7 Modeling the Covariance

7.1 INTRODUCTION

Since one of the defining features of longitudinal data is that they are correlated, we must consider approaches for appropriately modeling the covariance or time dependence among the repeated measures obtained on the same individuals. When an appropriate model for the covariance has been adopted, correct standard errors are obtained and valid inferences about the regression parameters can be made. Accounting for the covariance among repeated measures usually increases efficiency or the precision with which the regression parameters can be estimated; that is, the positive correlation among the repeated measures reduces the variability of the estimate of change over time within individuals. Thus in a longitudinal study the positive correlation among repeated measures can be used to advantage in the study of change over time. In addition, when there are missing data, correct modeling of the covariance is often a requirement for obtaining valid estimates of the regression parameters. In general, failure to take account of the covariance among the repeated measures will result in incorrect estimates of the sampling variability and can lead to misleading scientific inferences.

Longitudinal data present us with two aspects of the data that require modeling: the conditional mean response over time and the conditional covariance among repeated measures on the same individuals. Although these two aspects of the data can, in a certain sense, be modeled separately, they are also interrelated. That is, the choice of models for the mean response and the covariance are interdependent. This interdependence arises because the vector of residuals (observed responses minus fitted responses) depends on the specification of the model for the conditional mean.

Applied Longitudinal Analysis, Second Edition. By Garrett M. Fitzmaurice, Nan M. Laird, and James H. Ware. Copyright © 2011 John Wiley & Sons, Inc.

Put more formally, the covariance between any pair of residuals, say $\{Y_{ij} - \mu_{ij}(\beta)\}$ and $\{Y_{ik} - \mu_{ik}(\beta)\}$, depends on the model for the conditional mean (i.e., depends on β). A model for the covariance must be chosen on the basis of some model for the mean response; it represents an attempt to account for the covariance among the residuals that results from a specific model for the mean. A different choice of model for the mean, or moreover any misspecification of the model for the mean, can potentially result in a different choice of model for the covariance. As a result of this interdependence between the models for the mean and covariance, we will need to develop an overall modeling strategy that takes this interdependence into account.

7.2 IMPLICATIONS OF CORRELATION AMONG LONGITUDINAL DATA

Before considering approaches for modeling the covariance or correlation among repeated measures, it is worth stepping back and considering some of the implications of the correlation among longitudinal data. First, it should be kept in mind that longitudinal data are not only correlated, for the most part they are also positively correlated. Moreover the positive correlation among repeated measures can be used to advantage in the study of change over time. That is, we can capitalize on the positive correlation among longitudinal data when the main focus of the analysis is on change in the mean response over time.

Consider a simple longitudinal study design where it is of interest to measure change in a health outcome "before" and "after" receiving some health intervention. With only two repeated measures of the outcome, the statistical analysis of these data will focus on the difference score, say $Y_{i2} - Y_{i1}$, for each individual. Note that the variability of the difference score is given by

$$\begin{aligned} \text{Var}(Y_{i2} - Y_{i1}) &= \text{Var}(Y_{i1}) + \text{Var}(Y_{i2}) - 2\,\text{Cov}(Y_{i1}, Y_{i2}) \\ &= \sigma_1^2 + \sigma_2^2 - 2\,\sigma_{12} \\ &= \sigma_1^2 + \sigma_2^2 - 2\,\rho_{12}\sigma_1\sigma_2, \end{aligned}$$

where ρ_{12} is the correlation among the pair of responses, Y_{i1} and Y_{i2}. On the other hand, suppose that an alternative study design is adopted to assess the impact of the health intervention. Rather than using a longitudinal design, a cross-sectional design is adopted where study participants are randomly assigned to two groups, a group that receives the intervention and a control group that does not. Then the variance of the difference between the responses of any two individuals, when one individual is randomly selected from the intervention group and the other from the control group, is given by

$$\begin{aligned} \text{Var}(Y_{i2} - Y_{i1}) &= \text{Var}(Y_{i1}) + \text{Var}(Y_{i2}) \\ &= \sigma_1^2 + \sigma_2^2 \end{aligned}$$

(where Y_{i1} and Y_{i2} now denote the responses from two different individuals from the control and intervention groups, respectively).

Thus, provided that the correlation among repeated measures is positive, the variability of the within-individual differences is always smaller than the variability of the between-individual differences. If in this simple illustration we further assume that the variance of the response is constant (over time in the longitudinal study design, and across groups in the cross-sectional study design), with $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then the variance of the within-individual differences is simply $2\sigma^2(1 - \rho)$, while the variability of the between-individual differences is $2\sigma^2$. The ratio of these two variances provides an index of the precision of within-individual differences when compared to between-individual differences. Their ratio (within-individual variance / between-individual variance) can be expressed as $(1 - \rho)$. Thus, when the correlation is relatively large and positive, the variability of the within-individuals differences (or within-individual changes) can be substantially smaller than that for the corresponding between-individual differences. It is in this sense that a longitudinal study can provide a more precise (i.e., less variable) estimate of change in the mean response than a cross-sectional study with the same number and pattern of observations.

Finally, it must be emphasized that failure to adequately account for the correlation among repeated measures can result in misleading inferences. For instance, if it is assumed that the repeated measures are uncorrelated, when in fact there is strong positive correlation, the nominal standard errors (resulting from the naive assumption of independent or uncorrelated repeated measures) will be incorrect. Specifically, for contrasts that estimate change in the mean response over time, the nominal standard errors will be too large. In this case, one fails to get the full benefit of longitudinal data. With incorrect standard errors, test statistics and p-values will also be incorrect and thus can lead to incorrect inferences about patterns of change and their relation to covariates. In addition, when there is missingness that is MAR, but not MCAR, likelihood-based estimation of the regression parameters, β, requires that the entire joint distribution of the vector of responses be correctly specified. As a result the model for the covariance must be correctly specified to ensure that valid estimates of β are obtained. In general, when there are missing data, greater care must be exercised when modeling the covariance among the responses (see Chapters 17 and 18).

In summary, the positive correlation among repeated measures is an inescapable feature of longitudinal data that must be accounted for in the analysis in order to make appropriate inferences. Although the correlation, or more generally, the covariance among the repeated responses, is not usually of intrinsic interest, it cannot simply be ignored. Moreover the positive correlation among longitudinal data enables us to estimate changes in the mean response, and their relation to covariates, with far greater precision than would be possible if the data were uncorrelated. Recognizing that the covariance is an important aspect of the data that must be properly accounted for to complete the specification of any regression model for longitudinal data, there are three broad approaches to modeling the covariance that can be distinguished. The first is to allow any arbitrary pattern of covariance among the repeated measures; this approach results in an "unstructured" covariance and is the topic of Section 7.3. Alternatively, structure can be placed on the covariance matrix and there are two main strategies for doing so. The first modeling approach borrows ideas from the time series literature and assumes that the variances and covariances are not arbitrary but follow distinctive patterns. As a result we refer to these models as covariance pattern models, and they are the topic of Section 7.4. Finally, in a somewhat less direct way, structure can be imposed on the covariance through the introduction of *random effects* in the model for the mean response. That is, by assuming that the mean response depends on a combination of population parameters, β (also known as fixed effects), and individual-specific random effects, a very distinctive structure can be imposed on the covariance matrix. Because of the important role of the random effects structure in modeling the covariance in longitudinal data, discussion of these models will be the topic of Chapter 8.

7.3 UNSTRUCTURED COVARIANCE

When the number of measurement occasions is relatively small and all individuals are measured at the same set of occasions, it may be reasonable to allow the covariance matrix to be arbitrary, with all of its elements unconstrained. The only formal requirement is that the covariance matrix be symmetric and positive-definite (recall that the latter condition ensures that while the repeated measures can be highly correlated, there must be no redundancy; that is, none of the repeated measures can be expressed as a linear combination of the others). When no explicit structure is assumed for the covariance among the repeated measures (other than the homogeneity of covariance across different individuals, $\text{Cov}(Y_i) = \Sigma_i = \Sigma$), the resulting covariance is referred to as an "unstructured" covariance. The chief advantage of an "unstructured" covariance is that no assumptions are made about the variances and covariances. The absence of restrictions on the variances is especially important since our practical experience with many longitudinal studies has led to the empirical observation that the variances are rarely constant over time. For example, the variability of baseline measurements is often discernibly different from the variability of post-baseline measurements.

With n measurement occasions, the "unstructured" covariance matrix has $\frac{n\times(n+1)}{2}$ parameters: the n variances at each occasion and the $n \times (n-1)/2$ pairwise covariances (or correlations),

$$\text{Cov}(Y_i) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix}.$$

Herein lies one of the potential drawbacks of assuming an unstructured covariance: the number of covariance parameters to be estimated grows rapidly with the number of measurement occasions. For example, when there are three occasions ($n = 3$), the number of covariance parameters is 6 (3 variances and 3 pairwise covariances). However, when $n = 5$, the number of covariance parameters has grown to 15, while when $n = 10$, the number of covariance parameters is 55 (and may be fast approaching the number of subjects enrolled in some longitudinal studies!). When the number of covariance parameters that need to be estimated is large, relative to the sample size, estimation is likely to be very unstable. Thus the use of an unstructured covariance will be appealing only in cases where the number of subjects, N, is large relative to the number of covariance parameters, $\frac{n\times(n+1)}{2}$.

Setting aside the issue of the potentially large number of covariance parameters that may need to be estimated, the use of an unstructured covariance matrix is problematic when there are mistimed measurements or, more generally, measurement made at grossly irregular intervals. Even the most carefully designed longitudinal study will frequently suffer from deviations from the measurement protocol, resulting in measurements made at arbitrary, irregularly timed intervals. When this problem arises, as it frequently does in studies in the health sciences, the resulting mistimed repeated measurements cannot be accommodated in an unstructured covariance. Thus, when the longitudinal data are inherently unbalanced and/or when the sample size is not sufficiently large to estimate an unstructured covariance, it is usually desirable to impose some structure on the covariance matrix.

7.4 COVARIANCE PATTERN MODELS

When attempting to impose some structure on the covariance, a subtle balance needs to be struck. If too little structure is imposed, there may be too many parameters to estimate with the limited amount of data at hand. This was one of the main drawbacks of the unstructured covariance considered in the previous section; by imposing no structure on the covariance, the number of parameters to be estimated grows rapidly with the number of measurement occasions. In a certain sense any given data set contains but a fixed amount of longitudinal information. If too little structure is imposed on the covariance, there will be too many covariance parameters to be estimated from the limited amount of data available, and this will adversely affect the precision with which the main parameters of interest, β, can be estimated. As a result, imposing too little structure on the covariance can result in weaker inferences concerning β. When structure is imposed on the covariance, it is possible to improve the precision with which β can be estimated. However, if too much structure is imposed, there is a potential risk of model misspecification that could ultimately result in misleading inferences concerning β. Once again, this is the classic trade-off between bias and precision. In modeling the covariance, a balance must be struck between these two competing forces.

Structure can be built into the covariance by adopting a covariance pattern model. Covariance pattern models for longitudinal data have their basis in models for serial correlation that were originally developed for time series data. While time series data have a structure that is somewhat different than longitudinal data, being composed of a small number of replications or individuals (in some cases only a single replication) and a large number of repeated measures, they share a common characteristic: the repeated measures are positively correlated and measures taken closer together in time are expected to be more highly correlated than measures further apart in time. Because there are few, if any, replications, much of the statistical literature on the analysis of time series data has focused on parametric models that can describe the covariance structure among the repeated measures with only a few parameters. Many of the models for time series data result in relatively parsimonious models for the covariance that can also be adopted for longitudinal data. Here we describe some of the most widely used covariance pattern models for longitudinal data. Many of these covariance pattern models are available as options in standard statistical software packages for analyzing longitudinal data (e.g., PROC MIXED in SAS).

Compound Symmetry

Historically one of the first covariance pattern models used for the analysis of repeated measures data was compound symmetry. With a compound symmetry covariance it is assumed that the variance is constant across occasions, say σ^2, and $\text{Corr}(Y_{ij}, Y_{ik}) = \rho$ for all j and k. That is,

$$\text{Cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{pmatrix},$$

with the constraint that $\rho \geq 0$.

The compound symmetry covariance has a randomization justification in certain repeated measures designs (e.g., split-plot designs). In an experiment where the within-subject factor is randomly allocated to subjects, randomization arguments can be made to show that the constant variance and constant correlation conditions hold. (See Chapter 21 for a more detailed discussion of the randomization argument.) However, the randomization argument is simply not justifiable in the longitudinal data setting since measurement occasions cannot be randomly allocated to subjects.

As mentioned in Chapter 3, the compound symmetry covariance does have a theoretical justification when the mean response is thought to depend on a combination of population parameters, β, and a single individual-specific random effect. When the model for the longitudinal responses is expressed as

$$Y_{ij} = X'_{ij}\beta + b_i + \epsilon_{ij},$$

where b_i is a random effect and ϵ_{ij} is a within-individual measurement error, this induces marginally (or averaged over the random effect) a compound symmetry structure on the covariance matrix (with the constraint that $\rho \geq 0$). A more detailed discussion of random effects structures for the covariance will be given in Chapter 8.

The compound symmetry covariance is very parsimonious, with only two parameters regardless of the number of measurement occasions. However, it does make the rather strong assumption that the correlation between any pair of measurements is the same regardless of the time interval between the measurements. This latter aspect of the compound symmetry covariance, the constraint on the correlation among repeated measurements, is somewhat unappealing for most longitudinal data, where the correlations are expected to decay with increasing separation in time. Also the assumption of constant variance across time is unrealistic in many settings. As mentioned earlier, our practical experience with many longitudinal studies has led to the empirical observation that the variances are rarely constant over time. In many settings the assumption of constant variance is the one that is not valid with longitudinal data.

Toeplitz

The Toeplitz covariance pattern makes the assumption that any pair of responses that are equally separated in time have the same correlation. When the covariance has a Toeplitz form, it is assumed that the variance is constant across occasions, say σ^2, and $\text{Corr}(Y_{ij}, Y_{ij+k}) = \rho_k$ for all j and k. That is,

$$\text{Cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{pmatrix}.$$

Because a Toeplitz covariance assumes that the correlation among responses at adjacent measurement occasions is constant, ρ_1, this structure is only appropriate when the measurements are made at equal (or approximately equal) intervals of time. Note that the Toeplitz covariance has n parameters (1 variance parameter, and $n - 1$ correlation parameters). A special case of the Toeplitz covariance is the (first-order) autoregressive covariance.

Autoregressive

In the autoregressive model for the covariance it is assumed that the variance is constant across occasions, say σ^2, and $\text{Corr}(Y_{ij}, Y_{ij+k}) = \rho^k$ for all j and k, and $\rho \geq 0$. That is,

$$\text{Cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix}.$$

The autoregressive covariance is very parsimonious and has only two parameters, regardless of the number of measurement occasions. Because the autoregressive covariance has a Toeplitz form, this structure is only appropriate when the measurements are made at equal (or approximately equal) intervals of time. Note that the correlations decline over time as the separation between pairs of repeated measures increases. However, as mentioned in Section 2.5, in many settings the correlations among repeated measures on the same individual rarely decay that quickly over time.

The autoregressive covariance has a theoretical justification when the errors, e_{ij}, are thought of as arising from the following first-order autoregressive process:

$$e_{ij} = \rho\, e_{ij-1} + w_{ij},$$

where $w_{ij} \sim N\left(0, \sigma^2\left[1-\rho^2\right]\right)$ and the process is initiated by an error, say e_{i0}, where $e_{i0} \sim N\left(0, \sigma^2\right)$. The autoregressive process is said to be "first-order" because there is only dependence on the previous error; dependence on the two previous errors would yield a "second-order" autoregressive process. Thus the autoregressive covariance can be thought of as resulting from a process where the error term at the j^{th} occasion is a deterministic function of the error at the previous occasion, $\rho\, e_{i,j-1}$ (i.e., the recent past predicts the present), plus an additional (and independent) source of random error, w_{ij}. For such a process, it can be shown that

$$\text{Var}\left(e_{ij}\right) = \sigma^2$$

and

$$\text{Cov}\left(e_{ij}, e_{ik}\right) = \sigma^2 \rho^{|j-k|}.$$

Finally, the compound symmetry, Toeplitz, and autoregressive covariances assume that the variances are constant across time. This assumption can easily be relaxed, and it is possible to consider versions of these three covariance pattern models with heterogeneous variances, $\text{Var}(Y_{ij}) = \sigma_j^2$. Thus a heterogeneous (variances) autoregressive covariance pattern model is given by

$$\text{Cov}(Y_i) = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho^2\sigma_1\sigma_3 & \cdots & \rho^{n-1}\sigma_1\sigma_n \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \cdots & \rho^{n-2}\sigma_2\sigma_n \\ \rho^2\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 & \cdots & \rho^{n-3}\sigma_3\sigma_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1}\sigma_1\sigma_n & \rho^{n-2}\sigma_2\sigma_n & \rho^{n-3}\sigma_3\sigma_n & \cdots & \sigma_n^2 \end{pmatrix},$$

and has $n+1$ parameters (n variance parameters and 1 correlation parameter).

Banded

The banded covariance patterns make the assumption that the correlation is zero beyond some specified interval. For example, a banded covariance pattern with a band size of 3 assumes that $\text{Corr}(Y_{ij}, Y_{ij+k}) = 0$ for $k \geq 3$. It is possible to apply a banded pattern to any of the covariance pattern models considered so far. Thus a banded Toeplitz covariance pattern with a band size of 2 is given by

$$\text{Cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho_1 & 0 & \dots & 0 \\ \rho_1 & 1 & \rho_1 & \dots & 0 \\ 0 & \rho_1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix},$$

where $\rho_2 = \rho_3 = \cdots = \rho_{n-1} = 0$.

Banding makes a very strong assumption about how quickly the correlation decays to zero with increasing separation between the repeated measurements. In our experience with longitudinal studies in the health sciences, it is rare for the correlation to decay to zero, even in studies where there is a lengthy period of follow-up.

Exponential

When the measurement occasions are not equally spaced over time, the formulation of the autoregressive covariance model can be generalized as follows: Let $\{t_{i1}, \dots, t_{in}\}$ denote the observation times for the i^{th} individual, and assume that the variance is constant across all measurement occasions, say σ^2, and

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho^{|t_{ij} - t_{ik}|},$$

for $\rho \geq 0$. That is, the correlation between any pair of repeated measures decreases exponentially with the time separations between them. This structure is referred to as an "exponential" covariance model because it can be re-expressed as

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{ik}) &= \sigma^2 \rho^{|t_{ij} - t_{ik}|} \\ &= \sigma^2 \exp\left(-\theta \left|t_{ij} - t_{ik}\right|\right), \end{aligned}$$

where $\theta = -\log(\rho)$ or $\rho = \exp(-\theta)$ for $\theta \geq 0$. Also note that the exponential covariance model is invariant under linear transformation of the time scale. If we replace t_{ij} by $(a + bt_{ij})$ (e.g., if we replace time measured in "weeks" by time measured in "days"), the same form for the covariance matrix holds.

A distinctive feature of the exponential model is that it assumes that the correlation is one if measurements are made repeatedly at the same occasion (or replicate measurements on an individual can be obtained at the same occasion), and that the correlation decreases rapidly to zero as the time separation between measurements increases. This first aspect of the exponential covariance model corresponds to an assumption that the responses are measured without error, an unrealistic assumption in most longitudinal studies in the health sciences. The latter feature, correlations among repeated measurements that decay to zero, is rarely observed in longitudinal studies.

Hybrid Models

Finally, by combining the autoregressive and the compound symmetry models, it is possible to overcome many of the unappealing aspects of each of these models for longitudinal data. Consider a model for the covariance where

$$\text{Cov}(Y_i) = \Sigma_1 + \Sigma_2,$$

where

$$\Sigma_1 = \sigma_1^2 \begin{pmatrix} 1 & \rho_1 & \rho_1 & \cdots & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_1 \\ \rho_1 & \rho_1 & 1 & \cdots & \rho_1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_1 & \rho_1 & \rho_1 & \cdots & 1 \end{pmatrix}$$

and

$$\Sigma_2 = \sigma_2^2 \begin{pmatrix} 1 & \rho_2^{|t_{i1}-t_{i2}|} & \rho_2^{|t_{i1}-t_{i3}|} & \cdots & \rho_2^{|t_{i1}-t_{in}|} \\ \rho_2^{|t_{i2}-t_{i1}|} & 1 & \rho_2^{|t_{i2}-t_{i3}|} & \cdots & \rho_2^{|t_{i2}-t_{in}|} \\ \rho_2^{|t_{i3}-t_{i1}|} & \rho_2^{|t_{i3}-t_{i2}|} & 1 & \cdots & \rho_2^{|t_{i3}-t_{in}|} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_2^{|t_{in}-t_{i1}|} & \rho_2^{|t_{in}-t_{i2}|} & \rho_2^{|t_{in}-t_{i3}|} & \cdots & 1 \end{pmatrix}.$$

In this model

$$\text{Var}(Y_{ij}) = \sigma_1^2 + \sigma_2^2,$$

$$\text{Cov}(Y_{ij}, Y_{ik}) = \rho_1\sigma_1^2 + \rho_2^{|t_{ij}-t_{ik}|}\sigma_2^2,$$

$$\text{Corr}(Y_{ij}, Y_{ik}) = \frac{\rho_1\sigma_1^2 + \rho_2^{|t_{ij}-t_{ik}|}\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

This implies that the correlation between replicate measurements on an individual obtained at the same occasion is

$$\frac{\rho_1\sigma_1^2 + \sigma_2^2}{\sigma_1^2 + \sigma_2^2},$$

which is less than one when $\rho_1 < 1$. Furthermore, as the time separation increases, the correlation no longer decays to zero but has a minimum of

$$\frac{\rho_1\sigma_1^2}{\sigma_1^2 + \sigma_2^2},$$

which is greater than zero provided $\rho_1 > 0$. As noted, the compound symmetry model is also a random effects model, so that Σ_1 can be written as

$$\Sigma_1 = \begin{pmatrix} \sigma_b^2 + \sigma_\epsilon^2 & \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_\epsilon^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma_\epsilon^2 & \cdots & \sigma_b^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 + \sigma_\epsilon^2 \end{pmatrix},$$

so that $\sigma_1^2 = \sigma_b^2 + \sigma_\epsilon^2$, and $\rho_1 = \sigma_b^2/(\sigma_b^2 + \sigma_\epsilon^2)$. Thus we can think of the total variance, $\text{Var}(Y_{ij})$, as the sum of the autoregressive variance, σ_2^2, subject-to-subject variability, σ_b^2, and measurement error variability, σ_ϵ^2.

7.5 CHOICE AMONG COVARIANCE PATTERN MODELS

As mentioned at the beginning of this chapter, the choices of models for the covariance and for the mean are interdependent. As a result it is important to follow a modeling strategy that will result in a sensible choice of models for both aspects of the data. Since model selection criteria for the mean response depend on the correct specification of the model for the covariance (e.g., confidence intervals and tests of hypotheses concerning components of β depend critically on the correct model for the covariance), the first step is to choose a suitable model for the covariance.

It must be recognized that any model for the covariance depends on the assumed model for the mean. A model for the covariance tries to account for the covariance among the residuals, say $\{Y_{ij} - \mu_{ij}(\beta)\}$ and $\{Y_{ik} - \mu_{ik}(\beta)\}$, that result from a specific model for the mean. Therefore the choice of model for the covariance should be based on a "maximal" model for the mean that minimizes any potential misspecification of the model for the mean. Recall that any misspecification of the model for the mean can result in a certain amount of spurious covariance among the residuals, and can induce spurious dependence of the covariance on the covariates.

In longitudinal studies with balanced designs and a very small number of discrete covariates that can be classified as between-subject factors (e.g., treatment assignments, exposure levels, or some characteristic of the subjects), the choice of maximal model is relatively straightforward, since it is possible to choose as the maximal model one that includes the main effects of time (regarded as a within-subject factor) and all other main effects, in addition to their two-way and higher-way interactions. For example, with n measurement occasions and a single grouping factor with G levels (e.g., treatment versus control), it is possible to fit a saturated model for the mean response with separate parameters for the $G \times n$ means. This corresponds to a model with main effects for both the grouping factor and time, in addition to their interaction. This strategy of fitting saturated models for the mean response will be appropriate for longitudinal studies with balanced designs and where the number of qualitatively different levels of the covariates is relatively small. A saturated model for the mean response allows an arbitrary pattern for the mean response profile at every different level of the covariates and thereby minimizes any potential concerns about the impact of misspecification of the model for the mean.

However, in longitudinal studies where there are many covariates (some of which may be quantitative, rather than discrete), the choice of a maximal model is somewhat more difficult. In this case it is not realistic to consider a saturated model for the mean response; instead, a maximal model should be in a certain sense the most elaborate or complex model for the mean response that we would consider from a subject-matter point of view. Such a model may need to distinguish treatment covariates (e.g., treatment groups in experiments) or quasi-treatment covariates (e.g., exposure groups in observational studies) that are the main focus of the study from other covariates that are regarded as potential confounders or effect modifiers. The maximal model will ordinarily include the main effects of the treatment or quasi-treatment covariates and their interactions with time, since the latter effects characterize how changes in the mean response depend on these covariates. The choice of whether to include additional interactions, and so on, must be made on subject-matter grounds. In summary, when there are many potential covariates that can be included in the model for the mean, it is not straightforward to give a simple prescription for choosing the maximal model. The choice of maximal model, it must be recognized, cannot be made through any automatic procedure but must, rather, reflect substantive subject-matter considerations. The maximal model for the mean is a model that excludes certain higher-order interactions among the potential covariates and usually is more complex than any of the sequence of models for the mean response under consideration from a subject-matter point of view. In a sense the reader should envisage a model that, in its degree of complexity, goes a step beyond any model for which empirical researchers in the field would care to provide a specific rationale. Once a maximal model has been chosen, the residual variation and covariation can then be used to select an appropriate model for the covariance.

Given a maximal model for the mean, a sequence of covariance pattern models can be fit to the data at hand. The choice among models can be made by comparing the maximized likelihoods for each of the covariance pattern models. That is, when any pair of models is nested, a likelihood ratio test statistic can be constructed that compares the "full" and "reduced" models. Recall that two covariance models are said to be nested when the "reduced" model is a special case of the "full" model, so that, when the reduced model holds, the full model must necessarily hold. For example, the compound symmetry model is nested within the Toeplitz model since, if the compound symmetry model holds, then the Toeplitz model must necessarily hold, with $\rho_1 = \rho_2 = \cdots = \rho_{n-1}$. The likelihood ratio test for two nested covariance model can be constructed by comparing the maximized REML log-likelihoods, say $\widehat{l}_{\text{full}}$ and $\widehat{l}_{\text{red}}$, for the full and reduced models, respectively. The use of REML, as an alternative to ML, is preferred because it reduces the well-known finite sample bias in the estimation of the covariance. The likelihood ratio test is obtained by taking twice the difference in the respective maximized REML log-likelihoods,

$$G^2 = 2(\widehat{l}_{\text{full}} - \widehat{l}_{\text{red}}),$$

and comparing the statistic to percentiles from a chi-squared distribution with degrees of freedom equal to the difference between the number of covariance parameters in the full and reduced models.

In general, the likelihood ratio test provides a valid method for comparing nested models for the covariance. However, in certain cases the likelihood ratio test may not be valid, depending on the nature of the null hypothesis that is being tested. In particular, when the likelihood ratio test is testing a null hypothesis that is "on the boundary of the parameter space," the usual conditions required for classical likelihood theory no longer apply. What is meant by testing a null hypothesis that is "on the boundary of the parameter space"? This rather technical point is best illustrated by considering variances. Recall that variances cannot be negative, they must be positive. As a result variances are considered to be bounded from 0 to ∞. Thus a likelihood ratio test of the null hypothesis that a variance is zero is testing a null hypothesis that is "on the boundary of the parameter space" for a variance. One consequence is that the usual null distribution for the likelihood ratio test is no longer valid. That is, 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; instead, the null distribution is a mixture of chi-squared distributions. To illustrate the problem, consider the following model where the mean depends on a combination of population parameters, β, and a single individual-specific random effect:

$$Y_{ij} = X'_{ij}\beta + b_i + \epsilon_{ij},$$

where b_i is a random effect and ϵ_{ij} is a within-individual measurement error, with variances σ_b^2 and σ^2, respectively. This model induces marginally a compound symmetry covariance subject to the constraint that

$$\rho = \text{Corr}(Y_{ij}, Y_{ik}) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2} \geq 0.$$

Note that a test of the null hypothesis, H_0: $\sigma_b^2 = 0$ versus H_A: $\sigma_b^2 > 0$, is equivalent to a test of the null hypothesis, H_0: $\rho = 0$ versus H_A: $\rho > 0$, (subject to the constraint that ρ is non-negative). Under the null hypothesis, the repeated measures are assumed to be uncorrelated; under the alternative hypothesis, they are assumed to be positively correlated. In both instances, the null hypothesis is testing on the boundary of the parameter space (i.e., testing that a variance is zero or testing that a non-negative correlation is zero). As a result the null distribution of the likelihood ratio test is not a chi-squared distribution with 1 degree of freedom. Instead, it is an equally weighted mixture of chi-squared distributions with 0 and 1 degrees of freedom (a chi-squared distribution with 0 degrees of freedom has all of its mass or probability at zero). Some intuition for why it is a mixture of chi-squared distributions with 0 and 1 degrees of freedom can be obtained by considering the fit of the model to the data under the null hypothesis. When H_0: $\rho = 0$ is true, the fit of the model to the data is equally likely to show some evidence of positive or negative correlation among the responses due to sampling variability. When there is evidence of positive correlation, $\widehat{\rho}$ will be positive, but when there is evidence of negative correlation, $\widehat{\rho}$ will be zero (since under the alternative hypothesis, H$_A$: $\rho > 0$, ρ is constrained to be non-negative). When H$_0$: $\rho = 0$ is true, there is a 50:50 chance that $\widehat{\rho} > 0$ (or $\widehat{\rho} = 0$). As a result $\widehat{\rho}$ only makes contributions to the likelihood ratio test statistic approximately half of the time, when $\widehat{\rho}$ is positive. The distribution of the likelihood ratio test statistic can be thought of as chi-squared with 1 degree of freedom half of the time, when $\widehat{\rho}$ is positive (and chi-squared with 0 degrees of freedom, when $\widehat{\rho}$ is zero, the other half of the time).

A more detailed discussion of the null distribution of the likelihood ratio test under non-standard conditions is beyond the scope of this book. However, the reader should be aware that the comparison of models for the covariance can sometimes be a non-standard problem. In general, when testing a null hypothesis that is on the boundary of the parameter space, the usual null distribution for the likelihood ratio test is no longer valid. If this problem is simply ignored, and the standard null distribution is naively used, the resulting p-value for the likelihood ratio test will be overestimated (i.e., a p-value that is too large will be obtained). Consequently failure to account for this problem can lead to the selection of a model for the covariance that is too parsimonious. That is, there is a danger that the model for the covariance is too simple and ignores some inherent structure in the covariance. Because it is not straightforward to determine the correct null distribution for the likelihood ratio test in these non-standard settings, we recommend the use of $\alpha = 0.1$, instead of $\alpha = 0.05$, when judging the statistical significance of the likelihood ratio test. Use of the $\alpha = 0.1$ level is a somewhat *ad hoc* solution but protects against selection of a model for the covariance that is too parsimonious. Alternatively, for cases where the null distribution is a known 50:50 mixture of chi-squared distributions, the critical values given in Table C.1 in Appendix C can be used (see Section 8.5 for additional discussion of this topic).

Often it is of interest to compare non-nested models for the covariance. To compare non-nested models, an alternative approach is the Akaike Information Criterion (AIC). According to the AIC, given a set of competing models for the covariance, one should select the model that minimizes

$$\begin{aligned}\text{AIC} &= -2(\text{maximized log-likelihood}) + 2(\text{number of parameters})\\ &= -2(\widehat{l} - c),\end{aligned}$$

where $\widehat{l}$ is the maximized REML log-likelihood and c is the number of covariance parameters. Note that AIC can similarly be defined as selecting the model that minimizes

$$\text{AIC} = -\widehat{l} + c,$$

or the model that maximizes

$$\text{AIC} = \widehat{l} - c.$$

Although AIC can be defined in a number of different ways, the basic underlying idea behind AIC is to strike a balance between two competing objectives: the covariance model must be sufficiently complex to provide a good fit to the data, but at the same time a premium is attached to a parsimonious model. This is achieved by extracting a penalty for the estimation of each additional covariance parameter. With these definitions of AIC, it can be used to compare models with the same fixed effects (i.e., the same model for the mean), but different models for the covariance. Note that expanding the definition of c to include the number of fixed effects parameters, β, would not alter the selection of the model for the covariance provided the model for the mean is held constant.

We note that AIC is but one of a variety of different "information criteria" that have been proposed. Another criterion is the Bayesian Information Criterion (BIC). According to the BIC, given a set of competing models for the covariance, one should select the model that minimizes

$$\begin{aligned} \text{BIC} &= -2(\text{maximized log-likelihood}) + \log N^*(\text{number of parameters}) \\ &= -2(\widehat{l} - \log \sqrt{N^*}\, c), \end{aligned}$$

where N^* is the number of subjects. The BIC is sometimes defined where N^* is the number of "effective subjects," N in the case of ML estimation and $N - p$ in the case of REML estimation (where p is the dimension of β). The main idea underlying BIC requires some understanding of the Bayesian approach to model selection where the objective is to choose the model that has the highest posterior probability (or largest Bayes factor). While this is a legitimate model selection criterion, it must be emphasized that BIC only approximates this Bayesian criterion; furthermore the BIC extracts a very large penalty for the estimation of each additional covariance parameter. In general, we do not recommend the use of BIC for covariance model selection as it entails a high risk of selecting a model that is too simple or parsimonious for the data at hand.

Finally, as mentioned earlier, inferences about β depend on the correct specification of the model for the covariance. Recall that confidence intervals and tests of hypotheses concerning components of β rely on standard errors that are obtained by substituting the REML estimate of Σ_i in the expression for $\text{Cov}(\widehat{\beta})$ (see Eq. (4.5) in Section 4.2). Any misspecification of the model for the covariance has negligible impact on the estimates of the regression coefficients; that is, the regression parameter estimates are unbiased even when the covariance has been misspecified. However, misspecification of the covariance results in incorrect standard errors, and this can lead to potentially misleading inferences concerning β (e.g., due to confidence intervals that are too narrow or wide and p-values that are too small or large). Fortunately, in many cases, valid standard errors for $\widehat{\beta}$ can be obtained when there is concern about misspecification of the covariance. In particular, valid standard errors for $\widehat{\beta}$ can be based on the "sandwich" estimator of $\text{Cov}(\widehat{\beta})$; these standard errors are robust to any misspecification of the covariance. Although the "sandwich" estimator is more widely used in the marginal models for discrete longitudinal data that are the focus of Chapters 12 and 13, we note that the "sandwich" estimator of $\text{Cov}(\widehat{\beta})$ can be applied also in the linear models for longitudinal continuous data described in Part II. The "sandwich" estimator of $\text{Cov}(\widehat{\beta})$ will be discussed in greater detail in Chapter 13.

Table 7.1 Estimated unstructured covariance matrix for the strength data at baseline (day 0), day 4, day 6, day 8, and day 12 from the exercise therapy trial.

Day	0	4	6	8	12
0	9.668	10.175	8.974	9.812	9.407
4	10.175	12.550	11.091	12.580	11.928
6	8.974	11.091	10.642	11.686	11.101
8	9.812	12.580	11.686	13.990	13.121
12	9.407	11.928	11.101	13.121	13.944

7.6 CASE STUDY

Next we illustrate the main ideas by considering covariance pattern models for data from a trial examining the effectiveness of two different exercise therapy regimens.

Exercise Therapy Trial

In this study, subjects were assigned to one of two weightlifting programs to increase muscle strength. In the first program, hereafter referred to as treatment 1, the number of repetitions of the exercises was increased as subjects became stronger. In the second program, hereafter referred to as treatment 2, the number of repetitions was held constant but the amount of weight was increased as subjects became stronger. Measurements of muscle strength were taken at baseline and on days 2, 4, 6, 8, 10, and 12. However, to illustrate some of the main differences among the covariance models considered earlier, we focus only on measures of strength obtained at baseline (or day 0) and on days 4, 6, 8, and 12.

Before considering models for the covariance, it is necessary to choose a maximal model for the mean response. Here, with a balanced design on time and only two groups, we chose the maximal model to be the saturated model for the mean, with a total of 10 parameters for the response profiles for the two treatment groups.

First, we consider an unstructured covariance matrix, with all 15 of its elements unconstrained. The estimated covariance and correlation matrices are displayed in Tables 7.1 and 7.2, respectively. Note that the variance is larger by the end of the study when compared to the variance at baseline; this is a characteristic pattern observed in many longitudinal studies. Furthermore, from examination of Table 7.2, the correlations decrease as the time separation between the repeated measures increases.

Despite the apparent increase in the variance over time, we consider an autoregressive model for the covariance. This model is very parsimonious, with only two

Table 7.2 Estimated unstructured correlation matrix for the strength data at baseline (day 0), day 4, day 6, day 8, and day 12 from the exercise therapy trial.

Day	0	4	6	8	12
0	1.0000	0.9237	0.8847	0.8437	0.8102
4	0.9237	1.0000	0.9597	0.9494	0.9017
6	0.8847	0.9597	1.0000	0.9577	0.9113
8	0.8437	0.9494	0.9577	1.0000	0.9394
12	0.8102	0.9017	0.9113	0.9394	1.0000

parameters, one describing the variance, σ^2, the other the correlation, ρ. When a first-order autoregressive model is fit to the data, it results in the following estimates of the variance and correlation parameters, $\widehat{\sigma}^2 = 11.87$ and $\widehat{\rho} = 0.94$. The resulting estimated pairwise correlations among the five repeated measurements are given in Table 7.3. This model was fit primarily for illustrative purposes; the model is not very appropriate for these data as they are unequally spaced over time (i.e., there is a four-day interval between the first two repeated measures and the last two repeated measures, but all other adjacent repeated measurements were taken two days apart). In order to account for the unequal time interval, an exponential model for the covariance was considered, where

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho^{|t_{ij} - t_{ik}|}$$

for $t_{i1} = 0, t_{i2} = 4, t_{i3} = 6, t_{i4} = 8$, and $t_{i5} = 12$ for all subjects. This resulted in the following estimates of the variance and correlation parameters, $\widehat{\sigma}^2 = 11.87$ and $\widehat{\rho} = 0.98$. The resulting estimated pairwise correlations among the five repeated measurements are given in Table 7.4. Of note, the declines in the estimated correlations in Tables 7.3 and 7.4 are too fast when compared to the corresponding declines in Table 7.2.

Next we consider the choice among these covariance pattern models. The maximized REML log-likelihood and AIC for each of the covariance pattern models are displayed in Table 7.5. Note that there is a hierarchy among the models. The autoregressive and exponential models are both nested within the unstructured covariance. That is, if either the autoregressive or exponential model holds, then the unstructured covariance must necessarily hold. Comparisons of the autoregressive and exponential models with the unstructured covariance can be made using (REML) likelihood ratio tests. However, the autoregressive and exponential models are not nested models; indeed, both models have the same number of parameters. As a result any comparison between these two models can be made directly in terms of their maximized log-likelihoods, since any penalty extracted by information criteria will be the same in both cases (e.g., with AIC a penalty of 4 is extracted for the estimation of the two

Table 7.3 Estimated autoregressive correlation matrix for the strength data at baseline (day 0), day 4, day 6, day 8, and day 12 from the exercise therapy trial.

Day	0	4	6	8	12
0	1.0000	0.9402	0.8839	0.8311	0.7813
4	0.9402	1.0000	0.9402	0.8839	0.8311
6	0.8839	0.9402	1.0000	0.9402	0.8839
8	0.8311	0.8839	0.9402	1.0000	0.9402
12	0.7813	0.8311	0.8839	0.9402	1.0000

Table 7.4 Estimated exponential correlation matrix for the strength data at baseline (day 0), day 4, day 6, day 8, and day 12 from the exercise therapy trial.

Day	0	4	6	8	12
0	1.0000	0.9169	0.8780	0.8408	0.7709
4	0.9169	1.0000	0.9576	0.9169	0.8408
6	0.8780	0.9576	1.0000	0.9576	0.8780
8	0.8408	0.9169	0.9576	1.0000	0.9169
12	0.7709	0.8408	0.8780	0.9169	1.0000

covariance parameters). The likelihood ratio test, comparing the autoregressive and unstructured covariance, yields

$$G^2 = 621.1 - 597.3 = 23.8,$$

and can be compared to a chi-squared distribution with 13 (or 15 − 2) degrees of freedom. On the basis of the likelihood ratio test there is evidence that the autoregressive model does not provide an adequate fit to the covariance, when compared to the unstructured covariance ($p < 0.05$). On the other hand, the likelihood ratio test, comparing the exponential and unstructured covariance, yields

$$G^2 = 618.5 - 597.3 = 21.2,$$

Table 7.5 Comparison of the maximized (REML) log-likelihoods and AIC for the covariance pattern models for the strength data from the exercise therapy trial.

Covariance Pattern Model	−2 (REML) Log-Likelihood	AIC
Unstructured	597.3	627.3
Autoregressive	621.1	625.1
Exponential	618.5	622.5

and when compared to a chi-squared distribution with 13 degrees of freedom, $p > 0.05$. Thus the exponential covariance provides an adequate fit to the data. Also, in terms of AIC, the exponential model minimizes this criterion.

7.7 DISCUSSION: STRENGTHS AND WEAKNESSES OF COVARIANCE PATTERN MODELS

The defining feature of covariance pattern models is that they attempt to account for all the potential sources of variability that have an impact on the covariance among repeated measures on the same individual. That is, they do not distinguish between-subject and within-subject sources of variability. Covariance pattern models characterize the covariance among longitudinal data with a relatively small number of parameters. Many of the models (e.g., autoregressive, Toeplitz, and banded) are only appropriate when the repeated measurements are obtained at equal intervals and cannot handle irregularly timed measurements. Although there is a large selection of models for the correlations, the choice of models for the variances is somewhat limited. Covariance pattern models either make the strong assumption that the variances are constant over time, or relax this assumption entirely and allow the variances to depend arbitrarily on time.

For the most part, covariance pattern models are appropriate for balanced longitudinal designs, and many require that the repeated measurements are obtained at equal intervals. Although these models can handle imbalance due to missing data at any of the fixed occasions, they are not well suited for modeling data from inherently unbalanced longitudinal designs. With inherently unbalanced designs, many of the covariance pattern models are not well defined. In an attempt to overcome the latter limitation, a few covariance pattern models have been developed that allow for irregularly timed measurements (e.g., the exponential covariance pattern model). In these models the correlation is assumed to depend on the time separation between pairs of repeated measurements. However, a potential problem with these models is that they assume the correlation decays rapidly with increasing time separation and that the correlation between two measurements taken at the same occasion is one. As

Table 7.6 Covariance pattern modeling options using PROC MIXED in SAS.

TYPE =	<pattern>	Specifies the covariance pattern
	UN	Unstructured
	CS	Compound symmetry
	AR(1)	First-order autoregressive
	TOEP	Toeplitz
	UN(n)	Banded unstructured, with n bands
	CSH	Heterogeneous compound symmetry
	ARH(1)	Heterogeneous first-order autoregressive

mentioned earlier, in our experience with longitudinal studies in the health sciences, the correlation among repeated measures rarely exhibits either of these two characteristics. Furthermore, although these covariance pattern models allow the correlation to depend on the time separation between repeated measurements, they do not allow the variances to depend on time. As a result they make the strong and often unrealistic assumption that the variance remains constant over time.

In conclusion, covariance pattern models are appropriate for balanced longitudinal designs and many models require that the repeated measurements are obtained at equal intervals. In general, we do not recommend the use of covariance pattern models that make the strong assumption that the variances are constant over time. As mentioned earlier, our practical experience with many longitudinal studies has led to the empirical observation that the variances are rarely constant over time. Because the assumption of constant variance is the one that is not valid in many setting, we recommend that covariance pattern models with heterogeneous variances, allowing the variances to depend arbitrarily on time, should generally be adopted.

7.8 COMPUTING: FITTING COVARIANCE PATTERN MODELS USING PROC MIXED IN SAS

In the following we assume that there is a single group factor and the maximal model is the saturated model for the mean. Different patterns can be fit to the covariance matrix among the residuals, denoted R in PROC MIXED in SAS, by using the TYPE= option on the REPEATED statement. Table 7.6 provides a summary of some of the commonly used covariance pattern models; a full description of all of the options can be found in the SAS documentation.

For example, to fit an autoregressive model for the covariance we can use the illustrative SAS commands given in Table 7.7. The options R and RCORR on the

Table 7.7 Illustrative commands for an autoregressive model using PROC MIXED in SAS.

```
PROC MIXED;
    CLASS id group time;
    MODEL y=group time group*time / S CHISQ;
    REPEATED time / TYPE=AR(1) SUBJECT=id R RCORR;
```

Table 7.8 Illustrative commands for an exponential model using PROC MIXED in SAS.

```
PROC MIXED;
    CLASS id group time;
    MODEL y=group time group*time / S CHISQ;
    REPEATED time / TYPE=SP(EXP)(ctime) SUBJECT=id R RCORR;
```

REPEATED statement request that the estimated covariance matrix (R) and the corresponding correlation matrix be displayed as part of the output. By default, the covariance and correlation matrices are displayed for the first subject and will have row and column dimensions corresponding to the number of repeated measures obtained on the first subject. When the vector of responses on the first subject is incomplete, it may be preferable to display the covariance and correlation matrices for a subject with complete responses. The options R=1, 5, 7 and RCORR=1, 5, 7 request that the estimated covariance and correlation matrices be displayed for the first, fifth, and seventh subjects.

To fit covariance pattern models to inherently unbalanced data requires the use of the "spatial" covariance pattern options in PROC MIXED. These are covariance pattern models developed for spatial data that are defined in terms of "distances" in two-dimensional space. However, these options can also be used where "distance" (or time separation) is defined along the single dimension of time. For example, to fit an exponential covariance pattern model, the following option is used:

```
TYPE = SP(EXP)(list)
```

where list is the name of the variable used to construct "distances" or time separation between repeated measurements. Table 7.8 contains illustrative commands for fitting an exponential covariance pattern model. Note that the variable `ctime` is simply an additional copy of `time` that is treated as a continuous covariate for the purpose of constructing the time separation between repeated measurements.

Finally, when the EMPIRICAL option is included on the PROC MIXED statement standard errors for $\widehat{\beta}$ are based on the "sandwich" estimator of $\text{Cov}(\widehat{\beta})$. As mentioned earlier, these standard errors are robust to any misspecification of the model for the covariance. The "sandwich" estimator of $\text{Cov}(\widehat{\beta})$ will be discussed in Chapter 13.

7.9 FURTHER READING

Additional discussion of covariance pattern models can be found in Chapter 6, Section 6.2, of Brown and Prescott (1999) and in the tutorial by Littell et al. (2000).

Bibliographic Notes

Jennrich and Schluchter (1986) describe covariance pattern models for longitudinal data. For a more recent and comprehensive overview of this topic, see the review article by Zimmerman and Nunez-Anton (2001), and the references therein. Finally, Pourahmadi (1999) presents a flexible approach for parametric modeling of the covariance structure.

Altham (1984) discusses the advantages, in terms of increased precision of estimation of the parameters of interest, that can result from fitting a parsimonious model to complex data. Altham's (1984) general discussion of this issue has great relevance for the modeling of the covariance in longitudinal data.

The large-sample distribution theory for testing a null hypothesis that is "on the boundary of the parameter space" (e.g., testing that a variance is zero) is discussed in Miller (1977), Self and Liang (1987), Stram and Lee (1994, 1995), Silvapulle and Silvapulle (1995), Silvapulle (1996), and Verbeke and Molenberghs (2003).

Problems

7.1 In a study of dental growth, measurements of the distance (mm) from the center of the pituitary gland to the pteryomaxillary fissure were obtained on 11 girls and 16 boys at ages 8, 10, 12, and 14 (Potthoff and Roy, 1964).

The raw data are stored in an external file: `dental.dat`

Each row of the data set contains the following six variables:

ID Gender Y_1 Y_2 Y_3 Y_4

Note: The categorical (character) variable Gender is coded F = Female, M = Male. The third measure (at age 12) on subject ID = 20 is a potential outlier.

7.1.1 On a single graph, construct a time plot that displays the mean distance (mm) versus age (in years) for boys and girls. Describe the time trends for boys and girls.

7.1.2 Read the data from the external file and put the data in a "univariate" or "long" format, with four "records" per subject.

7.1.3 For the "maximal" model, assume a saturated model for the mean response. Fit the following models for the covariance:

(a) unstructured covariance

(b) compound symmetry

(c) heterogeneous compound symmetry

(d) autoregressive

(e) heterogeneous autoregressive

Choose a model for the covariance that adequately fits the data.

7.1.4 Given the choice of model for the covariance from Problem 7.1.3, treat age (or time) as a categorical variable and fit a model that includes the effects of age, gender, and their interactions. Determine whether the pattern of change over time is different for boys and girls.

7.1.5 Show how the *estimated* regression coefficients from Problem 7.1.4 can be used to estimate the means in the two groups at ages 8 and 14.

7.1.6 Given the choice of model for the covariance from Problem 7.1.3, treat age as a continuous variable and fit a model that includes the effects of a linear trend in age, gender, and their interaction. Compare and contrast the results with those obtained in Problem 7.1.4.

7.1.7 On a single graph, construct a time plot that displays the *estimated* mean distance (mm) versus age (in years) for boys and girls from the results generated from Problem 7.1.6.

7.1.8 Show how the regression coefficients from Problem 7.1.6 can be used to estimate the means in the two groups at ages 8 and 14.

7.1.9 Does a model with only a linear trend in age adequately account for the pattern of change in the two groups?

7.1.10 The third measure (at age 12) on subject ID = 20 is a potential outlier. Repeat the analyses in Problems 7.1.3, 7.1.4, 7.1.6 and 7.1.9 excluding the third measure on subject ID = 20. Do the substantive conclusions change?

7.1.11 Given the results of all the previous analyses, what conclusions can be drawn about gender differences in patterns of dental growth?