C H A P T E R 2

LITERATURE REVIEW

2.1 Basic Aspects of Longitudinal Studies

Repeated measures models are designed to take into account the stochastic dependence in longitudinal data. Two types of stochastic dependence exist between the responses: the homogeneity of the responses on the same unit and the heterogeneity across units; and the distance (in time or space) among responses on the same unit (Lindsey, 1993, p. 6). The correct specification of the stochastic interdependence model is important because the model of dependence among responses can have a great influence on the ability of the complete model to describe the observations (Lindsey, 1993).

The response variable in a repeated measures design can be in the form of count data, such as the number of eggs laid; binary, such as absence and presence of eggs; categorical responses, such as the type of damage to a leaf, which can be aggregated into counts; or in the form of continuous data, such as the growth in height of a plant. These responses may have come from a study where the subjects have undergone some treatment or treatments, or accompanying covariates may have been measured (Lindsey, 1993). In some studies, measurements may have been taken through space rather than time. Only continuous responses taken through time will be considered in this study.

Randomisation is required to allocate subjects to treatment groups so that bias is avoided. Lindsey (1993, p. 9) notes that randomisation allows for statements of causality, since which treatment a subject receives is not influenced by the response that the subject gives. It also minimises the effects of inter-response variability by distributing it randomly over treatments, thereby ensuring homogeneity of variability. In order to attribute causality, the relationship between the cause and the effect needs to be strong, and the relationship should be consistent in different populations and under different circumstances. In addition, the cause needs to lead to a single effect (specificity) and the cause must precede the effect in time (temporality). To conclude that a cause and effect relationship exists there needs to be experimental evidence and theoretical (e.g. biological) plausibility (Twisk, 2003, p.2).

2.2 Repeated Measures Models

There are a number of different types of linear models that can be used to analyse repeated measures. Some of these models will be described, with the main emphasis falling on linear mixed effects models.

2.2.1 Fixed effects models

A number of relatively different types of models for longitudinal data fall into this category, including the simpler models discussed in the Chapter one. More sophisticated methods will now be discussed.

Potthoff and Roy (1964) were the first to propose an extension to the standard multivariate ANOVA (MANOVA) model for growth curve analysis. Davis (2002) gives detail on this type of analysis. Suppose there are s treatment groups, and let N_h denote the number of individuals in group h, $h = 1,\ldots,s$, so that $N = \sum_{h=1}^{s} N_h$. Let y_{hij} denote the responses of the i^{th} subject in group h at the j^{th} measurement occasion, where $i = 1,\ldots,N_h$ and $j = 1,\ldots,t$. It is assumed that repeated measurements have been obtained from each individual at t equally spaced time points. In growth curve analysis it is assumed that the time trend in each group can be described by a $(\nu-1)$-degree polynomial, with $\nu \leq t$. The formulation of the growth curve model is then

$$y_{hij} = \beta_{h0} + \beta_{h1} j + \beta_{h2} j^2 + \ldots + \beta_{h,\nu-1} j^{\nu-1} + e_{hij}$$

where e_{hij} is the error for the i^{th} individual in group h at the j^{th} time point. This formulation of the time trend is assumed to be the same for each group, but the different parameter values may differ over the groups, leading to a total of $s\nu$ parameters (Davis, 2002).

Let $\mathbf{y}_{hi} = (y_{hi1},\ldots, y_{hit})'$ and $\mathbf{e}_{hi} = (e_{hi1},\ldots,e_{hit})'$ denote the vectors of responses and errors respectively of the i^{th} individual in group h, and let $\mathbf{Y} =$

$$\begin{pmatrix} y_{111} & y_{112} & \cdots & y_{11j} & \cdots & y_{11t} \\ y_{121} & y_{122} & \cdots & y_{12j} & \cdots & y_{12t} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{1N_1 1} & y_{1N_1 2} & \cdots & y_{1N_1 j} & \cdots & y_{1N_1 t} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{hi1} & y_{hi2} & \cdots & y_{hij} & \cdots & y_{hit} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{sN_s 1} & y_{sN_s 2} & \cdots & y_{sN_s j} & \cdots & y_{sN_s t} \end{pmatrix}$$

denote the $N\times t$ matrix of responses and $\mathbf{E}$ the corresponding $N\times t$ matrices of errors. The growth curve model would then be written as $\mathbf{Y} = \mathbf{XBT} + \mathbf{E}$ where $\mathbf{X}$ is an $N\times s$ across-individual design matrix indicating an

individual's group, $\mathbf{B} = \begin{pmatrix} B_{10} & B_{11} & \cdots & B_{1,\nu-1} \\ B_{20} & B_{21} & \cdots & B_{2,\nu-1} \\ \vdots & \vdots & & \vdots \\ B_{s0} & B_{s1} & \cdots & B_{s,\nu-1} \end{pmatrix}$ is an $s \times \nu$ parameter matrix, and $\mathbf{T} = \begin{pmatrix} T_{01} & T_{02} & \cdots & T_{0t} \\ T_{11} & T_{12} & \cdots & T_{1t} \\ \vdots & \vdots & & \vdots \\ T_{\nu-1,1} & T_{\nu-1,2} & \cdots & T_{\nu-1,t} \end{pmatrix}$ is a $\nu \times t$ within-individual design matrix. The rows of **Y**, $\mathbf{y}_{hi}'$, are assumed to be independent and distributed as multivariate normal with covariance matrix **ω** (Davis, 2002).

The PR data set can be used to illustrate the above formulation. The response matrix, **Y**, would contain 27 rows and four columns, each row representing an individual's observations at ages 8, 10, 12 and 14. The corresponding design matrix, **X**, would contain 27 rows and two columns, with values 1 and 0, where a 1 in the first column indicates a girl and a 1 in the second column indicates a boy. The parameter matrix, **B**, would contain two rows and ν columns, the first row containing the parameter values for girls and the second row containing the parameter values for boys. The design matrix, **T**, would contain ν rows and four columns, with the columns representing ages 8, 10, 12 and 14, and the rows corresponding to the parameter estimates. These matrices are shown in Fig. 2.1.

$$\mathbf{Y} = \begin{pmatrix} 21 & 20 & 21.5 & 23 \\ 21 & 21.5 & 24 & 25.5 \\ 20.5 & 24 & 24.5 & 26 \\ 23.5 & 24.5 & 25 & 26.5 \\ 21.5 & 23 & 22.5 & 23.5 \\ 20 & 21 & 21 & 22.5 \\ 21.5 & 22.5 & 23 & 25 \\ 23 & 23 & 23.5 & 24 \\ 20 & 21 & 22 & 21.5 \\ 16.5 & 19 & 19 & 19.5 \\ 24.5 & 25 & 28 & 28 \\ 26 & 25 & 29 & 31 \\ 21.5 & 22.5 & 23 & 26.5 \\ 23 & 22.5 & 24 & 27.5 \\ 25.5 & 27.5 & 26.5 & 27 \\ 20 & 23.5 & 22.5 & 26 \\ 24.5 & 25.5 & 27 & 28.5 \\ 22 & 22 & 24.5 & 26.5 \\ 24 & 21.5 & 24.5 & 25.5 \\ 23 & 20.5 & 31 & 26 \\ 27.5 & 28 & 31 & 31.5 \\ 23 & 23 & 23.5 & 25 \\ 21.5 & 23.5 & 24 & 28 \\ 17 & 24.5 & 26 & 29.5 \\ 22.5 & 25.5 & 25.5 & 26 \\ 23 & 24.5 & 26 & 30 \\ 22 & 21.5 & 23.5 & 25 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} B_{10} & B_{11} & B_{12} & B_{13} \\ B_{20} & B_{21} & B_{22} & B_{23} \end{pmatrix}, \quad \mathbf{T} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 8 & 10 & 12 & 14 \\ 8^2 & 10^2 & 12^2 & 14^2 \\ 8^3 & 10^3 & 12^3 & 14^3 \end{pmatrix}$$

Fig 2.1: Growth curve analysis matrices for the PR data set.

Hypothesis tests of the form **ABC** = **D** can be tested. For example, for the PR data set, to test if there is a difference between girls and boys assuming parallelism between the growth curves of girls and boys, the hypothesis **ABC** = 0 could be tested, where **A** = (1, -1) and **C** = (1, 1, 1, 1)'. If one does not want to assume parallelism, then the test becomes **ABC** = $\mathbf{0}_4$', where **A** = (1, -1) and **C** = $\mathbf{I}_4$. Davis (2002) gives a detailed discussion on the application of growth curve analysis with respect to the PR data set.

The above formulation for growth curve analysis is not used very often in practice as software for this technique is not readily available. More flexible methods, such as linear mixed effects models, which encompass the types of comparisons available from growth curve analysis, have since been developed (Davis, 2002).

Response profile analysis is a method whereby the mean is estimated at each time point, stratified according to time, and the sequence of means over time is referred to as the mean response profile for a particular level of the group factor (Crowder & Hand, 1990; Fitzmaurice *et al*. 2004). Taking the PR data set as an example, the model tested could be of the form: $\mu = \beta_0 + \beta_1 gender + \beta_2 age + \beta_3 gender \times age$, where μ is the mean response. Fitzmaurice *et al*. (2004, p. 105) note that there are three main hypotheses that can be tested. In the context of the PR data set, the first null hypothesis to be tested would be "the mean response profiles are parallel" which would concern the *gender*×*age* interaction effect. If the response profiles are parallel, then the next hypothesis to test would be "the response profiles are flat" and this would concern the *age* effect. Also on the condition that the response profiles are parallel, the third hypothesis would be "the response profiles coincide" and this would relate to the *gender* effect. The first hypothesis, testing if the slopes of the response profiles are parallel, is generally the main interest. If the response profiles are parallel, then testing if the slope is flat is equivalent to testing if the growth rate is equal to zero, and testing if the lines coincide would mean testing if there is a *gender* effect.

Fitzmaurice *et al*. (2004, p.132) note further that the response profile method is straightforward when the design is balanced and the timing of the repeated measures is common for all subjects, and when all the covariates are discrete. It can be adapted

to accommodate unbalanced data, where there may be missing values. Since this method can accommodate arbitrary patterns in the mean response and in the covariance of the responses, it has some robustness against misspecification of the models for the mean and covariance (Fitzmaurice *et al*., 2004). Problems with this method include that the model cannot handle mistimed measurements (i.e. where the measurements taken from individuals do not occur at the same time); the response profiles produce an overall test of effects and therefore may have low power when testing for group differences; and lastly, the number of covariance parameters that need to be estimated grows rapidly as the number of measurement occasions increases (Fitzmaurice *et al*., 2004).

2.2.2 Random effects models

Vittinghoff *et al.* (2005, p. 274) note that in random effects modelling, one or more variables are declared as random factors. If a model also contains fixed factors, then the model is referred to as a mixed model. Random factors have a distribution assumed for the different levels, such as identifiers of different individuals. The values for the levels of a fixed factor are fixed, known values which are chosen at the beginning of the experiment, and the effects of each level on the response are estimated as model coefficients.

When a factor is declared to be a random factor, then inferences can be made on a statistical basis on the population from which the levels of the random factor have been chosen. Correlation can also be incorporated into the model, as observations that share the same level of the random effect are modelled as correlated. More

assumptions need to be made when using random effects, which can lead to more accurate estimates (provided the correct assumptions are made), and different estimation methods can be used (Crowder & Hand, 1990; Davis, 2002; Fitzmaurice *et al*., 2004).

Random effects modelling is one of the oldest methods used to analyze longitudinal data (Fitzmaurice *et al*., 2004). In a repeated measures ANOVA, a random effect for the individuals in the study can be included in the model. By including random effects in a model, positive correlation is induced between repeated measurements through the covariance matrix of the random effects (Fitzmaurice *et al*., 2004). In terms of the mean structure, random effects can be thought of as randomly varying intercepts which account for all unmeasured factors which make some individuals "high responders" and others "low responders" (Fitzmaurice *et al*., 2004).

The repeated measures ANOVA model can be written as:

$$y_{ij} = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + b_i + e_{ij}$$

where b_i is a random individual-specific effect and e_{ij} is a within-individual measurement of error (Crowder & Hand, 1990; Fitzmaurice *et al*., 2004).

Three standard assumptions are made when using ANOVA for repeated measures (Crowder & Hand, 1990; Twisk, 2003). Firstly, that the observations on different subjects at each of the repeated measurement times are independent, and secondly, that these observations are distributed as multivariate normal. Therefore the b_i are assumed to be normally distributed with mean zero and var(b_i) = σ_b^2 and the e_{ij} are assumed to be normally distributed with mean zero and var(e_{ij}) = σ_e^2. Thus repeated

measures ANOVA distinguishes between two different sources of variability: between subject variability (σ_b^2) and within subject variability (σ_e^2). It is also assumed that the *b*-profiles of the different individuals are uncorrelated and that the errors, e_{ij}, are uncorrelated for different time points and for different individuals. Lastly, it is assumed that all the correlations in the outcome variable between repeated measurements are equal and variances of the outcome variable are the same at each of the repeated measurements (which is known as sphericity). An example of a covariance matrix that satisfies the sphericity condition is the compound symmetric (CS) covariance matrix:

$$\begin{pmatrix} \sigma_b^2+\sigma_e^2 & \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2+\sigma_e^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma_e^2 & \cdots & \sigma_b^2 \\ \vdots & \vdots & \vdots & \ddots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2+\sigma_e^2 \end{pmatrix}$$ (Hand & Crowder,1996, p. 41).

Since the means for b_i and e_{ij} are both equal to zero, the mean response can then be written as:

$$E(y_{ij}) = \mu_{ij} = \mathbf{x}_{ij}^{'}\boldsymbol{\beta}$$ (Crowder & Hand,1990).

Repeated measures ANOVA is only a part of a more flexible and general "regression paradigm" (Fitzmaurice *et al*., 2004, p. 16). Fitzmaurice *et al*. (2004, p. 14) note that regression models have a wide range of uses. Regression models include linear regression, linear logistic regression, and Poisson or log-linear regression models. Linearity means that all of these models for the mean, or a transformation of the mean, are linear in the regression parameters. The regression parameters in the model

express how the covariates are related to the mean of the response variable. The covariates can be quantitative or categorical (such as gender or treatment group). Models which only include categorical covariates are actually ANOVA models.

2.2.3 Linear mixed effects models

The general linear mixed effects model can be written as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \quad \text{for } i = 1,...N$$
$$\mathbf{b}_i \sim \mathbf{N(0,\Sigma)}, \ \boldsymbol{\varepsilon}_i \sim \mathbf{N(0,\omega}_i\mathbf{)}$$

where $\mathbf{y}_i$, $\mathbf{X}_i$, $\mathbf{Z}_i$ and $\mathbf{b}_i$ are as defined in Chapter one, and the random errors, $\boldsymbol{\varepsilon}_i$, have a covariance matrix of arbitrary structure, $\boldsymbol{\omega}_i$. In order to make inferences on $\mathbf{y}_i$ it is assumed that, conditional on the random effect $\mathbf{b}_i$, $\mathbf{y}_i$ is normally distributed with mean vector $\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$ and with covariance matrix $\boldsymbol{\omega}_i$. If $f(y_i|\mathbf{b}_i)$ and $f(\mathbf{b}_i)$ are the corresponding density functions, then the marginal density function of $\mathbf{y}_i$ can be calculated by

$$f(\mathbf{y}_i) = \int f(\mathbf{y}_i|\mathbf{b}_i)f(\mathbf{b}_i)\, d\mathbf{b}_i$$

which can be shown to be the density function of a n_i dimensional normal distribution with mean vector $\mathbf{X}_i\boldsymbol{\beta}$ and with covariance matrix $\mathbf{V}_i = \mathbf{Z}_i\boldsymbol{\Sigma}\mathbf{Z}_i' + \boldsymbol{\omega}_i$. Since this linear mixed model is defined through $f(y_i|\mathbf{b}_i)$ and $f(\mathbf{b}_i)$, it can be referred to as the hierarchical formulation or conditional modelling approach of the linear mixed model, and assumes that both Σ and $\boldsymbol{\omega}_i$ are positive (semi-) definite (Verbeke & Molenberghs, 2000).

Under the general linear model $\mathbf{y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\boldsymbol{\Sigma}\mathbf{Z}_i' + \boldsymbol{\omega}_i)$. In practice, inference is based on the marginal distribution for the response $\mathbf{y}_i$, where the hierarchical structure

of the original model is not taken into account (Verbeke & Molenberghs, 2000). This type of analysis usually only ensures that the estimated variance of the $\mathbf{y}_i$, $\mathbf{Z}_i\boldsymbol{\Sigma}\mathbf{Z}_i' + \boldsymbol{\omega}_i$, is positive (semi-) definite, but not the positive (semi-) definiteness of the separate components, $\boldsymbol{\Sigma}$ and $\boldsymbol{\omega}_i$.

Let $\boldsymbol{\tau}$ denote the vector of all variance and covariance parameters (known as the variance components) found in $\mathbf{V}_i$ so that $\mathbf{V}_i = \mathbf{V}_i(\boldsymbol{\tau}) = \mathbf{Z}_i\boldsymbol{\Sigma}\mathbf{Z}_i' + \boldsymbol{\omega}_i$, i.e. $\boldsymbol{\tau}$ consists of the $q(q+1)/2$ different elements in $\boldsymbol{\Sigma}$ and of all parameters in $\boldsymbol{\omega}_i$, and let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\tau}')'$ denote the vector of all parameters in the marginal model for $\mathbf{y}_i$. The classical approach to inference is based on estimators obtained from maximizing the marginal likelihood function

$$L_{ML}(\boldsymbol{\theta}) = \prod_{i=1}^{N}\{(2\pi)^{-n_i/2} \mid \mathbf{V}_i(\boldsymbol{\tau}) \mid^{-\frac{1}{2}} \times \exp(-\tfrac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{V}_i^{-1}(\boldsymbol{\tau})(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\}$$

with respect to $\boldsymbol{\theta}$ (Verbeke & Molenberghs, 2000). If $\boldsymbol{\tau}$ is assumed to be known, then the maximum likelihood (ML) estimator of $\boldsymbol{\beta}$, obtained from maximising the marginal likelihood function, conditional on $\boldsymbol{\tau}$, is then given by

$$\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{W}_i\mathbf{X}_i)^{-1}\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{W}_i\mathbf{y}_i$$

and its variance-covariance matrix then equals

$$\begin{aligned}\text{var}(\hat{\boldsymbol{\beta}}) &= (\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{W}_i\mathbf{X}_i)^{-1}(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{W}_i\ \text{var}(\mathbf{y}_i)\mathbf{W}_i\mathbf{X}_i)(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{W}_i\mathbf{X}_i)^{-1} \\ &= (\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{W}_i\mathbf{X}_i)^{-1}\end{aligned}$$

where $\mathbf{W}_i$ equals $\mathbf{V}_i^{-1}(\boldsymbol{\tau})$.

In practice, linear mixed models often contain many fixed effects and in such cases, it may be important to estimate the variance components, explicitly taking into account the loss of the degrees of freedom involved in estimating the fixed effects (Verbeke & Molenberghs, 2000). This can be done using restricted maximum likelihood (REML). The REML estimators for $\boldsymbol{\tau}$ and for $\boldsymbol{\beta}$ can be found by maximizing the function known as the REML likelihood function

$$L_{REML}(\boldsymbol{\theta}) = \left| \sum_{i=1}^{N} \mathbf{X}_i' \mathbf{W}_i \mathbf{X}_i \right|^{-\frac{1}{2}} L_{ML}$$

with respect to all parameters simultaneously ($\boldsymbol{\tau}$ and $\boldsymbol{\beta}$), where N is the number of individuals (Verbeke & Molenberghs, 2000).

Jennrich and Schluchter (1986) developed a general linear modelling approach, first proposed by Liang and Zeger (1986), which extended the linear mixed effects model of Laird and Ware (1982) by incorporating it into the framework of a general linear model with an arbitrary covariance structure. They proposed the model $\mathbf{y_i} = \mathbf{X_i}\boldsymbol{\beta} + \mathbf{e}_i$ where $\mathbf{e}_i \sim N(\mathbf{0}, \mathbf{V}_i)$. In the simplest case $\mathbf{V}_i = \boldsymbol{\sigma}^2\mathbf{I}_i$. Other structures could be chosen, such as autoregressive (AR), Toeplitz (TOEP) or CS. Random effects can be included in the model by letting $\mathbf{e}_i = \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i$, where $\mathbf{b}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ unstructured (UN), and $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{I})$. This is equivalent to the Laird and Ware (1982) model. The model of Liang and Zeger (1986) is further discussed in Section 2.3.1.

2.2.4 Conditional linear mixed effects models

Conditional linear mixed effects models have been developed in an attempt to get round the problem of specifying the relationship of the random effects with time

(Verbeke & Molenberghs, 2000). In general, the main interest is in the fixed effects and the parameters of the random effects are viewed as nuisance parameters. Using this approach, parameters are estimated in two steps. Firstly, the linear mixed effects model is conditioned on sufficient statistics for the random effects parameters related to time. Then by means of maximum likelihood or restricted maximum likelihood, the remaining parameters are estimated through the conditional distribution of the $\mathbf{y}_i$, given the sufficient statistics (Verbeke & Molenberghs, 2000). The formulation of the model using this approach would be:

$$y_i = \mathbf{1}_{n_i} b_i^* + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i$$

where n_i, $\boldsymbol{\beta}$, $\mathbf{b}_i$, $\mathbf{X}_i$ and $\mathbf{Z}_i$ are as specified previously, excluding those time-dependent elements of the random effects, and b_i^* is a parameter for the time-dependent random effects.

The same parameter estimates will be obtained if the nuisance parameters, b_i^*, are included in the standard linear mixed effects model as fixed effects. This means that for each subject in an experiment a subject-specific intercept will be estimated. Where a large number of subjects have been included, this would not be computationally feasible (Verbeke & Molenberghs, 2000).

The model on which a paired t-test is based is a very simple case of a conditional linear mixed effects model (Verbeke & Molenberghs, 2000). In this case the data are perfectly balanced with two observations per subject. The only time-varying covariate of interest is the binary indicator variable representing the measurement occasion. The conditional linear mixed effects approach is equivalent to analysing the difference

between the two observations taken from each subject (Verbeke & Molenberghs, 2000).

The advantage of using conditional linear mixed effects models is that inference is available for the parameters of interest without having to specify the time dependency of the random effects. Due to the simpler form of the model during the fitting step, the numerical complexity of the fitting algorithms is reduced. This method would not be appropriate if the user were interested in the relationship of the random effects with time, as this information would be masked using this procedure (Verbeke & Molenberghs, 2000).

Conditional linear mixed effects will not be considered any further in this study because the research concerns the consequences of specifying time-dependent random effects.

2.3 Hierarchical Versus Marginal Modelling Approaches

There are two modelling approaches that incorporate correlation into a statistical model. The first is the marginal modelling approach which assumes a model which holds averaged over all the clusters (also referred to as population averaged). The coefficients can then be interpreted as the average change in the response for a unit change in the predictor over the entire population. The second is the hierarchical, or conditional, modelling approach which assumes a model specific to each cluster (also referred to as subject specific). Coefficients can then be interpreted as the change in the response in each cluster in the population for a unit change in the predictor, and

the marginal information can be obtained by averaging over all the clusters. It is important to note which type of modelling approach is being used so that the results can be properly interpreted and compared (Vittinghoff *et al*., 2005).

A summary of the explanation by Verbeke and Molenberghs (2000) on hierarchical modelling approach follows. Hierarchical modelling implies a two-stage process. During the first stage of the analysis it is assumed that the following linear regression relationship holds:

$$\mathbf{y}_i = \mathbf{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i$$

where $\mathbf{Z}_i$ ($n_i \times q$) is a matrix of known covariates, $\boldsymbol{\beta}_i$ ($q \times 1$) is a vector of unknown subject-specific regression coefficients, and $\boldsymbol{\varepsilon}_i$ is the vector of residuals of length n_i. This regression equation models how the i^{th} subject's response evolves over time.

In the second stage a multivariate regression model for the subject-specific regression coefficients, $\boldsymbol{\beta}_i$, is assumed to be of the form:

$$\boldsymbol{\beta}_i = \mathbf{K}_i\boldsymbol{\beta} + \mathbf{b}_i$$

where $\mathbf{K}_i$ is a matrix of known covariates, $\boldsymbol{\beta}$ ($p \times 1$) is a vector of unknown regression coefficients, and $\mathbf{b}_i$ is a vector of independent elements of length q. Therefore

$$\begin{aligned}\mathbf{y}_i &= \mathbf{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \\ &= \mathbf{Z}_i(\mathbf{K}_i\boldsymbol{\beta} + \mathbf{b}_i) + \boldsymbol{\varepsilon}_i \\ &= \mathbf{Z}_i\mathbf{K}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \\ &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i\end{aligned}$$

where $\mathbf{X}_i$ is the fixed effects regressor matrix.

During the first stage of this process all $\boldsymbol{\beta}_i$ estimates for the observed $\mathbf{y}_i$ for each subject are obtained separately. This can be interpreted as the calculation stage. In the

second stage the estimates $\hat{\boldsymbol{\beta}}_i$ are used to provide inferences for $\boldsymbol{\beta}$. This can be interpreted as the analysis stage.

There are some drawbacks to using this two-stage process. Information is lost in summarising the $\mathbf{y}_i$ by the estimated vector of subject-specific regression coefficients, $\hat{\boldsymbol{\beta}}_i$. Random variability is brought into the modelling process by replacing $\boldsymbol{\beta}_i$ with $\hat{\boldsymbol{\beta}}_i$ in the model. Additionally there is the problem that the covariance matrix of $\hat{\boldsymbol{\beta}}_i$ is highly dependent on the number of measurements available for each subject and also when the measurements were taken.

Marginal models are mostly used to make inferences about population means, and therefore marginal models for longitudinal data model the mean response and the within-subject association among repeated responses obtained separately (Davis, 2002; Fitzmaurice *et al*., 2004). In order to use the marginal modelling approach, it needs to be assumed that the marginal expectation ($E(y_{ij}) = \mu_{ij}$) can be related to the covariates through a known link function (g):

$$g(\mu_{ij}) = \mathbf{x}_{ij}^{'}\boldsymbol{\beta}\,.$$

Secondly, it is assumed that the conditional variance of each y_{ij}, given the covariates, depends on the mean in the following way:

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

where $v(\mu_{ij})$ is a known variance function of the mean and ϕ is a scale parameter (Davis, 2002; Fitzmaurice *et al*., 2004).

Lee and Nelder (2004) argue that the conditional modelling approach is preferable to the marginal modelling approach since both marginal inferences and conditional

inferences can be obtained from models obtained through the conditional modelling approach, i.e. both $E(\mathbf{y}_i) = \mathbf{X}_i\boldsymbol{\beta}$ and $E(\mathbf{y}_i \mid \mathbf{b}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$ can be obtained. Since the expected value for the mean of the random effects is constrained to equal zero, this means that the fixed effects estimates of a conditional model have the same meaning as those of the marginal model. The authors show that if the individuals in a study have significant random treatment effects (e.g. random time effects), these will be confounded with the fixed treatment effects in a marginal model, whereas for a conditional model these two different treatment effects will have separate estimates. The marginal estimates for the fixed effects are then only useful if there is no interaction effect between the subject and the treatment, and this can only be checked by means of a conditional model. In addition, the authors conclude that conditional models allow for the estimation of two different types of error: random error and subject-specific error, which is not possible through the marginal modelling approach.

2.3.1 Generalised estimating equations (GEE)

Generalised estimating equations (GEE) is an alternative approach to linear mixed effects models for modelling repeated measures. This approach extends generalised linear models to longitudinal data. The biggest difference between the GEE and the linear mixed effects approach is that the mean response, when using GEE, does not depend on the random effects, as it does in the context of linear mixed effects models. Therefore the GEE approach is a marginal modelling approach. Marginal models do not require that a distribution is specified for the observations, only that a regression model is specified for the mean response, and their primary purpose is to make inferences about population means. This means that the regression parameters

estimates obtained will have population-averaged interpretations (Fitzmaurice *et al.*, 2004).

Liang and Zeger (1986) were the first to propose the extension of the generalised linear models to form GEE. They proposed using a "working correlation matrix", **R(α)**, defined so that:

$$\mathbf{V}_i = \mathbf{A}_i^{\frac{1}{2}}\mathbf{R}(\boldsymbol{\alpha})\mathbf{A}_i^{\frac{1}{2}} / \phi$$

will equal cov($\mathbf{y}_i$) if **R(α),** where **α** is a vector which fully characterises **R(α)** and $\mathbf{A}_i$ = diag{ϕvar($\mathbf{y}_i$))**,** is the true correlation matrix of the $\mathbf{y}_i$'s. The GEE are then defined as:

$$\sum_{i=1}^{N} \mathbf{d}_i' \mathbf{V}_i \mathbf{s}_i = 0$$

where $\mathbf{d}_i = \frac{\partial E(\mathbf{y}_i)}{\partial \boldsymbol{\beta}}$ and $\mathbf{s}_i = \mathbf{y}_i - E(\mathbf{y}_i)$ (Liang & Zeger, 1986).

The covariance structure is treated as a nuisance in the GEE methodology, which rather focuses on the regression of **y** on **X**. In this way the estimates obtained for the regression coefficients **β** are consistent and asymptotically normal, even when the covariance structure is misspecified. But a working correlation structure still needs to be specified (Hedeker & Gibbons, 2006).

The simplest of these structures is that of independence where it is assumed **R(α)** = **I**. This is the same as assuming that the longitudinal observations are independent, but this is generally inappropriate and can lead to large efficiency loss for time-varying covariates. The exchangeable structure is another simple form that can be specified,

and assumes that $\mathbf{R}(\boldsymbol{\alpha}) = \rho$, so that the correlations between the longitudinal data are all the same. This is equivalent to assuming a CS covariance structure for the linear mixed effects model. More advanced structures, such as an AR lag 1 (AR(1)) structure, where it is assumed $\mathbf{R}(\boldsymbol{\alpha}) = \rho^{|j-j'|}$, or a TOEP structure, where it is assumed $\mathbf{R}(\boldsymbol{\alpha}) = \rho_{|j-j'|}$ when $j - j' \leq m$ and $\mathbf{R}(\boldsymbol{\alpha}) = 0$ otherwise, or UN, where $\mathbf{R}(\boldsymbol{\alpha})$ will have $n(n - 1)/2$ correlations to estimate, can be specified as well (Hedeker & Gibbons, 2006).

A criticism of GEE methods is that they may not always correspond to a completely specified sampling model for the data, meaning that the assumptions being made by using the GEE method may not always be apparent. Inference using this method is entirely dependent on asymptotics (Weiss, 2005).

2.3.2 Robust inference

Going hand-in-hand with GEE methodology is estimation using robust standard errors. Robust standard errors make use of a temporary or working assumption as to the correlation structure in order to form the estimates, and these are then adjusted for the correlation in the data. Once the model coefficients have been estimated using the temporary correlation structure, within-subject residuals are used to compute robust standard errors for the coefficient estimates. Since these standard errors are based on the data, by use of the residuals, and not on the assumed working correlation structure, they give more robust inferences for large sized samples as long as the other specifications of the model (distribution, link and form of predictors) are correct, regardless of whether the working correlation structure is correct or not. It can be advantageous to avoid the calculation of a large number of correlations, and in cases

where both using an UN correlation structure and using robust standard errors can be used, the two methods produce similar results (Vittinghoff *et al*., 2005).

The covariance matrix is estimated by means of the "sandwich estimator", which is used to increase the efficiency of the covariance parameter estimates (Verbeke & Molenberghs, 2000; Crowder, 2001). This estimator is obtained by replacing the term var($\mathbf{y}_i$) in the variance estimator of $\boldsymbol{\beta}$ with $\frac{\mathbf{r}_i\mathbf{r}_i'}{\nu}$, where $\mathbf{r}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}$ and ν is the residual degrees of freedom (Verbeke & Molenberghs, 2000). Although the parameters of the covariance structure are treated as nuisance parameters, a covariance structure still needs to be specified. The estimated standard errors will be poor in comparison to those obtained if the correct covariance structure had been specified, therefore it is advantageous to model the covariance structure correctly (Crowder, 2001). Specifying the correct covariance structure will also lead to more efficient estimates of the standard errors, and in the case of missing data, valid estimates will only be obtained via the sandwich estimator if the "missing-ness" of the data follows very strict assumptions (Verbeke and Molenberghs, 2000).

Since the interest of this study is specifically related to the assumption of the covariance model and its consequence for the estimates and inference about the fixed effects, robust standard errors will not be considered further in this study, although it is a reasonable approach for any investigator only interested in the mean response. The model of interest in this study is the linear mixed effects model, so the GEE approach will also not be considered any further, but a similar study on the performance of the GEE model under difference covariance assumptions could be considered.

2.3.3 Mixed models as a compromise between Bayesian and frequentist approaches

The view of mixed models as a combination of frequentist and Bayesian approaches is based on the hierarchical formulation of the model (Verbeke, 1997, Demidenko, 2004). If the Bayesian approach were used on its own then values for the parameters would need to be specified, whereas for the mixed models approach these parameters are estimated from the data (Demidenko, 2004).

The mixed model approach lends itself to a Bayesian interpretation because the random effects are assumed to be random variables. The response variables, $\mathbf{y}_i$, can be written conditional on the random effects, $\mathbf{b}_i$, and the prior density function of $\mathbf{b}_i$ can be written as $f(\mathbf{y}_i|\,\mathbf{b}_i)$ and $f(\mathbf{b}_i)$ respectively. In the Bayesian framework

$$A = \int f(\mathbf{y}_i \mid \mathbf{b}_i) f(\mathbf{b}_i)\, d\mathbf{b}_i$$

is referred to as the normalising constant (Demidenko, 2004). By means of Bayes Theorem, the posterior density function can be written as:

$$f(\mathbf{b}_i \mid \mathbf{y}_i) \equiv f(\mathbf{b}_i \mid \mathbf{Y}_i = \mathbf{y}_i) = \frac{f(\mathbf{y}_i \mid \mathbf{b}_i) f(\mathbf{b}_i)}{\int f(\mathbf{y}_i \mid \mathbf{b}_i) f(\mathbf{b}_i)\, d\mathbf{b}_i}$$

Through the theory of Bayesian linear models, it can be shown that this posterior density is distributed multivariate normal. The $\mathbf{b}_i$ can then be estimated by the posterior mean as:

$$\begin{aligned} \hat{\mathbf{b}}_i(\theta) &= E(\mathbf{b}_i \mid \mathbf{Y}_i = \mathbf{y}_i) \\ &= \int \mathbf{b}_i f(\mathbf{b}_i \mid \mathbf{y}_i)\, d\mathbf{b}_i \\ &= \boldsymbol{\Sigma}\mathbf{Z}_i' \mathbf{W}_i(\boldsymbol{\theta})(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \end{aligned}$$

and with estimated covariance

$$\text{var}(\hat{\mathbf{b}}_i) = \boldsymbol{\Sigma}\mathbf{Z}_i'\left\{\mathbf{W}_i - \mathbf{W}_i\mathbf{X}_i\left(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{W}_i\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\mathbf{W}_i\right\}\mathbf{Z}_i\boldsymbol{\Sigma}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the random effects, $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\tau}')'$ and $\boldsymbol{\tau}$ is the vector of the parameters of the covariance matrix $\mathbf{V}_i = \mathbf{W}_i^{-1}$ (Verbeke, 1997). This would underestimate the variability in $\hat{\mathbf{b}}_i - \mathbf{b}_i$ as the variation is $\mathbf{b}_i$ is ignored, so $\text{var}(\hat{\mathbf{b}}_i - \mathbf{b}_i) = \boldsymbol{\Sigma} - \text{var}(\hat{\mathbf{b}}_i)$ would be used to assess this variation (Laird & Ware, 1982; Verbeke, 1997).

The unknown parameters $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ would be estimated by means of maximum or restricted maximum likelihood estimation of the marginal distribution:

$$f(\mathbf{y}_i) = \int f(\mathbf{y}_i \mid \boldsymbol{\beta}, \mathbf{b}_i) f(\mathbf{b}_i)\, d\mathbf{b}_i$$

Therefore the normalising constant plays the role of the likelihood in the mixed model (Demidenko, 2004).

2.4 Model Structure

A substantial amount of literature exists on the analysis of repeated measures. Some of these texts include Crowder and Hand (1990), Verbeke & Molenberghs (2000), Demidenko (2004), Fitzmaurice *et al*. (2004), and Weiss (2005). The rest of this chapter discusses some of the issues that need to be considered when addressing any repeated measures problem, and also more specific considerations related to the estimation of the linear mixed effects model and its covariance matrix.

Before a repeated measures model can be fitted to longitudinal data, certain considerations need to be taken into account, including the choice of covariance structure, the role of time in the model and whether to include predictor variables as fixed or random. The choices for these model considerations will be discussed in this section, together with the potential consequences of misspecifying these components of the model.

There are a number of reasons for modelling the covariance matrix of a regression model. Firstly, good estimates of the covariance matrix result in more efficient, and more precise, fixed effects estimates and more accurate confidence intervals and hypothesis tests (Fitzmaurice *et al*., 2004; Weiss, 2005). An appropriate choice for the covariance structure leads to correct standard errors and valid inferences about the regression parameters. The positive correlation found in repeated measures reduces the variability of the estimate of change over time within individuals; therefore this positive correlation can be looked at as an advantage to longitudinal study designs. Secondly, if predictions of future values, or imputation of missing values, are required from the model, then good covariance estimates are needed (Weiss, 2005). In fact, when there are missing data, the correct modelling of the covariance matrix is a requirement to obtain valid estimates of the regression parameters. Lastly, modelling the covariance matrix is part of modelling the science behind the data. For example, if the CD4 counts of HIV positive patients are modelled to fall about a subject-specific straight line with a negative slope, this implies that the eventual end is "fore-ordained" (Weiss, 2005). If the data is modelled so that the CD4 counts follow an AR(1) model, this implies that the data follows a random walk pattern, which could

potentially go up or down. Therefore, failing to correctly model the covariance matrix can lead to misleading scientific inferences (Fitzmaurice *et al*., 2004; Weiss 2005).

Verbeke & Molenberghs (2000, p.121) state that "An appropriate covariance model is essential to obtain valid inferences for the parameters in the mean structure". If the specifications of the covariance matrix are too restrictive then inferences may be invalid if the specifications don't hold. If the model is overparameterised, then this could lead to inefficient estimation and poor assessment of standard errors. Therefore, to make reliable predictions, efficient estimates of an appropriate covariance matrix are required.

2.4.1 Covariance structures

Covariance structures need to be chosen for both the random errors and the random effects. Selection of the covariance structure of the error components needs to be conditional on the selected structure for the covariance of the random effects, as together these will describe all the model variance. If the random effects are chosen correctly, one should be able to assume that the random effects account for most of the variability in the data and therefore simple, parsimonious models for both the error and random effects covariance structures can be chosen (Verbeke & Molenberghs, 2000). Fitting overly complicated covariance structures can lead to non-convergence (Verbeke & Molenberghs, 2000). SAS® (ver. 9.1) PROC MIXED has many available covariance structures to choose from, which can be applied to both the error and random effects covariance structures.

The simplest covariance structure is referred to by SAS (ver. 9.1) as the variance components (VC) covariance structure. This structure assumes that there is no correlation between observations and that there is constant variance, σ^2, across all measurement occasions. Only one parameter needs to be estimated for this covariance structure. In this case where no random effects are included in the model, this would be the assumed covariance structure for an OLS model.

When the correlations between all observations from the same subject can be assumed to be the same, the covariance structure is referred to as exchangeable or CS (Vittinghoff *et al.*, 2005). This type of covariance structure is suitable when there are no variables distinguishing one member of a level of a random effect from another, so can be used in the absence of any other data structure, but can be restrictive and unrealistic at times (Davis, 2002; Vittinghoff *et al.*, 2005). The form of this covariance structure was described in Section 2.2.2. Two parameters need to be estimated for this structure.

When measurements are taken through time, observations taken more closely together in time are likely to be more highly correlated, and in this case it may be more appropriate to use an AR structure, which exhibits this feature (Crowder & Hand, 1990). The most commonly used AR structure is an AR(1), which assumes that the variance across all occasions is the same and the correlation between two points one unit apart would be ρ, two units apart would be ρ^2, three units apart would be ρ^3, etc., and therefore the correlation value tends to zero as the observations get further and further apart (Fitzmaurice *et al.*, 2004; Vittinghoff *et al.*, 2005). Therefore it would not be appropriate to use this structure when the values for an individual are stable

over time, e.g. systolic blood pressure taken from a patient undergoing a single treatment regime over a long period of time (Vittinghoff *et al.*, 2005). The form of this covariance structure is $o^2\begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$ where ρ is the correlation coefficient, with $|\rho| < 1$ for stationarity, and σ^2 is the variance across all occasions. Two parameters need to be estimated for this covariance structure.

Another covariance structure related to time is the TOEP structure. This structure can be viewed as a moving-average structure of the same order as the dimensions of the covariance matrix. The TOEP structure assumes that any pair of responses that are separated by the same length of time have the same correlation and that the variance across all occasions is the same (Fitzmaurice *et al*., 2004). The form of this structure is

$\begin{pmatrix} \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 \end{pmatrix}$ where σ^2 is the variance across all occasions and σ_i for i = 1…(number of matrix dimensions – 1) is the variance for observations i units apart. For this structure t parameters need to be estimated, where t is the number of measurement occasions. It should be noted that this structure assumes that the correlation among responses at adjacent measurement occasions is constant, and therefore this structure is only appropriate when the measurement occasions are separated by equal units of time (Fitzmaurice *et al*., 2004).

In addition to these covariance structures, SAS (ver. 9.1) allows one to specify heterogeneous variants of these covariance structures. In particular, SAS allows you

to specify the heterogeneous CS structure (CSH), which allocates a different variance parameter for each diagonal element, and uses the square roots of these parameters in the off-diagonal entries. Therefore the form of this covariance structure can be displayed as follows:

$$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho & \sigma_1\sigma_4\rho \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho \\ \sigma_3\sigma_1\rho & \sigma_3\sigma_2\rho & \sigma_3^2 & \sigma_3\sigma_4\rho \\ \sigma_4\sigma_1\rho & \sigma_4\sigma_2\rho & \sigma_4\sigma_3\rho & \sigma_4^2 \end{pmatrix}$$ where σ_i^2 is the i^{th} variance component, for i = 1…t and ρ is the correlation parameter. For this structure $t + 1$ parameters need to be estimated.

Another heterogeneous structure is the AR structure (ARH), which is similar to the CSH structure, but the correlation parameter behaves as for the AR(1) structure. The form for this covariance structure is

$$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 & \sigma_1\sigma_4\rho^3 \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho^2 \\ \sigma_3\sigma_1\rho^2 & \sigma_3\sigma_2\rho & \sigma_3^2 & \sigma_3\sigma_4\rho \\ \sigma_4\sigma_1\rho^3 & \sigma_4\sigma_2\rho^2 & \sigma_4\sigma_3\rho & \sigma_4^2 \end{pmatrix}$$ where the parameters are as before. This model also requires that $t + 1$ parameters are estimated.

When the covariance structure is UN, so that the covariance between any two observations from the same subject could be estimated to have a different value, then $t(t+1)/2$ variance parameters will need to be estimated (Fitzmaurice *et al*., 2004). The loss of degrees of freedom due to the large number of estimated variance components would cause a decrease in the precision of the estimated parameters of interest or even a failure in fitting the model (Vittinghoff *et al*., 2005). Choosing the correct correlation structure can be very difficult in practice (Vittinghoff *et al*., 2005).

2.4.2 The effect of time

How time is included in the fixed effects part of a model is an important consideration when modelling longitudinal data. To summarise the discussion by Singer and Willett (2003): the first consideration with regards to time should be the metric of time used. The metric chosen should reflect the cadence that is expected to be the most useful of the outcome of interest. For example, when the outcome of interest is the wear of a tyre, the metric used for time should rather be the number of kilometres travelled than the age of the car, as this is more likely to have an influence on the wear of the tyre. Whether the individuals were measured with equal or unequal time spacing, whether the individuals were measured at different times and whether individuals had the same number of measurement occasions should also be included in the model. These considerations can affect parameter definition, model construction, estimation and model testing.

When time is not included as a predictor in the model, the model assumes a “no change” trajectory, i.e. that there in no change over time in the response. If a “linear change” model is assumed, time is included as a first-order polynomial, where each individual will then have a unique intercept and slope. When time is added as a second-order polynomial, it assumes a quadratic change over time. In this case, the coefficient of the first-order time term represents the instantaneous rate of change at any one specific moment. The coefficient of the second-order time term is the curvature parameter, which describes the changing rate of change. Higher order polynomials can be added, increasing the complexity of the trajectory. The greater the complexity of the polynomial, the more measurement occasions are required to fit the

model. The degree of the polynomial is limited by the number of observations taken per subject.

If there is more than one polynomial of time that seems suitable, such as when either a 2^{nd} or 4^{th} degree polynomial may be appropriate, an exploratory approach should be adopted to choose the correct polynomial, where individual-specific ordinary least squares (OLS) models are fitted to each individual's data (Singer & Willett, 2003). The highest order polynomial required to adequately summarise an individual's change should be chosen. Goodness-of-fit statistics can also be used to choose between models containing different polynomials in time. The multiple testing aspect of the data, and the effect of data inspection to choose the model, must be kept in mind.

2.4.3 Random effects

The random effects in a linear mixed effects model are generally restricted to only random intercepts and random coefficients for time-varying covariates (Verbeke, 1997). Only those covariates that have been included in the fixed effects part of the model (i.e. in $\mathbf{X}_i$) or covariates that form linear combinations of the columns of $\mathbf{X}_i$ will be considered as random effects covariates, as it is assumed that the random effects $\mathbf{b}_i$ have a mean of zero (Verbeke, 1997). It has been shown that, as for linear regression, polynomials of the time effect should not be included unless all the hierarchical inferior terms have also been included (Verbeke, 1997). By including random effects into the model, it implies that the covariance matrix of $\mathbf{y}_i$ is of the form $\mathbf{V}_i = \mathbf{Z}_i \boldsymbol{\Sigma} \mathbf{Z}_i' + \boldsymbol{\omega}_i$. Therefore the choice of the random effects covariates directly

impacts on the proposed covariance structure of the responses. If the diagonal elements of $\boldsymbol{\omega}_i$ are all equal, then the covariance matrix of the responses will only depend on time through the term $\mathbf{Z}_i\boldsymbol{\Sigma}\mathbf{Z}_i'$. It is then possible to perform an informal check on the appropriateness of the random effects by comparing the fitted variances to the residuals r_{ij} after grouping them into time intervals (Verbeke, 1997).

2.4.4 Fixed effect or random effect?

The decision whether to treat a factor as fixed or random can have important consequences for the manner in which the factor is treated in the model, and hence important consequences for the conclusions that can be drawn from the model. Duchateau and Janssen (1997) describe this by means of a simple example in the context of an ANOVA. In their example, eight batches of a particular antibiotic are chosen and the efficiency of these drugs is measured after a two-year period. If it can be considered that the batches were chosen at random from a population of available batches, then batch can be included in the model as a random effect. This model can be represented by the following equation:

$$y_{ij} = \mu + \mathrm{b}_i + \varepsilon_{ij}$$

where y_{ij} is the response for the j^{th} observation in the i^{th} batch, μ is the overall mean (a fixed parameter), b_i is the random effect associated with batch i, and ε_{ij} is the random error of the j^{th} observation in the i^{th} batch. It is then assumed that b_i and ε_{ij} are independently and identically normally distributed with zero mean and variance σ_b^2 and σ^2 respectively. Therefore it follows that $E[y_{ij}] = \mu$ and $Var[y_{ij}] = \sigma_b^2 + \sigma^2$.

If it is assumed that the eight batches chosen are the only batches available, then batch can be included as a fixed effect. The model will then be

$$y_{ij} = \mu + \beta_i + \varepsilon_{ij}$$

where y_{ij}, μ and ε_{ij} are as for the previous model, and β_i is the fixed effect parameter for the i^{th} batch. It then follows that $E[y_{ij}] = \mu + \beta_i$ and $Var[y_{ij}] = \sigma^2$. This is quite different from the previous model. Each batch now has its own mean value and the only source of variability is from the random errors.

There is a problem with the fixed effects parameterisation of the model. The model can be written in matrix notation as:

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\beta} = (\mu, \beta_1, ..., \beta_8)'$. The resulting design matrix, **X**, will then be such that the first column could be obtained by summing all the other columns. Therefore **X** is not of full rank, and this is due to the overparameterisation of the model (Duchateau & Janssen, 1997). This problem can be dealt with by treating one group as a base group and therefore reducing the number of regression coefficients by one, or by restricting the parameter values to sum to zero.

2.4.5 The effect of model misspecification

Misspecification can occur in various aspects of the linear mixed effects model. Some of these misspecification problems will be discussed in this section.

As discussed in Section 2.4.2, time is an important consideration in modelling longitudinal data. Misspecifying the metric of time, whether individuals are measured

at the same time, whether individuals have the same number of measurement occasions, or the spacing of measurements could result in incorrect parameter definition, incorrect model construction, incorrect estimation and incorrect model testing (Singer & Willett, 2003). How time is included in the model can also be misspecified. If a polynomial of time is chosen which is too simple, the true complexity of an individual's trajectory over time may not be fully explained by the model. If a polynomial of time is chosen which is too complex, then there may not be enough measurement occasions to fit the model (Singer & Willett, 2003).

Most papers addressing misspecification of linear mixed effects models refer to the distribution of the random effects or the random error structure. Verbeke and Lesaffre (1997) showed that the maximum likelihood estimators for fixed effects and variance components in linear mixed models are consistent and asymptotically normally distributed when obtained under the assumption of normally distributed random effects, even when the distribution of the random effects is not normal. Zewotir and Galpin (2004) studied the formal and informal assessments of the normality assumptions in linear mixed models and showed that probability plots of the residuals and tests based on these plots were not sensitive to non-normality of the random effects, but only to the non-normality of the error terms. They further showed that the shape of the probability plots could indicate specific transformations of the data that could improve the normality of the data.

Jacqmin-Gadda, Sibillot, Proust, Molina and Thiebaut (2007) showed that inference for the fixed effects under the assumption of independent normally distributed errors with constant variance is robust for errors which are not normally distributed or which

are heteroscedastic, but that estimates of the fixed effects are biased if the error variance depends on a covariate with interaction with time and if the errors are correlated. In the first case, they noted that although the inference for the estimated slopes may be biased, the inference for the difference between slopes is robust, and therefore the test for treatment effect is robust. If the errors terms are correlated, then using a model that includes both a random intercept and a random slope is more robust that using a model with only a random intercept.

Lange and Laird (1989) studied the effect of misspecification of the number of random effects in a linear growth curve models. By misspecification of the random effects, they refer to misspecification of the maximum polynomial in time to include in the random effects. They showed that the variance of the estimators depended strongly on the assumed number of random effects in the model, but that the random intercept and slope model gave conservative estimates of the variance, even when the true number of random effects in the model was more than two. Taylor and Law (1998) showed by means of a simulation study that misspecification of the model covariance structure, where they considered four models with closed form solutions for the model parameters, leads to reduced coverage probabilities for estimates of individuals' future observations. The models they considered include a linear mixed effects model with an added integrated Ornstein-Uhlenbeck stochastic process, a linear mixed effects model with an added Brownian Motion process, a random intercept and slope model, and a quadratic random effects model. They assumed that the error term had independent errors. They showed that of these four models, if the quadratic random effects model was mistakenly specified, it obtained the best coverage probabilities, but was the least efficient.

Both Richardson and Welsh (1995) and Copt and Victoria-Feser (2006) have studied linear mixed effects models in the presence of outliers and have found that if contaminated errors were inserted in the data with a null mean, the estimates for the fixed effects in linear mixed effects are comparatively robust, whereas the estimates for the variance parameters were biased. If the mean of the contaminated error distribution was different from zero, then the fixed effects estimates were biased as well. These authors propose more robust methods compared to ML and REML methods for estimating the parameters. Jacqmin-Gadda *et al*. (2007) state that their analyses on linear mixed effects models show the same pattern in bias of parameter estimates when outliers are included.

Ugrinowitsch *et al*. (2004) compared the performance of the repeated measures ordinary least squares (OLS) method, the generalized least squares method assuming a compound symmetric covariance structure and an AR(1) covariance structure, and the random coefficients growth curve model. They found that if a compound symmetrical covariance structure was incorrectly assumed for either the GLS or OLS method, then high probabilities of Type 1 error would occur, whereas the random coefficients growth curve model performed well under any covariance structure. Demidenko (2004) discuss the case when random effects are ignored, and instead OLS estimators are used. He showed that the OLS estimator of the variance was positively biased when the random effects were ignored.

In summary, these studies conclude that the estimates for the fixed effects are unbiased to model misspecification or to outliers, but that the model covariance structure is not. And therefore inferences about the fixed effects may not be accurate

if there is model misspecification, such as if the covariance structure for either the random effects or random errors is misspecified.

The purpose of this study was to investigate the robustness of linear mixed effects models when the covariance structure is misspecified. These covariance structures would then be the best option for a researcher to examine if the researcher were uncertain about the correct covariance structure.

2.5 Parameter Estimation

There have been a number of methods proposed in the literature to obtain the parameter estimates of a repeated measures model. Laird and Ware (1982) described a non-Bayes Expectation-Maximisation (EM) algorithm approach to obtaining the ML estimates of the linear mixed effects model. They expanded this method to one that used a combination of ML estimation and an empirical Bayes approach to obtaining REML estimates for the variance parameters.

Jennrich and Schluchter (1986) developed Newton-Raphson (NR) and Fisher scoring algorithms to estimate the parameters of the general linear model for longitudinal data. They also described a hybrid EM scoring algorithm to obtain the REML estimates. Lindstrom and Bates (1988) further developed the NR algorithm to be an efficient and effective means of estimating the parameters of the mixed effects model. They proposed improvements to the algorithm of Jennrich and Schluchter (1986) to improve the speed of convergence and to ensure the covariance matrix for the random effects is positive definite at each iteration, and derived all the necessary derivatives

and second derivatives required to implement the algorithm. They also compared the implementation of the NR algorithm to that of the EM algorithm of Laird and Ware (1982), and found that the algorithms of these two methods had similar computing times, but that the EM algorithm generally needed more iterations before convergence compared to the NR algorithm. In terms of convergence, Lindstrom and Bates (1988) concluded that the EM algorithm was guaranteed to converge to a local maximum, even if many iterations are required, but that the NR algorithm, if implemented as suggested by these authors, would converge very consistently, and have the advantage of producing the Hessian matrix for the parameter vector, and has an objective convergence criterion available, unlike for the EM algorithm.

Wolfinger (1993) describes a unified framework for likelihood-based approaches for parameter estimation, using both ML and REML estimates. Wolfinger, Tobias and Sall (1994) also develop NR algorithms for the estimation of the ML and REML estimates, including estimation of arbitrary covariance structures for both the errors and the random effects, and provide the derivation of the derivatives and second derivatives required for the implementation of this method.

Verbyla (1990) derived the REML equations of the mixed effects model by partitioning the full likelihood into two independent parts, one relating to the fixed effects contrasts and the other to the residual contrasts. Maximisation of the first set of contrasts leads to estimates of the fixed effects and maximisation of the second set of contrasts leads to the REML estimates of variance parameters (Verbyla & Cullis, 1990; Cullis, Smith & Thompson, 2004). Pourahmadi (2000) showed that the loglikelihood of the general linear model has three representations which correspond

to the submodels for the means, variances and correlations. Pourahmadi (2000) derived closed form solutions for the fixed effects and for the correlation parameters, and developed a NR algorithm for the variance parameter estimates. This method was developed to ensure the positive definiteness of the covariance matrix.

In the past, the EM algorithm was a popular method of obtaining the ML or REML estimates for the linear mixed effects model, although Laird and Ware (1982) have shown that the algorithm is slow to converge for estimates of the covariance when the maximum likelihood is close to the boundary space of the parameters (Verbeke & Molenberghs, 2000). Modern software for linear mixed effects estimation, such as proc mixed of SAS (ver. 9.1), use NR based procedures (Verbeke & Molenberghs, 2000). In general, REML estimates are obtained rather than ML estimates, as ML estimators for the covariance parameters tend to be biased downwards, as they do not take into consideration the loss of degrees of freedom from the estimation of the fixed effects (Lindstrom & Bates, 1988; Verbeke & Molenberghs, 2000).

SAS (ver 9.1) PROC MIXED, by default, uses a ridge-stabilised NR algorithm to minimise -2×loglikelihood for the ML approach, and -2×restricted loglikelihood for the REML approach. This procedure does not optimise the likelihoods directly, but rather optimises the profile likelihoods, which have one less parameter, and therefore are optimised more efficiently. These algorithms are based on work by Wolfinger, Tobias and Sall (1994) who developed algorithms for computing the Gaussian likelihood and restricted likelihood for the general linear mixed model through NR. The authors make use of Cholesky decomposition, the sweep operator, and the *W*-transformation in these algorithms, and also discuss the use of profile likelihood to

obtain the variance parameter estimates. The minimum variance quadratic unbiased estimator (MIVQUE) method is implemented in the procedure in order to obtain the starting values. (Littell, Milliken, Stroup & Wolfinger, 1996). A detailed explanation of this procedure can be found in Wolfinger *et al*. (1994). Swallow and Searle (1978) explain how to obtain the MIVQUE estimates for variance components.

2.5.1 Problems with parameter estimation

In practice, estimates for the linear mixed model are obtained from the less restrictive marginal model, rather than from the hierarchical model (Verbeke & Molenberghs, 2000). Since the marginal model does not imply the hierarchical model, this could result in estimates of the parameters of the hierarchical model not converging due to negative variance component estimates (Verbeke & Molenberghs, 2000). Therefore it is important to run exploratory data analyses prior to fitting the linear mixed model in order to ensure that valid estimates are obtained for the model (Verbeke & Molenberghs, 2000, p. 54).

Problems, which are not uncommon, can occur during the estimation of parameter estimates. As discussed earlier, the NR algorithm, which is most commonly implemented, is not guaranteed of converging (Lindstrom & Bates 1988). If non-convergence occurs, one of the easiest ways of solving this problem is to specify better starting values, or even to change the numerical optimisation procedure (Verbeke & Molenberghs, 2000). Demidenko (2004) suggests using methods such as minimum norm quadratic unbiased estimation (MINQUE), method of moments or variance least squares to obtain good starting values for the covariance matrix. In this

study, for the purpose of comparison, the optimisation method was not changed for any of the models. If a non-convergence message is reported, it is very important to check the parameter estimates to ensure that they are reasonable and within bounds (Weiss, 2005). For example, the standard deviation is expected to be smaller than one quarter of the range of the data, and therefore this rule of thumb can be checked against the estimates of the variance parameters (Weiss, 2005). Weiss (2005) also notes that if estimates are obtained without standard errors, this is usually an indication that the algorithm has not converged.

Non-convergence can sometimes result when the estimates of the variance components tend too closely to zero. By rescaling the time variable, for example to decades instead of years or months, it can be possible to solve this problem by artificially enlarging the variance components (Verbeke & Molenberghs, 2000).

If it is reported that the "Hessian matrix is non-positive definite", this means that a saddle point has been reached (Weiss, 2005). It should be expected that the Hessian matrix becomes positive definite in the neighbourhood of the maximum if the NR algorithm is implemented, but if the iterations are far from the maximum, then the Hessian matrix may not become positive definite, and NR algorithm may fail (Demidenko, 2004). The more parameters in the model, the more likely it is that problems will be encountered during the maximisation procedure (Weiss, 2005). Weiss (2005) recommends that if this occurs, that parameters should be removed from the model until the problem is solved, and then model diagnostics should be used to assess how well this model performs. If parameter estimates are obtained that go out

of their bounds, Crowder and Hand (1990) recommend parametrising the various quantities in such a way that any constraints are satisfied automatically.

2.6 Assessing the Fit of a Model

Information criteria can be used to assess the fit of repeated measures models, and outlier and influence diagnostics and residual analyses can be used to assess model misfit or appropriateness, but the OLS forms of these analyses cannot always be used. These criteria and adjustments required for repeated measures models will be discussed in this section. The choice of model will be dependent on the criteria chosen for assessing model adequacy.

2.6.1 Information criteria

For selection between different models, the selection criteria Akaike's information criterion (AIC) and Schwarz's Bayesian information criterion (BIC) can be used (Verbeke & Molenberghs, 2000; Davis, 2002; Fitzmaurice *et al.*, 2004). The AIC is defined as:

AIC = -2(maximised log-likelihood) + 2(number of parameters)

= $-2\hat{l} + 2c$

and the BIC is defined as:

BIC = -2(maximised log-likelihood) + log$N$ (number of parameters)

= $-2\hat{l} + \log N \times c$

where N is the number of subjects and c is the number of parameters (Fitzmaurice *et al.*, 2004). In order to select between models, these two selection criteria need to be

minimised. Duong (1984) states that models that are within two units of the lowest AIC can be considered as competitive for the best model. Burnham and Anderson (2002) and Jones (1993) are just two of a large number of texts which concur with this criterion. Models can then be selected according to those that have the fewest parameters, and this will be achieved by comparing the BIC of these models, as it penalises the number of parameters more strictly (Fitzmaurice *et al.*, 2004).

2.6.2 Outlier and influence diagnostics

Zewotir and Galpin (2005) derived influence diagnostics for linear mixed models, where the ordinary linear regression influence diagnostics have been extended to linear mixed models. The statistics which have been extended included Cook's distance (Cook, 1977), the likelihood distance (Cook & Weisberg, 1982), the variance (information) ratio (Belsley, Kuh & Welsch, 1980), the Cook-Weisberg statistic (Cook & Weisberg, 1980) and the Andrews-Pregibon statistic (Andrews & Pregibon, 1978). Zewotir and Galpin (2005) show that a one step form of the diagnostics, which is computationally inexpensive as opposed to the full iteration, adequately provide informoration on the influence of the data on various aspects of model fit. These statistics were tested using clustered, but not longitudinal, data. The sensitivity of these statistics have been studied by Zewotir and Galpin (2006) and they found that these statistics are capable of detecting influential points in **X**, **Z** and **y**, but masking effects could occur under certain circumstances, for example when there are multiple outliers in the same observation. Demidenko and Stukel (2005) have also proposed extensions to the leverage, infinitesimal influence, case deletion diagnostics, Cook's

distance, and local influence used for regression models to accommodate linear mixed effects models, which are in the form of explicitly defined functions.

In SAS (ver. 9.1), experimental code has been included for outlier and influence diagnostics. The basis on which these diagnostics are obtained is through computing parameter estimates based on all data points, removing the cases in question from the data, refitting the model, and computing statistics based on the change between full-data and reduced-data estimation. Refitting the model to the reduced data set involves going through the iterative maximisation procedure in order to obtain the parameter estimates if covariance parameters are not known. The computing time when this option is included does increase to about one minute for the data sets used in this study, but this is an acceptable amount if time in these circumstances. For very large data sets the computing time may pose problems. The diagnostics that are included are: the restricted likelihood distance, which measures the overall influence of an observation (Cook & Weisberg, 1982); Cook's distance, which measures the influence of an observation on all predicted values (Cook, 1977); the covariance ratio and trace, which measures the influence of an observation on the precision of the estimates (Belsley *et al.*, 1980); the PRESS residuals (Allen, 1974) and the DFFITs, which measure the influence of an observation on its own predicted value; and MDFFITS, which measures the influence of an observation on the parameter estimates (Belsley *et al.*, 1980). These influence and outlier diagnostics, as available in SAS (ver. 9.1), are described by Schabenberger (2004).

2.6.3 Analysis of residuals

The residuals can be used to assess the fit of a model. Firstly, the type of residuals to be used needs to be chosen. For example, the marginal residuals $\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}$ can be used, and these reflect how a specific individual's mean profile deviates from that of the population. If the conditional residuals, $\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} - \mathbf{Z}_i\hat{\mathbf{b}}_i$, are used this reflects how much the observed values deviate from an individual's own predicted values. Even the random effects $\hat{\mathbf{b}}_i$ can be considered as residuals as these values reflect how an individual's profile deviates from that of the population (Verbeke & Molenberghs, 2000).

Haslett & Haslett (2007) give a comprehensive review of the three different types of residuals available, describing where they are used, how they relate to each other, and their role in model fit analysis. They note that residuals should be interpretable in the context of the data, and that conditional residuals are estimates of the pure error or measurement error, and are the most useful out of the three different types of residuals. The marginal residuals in conjunction with the conditional residuals give the best description of the error in the model.

The following section summarises the methods of residual analyses that are described in Fitzmaurice *et al*. (2004), which are based on the marginal residuals. Many of the standard methods of residual analysis used for regression methods can be extended to longitudinal models.

For a longitudinal model a vector of residuals can be defined for each individual:

$$\mathbf{r}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}$$

which has a mean vector of zeroes. These residuals can then be used to check for systematic departures from the model, the presence of outliers, as well as the adequacy of the chosen covariance structure. A correctly specified model will have residuals which, if plotted against the fitted values, will be randomly scattered around the zero line and display no systematic pattern. Similarly, plots of the residuals against selected covariates can indicate missing quadratic terms or the need for a transformation.

Due to the properties of longitudinal data, the components of the residual vector will be correlated and may not have constant variance. The covariance of the marginal residuals can be approximated by the estimated marginal covariance matrix, $\hat{\mathbf{V}}_i$, and this has important implications for the analysis of the residual plots. Firstly, due to the covariance of the residuals not necessarily being constant, standard methods used to test for homogeneity of residual variance or autocorrelation among residuals should be avoided. Secondly, the residuals may be correlated with the covariates, unlike in the univariate case, and therefore systematic trend in plots of residuals versus a covariate may be due to this correlation.

To circumvent these problems, the marginal residuals can be transformed. The transformed residuals should then have zero correlation and unit variance, and this can be achieved through the Cholesky decomposition method (also known as Cholesky factorisation).

Since the approximate covariance matrix of the marginal residuals should equal $\hat{\mathbf{V}}_i$, using the Cholesky decomposition method, a lower triangular matrix, $\mathbf{L}_i$, can be created such that:

$$\hat{\mathbf{V}}_i = \mathbf{L}_i \mathbf{L}_i'$$

and $\mathbf{L}_i^{-1}$ can then be used to transform the residuals:

$$\mathbf{r}_i^* = \mathbf{L}_i^{-1}\mathbf{r}_i = \mathbf{L}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})$$

which will then have zero correlation and unit variance.

These transformed marginal residuals have useful interpretations in the longitudinal setting due to the temporal ordering of the observations for each subject. The first element of

$$\mathbf{r}_i^* = \mathbf{L}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})$$

is the standardised residual for the first repeated measurement, and the elements that follow represent standardised deviations from the conditional mean of the response given all previous observations. Therefore the k^{th} transformed residual is an estimate of

$$\frac{y_{ik} - E(y_{ik} \mid y_{i1},\ldots,y_{ik-1})}{\sqrt{\mathrm{Var}(y_{ik} \mid y_{i1},\ldots,y_{ik-1})}}.$$

Using the transformed residuals the usual residual diagnostics for standard regression methods can be applied. For example, the transformed residuals, r_{ij}^*, can be plotted against the transformed predicted values, $\hat{\mu}_{ij}^*$, where

$$\hat{\mu}_{ij}^* = \mathbf{L}_i^{-1}\hat{\boldsymbol{\mu}}_i = \mathbf{L}_i^{-1}\mathbf{X}_i\hat{\boldsymbol{\beta}}.$$

This plot should show a random scatter around zero with constant variance. Similarly the transformed residuals can be plotted against the individual transformed covariates. Plotting the transformed residuals against transformed time can be used to assess the model assumptions about the patterns of change in mean response over time. Using the transformed residuals makes it easier to identify skewness and potential outliers.

The transformed response vector and covariate matrix can be obtained as follows:

$$\mathbf{y}_i^* = \mathbf{L}_i^{-1}\mathbf{y}_i; \quad \mathbf{X}_i^* = \mathbf{L}_i^{-1}\mathbf{X}_i .$$

The generalised least squares (GLS) estimate of $\boldsymbol{\beta}$ from the regression of $\mathbf{y}_i$ on $\mathbf{X}_i$ can now be re-estimated using OLS regression of $\mathbf{y}_i^*$ on $\mathbf{X}_i^*$, and the standard residual diagnostics of this model can then be used to check model adequacy.

Houseman, Ryan and Coull (2004) give a detailed derivation of these "Cholesky residuals", and demonstrate the use of these residuals via simulation studies. The authors note that the transformed residuals are not always appropriate, in particular when the normality of the random effects is in question.

These scaled marginal residuals can be obtained from SAS PROC MIXED by specifying the option VCIRY in the model statement. The transformed residuals are only available for the marginal residuals. The untransformed marginal and conditional residuals, i.e. $\mathbf{r}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}$ and $\mathbf{r}_{ci} = \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} - \mathbf{Z}_i\hat{\mathbf{b}}_i$ can also be specified (SAS PROC MIXED, 2003). The residuals available in SAS (ver. 9.1), as well as the Cholesky residuals, are discussed by Schabenberger (2004).

2.6.4 The process of choosing the best model

In order to select the best fitting covariance structure for a linear mixed effects model, for both the random errors and the random effects, the literature recommends using likelihood ratio tests to select between nested models, and information criteria, such as the AIC or BIC to select between non-nested models (Verbeke & Molenberghs, 2000; Demidenko, 2004; Hedeker & Gibbons, 2006).

Verbeke and Molenberghs (2000) recommend fitting an overparameterised model for $E(\mathbf{y}_i)$ as a first step. This will result in consistent estimators of the covariance structure in future steps. Using this mean structure, the OLS method can be used to estimate $\boldsymbol{\beta}$. This can be shown to be a consistent estimator for $\boldsymbol{\beta}$. The OLS residuals can then be used to study the dependence among the repeated measures. These plots can be used to select the random effects to be included in the model. They further note that including high-dimensional random effects with an unconstrained covariance matrix $\boldsymbol{\Sigma}$ can lead to complicated covariance structures and may result in divergence of the maximisation procedure. If most of the variability can be assumed to be contained within the random effects, the parsimonious structures can be chosen for error covariance structure, $\boldsymbol{\omega}_i$.

The difference between the -2 REML loglikelihoods of the more complex model and the simpler model is distributed as chi-squared with degrees of freedom equal to the difference in the number of parameters between the two models. The covariates of these two models need to be the same. This can be used to see if the change of covariance structure, where one covariance structure is nested within another, is

suitable or not, but not to determine if the inclusion of random effects is necessary or not. This is because the null hypotheses of interest are on the boundary of the parameter space, which implies that the likelihood ratio statistic does not have the classical asymptotic chi-square distribution. This same method can be used in the case of maximum likelihood estimation, but both models need to have the same likelihood method. When the models are not nested, then the AIC or BIC should be used (Verbeke & Molenberghs, 2000; Demidenko, 2004; Hedeker & Gibbons, 2006). Davis (2002) notes that, to use the likelihood ratio test to compare the fit of the different models, the number of time points must be small and the time points must be equally spaced. If the number of time points is large, or vary from subject to subject, then the choice of covariance model can have a substantial effect on the results of the analysis, and likelihood ratio tests cannot be used.

Since the covariance parameters are not always defined as elements of $(-\infty, \infty)$, in particular the diagonal elements of the covariance matrix need to elements of $(0, \infty)$, standard tests for the parameter values which assume this may be misleading. When the likelihood ratio is used to test variance parameters to see if they are equal to a value on the parameter's boundary space, then this test may not be valid. The null distribution for the likelihood ratio test is no longer a chi-squared distribution with degrees of freedom equal to the difference between the number of parameters in the full and reduced models. The null distribution will be a mixture of chi-square distributions (Demidenko, 2004, Fitzmaurice *et al*., 2004). Under these circumstances, the p-value of the test may be adjusted. An approximate adjustment is to divide the p-value by two, which has been shown to work well (Hedeker & Gibbons, 2006).

When the AIC and BIC are used to compare between models, the BIC will almost always give a greater penalty for extra parameters, and can under certain circumstances heavily penalise for additional parameters, therefore Fitzmaurice *et al*. (2004) recommend against using the BIC. To demonstrate this, $(-2\hat{l} + 2c) < (-2\hat{l} + \log N \times c)$ when $2 < \log N$. Therefore for any sample size about eight, the BIC will give a greater penalty for more parameters compared to the AIC (McQuarrie & Tsai, 1998).

A limitation to using the AIC is that it does not work well when multicollinearity is present, and although it has been shown to be an asymptotically unbiased estimator of the Kullback-Leibler information, it can have significant bias under sufficiently small sample sizes. Therefore bias-corrected AIC (AICc), which estimates the Kullback-Leibler information directly instead of its approximation as estimated by the AIC (McQuarrie & Tsai, 1998), and Healthy AIC (HAIC) variants have been proposed in order to obtain more reliable criteria for selecting the best fitting model (Demidenko, 2004). The AICc is defined as

$$\text{AICc} = -2l + \frac{2cN}{N - c - 1} \text{ (McQuarrie \& Tsai, 1998).}$$

To define the HAIC, let $\mathbf{y}$ be an N-dimensional vector of observations whose distribution depends on c-dimensional vector of parameters $\boldsymbol{\theta}$ and has loglikelihood $l(\boldsymbol{\theta}; \mathbf{y})$. The penalised loglikelihood will then be

$$l = -\frac{c}{2}\ln\lambda^2 + l(\boldsymbol{\theta};\mathbf{y}) - \frac{1}{2\lambda^2}\| \boldsymbol{\theta} \|^2$$

and the maximum of the loglikelihood function over the variance is attained at $\lambda^2 = \frac{\| \boldsymbol{\theta} \|^2}{2}$. The HAIC is defined as

$$\text{HAIC} = H - 2\hat{l} + 2c$$

where $H = c(\ln(\| \hat{\boldsymbol{\theta}} \|^2 / c) - 1)$ (Demidenko, 2004). Demidenko (2004) justifies the use of the HAIC by demonstrating the multicollinearity problem of the AIC in the context of a linear regression model. Let the regression model have *c* explanatory variables and variance $\hat{\sigma}^2$. If a variable is added that is highly correlated with the other explanatory variables, this may not be well reflected in the AIC because $\hat{\sigma}^2$ will not change, due to multicollinearity. The OLS parameter estimate will become unstable due to $|\mathbf{X'X}| \approx 0$, which will lead to large values of $\| \hat{\boldsymbol{\beta}}_{LS} \|^2$. This will result in a large *H* value and therefore the instability in the model will be picked up by the HAIC. The HAIC is not implemented in standard statistical software for linear mixed effects models and will not be considered further in this study. The AICc is available as an output for SAS PROC MIXED (ver. 9.1) and will be included in the results of this study, together with the AIC and BIC. The AICc is asymptotically equivalent to the AIC for large samples (McQuarrie & Tsai, 1998). As discussed earlier, the BIC penalises more heavily for additional parameter values as the sample size increases, and this implies that as the sample size increases, the AICc will approach the AIC value and deviate more from the BIC value.

2.6.5 Graphical methods for comparing models with different covariance structures

Various methods of checking model fit or model assumptions are described in the literature. Outlying observations as well as outlying individuals can be identified using plots of the transformed residuals, as described in Section 2.6.3. A summary

measure of the multivariate distance between the observed and fitted responses can be calculated for each individual, based on the the Mahalanobis distance:

$$d_i = r_i^{*'} r_i^{*}$$

which should then have a chi-square distribution with degrees of freedom (df) equal to the dimension of r_i^* (which is equal to the number of repeated measures on individual *i*) if the model is correctly specified. Individuals that have d_i's with significant *p*-values would be possible outlying individuals (Fitzmaurice *et al*., 2004).

In order to check the adequacy of the variance assumption, the transformed residuals can be plotted against the transformed predicted values or against transformed time, and if the variance has been correctly specified, a random scatter around the zero line will be observed. The absolute values of the transformed residuals, $|r_{ij}^*|$, can be plotted against the transformed predicted values or against transformed time, and if the assumed variance structure is correct, then no systematic trend should be observed in the plot. To check for trend, a lowess (locally weighted scatterplot smoothing) curve can be fitted, centered at approximately 0.8, as the mean of the absolute values of the residuals should be 0.798, if the residuals follow the standard normal distribution (Fitzmaurice *et al*., 2004).

The empirical semi-variogram (also called the sample semi-variogram) is defined as half the average squared difference between pairs of residuals on the same individual whose corresponding observations are *h* units apart. A plot of the empirical semi-variogram provides an informal check on the overall adequacy of the model in terms

the covariance structure. For longitudinal data, the semi-variogram, $\gamma(h_{ijk})$, is given as:

$$\gamma(h_{ijk}) = \frac{1}{2}E(r_{ij} - r_{ik})^2$$

where h_{ijk} is the elapsed time between the j^{th} and k^{th} repeated measurement on the i^{th} individual. This can also be written as:

$$\begin{aligned}\gamma(h_{ijk}) &= \tfrac{1}{2}E(r_{ij} - r_{ik})^2 \\ &= \tfrac{1}{2}E(r_{ij}^2 + r_{ik}^2 - 2r_{ij}r_{ik})^2 \\ &= \tfrac{1}{2}\text{var}(r_{ij}) + \tfrac{1}{2}\text{var}(r_{ik}) - \text{cov}(r_{ij,}r_{ik})\end{aligned}$$

as the mean of the residuals is equal to zero. When the transformed residuals, r_{ij}^*, are used the semi-variogram simplifies to

$$\gamma(h_{ijk}) = \tfrac{1}{2}\text{var}(r_{ij}^*) + \tfrac{1}{2}\text{var}(r_{ik}^*) - \text{cov}(r_{ij}^*, r_{ik}^*) = \frac{1}{2}(1) + \frac{1}{2}(1) - 0 = 1$$

(Fitzmaurice *et al*., 2004).

Therefore a plot of the semi-variogram for the transformed residuals of a model with a correctly specified covariance matrix against the time elapsed between the corresponding observations should display a random scatter around the horizontal line centred at one. The empirical semi-variogram is very sensitive to outliers (Fitzmaurice *et al*., 2004).

Grady and Helms (1995) describe graphical techniques to aid in examining model fit of the variance-covariance structure. They suggest plotting the covariances or correlations as a function of the time between measures. This will help in investigating the assumed covariance structure.

2.7 The Potthoff and Roy Data Set

The data set used by Potthoff and Roy has become a classic data set in repeated measures models literature. Potthoff and Roy (1964) were the first to use the dental data set, which has become known as the Potthoff and Roy data set (PR data set), to investigate an approach to modelling longitudinal data. As explained in Section 2.2.1, their interest was in the implementation of an extension to the standard MANOVA model. Their approach involved appending a postmatrix, **P**, to the expectation equation of the standard MANOVA model, which is a within-individual design matrix. A positive definite matrix, **G**, whose form depends on the assumed structure of the covariance matrix of the responses, is required to transform the responses so that the usual MANOVA model can be used. Potthoff and Roy (1964) state that the choice of **G** is somewhat arbitrary, as the true structure of the covariance matrix is usually not known. In their study they found that the results were not sensitive to changes in the parameter values of **G**.

Pinheiro, Liu and Wu (2001) proposed a modification of the linear mixed effects model that would be more robust against outliers. They used the PR data set to investigate a robust hierarchical linear mixed effects model in which the random effects and the within-individual errors were multivariate *t*-distributed, which allowed for different numbers of observations per individual. They showed that by using a gamma-normal hierarchical structure, the model they proposed allows the identification and classification of outliers. By means of a simulation study based on the PR data set, they were able to show that their model based on a multivariate *t*-

distribution outperformed the standard Gaussian linear mixed effects model when outliers were present in the data.

Jennrich and Schluchter (1986) used the PR data set to compare the random effects model to general linear models under different structures of $\boldsymbol{\omega}_i$. Other literature which use the PR data set include the book by Davis (2002) who used this data to illustrate the repeated measures models discussed. As for Jennrich and Schluchter (1986), Davis (2002) compared linear models with CS, TOEP, UN, AR(1) and VC covariance specifications, as well as the Laird and Ware (1982) random intercept and slope model, and found that the CS and TOEP models fitted the data best. Verbeke & Molenberghs (2000) also use this data set to demonstrate various repeated measures models. They analysed this data under the same models as Davis (2002), in addition to the random effects model with $\boldsymbol{\omega}_i$ = VC and $\boldsymbol{\Sigma}$ = UN and the random intercept model with $\boldsymbol{\omega}_i$ = VC. They concluded that the random intercept model with $\boldsymbol{\omega}_i$ = VC (resulting in the same covariance matrix for the model as the linear model with CS error structure) best described the data. Pan and Fang (2002) use this data set to described different approaches to growth curve models. Therefore, as the PR data set is a landmark data set in the repeated measures literature, to allow for easy comparison to results obtained in previous studies, and as this data set is easily accessible, it presents itself as a logical choice for a data set on which to base the simulation study.