C H A P T E R 7

CONCLUSIONS

7.1 Choice of Covariance Structure

The consequence of misspecifying the covariance structure of a linear mixed effect model largely impacts on the inference around the estimates of the fixed effects rather than on the estimates of the fixed effects themselves. This will then impact on the probability of Type I and Type II errors. This was demonstrated by the results of the simulation study which showed that the coverage probability of the 95% confidence intervals of the fixed effects parameters was significantly lower than 95% in the case of most covariance specifications. This supports the literature (e.g Fitzmaurice *et al*., 2004 p. 163) which states that the covariance matrix needs to be correctly specified in order to obtain valid inferences about the mean. Correct specification of the covariance also increases the efficiency of the fixed effects estimates (Fitzmaurice *et al*., 2004, Weiss, 2005).

As discussed in Chapter two, if $\boldsymbol{\tau}$, the vector containing all the covariance parameters, is assumed to be known, then the ML estimator of $\boldsymbol{\beta}$, obtained from maximizing the marginal likelihood function, conditional on $\boldsymbol{\tau}$, is 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

$$\text{var}(\hat{\boldsymbol{\beta}}) = (\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{W}_i \mathbf{X}_i)^{-1} (\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{W}_i \, \text{var}(\mathbf{y}_i) \mathbf{W}_i \mathbf{X}_i) (\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{W}_i \mathbf{X}_i)^{-1} \quad ...(2)$$

$$= (\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{W}_i \mathbf{X}_i)^{-1} \quad ...(3)$$

where $\mathbf{W}_i$ equals $\mathbf{V}_i^{-1}(\boldsymbol{\tau})$ and $\mathbf{V}_i = \mathbf{Z}_i\boldsymbol{\Sigma}\mathbf{Z}_i' + \boldsymbol{\omega}_i$ = var($\mathbf{y}_i$) (Verbeke & Molenberghs, 2000). In order for $\hat{\boldsymbol{\beta}}$ to be a unbiased estimator of $\boldsymbol{\beta}$, it is sufficient that the mean $E(\mathbf{y}_i)$ be correctly specified as $\mathbf{X}_i\boldsymbol{\beta}$. In the case of the covariance matrix, the equivalence of equations (2) and (3) depends on the correct specification of the marginal covariance matrix $\mathbf{V}_i = \mathbf{Z}_i\boldsymbol{\Sigma}\mathbf{Z}_i' + \boldsymbol{\omega}_i$, and therefore inference based on equation (3) will not be robust if the covariance matrix is misspecified (Verbeke & Molenberghs, 2000). In cases where there are missing data, valid estimates of the covariance structure are required in order to impute values accurately, and therefore the estimates of the regression parameters may not be invariant to misspecification of the covariance structure (Fitzmaurice *et al*., 2004; Weiss, 2005).

Alternatively, robust standard errors may be obtained via the sandwich estimator (Liang & Zeger, 1986, Verbeke & Molenberghs, 2000, Crowder, 2001), which can be shown to be consistent as long as the mean is correctly specified (Verbeke & Molenberghs, 2000). But even in this case, the estimates of the standard errors under an incorrectly specified covariance will be poor compared to those obtained from the correct model (Crowder, 2001). In addition, the sandwich estimator is only suitable for balanced designs where the number of subjects is relatively high and the number of observations relatively low (Fitzmaurice *et al*., 2004). From a scientific point of view, it is important to understand the covariance matrix operating behind the data, as this describes the sampling distribution of the data, and therefore, for complete understanding of the data, it is important to model the covariance as accurately as possible. Therefore some statisticians (e.g. Weiss, 2005) are against the use of robust standard errors as the only alternative.

The question then arises as to whether a covariance structure exists that has the flexibility to describe most covariance structures, but has sufficiently few parameters to ensure that model estimates are obtained from the fitting procedure. From the results of both the simulation study and the model fitting to the ecological data set, it appears that a model with a TOEP error structure performs relatively well. In the simulation study, the TOEP model obtained average AIC and BIC values very close to the minimum for all models, and obtained coverage probabilities for 95% confidence intervals that were not significantly different from 95%. When fitted to the ecological data set, the TOEP model obtained good model fitting criteria compared to models with simple covariance structures. The plots for assessing the fit of the covariance structure revealed that the TOEP error structure was able to capture the basic covariance structure estimated in the best fitting model. The data sets used in this study represent a special case where the number of observations on each subject is the same, and the time between observations on each subject is also the same. An assumption of the TOEP structure is that the time between measurements on each subject is the same, and therefore this structure is not appropriate in cases where this assumption does not hold, for example cases where subjects differ in their number of observations (Fitzmaurice *et al*. 2004).

A second covariance structure which performed well is the random intercept and slope model with $\boldsymbol{\omega}_i$ = VC and $\boldsymbol{\Sigma}$ = UN. Models with this covariance structure performed well in the simulation study, obtaining coverage probabilities that were close to 95%, and in forecasting exercise, model with this covariance structure obtained the best MAE and MSE values under both the linear and quadratic mean models. When fitting a model with this covariance structure to the data set, the

information criteria tended to be lower compared to other models, such as the TOEP model, and the covariance plots obtained for these models tended to be slightly worse. Therefore, although the mean structure was fitted well, the covariance structure tended not to fit as well compared to other models.

A clear outcome from the results in this study is that the OLS model (i.e. the model assuming a covariance matrix with the off-diagonal elements equal to zero and constant variance) is a very poor choice for repeated measurements. This model consistently obtained the worst model-fitting criteria. The coverage probabilities obtained for the OLS model in the simulation study were significantly higher than the required 95%, indicating that the standard errors estimated from this model tend to be biased upwards. Demidenko (2004) shows that the variance estimated from the OLS model has systematic positive bias. Therefore this finding is not unexpected. Standard errors from the OLS model do not necessarily always have to be larger compared to those estimated from other models, as this model obtained relatively small standard errors when fit to the ecological data set. Essentially, the findings of this study show that inferences based on the OLS model regarding estimates of the fixed effects should not be relied on when the data are longitudinal.

7.2 Parameter Estimation

One of the biggest problems encountered while fitting models to this data set was the estimation of non-positive definite random effects covariance matrices. Of the 29 models fitted to the ecological data set, ten resulted in random effects covariance matrices that were not positive definite. This could be due to the marginal modelling

approach of obtaining the parameter estimates, as described is Section 2.2.3. As shown by Verbeke and Molenberghs (2000, p. 52), the hierarchical model for a linear mixed model implies that the covariance matrices for both the errors and the random effects need to be positive definite, thereby implying that $\mathbf{V}_i = \mathbf{Z}_i \boldsymbol{\Sigma} \mathbf{Z}_i' + \boldsymbol{\omega}_i$ is positive definite as well. The marginal model implies only that $\mathbf{V}_i = \mathbf{Z}_i \boldsymbol{\Sigma} \mathbf{Z}_i' + \boldsymbol{\omega}_i$ is positive definite, which does not imply that the $\boldsymbol{\Sigma}$ and $\boldsymbol{\omega}_i$ will be positive definite. Therefore estimates obtained for the covariance parameters may not always converge to values in the parameter space implied by the hierarchical model (Verbeke & Molenberghs, 2000). Alternatively, this error may also be encountered if a saddle point is reached by the fitting algorithm (Weiss, 2005).

The problem of non-convergence or obtaining covariance parameters that were on the boundary of their parameter spaces was also encountered. This generally occurred more frequently the more complex the linear mixed effect model became. Specifying better starting values for the parameters or by specifying a different fitting procedure these problems can in some cases be avoided (Verbeke and Molenberghs, 2000). Divergence of the fitting procedure can be an indicator of problems with the parameterisation of the model, or in the assumptions made by the model (Verbeke and Molenberghs, 2000). Therefore it is the recommendation of this study that the user avoid overly complex covariance structures unless there are specific reasons for specifying these structures. It is also important that users of linear mixed effects models carefully scrutinise the estimates of the covariance parameters to ensure that all estimates are valid and that the individual covariance structures are positive definite.

Ways of solving convergence problems include changing the fitting procedure, changing the starting values, or changing the model. Currently, few options are available for model fitting procedures in standard linear mixed effects software. In SAS PROC MIXED (ver. 9.1). there is no alternative to the NR algorithm implemented by this software. Within the "PARMS" statement of PROC MIXED (SAS ver. 9.1) it is possible to specify different starting values. By changing the starting values for the fitted models it is also possible to check how stable the estimates are. If the estimates for the fixed effects change dramatically when different starting values are used, then there may be problems with the estimates obtained for these models. By default, this procedure uses MIVQUE(0) estimates as starting values of the covariance parameters. A reason for the algorithm not converging may also be due to too few data points to estimate the covariance parameters specified in the model. The observation to parameter ratio for the PR data set had a minimum value of close to 1:1, but the ecological data set had a minimum ratio of approximately 3:1. Therefore, for at least the ecological data set, the number of observations should not have limited the optimisation procedure, as there were sufficient degrees of freedom available. By specifying a simpler covariance structure in the same family as the desired covariance structure, the convergence problem may also be solved, but at the cost of reduced complexity in the model.

Alternatives to the NR algorithm include the EM algorithm (Laird & Ware, 1982; Jennrich & Schluchter, 1986), the Fisher scoring algorithm (Jennrich & Schluchter, 1986), and Bayesian methods (Weiss, 2005). Lindstrom and Bates (1988) compare the EM algorithm to the NR algorithm. They state that the EM algorithm will always converge to a local maximum of the likelihood surface, but the number of iterations

required may be very high, whereas the NR algorithm may not always converge. In the twenty years since this article was published, computing power has increased substantially and therefore a large number of iterations may pose only a minor problem compared to lack of convergence. Jennrich and Schluchter (1986) state that the Fisher scoring algorithm is more robust to poor starting values compared to the NR method, and suggest starting off the fitting procedure with Fisher scoring estimates and then continuing with the NR method. Weiss (2005) suggests that more interest be placed in the implementation of Bayesian methods to fit linear mixed effects models. Bayesian methods are useful since they do not depend on asymptotics and can handle complex models, even with small data sets (Weiss, 2005). These methods also allow for inference about complex functions of the parameters and can easily be extended to new models, and therefore Bayesian methods are becoming more popular in non-linear modelling, especially with easily available software such as WinBUGS and increasing computer power (Weiss, 2005). Individuals in the R statistical software community have implemented Bayesian procedures for fitting the parameters of a linear mixed effects model through a Monte Carlo Markov Chain (MCMC) algorithm (Tüchler & Frühwirth-Schnatter, 2006). Therefore alternative procedures do exist, but may be difficult to implement, particularly for individuals not familiar with statistical software coding, until macros or packages for these procedures become more readily available. Individuals, both statisticians and non-statisticians, not experts in linear mixed effects theory may have difficulty in implementing these optimisation procedures to fit linear mixed effects models.

During the simulation study, estimates were often not obtained for the models with more complex covariance structures, even in cases when the model fitted had the

same covariance structure specified as during the simulation of the data. Therefore failure to obtain estimates was due to small sample size or inability of the fitting procedure to reach the best parameters subset, and not due to the non-existence of the parameter values. This finding indicates that greater flexibility, in terms of having different fitting procedures available and being able to tweak these procedures, is needed in order to guarantee convergence and obtain estimates for a model.

7.3 Simple Versus Complex Covariance Structures

The complexity of the covariance structures considered should be based on a number of factors. Firstly, the hypothesis or hypotheses behind the sampling distribution of the data should be considered, and covariance structures chosen accordingly. Secondly, the amount of data available should also be considered, as this will limit the maximum complexity of the covariance structure which can be fitted by the maximisation procedure. Ideally, this consideration should take place before the data is collected so that the sample size can be increased or the experimental design changed if necessary. The researcher needs to be aware of the number of data points available and the number of parameters that will need to be estimated by the model, and take this into account when analysing the parameter estimates. In cases where the number of data points does not allow for complex covariance structures, it may be better to fit a model with a simpler structure, as this study has shown that selection of a relatively flexible, low parameter covariance structure, such as the TOEP structure, can still result in good estimates for both the mean and the covariance, even if the covariance structure of the data is complex. Lastly, thought should also be given to whether or not the covariance structure estimates will need to be interpreted.

Interpretation of very complex covariance structures may be more difficult compared to a simpler structure.

In cases where complex covariance structures are required, it may be necessary to change starting values, or even to try different maximisation procedures across different statistical software, in order to obtain parameter estimates for the desired model.

In the model fitting exercise it was found that if the mean model structure was correctly specified, then the covariance structures estimated tended to be simpler, and therefore covariance structures could be specified with fewer parameters. On the other hand, if an over-simplified mean structure was selected, then the estimates for the estimated covariance matrix became more elaborate, evident in the covariance plots for the more complex covariance structures. Therefore the covariance structure can compensate for misspecified mean structures by having inflated estimates for certain covariances, which then results in larger standard errors for the fixed effects estimates.

7.4 Model Fit

Model selection requires choosing both the best fitting mean structure and the best fitting covariance structure. These two aspects of the model are interrelated, so there is no easy answer as to which should be fit first or take first priority. Fitzmaurice *et al*. (2004) suggest choosing a maximal mean structure first and then determining then which covariance structure fits best. Once the covariance structure is chosen, the best

fitting mean structure can be selected through model diagnostics. Verbeke and Molenberghs (2000) suggest initially fitting an elaborate or saturated mean structure, and then using the OLS estimates for **β**, plot the residuals for the OLS fit, which should be consistent for **β**, to explore the dependence between repeated measurements and be used to determine what random effects should be included in the model. They then suggest selecting a covariance matrix for the errors conditional on the selected random effects. In the next step, the need for the random effects can be tested, although this should not be done by means of the likelihood ratio test, as explained in Section 2.6.4.

The approach used for fitting a model to the ecological data set was to fit the maximal mean structure under various covariance structures, then to reduce the mean structure if an interaction or fixed effect term was consistently non-significant. This process was then repeated once a term was removed from the mean structure. This procedure could be quite tedious, but it ensures that no potential models are left out. In the model fitting exercise carried out on the ecological data set, this procedure had to be repeated twice, as closer investigation of the residuals fitted to the best linear model revealed that a simple linear relationship with time was not sufficient to describe the mean model, nor were the residual variances homogeneous. Therefore selecting an appropriate mean structure and a covariance structure for a linear mixed effects model applied to longitudinal data can be a tedious task.

In order to choose the best fitting covariance structure, the basic texts on linear mixed effects models (e.g. Verbeke & Molenberghs, 2000; Fitzmaurice *et al.*, 2004; Weiss, 2005; Hedeker & Gibbons, 2006) suggest using likelihood ratio tests to choose

between nested covariance structures, and to use information criteria, such as the AIC and BIC to choose between non-nested covariance structures. These model selection criteria can only be used to compare between models if the fixed effects are held constant (Verbeke & Molenberghs, 2000; Fitzmaurice *et al*., 2004). The BIC will tend to pick models with fewer covariance parameters compared to the AIC (McQuarrie & Tsai, 1998; Fitzmaurice *et al*., 2004; Weiss, 2005). The AIC has been shown to work poorly in the presence of multicollinearity and to have small-sample overfitting tendencies (McQuarrie & Tsai, 1998; Demidenko, 2004). McQuarrie and Tsai (1998) suggest using the bias corrected AIC (AICc), which has been shown to outperform the AIC for small samples. This study showed that the models selected by these two criteria are very similar, and generally tended to have very similar values for the same model. Demidenko (2004) proposes a Healthy AIC (HAIC), which is shown to be more sensitive under multicollinearity and better able to distinguish between models when the number of parameters is the same. This criterion is not readily available in most statistical software packages.

Recently, Gomez, Schaalje and Fellingham (2005) studied the performance of the Kenward-Roger degrees of freedom calculation method when useing the AIC and BIC to choose the best covariance structure, by means of a simulation study. They concluded at the end of their study that the AIC and BIC criteria were very poor at selecting the correct covariance structure. Therefore, these information criteria may not be suitable for selecting between linear mixed effects models with different covariance structures. They suggested using the AICc instead, but this study showed that the AIC and AICc tended to select the same models, even for the smaller PR data set.

Since there is more than one aspect of the model that needs to be correctly specified, once the best fitting covariance structures have been selected, it is then necessary to further investigate model fit through the use of residual and influence diagnostics, and plots for the covariance structure such as the semi-variogram and the plot of the covariance terms versus the lag in time. The influence diagnostics will be used to determine if there are any observations in the data to which the model fit is particularly sensitive (Verbeke & Molenberghs, 2000; Demidenko, 2004). The residual diagnostics (based on transformed residuals as discussed in Chapter two) can be used to assess the adequacy of the fitted model, as well as to identify potential outliers (Fitzmaurice *et al*., 2004). In addition, to assess the fit of the covariance matrix, the semi-variogram of the transformed residuals can be plotted against the time lag between observations as a visual check of the proposed covariance matrix (Fitzmaurice *et al*., 2004). A plot of the estimated covariance terms against the lag in time can be plotted as a visual representation of the covariance matrix, which can be used to visually compare covariance structures between different models (Hedeker & Gibbons, 2006). By systematically going through this process, it should be ensured that the model finally chosen is the best representation of the data under analysis. Since the linear mixed effects model is more complicated compared to a normal linear model, these model assessment steps are comparatively more critical to determine any model perturbations, and should not be ignored.

7.5 Improvement and Further Research

This study did not explicitly consider the effect of different sample sizes on the robustness of the linear mixed effects model, therefore an improvement on this study

would be the inclusion of different sample sizes during the simulation study. By comparing the models successfully fit to the ecological data set, which had approximately twice as many subjects per group compared to the PR data set, to those successfully fit to the PR data set, it does appear that the more observations available the better the more complicated covariance structures perform.

Only data sets where the same number of observations was taken from all individuals were considered. This study could be extended by including a data set where the time lags between observations differed from subject to subject, or where a different number of observations were obtained on each subject. Under these circumstances, covariance structures assuming constant lags in time between observations, such as AR(1) or TOEP structures, would not be appropriate. It would be interesting to examine which covariance structures would be suitable under these conditions.

This study showed that the maximisation procedure was not always successful, even when the underlying data had the same covariance specified in the model. Therefore, an interesting study would be to compare different fitting procedures, under data simulated from known covariance structures, in terms of whether the model fitting procedures converge and how close the parameters estimates get to their true values. Examples of different fitting procedures would include the NR algorithm, the Fisher scoring, and the EM algorithm, as well as hierarchical Bayes methods.

In conclusion, the linear mixed effects model is a very useful tool in the analysis of repeated measurements, but careful consideration needs to go into the postulated mean and covariance structures.