6

Modeling the Mean: Parametric Curves

6.1 INTRODUCTION

In the previous chapter we described an approach to modeling longitudinal data that effectively imposed no structure on the underlying mean response trend over time. This approach has some appeal when all subjects are measured at the same set of occasions and the number of measurement occasions is relatively small (e.g., not more than 4 or 5). But as the number of occasions increases and/or when the repeated measures are irregularly timed, analyzing response profiles becomes much less appealing. Even in cases where the number of repeated measures is relatively small, there are two obvious drawbacks of the analysis of response profiles that limit its usefulness for the analysis of longitudinal data. The first is that a statistical test of the null hypothesis of no group × time interaction is an omnibus or global test and provides only a broad assessment of whether the mean response profiles are the same in the different groups. If the null hypothesis is rejected, this does not indicate the specific ways in which the mean response profiles differ. As a result, additional analyses are invariably required. Second, by completely ignoring the time-ordering of the repeated measurements, the analysis of response profiles fails to recognize that they can be considered as observations of some continuous, underlying response process over time. The mean response over time can very often be described by relatively simple parametric (e.g., linear or quadratic) or semiparametric (e.g., piecewise linear) curves. From a purely substantive point of view, it is unlikely that the pattern of change in the mean response over the duration of a longitudinal study will be so complicated that its description requires as many parameters as there are measurement occasions. The analysis of response profiles uses a saturated model for the mean response over

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

time, and thereby produces a perfect fit to the observed mean response profile. At first glance this might seem like a desirable feature of any analytic approach; namely that it fits the observed mean responses well. (In fact, not just well, but perfectly!) However, in doing so, the method fails to describe the most salient aspects of the changes in the mean response over time in terms of some pattern that can be given a substantive or theoretical interpretation. In summary, in the analysis of response profiles there is no reduction in complexity.

In contrast, the fitting of parametric or semiparametric curves to longitudinal data can be justified on both substantive and statistical grounds. Substantively, in many longitudinal studies the true underlying mean response process is likely to change over time in a relatively smooth, monotonically increasing or decreasing pattern, at least for the duration of the study. As a result simple parametric or semiparametric curves can be used to describe how the mean response changes over time. From a statistical perspective the fitting of parsimonious models for the mean response will result in statistical tests of covariate effects (e.g., treatment × time interactions) that have greater power than in an analysis of response profiles. The reason for the greater power is that the tests of covariate effects focus only on a relatively narrow range of alternative hypotheses. In contrast, the test statistics in the analysis of response profiles disperse their power over a much broader, but in many cases less substantively plausible or relevant, range of alternative hypotheses. For example, when trends in the mean response over time are assumed to be linear, and a linear trend actually provides a reasonable approximation to the true underlying shape of the mean response profile, the resulting tests of time trends and covariate effects will have greater power than the global tests in an analysis of response profiles. Note, however, that the tests based on parametric curves will only be more powerful at detecting changes in the mean response that exhibit a linear trend. They will not be more powerful, however, if the underlying shape of the mean response over time is U-shaped, rather than linear. Finally, simple parametric curves provide a parsimonious description of changes in the mean response over time in terms of a relatively small number of parameters. The results can be communicated easily to investigators and empirical researchers. In the following two sections, we describe two broad approaches for describing patterns of change in the mean response over time: polynomial trends and linear splines.

6.2 POLYNOMIAL TRENDS IN TIME

One widely adopted approach for analyzing longitudinal data is to describe the patterns of change in the mean response over time in terms of simple polynomial trends, for example, linear or quadratic trends. In this approach the means are modeled as an explicit function of time. This approach can handle highly unbalanced designs in a relatively seamless way. For example, mistimed measurements are easily incorporated in the model for the mean response.

LINEAR TRENDS OVER TIME

The simplest possible curve for describing changes in the mean response over time is a straight line. In this model the slope for time has direct interpretation in terms of a constant change in the mean response for a single-unit change in time. Consider the hypothetical two-group study comparing a novel *treatment* and a *control* discussed in Section 5.2. If the mean response changes in an approximately linear fashion over the duration of the study, we can adopt the following linear trend model:

$$E\left(Y_{ij}\right) = \beta_1 + \beta_2 \, \text{Time}_{ij} + \beta_3 \, \text{Group}_i + \beta_4 \, \text{Time}_{ij} \times \text{Group}_i, \qquad (6.1)$$

where $\text{Group}_i = 1$ if the i^{th} individual was assigned to the novel treatment, and $\text{Group}_i = 0$ otherwise; Time_{ij} denotes the measurement time for the j^{th} measurement on the i^{th} individual. Note that Group_i requires only a single index i, since individuals do not change treatment groups over the course of the study. Also, by using two indices for Time_{ij}, we are implicitly allowing for the fact that there may potentially be mistimed measurements (in the latter case, $\text{Time}_{ij} \neq \text{Time}_{i'j}$, where i and i' denote two different subjects).

In the linear model given by (6.1), the model for the mean for subjects assigned to the control group is

$$E\left(Y_{ij}\right) = \beta_1 + \beta_2 \, \text{Time}_{ij},$$

while for subjects assigned to the treatment group

$$E\left(Y_{ij}\right) = \left(\beta_1 + \beta_3\right) + \left(\beta_2 + \beta_4\right) \text{Time}_{ij}.$$

Thus each group's mean response is assumed to change linearly over time. This model with linear trends for two groups is depicted graphically in Figure 6.1, where the two groups have different intercepts and slopes. Here β_1 is the intercept in the control group (the "reference" group), while $(\beta_1 + \beta_3)$ is the intercept in the treatment group. The intercepts for each of the two groups have interpretation in terms of the mean response when $\text{Time}_{ij} = 0$; more generally, β_1 has interpretation as the mean response when all of the covariates are set to zero. Unless some care is taken with how the covariates are scaled (e.g., by centering all quantitative covariates prior to inclusion in the model), β_1 is not always readily interpretable and may represent an extrapolation beyond the data at hand. There can also be good reason, beyond issues of parameter interpretation, for centering the variable that denotes the time of measurement; this issue will be discussed later. Finally, the slope, or constant rate of change in the mean response per unit change in time, is β_2 in the control group, while the corresponding slope in the treatment group is $(\beta_2 + \beta_4)$. Ordinarily, in a longitudinal study the question of primary interest concerns a comparison of the changes in the mean response over time; this can be translated into a comparison of the slopes. Thus, if $\beta_4 = 0$, then the two groups do not differ in terms of changes in the mean response over time.

The model with linear trend over time is the simplest parametric "curve" that can be used to describe changes in the mean response over time. This model can easily incorporate both discrete (e.g., treatment or exposure group) and quantitative (e.g.,

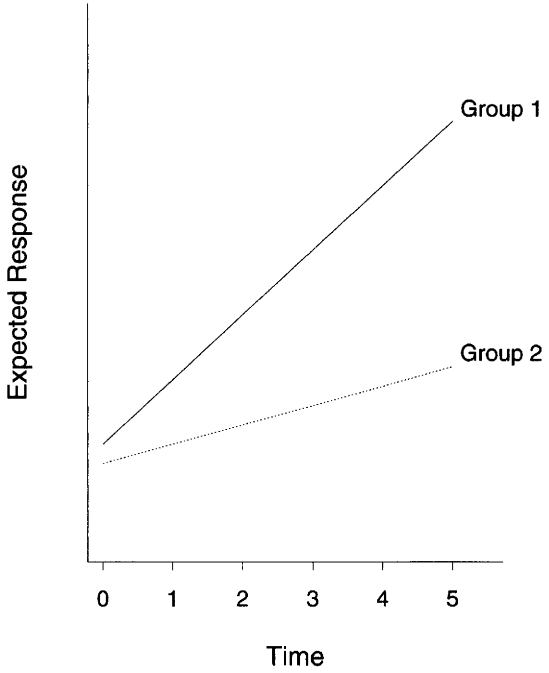

Fig. 6.1 Graphical representation of model with linear trends for two groups.

dose) covariates. Hypotheses about the dependence of changes in the mean response over time on covariates can be expressed in terms of hypotheses about whether the slope varies as a function of the covariates, that is, in terms of interactions between the covariates and the linear trend in time.

QUADRATIC TRENDS OVER TIME

When changes in the mean response over time are not linear, higher-order polynomial trends can be considered. For example, if the means are monotonically increasing or decreasing over the course of the study, but in a curvilinear way, a model with quadratic trends can be considered. In a quadratic trend model changes in the mean response are no longer constant (as in the linear trend model) throughout the duration of the study. Instead, the rate of change in the mean response depends on time; that is, the rate of change in the mean response depends on whether the focus is on changes that occur early or later in the study. As a result the rate of change must be expressed in terms of two parameters.

Consider the hypothetical two-group study comparing a novel *treatment* and a *control* discussed earlier. Assuming that the changes in the mean response can be approximated by quadratic trends, the following model can be adopted:

$$\begin{aligned} E\left(Y_{ij}\right) &= \beta_1 + \beta_2 \,\text{Time}_{ij} + \beta_3 \,\text{Time}_{ij}^2 + \beta_4 \,\text{Group}_i \\ &\quad + \beta_5 \,\text{Time}_{ij} \times \text{Group}_i + \beta_6 \,\text{Time}_{ij}^2 \times \text{Group}_i. \end{aligned} \tag{6.2}$$

In this model the mean response over time for subjects in the control group is given by

$$E\left(Y_{ij}\right) = \beta_1 + \beta_2 \,\text{Time}_{ij} + \beta_3 \,\text{Time}_{ij}^2;$$

while the corresponding mean response over time in the novel treatment group is given by

$$E\left(Y_{ij}\right) = \left(\beta_1 + \beta_4\right) + \left(\beta_2 + \beta_5\right) \text{Time}_{ij} + \left(\beta_3 + \beta_6\right) \text{Time}_{ij}^2.$$

This model with quadratic trends for two groups is depicted graphically in Figure 6.2, where the two groups have different intercepts (or mean response at time 0) and non-constant rates of change over time that differ between the two groups.

Note that in the quadratic trends model the mean response changes at a different rate, depending on Time_{ij}. For example, the rate of change in the control group is given by $\beta_2 + 2\,\beta_3 \,\text{Time}_{ij}$ (the derivation of this instantaneous rate of change requires some familiarity with calculus and is omitted). Thus early in the study when $\text{Time}_{\text{ij}} = 1$, the rate of change in the mean response is $\beta_2 + 2\,\beta_3$, whereas later in the study, say $\text{Time}_{\text{ij}} = 4$, the rate of change in the mean response is $\beta_2 + 8\,\beta_3$. The rate of change is different at the two occasions and the magnitude and sign of the regression coefficients β_2 and β_3 determine whether the mean response is increasing or decreasing over time and how the rate of change depends on time. The regression coefficients, $(\beta_2 + \beta_5)$ and $(\beta_3 + \beta_6)$, have similar interpretations for the treatment group.

When fitting polynomial trend models, one must take care to avoid extrapolation beyond the data at hand. While polynomial trend models can fit a flexible class of curves to the data, inferences beyond the measurement occasions should be avoided as these will be sensitive to the underlying model assumptions. While a quadratic trend might be a reasonable approximation for the data, recall that a quadratic trend necessarily has a turning point where the trend changes (e.g., from an increasing trend over time to a decreasing trend, or vice versa). In the absence of a strong theoretical rationale for the model, extrapolation beyond the data can produce nonsensical results and should be avoided.

With polynomial trend models there is a natural hierarchy of effects that has implications for testing hypotheses about linear, quadratic, and higher-order polynomial trends. That is, higher-order terms should be tested (and, if appropriate, removed from the model) before lower-order terms are assessed. Thus in the quadratic model

$$E\left(Y_{ij}\right) = \beta_1 + \beta_2 \,\text{Time}_{ij} + \beta_3 \,\text{Time}_{ij}^2,$$

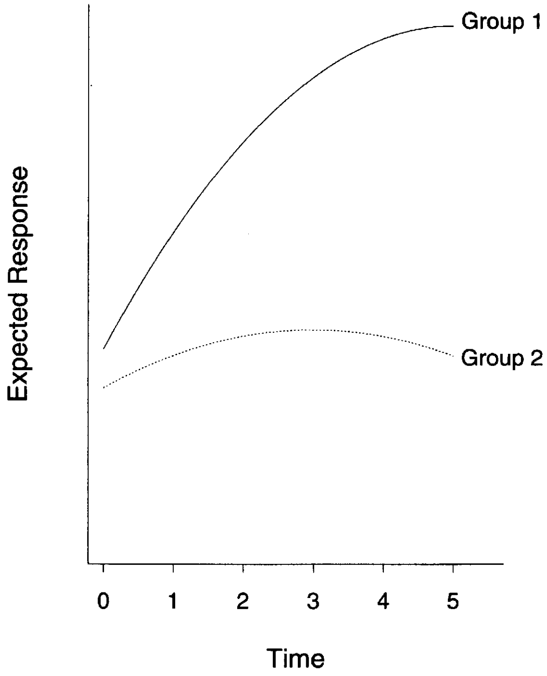

Fig. 6.2 Graphical representation of model with quadratic trends for two groups.

it is not meaningful or appropriate to test the coefficient for the linear trend, β_2, in a model that also includes a coefficient for the quadratic trend, β_3. Instead, a test for quadratic trend (versus linear trend) can be performed by testing the null hypothesis that $\beta_3 = 0$. If this null hypothesis cannot be rejected, it is then appropriate to remove the quadratic term from the model and consider the model with only linear trend. The test for linear trend is performed by testing the null hypothesis that $\beta_2 = 0$ in the model that only includes the linear term. This hierarchy is completely analogous to the testing of interactions; that is, tests of main effects (or attempts to interpret the main effects) are not meaningful in the presence of interaction. By the same token, tests of lower-order polynomials in time (e.g., linear trend) are not meaningful in the presence of higher-order polynomials in time (e.g., quadratic trend).

Finally, we return to the issue of "centering" variables. Although "centering" Time_j at zero leads to a simple interpretation of the intercept when Time_j represents time since baseline, to avoid potential problems of collinearity in the quadratic (or in any higher-order polynomial) trend model, it is advisable to "center" Time_j on its mean value. That is, prior to the analysis, replace Time_j by its deviation from the mean of $\text{Time}_1, \text{Time}_2, ..., \text{Time}_n$. To highlight the impact of this centering, consider the following example where $\text{Time}_j \in \{0, 1, 2, ..., 10\}$. The correlation between Time_j and Time_j^2 is 0.963. When two covariates are so highly correlated, computational problems associated with collinearity may arise in the estimation of β. Centering Time_j *before* quadratic (or any higher-order polynomial) terms are included in the model helps alleviate problems associated with collinearity. For example, when $\text{Time}_j \in \{0, 1, 2, ..., 10\}$ and we create a "centered" variable, say $\text{CTime}_j = (\text{Time}_j - 5.0)$, where 5.0 is the mean of $\{0, 1, 2, ..., 10\}$, then the correlation between CTime_j and CTime_j^2 is zero, thereby avoiding any potential problems associated with collinearity. With balanced longitudinal data (requiring only a single index j for Time_j), it is natural to center Time_j on the mean of $\text{Time}_1, \text{Time}_2, ..., \text{Time}_n$. However, with unbalanced longitudinal data, it is important to center Time_{ij} at some common value for all individuals. By centering at a common value, the regression intercept is interpretable as the mean response at that common value for time. In general, centering of Time_{ij} at individual-specific values (e.g., the mean of the n_i times of measurement for the i^{th} individual) should be avoided as these may vary considerably from one individual to another, thereby making the interpretation of the regression intercept meaningless.

6.3 LINEAR SPLINES

Simple parametric curves can provide a parsimonious description of longitudinal trends in the mean response. The simplest case is the linear trend model that characterizes change in the mean response over time in terms of a single slope parameter representing a constant rate of change. By introducing higher-order polynomials in time, various kinds of non-linearities in the longitudinal trends can also be accommodated. However, as the degree of the polynomial increases, the interpretation of the regression coefficients becomes more difficult. As a result the use of polynomials in time is most appealing when any non-linearity can be approximated by quadratic trends.

In some applications the longitudinal trends in the mean response cannot be characterized by first- and second-degree polynomials in time (i.e., linear or quadratic trends). In addition there are other applications where non-linear trends in the mean response cannot be well approximated by polynomials in time of any order. This will most often occur when the mean response increases (or decreases) rapidly for some duration and then more slowly thereafter (or vice versa). When this type of pattern of change arises, it can often by handled by using linear spline models.

If the simplest possible curve is a straight line, then one way to extend the curve is to have a sequence of joined or connected line segments that produces a piecewise linear pattern. Linear spline models provide a very useful and flexible way to accommodate many of the non-linear trends that cannot be approximated by simple polynomials in time. The basic idea behind linear spline models is remarkably simple: divide the time axis into a series of segments and consider a model for the trend over time that is comprised of piecewise linear trends, having different slopes within each segment but joined or tied together at fixed times. The locations where the lines meet or are tied

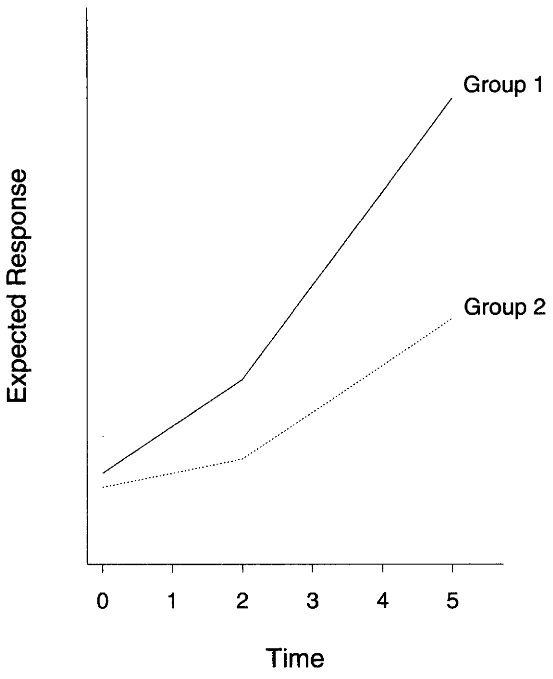

Fig. 6.3 Graphical representation of model with linear splines for two groups, with a common knot at Time = 2.

together are known as the "knots." This model allows the mean response to increase or decrease as time proceeds, depending on the sign and magnitude of the regression slopes for the line segments. The resulting piecewise linear curve is called a spline. Figure 6.3 provides an illustration of a linear spline model for two groups with a common knot at time 2. Note that the slopes of the two lines, before and after time 2, are different, with a greater increase in the mean response in the second time segment, and a more attenuated increase in the mean response in the first time segment. This spline model is sometimes referred to as a piecewise linear or "broken-stick" model.

The simplest possible spline model has only one knot and can be parameterized in a number of different ways. Returning to the hypothetical two-group study comparing a novel *treatment* and a *control* discussed earlier, if the mean response changes over time in a piecewise linear way, we can fit the following linear spline model with knot at t*:

$$\begin{aligned} E\left(Y_{ij}\right) &= \beta_1 + \beta_2 \,\text{Time}_{ij} + \beta_3 \left(\text{Time}_{ij} - \text{t}^*\right)_+ + \beta_4 \,\text{Group}_i \\ &\quad + \beta_5 \,\text{Time}_{ij} \times \text{Group}_i + \beta_6 \left(\text{Time}_{ij} - \text{t}^*\right)_+ \times \text{Group}_i, \end{aligned} \tag{6.3}$$

where $(x)_+$, known as a *truncated line function*, is defined as a function that equals x when x is positive and is equal to zero otherwise. Thus $(\text{Time}_{\text{ij}} - \text{t}^*)_+$ is equal to $(\text{Time}_{ij} - \text{t}^*)$ when $\text{Time}_{ij} > \text{t}^*$ and is equal to zero when $\text{Time}_{ij} \leq \text{t}^*$. In the model given by (6.3) the means for subjects in the control group are

$$E\left(Y_{ij}\right) = \beta_1 + \beta_2 \,\text{Time}_{ij} + \beta_3 \left(\text{Time}_{ij} - \text{t}^*\right)_+.$$

When expressed in terms of the mean response prior to and after t*,

$$\begin{aligned} E\left(Y_{ij}\right) &= \beta_1 + \beta_2 \,\text{Time}_{ij}, && \text{Time}_{ij} \leq \text{t}^*; \\ E\left(Y_{ij}\right) &= \left(\beta_1 - \beta_3 \text{t}^*\right) + \left(\beta_2 + \beta_3\right) \text{Time}_{ij}, && \text{Time}_{ij} > \text{t}^*. \end{aligned}$$

Thus, in the control group, the slope prior to t* is β_2 and following t* is $(\beta_2 + \beta_3)$. Similarly the means for subjects in the treatment group are given by

$$E\left(Y_{ij}\right) = \left(\beta_1 + \beta_4\right) + \left(\beta_2 + \beta_5\right) \text{Time}_{ij} + \left(\beta_3 + \beta_6\right) \left(\text{Time}_{ij} - \text{t}^*\right)_+.$$

When expressed in terms of the mean response prior to and after t*,

$$\begin{aligned} E\left(Y_{ij}\right) &= \left(\beta_1 + \beta_4\right) + \left(\beta_2 + \beta_5\right) \text{Time}_{ij}, && \text{Time}_{ij} \leq \text{t}^*; \\ E\left(Y_{ij}\right) &= \left\{\left(\beta_1 + \beta_4\right) - \left(\beta_3 + \beta_6\right) t^*\right\} \\ &\quad + \left(\beta_2 + \beta_3 + \beta_5 + \beta_6\right) \text{Time}_{ij}, && \text{Time}_{ij} > \text{t}^*. \end{aligned}$$

Then, in terms of group comparisons, the null hypothesis of no group differences in patterns of change over time can be expressed as H_0: $\beta_5 = \beta_6 = 0$. Comparisons of the groups before and after t* are also possible. For example, the null hypothesis of no group differences in patterns of change prior to t* can be expressed as H_0: $\beta_5 = 0$.

The simple spline model considered so far can be extended to include more than one knot or more than two joined line segments. More generally, a spline model with K knots or break-points will produce $K + 1$ line segments, and there will be $K + 1$ corresponding slopes; see Chapter 19 on "smoothing" longitudinal data for a detailed description of splines models with many knots. Thus, in principle, it is possible to accommodate quite complex non-linear patterns for changes in the mean response by including a sufficient number of variables, $(\text{Time}_{ij} - \text{t}_k^*)_+$, with knots located at t_k^* (for $k = 1, ..., K$). However, in practice, the data from many longitudinal studies can be well-approximated by simple piecewise linear models with at most one or two knots that are located at judiciously chosen time points.

Finally, our discussion thus far has avoided the thorny problem of the choice of location(s) for the knot(s). There is an extensive body of research in statistics on automated choices for the knot location, where the location is effectively determined by the data at hand. Ideally the choice of knot location should also incorporate subject-matter considerations. For example, in studies of growth, certain ages are associated with growth spurts. Similarly measures of hormonal response are known to change quite dramatically with the onset of puberty and menopause. In other settings, there may be a body of evidence that the response profile changes in a discernible way at certain time points. An example of the latter is in studies of HIV-infected patients. In early studies of the treatment of HIV-infected patients with AZT, it was increasingly recognized that CD4 counts, a measure of the body's immune response, increased sharply over a 4- to 6-week period following treatment with AZT, but then leveled off. In summary, the choice of knot location is a mixture of art and science. When it is available, subject-matter knowledge should be brought to bear on the empirical evidence for the most appropriate choice of knot location (see Chapter 19 for a more detailed discussion on choice of knot location).

6.4 GENERAL LINEAR MODEL FORMULATION

Below we demonstrate how both polynomial trend and spline models can be expressed in terms of the general linear model

$$E\left(Y_i|X_i\right) = \mu_i = X_i\beta,$$

for appropriate choices of X_i. Let n_i be the number of repeated measures on the i^{th} individual ($i = 1, ..., N$). To illustrate how the polynomial trend model can be expressed in terms of the general linear model, consider the hypothetical two-group study comparing a novel *treatment* and a *control* discussed earlier. Let us assume that the mean response changes over time in a quadratic trend. Then the design matrix X_i has the following form for the control group:

$$X_i = \begin{pmatrix} 1 & t_{i1} & t_{i1}^2 & 0 & 0 & 0 \\ 1 & t_{i2} & t_{i2}^2 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & t_{in_i}^2 & 0 & 0 & 0 \end{pmatrix},$$

while for the treatment group the design matrix is given by

$$X_i = \begin{pmatrix} 1 & t_{i1} & t_{i1}^2 & 1 & t_{i1} & t_{i1}^2 \\ 1 & t_{i2} & t_{i2}^2 & 1 & t_{i2} & t_{i2}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & t_{in_i}^2 & 1 & t_{in_i} & t_{in_i}^2 \end{pmatrix},$$

where t_{ij} denotes the time of the j^{th} measurement on the i^{th} individual. Then, in terms of the general linear model

$$E\left(Y_i|X_i\right) = \mu_i = X_i\beta,$$

where $\beta = (\beta_1, ..., \beta_6)'$ is a 6×1 vector of regression coefficients, the mean responses in the control group are

$$\mu_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{in_i} \end{pmatrix} = \begin{pmatrix} \beta_1 + \beta_2 t_{i1} + \beta_3 t_{i1}^2 \\ \beta_1 + \beta_2 t_{i2} + \beta_3 t_{i2}^2 \\ \vdots \\ \beta_1 + \beta_2 t_{in_i} + \beta_3 t_{in_i}^2 \end{pmatrix},$$

while the mean responses in the treatment group are

$$\mu_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{in_i} \end{pmatrix} = \begin{pmatrix} (\beta_1 + \beta_4) + (\beta_2 + \beta_5) t_{i1} + (\beta_3 + \beta_6) t_{i1}^2 \\ (\beta_1 + \beta_4) + (\beta_2 + \beta_5) t_{i2} + (\beta_3 + \beta_6) t_{i2}^2 \\ \vdots \\ (\beta_1 + \beta_4) + (\beta_2 + \beta_5) t_{in_i} + (\beta_3 + \beta_6) t_{in_i}^2 \end{pmatrix}.$$

For the spline model, let us assume that the mean response changes over time in a piecewise linear way, with knot at $t^* = 4$. Then the design matrix X_i has the following form for the control group:

$$X_i = \begin{pmatrix} 1 & t_{i1} & (t_{i1} - 4)_+ & 0 & 0 & 0 \\ 1 & t_{i2} & (t_{i2} - 4)_+ & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & (t_{in_i} - 4)_+ & 0 & 0 & 0 \end{pmatrix},$$

while for the treatment group the design matrix is given by

$$X_i = \begin{pmatrix} 1 & t_{i1} & (t_{i1} - 4)_+ & 1 & t_{i1} & (t_{i1} - 4)_+ \\ 1 & t_{i2} & (t_{i2} - 4)_+ & 1 & t_{i2} & (t_{i2} - 4)_+ \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & (t_{in_i} - 4)_+ & 1 & t_{in_i} & (t_{in_i} - 4)_+ \end{pmatrix}.$$

The spline model can then be expressed in terms of the general linear model,

$$E\left(Y_i | X_i\right) = \mu_i = X_i \beta,$$

where $\beta = (\beta_1, ..., \beta_6)'$ is a 6×1 vector of regression coefficients.

Given that both the polynomial trend and spline models can be expressed in terms of the general linear regression model,

$$E\left(Y_i | X_i\right) = \mu_i = X_i \beta,$$

restricted maximum likelihood estimation of β, and the construction of confidence intervals and tests of hypotheses, are possible once the covariance of Y_i has been specified. Unlike the analysis of response profiles, where the covariance of Y_i is assumed to be unstructured with no constraints on the covariance parameters other than the requirement that they yield a symmetric matrix (and one that is positive-definite), more parsimonious models for the covariance can be adopted. Indeed, the use of parametric curves for the mean response is most appealing in settings where the longitudinal data are inherently unbalanced over time. As a result an unstructured covariance matrix may not be well-defined, let alone estimated, when, in principle, each individual can have a unique sequence of measurement times. However, the discussion of models for the covariance is postponed until Chapter 7; here we simply assume that some appropriate model for the covariance has been adopted. Given models for both the mean and covariance, REML estimates of β, and their standard errors (based on the estimated covariance of $\widehat{\beta}$), can be obtained using the method of estimation described in Chapter 4.

6.5 CASE STUDIES

We illustrate the main ideas by considering polynomial trend models for data on lung function (FEV_1) from a longitudinal epidemiologic study of current and former smokers aged 36 and older. The application of spline models is illustrated using the blood lead data on the 100 children from the treatment and placebo groups of the Treatment of Lead-Exposed Children (TLC) Trial.

The Vlagtwedde–Vlaardingen Study

In an epidemiologic study conducted in two different areas in The Netherlands, the rural area of Vlagtwedde in the northeast and the urban, industrial area of Vlaardingen in the southwest, residents were followed over time to obtain information on the prevalence of and risk factors for chronic obstructive lung diseases (van der Lende et al., 1981; Rijcken et al., 1987). Here we focus on a sub-sample of men and women from the rural area of Vlagtwedde. The sample, initially aged 15 to 44, participated in follow-up surveys approximately every 3 years for up to 21 years. At each survey, information on respiratory symptoms and smoking status was collected by questionnaire and spirometry was performed. Pulmonary function was determined by spirometry and a measure of forced expiratory volume (FEV_1) was obtained every three years for the first 15 years of the study, and also at year 19.

In this study, FEV_1 was not recorded for every subject at each of the planned measurement occasions. That is, the data are unbalanced due to incompleteness. The number of repeated measurements of FEV_1 on each subject varied from 1 to 7. For the purpose of this illustration we focus on a subset of the data on 133 residents aged 36 or older at their entry into the study and whose smoking status did not change over the 19 years of follow-up. Each study participant was either a current or former smoker. Current smoking was defined as smoking at least one cigarette per day. The trends in the mean FEV_1 over time, for current and former smokers, are displayed in Figure

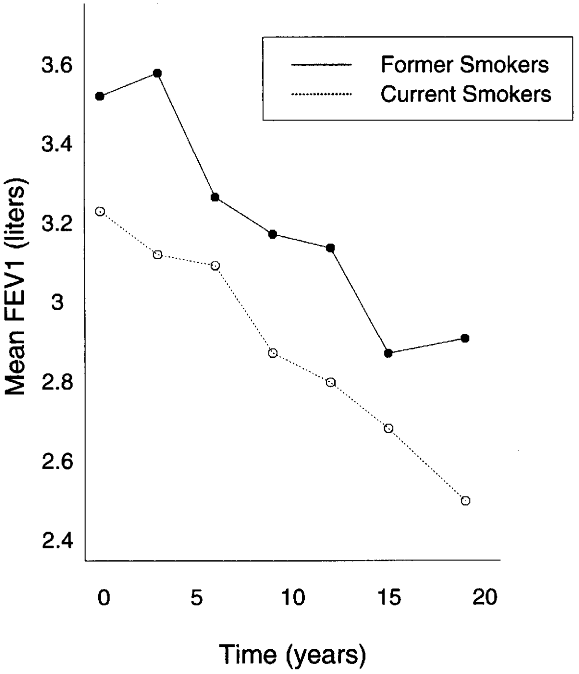

Fig. 6.4 Mean FEV_1 at baseline (year 0), year 3, year 6, year 9, year 12, year 15, and year 19 in the current and former smoking exposure groups.

6.4. The goal of our analysis is to describe changes in lung function over the 19 years of follow-up with parametric curves and to determine whether the time trends differ for current and former smokers. We summarize differences in mean change between current and former smokers, assuming that the change does not depend strongly on either age or gender (neither variable was available in the data set).

First we consider a linear trend in the mean response over time, with intercepts and slopes that differ for the two smoking exposure groups. For all of the analyses reported here, we assume an unstructured covariance matrix. Based on the REML estimates of the regression coefficients in Table 6.1, the mean response for participants who are former smokers is estimated to be

$$E\left(Y_{ij}\right) = 3.507 - 0.033 \text{ Time}_{ij},$$

while for participants who are current smokers

$$\begin{aligned} E\left(Y_{ij}\right) &= (3.507 - 0.262) - (0.033 + 0.005) \text{ Time}_{ij} \\ &= 3.245 - 0.038 \text{ Time}_{ij}. \end{aligned}$$

Table 6.1 Estimated regression coefficients and standard errors based on a model with linear trends for the FEV_1 data from the Vlagtwedde–Vlaardingen study.

Variable	Smoking Group	Estimate	SE	Z
Intercept		3.5073	0.1004	34.94
Smoke	Current	−0.2617	0.1151	−2.27
Time		−0.0332	0.0031	−10.84
Smoke × Time	Current	−0.0050	0.0035	−1.42

Table 6.2 Comparison of the maximized (ML) log-likelihoods for the model with linear and quadratic trends for the FEV_1 data from the Vlagtwedde–Vlaardingen study.

Model	−2 (ML) Log-Likelihood
Quadratic Trend Model	237.2
Linear Trend Model	238.5
−2 × Log-Likelihood Ratio: $G^2 = 1.3$, 2 df, $(p > 0.50)$	

Thus it would appear that both groups have a significant decline in mean FEV_1 over time, but there is no discernible difference between the two smoking exposure groups in the constant rate of change, since the $\text{Smoke}_i \times \text{Time}_{ij}$ interaction (i.e., the comparison of the two slopes) is not significant, with $Z = -1.42$, $p > 0.15$.

The adequacy of the linear trend model can be assessed by including higher-order polynomial trends. For example, we can consider a model that allows quadratic trends for changes in FEV_1 over time. Recall that the linear trend model is nested within the quadratic trend model. If the linear trend model is adequate for these data, the difference in maximized log-likelihoods (or the likelihood ratio test statistic) should not be large. The maximized log-likelihoods for the models with linear and quadratic trends are presented in Table 6.2. The likelihood ratio test statistic can be compared to a chi-squared distribution with 2 degrees of freedom (or 6, the number of parameters in the quadratic trend model, minus 4, the number of parameters in the linear trend model). Note that the likelihood ratio test is based on the ML, not REML, log-likelihood because we are comparing two different models for the mean; as a result both models were re-fit using ML estimation. For both models the polynomial trends over time are allowed to differ for the two smoking exposure groups. The likelihood

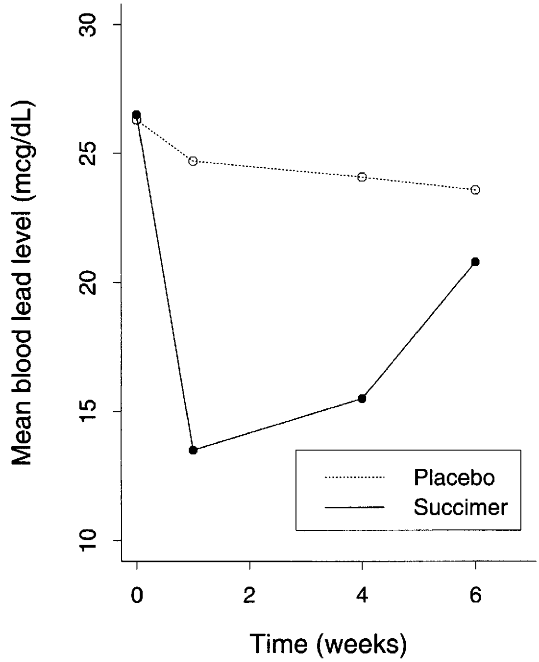

Fig. 6.5 Mean blood lead levels at baseline (week 0), week 1, week 4, and week 6 in the succimer and placebo groups.

ratio test (comparing the quadratic and linear trend models) produces $G^2 = 1.3$, with 2 degrees of freedom $(p > 0.50)$. Thus, when compared to the quadratic trend model, the linear trend model appears to be adequate for these data. Finally, for illustrative purposes we can make a comparison with a cubic trend model. This produces a likelihood ratio test statistic, $G^2 = 4.4$, with 4 degrees of freedom $(p > 0.35)$, indicating again that the linear trend model is adequate for these data.

Treatment of Lead-Exposed Children Trial

Recall that the TLC trial was a placebo-controlled, randomized trial of a chelating agent, succimer, in children with confirmed blood lead levels of 20 to 44 μg/dL. The children in the trial were aged 12 to 33 months and lived in deteriorating inner city housing. The following analyses are based on data on blood lead levels at baseline (or week 0), week 1, week 4, and week 6. Note that from the plot of the means in Figure 6.5, it would appear that only the mean blood lead levels in the placebo group can be described by a linear trend; the mean in the succimer group decreases from baseline to week 1, but then increases thereafter.

Table 6.3 Estimated regression coefficients and standard errors based on a piecewise linear model, with common knot at week 1, for the blood lead level data from the TLC trial.

Variable	Group	Estimate	SE	Z
Intercept		26.3422	0.4991	52.78
Week		−1.6296	0.7818	−2.08
$(\text{Week} - 1)_+$		1.4305	0.8777	1.63
Group × Week	S	−11.2500	1.0924	−10.30
Group × $(\text{Week} - 1)_+$	S	12.5822	1.2278	10.25

Given that there are non-linearities in the trends over time, higher-order polynomial models (e.g., a quadratic trend model) could be fit to the data. However, to illustrate the application of spline models, we accommodate the non-linearity with a piecewise linear model with common knot at week 1,

$$\begin{aligned} E\left(Y_{ij}\right) &= \beta_1 + \beta_2\,\text{Week}_{ij} + \beta_3\left(\text{Week}_{ij} - 1\right)_+ + \beta_4\,\text{Group}_i \times \text{Week}_{ij} \\ &\quad + \beta_5\,\text{Group}_i \times \left(\text{Week}_{ij} - 1\right)_+, \end{aligned}$$

where $\text{Group}_i = 1$ if assigned to succimer, and $\text{Group}_i = 0$ otherwise. Because of the randomization of children to the two treatment groups, the model does not contain a main effect of Group, and we assume a common mean blood lead level at baseline. In this piecewise linear model, the means for subjects in the placebo group are given by

$$E\left(Y_{ij}\right) = \beta_1 + \beta_2\,\text{Week}_{ij} + \beta_3\left(\text{Week}_{ij} - 1\right)_+,$$

while in the succimer group the means are given by

$$E\left(Y_{ij}\right) = \beta_1 + \left(\beta_2 + \beta_4\right)\text{Week}_{ij} + \left(\beta_3 + \beta_5\right)\left(\text{Week}_{ij} - 1\right)_+.$$

The REML estimates of the regression parameters from the piecewise linear model are given in Table 6.3. When expressed in terms of the mean response prior to and after week 1, the estimated means in the placebo group are

$$\begin{aligned} \widehat{\mu}_{ij} &= \widehat{\beta}_1 + \widehat{\beta}_2\,\text{Week}_{ij}, & \text{Week}_{ij} &\leq 1; \\ \widehat{\mu}_{ij} &= (\widehat{\beta}_1 - \widehat{\beta}_3) + (\widehat{\beta}_2 + \widehat{\beta}_3)\,\text{Week}_{ij}, & \text{Week}_{ij} &> 1. \end{aligned}$$

Thus in the placebo group the slope prior to week 1 is $\widehat{\beta}_2 = -1.63$ and, following week 1, is $(\widehat{\beta}_2 + \widehat{\beta}_3) = -1.63 + 1.43 = -0.20$. Similarly, when expressed in terms of the mean response prior to and after week 1, the estimated means for subjects in

Table 6.4 Estimated mean blood lead levels for the placebo and succimer groups from the piecewise linear model, with common knot at week 1; the observed means are in parentheses.

Group	Week 0	Week 1	Week 4	Week 6
Succimer	26.3	13.5	16.9	19.1
	(26.5)	(13.5)	(15.5)	(20.8)
Placebo	26.3	24.7	24.1	23.7
	(26.3)	(24.7)	(24.1)	(23.6)

the succimer group are given by

$$\begin{aligned}\widehat{\mu}_{ij} &= \widehat{\beta}_1 + (\widehat{\beta}_2 + \widehat{\beta}_4)\,\text{Week}_{ij}, && \text{Week}_{ij} \leq 1;\\ \widehat{\mu}_{ij} &= \widehat{\beta}_1 - (\widehat{\beta}_3 + \widehat{\beta}_5)\\ &\quad + (\widehat{\beta}_2 + \widehat{\beta}_3 + \widehat{\beta}_4 + \widehat{\beta}_5)\,\text{Week}_{ij}, && \text{Week}_{ij} > 1.\end{aligned}$$

The estimates of the mean blood lead levels for the placebo and succimer groups are presented in Table 6.4. The estimated means from the piecewise linear model appear to adequately fit the observed mean response profiles for the two treatment groups.

Note that the model with linear trends (and common intercept) is nested within the piecewise linear model (since the former can be obtained by setting $\beta_3 = \beta_5 = 0$ in the latter). When these two models are compared in terms of their maximized log-likelihoods (see Table 6.5), the likelihood ratio test statistic is $G^2 = 121.8$ and can be compared to a chi-squared distribution with 2 degrees of freedom (or 5, the number of parameters in the linear spline model, minus 3, the number of parameters in the linear trend model). Note that the likelihood ratio test is based on the ML, not REML, log-likelihood because we are comparing two different models for the mean; as a result both models were re-fit using ML. The magnitude of the likelihood ratio test statistic (with $p < 0.0001$) indicates that the piecewise linear model significantly improves the overall fit to the mean response over time when compared to a linear trend model. This simply confirms what was already obvious from the plot of the means in Figure 6.5. Although the piecewise linear and quadratic trend models (with common intercept for the two treatment groups) are not nested, they both have the same number of parameters and therefore their respective log-likelihoods can be directly compared (see Table 6.5). From a comparison of the maximized log-likelihoods it is apparent that the piecewise linear model fits these data discernibly better than the quadratic trend model (-2 ML log-likelihood = 2436.2 for the piecewise linear model versus -2 ML log-likelihood = 2551.7 for the quadratic trend model).

Table 6.5 Comparison of the maximized (ML) log-likelihoods for the models with linear and quadratic trends, and piecewise linear trend with common knot at week 1, for the blood lead level data from the TLC trial.

Model	−2 (ML) Log-Likelihood
Piecewise Linear (Spline) Model	2436.2
Quadratic Trend Model	2511.7
Linear Trend Model	2558.0

Table 6.6 Estimated unstructured covariance matrix for the linear trend model for the blood lead levels at baseline, week 1, week 4, and week 6 for the children from the TLC trial.

	Covariance Matrix		
25.5	13.8	16.1	21.4
13.8	111.2	81.2	38.4
16.1	81.2	78.3	36.8
21.4	38.4	36.8	59.4

Finally, it is quite instructive to examine the estimated unstructured covariance matrix for the linear trend model, a model that does not fit these data well. In Table 6.6 the REML estimates of the unstructured covariance matrix are presented. Note that the estimated variances at weeks 1 and 4 are approximately three to four times greater than at baseline. Moreover the estimated covariance matrix is discernibly different from that obtained in the analysis of response profiles in Section 5.4 that placed no structure on the means. The inflation of the variance at weeks 1 and 4 is indirectly an indication that the lack of fit to the means at these two occasions is being attributed to error variability, and hence the inflation of the variance at these two occasions. This highlights an important issue that will be discussed in greater detail in Chapter 7, namely that there is an interdependence between the mean response and the covariance that has important implications for how these two aspects of longitudinal data are jointly modeled.

In conclusion, one of the main aspects of summarizing trends in the mean response over time via parametric curves is that trends over time and their relation to covariates can be expressed as a function of a small number of parameters. That is, covariate effects on changes in the mean response over time can be captured in one or two regression parameters, leading to more powerful tests when the models are appropriate

Table 6.7 Illustrative commands for a linear trend model using PROC MIXED in SAS.

```
PROC MIXED;
    CLASS id group t;
    MODEL y=group time group*time / S CHISQ;
    REPEATED t / TYPE=UN SUBJECT=id R RCORR;
```

for the data at hand. Also the parametric curves define the conditional mean of Y_{ij}, $E(Y_{ij}|X_{ij})$, as an explicit function of the times of measurement, Time_{ij}. As a result there is no reason to require all individuals to have the same set of measurement times, nor even the same number of measurements. In our examples we have used only data sets where subjects are measured at the same set of occasions, but this is because we will require models for the covariance when subjects are measured at arbitrary points in time. Hence examples of this will be given in the next chapter.

6.6 COMPUTING: FITTING PARAMETRIC CURVES USING PROC MIXED IN SAS

To fit a linear trend model to data from two or more groups measured repeatedly over time, we can use the illustrative SAS commands given in Table 6.7. Note that this model assumes that the covariance matrix is unstructured. In principle, alternative assumption about the covariance can be considered. Indeed, when the data are unbalanced over time, it will be necessary to consider parametric models for the covariance; this topic will be discussed in greater detail in Chapter 7.

Note that the CLASS statement includes a variable `t`. This variable is an additional copy of the variable `time`. The difference is that while `t` is declared as a categorical variable on the CLASS statement, `time` is not and is treated as a quantitative covariate in the MODEL statement. The reason for having two versions, `time` and `t`, one quantitative and the other categorical, is that it is good practice to include, wherever possible, a REPEATED effect. This ensures that the covariance is estimated correctly when the design is balanced but incomplete due to missingness or when the study is balanced and complete but the repeated measures are not in the same order for each subject in the data set (e.g., this might arise when the data set has previously been sorted on another variable). With unbalanced data it will very often not be possible to include a REPEATED effect; instead, the covariance model will need to be defined explicitly in terms of the times of measurement. A further discussion of this point is postponed until Chapter 7.

Next we present illustrative commands for fitting a quadratic trend model in Table 6.8. The MODEL statement now includes both `time` and `timesqr`, the latter is simply an additional variable that is the square of `time` (i.e., `time`^{2}). Note that the MODEL

Table 6.8 Illustrative commands for a quadratic trend model using PROC MIXED in SAS.

```
PROC MIXED;
    CLASS id group t;
    MODEL y=group time timesqr group*time group*timesqr / S CHISQ;
    REPEATED t / TYPE=UN SUBJECT=id R RCORR;
```

Table 6.9 Illustrative commands for a spline model, with knot at `time` = 4, using PROC MIXED in SAS.

```
PROC MIXED;
    CLASS id group t;
    MODEL y=group time time_4 group*time group*time_4 / S CHISQ;
    REPEATED t / TYPE=UN SUBJECT=id R RCORR;
```

statement includes both main effects of `time` and `timesqr`, and their interactions with `group`.

Finally, we present illustrative commands for fitting spline models. In Table 6.9 we present commands in SAS for fitting a model with a single knot at `time` = 4. The MODEL statement includes `time` and `time_4`, where `time_4` is a derived variable for $(\texttt{time} - 4)_+$. The latter variable can easily be computed in SAS as

$$\texttt{time\_4} = \text{MAX}(\texttt{time} - 4, 0);$$

6.7 FURTHER READING

A concise and clear discussion of how to describe patterns of change over time using polynomial trends can be found in Section 3.5 of the book by Hand and Taylor (1987). A general discussion of splines and piecewise linear regression can be found in Chapter 11 (Section 11.5) of Neter et al. (1996).

Bibliographic Notes

The use of simple parametric curves to describe changes in the mean response over time has its origins in growth curve analysis. Methods for estimation and testing of growth curves were developed by Wishart (1938), Box (1950), and Rao (1958). Potthoff and Roy (1964 proposed an extension of the repeated measures analysis by MANOVA for growth curves; alternative formulations were developed by Rao (1965), Khatri (1966), and Grizzle and Allen (1969).

An excellent discussion of spline models can be found in Chapter 3 of Ruppert et al. (2003), and the references therein.

Problems

6.1 In a study of weight gain (Box, 1950) investigators randomly assigned 30 rats to three treatment groups: treatment 1 was a control (no additive); treatments 2 and 3 consisted of two different additives (thiouracil and thyroxin respectively) to the rats drinking water. Weight, in grams, was measured at baseline (week 0) and at weeks 1, 2, 3, and 4. Due to an accident at the beginning of the study, data on 3 rats from the thyroxin group are unavailable.

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

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

ID Group Y_1 Y_2 Y_3 Y_4 Y_5

Note: The variable Group is coded 1 = control, 2 = thiouracil, and 3 = thyroxin.

6.1.1 On a single graph, construct a time plot that displays the mean weight versus time (in weeks) for the three groups. Describe the general characteristics of the time trends for the three groups.

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

6.1.3 Assume that the rate of increase in each group is approximately constant throughout the duration of the study. Assuming an unstructured covariance matrix, construct a test of whether the rate of increase differs in the groups.

6.1.4 On a single graph, construct a time plot that displays the *estimated* mean weight versus time (in weeks) for the three treatment groups from the results generated from Problem 6.1.3.

6.1.5 Based on the results from Problem 6.1.3, what is the estimated rate of increase in mean weight in the control group (group 1)? What is the estimated rate of increase in mean weight in the thiouracil group (group 2)? What is the estimated rate of increase in mean weight in the thyroxin group (group 3)?

6.1.6 The study investigators conjectured that there would be an increase in weight, but that the rate of increase would level-off toward the end of the study. They also conjectured that this pattern of change may differ in the three treatment groups. Assuming an unstructured covariance matrix, construct a test of this hypothesis.

6.1.7 Compare and contrast the results from Problems 6.1.3 and 6.1.6. Does a model with only a linear trend in time adequately account for the pattern of change in the three treatments groups? Provide results that support your conclusion.

6.1.8 Given the results of all the previous analyses, what conclusions can be drawn about the effect of the additives on the patterns of change in weight?