C H A P T E R 1

INTRODUCTION

1.1 Background to Study

When measurements are taken from the same subject more than once, the responses are no longer independent. Longitudinal data is characterised by clusters of repeated measurements, each obtained from a single subject (Fitzmaurice, Laird & Ware, 2004). Longitudinal study designs are often used for environmental or ecological studies to measure trends over time. Although these studies can be expensive, time consuming and difficult to analyse, they allow the direct study of change over time and the factors which influence this change, as well as assessing within-subject changes (Lindsey, 1993; Twisk, 2003; Fitzmaurice *et al*., 2004). Analysing the growth curves of individuals over time, or determining the effects of the continued administration of treatments over time are examples of when longitudinal studies would be required (Lindsey, 1993).

A classic example of growth curve analysis is the data set of Potthoff and Roy (1964) (the PR data set). Their data set consists of measurements obtained during a dental study from 11 girls and 16 boys at the ages of 8, 10, 12 and 14. This study was conducted by researchers at the University of North Carolina Dental School, who measured the distance between the pituitary and pterygomaxillary fissure for each child using x-ray exposures of the side of the head taken every two years. The purpose of this study was to examine growth of this structure over time and to determine if this differs between girls and boys.

A second example is from an ecological study (Kirton, 2005) conducted on water hyacinth plants. Sixty plants were chosen at random and divided equally between two herbicide treatments, one using a sublethal dose of herbicide and the other using no herbicide. Three nutrient treatments were also used with varying amounts of phosphate and nitrate. Weekly measurements were taken from the plants over eight weeks. The length from the base of each plant until the tip of the second youngest petiole was measured each week. The purpose of this study was to examine the growth of the plants over time and to determine if this changed depending on whether the herbicide was applied or not, and if this depended on nutrient level.

These data sets could be analysed using several different models of varying sophistication. Taking the PR data set as an example, let y_{ij}, $i = 1,\ldots,N$ and $j = 1,\ldots,n_i$, be the length between the pituitary and pterygomaxillary fissure for the i^{th} individual at the j^{th} measurement occasion, where there are N individuals and n_i measurement occasions for the i^{th} individual ($n_i = 4$ for all individuals in this example). A simple approach to analysing these data would be to conduct a two sample t-test between the measurements from the girls and the measurements from the boys. This approach, although easy to implement, would be invalid and would ignore the time effect in the data. This is because more than one observation from each individual would be included in the data, thereby violating the assumption of independent observations. One way of getting around this assumption, provided the data could be assumed to be normally distributed and continuous, would be to perform multiple t-tests (Crowder & Hand, 1990; Davis, 2002). Therefore t-tests would be performed between the measurements of the girls and boys at each measurement occasion. The difficulty using this approach would be in deciding on an overall conclusion, since some of the

tests may show significant differences and others may not, leading to the possibility of subjective conclusions. Alternatively, a t-test could be performed on the data from the final measurement occasion only, but this would result in a huge amount of data wastage. In particular, this method would not allow for an analysis of growth trends.

To compare the measurements at different time points, paired t-tests could be performed between the data at two different ages. All possible paired combinations of ages could be considered. Because the test comparing *time 1* to *time 2* will be related to the test comparing *time 2* to *time 3* and *time 1* to *time 3*, these tests are not independent, and this can cause the probability of finding at least one test significant to increase spuriously (Crowder & Hand, 1990; Davis, 2002).

Subject, gender and time could be included in an analysis of variance (ANOVA) approach to analysing the data, resulting in the model $y_{ij} = \beta_0 + \beta_1\delta_i + \beta_{2i} + \beta_{3j} + \varepsilon_{ij}$, where δ_i is an indicator for gender, and β_{2i} and β_{3j} are adjustments to the mean response for the i^{th} individual and the j^{th} measurement occasion respectively, and ε_{ij} is the error term. Alternatively, time can be included as a continuous covariate, changing this to an analysis of covariance (ANCOVA). Since subject is included in the mean structure of this model, this approach would imply that the subjects included were the only subjects of interest and inference could not be made beyond these individuals. It also does not allow for the inclusion of variability arising from the random sampling process, and therefore underestimates the variability in the data (Allison, 2005).

The above analyses can be refined by subtracting a base value, usually the measurement taken at time zero for a particular subject, from the measurements of

each subject, thereby allowing each subject to be its own control (Crowder & Hand, 1990).

A different approach could be to summarise the vector of measurements for each individual into one summary measure (Crowder & Hand, 1990). For this method to be effective, a summary measure needs to be chosen that will adequately describe the subjects' data (Crowder & Hand, 1990; Davis, 2002). This method is referred to as response feature analysis. Examples of response features include the mean, maximum rate of increase, time to reach maximum rate of increase, half-life, or the slope of the least squares regression line. The data then simplifies to $y_i^+ = \beta_0 + \beta_1 \delta_i + \varepsilon_i^+$ and y_i^+ and ε_i^+ are the respectively the response feature and random error of the response feature for subject *i*. These methods require the assumption that the variance of the derived response feature be homoscedastic. This would be violated if there are different numbers of observations being summarised for each individual, implying that this can only be achieved when there are no missing values and the number and sequence of measurements are the same for each individual (Fitzmaurice *et al*., 2004).

All of the methods discussed so far result in information loss and make very strong assumptions about the data, such as homogeneity of variance (Crowder & Hand, 1990; Fitzmaurice *et al*., 2004). None of these methods consider the covariance between repeated measures on the same individual, which may contain much information about the total response of an individual. Therefore in order to take full advantage of the longitudinal study design, methods of analysis which explicitly include the covariance between repeated measures should be used.

1.2 Problems Related to Using Simple Techniques

Although methods, such as reducing the repeated measurements into a single summary measurement, can be useful for exploring the data (Davis, 2002), the use of overly simple analyses for repeated measures results in a loss of the richness of information inherent in longitudinal data. Additionally it can lead to efficiency loss, i.e. increasing the variability while not capitalising on the information available in the data, as well as biasing the results (Weiss, 2005).

Loss of efficiency can result from omitting subjects, e.g. because they contain missing data, or from omitting observations in order to accommodate a certain method of analysis. Using methods that can utilise all of the available information, thereby making better use of the data, will result in more efficient estimates (Weiss, 2005).

Bias can be introduced into the analysis in a number of ways, e.g. by means of inappropriate experimental designs, inappropriate analysis, or leaving out subjects for reasons related to the study. If the design of a study leads to subjects being sampled so that the true sampled population is different to the intended population of interest, then the results of the analysis will be biased in favour of the subset of the population that was sampled. Therefore appropriate randomisation is important to avoid bias. The same result will occur if a poorly selected subset of otherwise well collected data is chosen for analysis. This problem can result from the omission of subjects with missing data. If the “missing-ness” of the data is related to the outcome of the study, then this type of omission will bias the results of the analysis (Weiss, 2005).

Using simple analyses on data with missing values can lead to several problems. For example, it can result in subjects being compared that have been measured at entirely different points in time, resulting in differences in their averages and slopes over time simply because they were measured at different times and for no other reason. Once data belonging to these subjects is summarised into an average or slope, there is no way of identifying this problem (Weiss, 2005). When there are missing data for certain subjects it also means that the variance for the derived summary measures is no longer the same, thereby complicating the analysis by invalidating the assumption of homogeneity of variance required for standard parametric methods (Fitzmaurice *et al*., 2004)

Very different reasons for groups showing differences in a longitudinal study can result in exactly the same result using a simple statistical analysis. For example, two groups that have different means may have the same slope over time, or the slopes could be very different, yet in both cases the same difference in means may be found. Therefore simple analyses are very limited in the types of conclusions that can be drawn from them. Alternatively, it is also very possible that two groups with very different responses over time can result in a non-significant result. For example, two groups may have the same average over time, but their slopes could be very different. Therefore these groups respond differently over time, but their averages do not convey this information (Weiss, 2005; Fitzmaurice *et al*., 2004). Even more sophisticated means of analysis such as repeated measures ANOVA is too restrictive in the compound symmetry assumption for the covariance structure, which assumes equal covariance between all repeated measures, and can lead to overly conservative conclusions (Fitzmaurice *et al*., 2004).

Much of the loss of information resulting from overly simple methods of analysis is due to the disregard of the covariance between observations. Only by incorporating the covariance into the analysis is it possible to make predictions of the subjects' responses through time (Weiss, 2005).

1.3 Linear Mixed Effects Models

One of the most widely used methods of including the covariance matrix in the analysis is through the linear mixed effects model (Laird & Ware, 1982; Verbeke & Molenberghs, 2000; Davis, 2002; Fitzmaurice *et al*., 2004; Ugrinowitsch, Fellingham & Ricard, 2004; Vittinghoff, Glidden, Shiboski & McCulloch, 2005). Mixed effects models are those where the mean is modelled through both random and fixed effects.

Fixed effects are those factors in a model for which the designer of the experiment had deliberately chosen certain levels, and which are the only levels of interest, rather than randomly sampling levels from an infinite population of possible levels (Vittinghoff *et al*., 2005). An example of a fixed effect would be if a researcher were interested in the growth of a certain species of plant under different nitrogen levels, and then selected five different nitrogen levels under which the plants would be grown, say 1%, 2%, 3%, 4% and 5%, which are considered to be the only levels of interest. Here the researcher would be investigating the impact of increasing nitrogen levels on the growth of the plant. Nitrogen can be considered as a fixed effect in the model since each plant in a particular treatment group would receive the same amount of nitrogen. When a researcher chooses individuals for a study in such a way that

specifically both males and females are included, then gender can be considered as a fixed effect.

When the researcher does not explicitly choose the levels of a factor, but rather the levels are a sample of the possible levels available, then this is known as a random effect (Fitzmaurice *et al*., 2004). In the PR data set the children included in the study are an example of a random effect, as they were randomly selected from a larger population of children, whereas the gender of the children can be considered a fixed effect since all levels of interest are represented. Including individual specific random effects into a model can be used to account for correlation among repeated measurements (Fitzmaurice *et al*., 2004; Vittinghoff *et al*., 2005).

Linear mixed effects models are a special case of mixed effects models in which both the fixed and random effects occur linearly in the model function. The most common formulation of the model is that of Laird and Ware (1982):

$$\mathbf{y_i} = \mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{b_i} + \boldsymbol{\varepsilon_i} \quad \text{for } i = 1,\ldots, N$$
$$\mathbf{b_i} \sim \mathbf{N(0, \Sigma)}, \quad \boldsymbol{\varepsilon_i} \sim \mathbf{N(0, \sigma^2 I)}$$

where $\mathbf{y}_i(n_i \times 1)$ are independent and normally distributed, $\boldsymbol{\beta}$ is the p-dimensional vector of fixed effects, $\mathbf{b}_i$ is the q-dimensional vector of random effects, $\mathbf{X}_i(n_i \times p)$ and $\mathbf{Z}_i(n_i \times q)$ are known fixed effects and random effects regressor matrices respectively, and $\boldsymbol{\varepsilon}_i$ is the n_i-dimensional within-individual error vector with a spherical Gaussian distribution. It is assumed that $\mathbf{b}_i$ and $\boldsymbol{\varepsilon}_i$ are independent for different individuals and that they are independent of each other for the same individual. A structure needs to be chosen for the covariance matrix of $\mathbf{b}_i$, $\boldsymbol{\Sigma}$, and, in the more general formulation, for the covariance matrix of $\boldsymbol{\varepsilon}_i$. The consequences of these structural choices will be the

main consideration of this study. In Chapter two further details concerning this model will be discussed.

1.4 Study Objectives

My interest in repeated measures models stems from my interaction with postgraduate students in the life sciences and my observation of the types of methods these students use to analyse repeated measures data. I am particularly interested in (1) the consequences of using an over-simplified model, namely the ordinary linear regression model which assumes independence of repeated measurements, to analyse repeated measures data, and (2) if an appropriate model is chosen, what the consequences are of using an incorrect parameterisation of the covariance structure for the estimates of the fixed effects and inferences about these estimates.

The objective of my study was to investigate the use of linear mixed effects models to analyse repeated measures data, with an application to an ecological data set. I simulated models under various available covariance structures and determined if a covariance structure or structures exist that perform well under misspecification. The linear mixed effects model was fitted to the ecological data set, and by means of goodness-of-fit measures, the results of the simulation study were validated. The interpretability of the linear mixed effects models is also discussed in the context of the ecological study.

Using the PR data set, I carried out a simulation study to determine the consequences of incorrect covariance structure choice. I fitted linear mixed effects models with

different covariance structures and then, using the estimated parameters of each of these models, I simulated more data and then investigated the effect of fitting linear mixed effects models with different covariance structures to these data sets, with the intention of determining how robust these methods are to misspecification of the covariance structure. This was done through the use of goodness-of-fit measures and measures of robustness.

For the ecological data set, I fitted the various repeated measures models to the data set and obtained the best fitting repeated measures models for the data. The fit of the models was assessed using goodness-of-fit tests and residual diagnostics. These results were then contrasted with the results from the simulation study. I then compared these analyses against a simpler, but invalid, method which may have been used by an inexperienced researcher to analyse this data, and determined if there were any differences in the conclusions drawn from the different types of analyses.

1.5 Other Issues

Complications can occur in a longitudinal study including missing values for particular individuals, and responses that are not continuous. These issues will not be explicitly considered in this study. The researcher should be aware that missing data is a common problem in some longitudinal studies. In longitudinal studies, specifically clinical trials, observations for different subjects are usually missing, leading to differences in sample sizes between individuals, for some reason related to the outcome of the study, as discussed by Vittinghoff *et al*. (2005). Missing data during an out-patient clinical trial can lead to overly optimistic estimates for the sicker

patients, as they may have arrived for fewer of the follow-up visits, and those visits where they were present may well have been days when they were feeling relatively well, thereby biasing the results towards healthier patients. During an in-hospital trial, there would be more data for the sicker patients, as the healthier patients would have been discharged, thereby biasing the results towards the sicker patients. If the "missing-ness" of the data is due to a factor already included in the model, then the "missing-ness" will not bias the results, but if the "missing-ness" is not related to one of the factors in the model then the results will be especially misleading. This is not easily dealt with and methods that attempt to correct for such missing data based on assumptions of the missing data mechanism (such as informative missing data or not missing at random methods) need to be used (Davis, 2002, p. 22; Vittinghoff *et al*., 2005, p. 286)