Hosmer and Lemeshow (1999) state that, the distribution of the survival time variable T can be incorporated by modelling the density function of a parametric distribution for T, or to model the hazard function as a function of risk factors. It is advantageous to model the hazard function directly because of the relationship between the survivor function and the hazard function. The advantage is that, an estimate of the survivor function can be found as well as estimates of other quantities such as the median survival time. The disadvantage of this approach is the use of scatter plots is not viable to motivate regression models.

In 1972, Cox proposed a model that famously came to be known as the *Cox Regression Model* and later on as the *Proportional Hazards Model*. This model is in the class of a semi-parametric models since no particular form

of probability distribution is assumed for the survival times. This model is based soley on the assumption of proportional hazards.

10.4.1 The Theory of the Cox Regression Model

Suppose, for simplicity, that x is the single known covariate and β the corresponding unknown coefficient. Recall that the hazard function defined earlier, is the probability that an individual dies or experiences an event at time t, given that they have survived up until that point. In general, the hazard function can be specified as a function of time and the covariates in the form

$$h(t, x, \beta) = h_0(t)\phi(x, \beta) \tag{10.7}$$

where $\phi(x, \beta)$ is a function of the covariates only. The above equation has to be strictly positive. Here $h_0(t)$ characterizes how the hazard function changes as a function of survival time and is also known as the baseline hazard since $h(t, x, \beta) = h_0(t)$ when $x = 0$. The term $\phi(x, \beta)$ explains how the hazard function changes as function of specific covariates. The proportional hazard concept arises because the r.h.s of Eq (10.7) is expressed as a product of a function of time only and $\phi(x, \beta)$, a function of the covariates.

Consider then the simplest case where patients are randomly allocated to two groups for comparison, such as a treatment and control group in clinical trials. Let the hazard for the control group be $h(t, x_0, \beta)$ and for the treatment group be $h(t, x_1, \beta)$. Suppose that the ratio of the two hazard functions for the treatment and control groups is

$$\psi = \frac{h(t, x_1, \beta)}{h(t, x_0, \beta)}. \tag{10.8}$$

Using Eq(10.8), it follows that ψ simplifies to

$$\begin{aligned}\psi &= \frac{h_0(t)\phi(x_1,\beta)}{h_0(t)phi(x_0,\beta)} \\ &= \frac{\phi(x_1,\beta)}{\phi(x_0,\beta)}.\end{aligned}$$

Hosmer and Lemeshow (1999) state that if the hazard ratio, ψ is easily interpreted then the actual form of the baseline hazard is of little importance. The ratio ψ measures the risk of death at time t for an individual on treatment relative to a person on the control. If $\psi < 1$ the hazard for an individual on treatment is said to be smaller than for an individual on the control, and the treatment is thus an improvement. If $\psi > 1$ then the hazard is smaller for the person in the control group than for a person in the treatment group and it cannot be concluded that the treatment is effective in increasing survival time, (Collett, 1994, p. 55).

From the proportional hazards assumption the following important relationship between the survivor function of the treatment group and that of the control group emerges,

$$\begin{aligned}S_1(t) &= e^{-H(t,x_1,\beta)} \\ &= \exp\left\{-\int_0^t h(u,x_1,\beta)du\right\} \\ &= \exp\left\{-\int_0^t \psi h_0(u)r(x_0,\beta)du\right\} \\ &= \exp\left\{-\psi\int_0^t h_0(u)r(x_0,\beta)du\right\} = [S_0(t)]^{\psi}.\end{aligned}$$

Cox (1972) proposed a model that uses $\phi(x,\beta) = e^{x\beta}$ in order to ensure the hazard function $h(t)$ is non-negative. The proportional hazards model assumption then becomes

$$h(t,x,\beta) = h_0(t)e^{x\beta}$$

and the hazard ratio for comparing the two groups is

$$\psi = e^{\beta(x_1 - x_0)}.$$

In the case where the single covariate x is dichotomous i.e. assuming either a 0 or 1, the hazard ration can be seen as a type of "relative risk" (Hosmer and Lemeshow, 1999). Hence if $\beta = \ln(2)$, then those with $x = 1$ are dying at twice the rate of those with $x = 0$. Under Cox's regression model the survivor function can be rewritten as,

$$S_1(t) = [S_0(t)]^{\exp(x\beta)}.$$

10.4.2 The General Proportional Hazards Model

As a generalization of the above concept (Collett, 1994, pp 55-56), suppose now that the hazard of death at a particular time depends on the values $x_1, x_2, \ldots, x_p$ of p explanatory variables $X_1, X_2, \ldots, X_p$ where $x_1, x_2, \ldots, x_p$ are assumed to be recorded at the outset of the study for a given individual. This constitutes what is known as baseline data. Let $x_{i1}, x_{i2}, \ldots, x_{ip}$ denote the measured values of the p covariates for individual i. Thus the set of variable values can be denoted by the vector $\boldsymbol{x_i}$. Let $h_0(t)$ be the hazard function for an individual whose set of covariates, $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})$, are equal to zero, that is $h_0(t)$ gives the baseline hazard function. The hazard function for the i^{th} individual can then be written as

$$\begin{aligned} h_i(t|\boldsymbol{x_i}) &= \lim_{\delta t \to 0} \frac{P(t \leq T \leq t + \delta t)}{\delta P(T \geq t; \boldsymbol{x_i})} \\ &= \psi(\boldsymbol{x_i}) h_0(t) \end{aligned}$$

where $\psi(\boldsymbol{x}_i)$ is a function of the explanatory variables for the i^{th} person. Now $\psi(\boldsymbol{x}_i)$ can be interpreted as the hazard at time t for an individual whose

vector of explanatory variables is $\boldsymbol{x}_i$, relative to the baseline hazard, that is,

$$\psi(\boldsymbol{x_i}) = \frac{h(t|\boldsymbol{x_i})}{h_0(t)}.$$

Extending Cox's definition for the proportional hazards model, the following expression holds:

$$\begin{aligned}
\psi(\boldsymbol{x_i}) &= \exp(\eta_i) \\
&= \exp(x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p) \\
&= \exp(\sum_{j=1}^{p} x_{ij}\beta_j) \\
&= \exp(\boldsymbol{x_i}^T\beta).
\end{aligned}$$

where $\boldsymbol{\beta^T} = (\beta_1, \ldots, \beta_p)$ is a vector of regression coefficients and η_i is the linear component of the model, where η_i is also known as the risk score or prognostic index for the i^{th} individual (Collett , 1994). The general proportional hazards model then becomes

$$h_i(t|x_i) = \exp(x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p)h_0(t). \tag{10.9}$$

The equation can be linearized by dividing through by the baseline hazard and taking logs on both sides of the above equation to give

$$\log\left[\frac{h_i(t)}{h_0(t)}\right] = x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p. \tag{10.10}$$

The structure of the model can be seen as a logistic regression model without the constant term β_0.

10.4.3 Fitting the Proportional Hazards Model

Collett (1994, p.61) gives the following summary as to how to fit the proportional hazards model. Firstly the unknown coefficients $\beta_1, \ldots, \beta_p$ need to

be estimated. Let $\boldsymbol{\beta^T} = (\beta_1, \ldots, \beta_p)$ denote a vector of unknown coefficients corresponding to the p known variates. In some cases it might be necessary that the baseline hazard $h_0(t)$ be estimated. The most common method of estimating the regression coefficients is the Maximum Likelihood (ML) approach. Thus the first step is to construct the likelihood of the observed survival data.

Suppose that there are n individuals each with the triplet $(t_i, \boldsymbol{x_i}, c_i)$ where t_i is the observed survival time, $\boldsymbol{x}_i$ is the covariate vector and c_i is the censoring indicator variable for individual i where $c_i = 1$ if a survival time is uncensored and $c_i = 0$ if the time is censored. Suppose that there are r ordered distinct death time such that $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$. This implies that there are $n - r$ right censored survival times. We will not consider the treatment of ties for now. Suppose that the set of individuals that are at risk at time $t_{(j)}$ are denoted by $R(t_{(j)})$ which is also known as the risk set. The risk set consists of all the subjects with survival or censored times greater than or equal to the specified time. Now, consider the result relating the three functions, the hazard, survivor and probability density functions (Equation (10.3)). It follows we can write

$$f(t, \boldsymbol{x}, \beta) = h(t, \boldsymbol{x}, \beta) \times S(t, \boldsymbol{x}, \beta). \tag{10.11}$$

It follows also that the likelihood function for the regression models is

$$L(\beta) = \prod_{i=1}^{n} [f(t_i, \boldsymbol{x}_i, \beta)]^{c_i} \times [S(t_i, \boldsymbol{x}_i, \beta)]^{1-c_i}. \tag{10.12}$$

Substituting Eq.(10.11) into Eq.(10.12) above, yields

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} [h(t_i, \boldsymbol{x}_i, \beta) \times S(t_i, \boldsymbol{x}_i, \beta)]^{c_i} \times [S(t_i, \boldsymbol{x}_i, \beta)]^{1-c_i} \\ &= \prod_{i=1}^{n} [h(t_i, \boldsymbol{x}_i, \beta)]^{c_i} \times S(t_i, \boldsymbol{x}_i, \beta). \end{aligned}$$

Note the components of the likelihood expression above can be interpreted as follows. An individual has to survive up to time t_i with probability $S(t_i, \boldsymbol{x}_i, \beta)$. At this time either the individual experiences an event ($c_i = 1$) or does not ($c_i = 0$). Thus the product of these two terms gives the required joint probability and the product of n terms gives the full likelihood. Substituting $h(t_i, \boldsymbol{x}_i, \beta) = h_0(t_i)e^{x_i^T\beta}$ and $S(t_i, \boldsymbol{x}_i, \beta) = [S_0(t_i)]^{\exp(x_i^T\beta)}$ into the above equation gives

$$L(\beta) = \prod_{i=1}^{n} [h_o(t_i)e^{x_i^T\beta}]^{c_i} \times [S_0(t_i)]^{\exp(x_i^T\beta)} \tag{10.13}$$

Hence

$$\ln[L(\beta)] = \ell(\beta) = \sum_{i=1}^{n} c_i \ln[h_0(t_i)] + c_i\boldsymbol{x}_i^T\beta + e^{x_i^T\beta} \ln[S_0(t_i)]. \tag{10.14}$$

The ML estimation method requires that the above Eq.(10.13) be maximized with respect to the unknown parameters, β and a parametric model for the baseline hazard be specified. However, the proportional hazards model is adopted in order to avoid explicitly defining the baseline hazard function.
Cox (1972) constructed a partial likelihood (depending only on the parameters of interest) that can be maximized in order to obtain estimates for the unknown parameters. He showed that the resulting parameter estimates from the partial likelihood function would have the same distributional properties as the ML estimators. Suppose that $x_{(j)}$ is the vector of covariates for a subject with observed ordered survival time $t_{(j)}$. Then the partial likelihood which can be derived using the counting process approach as given in Fleming and Harrington (1991) and Collett (1994, p.62) is

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{x}_{(j)}^T\beta)}{\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{x}_{(l)}^T\beta)}. \tag{10.15}$$

The above derivation is based on a conditional probability argument and the partial likelihood is given for participants who experience an event. However,

if c_i is the censoring variable that takes on the value 0 when an observation is censored and 1 when the observation is not censored, then the above likelihood can be written as

$$L(\beta) = \prod_{j=1}^{n} \left[\frac{\exp(\boldsymbol{x}_{(j)}^T \beta)}{\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{x}_{(l)}^T \beta)} \right]^{c_j}. \tag{10.16}$$

Note that in Eq. (10.15) the baseline hazard tern $h_0(t_{(j)})$ will automatically cancel out and therefore will not feature in the partial likelihood. When we take the log of the partial likelihood, we have

$$\ln L(\beta) = \ell(\beta) = \sum_{i=1}^{n} c_i \left[\boldsymbol{x}_i^T \beta - \ln \sum_{l \in R(t_{(j)})} \exp(\boldsymbol{x}_l^T \beta) \right].$$

Differentiating this log likelihood with respect to, the unknown coefficients β, we have p equations given by

$$\begin{aligned} \frac{\partial \ell(\beta)}{\partial \beta_k} &= \sum_{j=1}^{r} \left[x_{(jk)} - \frac{\sum_{l \in R(t_{(j)})} x_{lk} \exp(\boldsymbol{x}_l^T \beta)}{\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{x}_l^T \beta)} \right] \\ &= \sum_{j=1}^{r} \{ x_{(jk)} - \overline{x}_{wjk} \} \end{aligned} \tag{10.17}$$

where

$$\overline{x}_{wjk} = \sum_{l \in R(t_{(j)})} w_{jl} x_{lk} \tag{10.18}$$

and

$$w_{jl} = \frac{\exp(\boldsymbol{x}_l^T \beta)}{\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{x}_l^T \beta)}. \tag{10.19}$$

Expression 10.19 can be viewed as a weight for an individual contributing to the covariate vector $\boldsymbol{x}_l^T$. Here $x_{(jk)}$ is the value of the covariate x_k for a subject with observed ordered survival time $t_{(j)}$. The estimator is obtained by setting the derivatives equal to 0 and solving for $\beta_k,\ k = 1, \ldots, p$. In general, an iterative technique needs to be employed in order to solve for the unknown

parameters because of the intractability of the estimating equation (10.17). The variance of the estimator β is obtained by taking the inverse of the negative of the second derivative of the, partial likelihood at the value of the estimator. Namely,

$$\text{Var}(\hat{\beta}) = I(\hat{\beta})^{-1}$$

where

$$I(\beta) = -\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T}$$

The diagonal elements of the information matrix are given by

$$\frac{\partial^2 \ell(\beta)}{\partial \beta_k^2} = -\sum_{j=1}^{r} \sum_{l \in R(t_{(j)})} w_{jl}(x_{ll} - \overline{x}_{wjk})^2$$

and the off diagonal elements are

$$\frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_h} = -\sum_{j=1}^{r} \sum_{l \in R(t_{(j)})} w_{jl}(x_{ll} - \overline{x}_{wjk})(x_{lh} - \overline{x}_{wjh})$$

where $\overline{x}_{wjk}$ and w_{jl} are defined in previous equations (10.18) and (10.19).

10.4.4 Tests of Significance

In almost all statistical procedures the significance of the estimated coefficients needs to be assessed, and it is usual practice to form the confidence intervals for these estimates. The three tests that are commonly used to do so in the context of fitting the proportional hazards model are the partial likelihood ratio test, the Wald test and the score test. These three tests are briefly reviewed below.

Partial Likelihood Ratio Test

Hosmer and Lemeshow (1999) propose that the partial likelihood ratio test is the easiest to compute, and the best of the three above mentioned tests

for assessing the significance of the fitted model. The test statistic is given by

$$G = 2\ell_p(\hat{\beta}) - \ell_p(0)$$

where $\ell_p(\hat{\beta})$ is the log partial likelihood evaluated at $\hat{\beta}$ and $\ell_p(0) = -\sum_{i=1}^{m} n_i$ where n_i is the number of subjects in the risk set at the observed survival time $t_{(j)}$ and m is the number of distinct times. This statistic tests whether all coefficients are equal to zero versus that at least one of the coefficients is non-zero. Under the null hypothesis, $H_0 : \beta = 0$, G follows the χ^2 distribution with degrees of freedom equal to the number of parameters estimated in the model.

Score Test

Hosmer and Lemeshow (1999) further state that as opposed to the partial likelihood ratio test, the Wald and Score tests require matrix calculations and formulations. Let the vector of the first order partial derivatives of the partial log-likelihood be denoted by $\boldsymbol{u(\beta)}$. Under the null hypothesis that all the coefficients are equal to zero, the vector of scores $\boldsymbol{u(0)}$=$\boldsymbol{u(\beta)\rfloor_{\beta=0}}$. The score statistic then becomes

$$\boldsymbol{u^T(0)[I(0)]^{-1}u(0)}$$

which is, under H_0 approximately χ^2 distributed with degrees of freedom equal to the number of parameters in the model. In the case of one covariate, the score test is given by

$$z^* = \left.\frac{d\ell_p/d\beta}{\sqrt{I(\beta)}}\right\rfloor_{\beta=0}$$

and under H_0, $z^* \sim N(0,1)$ or $(z^*)^2 \sim \chi^2(1)$.

Wald Test

Hosmer and Lemeshow (1999) state that the Wald test statistic is obtained from the theory which states that under the null hypothesis the estimator of the vector of coefficients, $\hat{\beta}$, will be asymptotically normally distributed with mean vector $\boldsymbol{\mu} = \mathbf{0}$ and covariance matrix estimated by $\hat{\text{var}}(\hat{\beta}) = I(\hat{\beta})^{-1}$. Thus the multiple variable Wald statistic is given by

$$\hat{\beta}^T I(\hat{\beta})\hat{\beta}$$

which under the null hypothesis is χ^2 distributed with degrees of freedom equal to the number of parameters fitted in the model. In the case of a single covariate the square root of the test statistic reduces to

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

where $\text{se}(\hat{\beta}) = \sqrt{\text{var}(\hat{\beta})}$. Thus z is standard normally distributed under H_0 or $z^2 \sim \chi^2(1)$. The confidence interval of $\hat{\beta}$ is based on the Wald statistic and can be found from the usual expression

$$\hat{\beta} \pm Z_{\alpha/2} \times \text{se}(\hat{\beta})$$

In practice the three statistics, $\sqrt{(G)}, z, z^*$ should all be quite similar, resulting in the same conclusion. However, the partial likelihood ratio test is the preferred choice. Hosmer and Lemeshow (1999) comment that an advantage of using the score statistic is that the statistic can be computed without evaluating the maximum partial likelihood estimates of the parameters. It is useful as a test to use in model building applications in which evaluation of the estimator is computationally intensive.

10.5 Fitting the Proportional Hazards Model with Tied Survival Times

In practice, tied survival times do occur, while the partial likelihood function is based on the assumption that there are no tied survival times. However the partial likelihood function can be adapted to accommodate tied survival times. The exact expression for the modified partial likelihood function is derived by Kalbfleish and Prentice (1980) and approximations are due to Breslow (1974) and Efron (1977).

For simplicity we will assume only one covariate. The basis for construction of the exact partial likelihood is to assume that the d ties at a particular survival time are due to the lack of precision in measuring the survival time, such as recording the time in days ignoring the fractional days. Hosmer and Lemeshow (1999) state that tied survival times could have actually been observed in any one of the $d!$ possible arrangements of their values. The exact partial likelihood is obtained by modifying the denominator of

$$L_p(\beta) = \prod_{i=1}^{m} \frac{e^{x_{(i)}\beta}}{\sum_{j \in R(t_{(i)})} e^{x_j \beta}} \tag{10.20}$$

to include each of these arrangements. Approximations derived by Breslow (1974) and Efron (1977) are designed to provide expressions that are easier to compute than the exact partial likelihood, and yet still account for the fact that ties are present among the observed values of survival time. The Breslow (1974) approximation uses the partial likelihood given by

$$L_{p1}(\beta) = \prod_{i=1}^{m} \frac{e^{x_{(i)+}\beta}}{[\sum_{j \in R(t_{(i)})} e^{x_j \beta}]^{d_i}} \tag{10.21}$$

where d_i denotes the number of subjects with survival time $t_{(i)}$ and $x_{(i)+}$ is the sum of the covariate over the d_i subjects namely, $x_{(i)+} = \sum_{j \in D(t_{(i)})} x_j$,

where $D(t_{(j)})$ represents subjects with survival times equal to $t_{(i)}$. The upper limit m is the number of distinct survival times.

The Efron (1977) approximation yields a slightly better approximation to the exact partial likelihood than the Breslow (1974) approximation, with the partial likelihood given as

$$L_{p2}(\beta) = \prod_{i=1}^{m} \frac{e^{x_{(i)+}\beta}}{\prod_{k=1}^{d_i}[\sum_{j\in R(t_{(i)})} e^{x_j\beta} - \frac{k-1}{d_i}\sum_{j\in D(t_{(i)})} e^{x_j\beta}]} \tag{10.22}$$

The modified partial likelihood function for β in the presence of ties is obtained in the same manner as in the non-tied case data, with the exception that the derivative are taken with respect to the unknown parameters in the natural logarithm of the above equations (10.21) and (10.22) relating to the Breslow and Efron approximations. The variance of the estimated coefficient is obtained from the second partial derivative evaluated at the estimated value of the parameter. In reality, there will be little practical difference between the Breslow and Efron estimators. The Breslow estimator is however the most frequently used approximation due to its simplicity and less complexity. Much of the above derivation can be found in Collett (1994, pp 65-66), not repeated here.

10.6 Multiple Events per Subject

A major and still growing interest in the statistics community is the application of survival analysis to data sets with multiple events per subject. These multiple events can be classified as either cases where the multiple events are of the same type or when these multiple events are of a different type. A good example of multiple events of the same type are multiple respiratory infections such as the RSV disease infection experienced by children during

their early childhood and an example of multiple events of a different type is the recurrent opportunistic infections in AIDS patients. Therneau and Grambsch (2000) give examples of the multiple events that are of a different type that admit the use of survival and recurrence information in cancer trials or multiple sequelae in the management of chronic disease. The whole idea of multiple events also lend itself to the concept of multi-state models (Therneau and Grambsch, 2000).

There are however certain issues that are a point of concern with respect to how we model these multiple events per subject. A major issue arises as to how to extend the proportional hazards regression models to where there is now intra-subject correlation. Other complicating factors include multiple time scales, stratum by covariate interaction, discontinuous intervals of risk and the structure of risk sets. As a results of these concerns, several approaches have been looked at:

(i). Time to first event, ignoring the multiplicity. This makes the analysis easy for interpretation but the risk is that information will be wasted.

(ii). Including a random per subject effect which is the Random effect or *frailty* models. Multiple outcomes are assumed to be independent conditional on the per subject effect. Oakes (1992) and Keiding et al. (1997) examine the frailty models at great length. Frailty is often defined to be the deviations from the proportional hazards model that can be explained by unaccounted random heterogeneity.

(iii). A marginal models approach similar to that of Generalized Estimating Equations (GEE).

(iv). A more ambitious plan is to model the per subject's correlation structure directly within the Cox framework. This method is due to Pren-

tice and Cai (1992). They use a sample of industrial failure data. This method is very computer intensive (Therneau and Grambsch, 2000, pp. 169-170).

The marginal models in option (ii.) above will be looked at as a means to model the RSV data set in the current work by focusing on the time duration between events $d_{ij} = t_{ij} - t_{ij-1}$. Therneau and Grambsch (2000) point out the variants of this approach are due to Anderson and Gill (AG)(1982), Wei, Lin and Weissfield (WLW) (1989) and Prentice, Williams and Peterson (PWP) (1981). This method also affords great flexibility in the formation of strata and risk sets, manipulation of the time scale and has a well developed estimator of variance. The analysis of these models is based on three steps, namely,

1. Decide on a model which include issues such as strata, time dependent covariates etc. and structure the data set accordingly

2. Fit the data using an ordinary Cox model, ignoring the possible correlation

3. Replace the standard variance estimate with one which is corrected for possible correlations (Therneau and Grambsch, 2000, p. 170)

We will now look at these three steps as outlined in Therneau and Grambsch (2000).

10.6.1 Selecting a model

The computation for the marginal model is not a difficult issue, but the creation and preparation of an appropriate data set is the difficult part, as well as the choice between alternative models. In SAS, the 'phlev' macro can be

used to set up the data. Setting up the model includes the choice of strata and membership within strata, time scales within strata, constructed time dependent covariates, stratum by covariate interactions and data organization. Some details of these issues will now be expounded:

- Stratification, if used, is based on external variables that may include, enrolling institution or disease subtype. These generally correspond to predictors for which we desire flexible adjustment, but not an estimate of the covariate effect. Each subject is in exactly one stratum.

- The time scale is almost invariably time since entry to the study although alternative time scales based on a counting process is possible.

- Time dependent covariates usually reflect directly measured data such as repeated lab tests. Strata by covariate interactions are infrequent.

- The counting process form may be used for a time dependent covariate, but normally the data set will consist of one observation per subject.

These four areas can be extended. The first aspect is to distinguish whether or not the data set for multiple events have a distinct ordering or not. The unordered events have simpler issues attached to them and they are discussed below.

Unordered events

Therneau and Grambsch (2000) consider 5 examples of correlated but unordered outcomes of multiple events. One of the examples involved a frailty model. The setup of unordered events is straightforward where each observation is entered into the data just as it would be, if correlation were not an issue. An appropriate software package such as SAS or S-Plus can be used to

fit the model to the data using a sandwich estimator to account for the correlation structure. The analysis is often stratified by the type of endpoint, for example if we to assume that the baseline hazard functions for time-to-death and time-to-progression may differ.

Ordered events

The three most common approaches are the independent increment (AG), marginal (WLW) or conditional (PWP) models. All three models are marginal regression models in the sense that $\hat{\beta}$ is determined from the fit that ignores correlation followed by a corrected variance $\tilde{D}'\tilde{D}$ (a sandwich estimate) but differ considerably in the creation of the risk sets. The implementation of the three approaches are now considered in detail before applying them to the RSV data being analyzed in the current thesis.

The AG method by Andersen-Gill (AG)(1982)

This approach is very close to a Poisson type regression. The method is easiest to perceive and set up but makes the strongest assumptions. Laird and Oliver (1981) used Poisson regression as an approximation to this method by using an ordinary single event Cox model. The AG method uses a counting process style of data input, where each subject is represented as a series of observations in rows of data, with time intervals (entry time, first event], (first event, second event], ..., (m^{th} event, last follow up]. A subject with zero events would have a single observation, one with one event would have one or two observations (depending on whether there was additional followup experience after the first event), and so on. Depending on the time scale the first observation may or may not begin at zero.

When the time scale is "time since entry" the intensity process or hazard

function is identical to the Cox model for survival data,

$$h(t, x_i(t)) = Y_i(t)\lambda_0 \exp(X_i(t)\beta) \tag{10.23}$$

However the difference arises in the definition of $Y_i(t)$. In the case of survival data, when an event occurs the individual ceases to be at risk and Y_i goes to zero, but for the AG model for recurrent events $Y_i(t)$ remains as one as events occur. If strata are used then they are based on the same considerations as for an ordinary single event model. No extra strata or strata by covariate interaction terms are induced by multiple events. The model is ideally suited to the situation of mutual independence of the observations within a subject. This assumption is equivalent to each individual counting process possessing independent increments where non-overlapping time intervals are independent, given the covariates. Alternative time scales include the sojourn- or gap-time scales with intervals of $(0, t_1], (0, t_2 - t_1], \ldots$ corresponding to "time since last entry or last event". The gap times form a renewal process and the lack-of-memory property from the Exponential distribution mean the gap times themselves are a counting process with independent increments. However in general , it should be noted that a counting process cannot possess both independent increments and independent gap times. If the above assumptions are met then the three variance estimators $\mathcal{I}^{-1}$, $D'D$ and $\tilde{D}'\tilde{D}$ estimate the same quantity.

The WLW method by Wei, Lin and Weissfield (1989)

Wei, Lin and Weissfeld (1989) used this method to analyze bladder cancer data set with multiple events. What happens here, is that one treats the ordered outcome data set as though it were an unordered competing risks problem. If there are a maximum of four events in the data set, then there

will be four strata in the analysis. Every subject will have four observations, one in each stratum (unless there are missing covariates). The time scale is "time from study entry", and since all the time intervals start at zero, the model can be fit without recourse to the counting style of input. The intensity or hazard function for the j^{th} event for the i^{th} subject is

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

This model allows a separate underlying hazard for each event and for strata by covariate interactions denoted by β_j. The at-risk indicator for the j^{th} event, $Y_{ij}(t)$ is one until the occurrence of the j^{th} event, unless there is censoring. When either of those occur, it becomes zero.

The PWP method or conditional model by Prentice, Williams and Peterson (1981)

This model assumes that a subject cannot be at risk for the 2^{nd} event until event 1 occurs so if we generalize, a subject cannot be at risk for event k until the individual has already experience event $k-1$. The counting style of input is used for this model as in the AG model, but each event is assigned to a separate stratum. The time scale may be time since entry or gap time. The underlying hazard function may vary from event to event due to the use of time dependent strata. The intensity (hazard) in the time since entry scale is identical to the WLW intensity, except for the definition of the at risk indicator, $Y_{ij}(t)$, which is zero until the $(j-1)^{st}$ event and only then does it become one. The conditional approach is favoured by certain authors over the marginal method. Allison (1995, pp. 242-243) looked at the WLW method in computation detail. The three models discussed above have been described in detail in Therneau and Grambsch (2000, pp. 185-187).

10.6.2 Robust variance and computation

In the case of multiple events, the assumption in the usual Cox model, where to estimate the variance of $\hat{\beta}$ treats each of the observations as independent will not hold. Hence a robust variance estimate is required and the jackknife provides the most valid estimate of variance in the case of correlated data whenever the the observations left out at any step are independent of the observations left in. In the case of multiple observations per subject, the correlation is restricted to disjoint groups, so a grouped jackknife estimate that leaves out one subject at a time rather than one observation at a time, is appropriate. In SAS the 'proc phreg' procedure can be used to obtain this robust estimate of variance.

Due to the fact that the sandwich estimate $\tilde{D}'\tilde{D}$ will be much larger in magnitude than the model based variance $\mathcal{I}^{-1}$, in the presence of correlated data, the usual tests that were discussed earlier, namely, the Wald, score and likelihood ratio tests will be anti-conservative. Hence an important extension, in the robust Wald test, is to replace the usual variance with a sandwich estimate, $\hat{\beta}'[\tilde{D}'\tilde{D}]^{-1}\hat{\beta}$.

The extension of the score test which accounts for correlated data is to define the per subject leverage matrix as $\tilde{D}_{m\times p} = B_{m\times n}D_{n\times p}$ where B is a matrix of 0's and 1's that sums the appropriate rows such that

$$\tilde{D}'\tilde{D} = \mathcal{I}^{-1}U'B'BU\mathcal{I}^{-1}.$$

Then write the usual score test statistic

$$T = [1'U]\mathcal{I}^{-1}[U'1]$$

as

$$[1'U\mathcal{I}^{-1}]\mathcal{I}[\mathcal{I}^{-1}U'1]$$

The insertion of the inverse of the sandwich estimate of variance for the central term, based on the starting estimate of β, gives the following robust score test statistic:

$$T_r = [1^{'}U][U^{'}B^{'}BU]^{-1}[U^{'}1].$$

It can be shown that $\tilde{D}^{'}\tilde{D}$ is equal to the working independence estimate of variance for generalized estimating equation (GEE) models (Liang and Zeger, 1986). The paper by Liang and Zeger (1986) is with respect to longitudinal data, so summation of each individual's contributions when going from D to $\tilde{D}$ is over observations at multiple time points.

10.7 SAS Software PROCEDURES

First we state the usual survival routines in SAS. These are:

- `lifetest`: This procedure computes the Kaplan-Meier curves, and the log-rank (Mantel-Haenszel), Gehan-Wilcoxon and other tests
- `lifereg`: Accelerated time failure models
- `phreg`: Cox proportional hazard model

The above procedures or routines are briefly described below (Allison, 1995 and the SAS/STAT version 9 User Guide):

PROC LIFETEST

This procedure is primarily designed for univariate analysis of the timing of events. It produces life tables and graphs of survival curves (survivor functions). Using several methods, this procedure tests whether survival curves

are the same in two or more groups. This procedure also tests for associations between event times and time constant covariates, but it does not produce estimates of parameters. Although the Kaplan-Meier estimate is the default, the option 'method=KM' ensures that we get this estimator. The 'outsurv' and 'outtest' statements names an output data set to contain survival estimates, confidence limits and association of survival time with covariance limits. PROC LIFETEST also has a 'missing' statement which allows missing values to be at stratum level noting that the default is that missing values are not used in the analysis. The 'plot' statement produces a high resolution graph whilst the 'time' option requires the failure time variable to be input here and the 'strata' statement indicates which variables determine strata levels for the computations. The 'id' statement labels the observations of the product-limit survival function estimates. The 'test' statement tests for the effects of covariates. The 'survival' statement creates an output data set containing the results of the estimation of the survivor function. In this procedure, we can also test hypotheses about the shape of the hazard function. A comprehensive account of PROC LIFETEST can be found in the SAS/STAT User Guide , Volume 3 chapter 40, pp. 2148-2215.

PROC LIFEREG

PROC LIFEREG estimates regression models with censored, continuous-time data under several alternative distributional assumptions. The procedure allows for several types of censoring, but it does not allow for time-dependent covariates. PROC LIFEREG accommodates left censoring and interval censoring. If the shape of the survival distribution is known, then more efficient estimates with smaller standard errors are produced. PROC LIFEREG automatically creates sets of dummy (indicator) variables to repre-

sent categorical variables with multiple values. The 'model' statement allows the mathematical or equivalently statistical model to be specified. The 'dist' statement allows for different distributions that range from Weibull, Exponential, Gamma, Log-logistic and Log-normal distributions to be fitted. The 'class' statement allows for categorical variables. Covariance matrices can be requested using the 'covb' option in the model statement. The 'output' statement creates a new SAS data set containing the statistics calculated after fitting the model. PROC LIFEREG also has a 'weight' statement that can be used for weights in the analysis. Hypothesis tests using Wald (Chi-square) statistics can also be calculated within this procedure. Graphical options are also available using the 'probplot' or 'pplot' statements within the PROC LIFEREG procedure. Left censoring and interval censoring can be specified in the 'model' statement as model (lower, upper). The following options are present.

Uncensored: lower and upper are present and equal

Interval censored: lower and upper are present and are different

Right censored: lower is present and upper is missing

Left censored: lower is missing but upper is present

A comprehensive account of PROC LIFEREG can be found in the SAS/STAT User Guide, Volume 3, chapter 39, pp. 2090-2148.

PROC PHREG

This procedure uses Cox's partial likelihood method to estimate regression models with censored data. The model is somewhat less restrictive than the other models in PROC LIFEREG, and the estimation method allows for time dependent covariates. Furthermore, the semi-parametric regression analysis is done using the partial likelihood method. PROC PHREG handles both

continuous-time and discrete-time data. PROC PHREG only allows for right censoring and only gives nonparametric estimates of the survivor function which can be difficult to interpret. PROC PHREG also enables one to:

- include an offset variable in the model
- weight the observations in the data set
- test linear hypotheses about the regression parameters
- perform conditional logistic regression analysis for macthed case control studies
- create a SAS data set with residuals, survivor function estimates and regression diagnostics
- create a SAS data set containing survival distribution estimates and confidence interval for the survivor function at each time for a given realization of the explanatory variables

It has however the very powerful built-in 'stratification' option and also has by far the most powerful capability for incorporating the already mentioned time-dependent covariates. The 'model' statement allows for the specification of the model. The three alternative chi-square tests, namely, the likelihood ratio test, a score test and a Wald test are given. The 'assess' statement performs the graphical methods of Lin, Wei and Ying (1993) for checking the adequacy of the Cox regression model. The PHREG procedure also is able to handle ties using the 'ties' option within the 'model' statement. The 'ties' option allows for 'ties=Breslow (default), discrete, Efron or exact methods'. The exact option assumes that time is continuous and will assume that there is a true but unknown ordering for the tied event times whilst the discrete

option assumes that the events really occurred at exactly the same time. The Efron and Breslow methods use the approximate likelihoods of Efron (1977) and Breslow (1974). The 'strata' statement names the variables that determine stratification. The 'weight' and 'test' statements are also present on PROC PHREG. Time dependent covariates that change at irregular intervals can also be incorporated into PROC PHREG. The 'freq' statement identifies the variable containing the frequency of occurrence of each observation. The 'output' option can produce a data set containing statistics calculated for each observation, influence statistics, the linear predictor, standard error, survival distribution estimates and different diagnostic statistics for each individual with a wide range of residuals. The 'baseline' statement creates a SAS data set that contains the survivor function estimates at the event time for each stratum for every pattern of explanatory variables given in the covariates. The 'test' statement tests for linear hypotheses in conjunction with the 'class' statement. These tests of linear hypotheses tests that all the coefficients corresponding to a categorical covariate are equal to 0. The ODS Graphics part is still in the experimental stage but can be used for graphs. A comprehensive account of PROC PHREG can be found in the SAS/STAT User Guide , Volume 5 chapter 54, pp. 3215-3332.

10.8 Fitting the AG, WLW and PWP models to the RSV data

The fitting of the AG, WLW and PWP models can be done by using the 'PROC PHREG' statement in SAS. Details of these models and how they are specifically fitted, with illustrative examples can be found in SAS/STAT User Guide , Volume 5 chapter 54, pp. 3247-3257. The models are explained

very clearly as to how they can be fitted using the 'PROC PHREG' procedure in SAS and the manipulation of the data into intervals are also explained. For the WLW model the data was set out according to the following format. Here shown for a random individual with id= 4.

id	rsvpos	d_{ij}	age	prev
4	1	0	1	0.0135
4	0	10	1	0.0524
4	1	7	1	0.0111
4	0	20	1	0.0524
⋮	⋮	⋮	⋮	⋮

Table 10.2: Data description of the WLW model

However for the AG and PWP models the data was reorganized as follows:

id	rsvpos	start	end	age	prev
4	1	0	10	1	0.0135
4	0	10	17	1	0.0524
4	1	17	37	1	0.0111
⋮	⋮	⋮	⋮	⋮	⋮

Table 10.3: Data description of the AG and PWP models

The Cox model that was fitted was

$$h_{ij}(t) = h_0(t)\exp(\beta_1 age + \beta_2 prev + \beta_3 actipass + \beta_4 timemonth)$$

A summary of the results are given below: The Likelihood ratio, score and

Model	Variable	Parameter Estimate	Standard Error	Chi-Sq	Pr>Chi-Sq
WLW	Age	0.0576	0.081	0.0051	0.9430
	Prev	47.1980	4.997	89.2308	$< .0001$
	Actipass	-2.4346	0.175	193.238	$< .0001$
	Timemonth	-0.1615	0.076	4.468	0.0345
PWP	Age	0.0343	0.104	0.120	0.741
	Prev	37.050	6.568	31.825	$< .0001$
	Actipass	-2.219	0.180	151.605	$< .0001$
	Timemonth	-0.020	0.085	0.0522	0.8192
AG	Age	0.0343	0.113	0.093	0.760
	Prev	37.050	7.689	23.213	$< .0001$
	Actipass	-2.219	0.176	159.621	$< .0001$
	Timemonth	-0.020	0.099	0.039	0.8424

Table 10.4: Estmate of the WLW, PWP and AG models

Wald statistics were significant at the 5% level with small p-values. The PWP gap time model with common regression coefficients gave exactly the same results as the WLW model and will not be repeated. From the above table we see that in the WLW model the 'Prev', 'Actipass' and 'Timemonth' variables are significant in contributing towards the model, at the 5% level whilst the 'Age' variable is not. The PWP and AG models reveal only 'Prev' and 'Actipass' to be significant at the 5% level. Furthermore the AG and PWP gave similar results, except their standard errors of the estimates differing slightly. The objective of these marginal models is to assess the significance

of explanatory variables and they have done simply that. We will now look at extensions of these models by incorporating the 'child' as the random effect.

10.9 Frailty Models

Most real life situations are those where an individual experiences more than one event, that is, most events are repeatable, for example, births, marriages, job changes, promotions, tumor recurrences, seizures, urinary infections and hospitalizations, just to name a few. Allison (1995, p. 237) gives the following example to illustrate problems encountered with conventional methods when handling repeated events. There are basically two approaches to analyzing repeated events. Firstly you can do a separate analysis for each successive event. Suppose that you have reproductive histories for a sample of ever married women, and you want to construct and analyze a model for the birth intervals. You can start with an analysis for the interval between marriage and the first birth. First, all those woman that had a first birth, you then do a second analysis for the interval between the first birth and the second birth. You could continue in this fashion until the number of women gets too small to reliably estimate a model.

The second general approach to repeated events is by treating each interval as a distinct observation, pooling all the intervals together, and estimating a single model. This second method poses a problem in that, it does not account for dependence among multiple observations. Not taking this dependence into account, can lead to standard errors that are biased downward and test statistics that are biased upward leading to unnecessary significance of test statistics. Dependence among observations can be thought of as arising from unobserved heterogeneity. Hence models that deal with the problem of

dependence, must not only correct for the standard errors and test statistics but also for some or all of the bias in the coefficients caused by unobserved heterogeneity.

The basic idea in frailty modelling is to formulate a model that explicitly introduces a disturbance (random) term representing unobserved heterogeneity. These models are called *Frailty Models* (sometimes conditional or subject-specific) (Keiding et al., 1997). The random term is incorporated in the hazard function, under the assumption that frailty is independent of any censoring that might take place and this random term acts multiplicatively on the hazard function.

We will now use the introductory work on frailty by Vaupel, Manton and Stallard (1979): Let $h(t, \mathbf{x}, z)$ be the hazard function for an individual in population i with a vector of covariates $\mathbf{x}$, at some time t, and with a frailty of $z > 0$. The definition of frailty as defined by Vaupel et al. (1979) states that the ratio of the hazards for two different individuals in a population group i is equal to the ratio of their frailties. This is expressed as

$$\frac{h(t, \mathbf{x}, z)}{h(t, \mathbf{x}, z')} = \frac{z}{z'}$$

or

$$h_i(t, \mathbf{x}, z) = z h_i(t, \mathbf{x}, 1) \tag{10.24}$$

where an individual with a frailty of 1 might be viewed as a 'standard' individual. If an individual has a frailty of 2, then that person is twice as likely to die at any particular age, at any particular time, than a standard individual. On the other hand, a person with a frailty of 0.5 is only half as likely to die. In other words, if $z > 1$, then an individual is more 'frail' than a standard individual, if $z < 1$ the subject is less 'frail' than an average individual. Thus

frailty can be interpreted as relative risk.

The above definition of frailty assumes that each individual maintains a constant level of frailty, from birth to death. However, it does not imply that an individuals with the same frailty are identical. Also, it is more convenient to define frailty in terms of the hazard, rather than the age-specific probability of death q_x for the following reasons.

1. q_x is bounded above and thus the range of frailty would also be bounded above.

2. q_x is a nonlinear function of the size of the age interval used.

For the purposes of simplicity, let $h(t, \mathbf{x}, z)$ and $h_i(t, \mathbf{x}, z^{'})$ be denoted as $h(z)$ and h so that

$$h(z) = zh$$

The required relationships now follow

$$\begin{aligned} H(z) &= zH && (10.25) \\ S &= e^{-H} && (10.26) \\ \Rightarrow S(z) &= S^z && (10.27) \end{aligned}$$

where $S = S(t, \mathbf{x}, 1)$ for some vector $\mathbf{x}$ and time t.

10.9.1 The Distribution of Frailty

Let $\overline{h}_i(t, x)$ be the hazard for a cohort of individuals from a population group i at age x at time t. For simplicity assume that only one covariate is measured, in this case, age, measured by x. Note that $\overline{h}_i(t, x)$ is analogous to the average hazard in a group of individuals. Then

$$\overline{h}_i(t, x) = \int_0^\infty h_i(t, x, z) f_x(z) dz$$

where $f_x(z)$ is the p.d.f. of the frailty at age x among the surviving individuals in the cohort. Average frailty in the cohort $\overline{z}$, is defined by

$$\overline{z}_i(t,x) = \int_0^\infty z f_x(z) dz$$

Hence we have

$$\begin{aligned}
\overline{h}_i(t,x) &= \int_0^\infty h_i(t,x,z) f_x(z) dz \\
&= \int_0^\infty z h_i(t,x,1) f_x(z) dz \\
&= h_i(t,x,1) \int_0^\infty z f_x(z) dz \\
&= h_i(t,x,1) \overline{z}(t,x)
\end{aligned}$$

or alternatively $\overline{h} = h\overline{z}$.

The interpretation is that frail individuals with high values of z will tend to die first. This implies that $\overline{z}$ (which is the average frailty of the surviving cohort) will decline with age. The equation $\overline{h} = h\overline{z}$ also indicate that the hazard for individuals increases more swiftly than for the cohort in which the individuals belong (in other words "age" faster than cohorts). The relationship between the individual and cohort aging depends on the distribution of frailty among individuals.

Many literature papers and texts (Nguti, 2003; Zuma and Lurie, 2005; Bolstad and Manda, 2001) assume that the frailty is Gamma distributed with p.d.f.

$$f(z) = \frac{\lambda^k z^{k-1} e^{-\lambda z}}{\Gamma(k)}$$

where λ and k are the scale and shape parameters respectively therefore the mean and variance are given by

$$\overline{z} = \frac{k}{\lambda}$$

and

$$\sigma_z^2 = \frac{k}{\lambda^2}.$$

It is common to set the mean equal 1 so that $\lambda = k$ and $\sigma^2 = 1/k$.

There are a few reasons why the Gamma distribution is chosen for frailty. It is analytically tractable, readily computable, and is one of the few distributions that is able to model variables that are positive, and since frailty cannot be negative, it is thus suitable. It is also a flexible distribution that can take on a variety of shapes as k varies. When $k = 1$ then the p.d.f simplifies to the Exponential distribution. When k becomes large, it assumes a bell-shaped distribution similar to the Normal distribution. As k increases, and thus as variability in frailty decreases, mortality rates for standard individuals become more like the observed cohort rates. Thus there are two useful mathematical results noted that arise from the assumption that frailty at birth is Gamma distributed:

1. Frailty among those who have not yet died is Gamma distributed with the same value of shape parameter as at birth but now,

   $$\lambda(x) = \lambda + H(x)$$

   and the mean frailty is

   $$\overline{z}(x) = \overline{z}\frac{k}{k + \overline{z}H(x)}$$

   where $\overline{z}$ is the average frailty of the cohort at birth. When $k = 1$ and $\overline{z} = 1$, the mean frailty reduces to

   $$\overline{z}(x) = \frac{1}{1 + H(x)}.$$

   It is obvious from the above equation that as the cumulative hazard $H(x)$, increases then the average frailty of the remaining cohort decreases.

2. Frailty among those who die at any age x is also Gamma distributed, with the same scale parameter $\lambda(x)$ as among those surviving to age x, but with shape parameter $k+1$. This implies that the mean frailty of those who die at age x, denoted by $\overline{z}'(x)$, is greater than the mean frailty of the survivors

$$\overline{z}'(x) = \overline{z}\frac{k+1}{k}$$

Thus Vaupel et al. (1979) concluded that ignoring the frailty in a survival model may lead to biased estimates. Individual ageing rates, past and future progress in reducing mortality, and mortality differentials between populations may be underestimated, and current life expectancy and potential gains in life expectancy from averting specific causes of death may be overestimated.

10.9.2 Multivariate Semi-Parametric Frailty Models

In the case of univariate data , the hazard function is completely specified by the baseline hazard function and other covariates. There could be other factors that significantly affect the distribution of survival time, other than the covariates. We have already defined these factors to be the source of heterogeneity between subjects. Keiding et al. (1997) and Struthers and Kalbfleish (1986) demonstrate the effect of leaving out important covariates and the consequences of this on the results. The proportional hazards frailty model assumes that for a given frailty variable z_i and covariates $\boldsymbol{x}$, individual i has a hazard function given by

$$h_i(t|z_i, \boldsymbol{x}) = h_0(t)e^{\mathbf{x_i^T}\boldsymbol{\beta}+w_i} = z_i h_0(t)e^{\boldsymbol{x}_i^T\boldsymbol{\beta}} \tag{10.28}$$

where $z_i = e^{w_i}$ and w_i is the random effect for the i^{th} individual where $-\infty < w_i < \infty$. The participants who experience an event contribute the product

of their conditional hazards function and conditional survival function to the likelihood whereas those who do not experience an event, implying that they are right censored, contribute only their conditional survival function to the likelihood. The conditional survival function is given as

$$\begin{aligned} S(t|z_i, \boldsymbol{x}_i) &= \exp[-H(t|z_i, \boldsymbol{x}_i)] \\ &= \exp[-z_i \Lambda_0(t) \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})] \end{aligned} \tag{10.29}$$

where $\Lambda_0(t)$ is the integrated or cumulative baseline hazard. It may not always be the case that the event times among individuals are independent, as the failure times of certain individual may be correlated, for example individuals from the same family or community may be correlated violating the independence assumption, and these data are referred to as multivariate or repeated survival data. Time-to-event data that are not correlated are known as univariate survival data. Parallel or longitudinal data are the two main types or survival data. Parallel data consist of different clusters which have a number of items or individuals contained in them. Longitudinal data are a result of a stochastic process of events, namely an individual experiences a number of the same event over time which results in recurrent or short time series data. The cluster is now the individual, and within that individual events are observed. In both types of data, the events within a cluster are correlated. There is a school of thought that stipulates that there are unobserved risk factors that explain the dependence and these factors are generally assumed to be constant over time, and using the standard approach of modelling the survival data, say for example Cox regression would lead to biased estimates.

The Shared Frailty Model

A common approach applied to explain such unexplained risk factors is the shared frailty model. The shared frailty model assumes that all individuals within a cluster or group have the same frailty. To account for the heterogeneity among groups, a random term is included in the hazard function to account for the correlation of failure times within a cluster. The frailty model can be seen as linear mixed effects model with the frailty terms acting multiplicatively on the hazard function. The frailties are assumed to be independent between or across clusters, whilst the failure times within a cluster are dependent. However, conditional on frailties, the failure times are independent. In the univariate case, if we add frailty effects for each individual, we induce the heterogeneity among individuals after taking into account any measured covariates.

10.9.3 Frailty Model Formulation

Suppose that there are n individuals assigned to I groups where the i^{th} group has n_i individuals such that $\sum_{i=1}^{I} n_i = 1$. It should be said that in the current RSV data set, an individual is taking on the role of a cluster. Suppose that the number of events experienced by the i^{th} group is given by $D_i = \sum_{j=1}^{n_i} \delta_{ij}$, where δ_{ij} is the censoring indicator which takes on the value 1 when an event occurs and 0 when it does not. Then the hazard for the j^{th} individual from the i^{th} group is given by

$$h_{ij}(t) = h_0(t) \exp(\boldsymbol{x_{ij}}^T \boldsymbol{\beta} + w_i) \tag{10.30}$$

where $\boldsymbol{x}_{ij}$ is a vector of p covariates for individual j in group i, $h_0(t)$ is the baseline hazard and w_i is the random effect for the i^{th} group. The w_i's are i.i.d random sample from a density $f_W(.)$. We can then rewrite the model (10.30)

as:

$$\begin{aligned} h_{ij}(t) &= h_0(t)\exp(w_i)\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}) \\ &= z_i h_0(t)\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}) \end{aligned} \tag{10.31}$$

where $z_i = \exp(w_i)$ is the frailty term. The z_i's are independent with a common density $f_Z(.)$. The two commonly used densities for the frailties typically chosen are:

1. The zero-mean normal density for W which transforms to the log-normal density for Z, that is,

$$f_Z(z) = \frac{1}{z\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\log z)^2}{2\sigma^2}\right)$$

   with mean $e^{\sigma^2/2}$ and variance $e^{\sigma^2}(e^{\sigma^2}-1)$

2. The one parameter gamma density for Z with density given by

$$f_Z(z) = \frac{\alpha^\alpha z^{\alpha-1}e^{-\alpha z}}{\Gamma(\alpha)}$$

   which corresponds to a log-gamma density for W. The mean and variance for Z are given by

$$\begin{aligned} E[Z] &= 1 \\ \text{Var}[Z] &= \frac{1}{\alpha} \end{aligned}$$

Since z_i in the equation (10.31) can be thought of as a mixing term, its corresponding density $f_Z(.)$ is also referred to as a mixing distribution. When using the log-normal density for $f_Z(.)$, $\text{Var}[Z] = \sigma_z^2$ is used to describe the heterogeneity among the groups. For a gamma density describing frailties, $\text{Var}[Z] = \frac{1}{\alpha}$. In general heterogeneity is commonly described by a parameter

θ. This is σ_z^2 in the case of the log-normal distribution and $\frac{1}{\alpha}$ for the gamma density. If $\frac{1}{\alpha}$ is small then the gamma and log-normal distributions are similar (Kalbfleish and Prentice, 1980).

For creating correlated frailties, the log-normal distribution is preferred to the gamma distribution and is therefore extremely useful in modelling multivariate frailty models. Hougaard (2000) states that other distributions that are used for the frailties are the stable distribution and the power variance functions. The power variance function is a large family of distributions that comprise of the gamma and other positive stable distributions, making it a less restrictive function to use. However, the calculations for these functions are more difficult rendering it being less frequently used.

10.9.4 Estimation in the Frailty Model

The baseline hazard $h_0(t)$ can be specified explicitly or left unspecified. If it is specified explicitly, a parametric assumption for $h_0(t)$ means that parameters in the resulting model can be estimated using the maximum likelihood estimation (MLE). However if $h_0(t)$ is left unspecified then the unknown parameters in the shared frailty model have to estimated by other approaches and methods such as:

1. Expectation Maximization (EM) Algorithm (Klein, 1992)
2. Penalized Partial Likelihood (PPL) Approach (Therneau and Grambsch, 2000)
3. Markov Chain Monte Carlo (MCMC) methods (Vaida and Xu, 2000)
4. Monte Carlo EM (MCEM) approach (Ripatti et al. 2002)

5. Different methods using Laplace approximation (Ripatti and Palmgren, 2000; Cortinas Abrahantes and Burzykowski, 2004)

The choice of estimation methods depends on the choice of frailty distribution. When a gamma frailty is used, the EM algorithm is commonly used. However, when a log-normal frailty is used, the estimation procedures are based on numerical integration methods such as the Laplace approximation methods. We will look at the EM Algorithm and the Penalized Partial Likelihood approaches.

The Expectation Maximization (EM) Algorithm

The theory of the EM algorithm has already been covered in detail in Chapter 8 and will only be applied here. We will however look at how the EM Algorithm is applied to Gamma frailty models. For simplicity consider a univariate analysis with the following hazard function for individual i

$$h_i(t|z_i, \boldsymbol{x}_i) = z_i h_0(t) e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}.$$

Suppose that the baseline hazard is constant, that is $h_0(t) = h_0$. Also assume that the frailty is Gamma(α, α) distributed. The individuals who experienced an event contributes the product of their hazard and survival function, whereas censored individuals contribute only the survival function to the likelihood. The relationship between the hazard and survival function implies that the survival function is

$$\begin{aligned} S(t) &= e^{-\int_0^t h(u)} du \\ &= e^{-\int_0^t z_i h_0 e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}} du \\ &= e^{-z_i h_0 t e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}} \end{aligned}$$

The complete data likelihood contribution from individual i is given by

$$\begin{aligned} L_i(z_i, t_i; \alpha) &= f(z) \times [S_i(t_i)h_i(t_i)]^{\delta_i}[S_i(t_i)]^{1-\delta_i} \\ &= \frac{\alpha^\alpha}{\Gamma(\alpha)} z_i^{\alpha-1} e^{-\alpha z_i} \\ &\times [e^{-z_i h_0 t_i e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}} z_i h_0 e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}]^{\delta_i} [e^{-z_i h_0 t_i e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}]^{1-\delta_i} \end{aligned} \tag{10.32}$$

The associated complete-data log likelihood is then

$$\begin{aligned} \ell(z_i, t_i; \alpha) &= \alpha \ln \alpha - \ln \Gamma(\alpha) + (\alpha - 1) \ln z_i - \alpha z_i \\ &+ \delta_i [-z_i h_0 t_i e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} + \ln z_i + \ln h_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}] \\ &- (1 - \delta_i) z_i h_0 t_i e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}. \end{aligned}$$

Zuma and Lurie (2005) show that the observed data likelihood is attained by integrating out unobserved data from the likelihood in Eq. (10.32) as

$$\begin{aligned} L_{obs,i}(t_i; \alpha) &= \int_0^\infty \frac{\alpha^\alpha}{\Gamma(\alpha)} z_i^{\alpha-1} e^{-\alpha z_i} \qquad (10.33) \\ &\times [e^{-z_i h_0 t_i e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}} z_i h_0 e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}]^{\delta_i} [e^{-z_i h_0 t_i e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}]^{1-\delta_i} dz_i \\ &= \frac{\alpha^\alpha}{\Gamma(\alpha)} \int_0^\infty z_i^{\alpha-1} e^{-\alpha z_i} e^{-\delta_i z_i h_0 t_i e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}} \\ &\times z_i^{\delta_i} h_0^{\delta_i} e^{\delta_i \boldsymbol{x}_i^T \boldsymbol{\beta}} e^{-(1-\delta_i) z_i h_0 t_i e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}} dz_i \\ &= \frac{\alpha^\alpha}{\Gamma(\alpha)} h_0^{\delta_i} e^{\delta_i \boldsymbol{x}_i^T \boldsymbol{\beta}} \int_0^\infty z_i^{\alpha+\delta_i-1} e^{-z_i(\alpha + h_0 t_i e^{\boldsymbol{x}_i^T \boldsymbol{\beta}})} dz_i \qquad (10.34) \end{aligned}$$

The integral here appears to be the kernel of a Gamma$(\alpha + \delta_i, \frac{1}{\alpha + h_0 t_i e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}})$ function. The resulting integrand of Eq. (10.34) is then

$$L_{obs,i}(t_i; \alpha) = \frac{\alpha^\alpha (h_0 e^{\boldsymbol{x}_i^T \boldsymbol{\beta}})^{\delta_i} \Gamma(\alpha + \delta_i)}{\Gamma(\alpha)(\alpha + h_0 t_i e^{\boldsymbol{x}_i^T \boldsymbol{\beta}})^{\alpha+\delta_i}}.$$

To estimate the parameters, the log-likelihood of the observed data likelihood needs to be maximized

$$\begin{aligned} \ell_{obs,i}(\alpha; t_i) &= \alpha \ln \alpha + \delta_i \ln h_0 + \delta_i \boldsymbol{x}_i^T \boldsymbol{\beta} + \ln \Gamma(\alpha + \delta_i) \\ &- \ln \Gamma(\alpha) - (\alpha + \delta_i) \ln(\alpha + h_0 t_i e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}) \end{aligned} \tag{10.35}$$

This log-likelihood is difficult to maximize as it contains an unspecified baseline hazard; thus the EM Algorithm needs to be implemented to solve for the unknown parameters. In order to run the EM Algorithm the complete data likelihood and the marginal distribution of the unobserved data is needed. The marginal distribution of the unobserved data is found to be

$$\begin{aligned} g(z_i|x_i;\alpha) &= \frac{L_i(z_i,t_i;\alpha)}{L_{obs,i}(t_i,\alpha)} \\ &= \frac{(\alpha+h_0t_ie^{\boldsymbol{x}_i^T\boldsymbol{\beta}})^{\alpha+\delta_i}}{\Gamma(\alpha+\delta_i)}Z_i^{\alpha+\delta_i-1}e^{-z_i(\alpha+h_0t_ie^{\boldsymbol{x}_i^T\boldsymbol{\beta}})} \end{aligned} \tag{10.36}$$

Clearly this is a two parameter gamma distribution with parameters $(\alpha+\delta_i, \alpha+h_0t_ie^{\mathbf{x_i^T}\beta})$. The EM Algorithm can now be used to solve for the unknown parameters. It should be noted that if the baseline hazard is not constant, then

$$\int_0^t h_0(t)du = \Lambda(t).$$

Penalized Partial Likelihood Approach

We shall use most of the work by Therneau and Grambsch (2000, pp 232-233) and Nguti (2003) in this section. The whole concept of penalized partial likelihood estimation originates from the cubic splines regression in the Cox proportional hazards model. When using the penalized partial likelihood approach for this estimation, the random effects w_i are used rather than the frailties z_i. We will assume the univariate frailty model with the corresponding equations

$$h_i(t) = h_0(t)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}+w_i) \tag{10.37}$$

$$S_i(t) = \exp[\Lambda_0(t)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}+w_i)] \tag{10.38}$$

$$L_i(w_i,t_i;\alpha) = f_w(w_i)\times h_i(t)^{\delta_i}S_i(t) \tag{10.39}$$

$$\ell_i(\alpha;w_i,t_i) = \ln f_w(w_i)+\delta_i\ln h_i(t)+\ln S_i(t) \tag{10.40}$$

Substituting Eq. (10.37) into the full data log-likelihood Eq. (10.40) gives the contribution for individual i as

$$\begin{aligned}\ell_i(\alpha; w_i, t_i) &= \ln f_w(w_i) + \delta_i(\ln[h_0(t)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta} + w_i)]) - \Lambda_0(t)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta} + w_i) \\ &= \ln f_w(w_i) + \delta_i(\ln h_0(t) + \boldsymbol{x}_i^T\boldsymbol{\beta} + w_i) - \Lambda_0(t)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta} + w_i).\end{aligned}$$

Thus the full data log-likelihood can be written as

$$\tilde{\ell}_{full}(\beta, \alpha, h_0) = \tilde{\ell}_{full,1}(\boldsymbol{\beta}, h_0) + \tilde{\ell}_{full,2}(\alpha)$$

where

$$\begin{aligned}\tilde{\ell}_{full}(\boldsymbol{\beta}, \alpha, h_0) &= \sum_{i=1}^{I}[\delta_i(\ln h_0(t) + \boldsymbol{x}_i^T\boldsymbol{\beta} + w_i) - \Lambda_0(t)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta} + w_i)] \\ \tilde{\ell}_{full,2}(\alpha) &= \sum_{i=1}^{I}\ln f_w(w_i)\end{aligned}$$

where $\tilde{\ell}_{full,2}(\alpha)$ can be seen as the penalty term, where the mean for the w_i's is 0. For $w_i << 0$ or $w_i >> 0$, $f_w(w_i)$ is small and thus $\log f_w(w_i)$ takes on a large negative value which in turn decreases the likelihood, in other words it acts like a penalty. We therefore take

$$\tilde{\ell}_{full,2}(\alpha) = -\ell_{pen}(\alpha)$$

with

$$\ell_{pen}(\alpha) = -\sum_{i1=}^{I}\ln f_w(w_i)$$

In order to apply semi-parametric ideas, consider the w_i's in $\tilde{\ell}_{full,1}(\boldsymbol{\beta}, h_0)$ as 'parameters' with corresponding covariates similar to that of a design matrix Z in the equation

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{w}$$

where $\boldsymbol{Z}$ is the design matrix (Janssen, 2005). Using the partial likelihood ideas, $\tilde{\ell}_{full,1}(\boldsymbol{\beta}, h_0)$ is replaced by

$$\ell_{part}(\boldsymbol{\beta}, w) = \sum_{i=1}^{I} \delta_i \left[\eta_i - \ln \left(\sum_{qw \in R(t_i)} \exp(\eta_{qw}) \right) \right]$$

where $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + w_i$. Thus in order to make inference on the parameters β and α, the following penalized partial likelihood is used

$$\ell_{ppl}(\beta, \alpha, w) = \ell_{part}(\boldsymbol{\beta}, w) - \ell_{pen}(\alpha, w)$$

Application to the log gamma density

If the frailties are assumed to be gamma distributed, then the random terms, the w_i's are log gamma distributed, with probability density function

$$f_w(w) = \frac{\alpha^\alpha (\exp(w)^\alpha) \exp[-\alpha \exp(w)]}{\Gamma(\alpha)}.$$

Taking the natural log of this p.d.f results in

$$\ln f_w(w) = \alpha(w - \exp(w)) - (\alpha \ln \alpha + \Gamma(\alpha)).$$

Hence $\ell_{pen}(\alpha)$ is given by

$$\ell_{pen}(\alpha) = -\sum_{i=1}^{I} (\alpha(w_i - \exp(w_i))) + I(\alpha \ln \alpha + \Gamma(\alpha)).$$

Nguti (2003) and Therneau and Grambsch (2000) state that the maximization of the penalized partial likelihood consists of an inner loop and an outer loop. In the inner loop the rule is given a provisional value of α, the Newton-Raphson procedure is employed to maximize $\ell_{ppl}(\boldsymbol{\beta}, \alpha, \mathbf{w})$ for $\boldsymbol{\beta}$ and $\mathbf{w}$ to obtain the best linear unbiased predictors (BLUP). In the outer loop, a log likelihood similar to $\ell_{obs}(.)$ is maximized for α as in the case of the EM algorithm. Let ℓ denote the outer loop index, and k the inner loop index. Let $\alpha^{(\ell)}$

be the estimate of the ℓ^{th} iteration of the outer loop. Then $\boldsymbol{\beta}^{(\ell,k)}$ and $\mathbf{w}^{(\ell,\mathbf{k})}$ are the estimates and predictions for $\boldsymbol{\beta}$ and $\mathbf{w}$ at the k^{th} iterative step, given $\alpha^{(\ell)}$. The starting value of $\boldsymbol{\beta}$ is obtained from the estimates from fitting a normal Cox model, and for starting values of $\mathbf{w}^{(\mathbf{1},\mathbf{0})}$ and $\alpha^{(1)}$ the k^{th} iterative step given by

$$\begin{bmatrix} \boldsymbol{\beta}^{(\ell,k)} \\ \mathbf{w}^{(\ell,\mathbf{k})} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}^{(\ell,k-1)} \\ \mathbf{w}^{(\ell,\mathbf{k-1})} \end{bmatrix} V^{-1} \begin{bmatrix} \mathbf{0} \\ (\alpha^{(\ell)})^{-1}\mathbf{w}^{(\ell,\mathbf{k-1})} \end{bmatrix} + V^{-1} \begin{bmatrix} X & Z \end{bmatrix} \frac{d\ell_{part}(\boldsymbol{\beta},\mathbf{w})}{d\boldsymbol{\eta}}$$

where

$$\begin{aligned} V &= \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \\ &= \begin{bmatrix} X^T \\ Z^T \end{bmatrix} \left(\frac{-\partial^2 \ell_{part}(\boldsymbol{\beta},\mathbf{w})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T} \right) \begin{bmatrix} X & Z \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & (\alpha^{(\ell)})^{-1} I_I \end{bmatrix} \end{aligned}$$

Note that $X = (\mathbf{x_{11}}, \ldots, \mathbf{x_{In_I}})^{\mathbf{T}}$ is a $n \times p$ covariate matrix with $n = \sum_{i=1}^{I} n_i$, $Z = \text{diag}(\mathbf{1_{n1}}, \ldots, \mathbf{1_{nI}})$ with $\mathbf{1}_{ni}$ as a column vector of size n_i with all entries one, and $\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zw}$ such that $\boldsymbol{\eta^T} = (\boldsymbol{\eta_{11}}, \ldots, \boldsymbol{\eta_{In_I}})$ as in Nguti (2003). Once the Newton Raphson procedure has converged for the current value of $\alpha^{(\ell)}$, the procedure moves to the outer loop of the algorithm. In the outer loop of the algorithm, a golden section search method (Press et al., 1992), described in the appendix, is applied to a modified version of the log-likelihood in order to update the estimate of α. The likelihood is

$$\begin{aligned} \ell_{part,obs}(\boldsymbol{\beta},\mathbf{w}) &= \ell_{part}(\boldsymbol{\beta},\mathbf{w}) \\ &= \sum_{i=1}^{I} \left[\ln\left(\frac{\Gamma(D_i+\alpha)}{\Gamma} \right) + \alpha \ln\left(\frac{\alpha}{\Lambda+\alpha} \right) - D_i \ln(D_i+\alpha) + D_i \right] \end{aligned}$$

and the details of how this is derived are found in Therneau and Grambsch (2000) and Nguti (2003). The algorithm continues until the stopping criterion given by

$$|\ell_{part,obs}(\hat{\beta}^{(\ell)}, \alpha^{(\ell)}, \hat{w}^{(\ell)}) - \ell_{part,obs}(\hat{\beta}^{(\ell-1)}, \alpha^{(\ell-1)}, \hat{w}^{(\ell-1)})| < \varepsilon^*$$

is reached. We do not consider the Bayesian and the MCMC approaches although details of it can be found in Zuma and Lurie (2005) and Bolstad and Manda (2000).

10.10 Fitting the frailty model to the RSV data using the PPL approach

We apply the penalized partial likelihood approach (PPL) to the RSV data even though consideration was given to the EM Algorithm. The reason for this approach is that the EM Algorithm and the PPL approach produce the same estimates and there was no readily available software to run the frailty model using the EM Algorithm. Further to this point, the EM algorithm can take up to ten times longer to converge than the PPL approach, rendering it inefficient. The Cox model was fitted to include the explanatory variables of age, prev, visit and timemonth but STATA could not compute the results of this model since the likelihood was found to go to infinity. Hence the Cox model that was fitted was

$$h_{ij}(t) = h_0(t) \exp(\beta_1 age + \beta_2 prev + w_i)$$

where $i = 1, \ldots, 338$ and w_i is the random effect for child i. The model was fitted in STATA. The commands that were given in STATA to fit the Cox proportional hazards frailty model are:

```
stset dt, failure(rsvpos) set matsize 336 stcox age
prev,shared(id) nohr effects(logfr)
```

The option 'nohr' requests that parameter estimates instead of hazard ratios to be given in the output. If one exponentiates the parameter esti-

mates then the hazard ratios are recovered. The matsize of 336 owes itself to a large data set and is calculated as 334 children plus 2 covariates. The default size in STATA for matsize is 200 but can be expanded to a maximum of 800. It is apparent that matsize depends on the model and its parameters being fitted. The 'effects()' component of the 'stcox' statement requests that the log-frailties be printed for each child/indivivdual/cluster and this variable that STATA will print in the original data set should be called 'logfr'. Obvioulsy one can specify any name that one wishes to name the log frailties variable. It took 4 iterations to estimate the frailty variance and 3 iterations to fit the final Cox model. The final log-likelihood was worked out as -1201.4732. As a default, STATA uses the Breslow (1974) method for handling ties. For the random effect of the children denoted as 'id', a shared gamma frailty model was used. A summary of the results are given below: The likelihood ratio test for $\theta = 0$ has a χ^2 statistic of 0.14 with a probability

Variable	Parameter Estimate	Standard Error	z	Pr$> z$	Hazard Ratio
Age	-0.0769973	0.0242813	-3.17	0.002	0.9259
Prev	51.22724	5.429363	9.44	0.000	1.76e22
θ	0.06121	0.17165			

Table 10.5: Parameter Estimates for gamma shared frailty model

of 0.355. Thus the child specific random effect is not significant and models without the random effect seem to be better to use. However the analysis has shown how such frailty models can be fitted. The 'age' and 'prev variables are significant in this model, with 95% confidence intervals given as (-0.12459, -0.0294) for 'age' and (40.58589, 61.8686) for 'prev'. The frailty variance is 0.06121 implying that there is not much variability from child to child in this data set. This makes sense since most children were within one year of age thus the sample was more homogenous than heterogenous. The

log frailties for every child is also printed upon request. There is a log frailty for every child and this list will not be printed, however for the purposes of interpretation, 2 children will be looked at: The log frailty for individ-

id	rsvpos	d_{ij}	age	prev	log-frailty
1	0	0	1	0.018182	0.027906
1	1	40	2	0.041276	0.027906
1	0	7	3	0.025424	0.027906
⋮	⋮	⋮	⋮	⋮	⋮
5	1	0	1	0.018182	-0.03287
5	1	33	2	0.041276	-0.03287
5	1	5	2	0.041276	-0.03287
⋮	⋮	⋮	⋮	⋮	⋮

Table 10.6: Data description showing log frailty

ual/child 1 is 0.027906. The frailty is worked out as $e^{0.027906} = 1.028$. Since this value is greater than 1, individual/child 1 is more frail than a standard individual/child. For individual/child 5, the log frailty is -0.03287, so the frailty is worked out again as $e^{-0.03287} = 0.9677$. Thus individual 5 is less frail than a standard individual/child. However these values are not significantly different from 1.

Two other models were fitted, one by adding in the 'actipass' variable, whether a child was actively or passively sampled, and the other by adding in the time in months since the beginning of the study, 'timemonth'. The results for the models are summarized below:

The final log-likelihood was worked out as -1107.79. A summary of the results are given below: The likelihood ratio test for $\theta = 0$ has a χ^2 statistic of $1.9e-05$ with a probability of 0.498. Thus the child specific random effect is not significant and models without the random effect seem to be better to use. The 'age','prev' and 'actipass' variables are significant in this model,

Variable	Parameter Estimate	Standard Error	z	Pr> z	Hazard Ratio
Age	-0.1430	0.0254	-5.61	0.000	0.8668
Prev	48.1380	4.961	9.70	0.000	7.99e22
Actipass	-2.434	0.1784	-13.64	0.000	0.0877
θ	1.13e-07	0.000018			

Table 10.7: Parameter Estimates for gamma shared frailty model

with 95% confidence intervals given as (-0.193, -0.093) for 'age', (38.41, 57.86) for 'prev' and (-2.78,-2.08) for 'actipass'. The frailty variance is $1.13e-07$ implying that there is not much variability from child to child in this data set as discussed in the first model above. The frailty variance tends to zero, hence there is no apparent frailty. This explains the results now to follow. Here again we look at the log frailty for, 2 children: The log frailty for individ-

id	rsvpos	d_{ij}	age	prev	log-frailty
1	0	0	1	0.018182	-1.84e-08
1	1	40	2	0.041276	-1.84e-08
1	0	7	3	0.025424	-1.84e-08
⋮	⋮	⋮	⋮	⋮	⋮
5	1	0	1	0.018182	-4.12e-08
5	1	33	2	0.041276	-4.12e-08
5	1	5	2	0.041276	-4.12e-08
⋮	⋮	⋮	⋮	⋮	⋮

Table 10.8: Data description showing log frailty

ual/child 1 is $-1.84e08$. The frailty is worked out as $e^{-1.84e08} = 0.9999 \approx 1$. Since this value is equal to 1, individual/child 1 is equally as frail as a standard individual/child. For individual/child 5, the frailty is worked out again as $e^{-4.12e08} = 0.99999 \approx 1$. Thus individual 5 is equally as frail as a standard individual/child. Finally the final model is summarized below. The final log-

likelihood was worked out as -1105.64. A summary of the results are given below: The likelihood ratio test for $\theta = 0$ has a χ^2 statistic of $2.4e-05$ with

Variable	Parameter Estimate	Standard Error	z	Pr$> z$	Hazard Ratio
Age	0.016	0.0817	2.00	0.845	1.016
Prev	46.268	5.0447	9.17	0.000	1.241e22
Actipass	-2.422	0.17838	-13.58	0.000	0.0877
Timemonth	-0.158	0.0775	-2.04	0.041	0.8538
θ	1.14e-07	9.68e-06			

Table 10.9: Parameter Estimates for gamma shared frailty model

a probability of 0.498. Thus the child specific random effect is not significant and models without the random effect seem to be better to use. We find that 'prev', 'actipass' and 'timemonth' are all significant at the 5% level. Here again we look at the log frailty for, 2 children:

id	rsvpos	d_{ij}	age	prev	log-frailty
1	0	0	1	0.018182	-5.02e-08
1	1	40	2	0.041276	-5.02e-08
1	0	7	3	0.025424	-5.02e-08
⋮	⋮	⋮	⋮	⋮	⋮
5	1	0	1	0.018182	-5.14e-08
5	1	33	2	0.041276	-5.14e-08
5	1	5	2	0.041276	-5.14e-08
⋮	⋮	⋮	⋮	⋮	⋮

Table 10.10: Data description showing log frailty

The log frailty for individual/child 1 is $-5.02e08$. The frailty is worked out as $e^{-5.02e08} = 0.9999 \approx 1$. Since this value is equal to 1, individual/child 1 is equally as frail as a standard individual/child. For individual/child 5, the frailty is worked out again as $e^{-5.14e08} = 0.99999 \approx 1$. Thus individual 5 is also equally as frail as a standard individual/child. It seems that perhaps the marginal models are better suited to model the RSV disease from the survival analysis approach.

10.11 Conclusion

The marginal models of AG, PLP and WLW models all gave similar results. The frailty models showed the inclusion of the child effect to be negligible. The frailty model took longer to converge in STATA. The infected and uninfected child seem to be equally frail. It could be due to the large number of uninfections as compared to the small number of infections in this scientific setting. It seems in the survival analysis setting that the marginal models are better suited to model the RSV data set.

Chapter 11

Concluding Remarks

The aim of this research study was to develop methods of modelling disease outcome data with reference to the RSV as the particular disease process. In addition to the work, the force of infection and the recovery rate were estimated using frequency based approaches. Both of these aims were successfully met. The longitudinal structure of the data presented one with three families of modelling approaches namely the marginal, random effects and transition models. All three models were investigated as well as Survival analysis type of models. All models have merits and demerits, depending on the nature of the problem and assumptions imposed.

In this scientific setting, there were not many infected state transitions as compared to the overwhelmingly large number of uninfected to infected state transitions. The use of generalised linear modelling combined with likelihood estimation was used to estimate the force of infection and the recovery rate of the childhood respiratory viral disease (RSV). Construction of the full likelihood was not possible therefore a form of conditional likelihood was used to model the data. The generalised modelling approach was modified to esti-

mate stepwise monthly specific force of infection for the disease thus allowing the model to capture the temporal trends of disease incidence. Assuming time dependence the force of infection is estimated as $\hat{\lambda} = 0.0012$ and the rate of recovery is estimated as $\hat{\nu} = 0.4550$ using the direct maximum likelihood estimation method. Corresponding estimates using the generalized linear modelling approach were 0.0021 and 0.5032. These two approaches gave quite similar sets of parameter estimates. However the latter approach is preferred because of its flexibility in allowing the estimation of monthly piecewise parameter estimates. It is also seen from the estimation of the monthly parameters that RSV peaks at particular months in the year namely May, January and February. This result demonstrates that RSV like many child infections is seasonal in nature.

The marginal models allowed us to use GEE with different correlation structures where the exchangeable as well as the independent structures were deemed most suitable to model the disease at successive time intervals. Age, prevalence of RSV in the blood as well as whether or not a child was actively or passively sampled was related to the current RSV status of the child. When the random effect of the child was added into the model, it was found that the data did not exhibit much variability from child to child. It must also be said that SAS allows a good flexibility in using Proc GLIMMIX and Proc NLMIXED to fit these models. The transition models, which are also heavily criticized for using a the history term Y_{ij-1}, the previous response for child i at time occasion j, gave odds ratios. The plausibility of the second and third order Markov models must be interpreted with caution because the second and third order history terms when compared to the present response term were not found to be significant. The odds ratios also increased as the

history variables are increased to higher orders. The decreasing pattern of the odds ratios was evident seen across all the variables: 'age', 'dt', 'prev', 'actipass' and 'timemonth' in all three conditional models involving the first, second and third order history terms. It is also important to emphasize that a difference exists with respect to the interpretation of the fixed effects $\boldsymbol{\beta}$, for example the $E(Y_i)$ has different estimates in terms of the classical linear mixed model and the non-linear mixed model. In the general case the fixed effects under the marginal model, say, $\boldsymbol{\beta}^{marginal}$ and the random effects models, $\boldsymbol{\beta}^{randomeffects}$ are different to each other in the sense that when the random effects model is considered, the marginal mean profile can be derived but will not produce a simple parametric form. One needs to be careful of this in the interpretation of these fixed effects under the different families of models.

The RSV data set also had two issues of missingness associated with it namely the 85 missing values in the response variable, $\boldsymbol{Y}$ and the dropout process. The intermittent missingness is not estimated. The model was estimated using different methods such as the EM algorithm and LOCF. Thereafter different models were refitted showing not much difference when compared to the original data estimates. The dropout process was best handled using WGEE and GLMM which both carry the relaxed assumption of MAR. The LOCF method was also implemented and compared to the other methods. LOCF produced inflated and deflated estimates in its estimation of the key disease parameters and the model parameters. LOCF is clearly an unreasonable analysis to make in the RSV data set, especially if we assume that a child's disease state will continue to be the same as the last state prior to dropout. LOCF has attracted a lot of criticism from several authors such as Kenward and Molenberghs (1998), Jansen et al. (2006) and Siddiqui and Ali

(1998), just to name a few. It continues to be a flawed technique of handling the dropout in many scientific settings. The survival modelling approach allowed us the flexibility of taking into account the interarrival times between the child's disease states and build the relevant models.

The survival analysis approach involved fitting marginal models and frailty models to the RSV data, which in the survival analysis setting could be treated as multiple events per subject. The marginal and frailty models can be thought of as an extension of the Cox proportional hazards model. The marginal type models involved the fitting of the AG, PLP and WLW models. The three models gave similar results and are useful in modelling the RSV data set. The frailty model includes the frailty term in the model and is thought to account for subject or individual random heterogeneity and any possible clustering that may have occurred in the data set. The frailty term in the RSV data set is the child effect. There are different types of frailty modelling, namely, univariate, multivariate and shared frailty modelling. Univariate frailty modelling occurs when the frailty term may be at an individual level, where every individual is assumed to have a different frailty due to unmeasured covariates. Multivariate frailty modelling is such that each cluster is assigned a frailty term and all individuals within a cluster are assumed to have the same frailty. The frailty term is thought to account for the correlated nature of the data. Thus, it is extremely important to consider including frailty in a survival model, and is particularly useful in scientific settings where clustering needs to be accounted for. The child effect in our frailty model was not significant, perhaps due to the fact that RSV is a rare disease and the data set does not exhibit much variation from one child to another in terms of the disease status. There are still avenues of

research especially in multi-level frailty modelling.

The modelling of the RSV disease is a varied and complex one. However one has got to take into account of the nature of the data as well as the scientific settings. This project highlighted and addressed these questions in a meaningful and satisfying way. There is definitely more avenues of research for modelling diseases and estimating their parameters. More sophisticated models that carry the MNAR assumption for handling the dropout are also pathways of research. Kenward and Carpenter (2007) point out that a more rigorous theoretical framework needs to be developed for multiple imputation via chained equations There is still much to be accomplished as far as disease modelling is concerned but this area is relevant and imperative to biostatistics.

Bibliography

Aaerts, M.J., Geys, H., Molenberghs, G., and Ryan, L.M. (2002) *Topics in Modelling of Clustered Binary Data*. London: Chapman and Hall.

Afifi, A. and Elashoff, R. (1966) Missing observations in multivariate statistics I: Review of the literature. *Journal of the American Statistical Association,* **61**, 595-604.

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

Albert, P.S. and McShane, I.M. (1995) A generalized estimating equations approach for spatially corelated data binary data: application to the analysis of neuroimaging data. *Biometrics,* 51, 627-638.

Allison, P.D. (1995) *Survival Analysis using SAS: A Practical Guide.* SAS Institute Inc., Cary, NC, USA.

Andersen, P.K., and Gill, R.D. (1982) Cox's regression model for

counting processes: A large sample study. *Annals of Statistics,* **10**, 1100-1120.

Ashford, J.R. and Sowden, R.R. (1970) Multivariate probit analysis. *Biometrics,* **26**, 535-546.

Ayis, S.A.M. (1995) Modelling unobserved heterogenity : theoretical and practical aspects. *Ph.D. Thesis.* University of Southampton.

Bahadur, R.R. (1961) A representation of the joint distribution of responses to n dichotomous outcomes. In:*Studies in Item Analysis and Prediction,* H. Solomon (Ed.). Stanford Mathematical Studies in the Social Sciences VI. Stanford, CA: Stanford Univeristy Press.

Baker, S.G. (1992) A simple method for computing the observed information matrix when using the EM algorithm with categorical data. *Journal of Computational and Graphical Statistics,* **1**, 63-76.

Beale, E.M.L. and R.J.A. Little (1975) Missing values in multivariate analysis. *Journal of the Royal Statistical Society, Series B,* **37**, 129-145.

Bolstad, W.M. and Manda, S.O. (2001) Investigating child mortality in Malawi using family and community effects: A Bayesian analysis. *Journal of the American Statistical Associa-*

tion, **96**, 12-19.

Breslow, N.E. (1974) Covariance analysis of censored survival data . *Biometrics,* **30**, 89-100.

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

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

Browne, M.W. (1974) Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal,* **8**, 1-24.

Buck, S.F. (1960) A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, Series B,* **22**, 302-306.

Burton, P., Gurrin, L., and Sly, P. (1998) Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Statistics in Medicine,* **17**, 1261-1291.

Cane, P. (2001) Molecular and epidemiology of respiratory

syncytial virus. *Reviews in Medical Virology,* **11**, 103-116.

Cane, P. and Pringle, C. (1992) Molecular epidemiology of respiratory syncytial (RSV) virus: rapid identification of subgroup A isolates. *Journal of Virological Methods,* **40**, 297-306.

Carey, V.C., Zeger, Z.L. and Diggle, P.J. (1993) Modelling multivariate binary data with alternating logistic regressions. *Biometrika,* **80**, 517-526.

Chananty, N.R. (1997) An alternative approach to the analysis of longitudinal data via generalized estimating equations. *Journal of Statistical Planning and Inference,* **63**, 39-54.

Chew, F., Doraisingham, S., Ling, A., Kumarsinghe, G., and Lee, B. (1998) Seasonal trends of viral respiratory tract infections in the tropics. *Epidemiology and Infection,* **121**, 121-128.

Collett, D. (1994) *Modelling Survival Data in Medical Research.* London: Chapman and Hall.

Collins, P.L., McIntosh, K., and Channock, R.M. (1996) Respiratory Syncytial Virus. *Fields' Virology. 3rd ed.*, 1313-1351.

Cortinas Abrahantes, J. and Burzykowski, T. (2004) A version of the EM algorithm for proportional hazard model with random effects. *Biometrical Journal,* **47**, 847-862.

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

Cox, D.R. (1972) Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B,* **74**, 187-200.

Cox, D.R. (1974) Estimation of the correlation between a continuous and a discrete variable. *Biometrics,* **30**, 171-178.

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

Cox, D.R. and Wermuth, N. (1994) *Multivariate Dependencies: Models, Analysis and Interpretation.* London: Chapman and Hall.

Crowder, M.J. (1995) On the use of a working correlation matrix in using genrealized linear models for repeated measurements. *Biometrika,* **82**, 407-410.

Crowder, M.J. and Hand, D.J. (1990) *Analysis of Repeated Measures.* (1st edition) London: Chapman and Hall.

Cytel Software Corporation (2000) *EGRET for Windows, User Manual.*

Dale, J.R. (1986) Global cross ratio models for bivariate, discrete, ordered responses. *Biometrics,* **42**, 721-727.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B,* **39**, 1-38.

Desmond, A. (1997) Optimal Estimating Functions, Quasi-likelihood and Statistical Modelling. *Journal of Statistical Planning and Inference,* **60**, 77-121.

Diggle, D.J. and Donnelly, J.B. (1989) A selected bibliography on the analysis of repeated measurements and related areas. *Australian Journal of Statistics,* **31**, 183-193.

Diggle, D.J., Heagerty, P.J, Liang, K.-Y., and Zeger, S.L. (2002) *Analysis of Longitudinal Data.* Oxford Science Publications. Oxford:Claredon Press.

Dixon, W.J. (1983) *BMDP Statistical Software.* Berkeley: University of California Press.

Duchateau, L., Janssen, P., and Rowlands, J.G. (1998) *Linear Mixed Models: An Introduction with Applications in Veterinary research.* ILRI (International Livestock Research Institute)

Nairobi, Kenya.

Efron, B. (1977) The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association,* **72**, 557-565.

Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Annals of Statistics,* **7**, 1-26.

Efron, B. (1994) Missing data, imputation and the bootstrap (with discussion). *Journal of the American Statistical Association,* **89**, 463-479.

Efron, B. and Tibsharani, R.J. (1993) *An introduction to Bootstrap.* New York: Chapman and Hall.

Engel, B. (1998) A simple illustration of the failure of PQL, IRREML and APHL as approximate ML methods for mixed models for binary data. *Biometrical Journal,* **40**, 141-154.

Engel, B. and Buist, W. (1996) Analysis of a generalized linear mixed model: a case study and simulation results. *Biometrical Journal,* **38**, 61 - 80.

Engel, B. and Buist, W. (1998) Bias reduction of approximate maximum likelihood estimates for heritability in threshold models. *Biometrics,* **54**, 1155 - 1164.

Engel, B. and Busit, W., and Visscher, A. (1995) Inference for threshold models with variance components from the generalized linear mixed model perspective. *Genetics Selection Evolution,* **27**, 15 - 32.

Engel, B. and Keen, A. (1994) A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica,* **48**, 1-22.

Engel, B. and Keen, A. (1996) An introduction to generalized linear mixed models. Invited paper. *Proceedings XIIIth International Biometric Conference,* Amsterdam.

Fahemeir, L. and Gerhard, T. (1994) *Multivariate Statistical Modelling Based on Generalized Linear Models* USA: Springer Verlag.

Feller, W. (1968) *An Introduction to Probability Theory and Its Applications* (3rd edition) New York: John Wiley .

Fessler, J.A. and Hero, A.O. (1994) Space-alternating generalized expectation maximization algorithm. *IEEE Transactions and Signal Processing,* **42**, 2664-2677.

Fitzmaurice, G.M. (1995) A caveat concerning independence estimating equations with multivariate binary data. *Biometrics,*

51, 309-317.

Fitzmaurice, G.M. and Laird, N.M. (1993) A likelihood based method for analysing longitudinal binary responses. *Biometrika,* **80**, 141-151.

Fitzmaurice, G.M., Laird, N.M. and Tosteson, T. (1996) Polynomial exponential models for clustered binary outcomes. *unpublished manuscript.*

Fitzmaurice, G.M., Laird, N.M. and Ware, J.H (2004) *Applied Longitudinal Analysis.* New York: John Wiley and Sons.

Fraley, C. (1999) On computing the largest fraction of missing information for the EM algorithm and the worst linear function for data augmentation. *Computational Statistics and Data Analysis,* **31**, 13-26.

Harville, D. (1974) Bayesian inference for variance components using only error contrasts. *Biometrika,* **61**, 383-385.

Gelman, A.E., and Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science,* **7**, 457-472.

Geyer, C.J. (1992) Practical Markov chain Monte Carlo (with discussion). *Statistical Science,* **7**, 473-503.

Geys, H., Molenberghs, G., and Lipsitz, S.R. (1998) A note on the comparison of pseudo-likelihood and generalized estimating equations for marginal odds ratio models. *Journal of Statistical Computation and Simulation,* **62**, 45-72.

Geys, H., Molenberghs, G., and Ryan, L.M. (1997) Pseudo-likelihood inference for clustered binary data. *Communications in Statistics: Theory and Methods,* **26**, 2743-2767.

Geys, H., Molenberghs, G., and Ryan, L.M. (1997) Pseudo-likelihood modelling of multivariate outcomes in developmental toxicology . *Journal of the American Statistical Association,* **94**, 734-745.

Glonek, G.F.V. and McCullagh, P. (1995) Multivariate logistic models. *Journal of the Royal Statistical Society, Series B,* **81**, 477-482.

Goldstein, H. (1991) Nonlinear multilevel models with an application to discrete response data. *Biometrika,* **78**, 45-51.

Goldstein, H. (1995) *Multilevel Statistical Models (2nd edition).* London: Arnold.

Goldstein, H. and Rasbash, J. (1996) Improved approximations for multilevel models with binary responses. *Journal of the*

Royal Statistical, Society A, **159**, 505 - 513.

Greenhalgh, D., Deikmann, O., and de Jong, M.C.M. (2000) Subcritical endemic steady states in mathematical models for animal infections with incomplete immunity. *Mathematical Biosciences,* **165**, 1-25.

Hall, D. (2001) On the application of extended quasi-likelihood to the clustered data case. *The Canadian Journal of Statistics,* **29**, 77-97.

Hall, D. and Severini, T.A. (1998) Extended generalized estimating equations for clustered data. *Journal of the American Statistical Association,* **93**, 1365-1375.

Hannan, E.J. and Tate, R.F. (1965) Estimation of the parameters for a multivariate normal distribution when one variable is dichotomized. *Biometrika,* **52**, 664-668.

Hartley, H.O. (1958) Maximum likelihood estimation from incomplete data. *Biometrics,* **14**, 174-194.

Hartley, H.O. and Hocking, R. (1971) The analysis of incomplete data. *Biometrics,* **27**, 7783-7808.

Healy, M.J.R. and Westmacott, M. (1956) Missing values in experiments analyzed on automatic computers. *Applied Statistics,*

5, 203-206.

Heckman, J.J. (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement,***5**, 475-492.

Hedeker, D. and Gibbons, R.D. (1994) A random effects ordinal regression model for multilevel analysis. *Biometrics,* 50, 933-944.

Hedeker, D. and Gibbons, R.D. (1996) MIXOR: a computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine,* **49**, 157-176.

Heitjan, D.F. (1997) Annotation: What can be done about missing data? Approaches to imputation. *American Journal of Public Health,* **87**(4), 548-550.

Heyting, A, Tolboom, J.T.B.M., and Essers, J.G.A. (1992) Statistical handling of dropouts in longitudinal clinical trials. *Statistics in Medicine,* **11**, 2043-2061.

Hogan, J.W. and Laird, N.M. (1997) Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine,* **16**, 239-258.

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

Hougaard, P. (2000) *Analysis of Multivariate Survival Data. Statistics for biology and health.* New York: Springer Verlag.

Huisman, M. (2000) Imputation of missing item responses: Some simple techniques. *Quality and Quantity,* **34**, 331-351.

Jackson, R.W.B. (1939) Reliability of mental tests. *British Journal of Psychology,* **29**, 267 - 287.

Jamshidian, M. and Jennrich, R.I. (1993) Conjugate gradient acceleration of the EM algorithm. *Journal of the American Statistical Association,* **88**, 221-228.

Jansen, I., Beunckens, C., Molenberghs, G., Verbeke, G., and Mallinckrodt, C. (2006) Analyzing incomplete discrete longitudinal cllinical trial data. *Statistical Science,* **21**, 52-69.

Kalbfleisch, J.D. and Prentice, R.L. (1980) *The Statistical Analysis of Failure Time Data.* New York: John Wiley and Sons.

Keiding, N., Andersen, P.K., and Klein, J.P. (1997) The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in*

Medicine, **16**, 215-224.

Kenward, M.G. (1994) Computing the generalized estimating equations with quadratic covariance estimation for repeated measurements. *Genstat Newsletter,* **32**, 50-62.

Kenward, M.G. (2006) Multiple Imputation.*South African Statistical Association (SASA) Conference Workshop,* Stellenbosch.

Kenward, M.G. and Carpenter, J. (2007) Multiple Imputation: current perspectives. *Statistical Methods in Medical Research,* **16**, 199-218.

Kenward, M.G. and Molenberghs, G. (1998) Likelihood based frequentist inference when data are missing at random . *Statistical Science,* **12**, 236-247.

Kenward, M.G. and Molenberghs, G. (1999) Parametric models for incomplete continuous and categorical longitudinal studies data.*Statistical Methods in Medical Research,* **8**, 51-83.

Kenward, M.G., Molenberghs, G., and Lesaffre, E. (1994) An application of maximum likelihood and estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics,* **50**, 945-953.

Klein, J.P. (2000) Semiparametric estimation of random effects using the Cox Model based on the EM algorithm. *Biometrics,* **48**, 795-806.

Kleinman, J. (1973) Proportions with extraneous variance:single and independent samples. *Journal of the American Statistical Association,* **68**, 46-54.

Koch, G.G. and Freeman, D.H., and Freeman, J.L. (1975) Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review,* **43**, 59-78.

Korn, E.L. and Whittemore, A.S. (1979) Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics,* **35**, 795-802.

Krzanowski, W.J. (1988) *Principles of Multivariate Analysis.* Oxford: Clarendon Press.

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

Laird, N.M. and Oliver, D. (1981) Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association,* **76**, 231-240.

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

Lang, J.B. and Agresti, A. (1994) Simultaneously modelling joint and marginal distributions of multuivariate categorical responses. *Journal of the American Statistical Association,* **89**, 625-632.

Lavori, P.W., Dawson, R., and Shera, D. (1995) An imputation strategy for clinical trials with truncation of patient data. *Statistics in Medicine,* **14**, 1913-1925.

Le, C.T. (1997) *Applied Survival Analysis.* New York: John Wiley and Sons.

Le Cressie, S. and Van Houwelingen, J.C. (1994) Logistic regression for correlated binary data. *Applied Statistics,* **43**, 95-108.

Lee, Y., Nelder, J.A. and Pawitan, Y. (2006) *Generalized Linear Models with Random Effects.*(1st edition) London: Chapman and Hall.

Levin, J.B. (1999) *The use of variance component models in the analysis of complex sureys.* Unpublished Ph.D Dissertation, University of Natal.

Li, K.H., Raghunathan, T.E., and Rubin, D.B. (1991) Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica,* **1**, 65-92.

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

Liang, K.-Y., Zeger, S.L. and Qaqish, B. (1992) Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B,* **54**, 3-40.

Lin, D., Wei, L., and Ying, Z. (1993) Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika,* **80**, 557-552.

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

Lindsey, J.K. (1999) *Models for Repeated Measurements.* Oxford University Press, New York.

Lipsitz, S.R., Laird, N.M. and Harrington, D.P. (1991) Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika,* **78**, 153-160.

Littell, R.C. and Milliken, G.A., Stroup, W.W., and Wolfin-

ger, R.D. (1996) *SAS System for Mixed Models.* Cary, NC : SAS Institute Inc.

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

Little, R.J.A. (1988) Robust estimation of the mean and covariance matrix from data with missing values. *Applied Statistics,* **37**, 23-38.

Little, R.J.A. (1992) Regression with missing X's: a review. *Journal of the American Statistical Association,* **87**, 1227-1237.

Little, R.J.A. (1993) Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association,* **88**, 125-134.

Little, R.J.A. (1994) A class of pattern mixture models for normal missing data. *Biometrika,* **81**, 471-483.

Little, R.J.A. (1995) Modeling the drop-out mechanism in longitudinal studies. *Journal of the American Statistical Association,* **90**, 1112-1121.

Little, R.J.A., and Rubin, D.B. (1987) *Statistical Analysis with Missing Data.*New York: John Wiley and Sons.

Little, R.J.A., and Rubin, D.B. (2002) *Statistical Analysis with Missing Data.*New York: John Wiley and Sons.

Little, R.A. and Schluchter, M.D. (1985) Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika,* **72**, 497-512.

Liu, C.H., and Rubin, D.B. (1994) The ECME algorithm: a simple extension of the EM and ECM with fast monotone convergence. *Biometika,* **81**, 633-648.

Liu, C.H., Rubin, D.B. and Wu, Y.N. (1998) Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometika,* **85**, 755-770.

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

Longford, N.T. (1993) *Random Coefficient Models, 2nd ed.* New York: Oxford University Press.

Louis, T.A. (1982) Finding the observed information when using the EM algorithm. *Journal of the Royal Statistical Society, Series B,* **44**, 226-233.

Mallinckrodt, C.H., Sanger, T.M., Dube, S., DeBrota, D.J.,

Molenberghs, G., Carroll, R.J., Potter, W.Z., Tollefson, G.D. (2003) Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biological Psychiatry,* **53**(8), 754-760.

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

McKendrick, A.G. (1926) Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society,* **44**, 98-130.

McLachlan, G.J. and Krishnan, T. (1997) *The EM Alogorithm and Extensions.*New York: John Wiley and Sons.

Meilijson, I. (1989) A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society, Series B,* **51**, 127-138.

Meng, X.-L. (1994) Multiple imputation inferences with uncongenial sources of input. *Statistical Science,* **9**, 538-558.

Meng, X.-L. (1997) The EM alogorithm and medical studies: a historical link. *Statistical Methods in Medical Research,* **6**, 3-23.

Meng, X.-L., and Pedlow, S. (1992) EM: a bibliographic review

with mssing articles. *Proc. Statist. Computing Sec., American Statistical Association,* 24-27.

Meng, X.-L., and Rubin, D.B. (1991) Using the EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association,* **86**, 899-909.

Meng, X.-L., and Rubin, D.B. (1993) Maximum likelihood estimation via the EM algorithm: a general framework. *Biometrika,* **80**, 267-278.

Meng, X.-L., and van Dyk, D. (1997) The EM algorithm-an old folk song sung to a fast new tune. *Journal of the Royal Statistical Society, Series B,* **3**, 511-567.

Molenberghs, G. and Danielson, L. (1999) Simple methods for the analysis of multivariate and longitudinal categorical data. In:*Proceedings of the* 7^{th} *International Conference on Probability Theory and Mathematical Statistics and Vilnius Conference (1998),* B. Grigelionis et al (Eds.), TEV, Vilnius/VSP, Utecht, pp. 499-514.

Molenberghs, G. and Kenward, M.G. (2007) *Missing Data in Clinical Studies.* New York: John Wiley and Sons.

Molenberghs, G., Kenward, M.G. and Lesaffre, E. (1997) The analysis of longitudinal ordinal data with non-random

dropout. *Biometrika,* **84**, 33-44.

Molenberghs, G. and Lesaffre, E. (1994) Marginal modelling of correlated ordinal data using a multivariate Plackett distribution *Journal of the American Statistical Association,* **89**, 633-644.

Molenberghs, G. and Lesaffre, E. (1999) Marginal modelling of multivariate categorical data. *Statistics in Medicine,* **18**, 2237-2255.

Molenberghs, G. and Ritter, L. (1996) Likelihood and quasi-likelihood based methods for analysing multivariate categorical data, with the association between outcomes of interest. *Biometrics,* **52**, 1121-1133.

Molenberghs, G. and Ryan, L.M. (1999) Likelihood inference for clustered multivariate binary data. *Envirometrics,* **10**, 279-300.

Molenberghs, G. and Verbeke, G. (2005) *Discrete Models for Longitudinal Data.*(1st edition) USA: Springer Verlag.

Molenberghs, G. and Verbeke, G. (2006) *Lecture Notes: Longitudinal Data Analysis.*

Morris, J.A., Blount, R.E. Jnr., and Savage, R.E. (1956) Recovery of cytopathogenic agent from chimpanzees with coryza *Proc. Soc. Exp. Biol. Med.,* **92**, 544-549.

Nagelkereke, G., Chunge, R.N. and Kinoti, S.N. (1990) Estimation of parasite infection dynamics when dectectability is imperfect. *Statistics in Medicine,* **9**, 1211-1219.

Neuhaus, J.M., Kalbfleish, J.D., and Hauck, W.W. (1991) A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review,* **59**, 25-35.

Nguti, R. (2003) *Random Effects Survival Models Applied to Animal Breeding Data.* Unpublished Ph.D Dissertation, Limburgs Universitair Centrum, Belgium.

Nokes, D.J., Okiro, E., Ngama, M.J., White, L.J. ,Ochola, R., Scott, P.D., Cane, P.A. and Medley, G.F. (2004) RSV epidemiology in a birth cohort in Kilifi District, Kenya: infection in the first year of life. *Journal of Infectious Diseases,* **190**, 1828-1832.

Oakes, D. (1992) *"Frailty models for multiple event times"*, J. P. Klein and P. K. Goel (eds.), Survival Analysis: State of the Art, 371-379, Kluwer Academic Publisher: Netherladns.

Oakes, D. (1999) Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society, Series B,* **61**, 479-482.

O'Hara-Hines, R. (1998) Comparison of two covariance structures in the analysis of clustered polytomous data using generalized estimating equations. *Biometrics,* **54**, 312-316.

Olkin, I. and Tate, R.F. (1961) Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics,* **32**, 448-465 (with correction in **36**, 343-344).

Orchard, T and Woodbury, M.A. (1972) A missing information principle: Theory and applications. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics,* **1**, 697-715.

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

Pinheiro, J.C. and Bates, D.M. (1995) Approximations to the log-likelihood function in the non-linear mixed effects model. *Journal of Computational and Graphical Statistics,* **4**, 12-35.

Pinheiro, J.C. and Bates, D.M. (2000) *Mixed effects models in S and S-Plus.* Springer-Verlag: New York.

Plackett, R.L. (1965) A class of bivariate distrbutions. *Journal of the American Statistical Association,* **60**, 516-522.

Prentice, R.L. and Cai, J. (1992) Covariance and survivor function estimation using multivariate failure time data. *Biometrika,* **79**, 495-512.

Prentice, R.L., Williams, B.J., and Petersen, A.V. (1981) On the regression analysis of multivariate failure time data. *Biometrika,* **68**, 373-379.

Prentice, R.L. and Zhao, L.P. (1991) Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics,* **47**, 825-839.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1992) *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, Cambridge UK.

Press, W.L, Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. (1992) *Numerical Recipes in FORTRAN,* Cambridge University Press (2nd edition).

Rabe-Hesketh, S.,Pickles, A. and Skrondal, A. (2001) Gllamm manual, Technical report, Department of Biostatistics and Computing, Institute of Psychiatry, King's College, University of London [Report no. 2001/01].

Raudenbush, S.W., Bryk, A.S., Cheong, Y.F., Fai, Y., and Congdon, R. (2001) *HLM5: Hierarchical linear and non-linear*

modelling, Lincolnwood, IL: Scientific Software International.

Raudenbush, S.W., Yang, M.L., and Yosef, M. (2000) Maximum llikelihood for generalized linear models with nested random effects via high-order, multivariate laplace approximations. *Journal of Computational and Graphical Statistics,* **9**, 141-157.

Ripatti, S., Larsen, K., and Palmgren, J. (2002) Maximum likelihood inference for multivariate frailty models using an automated Monte Carlo EM algorithm. *Lifetime Data Analysis,* **8**, 349-360.

Ripatti, S. and Palmgren, J. (2000) Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics,* **56**, 1016-1022.

Robins, J.M, Rotnizky, A., and Zhao, L.P. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association,* **90**, 106-121.

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

Rosenbaum, P.R., and Rubin, D.B. (1983) The central role of the propensity score in observational studies for causal effects.

Biometrika, **70**, 41-55.

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

Ross, R. (1915) Some a priori pathometric equations. *British Medical Journal,* **1**, 546.

RSV information website http:www.rsvinfo.com.

Rubin, D.B. (1976) Inference with missing data. *Biometrika,* **63**, 581-592.

Rubin, D.B. (1977) Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association,* **77**, 538-543.

Rubin, D.B. (1997) Comment on 'The EM algorithm- An old folk song sung to a fast new tune (with discussion)' by Meng, S.L., and van Dyk, D. *Journal of the Royal Statistical Society, Series B,* **59**, 511-567.

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley and Sons.

Rubin, D.B. (1996) Multiple imputation after 18+ years (with

discussion) . *Journal of the Aamerican Statistical Association,* **91**, 473-489.

Rubin, D.B., and Schafer, J.L. (1990) Efficiently creating mulitple imputations for incomplete multivariate normal data. *Proceedings of the Statistical Computing Section, American Statistical Association,* 83-88.

Ryan, T.P. and Woodall, W.H. (2005) The most-cited statistical papers. *Journal of Applied Statistics,* **32**, 461-474.

Sammel, M.D., Ryan, L.M., and Legler, J.M. (1997) Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society, Series B,* **59**, 667-678.

SAS Institute Inc. (2004) *SAS/STAT User's Guide* Version 9.1, Volumes 3-5 . SAS Institute Inc., Cary, NC.

Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data.* London: Chapman and Hall.

Schafer, J.L. (1999) Multiple imputation: a primer. *Statistical Methods in Medical Research,* **8**, 3-15.

Schafer, J.L. (2001) NORM Version 2.03 for Windows 95/98/NT. Multiple imputation of incomplete multivariate data under a normal model available from the URL:

http://www.stat.psu.edu/jls/misoftwa.html.

Schafer, J.L. (2002) Discussion on "A nonparametirc approach to matched pairs with missing data". *Sociological Methods and Research,* **30**, 458-459.

Schafer, J.L. and Olsen, M.K (1998) Multiple Imputation for Multivariate Missing-Data Problems:A Data Analyst's Perspective. *Multivariate Behaviourial Research,* **33**, 545-571.

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

Searle, S.R. (1988) Mixed Models and Unbalanced Data: Wherefrom, whereat, whereto? *Communication in Statististics: Theory Methods,* **17**, 935-968.

Searle, S.R., Casella, G. and McCulloch, C.E. (1992) *Variance Components.*New York: John Wiley & Sons, Inc.

Segal, M.R., Neuhaus, J.M. and and James, I.R. (1997) Dependence estimation for marginal models of multivariate survival data. *Lifetime Data Analysis,* **3**, 251-268.

Self, S.G. and Liang, K.-Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *Journal of the American Statistical*

Association, **82**, 605-610.

Shults, J. and Changanty, N.R. (1998) Analysis of serially correlated data using quasi-likelihood. *Biometrics,* **54**, 1622-1630.

Siddiqui, O and Ali, M.W. (1998) A comparison of the random effects pattern-mixture model with last-observation-carried-forward (LOCF) analysis in longitudinal clinical trials with dropouts. *Journal of Biopharmaceutical Statistics,* **8**, 545-563.

Simoes, E. (1999) Respiratory syncytial virus infection.*Lancet.,* **354**, 847-852.

Skellam, J.G. (1948) A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society, Series B,* **10**, 257-261.

Skrondal, A. and Rabe-Hesketh, S. (2004) *Generalized latent variable modelling.* London: Chapman and Hall.

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

Stigler, S.M. (1994) Citation patterns in the journals of statistics and probablility. *Statistical Science,* **9**, 94-108.

Stram, D.O. and Lee, J.W. (1994) Variance components testing in the longitudinal mixed effects model. *Biometrics,* **50**, 1171-1177.

Stram, D.O. and Lee, J.W. (1995) Correction to "Variance components testing in the longitudinal mixed effects model". *Biometrics,* **51**, 1196.

Struthers, C.A. and Kalbfleisch, J.D. (1986) Misspecified proportional hazards models. *Biometrika,* **73**, 363-369.

Sutradhar, B.C. and Das, K. (1999) On the efficiency of regression estimators in genrealized linear models for longitudinal data. *Biometrika,* **86**, 459-465.

Tanner, M.A (1993) *Tools for statistical inference.* New York: Springer-Verlag.

Therneau, T.M. and Grambsch, P.M. (2000) *Modelling Survival Data: Extending the Cox model.* New York: Springer-Verlag.

Thiébaut, T., Hélène, J.-G., Chêne, G., Leport, C and Commenges, D. (2002) Bivariate linear mixed models using SAS proc MIXED. *Computer Methods and Programs in Biomedicine,* **69**, 249-256.

Tuerlinckx, F., Rijmen, F., Molenberghs, G., Verbeke, G., Briggs, D., Noortgate, W., Van den Meulders, M., and Boeck,

P. De (2004) Estimation and software, in P. De Boeck and M. Wilson, editors, *Explanatory item response models,* Statistics for Social Science and Public Policy, chapter 12, 343-373, Springer Verlag: New York.

Vaida, F. and Xu, R. (2000) Proportional hazards model with random effects. *Statistics in Medicine,* **19**, 3309-3324.

Van Buuren, S. (2007) Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research,* **16**, 219-242.

Vaupel, J.W., Manton, K.G., and Stallard, E. (1979) The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography,* **16**, 439-454.

Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data.* USA: Springer Verlag.

Verbeke, G. and Molenberghs, G. (2003) The use of score tests for inference on variance components. *Biometrics,* **59**, 254-262.

Waddington, D., Welham, S.J., Gilmour, A.R., and Thompson, R. (1994) Comparisons of some GLMM estimators for a simple binomial model. *Genstat Newsletter,* **30**, 13 - 24.

Wang, Y.-G. and Carey, V.J. (2004) Unbiased estimating equations from working correlation models for irregularly timed repeated measures. *Journal of the American Statistical Association,* **99**, 845-853.

Wei, G.C.G. and Tanner, M.A. (1990) A Monte Carlo implmentation of the EM algorithm and the Poor Man's Data Augmentation algorithms. *Journal of the American Statistical Association,* **85**, 699-704.

Wei, L.J., Lin, D.Y., and Weissfeld, D. (1989) Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. *Journal of the American Statistical Association,* **84**, 1065-1073.

Weiss, G.H. and Dishon, M. (1971) On the asymptotic behaviour of the stochastic and deterministic models of an epidemic. *Mathematical Biosciences,* **11**, 261-265.

Welham, S.J. (1993) *Procedure GLMM. In: Genstat 5 procedure Library Manual, Release 2[3]* editors. R.W. Payne and G.M. Arnold. Oxford: Numerical Algorithms Group.

White, L.J., Nokes, D., Ngama, N.J., Okiro, E.A., Medley, G.F. and Nokes, D.J. (2003) Mechanistically based probability models for RSV transimission and determinants of the rate of infection.*In preparation.*

Wolfinger, R. (1993) Laplace's approximation for nonlinear mixed models. *Biometrika,* **80**, 791-795.

Wolfinger, R. (1998) Towards practical appplication in generalized linear mixed models, in B. Marx and H. Friedl, editors, *Proceedings of the* 13^{th} *International Workshop on Statistical Modelling,* 388-395, New Orleans, Louisiana, USA.

Wolfinger, R. and O'Connell, M. (1993) Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation,* **48**, 233-243.

Wu, M.C. and Bailey, K.R. (1988) Analysing changes in the presence of informative right censoring cause by death and withdrawal. *Statistics in Medicine,* **7**, 337-346.

Wu, M.C. and Bailey, K.R. (1989) Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics,* **45**, 939-955.

Wu, M.C. and Carroll, R.J. (1988) Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process. *Biometrics,* **44**, 175-188.

Zeger, S.L., Liang, K.-Y., and Albert, P.S. (1988) Models for longitudinal data: a generalized estimating equation approach.

Biometrics, **44**, 1049-1060.

Zeger, S.L., Liang, K.-Y., and Self, S.G. (1985) The analysis of binary longitudinal data with time dependent covariates . *Biometrika,* **72**, 31-38.

Zhao, L.P. and Prentice, R.L. (1990) Correlated binary regression using a quadratic exponential model. *Biometrika,* **77**, 642-648.

Zhao, L.P., Prentice, R.L. and Self, S.G. (1991) Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society, Series B,* **54**, 805-811.

Zuma, K. and Lurie, M.N. (2005) Application and comparison methods for analysing correlated interval censored data from sexual partnerships. *Journal of Data Science,* **3**, 241-256.

Appendix A

Some SAS Proc IML Programs

```
/* this is a programme to get individual transition matrices*/
proc iml;
 use lisa;
read all into xx; x=xx[,7]; id=xx[,1]; y=j(2,736,0); do j=1 to
368; do i=1 to nrow(x)-1; if ((id[i]=j) & (id[i+1]=j)) then do; if
((x[i]=1) & (x[i+1]=1)) then y[1,(2*j)-1]=y[1,(2*j)-1]+1; else if
((x[i]=1) & (x[i+1]=2)) then y[1,2*j]=y[1,2*j]+1; else if
((x[i]=2) & (x[i+1]=1)) then y[2,(2*j)-1]=y[2,(2*j)-1]+1; else if
((x[i]=2) & (x[i+1]=2)) then y[2,2*j]=y[2,2*j]+1; end; end; end;
yy=y‘; print yy;

/* this was the programme to calculate the visits by months*/ proc
iml; use lisa; read all into xx; tt={1 3 7}; xx=xx[,tt];
mid=max(xx[,1]); m=j(mid,450,0); do id=1 to mid;
help=j(1,ncol(xx),0); do i=1 to nrow(xx); if xx[i,1]=id then
help=help//xx[i,]; end; nid=nrow(help); if nid>1 then tt1=2:nid;
else tt1=1; tt1=tt1‘; help0=help[tt1,]; help1=cusum(help0[,2]); do
```

```
i=1 to nid-1;
 if i=1 then do;
   mm=1;
   jj=1;
 end;
else do; jj=help1[i]; mm=help0[i,3]; end; m[id,jj]=mm; end; end;
v=j(2,14,0); do i=1 to 14; do c=(30*i)-28 to (30*i)+1; do j=1 to
368; if m[j,c]=1 then v[1,i]=v[1,i]+1; if m[j,c]=2 then
v[2,i]=v[2,i]+1; end; end; end; print v;

/* this is a programme to estimate the force of infection and the
rate of recovery using a GLM */
proc sort data=lisa out= rsv1; by
id visit; run; data rsv2; set rsv1; x1=lag(id); x2=lag(rsvpos); if
x1 ne id then x2=.; run; proc sort data=rsv2 out=rsv3; by  id
visit; run; proc print data=rsv3; var id visit dt actipass
symptoms age rsvpos x2 timemonth prev; run; /* The x2 variable is
the Y_t-1 and the rsvpos is the Y_t variable*/ data rsv4;
 set rsv3;
 if x2 ^=.;
 run;

data rsv4;
 set rsv4;
 yt=rsvpos-1;
 yt1=x2-1;
 ni=1;
```

```
 timeit=timemonth;
 run;
 proc print data=rsv4;
var id visit ni dt rsvpos yt x2 yt1; run;
 proc freq data=rsv4;
table yt*yt1; run; data rsv5; set rsv4; z=0; if yt=0 and yt1=0
then z=1; if yt=1 and yt1=1 then z=1; index1=yt1; index2=(1-yt1);
ldt=log(dt); run;
 proc freq data=rsv5; table z; table z*yt*yt1;
run; proc freq data=rsv5; table index1*index2; run;

proc print data=rsv5; run;
*************************************************************************;
* the model with constant lambda and nu                                 *;
* the model was fitted with cloglog link for P(Z=0)                     *
*                                                                       *
* lambda=exp(coeff index2)                                               *;
* nu=exp(coeff index2)                                                   *;
* ldt=log(dt)                                                           *;
* C.I exp(C.I)                                                          *
*************************************************************************;

proc genmod data=rsv5 ; model z= index1 index2/dist=bin
link=cloglog offset=ldt noint; run;
 data rsv51; set rsv5; if
yt1=0; run;
```

```
*************************************************************************;
* time dependent force of infection with constant recovery rate       *
* the model was fitted with cloglog link for P(Z=0)                   *
*                                                                     *
* lambda=exp(coff index1)                                             *;
* nu=exp(coff index2)                                                 *;
* ldt=log(dt)                                                         *;
* C.I exp(C.I)                                                        *
*************************************************************************;

proc freq data=rsv5; table timemonth; run;
 data rsv51; set rsv5;
if 1< timemonth <= 13; run;
 proc genmod data=rsv51; class
timemonth; model z= index2*timemonth index1/dist=bin link=cloglog
offset=ldt noint; run;
```