4 Introduction to modeling longitudinal data

We are now in a position to introduce a basic statistical model for longitudinal data. The models and methods we discuss in subsequent chapters may be viewed as modifications of this model to incorporate specific assumptions on sources of variation and the form of mean vectors.

We restrict our discussion here to the case of **balanced data**; i.e., where all units have repeated measurements at the same n time points. Later, we will extend our thinking to handle the case of **unbalanced data**.

4.1 Basic Statistical Model

Recall that the longitudinal (or more general repeated measurement data) situation involves observation of the same response repeatedly over time (or some other condition) for each of a number of units (individuals).

- In the simplest case, the units may be a **random sample** from a **single population**.
- More generally, the units may arise from **different populations**. Units may be randomly assigned to different treatments or units may be of different types (e.g. male and female).
- In some cases, additional information on individual-unit characteristics like age and weight may be recorded.

We first introduce a fundamental model for balanced longitudinal data for a single sample from a common population, and then discuss how it may be adapted to incorporate these more general situations.

MOST BASIC MODEL FOR BALANCED DATA: Suppose the response of interest is measured on each individual at n times $t_1 < t_2 < \cdots < t_n$. The dental study ($n = 4$; $t_1, \ldots, t_4 = 8, 10, 12, 14$) and the guinea pig diet data ($n = 6$; $t_1, \ldots, t_6 = 1, 3, 4, 5, 6, 7$) are balanced data sets (with units coming from more than one population).

Consider the case where all the units are from a **single population** first. Corresponding to each t_j, $j = 1, \ldots, n$, there is a random variable Y_j, $j = 1, \ldots, n$, with a probability distribution that summarizes the way in which responses at time t_j among all units in the population take on their possible values.

As we discuss in detail shortly, values of the response at any time t_j may **vary** due to the effects of relevant **sources of variation**.

We may think of the generic **random vector**

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \tag{4.1}$$

where the variables are arranged in **increasing time order**.

- $\boldsymbol{Y}$ in (4.1) has a multivariate probability distribution summarizing the way in which all responses at times $t_1, \ldots, t_n$ among all units in the population take on their possible values jointly.
- This probability distribution has mean vector $E(\boldsymbol{Y}) = \boldsymbol{\mu}$ with elements $\mu_j = E(Y_j)$, $j = 1, \ldots, n$, and covariance matrix $\text{var}(\boldsymbol{Y}) = \boldsymbol{\Sigma}$.

CONVENTION: Except when we discuss "classical" methods in the next two chapters, we will use i as the subscript indexing units and j as the subscript indexing responses in time order within units.

We will also use m to denote the total number of units (across groups where relevant). E.g. for the dental study and guinea pig diet data, $m = 27$ and $m = 15$, respectively.

Thus, in thinking about a random sample of units from a single population of interest, just as we do for scalar response, we may thus think of m $(n \times 1)$ random vectors

$$\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_m,$$

corresponding to each of m individuals, each of which has features (e.g. multivariate probability distribution) identical to $\boldsymbol{Y}$ in (4.1).

For the ith such vector,

$$\boldsymbol{Y}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix},$$

such that

$$E(\boldsymbol{Y}_i) = \boldsymbol{\mu}, \quad \text{var}(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}.$$

- It is natural to be concerned that components Y_{ij}, $j = 1, \ldots, n$, are **correlated**.
- In particular, this may be due to the simple fact that observations on the same unit may tend to be "more alike" than those compared across different units; e.g. a guinea pig with "low" weight at any given time relative to other pigs will likely be "low" relative to other pigs at any other time.
- Alternatively, correlation may be due to biological "fluctuations" within a unit, as in the pine seedling example of the last chapter.

We will discuss these sources of variation for longitudinal data shortly. For now, it is realistic to expect that

$$\text{cov}(Y_{ij}, Y_{ik}) \neq 0 \;\text{ for any }\; j \neq k = 1, \ldots, n.$$

in general, so that $\boldsymbol{\Sigma}$ is unlikely to be a diagonal matrix.

INDEPENDENCE ACROSS UNITS: On the other hand, if each $\boldsymbol{Y}_i$ corresponds to a different individual, and individuals are not related in any way (e.g. different children or guinea pigs, treated and handled separately), then it seems reasonable to suppose that the way any observation may turn out at any time for unit i is unrelated to the way any observation may turn out for another unit $\ell \neq i$; that is, observations from different vectors are independent.

- Under this view, the random vectors $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_m$ are all mutually independent.
- It follows that if Y_{ij} is a response from unit i and $Y_{\ell k}$ is a response from unit ℓ, $\text{cov}(Y_{ij}, Y_{\ell k}) = 0$ even if $j = k$ (same time point but different units).

BASIC STATISTICAL MODEL: Putting all this together, we have m mutually independent random vectors $\boldsymbol{Y}_i$, $i = 1, \ldots, m$, with $E(\boldsymbol{Y}_i) = \boldsymbol{\mu}$ and $\text{var}(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}$.

- We may write this model equivalently similarly to the univariate case; specifically,

$$\boldsymbol{Y}_i = \boldsymbol{\mu} + \boldsymbol{\epsilon}_i, \quad E(\boldsymbol{\epsilon}_i) = \mathbf{0}, \;\; \text{var}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}, \tag{4.2}$$

  where the $\boldsymbol{\epsilon}_i$, $i = 1, \ldots, m$, are mutually independent.

- $\boldsymbol{\epsilon}_i$ are **random vector deviations** such that $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{in})'$, where each ϵ_{ij}, $j = 1, \ldots, n$, $E(\epsilon_{ij}) = 0$ represents how Y_{ij} deviates from its mean μ_j due to aggregate effects of sources of variation.

- In addition, the ϵ_{ij} are **correlated**, but $\boldsymbol{\epsilon}_i$ are mutually independent across i.

Questions of scientific interest are characterized as questions about the elements of $\boldsymbol{\mu}$, as will be formalized in later chapters.

MULTIVARIATE NORMALITY: If the response is continuous, it may be reasonable to assume that the Y_{ij} and ϵ_{ij} are normally distributed. In this case, adding the further assumption that $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, (4.2) implies

$$\boldsymbol{Y}_i \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad i = 1, \ldots, m,$$

where the $\boldsymbol{Y}_i$ are mutually independent.

EXTENSION TO MORE THAN ONE POPULATION: Suppose that individuals may be thought of as sampled randomly from q different populations; e.g. $q = 2$ (males and females) in the dental study.

- We may again think of $\boldsymbol{Y}_i$, m independent random vectors, where, if $\boldsymbol{Y}_i$ corresponds to a unit from group ℓ, $\ell = 1, \ldots, q$, then $\boldsymbol{Y}_i$ has a multivariate probability distribution with

  $$E(\boldsymbol{Y}_i) = \boldsymbol{\mu}_\ell, \quad \text{var}(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}_\ell.$$

  That is, each population may have a different mean vector and covariance matrix.

- Equivalently, we may express this as

  $$\boldsymbol{Y}_i = \boldsymbol{\mu}_\ell + \boldsymbol{\epsilon}_i, \quad E(\boldsymbol{\epsilon}_i) = \mathbf{0}, \quad \text{var}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_\ell \text{ for } i \text{ from group } \ell = 1, \ldots, q.$$

- We might also assume $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\ell)$ for units in group ℓ, so that

  $$\boldsymbol{Y}_i \sim \mathcal{N}(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell)$$

  for i from group ℓ.

- If furthermore it is reasonable to assume that all sources of variation act similarly in each population, we might assume that $\boldsymbol{\Sigma}_\ell = \boldsymbol{\Sigma}$, a **common covariance matrix** for all populations.

  With univariate responses, it is often reasonable to assume that population membership may imply a change in mean response but not affect the nature of variation; e.g. the primary effect of a treatment may be to shift responses on average relative to those for another, but to leave variability unchanged. This reduces to the assumption of **equal variances**.

  For the longitudinal case, such an assumption may also be reasonable, but is more involved, as assuming the same "variation" in all groups must take into account both **variance** and **covariation**.

- Under this assumption, the model becomes

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_\ell + \boldsymbol{\epsilon}_i, \quad E(\boldsymbol{\epsilon}_i) = \mathbf{0}, \quad \text{var}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma} \text{ for } i \text{ from group } \ell = 1, \ldots, q,$$

  for a covariance matrix $\boldsymbol{\Sigma}$ common to all groups.

- Note that even though $\boldsymbol{\Sigma}$ is common to all populations, the diagonal elements of $\boldsymbol{\Sigma}$ may be different across $j = 1, \ldots, n$, so that variance may be different at different times; however, at any given time, the variance is the same for all groups.

- Similarly, the covariances in $\boldsymbol{\Sigma}$ between the jth and kth elements of $\boldsymbol{Y}_i$ may be different for different choices of j and k, but for any particular pair (j, k), the covariance is the same for all groups.

EXTENSION TO INDIVIDUAL INFORMATION: We may extend this thinking to take into account other individual **covariate** information besides population membership by analogy to regression models for univariate response.

- E.g., suppose age a_i at the first time point is recorded for each unit $i = 1, \ldots, m$.

- We may envision for each age a_i a multivariate probability distribution describing the possible values of $\boldsymbol{Y}_i$. The **mean vector** of this distribution would naturally depend on a_i.

- We write this for now as $E(\boldsymbol{Y}_i) = \boldsymbol{\mu}_i$, where $\boldsymbol{\mu}_i$ is the mean of random vectors from the population corresponding to age a_i, and the subscript i implies that the mean is "unique" to i in the sense that it depends on a_i somehow.

- Assuming that variation is similar regardless of age, we may write

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i, \quad E(\boldsymbol{\epsilon}_i) = \mathbf{0}, \quad \text{var}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}.$$

  We defer discussion of how dependence of $\boldsymbol{\mu}_i$ on a_i (and other factors) might be characterized to later chapters.

All of the foregoing models represent random vectors $\boldsymbol{Y}_i$ in terms of a **mean vector** plus a **random deviation vector** $\boldsymbol{\epsilon}_i$ that captures the aggregate effect of all sources of variation. This emphasizes the two key aspects of modeling longitudinal data:

(1) Characterizing mean vectors in these models in a way that best captures how mean response changes with time and depends on other factors, such as group or age, in order to address questions of scientific interest;

(2) Taking into account important sources of variation by characterizing the nature of the random deviations $\boldsymbol{\epsilon}_i$, so that these questions may be addressed by taking faithful account of all variation in the data.

Models we discuss in subsequent chapters may be viewed as particular cases of this representation, where (1) and (2) are approached differently.

We first take up the issue in (2), that of the sources of variation that $\boldsymbol{\epsilon}_i$ may reflect.

4.2 Sources of variation in longitudinal data

For longitudinal data, potential sources of variation usually are thought of as being of two main types:

- **Among-unit** variation
- **Within-units** variation.

It is useful to conceptualize the way in which longitudinal response vectors may be thought to arise. There are different perspectives on this; here, we consider one popular approach. For simplicity, consider the case of a single population and the model

$$\boldsymbol{Y}_i = \boldsymbol{\mu} + \boldsymbol{\epsilon}_i.$$

The ideas are relevant more generally.

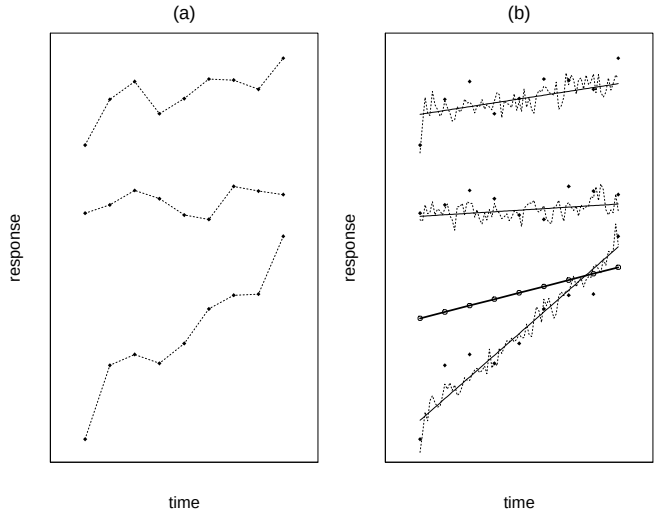
Figure 1 provides a convenient backdrop for thinking about the sources that might make up $\boldsymbol{\epsilon}_i$.

- Panel (a) shows the values actually observed for $m = 3$ units; these values include the effects of all sources of variation.
- Panel (b) is a conceptual representation of possible underlying features of the situation.

  The open circles on the thick, solid line represent the elements of $\boldsymbol{\mu}$ at each of the $n = 9$ time points. E.g., the leftmost circle represents the mean μ_1 of all possible values that could be observed at t_1, thus averaging all deviations ϵ_{i1} due to all among- and within-unit sources over all units i. The means over time lie on a straight line, but this need not be true in general.

  The solid diamonds represent the actual observations for each individual. If we focus on the first time point, for example, it is clear that the observations for each i vary about μ_1.

Figure 1: *(a) Hypothetical longitudinal data from $m = 3$ units at $n = 9$ time points. (b) Conceptual representation of sources of variation. The open circles connected by the thick solid line represent the means μ_j, $j = 1, \ldots, n$ for the populations of all possible observations at each of the n time points. The thin solid lines represent "trends" for each unit. The dotted lines represent the pattern of error-free responses for the unit over time, which fluctuate about the trend. The diamonds represent the observations of these responses, which are subject to measurement error.*



- For each individual, we may envision a "trend," depicted by the solid lines (the trend need not follow a straight line in general). The "trend" places the unit in the population.

  The vertical position of this trend at any time point dictates whether the individual is "high" or "low" relative to the corresponding mean in $\boldsymbol{\mu}$. Thus, these "trends" highlight (biological) variation **among** units.

  Some units may be consistently "high" or "low," others may be "high" at some times and "low" at others relative to the mean.

- The dotted lines represent "fluctuations" about the smoother (straight-line) trend, representing variation in how responses for that individual may evolve. In the pine seedling example cited earlier, with response height of a growing plant over time, although the overall pattern of growth may "track" a smooth trend, natural variation in the growth process may cause the responses to **fluctuate** about the trend.

  This phenomenon necessarily occurs **within** units; (biological) fluctuations about the trend are the result of processes taking place only **within** that unit.

Note that values on the dotted line that are very close in time tend to be "larger" or "smaller" than the trend together, while those farther apart seem just as likely to be larger or smaller than the trend, with no relationship.

- Finally, the observations for a unit (diamonds) do not lie exactly on the dotted lines, but vary about them. This is due to **measurement error**. Again, such errors take place **within** the unit itself in the sense that the measuring process occurs at the specific-unit level.

We may formalize this thinking by refining how we view the basic model $\boldsymbol{Y}_i = \boldsymbol{\mu} + \boldsymbol{\epsilon}_i$. The jth element of $\boldsymbol{Y}_i$, Y_{ij}, may be thought of as being the sum of several components, each corresponding to a different source of variation; i.e.

$$Y_{ij} = \mu_j + \epsilon_{ij} = \mu_j + b_{ij} + e_{ij} = \mu_j + b_{ij} + e_{1ij} + e_{2ij}, \tag{4.3}$$

where $E(b_{ij}) = 0$, $E(e_{1ij}) = 0$, and $E(e_{2ij}) = 0$.

- b_{ij} is a deviation representing **among unit** variation at time t_j due to the fact that unit i "sits" somewhere in the population relative to μ_j due to **biological variation**.

  We may think of b_{ij} as dictating the "inherent trend" for i at t_j.

- e_{1ij} represents the additional deviation due to **within-unit fluctuations** about the trend.

- e_{2ij} is the deviation due to measurement error (**within-units**).

- The sum $e_{ij} = e_{1ij} + e_{2ij}$ denotes the aggregate deviation due to all **within-unit sources**.

- The sum $\epsilon_{ij} = b_{ij} + e_{1ij} + e_{2ij}$ thus represents the **aggregate** deviation from μ_j due to all sources. Stacking the ϵ_{ij}, b_{ij}, and e_{ij}, we may write

  $$\boldsymbol{\epsilon}_i = \boldsymbol{b}_i + \boldsymbol{e}_i = \boldsymbol{b}_i + \boldsymbol{e}_{1i} + \boldsymbol{e}_{2i},$$

  which emphasizes that $\boldsymbol{\epsilon}_i$ includes components due to among- and within-unit sources of variation.

SOURCES OF CORRELATION: This representation provides a framework for thinking about assumptions on among- and within-unit variation and how correlation among the Y_{ij} (equivalently, among the ϵ_{ij}) may be thought to arise.

- The b_{ij} determines the "inherent trend" in the sense that $\mu_j + b_{ij}$ represents position of the "inherent trajectory" for unit i at time j. The Y_{ij} thus all tend to be in the vicinity of this trend across time (j) for unit i. As can be seen from Figure 1, this makes the observations on i "more alike" relative to observations from units.

Accordingly, we expect that the elements of $\boldsymbol{\epsilon}_i$ (and hence those of $\boldsymbol{Y}_i$) are **correlated** due to the fact that they share this common, underlying trend. We may refer to correlation arising in this way as **correlation due to among-unit sources**.

In subsequent chapters, we will see that different longitudinal data models may make specific assumptions about terms like b_{ij} that represent among-unit variation and hence this source of correlation.

- Because e_{1ij} are deviations due to the "fluctuation" process, it is natural to think that the e_{1ij} might be **correlated** across j. If the process is "high" relative to the inherent trend at time t_j (so e_{1ij} is positive), it might be expected to be "high" at times $t_{j'}$ close to t_j ($e_{1ij'}$ positive) as well. Thus, we might expect the elements of $\boldsymbol{\epsilon}_i$ and thus $\boldsymbol{Y}_i$ to be **correlated** as a consequence of such fluctuations (because the elements of $\boldsymbol{e}_{1i}$ are correlated).

  We may refer to correlation arising in this way as **correlation due to within-unit sources**.

  Note that if the fluctuations occur in a very short time span relative to the spacing of the t_j, whether the process is "high" at t_j may have little or no relation to whether it is high at adjacent times. In this case, we might believe such within-unit correlation is **negligible**. As we will see, this is a common assumption, often justified by noting that the t_j are far apart in time.

- The **overall** pattern of correlation for $\boldsymbol{\epsilon}_i$ (and hence $\boldsymbol{Y}_i$) may be thought of as resulting from the **combined effects** of these two sources (among- and within-units).

- As measuring devices tend to commit "haphazard" errors every time they are used, it may be reasonable to assume that the e_{2ij} are **independent** across j. Thus, we expect no contribution to the overall pattern of correlation.

To complete the thinking, we must also consider the **variances** of the b_{ij}, e_{1ij}, and e_{2ij}. We defer discussion of this to later chapters in the context of specific models.

4.3 Exploring mean and covariance structure

The aggregate effect of all sources of variation, such as those identified in the conceptual scheme of Section 4.2, dictates the form of the covariance matrix of $\boldsymbol{\epsilon}_i$ and hence that of $\boldsymbol{Y}_i$.

As was emphasized earlier in our discussion of weighted least squares, if observations are correlated and have possibly different variances, it is important to acknowledge this in estimating parameters of interest such as population means so that differences in data quality and associations are taken into adequate account. Thus, an accurate representation of $\text{var}(\boldsymbol{\epsilon}_i)$ is critically important.

A first step in an analysis is often to examine the data for clues about the likely nature of the form of this covariance matrix as well as the structure of the means and how they change over time.

Consider first the model for a single population

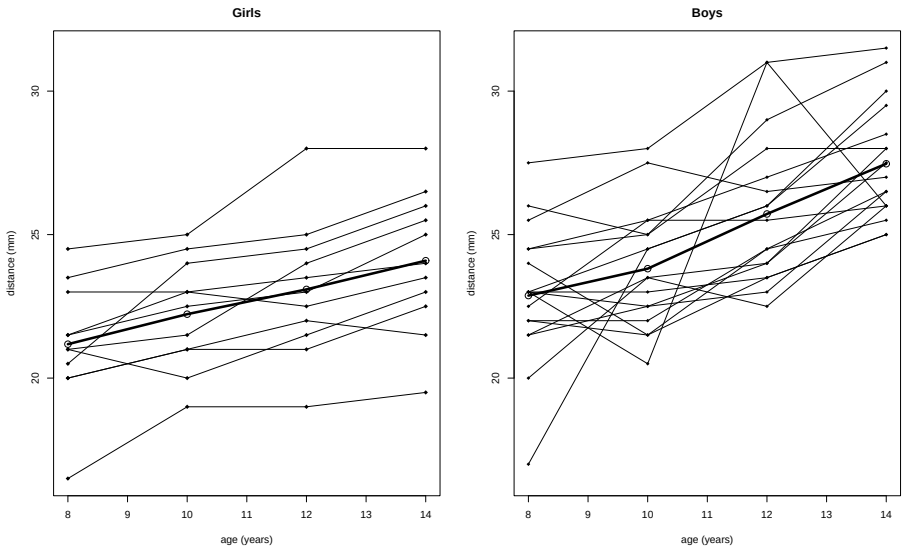
$$\boldsymbol{Y}_i = \boldsymbol{\mu} + \boldsymbol{\epsilon}_i, \quad E(\boldsymbol{\epsilon}_i) = \mathbf{0}, \quad \text{var}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}.$$

Based on observed data, we would like to gain insight into the likely forms of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

- We illustrate with the data for the 11 girls in the dental study, so for now take $m = 11$ and $n = 4$.
- Thus, the μ_j, $j = 1, \ldots, 4$, of $\boldsymbol{\mu}$ are the population mean distance for girls at ages 8, 10, 12, and 14, the diagonal elements of $\boldsymbol{\Sigma}$ are the population variances of distance at each age, and the off-diagonal elements of $\boldsymbol{\Sigma}$ represent the covariances among distances at different ages.

Spaghetti plots for both the boys and girls are given in in Figure 2.

Figure 2: *Spaghetti plots of the dental data. The open circles represent the sample mean distance at each age; these are connected by the thick line to highlight the relationship among means over time.*



SAMPLE MEAN VECTOR: As we have discussed, the natural **estimator** for the mean μ_j at the jth time point is the **sample mean**

$$\overline{Y}_{\cdot j} = m^{-1} \sum_{i=1}^{m} Y_{ij},$$

where the "dot" subscript indicates averaging over the first index i (i.e. across units). The sample mean may be calculated for each time point $j = 1, \ldots, n$, suggesting that the obvious estimator for $\boldsymbol{\mu}$ is the vector whose elements are the $\overline{Y}_{\cdot j}$, the **sample mean vector** given by

$$\overline{\boldsymbol{Y}} = m^{-1} \sum_{i=1}^{m} \boldsymbol{Y}_i = \begin{pmatrix} \overline{Y}_{\cdot 1} \\ \vdots \\ \overline{Y}_{\cdot n} \end{pmatrix}.$$

- It is straightforward to show that the random vector $\overline{\boldsymbol{Y}}$ is an unbiased estimator for $\boldsymbol{\mu}$; i.e.

$$E(\overline{\boldsymbol{Y}}) = \boldsymbol{\mu}.$$

We may apply this estimator to the dental study data on girls to obtain the estimate (rounded to three decimal places)

$$\overline{\boldsymbol{y}} = \begin{pmatrix} 21.182 \\ 22.227 \\ 23.091 \\ 24.091 \end{pmatrix}.$$

In the left panel of Figure 2, these values are plotted for each age by the open circles.

- The thick solid line, which connects the $\overline{Y}_{\cdot j}$, gives a visual impression of a "smooth," indeed straight line, relationship over time among the μ_j.
- Of course, we have no data at ages intermediate to those in the study, so it is possible that mean distance in the intervals between these times deviates from a straight line relationship. However, from a biological point of view, it seems sensible to suppose that dental distance would increase steadily over time, at least on average, rather than "jumping" around.

Graphical inspection of sample mean vectors is an important tool for understanding possible relationships among means over time. When there are $q > 1$ groups an obvious strategy is to carry this out separately for the data from each group, so that possible differences in means can be evaluated.

For the dental data on the 16 boys, the estimated mean turns out to be $\overline{\boldsymbol{y}} = (22.875, 23.813, 25.719, 27.469)'$; this is shown as the thick solid line with open circles in the right panel of Figure 2. This estimate seems to also look like a "straight line," but with steepness possibly different from that for girls.

SAMPLE COVARIANCE MATRIX: Gaining insight into the form of $\mathbf{\Sigma}$ may be carried out both graphically and through an unbiased estimator for $\mathbf{\Sigma}$ and its associated correlation matrix.

- The diagonal elements of $\mathbf{\Sigma}$ are simply the variances σ_j^2 of the distributions of Y_j values at each time $j = 1, \ldots, n$. Thus, based on m units, the natural estimator for σ_j^2 is the **sample variance** at time j,

$$S_j^2 = (m-1)^{-1} \sum_{i=1}^{m} (Y_{ij} - \overline{Y}_{\cdot j})^2,$$

  which may be shown to be an **unbiased estimator** for σ_j^2.

- The off-diagonal elements of $\mathbf{\Sigma}$ are the covariances

$$\sigma_{jk} = E\{(Y_j - \mu_j)(Y_k - \mu_k)\}.$$

  Thus, a natural estimator for σ_{jk} is

$$S_{jk} = (m-1)^{-1} \sum_{i=1}^{m} (Y_{ij} - \overline{Y}_{\cdot j})(Y_{ik} - \overline{Y}_{\cdot k}),$$

  which may also be shown to be **unbiased**.

- The obvious estimator for $\mathbf{\Sigma}$ is thus the matrix in which the variances σ_j^2 and covariances σ_{jk} are replaced by S_j^2 and S_{jk}. It is possible to represent this matrix succinctly (verify) as

$$\widehat{\mathbf{\Sigma}} = (m-1)^{-1} \sum_{i=1}^{m} (\boldsymbol{Y}_i - \overline{\boldsymbol{Y}})(\boldsymbol{Y}_i - \overline{\boldsymbol{Y}})'.$$

  This is known as the **sample covariance matrix**.

- The sum $\sum_{i=1}^{m} (\boldsymbol{Y}_i - \overline{\boldsymbol{Y}})(\boldsymbol{Y}_i - \overline{\boldsymbol{Y}})'$ is often called the **sum of squares and cross-products** (SS&CP) matrix, as its entries are the sums of squared deviations and cross-products of deviations from the sample mean.

- The sample covariance matrix is exactly as we would expect; recall that the covariance matrix itself is defined as

$$\mathbf{\Sigma} = E\{(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})'\}.$$

The sample covariance matrix may be used to estimate the covariance matrix. However, although the diagonal elements may provide information on the true variances at each time point, the off-diagonal elements may be difficult to interpret. Given the unitless nature of correlation, it may be more informative to learn about associations from estimates of **correlation**.

SAMPLE CORRELATION MATRIX: If $\widehat{\mathbf{\Sigma}}$ is an estimator for a covariance matrix $\mathbf{\Sigma}$ with elements $\widehat{\Sigma}_{jk}$, $j, k = 1, \ldots, n$, then the natural estimator for the associated correlation matrix $\mathbf{\Gamma}$ is $\widehat{\mathbf{\Gamma}}$, the $(n \times n)$ matrix $\widehat{\mathbf{\Gamma}}$ with ones on the diagonal (as required for a correlation matrix) and (j, k) off-diagonal element

$$\frac{\widehat{\Sigma}_{jk}}{\sqrt{\widehat{\Sigma}_{jj}\widehat{\Sigma}_{kk}}}.$$

- For a single population, where $\widehat{\mathbf{\Sigma}}$ is the sample covariance matrix, the off-diagonal elements are

$$\frac{S_{jk}}{S_j S_k}, \tag{4.4}$$

  which are obvious estimators for the correlations

$$\rho_{jk} = \frac{\sigma_{jk}}{\sigma_j \sigma_k}.$$

- In this case, the estimated matrix $\widehat{\mathbf{\Gamma}}$ is called the **sample correlation matrix**, as it is an estimate of the correlation matrix corresponding to the sample covariance matrix for the single population.

- The expression in (4.4) is known as the **sample correlation coefficient** between the observations at times t_j and t_k, as it estimates the correlation coefficient ρ_{jk}.

Shortly, we shall see how to estimate common covariance and correlation matrices based on data from several populations.

For the 11 girls in the dental study, we obtain the estimated covariance and correlation matrices (rounded to three decimal places)

$$\widehat{\mathbf{\Sigma}}_G = \begin{pmatrix} 4.514 & 3.355 & 4.332 & 4.357 \\ 3.355 & 3.618 & 4.027 & 4.077 \\ 4.332 & 4.027 & 5.591 & 5.466 \\ 4.357 & 4.077 & 5.466 & 5.941 \end{pmatrix}, \quad \widehat{\mathbf{\Gamma}}_G = \begin{pmatrix} 1.000 & 0.830 & 0.862 & 0.841 \\ 0.830 & 1.000 & 0.895 & 0.879 \\ 0.862 & 0.895 & 1.000 & 0.948 \\ 0.841 & 0.879 & 0.948 & 1.000 \end{pmatrix}.$$

- The diagonal elements of $\widehat{\mathbf{\Sigma}}_G$ suggest that the aggregate variance in dental distances roughly increases over time from age 8 to 14.

  However, keep in mind that the values shown are estimates of the corresponding parameters based on only $m = 11$ observations; thus, they are subject to the usual uncertainty of estimation. It is thus sensible to not "over-interpret" the numbers but rather to only examine them for suggestive features.

- The off-diagonal elements of $\mathbf{\Gamma}$ represent the aggregate pattern of correlation due to **among- and within-girl sources**. Here, the estimate of this correlation for any pair of time points is positive and close to one, suggesting that "high" values at one time are strongly associated with "high" values at another time, regardless of how far apart in time the observations occur.

  In light of Figure 2, this is really not surprising. The data for individual girls in the figure show pronounced trends that for the most part place a girl's trajectory above or below the estimated mean profile (thick line). Thus, a girl such as the topmost one is "high" throughout time, suggesting a strong component of among-girl variation in the population, and the estimates of correlation are likely reflecting this.

- Again, it is not prudent to attach importance to the numbers and differences among them, as they are estimates from a rather small sample, so the observed difference between 0.948 and 0.830 may or may not reflect a real difference in the true correlations.

SCATTERPLOT MATRICES: A useful supplement to numerical estimates is a graphical display of the observed data known as a **scatterplot matrix.**

As correlation reflects associations among observations at different time points, initially one would think that a natural way of graphically assessing these associations would be to make the following plot.

- For each pair of times t_j and t_k, graph the observed data values (y_{ij}, y_{ik}) for all $i = 1, \ldots, m$ units, with y_{ij} values on the horizontal axis and y_{ik} values on the vertical axis. The observed pattern might be suggestive of the nature of association among responses at times t_j and t_k.

- This is not exactly correct; in particular, if the means μ_j and μ_k and variances σ_j^2 and σ_k^2 are **not the same**, the patterns in the pairwise plots will in part be a consequence of this. It would make better sense to plot the "centered" and "scaled" versions of these; i.e. plot the pairs

$$\left(\frac{y_{ij} - \mu_j}{\sigma_j}, \frac{y_{ik} - \mu_k}{\sigma_k} \right).$$

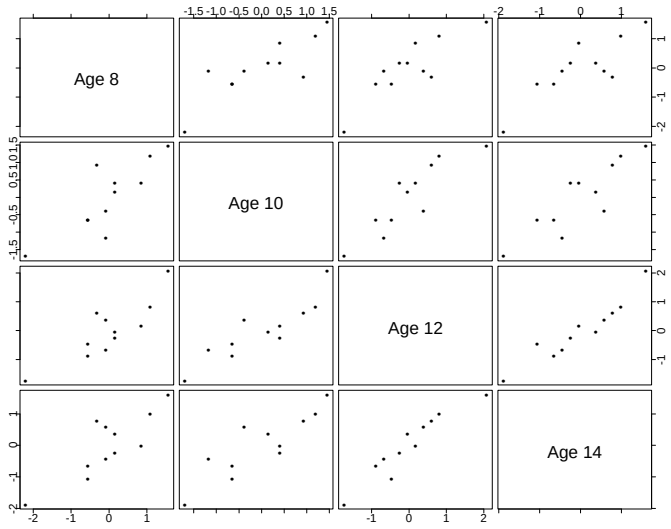
- Given we do not know the μ_j or σ_j, a natural strategy is to **replace** these by estimates and instead plot the pairs

$$\left(\frac{y_{ij} - \overline{y}_{\cdot j}}{s_j}, \frac{y_{ik} - \overline{y}_{\cdot k}}{s_k} \right).$$

Following this reasoning, it is common to make these plots for all pairs (j, k), where $j \neq k$.

Figure 3 shows the scatterplot matrix for the girls in the dental study.

Figure 3: *Scatterplot matrix for the girls in the dental study.*



In each panel, the apparent association among centered and scaled distance observations appears strong. The fact that the trend is from lower left to upper right in each panel, so that large centered and scaled values at one time correspond to large ones at another time, indicates that the association is **positive** for each pair of time points. Moreover, the nature of the association seems fairly similar **regardless** of the separation in time; i.e. the pattern of the plot corresponding to ages 8 and 14 shows a similar qualitative trend to those corresponding to ages 8 and 10, ages 8 and 12, and so on.

The evidence in the plots coincides with the numerical summary provided by the sample correlation matrix, which suggests that correlation is of similar magnitude and direction for any pair of times.

Some remarks:

- Visual display offers the data analyst another perspective on the likely pattern of aggregate correlation in the data in addition to that provided by the estimated correlation matrix. This information taken with that on variance in the sample covariance matrix can help the analyst to identify whether the pattern of variation has **systematic features**. If such systematic features are identified, it may be possible to adopt a **model** for $\text{var}(\boldsymbol{\epsilon}_i)$ that embodies them, allowing an accurate characterization. We take up this issue shortly.

- The same principles may be applied in more complicated settings; e.g. with more than one group. Here, one could estimate the covariance matrix $\boldsymbol{\Sigma}_\ell$ and associated correlation matrix $\boldsymbol{\Gamma}_\ell$, say, for each group ℓ separately and construct a separate scatterplot matrix.

- In the case of $q > 1$ groups, a natural objective would be to assess whether in fact it is reasonable to assume that the covariance matrix is the same for all groups.

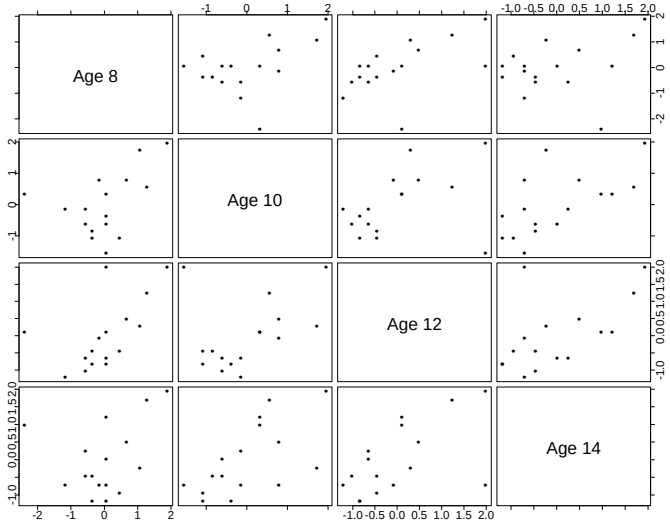
POOLED SAMPLE COVARIANCE AND CORRELATION MATRICES: To illustrate this last point, consider the data for boys in the dental study. It may be shown that the sample covariance and correlation matrices are

$$\widehat{\mathbf{\Sigma}}_B = \begin{pmatrix} 6.017 & 2.292 & 3.629 & 1.613 \\ 2.292 & 4.563 & 2.194 & 2.810 \\ 3.629 & 2.194 & 7.032 & 3.241 \\ 1.613 & 2.810 & 3.241 & 4.349 \end{pmatrix}, \quad \widehat{\mathbf{\Gamma}}_B = \begin{pmatrix} 1.000 & 0.437 & 0.558 & 0.315 \\ 0.437 & 1.000 & 0.387 & 0.631 \\ 0.558 & 0.387 & 1.000 & 0.586 \\ 0.315 & 0.631 & 0.586 & 1.000 \end{pmatrix}.$$

- Comparing to $\widehat{\mathbf{\Sigma}}_G$ for girls, aggregate variance does not seem to increase over time and seems larger than that for girls at all but the last time. (These estimates are based on small samples, 11 and 16 units, so should be interpreted with care.)
- Comparing to $\widehat{\mathbf{\Gamma}}_G$ for girls suggests that correlation for boys, although positive, is of smaller magnitude. Moreover, the estimated correlations for boys tend to "jump around" more than those for girls.

Figure 4 shows the scatterplot matrix for boys.

Figure 4: *Scatterplot matrix for the boys in the dental study.*



Comparing this figure to that for girls in Figure 3 reveals that the trend in each panel seems less profound for boys, although it is still positive in every case.

Overall, there seems to be **informal evidence** that both the mean and pattern of variance and correlation in the populations of girls and boys may be different. We will study longitudinal data models that allow such features to be taken into account.

Although this seems to be the case here, in many situations, the evidence may not be strong enough to suggest a difference in variation across groups, or scientific considerations may dictate that an assumption of a common pattern of overall variation is reasonable.

Under these conditions, it is natural to **combine** the information on variation across groups in order to examine the features of the assumed common structure. Since ordinarily interest focuses on whether the $\boldsymbol{\mu}_\ell$ are the same, as we will see, such an assessment continues to assume that the $\boldsymbol{\mu}_\ell$ may be different.

The assumed common covariance matrix $\boldsymbol{\Sigma}$ and its corresponding correlation matrix $\boldsymbol{\Gamma}$ from data for q groups may be estimated as follows. Assume that there are r_ℓ units from the ℓth population, so that m, the total number of units, is such that $m = r_1 + \cdots + r_q$.

- As we continue to believe the $\boldsymbol{\mu}_\ell$ are different, estimate these by the sample means $\overline{\boldsymbol{Y}}_\ell$, say, for each group.
- Let $\widehat{\boldsymbol{\Sigma}}_\ell$ denote the sample covariance matrix calculated for each group separately (based on $\overline{\boldsymbol{Y}}_\ell$).
- A natural strategy if we believe that there is a common covariance matrix $\boldsymbol{\Sigma}$ is then to use as an estimator for $\boldsymbol{\Sigma}$ a **weighted average** of the $\widehat{\boldsymbol{\Sigma}}_\ell$, $\ell = 1, \ldots, q$, that takes into account the differing amount of information from each group:

  $$\widehat{\boldsymbol{\Sigma}} = (m-q)^{-1}\{(r_1 - 1)\widehat{\boldsymbol{\Sigma}}_1 + \cdots + (r_q - 1)\widehat{\boldsymbol{\Sigma}}_q\}.$$

  This matrix is referred to as the **pooled sample covariance matrix**.
- If the number of units from each group is the **same**, so that $r_\ell \equiv r$, say, then $\widehat{\boldsymbol{\Sigma}}$ reduces to a simple average; i.e. $\widehat{\boldsymbol{\Sigma}} = (1/q)(\widehat{\boldsymbol{\Sigma}}_1 + \cdots + \widehat{\boldsymbol{\Sigma}}_q)$.
- The quantity in braces is often called the **Error SS&CP matrix**, as we will see later.
- The **pooled sample correlation matrix** estimating the assumed common correlation matrix $\boldsymbol{\Gamma}$ is naturally defined as the estimated correlation matrix corresponding to $\widehat{\boldsymbol{\Sigma}}$.

From the definition, the diagonal elements of the pooled sample covariance matrix are weighted averages of the sample variances from each group. That is, if $S_j^{(\ell)2}$ is the sample variance of the observations from group ℓ at time j, then the (j,j) element of $\widehat{\boldsymbol{\Sigma}}$, $\widehat{\Sigma}_{jj}$, say, is equal to

$$\widehat{\Sigma}_{jj} = (m-q)\{(r_1 - 1)S_j^{(1)2} + \cdots + (r_q - 1)S_j^{(q)2}\},$$

the so-called **pooled sample variance** at time t_j.

If the analyst is willing to adopt the assumption of a **common covariance matrix** for all groups, then inspection of the pooled estimate may be carried out as in the case of a single population. Similarly, a pooled scatterplot matrix would be based on centered and scaled versions of the y_{ij}, where the "centering" continues to be based on the sample means for each group but the "scaling" is based on the common estimate of variance for y_{ij} from $\widehat{\mathbf{\Sigma}}$. In particular, one would plot the observed pairs

$$\left(\frac{y_{ij} - \overline{y}_{\cdot j}^{(\ell)}}{\sqrt{\widehat{\Sigma}_{jj}}}, \frac{y_{ik} - \overline{y}_{\cdot k}^{(\ell)}}{\sqrt{\widehat{\Sigma}_{kk}}} \right)$$

for all units $i = 1, \ldots, m$ from all groups $\ell = 1, \ldots, q$ on the same graph for each pair of times t_j and t_k.

DENTAL STUDY: Although we are not convinced that it is appropriate to assume a common covariance matrix for boys and girls in the dental study, for illustration we calculate the pooled sample covariance and correlation matrix to obtain:

$$\widehat{\mathbf{\Sigma}} = (1/25)(10\widehat{\mathbf{\Sigma}}_G + 15\widehat{\mathbf{\Sigma}}_B) = \begin{pmatrix} 5.415 & 2.717 & 3.910 & 2.710 \\ 2.717 & 4.185 & 2.927 & 3.317 \\ 3.910 & 2.927 & 6.456 & 4.131 \\ 2.710 & 3.317 & 4.131 & 4.986 \end{pmatrix}$$

and

$$\widehat{\mathbf{\Gamma}} = \begin{pmatrix} 1.000 & 0.571 & 0.661 & 0.522 \\ 0.571 & 1.000 & 0.563 & 0.726 \\ 0.661 & 0.563 & 1.000 & 0.728 \\ 0.522 & 0.726 & 0.728 & 1.000 \end{pmatrix}.$$

- Inspection of the diagonal elements shows that the pooled estimates seem to be a "compromise" between the two group-specific estimates. This in fact illustrates how the pooled estimates combine information across groups.

- For brevity, we do not display the combined scatterplot matrix for these data. Not surprisingly, the pattern is somewhere "in between" those exhibited in Figures 3 and 4.

We have assumed throughout that we have **balanced data**. When the data are not balanced, either because some individuals are missing observations at intended times or because the times are different for different units, application of the above methods can be misleading. Later in the course, we consider methods for unbalanced data.

4.4 Popular models for covariance structure

As we have noted previously, if estimated covariance and correlation matrices show **systematic features**, the analyst may be led to consider **models** for covariance and associated correlation matrices. We will see later in the course that common models and associated methods for longitudinal data either explicitly or implicitly involve adopting particular models for $\text{var}(\boldsymbol{\epsilon}_i)$.

In anticipation this, here, we introduce some popular such covariance models that embody different systematic patterns that are often seen with longitudinal data. Each covariance model has a corresponding correlation model. We consider these models for **balanced data** only; modification for unbalanced data is discussed later.

UNSTRUCTURED COVARIANCE MODEL: In some situations, there may be no evidence of an apparent systematic pattern of variance and correlation. In this case, the covariance matrix is said to follow the **unstructured** model. The unstructured covariance model was adopted in the discussion of the last section as an initial assumption to allow assessment of whether a model with more structure could be substituted.

The unstructured covariance matrix allows n different variances, one for each time point, and $n(n-1)/2$ **distinct** off-diagonal elements representing the possibly different covariances for each pair of times, for a total of $n + n(n-1)/2 = n(n+1)/2$ variances and covariances. (Because a covariance matrix is **symmetric**, the off-diagonal elements at positions (j,k) and (k,j) are the same, so we need only count each covariance once in totaling up the number of variances and covariances involved.)

Thus, if the unstructured model is assumed, there are numerous **parameters** describing variation that must be estimated, particularly if n is large. E.g., if $n = 5$, which does not seem that large, there are $5(6)/2 = 15$ parameters involved. If there are q different groups, each with a different covariance matrix, there will be q times this many variances and covariances.

If the pattern of covariance does show a systematic structure, then not acknowledging this by maintaining the unstructured assumption involves estimation of many more parameters than might otherwise be necessary, thus making inefficient use of the available data. We now consider models that represent things in terms of far fewer parameters.

As we will see in the following, it is sometimes easier to discuss the correlation model first and then discuss the covariance matrix models to which it may correspond.

COMPOUND SYMMETRIC COVARIANCE MODELS: For both the boys and girls in the dental study, the correlation between observations at any times t_j and t_k seemed similar, although the variances at different times might be different.

These considerations suggest a covariance model that imposes equal correlation between all time points but allows variance to differ at each time as follows. Suppose that ρ is a parameter representing the common correlation for any two time points. For illustration, suppose that $n = 5$. Then the correlation matrix is

$$\mathbf{\Gamma} = \begin{pmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{pmatrix};$$

the same structure generalizes to any n. Here, $-1 < \rho < 1$. This is often referred to as the **compound symmetric** or **exchangeable** correlation model, where the latter term emphasizes that the correlation is the same even if we "exchange" two time points for two others.

Two popular covariance models with this correlation matrix are as follows.

- If σ_j^2 and σ_k^2 are the overall variances at t_j and t_k (possibly different at different times), and σ_{jk} is the corresponding covariance, then it must be that

$$\rho = \frac{\sigma_{jk}}{\sigma_j \sigma_k} \quad \text{or } \sigma_{jk} = \sigma_j \sigma_k \rho.$$

  We thus have a covariance matrix of the form, in the case $n = 5$,

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho\sigma_1\sigma_3 & \rho\sigma_1\sigma_4 & \rho\sigma_1\sigma_5 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \rho\sigma_2\sigma_4 & \rho\sigma_2\sigma_5 \\ \rho\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 & \rho\sigma_3\sigma_4 & \rho\sigma_3\sigma_5 \\ \rho\sigma_1\sigma_4 & \rho\sigma_2\sigma_4 & \rho\sigma_3\sigma_4 & \sigma_4^2 & \rho\sigma_4\sigma_5 \\ \rho\sigma_1\sigma_5 & \rho\sigma_2\sigma_5 & \rho\sigma_3\sigma_5 & \rho\sigma_4\sigma_5 & \sigma_5^2 \end{pmatrix},$$

  which of course generalizes to any n. This covariance matrix is often said to have a **heterogeneous compound symmetric** structure – **compound symmetric** because it has corresponding correlation as above and **heterogeneous** because it incorporates the assumption of different, or heterogeneous, variances at each time point. Note that this model may be described with $n + 1$ parameters, the correlation ρ and the n variances.

- In some settings, the evidence may suggest that the overall variance at each time point is the same, so that $\sigma_j^2 = \sigma^2$ for some common value σ^2 for all $j = 1, \ldots, n$. Under this condition,

$$\rho = \frac{\sigma_{jk}}{\sigma^2} \quad \text{so that } \sigma_{jk} = \sigma^2\rho \quad \text{for all } j, k.$$

  Under these conditions, the covariance matrix is, in the case $n = 5$.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \sigma^2 \end{pmatrix} = \sigma^2\boldsymbol{\Gamma}.$$

  This covariance matrix for any n is said to have the **compound symmetric** or **exchangeable** structure with no qualification.

This model involves only two parameters, σ^2 and ρ, for any n.

Remarks:

- From the diagnostic calculations and plots for the dental study data, the heterogeneous compound symmetric covariance model seems like a plausible model for each of the boys and girls, although the values of ρ and the variances at each time may be potentially different in each group.
- The unstructured and compound symmetric models do not emphasize the fact that observations are collected over time; neither has "built-in" features that really only make sense when the n observations are in a particular order. Recall the two sources of correlation that contribute to the overall pattern: that arising from among-unit sources (e.g. units being "high" or "low") and those due to within-unit sources (e.g. "fluctuations" about a smooth trend and measurement error). The compound symmetric models seem to emphasize the among-unit component.

The models we now discuss instead may be thought of as emphasizing the within-unit component through structures that are plausible when correlation depends on the times of observation in some way. As "fluctuations" determine this source of correlation, these models may be thought of as assuming that the variation attributable to these fluctuations dominates that from other sources (among-units or measurement error). These models have roots in the literature on **time series analysis**.

ONE-DEPENDENT: Correlation due to within-unit fluctuation would be expected to be "stronger" the closer observations are taken in time on a particular unit, as observations close in time would be "more alike" than those far apart. Thus, we expect correlation due to within-unit sources to be largest in magnitude among responses that are **adjacent** in time, that is, are at consecutive observation times, and to become less pronounced as observations become farther apart. Relative to this magnitude of correlation, that between two nonconsecutive observations might be for all practical purposes be negligible.

A correlation matrix that reflects this (shown for $n = 5$) is

$$\mathbf{\Gamma} = \begin{pmatrix} 1 & \rho & 0 & 0 & 0 \\ \rho & 1 & \rho & 0 & 0 \\ 0 & \rho & 1 & \rho & 0 \\ 0 & 0 & \rho & 1 & \rho \\ 0 & 0 & 0 & \rho & 1 \end{pmatrix}.$$

Here, the correlation is the same, equal to ρ, $-1 < \rho < 1$, for any two consecutive observations. This model is referred to as the **one-dependent** correlation structure, as dependence is nonnegligible only for adjacent responses. Alternatively, such a matrix is also referred to as a **banded Toeplitz** matrix.

The one-dependent correlation model seems to make the most sense if observation times are **equally-spaced** (separate by the same time interval).

If the overall variances σ_j^2, $j = 1, \ldots, n$, are possibly different at each time t_j, the corresponding covariance matrix ($n = 5$) looks like

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & 0 & 0 & 0 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 & 0 & 0 \\ 0 & \rho\sigma_2\sigma_3 & \sigma_3^2 & \rho\sigma_3\sigma_4 & 0 \\ 0 & 0 & \rho\sigma_3\sigma_4 & \sigma_4^2 & \rho\sigma_4\sigma_5 \\ 0 & 0 & 0 & \rho\sigma_4\sigma_5 & \sigma_5^2 \end{pmatrix}$$

and is called a **heterogeneous** one-dependent or banded Toeplitz matrix, for obvious reasons. Of course, this structure may be generalized to any n.

If overall variance at each time point is the same, so that $\sigma_j^2 = \sigma^2$ for all j, then this becomes

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma^2 & \rho\sigma^2 & 0 & 0 & 0 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & 0 & 0 \\ 0 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & 0 \\ 0 & 0 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ 0 & 0 & 0 & \rho\sigma^2 & \sigma^2 \end{pmatrix} = \sigma^2\mathbf{\Gamma},$$

which is usually called a **one-dependent** or **banded Toeplitz** matrix without qualification.

It is possible to extend this structure to a **two-dependent** or higher model. For example, two-dependence implies that observations one or two intervals apart in time are correlated, but those farther apart are not.

The one-dependent correlation model implies that correlation "falls off" as observations become farther apart in time in a rather dramatic way, so that only consecutive observations are correlated. Alternatively, it may be the case that correlation "falls off" more gradually.

AUTOREGRESSIVE STRUCTURE OF ORDER 1: Again, this model makes sense sense when the observation times are equally spaced. The **autoregressive**, or AR(1), correlation model, formalizes the idea that the magnitude of correlation among observations "decays" as they become farther apart. In particular, for $n = 5$, the AR(1) correlation matrix has the form

$$\mathbf{\Gamma} = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix},$$

where $-1 < \rho < 1$.

- As ρ is less than 1 in magnitude as we take it to higher powers, the result is values closer and closer to zero. Thus, as the number of time intervals between pairs of observations increases, the correlation decreases toward zero.
- With equally-spaced data, the time interval between t_j and t_{j+1} is the same for all j; i.e., $|t_j - t_{j+1}| = d$ for $j = 1, \ldots, n-1$, where d is the length of the interval. Note then that the power of ρ corresponds to the number of intervals by which a pair of observations is separated.

As with the compound symmetric and one-dependent models, both **heterogeneous** and "standard" covariance matrices with corresponding AR(1) correlation matrix are possible. In the case of overall variances σ_j^2 that may differ across j, the heterogeneous covariance matrix in the case $n = 5$ has the form

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho^2\sigma_1\sigma_3 & \rho^3\sigma_1\sigma_4 & \rho^4\sigma_1\sigma_5 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \rho^2\sigma_2\sigma_4 & \rho^3\sigma_2\sigma_5 \\ \rho^2\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 & \rho\sigma_3\sigma_4 & \rho^2\sigma_3\sigma_5 \\ \rho^3\sigma_1\sigma_4 & \rho^2\sigma_2\sigma_4 & \rho\sigma_3\sigma_4 & \sigma_4^2 & \rho\sigma_4\sigma_5 \\ \rho^4\sigma_1\sigma_5 & \rho^3\sigma_2\sigma_5 & \rho^2\sigma_3\sigma_5 & \rho\sigma_4\sigma_5 & \sigma_5^2 \end{pmatrix}.$$

When the variance is assumed equal to the same value σ^2 for all $j = 1, \ldots, n$, the covariance matrix has the form ($n = 5$)

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 & \rho^3\sigma^2 & \rho^4\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 & \rho^3\sigma^2 \\ \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 \\ \rho^3\sigma^2 & \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho^4\sigma^2 & \rho^3\sigma^2 & \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 \end{pmatrix} = \sigma^2\mathbf{\Gamma},$$

The one-dependent and AR(1) models really only seem sensible when the observation times are spaced at equal intervals, as in the dental study data. This is not always the case; for instance, for longitudinal data collected in clinical trials comparing treatments for disease, it is routine to collect responses frequently at the beginning of therapy but then to take them at wider intervals later.

The following offers a generalization of the AR(1) model to allow the possibility of unequally-spaced times.

MARKOV STRUCTURE: Suppose that the observation times $t_1, \ldots, t_n$ are not necessarily equally spaced, and let

$$d_{jk} = |t_j - t_k|$$

be the length of time between times t_j and t_k for all $j, k = 1, \ldots, n$. Then the **Markov** correlation model has the form, shown here for $n = 5$,

$$\mathbf{\Gamma} = \begin{pmatrix} 1 & \rho^{d_{12}} & \rho^{d_{13}} & \rho^{d_{14}} & \rho^{d_{15}} \\ \rho^{d_{12}} & 1 & \rho^{d_{23}} & \rho^{d_{24}} & \rho^{d_{25}} \\ \rho^{d_{13}} & \rho^{d_{23}} & 1 & \rho^{d_{34}} & \rho^{d_{35}} \\ \rho^{d_{14}} & \rho^{d_{24}} & \rho^{d_{34}} & 1 & \rho^{d_{45}} \\ \rho^{d_{15}} & \rho^{d_{25}} & \rho^{d_{35}} & \rho^{d_{45}} & 1 \end{pmatrix}.$$

- Here, we must have $\rho \geq 0$ (why?).
- Comparing this to the AR(1) structure, the powers of ρ and thus the degree of decay of correlation are also related to the length of the time interval between observations. Here, however, because the time intervals d_{jk} are of unequal length, the powers are the actual lengths.

Corresponding covariance matrices are defined similarly to those in the one-dependent and AR(1) cases. E.g., under the assumption of common variance σ^2, we have

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma^2 & \sigma^2\rho^{d_{12}} & \sigma^2\rho^{d_{13}} & \sigma^2\rho^{d_{14}} & \sigma^2\rho^{d_{15}} \\ \sigma^2\rho^{d_{12}} & \sigma^2 & \sigma^2\rho^{d_{23}} & \sigma^2\rho^{d_{24}} & \sigma^2\rho^{d_{25}} \\ \sigma^2\rho^{d_{13}} & \sigma^2\rho^{d_{23}} & \sigma^2 & \sigma^2\rho^{d_{34}} & \sigma^2\rho^{d_{35}} \\ \sigma^2\rho^{d_{14}} & \sigma^2\rho^{d_{24}} & \sigma^2\rho^{d_{34}} & \sigma^2 & \sigma^2\rho^{d_{45}} \\ \sigma^2\rho^{d_{15}} & \sigma^2\rho^{d_{25}} & \sigma^2\rho^{d_{35}} & \sigma^2\rho^{d_{45}} & \sigma^2 \end{pmatrix} = \sigma^2\mathbf{\Gamma},$$

This model has two parameters, σ^2 and ρ, for any n.

These are not the only such models available, but give a flavor of the types of considerations involved. The documentation for the SAS procedure `proc mixed`, the use of which we will demonstrate in subsequent chapters, offers a rich catalog of possible covariance models.

If one believes that one of the foregoing models or some other model provides a realistic representation of the pattern of variation and covariation in the data, then intuition suggests that a "better" estimate of var($\boldsymbol{\epsilon}_i$) could be obtained by exploiting this information. We will see this in action shortly.

We will also see that these models may be used not only to represent var($\boldsymbol{\epsilon}_i$), but to represent covariance matrices of components of $\boldsymbol{\epsilon}_i$ corresponding to among- and within-unit variation.

4.5 Diagnostic calculations under stationarity

The one-dependent, AR(1), and Markov structures are popular models when it is thought that the predominant source of correlation leading to the aggregate pattern is from **within-individual** sources. All of these models are such that the correlation between Y_{ij} and Y_{ik} for any $j \neq k$ depends only on the time **interval** $|t_j - t_k|$ and not only the specific times t_j or t_k themselves. This property is known as **stationarity**.

- If stationarity is thought to hold, the analyst may wish to investigate which correlation structure (e.g. one-dependent, AR(1), or other model for equally-spaced data) might be the best model.

- Variance at each t_j may be assessed by examining the sample covariance matrix.
- If one believes in stationarity, an investigation of correlation that takes this into account may offer more refined information than one that does not, as we now demonstrate.

The rationale is as follows:

- When the t_j, $j = 1, \ldots, n$, are equally spaced, with time interval d, under stationarity, all pairs of observations corresponding to times whose subscripts differ by 1, e.g. j and $j + 1$, are d time units apart and are correlated in an identical fashion.
- Similarly, all pairs with subscripts differing by 2, e.g. j and $j + 2$ are $2d$ time units apart and correlated in the same way. In general, pairs with subscripts j and $j + u$ are ud time units apart and share the same correlation.
- The value of subscripts for n time points must range between 1 and n. Thus, when we write j and $j + u$, it is understood that the values of j and u are chosen so that all possible distinct pairs of unequal subscripts in this range are represented. E.g. if $j = 1$, then u may take on the values $1, \ldots, n - 1$ to give all pairs corresponding to time t_1 and all other times $t_2, \ldots, t_n$. If $j = 2$, then u may take on values $1, \ldots, n-2$, and so on. If $j = n-1$, then $u = 1$ gives the pair corresponding to times t_{n-1}, t_n.
- For example, under the AR(1) model, for a particular u, pairs at times t_j and t_{j+u} for satisfy

$$\text{corr}(Y_{ij}, Y_{i,j+u}) = \rho^u,$$

  suggesting that the correlation between observations u time intervals apart may be assessed using information from **all** such pairs.

AUTOCORRELATION FUNCTION: The **autocorrelation function** is just the correlation corresponding to pairs of observations u time intervals apart thought of as a **function** of the number of intervals. That is, for all $j = 1, \ldots, n - 1$ and appropriate u,

$$\rho(u) = \text{corr}(Y_{ij}, Y_{i,j+u}).$$

- This depends only on u and is the same for all j because of stationarity.
- The value of $\rho(0)$ is taken to be equal to one, as with $u = 0$ $\rho(0)$ is just the correlation between an observation and itself.

- The value u is often called the **lag**. The total number of possible lags is $n-1$ for n time points.
- The autocorrelation function describes how correlation changes as the time between observations gets farther apart, i.e. as u increases. As expected, the value of $\rho(u)$ tends to decrease in magnitude as u increases, reflecting the usual situation in which within-unit correlation "falls off" as observations become more separated in time.

In practice, we may **estimate** the autocorrelation function if we are willing to assume that stationarity holds. Inspection of the estimate can help the analyst decide which model might be appropriate; e.g. if correlation falls off gradually with lag, it may suggest that an AR(1) model is appropriate.

For data from a single population, it is natural to base estimation of $\rho(u)$ for each $u = 1, \ldots, n-1$ on all pairs of observations $(Y_{ij}, Y_{i,j+u})$ across all individuals $i = 1, \ldots, m$ and relevant choices of j.

- Care must be taken to ensure that the fact that responses have different means and overall variances at each t_j is taken into account, as with scatterplot matrices.
- Thus, we consider "centered" and "scaled" observations. In particular, $\rho(u)$ for a particular lag u may be estimated by calculating the **sample correlation coefficient** treating all pairs of the form
$$\frac{Y_{ij} - \overline{Y}_{\cdot j}}{S_j}, \frac{Y_{i,j+u} - \overline{Y}_{\cdot j+u}}{S_{j+u}}$$
as if they were observations on two random variables from a sample of m individuals, where each individual contributes more than one pair.
- The resulting estimator as a function of u is called the **sample autocorrelation function**, which we denote as $\widehat{\rho}(u)$.
- $\widehat{\rho}(u)$ may be calculated and plotted against u to provide the analyst with both numerical and visual information on the nature of correlation if the stationarity assumption is plausible.

We illustrate using the data from girls in the dental study. Here, the time interval is of length $d = 2$ years, and $n = 4$, so u can take on values $1, \ldots, n-1 = 3$.

- When $u = 1$, each girl has three pairs of values separated by d units (i.e. one time interval), the values at (t_1, t_2), (t_2, t_3), and (t_3, t_4). Thus, there is a total of 33 possible pairs from all 11 girls.
- When $u = 2$, there are two pairs per girl, at (t_1, t_3) and (t_2, t_4), or 22 total pairs.

- When $u = 3$, each girl contributes a single pair at (t_1, t_4), 11 pairs in total).

Thus, the calculation of $\widehat{\rho}(u)$ is carried out by calculating the sample correlation coefficient from 33, 22, and 11 observations for $u = 1, 2$, and 3, respectively, and yields

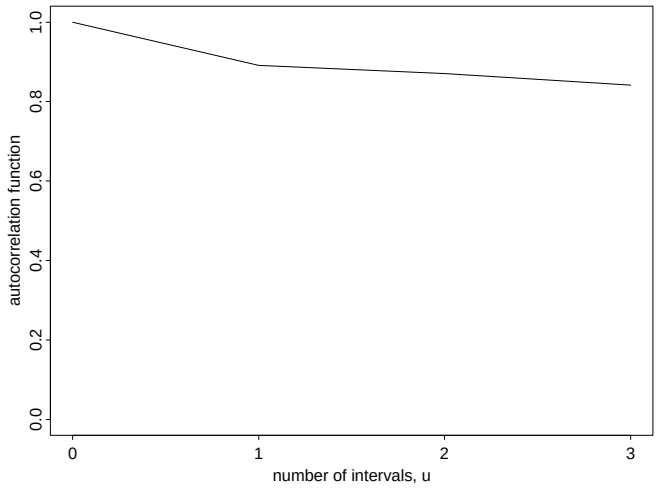
u	1	2	3
$\widehat{\rho}(u)$	0.891	0.871	0.841

Because each estimated value is based on a decreasing number of pairs, they are not of equal quality, so should be interpreted with care.

The estimates suggest that, if we are willing to believe stationarity, as observations become farther apart in time (u increasing), correlation seems to stay fairly constant. This agrees with the evidence from the calculation of the sample covariance matrix and the scatterplot matrix in Figure 3.

Figure 5 shows a plot of the sample autocorrelation function, displaying the same information graphically.

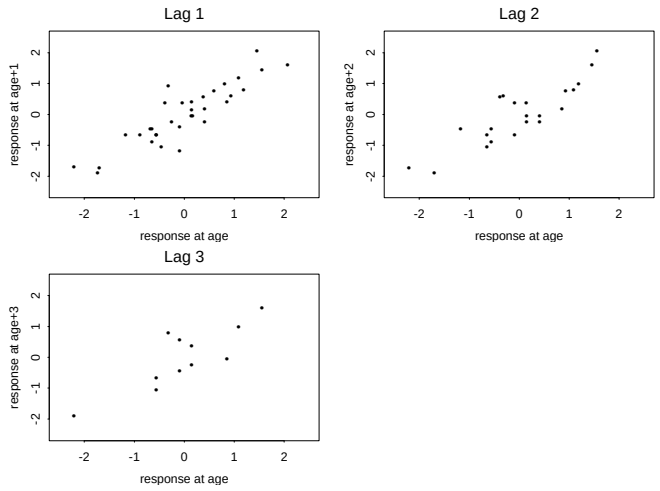
Figure 5: *Sample autocorrelation function for data from girls in the dental study.*



An alternative way of displaying information on correlation under the assumption of stationarity is to plot the pairs for each choice of lag u. From above, there are 33 pairs corresponding to lag $u = 1$, 22 for lag $u = 2$, and 11 for lag $u = 3$. In Figure 6, these pairs are plotted for each u. The plot gives a similar impression as the numerical estimate. An advantage of the plot is that it clearly shows that the information on correlation (total number of pairs) decreases as u increases.

For more than one group, these procedures may be carried out separately for each group.

Figure 6: *Lag plots for data from girls in the dental study for lags $u = 1$, 2, and 3.*



When data are not equally spaced, extensions of the method for estimating the autocorrelation function are available, but are beyond the scope of our discussion here. The reader is referred to Diggle, Heagerty, Liang, and Zeger (2002).

It is important to recognize that whether stationarity holds is an **assumption**. The foregoing procedures are relevant when this assumption is valid. Unfortunately, assessing with confidence whether stationarity holds is not really possible in longitudinal data situations where the number of time points is usually small. Because many popular models for correlation used in longitudinal data analysis embody the stationarity assumption, it is often assumed without comment, and it is often reasonable.

4.6 Implementation with SAS

We demonstrate the use of various SAS procedures on the dental data. In particular, we show how the following may be obtained:

- Sample mean vectors for each group (girls and boys)
- Group-specific sample covariance and correlation matrices
- Pooled sample covariance and correlation matrix
- Pairs for plotting scatterplot matrices for each group
- Autocorrelation functions for each gender and pairs for making lag plots

There are actually numerous ways to obtain the pooled sample covariance and correlation matrices. We show one way here, using SAS `PROC DISCRIM`. Additional ways can be found in the program on the course web site.

EXAMPLE 1 – DENTAL STUDY DATA: The data are in the file `dental.dat`.

PROGRAM:

```
/*****************************************************************

  EXAMPLE 1, CHAPTER 4

  Using SAS to obtain sample mean vectors, sample covariance
  matrices, and sample correlation matrices.

*****************************************************************/

options ls=80 ps=59 nodate; run;

/*****************************************************************

  The data are not in the correct from for use with the SAS procedures
  CORR and DISCRIM we use below.  These procedures require that the
  data be in the form of one record (line) per experimental unit.
  The data in the file dental.dat are in the form of one record per
  observation (so that each child has 4 data records).

  In particular, the data set looks like

  1 1 8 21 0
  2 1 10 20 0
  3 1 12 21.5 0
  4 1 14 23 0
  5 2 8 21 0
      .
      .
      .

  column 1    observation number
  column 2    child id number
  column 3    age
  column 4    response (distance)
  column 5    gender indicator (0=girl, 1=boy)

  We thus create a new data set such that each record in the data
  set represents all 4 observations on each child plus gender
  identifier.  To do this, we use some data manipulation features
  of the SAS data step.  The second data step does this.

  We redefine the values of AGE so that we may use AGE as an "index"
  in creating the new data set DENT2. The DATA step that creates
  DENT2 demonstrates one way (using the notion of an ARRAY) to
  transform a data set in the form of one observation per record
  (the original form) into a data set in the form of one record per
  individual.  The data must be sorted prior to this operation; we
  invoke PROC SORT for this purpose.

  In the new data set, the observations at ages 8, 10, 12, and 14
  are placed in variables AGE1, AGE2, AGE3, and AGE4, respectively.

  We use PROC PRINT to print out the first 5 records (so data for
  the first 5 children, all girls) using the OBS= feature of the
  DATA= option.

*****************************************************************/

data dent1; infile 'dental.dat';
  input obsno child age distance gender;
run;

data dent1; set dent1;
  if age=8 then age=1;
  if age=10 then age=2;
  if age=12 then age=3;
  if age=14 then age=4;
  drop obsno;
run;

proc sort data=dent1;
  by gender child;
run;
```

```
data dent2(keep=age1-age4 gender child);
  array aa{4} age1-age4;
  do age=1 to 4;
  set dent1;
  by gender child;
  aa{age}=distance;
  if last.child then return;
end;
run;

title "TRANSFORMED DATA -- 1 RECORD/INDIVIDUAL";
proc print data=dent2(obs=5); run;

/*****************************************************************

  Here, we use PROC CORR to obtain the sample means at each
  age (the means of the variables AGE1,...,AGE4 in DENT2 and to
  calculate the sample covariance matrix and corresponding sample
  correlation matrix separately for each group (girls and boys).
  The COV option in the PROC CORR statement asks for the sample
  covariance to be printed; without it, only the sample correlation
  matrix would appear in the output.

*****************************************************************/

proc sort data=dent2; by gender; run;

title "SAMPLE COVARIANCE AND CORRELATION MATRICES BY GENDER";
proc corr data=dent2 cov;
 by gender; var age1 age2 age3 age4;
run;

/*****************************************************************

  We now obtain the "centered" and "scaled" values
  that may be used for plotting scatterplot matrices such as that
  in Figure 3.  Here, we call PROC MEANS to calculate the sample
  mean (MAGE1,...,MAGE4) and standard deviation (SDAGE1,...,SDAGE4)
  for each of the variables AGE1,...,AGE4 for each gender.  These
  are output to the data set DENTSTATS, which has two records, one
  for each gender (see the output).  We then MERGE this data set
  with DENT2 BY GENDER, which has the effect of matching up the
  appropriate gender mean and SD to each child.  We print out the
  first three records of the resulting data set to illustrate.
  We use the NOPRINT option with PROC MEANS to suppress printing of
  its output.

  The variables CSAGE1,...,CSAGE4 contain the centered/scaled values.
  These may be plotted against each other to obtain plots like Figure 3.
  We have not done this here to save space.

*****************************************************************/

proc sort data=dent2; by gender child; run;

proc means data=dent2 mean std noprint; by gender;
  var age1 age2 age3 age4;
  output out=dentstats mean=mage1 mage2 mage3 mage4
         std=sdage1 sdage2 sdage3 sdage4;
run;

title "SAMPLE MEANS AND SDS BY GENDER FROM PROC MEANS";
proc print data=dentstats; run;

data dentstats; merge dentstats dent2; by gender;
  csage1=(age1-mage1)/sdage1;
  csage2=(age2-mage2)/sdage2;
  csage3=(age3-mage3)/sdage3;
  csage4=(age4-mage4)/sdage4;
run;

title "INDIVIDUAL DATA MERGED WITH MEANS AND SDS BY GENDER";
proc print data=dentstats(obs=3); run;

/*****************************************************************

  One straightforward way to have SAS calculate the pooled sample
  covariance matrix and the corresponding estimated correlation matrix
  is using PROC DISCRIM.  This procedure is focused on so-called
  discriminant analysis, which is discussed in a standard text on
  general multivariate analysis.  The data are considered as
  in the form of vectors; here, the elements of a data vector are
  denoted as AGE1,...,AGE4.

  Here, we only use PROC DISCRIM for its facility to print out the
  sample covariance matrix and correlation matrix "automatically,"
  and disregard other portions of the output.
```

```
*****************************************************************/

proc discrim pcov pcorr data=dent2;
  class gender;
  var age1 age2 age3 age4;
run;

/*****************************************************************

  Although it is a bit cumbersome, we may use some DATA step
  manipulations and PROC CORR to obtain the values of the autocorrelation
  function  for each gender.  We first drop variables
  no longer needed from the data set DENTSTATS.

  We create then three data sets, LAG1, LAG2, and LAG3, and describe
  LAG1 here; the other two are similar.  We create two new variables,
  PAIR1 and PAIR2.  For LAG1, PAIR1 and PAIR2 are the two values in (5.43)
  for u=1. As there are 4 ages, each child has 3 such pairs.  The output
  of PROC PRINT for LAG1 shows this for the first 2 children.
  We then sort the data by gender and call PROC CORR to find the
  sample correlation between the two variables for each gender.

  The same principle is used to obtain the correlation by gender for
  lags 2 and 3 [u=2,3].

  There are other, more sophisticated ways to obtain the values
  of the autocorrelation function; however, for longitudinal data sets
  where the number of time points is small, the "manual" approach
  we have demonstrated here is easy to implement and understand.

  PAIR1 versus PAIR2 may be plotted for each lag to obtain visual
  presentation of the results as in Figure 6.

*****************************************************************/

data dentstats; set dentstats;
  drop age1-age4 mage1-mage4 sdage1-sdage4;
run;

data lag1; set dentstats;
  by child;
  pair1=csage1; pair2=csage2; output;
  pair1=csage2; pair2=csage3; output;
  pair1=csage3; pair2=csage4; output;
  if last.child then return;
  drop csage1-csage4;
run;

title "AUTOCORRELATION FUNCTION AT LAG 1";
proc print data=lag1(obs=6); run;
proc sort data=lag1; by gender;

proc corr data=lag1; by gender;
  var pair1 pair2;
run;

data lag2; set dentstats;
  by child;
  pair1=csage1; pair2=csage3; output;
  pair1=csage2; pair2=csage4; output;
  if last.child then return;
  drop csage1-csage4;
run;

title "AUTOCORRELATION FUNCTION AT LAG 2";
proc print data=lag2(obs=6); run;
proc sort data=lag2; by gender;

proc corr data=lag2; by gender;
  var pair1 pair2;
run;

data lag3; set dentstats;
  by child;
  pair1=csage1; pair2=csage4; output;
  if last.child then return;
  drop csage1-csage4;
run;

title "AUTOCORRELATION FUNCTION AT LAG 3";
proc print data=lag3(obs=6); run;
proc sort data=lag3; by gender;

proc corr data=lag3; by gender;
  var pair1 pair2;
run;
```

OUTPUT: We have deleted some of the output of `PROC DISCRIM` that is irrelevant to our purposes here to shorten the presentation. The full output from the call to this procedure is on the course web page.

```
                TRANSFORMED DATA -- 1 RECORD/INDIVIDUAL                1

            Obs    age1    age2    age3    age4    child    gender

              1    21.0    20.0    21.5    23.0      1         0
              2    21.0    21.5    24.0    25.5      2         0
              3    20.5    24.0    24.5    26.0      3         0
              4    23.5    24.5    25.0    26.5      4         0
              5    21.5    23.0    22.5    23.5      5         0

          SAMPLE COVARIANCE AND CORRELATION MATRICES BY GENDER            2
---------------------------------- gender=0 ----------------------------------
                              The CORR Procedure

               4  Variables:    age1     age2     age3     age4

                          Covariance Matrix, DF = 10

                    age1              age2              age3              age4

  age1       4.513636364       3.354545455       4.331818182       4.356818182
  age2       3.354545455       3.618181818       4.027272727       4.077272727
  age3       4.331818182       4.027272727       5.590909091       5.465909091
  age4       4.356818182       4.077272727       5.465909091       5.940909091

                          Simple Statistics

 Variable          N        Mean     Std Dev         Sum     Minimum     Maximum

 age1             11    21.18182     2.12453   233.00000    16.50000    24.50000
 age2             11    22.22727     1.90215   244.50000    19.00000    25.00000
 age3             11    23.09091     2.36451   254.00000    19.00000    28.00000
 age4             11    24.09091     2.43740   265.00000    19.50000    28.00000

                   Pearson Correlation Coefficients, N = 11
                         Prob > |r| under H0: Rho=0

                       age1          age2          age3          age4

          age1       1.00000       0.83009       0.86231       0.84136
                                     0.0016        0.0006        0.0012

          age2       0.83009       1.00000       0.89542       0.87942
                      0.0016                       0.0002        0.0004

          age3       0.86231       0.89542       1.00000       0.94841
                      0.0006        0.0002                      <.0001

          age4       0.84136       0.87942       0.94841       1.00000
                      0.0012        0.0004        <.0001

          SAMPLE COVARIANCE AND CORRELATION MATRICES BY GENDER            3
---------------------------------- gender=1 ----------------------------------
                              The CORR Procedure

               4  Variables:    age1     age2     age3     age4

                          Covariance Matrix, DF = 15

                    age1              age2              age3              age4

  age1       6.016666667       2.291666667       3.629166667       1.612500000
  age2       2.291666667       4.562500000       2.193750000       2.810416667
  age3       3.629166667       2.193750000       7.032291667       3.240625000
  age4       1.612500000       2.810416667       3.240625000       4.348958333

                          Simple Statistics

 Variable          N        Mean     Std Dev         Sum     Minimum     Maximum

 age1             16    22.87500     2.45289   366.00000    17.00000    27.50000
 age2             16    23.81250     2.13600   381.00000    20.50000    28.00000
 age3             16    25.71875     2.65185   411.50000    22.50000    31.00000
```

```
age4            16     27.46875      2.08542    439.50000     25.00000     31.50000

                  Pearson Correlation Coefficients, N = 16
                          Prob > |r| under H0: Rho=0

                           age1          age2          age3          age4

          age1          1.00000       0.43739       0.55793       0.31523
                                       0.0902        0.0247        0.2343

          age2          0.43739       1.00000       0.38729       0.63092
                         0.0902                      0.1383        0.0088

          age3          0.55793       0.38729       1.00000       0.58599
                         0.0247        0.1383                      0.0171

          age4          0.31523       0.63092       0.58599       1.00000
                         0.2343        0.0088        0.0171

                  SAMPLE MEANS AND SDS BY GENDER FROM PROC MEANS                4

      g   _   _                                       s       s       s       s
      e   T   F     m       m       m       m       d       d       d       d
      n   Y   R     a       a       a       a       a       a       a       a
  O d P   E   E     g       g       g       g       g       g       g       g
  b e E   Q   Q     e       e       e       e       e       e       e       e
  s r _   _   1     2       3       4       1       2       3       4

  1 0 0 11 21.1818 22.2273 23.0909 24.0909 2.12453 1.90215 2.36451 2.43740
  2 1 0 16 22.8750 23.8125 25.7188 27.4688 2.45289 2.13600 2.65185 2.08542

              INDIVIDUAL DATA MERGED WITH MEANS AND SDS BY GENDER               5

Obs gender _TYPE_ _FREQ_  mage1   mage2  mage3   mage4   sdage1  sdage2  sdage3

 1    0      0      11   21.1818 22.2273 23.0909 24.0909 2.12453 1.90215 2.36451
 2    0      0      11   21.1818 22.2273 23.0909 24.0909 2.12453 1.90215 2.36451
 3    0      0      11   21.1818 22.2273 23.0909 24.0909 2.12453 1.90215 2.36451

Obs  sdage4 age1 age2 age3 age4 child  csage1   csage2   csage3   csage4

  1 2.43740 21.0 20.0 21.5 23.0   1   -0.08558 -1.17092 -0.67283 -0.44757
  2 2.43740 21.0 21.5 24.0 25.5   2   -0.08558 -0.38234  0.38447  0.57811
  3 2.43740 20.5 24.0 24.5 26.0   3   -0.32093  0.93196  0.59593  0.78325

              INDIVIDUAL DATA MERGED WITH MEANS AND SDS BY GENDER               6

                            The DISCRIM Procedure

              Observations      27          DF Total                26
              Variables          4          DF Within Classes       25
              Classes            2          DF Between Classes       1

                            Class Level Information

                Variable                                                 Prior
   gender       Name        Frequency       Weight    Proportion    Probability

        0       _0                 11      11.0000      0.407407       0.500000
        1       _1                 16      16.0000      0.592593       0.500000

              INDIVIDUAL DATA MERGED WITH MEANS AND SDS BY GENDER               7

                            The DISCRIM Procedure

                Pooled Within-Class Covariance Matrix,     DF = 25

Variable            age1              age2              age3              age4

age1          5.415454545       2.716818182       3.910227273       2.710227273
age2          2.716818182       4.184772727       2.927159091       3.317159091
age3          3.910227273       2.927159091       6.455738636       4.130738636
age4          2.710227273       3.317159091       4.130738636       4.985738636

              INDIVIDUAL DATA MERGED WITH MEANS AND SDS BY GENDER               8

                            The DISCRIM Procedure

             Pooled Within-Class Correlation Coefficients  /  Pr > |r|

           Variable          age1          age2          age3          age4

           age1           1.00000       0.57070       0.66132       0.52158
```

		0.0023	0.0002	0.0063
age2	0.57070 0.0023	1.00000	0.56317 0.0027	0.72622 <.0001
age3	0.66132 0.0002	0.56317 0.0027	1.00000	0.72810 <.0001
age4	0.52158 0.0063	0.72622 <.0001	0.72810 <.0001	1.00000

AUTOCORRELATION FUNCTION AT LAG 1 11

Obs	gender	_TYPE_	_FREQ_	child	pair1	pair2
1	0	0	11	1	-0.08558	-1.17092
2	0	0	11	1	-1.17092	-0.67283
3	0	0	11	1	-0.67283	-0.44757
4	0	0	11	2	-0.08558	-0.38234
5	0	0	11	2	-0.38234	0.38447
6	0	0	11	2	0.38447	0.57811

AUTOCORRELATION FUNCTION AT LAG 1 12

---------- gender=0 ----------

The CORR Procedure

2 Variables: pair1 pair2

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
pair1	33	0	0.96825	0	-2.20369	2.07616
pair2	33	0	0.96825	0	-1.88353	2.07616

Pearson Correlation Coefficients, N = 33
Prob > |r| under H0: Rho=0

	pair1	pair2
pair1	1.00000	0.89130 <.0001
pair2	0.89130 <.0001	1.00000

AUTOCORRELATION FUNCTION AT LAG 1 13

---------- gender=1 ----------

The CORR Procedure

2 Variables: pair1 pair2

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
pair1	48	0	0.97849	0	-2.39513	1.99154
pair2	48	0	0.97849	0	-1.55080	1.99154

Pearson Correlation Coefficients, N = 48
Prob > |r| under H0: Rho=0

	pair1	pair2
pair1	1.00000	0.47022 0.0007
pair2	0.47022 0.0007	1.00000

AUTOCORRELATION FUNCTION AT LAG 2 14

Obs	gender	_TYPE_	_FREQ_	child	pair1	pair2
1	0	0	11	1	-0.08558	-0.67283
2	0	0	11	1	-1.17092	-0.44757
3	0	0	11	2	-0.08558	0.38447
4	0	0	11	2	-0.38234	0.57811
5	0	0	11	3	-0.32093	0.59593
6	0	0	11	3	0.93196	0.78325

AUTOCORRELATION FUNCTION AT LAG 2 15

---------- gender=0 ----------

The CORR Procedure

2 Variables: pair1 pair2

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
pair1	22	0	0.97590	0	-2.20369	1.56184
pair2	22	0	0.97590	0	-1.88353	2.07616

Pearson Correlation Coefficients, N = 22
Prob > |r| under H0: Rho=0

	pair1	pair2
pair1	1.00000	0.87087 <.0001
pair2	0.87087 <.0001	1.00000

AUTOCORRELATION FUNCTION AT LAG 2 16

---------- gender=1 ----------

The CORR Procedure

2 Variables: pair1 pair2

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
pair1	32	0	0.98374	0	-2.39513	1.96044
pair2	32	0	0.98374	0	-1.21378	1.99154

Pearson Correlation Coefficients, N = 32
Prob > |r| under H0: Rho=0

	pair1	pair2
pair1	1.00000	0.59443 0.0003
pair2	0.59443 0.0003	1.00000

AUTOCORRELATION FUNCTION AT LAG 3 17

Obs	gender	_TYPE_	_FREQ_	child	pair1	pair2
1	0	0	11	1	-0.08558	-0.44757
2	0	0	11	2	-0.08558	0.57811
3	0	0	11	3	-0.32093	0.78325
4	0	0	11	4	1.09115	0.98839
5	0	0	11	5	0.14977	-0.24243
6	0	0	11	6	-0.55627	-0.65271

AUTOCORRELATION FUNCTION AT LAG 3 18

---------- gender=0 ----------

The CORR Procedure

2 Variables: pair1 pair2

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
pair1	11	0	1.00000	0	-2.20369	1.56184
pair2	11	0	1.00000	0	-1.88353	1.60380

Pearson Correlation Coefficients, N = 11
Prob > |r| under H0: Rho=0

	pair1	pair2

```
                    pair1       1.00000       0.84136
                                              0.0012

                    pair2       0.84136       1.00000
                                0.0012


                      AUTOCORRELATION FUNCTION AT LAG 3                    19

---------------------------------- gender=1 ----------------------------------

                            The CORR Procedure

                     2  Variables:    pair1    pair2


                            Simple Statistics

 Variable         N        Mean     Std Dev         Sum     Minimum     Maximum

 pair1           16           0     1.00000           0    -2.39513     1.88553
 pair2           16           0     1.00000           0    -1.18382     1.93307


                  Pearson Correlation Coefficients, N = 16
                        Prob > |r| under H0: Rho=0

                                      pair1         pair2

                    pair1       1.00000       0.31523
                                              0.2343

                    pair2       0.31523       1.00000
                                0.2343
```