ST 732

Longitudinal Data Analysis

Lecture Notes

M. Davidian

Department of Statistics

North Carolina State University

Contents

1 Introduction and Motivation

1.1 Objective of this course

OBJECTIVE: The goal of this course is to provide an overview of statistical models and methods for the analysis of ***longitudinal data***; that is, data in the form of ***repeated measurements*** over time or other factor on each ***individual*** or ***unit*** (human subject, plant, plot, sample, etc.) in a sample from a population of interest.

Data are collected routinely in this fashion in a broad range of applications, including agriculture and the life sciences, health sciences research, and physical science and engineering. For example

- In agriculture, a measure of growth is taken on experimental plots ***weekly*** over the growing season. Plots are assigned to be treated with different fertilizers at the start of the season.
- In a study of human immunodeficiency virus (HIV) infection, a measure of viral load (roughly, concentration of HIV present in the blood) is made ***monthly*** on infected patients. Patients are assigned to take different "cocktails" of antiretroviral treatments at the start of the study.

A defining characteristic of these examples is that the ***same*** response is measured ***repeatedly*** on each unit; e.g., viral load is measured repeatedly on the same subject. This particular type of data structure is the focus of this course.

The scientific questions of interest often involve not only the usual kinds of questions, such as how the ***mean response*** differs across treatments, but also how the ***change in mean response over time*** differs and other features of the relationship between response and time. Thus, it is necessary to represent the situation in terms of a ***statistical model*** that acknowledges the way in which the data were collected to address these questions. Complementing the models, specialized methods of analysis are required.

Because the study of change is fundamental across almost all scientific disciplines, studies in which longitudinal data are collected have become ubiquitous, and interest in the most appropriate ways to represent and interpret longitudinal data has grown tremendously. In this course, we will study approaches to modeling these data, and we will explore the associated classical and more modern approaches to analyzing them in detail.

TERMINOLOGY: Although the term ***longitudinal*** suggests that data are collected over ***time***, the models and methods we will discuss are more broadly applicable to any kind of ***repeated measurement*** data. That is, although repeated measurement most often takes place over time, this is not the only way that measurements can be taken repeatedly on the same individual or unit. For example,

- Individuals may be human subjects. On each subject, ***prothrombin time***, a measure of how long it takes blood to clot, is measured on several occasions, each involving administration of a different dose of an anti-coagulant agent. Thus, a subject is measured repeatedly over ***dose***.
- Units may be trees in a forest. For each tree, measurements of the diameter of the tree are made at several different points along the trunk of the tree. Thus, the tree is measured repeatedly over ***positions*** along the trunk.
- Individuals may be pregnant female rats. Each gives birth to a litter of pups, and the birthweight of each pup is recorded. Thus, the rat is measured repeatedly over each of her ***pups***.

The third example differs from the other two in that there is no natural ***order*** to the repeated measurements.

Thus, the methods apply more broadly than the strict definition of the term ***longitudinal data*** indicates – this term will mean, to us, data in the form of ***repeated measurements*** that might be over time, but might alternatively be over some other set of conditions. Because time is most often the condition of measurement, however, many of our examples will indeed involve repeated measurement over time, and we will often use the word ***time*** to refer generically to the repeated measurement condition.

We use the terms ***response*** or ***outcome*** to denote the repeated measurement or outcome of interest. Because longitudinal studies are frequently conducted with human or animal subjects, we use the terms ***unit***, ***individual***, and ***subject*** interchangeably.

1.2 Examples

To set the stage, we consider several data sets from a variety of applications. These not only provide concrete examples of longitudinal data situations, but serve to illustrate the range of ways that data are collected and the types of responses and questions that may be of interest.

EXAMPLE 1: The orthodontic study data of Potthoff and Roy (1964). This is a world famous data set that is used to introduce features of longitudinal data modeling and analysis. A study was conducted involving 27 children, 16 boys and 11 girls. On each child, the distance (mm) from the center of the pituitary to the pterygomaxillary fissure was made at ages 8, 10, 12, and 14 years of age. The pterygomaxillary fissure is a vertical opening in the human skull, depicted in Figure 1.1.

Figure 1.1: *Pterygomaxillary fissure.*

In Figure 1.2, the distance measurements are plotted against age for each child.

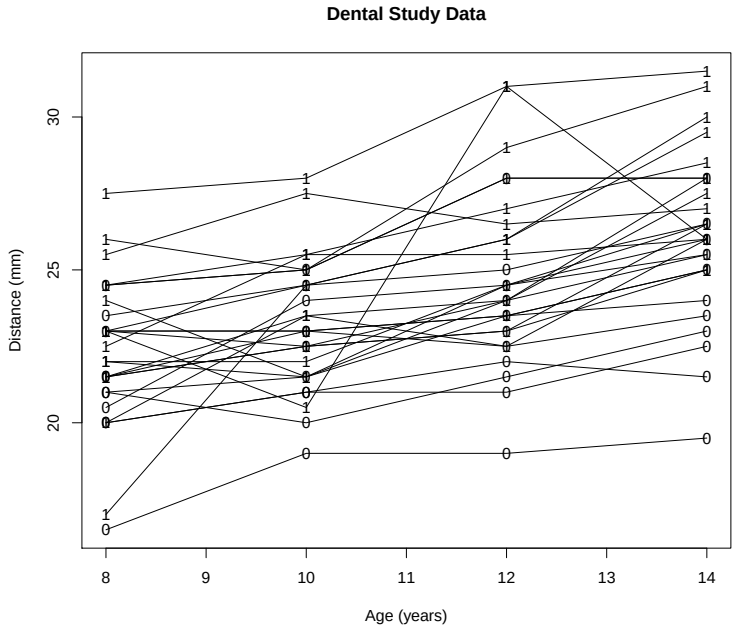

Figure 1.2: *Orthodontic distance measurements (mm) for 27 children at ages 8, 10, 12, 14. The plotting symbols are 0's for girls, 1's for boys.*

The plotting symbols denote girls (0) and boys (1), and the trajectory for each child is connected by a solid line so that individual child patterns can be seen. Plots like Figure 1.2 are often called ***spaghetti plots***, for obvious reasons.

The objectives of the study were to

- Determine if distances over time are larger on average for boys than for girls
- Determine if the ***rate of change*** of distance over time is different for boys and girls.

Several features are notable from the plot of the data:

- Each child has his/her own ***trajectory*** of distance as a function of age. For any given child, the trajectory looks roughly like a ***straight line***, with some fluctuations. But from child to child, features of the trajectory (e.g., its ***steepness***), vary. Thus, the trajectories are all of similar form, but vary in their specific characteristics among children. Note the one unusual boy whose pattern fluctuates more profoundly than those of the other children and the one girl who is much "lower" than the others.
- The ***overall trend*** is for the distance measurement to ***increase*** with age. The trajectories for some children exhibit strict increase with age, while others show some intermittent decreases, but still with an overall increasing trend across the entire 6 year period.
- The distance trajectories for boys seem for the most part to be "***higher***" than those for girls – most of the boy profiles involve larger distance measurements than those for girls. However, this is not uniformly true: some girls have larger distance measurements than boys at some of the ages.
- Although boys seems to have larger distance measurements, the ***rate of change*** of the measurements with increasing age seems similar. More precisely, the ***slope*** of the increasing (approximate straight-line) relationship with age seems roughly similar for boys and girls. However, for any ***individual*** boy or girl, the rate of change (slope) might be steeper or shallower than the apparent ***typical*** rate of change.

The foregoing observations are informal visual impressions. To address the questions of interest, it is clear that some formal way of representing these features is needed. Within such a representation, a formal way of stating the questions is required.

EXAMPLE 2: Vitamin E diet supplement and growth of guinea pigs. These data are reported by Crowder and Hand (1990, p. 27) and are from a study of the effect of a vitamin E diet supplement on the growth of guinea pigs. Fifteen guinea pigs were given a growth-inhibiting substance at the beginning of week 1 of the study (time 0, prior to the first measurement), and body weight was measured at the ends of weeks 1, 3, and 4. At the beginning of week 5, the pigs were randomized into 3 groups of 5, and vitamin E therapy was started. One group received zero dose of vitamin E, another received a low dose, and the third received a high dose. The body weight (g) of each guinea pig was measured at the end of weeks 5, 6, and 7.

In Figure 1.3, the data for the three dose groups are shown in spaghetti plots for each group; the plotting symbol is ID number (1–15) for each guinea pig.

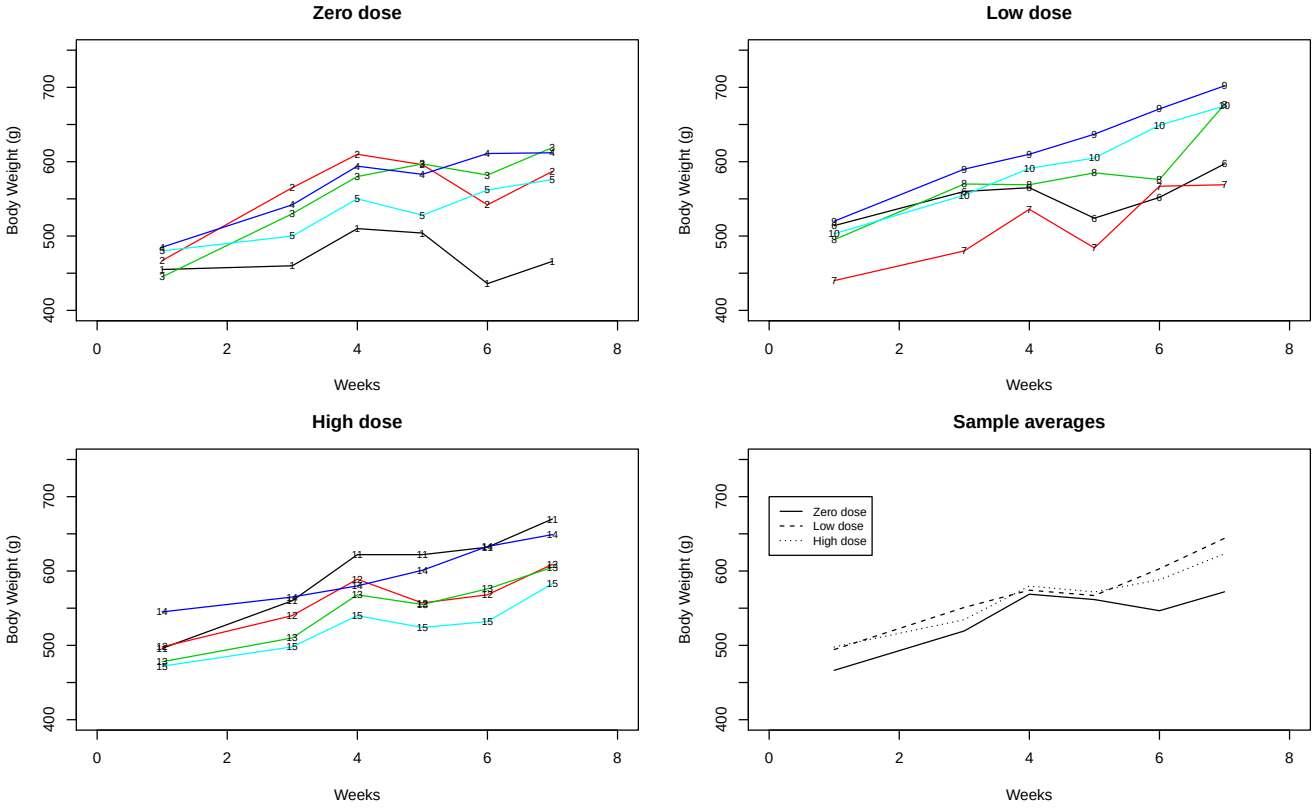

Figure 1.3: *Growth of guinea pigs receiving different doses of vitamin E diet supplement. Pigs 1–5 received zero dose, pigs 6–10 received low dose, pigs 11–15 received high dose.*

The primary objective of the study was to

- Determine if the ***growth patterns*** differ among the three groups.

As with the dental data, several features are evident:

- For the most part, the trajectories for individual guinea pigs seem to ***increase overall*** over the study period (although note pig 1 in the zero dose group). Different guinea pigs in the same dose group have different trajectories, some of which look like a straight line and others of which seem to have a "dip" at the beginning of week 5, the time at which vitamin E was added in the low and high dose groups.
- Trajectories for the zero dose group seem somewhat "***lower***" than those in the other groups.
- It is unclear whether or not the rate of change in body weight on average is similar or different across dose groups. In fact, it is not clear that the pattern for either individual pigs or "on average" is a ***straight line***, so the rate of change might not be constant. Because vitamin E therapy was not administered until the beginning of week 5, we might expect two "phases," before and after vitamin E, making things more complicated.

Again, some formal framework for representing this situation and addressing the primary research question is required.

EXAMPLE 3: Growth of two different soybean genotypes. This study was conducted by Colleen Hudak, a former student in the Department of Crop Science at North Carolina State University, and is reported in Davidian and Giltinan (1995, p. 7). The goal was to compare the growth patterns of two soybean genotypes, a commercial variety, Forrest (F) and an experimental strain, Plant Introduction #416937 (P).

Data were collected in each of three consecutive years, 1988–1990. In each year, 8 plots were planted with F, 8 with P. Over the course of the growing season, each plot was sampled at approximate weekly intervals. At each sampling time, 6 plants were randomly selected from each plot, leaves from these plants were mixed together and weighed, and an ***average leaf weight per plant*** (g) was calculated.

In Figure 1.4, the data from the 8 F plots and 8 P plots for 1989 are depicted.

The primary objective of the study was

- To compare ***growth characteristics*** of the two genotypes.

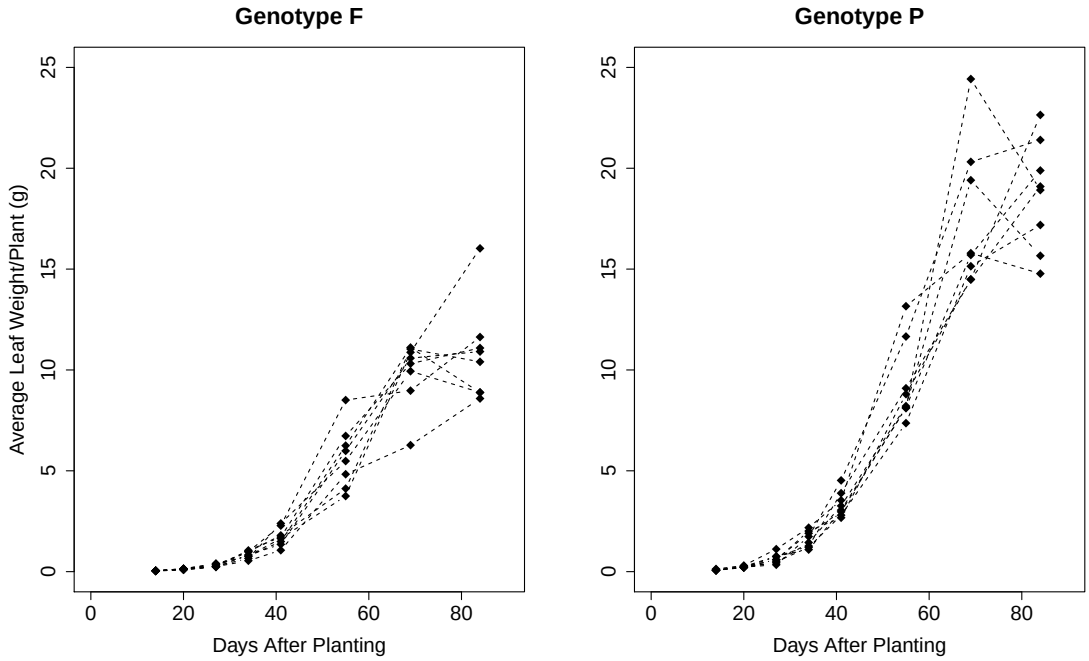

Figure 1.4: *Average leaf weight/plant profiles for 8 plots planted with Forrest and 8 plots planted with PI #416937 in 1989.*

From the figure, several features are notable:

- If we focus on the trajectory of a particular plot, we see that, typically, the growth begins ***slowly*** , with not much change over the first 3–4 observation times. Then, growth begins ***increasing*** at a faster rate in the middle of the season.
- Toward the end of the season, growth appears to begin ***leveling off***. This makes sense – soybean plants can only grow so large, so their leaf weight cannot increase without bound forever.
- Overall, then, the trajectory for any one plot does not appear to follow roughly a ***straight line*** as in the previous two examples, with an apparent constant rate of change over the observation period. Rather, the form of the trajectory seems more complicated, with almost an ***S shape***. It is thus clear that trying to characterize differences in growth characteristics will involve more than simply comparing a (constant) rate of change over the season.

In fact, the investigators expected that the growth pattern would not be as simple as an apparent straight line. They knew that growth would tend to level off toward the end of the season.

Thus, a more precise statement of the primary objective is

- To compare the ***limiting*** , or ***asymptotic*** , average leaf weight/plant between the 2 genotypes.
- To compare the way in which growth ***changes*** during the middle of the growing season.
- To compare the apparent ***initial*** average leaf weight/plant.

Several ***theoretical models*** have been postulated to describe growth processes exhibiting features like those observed for the soybean data. The most common such model is the ***logistic growth model*** , which says that ***growth rate relative to present size*** decreases ***linearly*** with increasing size. Formally, letting Y denote the growth value (average leaf weight/plant here) and t denote time, this may be expressed by the ***deterministic relationship***

$$\frac{dY}{dt} \Big/ Y = k\left(1 - \frac{Y}{a}\right),$$

where the right hand side is a linear function of present size Y, and $k > 0$ and $a > 0$. Upon integration, this model leads to

$$Y = \frac{a}{1 + c\exp(-kt)}, \tag{1.1}$$

where c is the value such that $a/(1 + c)$ represents growth value at time $t = 0$. As $t \to \infty$, $Y \to a$, so that a is a ***physically meaningful*** parameter characterizing ***asymptotic behavior*** , and, together with c, it characterizes the physically meaningful feature of "starting growth" at $t = 0$. The parameter $k > 0$ describes the ***change of growth with time***. A plot of the function (1.1) over a range of t for various choices of a, k, and c reveals that it has an ***S shape***.

This model appears to be a reasonable way to represent the growth profile for a ***given plot***. Figure 1.4 suggests that, although individual plot profiles have a similar shape, the parameters a and k describing asymptotic and change of growth might be ***different*** for different plots.

From Figure 1.4, it seems that average leaf weight/plant achieves "higher" limiting growth for genotype P relative to genotype F. That is, the ***asymptotic behavior*** seems to begin at lower values of the response for genotype F. The two genotypes seem to start off at roughly same value. It is difficult to make a simple statement about the relative rates of growth from the figure.

As it happened, ***weather patterns*** differed considerably over the three years: in 1988, conditions were unusually dry; in 1989, they were unusually wet; and in 1990 were relatively normal. Thus, comparison of growth patterns across the different weather patterns as well as how the weather patterns affect the comparison between genotypes, was also of interest.

Naturally, the investigators would like to be more formal about these observations and questions. This could be accomplished by incorporating a model like (1.1) within an appropriate statistical framework.

EXAMPLE 4: Pharmacokinetics of theopylline. A common objective is to investigate the ***pharmacokinetics*** (PK) of a drug; PK is, roughly speaking, the study of ***what the body does to the drug***. In a typical experimental PK study, a known dose of the drug is given to each of several (human or animal) subjects, and at several subsequent time points, blood samples are drawn from each and the ***concentration*** of drug in blood or plasma (the response) is determined for each sample.

The goal of such a study is

- To characterize the ***processes*** of drug ***absorption*** into the body, ***distribution*** throughout the body, and ***elimination*** from the body; the ***typical*** behavior of these processes; and how these processes ***vary*** in the population of subjects.

Armed with this information, ***pharmacokineticists*** develop ***dosing recommendations*** for the likely patient population taking the drug that will appear, e.g., in the ***labeling***.

Figure 1.5 shows concentration-time profiles for 4 of the 12 subjects in a PK study of the anti-asthmatic agent theophylline, each of whom received a single oral dose of theophylline at time 0 (given in units of mg/kg, so scaled to each individual's body weight in kg).

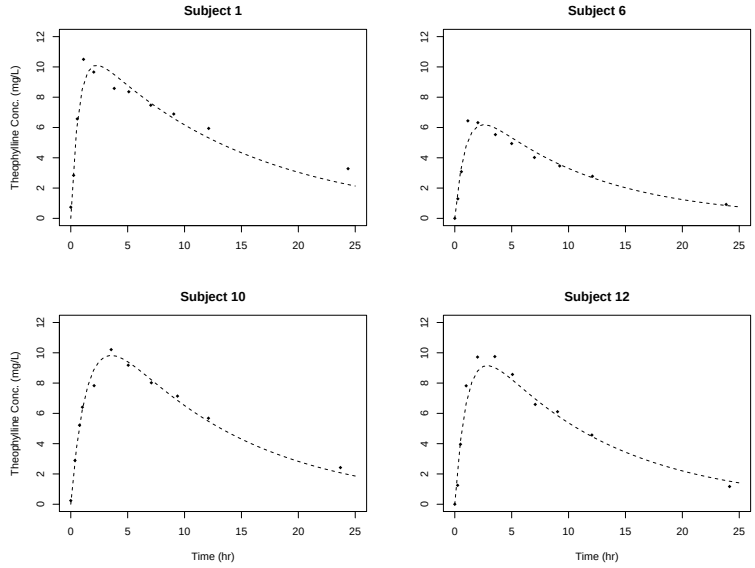

Figure 1.5: *Theophylline concentration-time profiles for 4 subjects receiving an oral dose of theophylline at time 0, with fits of model (*1.3) superimposed.

For each subject, 10 blood samples were drawn following the dose and assayed for theophylline concentration.

The profiles all exhibit the same general shape, with an early, steep rise in concentration that reaches a peak followed by what appears to be an ***exponential*** sort of decay. However, initial steepness, value and timing of the peak, and nature of the decay are ***different*** across subjects.

A statistician might be tempted to describe the concentration-time profile for an individual subject using a ***polynomial***. However, this is not a good approximation, nor is it a ***meaningful representation*** that addresses the questions of interest.

Instead, akin to the use of growth models like (1.1) in the soybean study, pharmacokineticists appeal to a deterministic ***theoretical representation*** based on representing the body as a system of ***compartments*** corresponding to components like "blood" and "deeper tissues." For orally-administered theophylline, a standard model is the ***one compartment open model with first order absorption and elimination***, represented pictorially as in Figure 1.6.

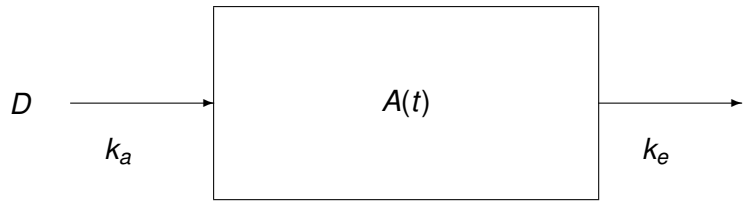

Figure 1.6: *One-compartment open model.*

In Figure 1.6, $A(t)$ is the amount of drug in the ***blood compartment*** at time t. Drug is ***absorbed*** through the gut into the blood at ***fractional absorption rate*** k_a and is ***eliminated*** (e.g., ***excreted*** by the kidneys and ***metabolized*** by the liver) at ***fractional elimination rate*** k_e.

The following ***system of differential equations***, with appropriate initial conditions, corresponds to Figure 1.6; see, for example, Gibaldi and Perrier (1982):

$$\begin{aligned} \frac{dA(t)}{dt} &= k_a A_a(t) - k_e A(t), \quad A(0) = 0 \\ \frac{dA_a(t)}{dt} &= -k_a A_a(t), \qquad\qquad A_a(0) = FD \end{aligned} \tag{1.2}$$

F = bioavailability, $A_a(t)$ = amount at absorption site

The system (1.2) can be solved for $A(t)$, and, dividing by V, the ***volume*** of the blood compartment, yields the following expression for ***concentration*** of drug at time t:

$$C(t) = \frac{A(t)}{V} = \frac{k_a DF}{V(k_a - k_e)}\{\exp(-k_e t) - \exp(-k_a t)\},\ k_e = Cl/V, \tag{1.3}$$

The parameter F, ***bioavailability*** , is usually taken to be equal to 1. The parameter Cl, the ***clearance*** rate, is a measure of the volume of blood cleared of drug per unit time, and is of primary importance in understanding how the drug is eliminated from the system. The ***volume of distribution*** parameter V reflects the extent to which the drug distributes throughout the system.

The compartment model and the ensuing model for concentration $C(t)$ in (1.3) pertain to ***individual*** subject behavior; that is, the model is a theoretical description of biological processes taking place over time ***within*** a given subject, as that subject processes the drug. From Figure 1.5, which shows certain ***fits*** of the model (1.3) to each subject's data superimposed, it seems clear that differences in steepness, peak, and decay we noted earlier reflect the fact that each subject has his/her own PK parameters k_a, V, and Cl governing his/her individual PK behavior.

Returning to the objective of the study, characterizing ***typical*** absorption, distribution, and elimination behavior and how these processes ***vary*** in the population clearly will involve deducing the average values of the PK parameters and how they vary across the population from the concentration-time data. More formally, the goal is to describe the ***distribution*** of these parameters across the population.

An appropriate statistical framework is required in which the theoretical compartment model can be incorporated and within which this distribution can be characterized.

SO FAR: In the four examples we have considered, the outcome of interest is ***continuous*** in nature. That is,

- Distance (mm) from the center of the pituitary to the pterygomaxillary fissure
- Body weight (g)
- Average leaf weight/plant (g)
- Drug concentration (mg/L)

all can in principle take on any possible value in a particular range. How precisely we observe the value of the response is limited only by the precision of the measuring device we use.

In some situations, the outcome of interest is ***not*** continuous; rather, it is ***discrete*** in nature. We consider two additional examples.

EXAMPLE 5: Epileptic seizures and chemotherapy. A common situation is where the measurements are in the form of ***counts***. A response in the form of a ***count*** is by nature ***discrete*** – counts (usually) take only nonnegative integer values $(0, 1, 2, 3, ...)$.

The following famous data are reported by Thall and Vail (1990). A clinical trial was conducted in which 59 subjects with epilepsy suffering from simple or partial seizures were assigned at random to receive either the anti-epileptic drug progabide (subjects 29–59) or a ***placebo*** (an inert substance, subjects 1–28) in addition to a standard chemotherapy regimen all were taking. Because each individual might be prone to different rates of experiencing seizures, the investigators first tried to get a sense of this by recording the number of seizures suffered by each subject over the 8-week period prior to the start of administration of the assigned treatment. It is common in such studies to record such ***baseline*** measurements, so that the effect of treatment for each subject can be measured relative to how that subject behaved ***prior to*** treatment.

Following initiation of treatment, the number of seizures for each subject was counted for each of 4 consecutive 2-week periods. The age of each subject at the start of the study was also recorded, as it was suspected that subject age might be associated with the effect of the treatment.

The data for the first 5 subjects in each treatment group are summarized in Table 1.1.

	Period						
Subject	*1*	*2*	*3*	*4*	*Trt*	*Baseline*	*Age*
1	5	3	3	3	0	11	31
2	3	5	3	3	0	11	30
3	2	4	0	5	0	6	25
4	4	4	1	4	0	8	36
5	7	18	9	21	0	66	22
				⋮			
29	11	14	9	8	1	76	18
30	8	7	9	4	1	38	32
31	0	4	3	0	1	19	20
32	3	6	1	3	1	10	30
33	2	6	7	4	1	19	18

Table 1.1: *Seizure counts for 5 subjects assigned to placebo (coded as 0) and 5 subjects assigned to progabide (coded as 1).*

The primary objective of the study was to

- Determine if progabide ***reduces*** the rate of seizures relative to placebo in subjects like those in the trial.

We have repeated measurements (counts) on each subject over four consecutive observation periods, and we would like to compare somehow the baseline seizure counts to post-treatment counts, where the latter are observed ***repeatedly*** over time following initiation of treatment. Clearly, an appropriate analysis would make the best use of this feature of the data in addressing the main objective.

Note that some of the counts are quite small; in fact, for some subjects, 0 seizures (none) were experienced in some periods. For example, subject 31 in the treatment group experienced only 0, 3, or 4 seizures over the 4 observation periods. Clearly, models and methods that are appropriate for ***continuous*** outcomes like those in the first four examples would be suspect in this situation.

A classical approach to handling data in the form of counts is to ***transform*** them to some other scale. The motivation is to make them seem more ***normally distributed*** with constant variance, and the ***square root*** transformation is used to (hopefully) accomplish this. The desired result is that methods that are usually used to analyze continuous measurements can then be applied.

The drawback of this approach is that one is no longer working with the data on the ***original scale*** of measurement, numbers of seizures in this case. The statistical models assumed by this approach describe "square root number of seizures," which is not particularly intuitive. Statistical methods that are designed to address questions on the basis of ***discrete*** repeated measurements like counts are required.

EXAMPLE 6: Maternal smoking and child respiratory health. Another common ***discrete data*** situation is where the outcome is ***binary***; that is, the outcome can take on only ***two*** possible values, which usually correspond to

- ***success*** or ***failure*** of a treatment to elicit a desired response
- ***presence*** or ***absence*** of some condition

The following data come from a very large public health study called the ***Six Cities Study***, which was undertaken in six small American cities to investigate a variety of public health issues. The full situation is reported in Lipsitz, Laird, and Harrington (1992). The portion we consider was focused on the association between maternal smoking and child respiratory health. Each of 300 children was examined once a year at ages 9–12. The outcome of interest was "***wheezing status***," a measure of the child's respiratory health, which was coded as either "no" (0) or "yes" (1), where "yes" corresponds to respiratory problems. Also recorded at each examination was a code indicating the mother's current level of smoking: 0 = none, 1 = moderate, 2 = heavy.

The data for the first 5 subjects are summarized in Table 1.2.

		Smoking at age				*Wheezing at age*			
Subject	*City*	9	10	11	12	9	10	11	12
1	Portage	2	2	1	1	1	0	0	0
2	Kingston	0	0	0	0	0	0	0	0
3	Portage	1	0	0	.	0	0	0	.
4	Portage	.	1	1	1	.	1	0	0
5	Kingston	1	.	1	2	0	.	0	1

Table 1.2: *Data for 5 children in the Six Cities study. Missing data are denoted by a "."*

The objective of an analysis of these data was to

- Determine how the ***typical wheezing response pattern*** changes with age
- Determine if there is an ***association*** between maternal smoking severity and child respiratory status (as measured by wheezing).

This study exemplifies the very common situation in medical and public health (and other) research of ***repeated measurements*** on a ***binary*** outcome. As with the count data, one might first think about trying to summarize and transform the data to allow (somehow) methods for continuous data to be used; however, this would clearly be inappropriate. As with the seizure study, statistical methods that are designed to address questions on the basis of ***discrete*** repeated measurements are required to address the questions of interest.

With binary outcome, spaghetti-type plots of repeated measures on the subjects as a function of age do not lend much insight. Informal inspection of individual subject data does suggest a possible association between wheezing and maternal smoking; e.g., subject 5 did not exhibit positive wheezing status until his/her mother's smoking increased in severity.

This highlights the fact that this situation is complex: over time (measured here by age of the child), an important characteristic, maternal smoking, ***changes***. Contrast this with the previous examples, where a main focus is to compare groups whose membership stays ***constant*** over time.

Another feature of these data is the fact that some data are ***missing*** for some subjects. Specifically, although the intention was to collect data for each of the four ages, this information is not available for some children and their mothers at some ages; for example, subject 3 has both the mother's smoking status and wheezing indicator missing at age 12. This pattern would suggest that the mother may have failed to appear with the child for this intended examination.

In the other examples, units (children, guinea pigs, plots, patients) were ***assigned*** to treatments; thus, these can be regarded as ***controlled experiments***, where the investigator has some control over how the factors of interest are "applied" or administered to the units (through ***randomization***). In contrast, in this study, the investigators did not decide which children would have mothers who smoke; instead, they could only ***observe*** smoking behavior of the mothers and wheezing status of their children. That is, this is an ***observational study***. Because it might be impossible or unethical to randomize subjects to potentially hazardous circumstances, studies of issues in public health and the social sciences are often ***observational***.

As in many observational studies, an additional difficulty is the fact that the ***exposure*** of interest, in this case maternal smoking, ***also changes*** with the response over time. This leads to complicated issues of interpretation in statistical modeling that are a matter of some debate. We discuss these issues in our subsequent development.

SUMMARY: These five examples illustrate the broad range of applications where data in the form of repeated measurements arise and the diverse range of questions of interest that are posed. The response of interest can be ***continuous*** or ***discrete***. The scientific questions might focus on very specific features of response trajectories, e.g., asymptotic behavior or PK processes; or might involve vague questions about the form of the ***typical response trajectory***. To complicate matters further, in studies where data were planned to be collected at certain points in time, it is possible for some responses to be ***missing***.

1.3 Statistical models for longitudinal data

In this course, we discuss a number of approaches for modeling data like those in the examples and describe different statistical methods for addressing questions of scientific interest within the context of these models.

STATISTICAL MODELS: Recall that a statistical model is a representation of the way in which data are thought to arise. Formally, a ***statistical model*** is a class of ***probability distributions*** that is assumed to have generated the data, where the data are represented by appropriately defined ***random variables***. Thus, a statistical model is a class of joint probability distributions for these random variables, which the analyst believes is a plausible representation of the true data generating mechanism.

The nature and features of an assumed statistical model dictate how questions of interest can be stated formally and unambiguously and how the data should be analyzed to address the questions. Different models embody different assumptions about how the data arise. Thus, the extent to which valid inferences on the questions of interest can be drawn under an assumed statistical model rests on how relevant its assumptions are to the situation at hand.

Thus, to appreciate the basis for techniques for data analysis and to use them appropriately, one must refer to and understand the associated statistical models. This connection is especially critical in the context of longitudinal data.

BASIC REPRESENTATION OF LONGITUDINAL RESPONSES: As in all of our examples, we consider a ***scalar response*** that is recorded on the same individual over time or some other set of conditions. In the specification of statistical models, it is convenient to think of all responses on the same individual ***together***, so that complex relationships among them can be summarized. Accordingly, we represent the responses at each time as ***random variables*** and collect these for each individual into ***random vectors***, as follows.

In general, define the random variable

$$Y_{ij} = \text{the } j\text{th response recorded on individual } i,$$

where $i = 1, \ldots, m$ indexes individuals, and $j = 1, \ldots, n_i$ indexes repeated measurements on the ith individual. Here, n_i, the number of repeated measurements on individual i, can be different for different individuals.

Let

$$t_{ij} = \text{the time at which } Y_{ij} \text{ is recorded for the } i\text{th individual.}$$

This notation allows further for the possibility of different times for different individuals.

For the ith individual, collect Y_{ij}, $j = 1, \ldots, n_i$, into the random vector

$$\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T.$$

For example, consider the dental study data in Figure 1.2. If we index all children regardless of gender by i, then $i = 1, \ldots, m = 27$. Each child was measured at ages 8, 10, 12, and 14 years; thus, $n_i = 4$ for all children, and $t_{i1} = 8$, $t_{i2} = 10$, and so on for all children in the study. The dental distances for the ith child can be summarized by the (4×1) random vector

$$\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{i4})^T.$$

We use this notation throughout to represent observed responses on each individual. In subsequent chapters, we will develop further notation to represent additional ***covariate*** information, such as gender in this example, dose in the theopylline pharmacokinetics study in ***EXAMPLE 4***, or maternal smoking status in the Six Cities study in ***EXAMPLE 6***, that may be available on each individual.

UNEQUAL n_i AND MISSING DATA: We emphasize that the foregoing notation is meant to refer to outcomes that are ***actually observed and recorded***. That is, $\boldsymbol{Y}_i$ is the $(n_i \times 1)$ vector of responses recorded at times $t_{i1}, \ldots, t_{in_i}$ on individual i and ***available to the data analyst***.

In many settings, such as the dental study or the Six Cities study, the intention of the investigators is to collect outcomes from each individual at the ***same*** times (or other conditions). E.g., in the dental study, children were to be evaluated at ages 8, 10, 12, and 14; and mother-child pairs were to be assessed in the Six Cities study when the child was ages 9, 10, 11, and 12.

- In the dental study, all children were seen at the intended times. Accordingly, for child i, the random vector $\boldsymbol{Y}_i$ of outcomes actually ***observed*** is the same as the random vector of ***intended*** outcomes.
- In contrast, in the Six Cities study, the intended wheezing outcome for some mother-child pairs is ***missing***; for example, subject $i = 5$ is missing the outcome at age 10. In this case, $n_i = 3$, $t_{i1} = 9$, $t_{i2} = 11$, and $t_{i3} = 12$, and $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3})^T$, with realized value $(0, 0, 1)^T$.

- In the sequel, when we discuss the implications of ***missing data***, we take care to distinguish the vector of ***intended*** outcomes from the vector of ***observed*** responses.

CLASSICAL NOTATION: Although summarizing repeated outcomes as random vectors for each individual is fundamental to modern longitudinal data models and methods, it was ***not*** generally used in the original formulation of ***classical models and methods*** for the analysis of (***continuous***) repeated measurements, namely, the ***univariate repeated measures analysis of variance*** methods we will discuss in Chapter 3. As we demonstrate, specification of these methods is through "***analysis of variance***" notation, where each scalar response is identified by subscripts corresponding to individual, group factors, and time; and time is viewed as a ***categorical factor*** rather than as potentially ***continuous***. The scalar responses for each individual are ***not*** collected into individual-specific random vectors in the standard presentation of the classical models.

We present classical models and methods using the ***original notation*** and then reexpress them in the notation given above so that they can be contrasted to the more modern methods.

CONTINUOUS RESPONSE: When the response is ***continuous***, not surprisingly, both classical and more modern methods are based on assuming each scalar response Y_{ij} (given covariate information) is ***normally distributed*** so that each vector $\boldsymbol{Y}_i$ (given covariate information) is taken to follow a ***multivariate normal distribution***. Moreover, models for the $\boldsymbol{Y}_i$ as a function of covariate information are often taken to be ***linear*** in parameters that characterize features of interest.

The methods we discuss in Chapters 3–6 are based on these assumptions. However, as we will demonstrate, even when normality does not hold, as long as m is "***large***," the methods have good properties that can be approximated via ***asymptotic theory***.

DISCRETE RESPONSE: When the outcome is ***discrete***, for example, binary or in the form of a count, things become more complicated. For a scalar discrete random variable Y_{ij}, probability distributions such as the Poisson or Bernoulli are standard models. However, in contrast to the normal, these distributions ***do not*** have immediate ***multivariate generalizations***. In addition, usual regression models for such scalar responses as a function of covariates, such as ***loglinear*** or ***logistic*** models, are ***not*** linear in parameters.

As a result, taking an approach analogous to that for continuous repeated outcomes is not directly possible. This challenge led to an enormous body of work in the 1980s and 1990s on approaches to modeling repeated discrete responses and associated inferential methods. In Chapters 7–9, we discuss these modeling strategies and associated methods. These are also applicable in the case of continuous responses for which linear models are ***not appropriate***, as in the soybean growth study in ***EXAMPLE 3*** and the pharmacokinetic study of ***EXAMPLE 4***.

1.4 Outline of the course

Given the considerations of the previous section, the course offers coverage of two main areas. First, methods for the analysis of continuous repeated measurements that are reasonably thought of as normally distributed and for which linear models may be appropriate are discussed. Later, methods for the analysis of repeated measurements that are not reasonably thought of as normally distributed, such as discrete outcomes, or for which linear models are not appropriate, are covered.

The course can be thought of as having five parts:

I. Preliminaries:

- Introduction and motivation (Chapter 1)
- Modeling longitudinal data (Chapter 2)

II. Classical methods:

- Repeated measures analysis of variance (Chapter 3)

III. Methods for continuous, normally distributed responses:

- Modern methods: preliminaries (Chapter 4)
- Population-averaged linear models for continuous responses (Chapter 5)
- Linear mixed effects models (Chapter 6)

IV. Methods for discrete responses and problems involving nonlinear models:

- Review of generalized linear and nonlinear models (Chapter 7)
- Population-averaged models and generalized estimating equations (Chapter 8)
- (Subject-specific) generalized linear and nonlinear mixed effects models (Chapter 9)

V. Advanced topics

It is important to stress that there are ***numerous approaches*** to the modeling and analysis of longitudinal data, and there is no strictly "***right***" or "***wrong***" way. It is true, however, that some approaches are more flexible than others, imposing less restrictions on the nature of the data and allowing questions of scientific interest to be addressed more directly. We discuss how various approaches compare as we proceed.

Missing data, often the result of ***dropout*** from studies in which individuals are to be evaluated at several time points, are a recurring challenge in longitudinal data analysis. Although this is not a course on analysis in the presence of missing data, because missing data are a fact of life in longitudinal studies, we discuss their implications.

2 Modeling Longitudinal Data

2.1 Introduction

Before we discuss specific modeling approaches and the associated inferential methods, we introduce ***notation*** that we will use throughout the course. We also describe a ***conceptual framework*** for thinking about longitudinal data that highlights the considerations underlying the different models and methods we discuss in subsequent chapters.

As we demonstrate, acknowledging and representing ***correlation*** among responses on the same individual over time is central to modeling and analysis of longitudinal data. The conceptual framework clarifies that correlation comes about because of phenomena acting both ***within*** and ***among*** individuals, which are represented in different ways within different modeling strategies. We review several popular models for correlation structure.

We introduce conceptually two main modeling strategies, ***population-averaged*** modeling and ***subject-specific*** modeling, instances of which we discuss in considerable detail in subsequent chapters. As we emphasize, the nature of the scientific questions of interest dictates which modeling approach and perspective is relevant in a given application.

The conceptual framework we present is relevant for ***continuous*** outcome. Indeed, the classical methods we consider in Chapter 3 are relevant to responses that are or that can be viewed approximately as continuous. We discuss some of the challenges involved in modeling ***discrete*** longitudinal outcome at the end of this chapter and return to these in subsequent chapters.

2.2 Data structure and notation

DATA STRUCTURE: As discussed in Section 1.3, we ***observe*** m response vectors $\boldsymbol{Y}_i$, $i = 1, \ldots, m$, where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$, so that the $\boldsymbol{Y}_i$ need not be of the same dimension. Each response vector corresponds to a different ***individual***, where individuals are drawn from some population(s) of interest. The Y_{ij} are recorded at times t_{ij}, $j = 1, \ldots, n_i$.

Thus, the data are comprised of a total of $N = \sum_{i=1}^{m} n_i$ scalar observations.

The numbers of observations n_i may be different by design or because, although the intention was to collect the same number of observations on each *i* at the ***same*** times, some intended responses are ***missing*** for some individuals.

Formally, we assume that the ***random vectors*** $\boldsymbol{Y}_i$ are ***statistically independent***. In applications like those in Section 1.2, this makes intuitive sense in that the outcome (e.g., dental distance, drug concentration, wheezing status) is reasonably thought of as evolving over time ***within*** an individual in a way that is ***unrelated*** to the way in which responses evolve for other individuals.

As we discuss in detail shortly, it is generally ***not reasonable*** to assume that responses from the ***same*** individual; i.e., the Y_{ij}, are independent across *j*. In particular, as noted above, responses on the same individual are expected to be ***correlated***.

Thus, in the longitudinal data situation we consider, it is natural to think of data from different individuals as ***independent*** , and we adopt this assumption without further comment henceforth.

It is important to recognize that things can be even more complicated. In the longitudinal data situation in the examples we have considered, the vectors of independent observations are ***obvious*** : each is from a different unit. In some settings, it may ***not*** be possible to identify random vectors that can be viewed as independent:

- For instance, a famous data set reported by McCullagh and Nelder (1989, Section 14.5) involves a number of male and female salamanders of two different varieties. The males and females were placed together to mate one another across and within varieties according to a rather complex design. The outcome was number of observed matings that took place during the pairing. It is obvious that responses involving the same female or male could potentially be correlated, as particular males or females may be more or less interested in mating! Clearly, in this ***crossed*** experiment, identifying independent data vectors is complicated, if not impossible.
- When observations are taken over a physical area in two or three dimensional space, it is reasonable to be concerned about correlation due to physical proximity – observations close together may be "***more alike***" than those far apart. Hopefully, the correlation "***dies off***" as they become farther apart in space, but there is no natural way to decide if some observations could be treated as independent from others, as in this setting, there are no "***individuals***" per se. Thus, there is no obvious way to represent the data as independent random vectors. Frameworks for this situation are the subject of study of ***spatial statistics***.

In this course, we limit our consideration of multivariate response to the situation where identification of independent outcome vectors can be made ***unambiguously***. This is of course the case for ***longitudinal repeated measurement*** data, where repeated responses, over time or some other factor, are recorded on independent units.

COVARIATE INFORMATION: As in the examples in Section 1.2, in addition to longitudinal outcomes, ***covariate information*** may be collected on each individual. It is important to distinguish between ***two types*** of covariates; to make the distinction clear, we adopt specific notation.

- ***Within-individual covariates***. These are covariates that describe conditions under which the Y_{ij} were collected on individual i. Such covariates would be important to know even if the ***focus of inference*** is restricted to individual i ***only***.

  To appreciate this, consider ***EXAMPLE 4*** in Section 1.2, the pharmacokinetic study of theophylline. Here, the dose administered to each subject at time 0 is ***different*** (scaled to the subject's body weight). Letting D_i denote the dose given to subject i, if we were interested in estimating the ***PK parameters*** k_a, Cl, and V pertaining to subject i in the ***one compartment model*** (1.3), it should be clear that we could do so based on the concentration-time data $\boldsymbol{Y}_i$ for subject i using suitable ***nonlinear regression*** methods. This would require knowledge of the dose D_i given to subject i.

  Generically, we use the notation $\boldsymbol{u}_i$ to denote the collection of such ***within-individual*** covariates on the ith individual. Thus, in the PK example, $\boldsymbol{u}_i = D_i$.

  For longitudinal data, we also have ***time*** or other condition of measurement that ***changes value*** for i over $j = 1, \ldots, n_i$; e.g., in the PK example, the times t_{ij} at which blood samples were drawn for subject i. As we will discuss in the next section when we view the responses on a given individual as coming about due to an underlying ***stochastic process***, it is not entirely appropriate to view time or other condition of measurement as a "covariate" in this same sense. It is indeed true that the t_{ij} ***operationally*** play the role of covariates from the perspective of fitting the model (1.3) to the concentration-time data for individual i. However, as will be clear shortly, time or other condition is tied up with ***serial correlation*** and should be regarded separately.

  For convenience when discussing model fitting, we use a single notation to refer to both "***true***" within-individual covariates and "***time***" and write $\boldsymbol{z}_{ij}$ to denote all such "conditions" associated with collecting Y_{ij} on i; e.g., $\boldsymbol{z}_{ij} = (D_i, t_{ij})$ in the pharmacokinetic example. Later, we are careful to distinguish time from other conditions $\boldsymbol{u}_i$ when it is necessary to do so.

- ***Among-individual*** or ***individual-level covariates*** These are covariates that ordinarily ***do not change value*** over $j = 1, \ldots, n_i$ and that can be viewed as characteristics of i or how i was treated.

  In ***EXAMPLES 1–3*** and ***EXAMPLE 5***, gender, vitamin E dose, soybean genotype, and treatment (progabide or placebo), respectively, are examples of such covariates. These covariates would be of no interest if we focused only on the data on individual i.

  To appreciate this, consider the soybean growth data on the ith plot. Suppose we are interested in estimating the parameters a, c, and k in the ***logistic growth model*** (1.1) for plot i by fitting this model to the average leaf weight/plant responses on plot i using suitable ***nonlinear regression*** methods. Clearly, the soybean genotype planted on plot i is not relevant to this objective; indeed, genotype does not even enter into the model.

  Among-individual covariates instead characterize questions of scientific interest at the level of the ***population(s)*** from which individuals $i = 1, \ldots, m$ are drawn. For example, in the dental study of ***EXAMPLE 1***, questions of interest focus on differences between genders. We denote the collection of such covariates as $\boldsymbol{a}_i$.

  The covariate ***maternal smoking*** in ***EXAMPLE 6*** is a bit troublesome. Its value ***does change*** over $j = 1, \ldots, n_i$, but it is ***different from*** a within-individual covariate like dose in the PK study of ***EXAMPLE 4*** in that it is reflects the way child i was ***treated*** and is involved in the ***population-level*** question of whether or not there is an ***association*** between maternal smoking severity and wheezing status (the response).

  We view covariates like maternal smoking in this example that ***do change*** over j as ***among-individual*** covariates that are included in $\boldsymbol{a}_i$. We defer further discussion of this type of covariate to later chapters and restrict attention to among-individual covariates that ***do not*** change value over j for now.

SUMMARY: The available data for individual i consist of pairs $(Y_{i1}, \boldsymbol{z}_{i1}), \ldots, (Y_{in_i}, \boldsymbol{z}_{in_i})$ along with the associated individual-level covariates $\boldsymbol{a}_i$. Writing $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}^T, \ldots, \boldsymbol{z}_{in_i}^T)^T$ to denote the collection of within-individual covariates over j, including "time," we can think of the data as the triplets $(\boldsymbol{Y}_i, \boldsymbol{z}_i, \boldsymbol{a}_i)$, $i = 1, \ldots, m$.

As above, we assume that $(\boldsymbol{Y}_i, \boldsymbol{z}_i, \boldsymbol{a}_i)$ are ***independent*** across $i = 1, \ldots, m$. However, independence among the components of $\boldsymbol{Y}_i$ is ***not*** assumed.

As shorthand, we define $\boldsymbol{x}_i = (\boldsymbol{z}_i^T, \boldsymbol{a}_i^T)^T$ to denote the full set of all covariates associated with $\boldsymbol{Y}_i$. Thus, we represent the data more succinctly as ***independent*** pairs $(\boldsymbol{Y}_i, \boldsymbol{x}_i)$, $i = 1, \ldots, m$.

MODELING MULTIVARIATE RESPONSE: In the case of ***univariate response***, questions of scientific interest are often cast within the framework of a classical ***regression model***.

For example, as we noted previously, in the pharmacokinetic study in ***EXAMPLE 4*** in Section 1.2, suppose we are interested in estimating the ***PK parameters*** k_a, Cl, and V in the ***one compartment model*** (1.3) for subject i, who received dose D_i at time 0. We would base this on the data for subject i, $(Y_{i1}, \boldsymbol{z}_{i1}), \ldots, (Y_{in_i}, \boldsymbol{z}_{in_i})$, where $\boldsymbol{z}_{ij} = (D_i, t_{ij})$, and the Y_{ij} are univariate drug concentrations measured at times t_{ij}, $j = 1, \ldots, n_i$.

The obvious approach is to consider the ***regression model***

$$E(Y_{ij}|\boldsymbol{z}_{ij}) = f(\boldsymbol{z}_{ij}, \beta_i) = \frac{k_{ai} D_i}{V_i(k_{ai} - Cl_i/V_i)}\{\exp(-Cl_i\, t_{ij}/V_i) - \exp(-k_{ai} t_{ij})\}, \quad \beta_i = (k_{ai}, Cl_i, V_i)^T, \qquad (2.1)$$

where we have added a subscript i to the parameters to emphasize that they are unique to subject i and collected them in the parameter vector β_i. Along with the ***conditional mean*** model (2.1), we would make some assumption on the ***conditional variance*** $\text{var}(Y_{ij}|\boldsymbol{z}_{ij})$; for example $\text{var}(Y_{ij}|\boldsymbol{z}_{ij}) = \sigma^2$, constant variance over time. A more relevant assumption in this application is that variance is ***proportional*** to the square of the mean; that is, exhibits ***constant coefficient of variation***, which can be expressed as

$$\text{var}(Y_{ij}|\boldsymbol{z}_{ij}) = \sigma^2 f^2(\boldsymbol{z}_{ij}, \beta_i). \qquad (2.2)$$

One would then use standard ***nonlinear regression*** techniques that accommodate a ***variance model*** like (2.2), such as ***iteratively reweighted least squares***, to estimate β_i based on the data on subject i; we review these methods in Chapter 7.

Standard such regression methods assume that $(Y_{ij}, \boldsymbol{z}_{ij})$ over $j = 1, \ldots, n_i$ are ***independent***; as we discuss shortly, this may ***not*** be the case here, given the ***time-ordered*** nature of data collection.

Putting this issue aside for the moment, the upshot is that, when the focus of inference involves phenomena leading to a univariate response (drug concentration here), there is a ***natural framework***, that of classical regression modeling, in which to cast the problem.

REMARK: In this example in (2.1) and (2.2), we condition on $\boldsymbol{z}_{ij}$, as would be conventional in standard regression analysis, treating t_{ij} as a covariate. Consistent with our previous discussion, we really mean conditioning on $\boldsymbol{u}_i = D_i$ In addition, because we are interested in inference at the level of individual i, we have implicitly regarded β_i as a ***fixed parameter***. In later chapters where we consider inference on the ***population*** of individuals, it will be natural to treat β_i as ***random*** and to condition on β_i as well. These points will become more clear in the next section.

In the examples in Section 1.2, the questions of interest are more complex. E.g., in ***EXAMPLE 4*** the focus is ***not*** on the PK parameters for individual subjects, but on the PK properties in the ***population*** of subjects. In the seizure study in ***EXAMPLE 5***, the focus is again on the ***population*** of patients suffering from epilepsy and comparison of how this population would fare if given progabide versus not in terms of the rate of seizures experienced in the population.

In each case, the available data are now the ***independent*** $(\boldsymbol{Y}_i, \boldsymbol{x}_i)$, $i = 1, \ldots, m$, where $\boldsymbol{Y}_i$ for individual i is a vector of repeated outcomes on i. There is no longer an obvious, single framework in which to address these questions; in fact, the very nature of the questions seems different in these two examples.

Indeed, in the more complex setting of ***multivariate response*** , there is more than one approach to statistical modeling, and the appropriate approach is dictated by the particular setting and questions of interest. In Section 2.4, we will discuss the two most popular and widely-used modeling strategies.

As we noted at the beginning of this chapter, a key feature of longitudinal data that must be acknowledged in any modeling strategy is ***correlation*** among the elements Y_{ij} of $\boldsymbol{Y}_i$. To appreciate how correlation is taken into account, one must understand how correlation in these data is thought to arise. We now consider this in detail.

2.3 Conceptual framework for continuous response

Recall the dental study data in ***EXAMPLE 1*** of Section 1.2 , which are shown again in Figure 2.1. Here, we plot the data separately by gender, and superimpose the sample means at each age for each gender, which are connected by a bold line in each panel.

We now consider a ***conceptual representation*** of the ***underlying mechanism*** giving rise to data like those in Figure 2.1.

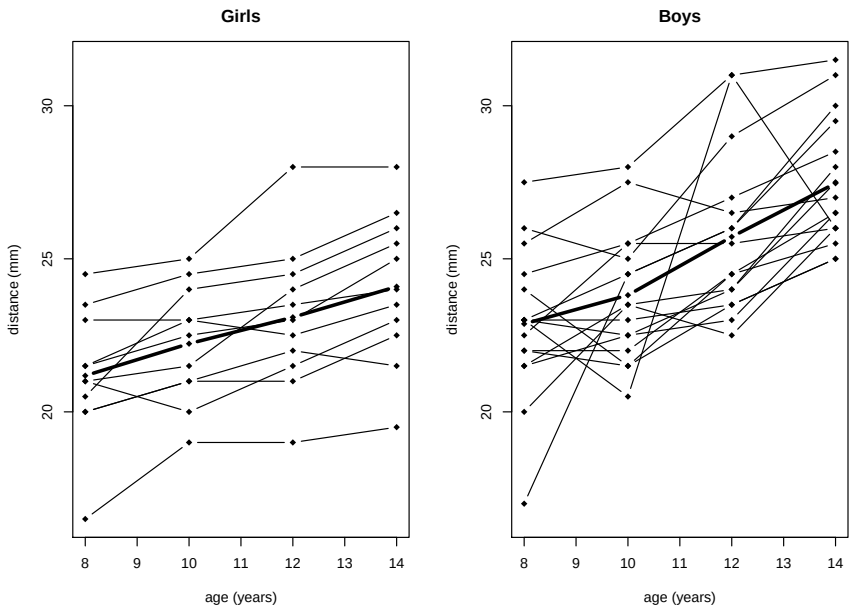

Figure 2.1: *Orthodontic distance measurements (mm) for 27 children at ages 8, 10, 12, 14. The left hand panel shows the data for the 11 girls; the bold line shows the sample mean distances for the girls at each age. The right hand panel shows the same for the boys.*

The conceptual representation demonstrates that ***correlation*** among the elements of a response vector $\boldsymbol{Y}_i$ can arise at ***two levels***:

(i) Due to ***within-individual*** sources

(ii) Due to ***among-individual*** , ***population-level*** sources.

To discuss how each of these phenomena contribute to the ***overall pattern*** of correlation among elements of $\boldsymbol{Y}_i$, we consider a generic conceptual depiction of how responses collected over time on each of several individuals can be thought to arise, which is shown in Figure 2.2.

Figure 2.2(a) depicts three hypothetical observed response vectors from different individuals; e.g., three children in the dental study. The plotting symbols (diamonds) represent the ***actual responses*** observed at each of several time points for each, and thus the figure corresponds to what we might see in practice if we were to create a spaghetti plot of such data.

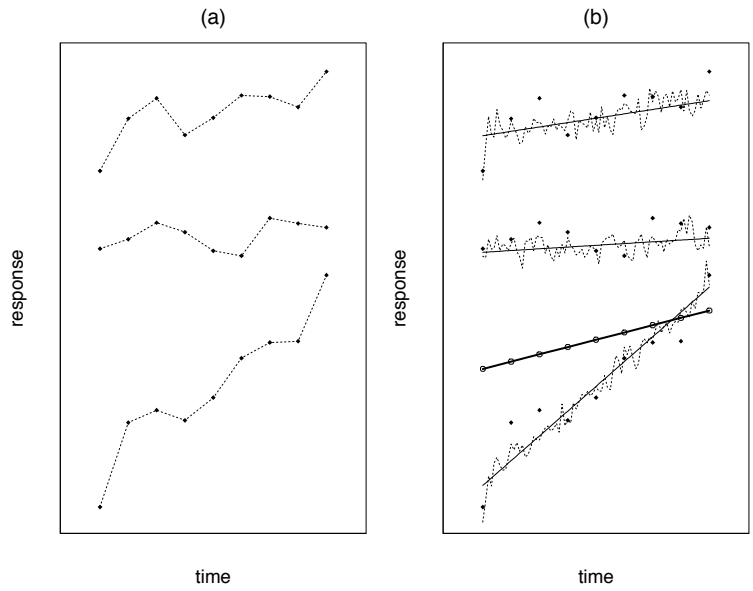

Figure 2.2: *Conceptual model for sources of variation/correlation in data collected over time. (a) Spaghetti plot of data actually observed on three individuals. (b) Conceptual representation of stochastic process giving rise to the data. The thick solid line represents the population mean response over continuous time, averaged over all individuals and possible realizations, with the open circles representing the population means at the observation times. The dotted lines represent the underlying error-free realization of the response over time for the three individuals. The diamonds represent the observed responses, which are subject to measurement error.*

Figure 2.2(b) is a conceptual representation of the ***underlying mechanism*** that might give rise to the actual responses (which, as in Figure 2.2(a), are all we get to see). For example, in the context of the dental study, Figure 2.2(b) can be interpreted as representing what underlies the observed data on three children in either panel of Figure 2.1.

For definiteness, focus on the topmost response vector and consider the mechanism for a single individual.

- Suppose that the response is blood pressure. As is well established, an individual's blood pressure varies throughout the day. If we could ***continually*** monitor blood pressure for an individual using a ***perfect*** measuring device (so with no error in measurement at all; see below), we might see something like the dotted, continuous line in Figure 2.2(b).

This dotted line can be thought of as representing how real phenomena can exhibit "***local***" patterns of change that follow a smooth trend over the long term. E.g., a person's blood pressure may follow a smooth trajectory over a period of week or months (in the figure, a ***straight line***), but the actual ***process*** of blood pressure from second to second ***fluctuates*** about such a trend, as a natural biological phenomenon and as a result of things the individual does, such as drink a cup of coffee.

Of course, the extent to which an actual process fluctuates depends on the outcome in question. Blood pressure might be expected to fluctuate fairly dramatically and locally, while a phenomenon like dental distance might barely fluctuate, if at all, and, if it does, might exhibit subtle changes over a much longer time interval.

The dotted continuous line thus depicts an actual ***realization*** of the underlying ***response process*** for this individual in continuous time; that is, a realization of the ***stochastic process*** of blood pressure taking place within this individual.

- The ***solid line*** can then be thought of as the "***inherent***" response profile for the individual, reflecting the smooth trend. More formally, at each time point the solid line represents the ***mean*** of all possible realizations of blood pressure that could arise for this individual, so that the line represents the ***mean response*** over time for this individual, averaging across all possible realization paths that could occur.

  Thus, if we were to consider regression modeling of the response for ***this individual***, we would model this mean response profile. Here, we have taken it to be a ***straight line*** for simplicity, but the same considerations apply to a more complicated relationships, such as that in that in the PK study of ***EXAMPLE 4***, where the ***mean response*** would instead be the smooth curve traced out by the one compartment model in (2.1).

- The responses ***actually observed*** in Figure 2.2(a) and depicted in Figure 2.2(b) by the diamonds do not lie exactly on this smooth line representing the ***mean response*** for this individual. Instead, they deviate from it in a positive or negative fashion. These deviations can be conceptualized to be the result of the combined effects of two phenomena.

  The first is obviously the fact that the observed responses reflect the underlying, ***actual realization*** shown in the dotted line. The second is that, although we wish to ascertain blood pressure ***perfectly***, the device used may be subject to ***measurement error*** so that the responses we actually observe at the intermittent times shown deviate from the realization somewhat, reflected in Figure 2.2(b) by the fact that the plotting symbols do not necessarily lie on the dotted line.

The deviation of the observed response at any time from the smooth mean response trend is thus the ***net result*** of the deviation from the smooth trend of the actual realization of blood pressure and the deviation of the observed response from the true realized blood pressure due to measurement error.

Summarizing, this conceptual representation for a ***single individual*** says that the response vector for the individual comprises ***intermittent observations*** , possibly subject to ***measurement error*** , on a ***stochastic process*** , whose ***realizations*** fluctuate about a smooth ***inherent*** trend,.

WITHIN-INDIVIDUAL SOURCES OF CORRELATION: We can summarize this perspective formally. For now, focus on a given individual i and consider the ***individual-level stochastic process***

$$\mathcal{Y}_i(t, \boldsymbol{u}_i) = \mu_i(t, \boldsymbol{u}_i) + e_{Pi}(t, \boldsymbol{u}_i). \tag{2.3}$$

- In (2.3), $\mathcal{Y}_i(t, \boldsymbol{u}_i)$ is the stochastic process of the actual, ***realized response*** if we were able to ascertain it ***perfectly*** in ***continuous time*** under the conditions dictated by the ***within-individual covariate*** $\boldsymbol{u}_i$. It can be ***decomposed*** into two components.
- $\mu_i(t, \boldsymbol{u}_i)$ represents the ***smooth inherent trend*** for individual i, which we have taken to be a straight line in Figure 2.2(b). As we are focusing ***only*** on individual i for now (i.e., ***conditioning*** on individual i), we view $\mu_i(t, \boldsymbol{u}_i)$ as a ***fixed*** feature of individual i.
- $e_{Pi}(t, \boldsymbol{u}_i)$, is the ***process*** associated with deviation of the (error-free) response from the inherent trend. Viewing $\mu_i(t, \boldsymbol{u}_i)$ as ***fixed*** for individual i, it is natural to take $E\{e_{Pi}(t, \boldsymbol{u}_i)|\boldsymbol{u}_i\} = 0$ for all t. As in regression analysis, we condition on $\boldsymbol{u}_i$, and, as above, implicitly on individual i. It follows that

  $$E\{\mathcal{Y}_i(t, \boldsymbol{u}_i)|\boldsymbol{u}_i\} = \mu_i(t, \boldsymbol{u}_i),$$

  consistent with the definition of $\mu_i(t, \boldsymbol{u}_i)$ as the mean response for individual i, averaging across all possible realizations.

Given the representation (2.3) of the actual response process $\mathcal{Y}_i(t, \boldsymbol{u}_i)$, we can write the responses Y_{ij} observed at intermittent times t_{ij}, $j = 1, \ldots, n_i$, as

$$\begin{aligned} Y_{ij} &= \mathcal{Y}_i(t_{ij}, \boldsymbol{u}_i) + e_{Mij} = \mu_i(t_{ij}, \boldsymbol{u}_i) + e_{Pi}(t_{ij}, \boldsymbol{u}_i) + e_{Mij} \\ &= \mu_i(t_{ij}, \boldsymbol{u}_i) + e_{Pij} + e_{Mij} = \mu_i(t_{ij}, \boldsymbol{u}_i) + e_{ij}, \quad e_{ij} = e_{Pij} + e_{Mij}. \end{aligned} \tag{2.4}$$

In (2.4), e_{Mij} is a deviation due to ***measurement error*** in the device or procedure used to ascertain the response at time t_{ij}. If we assume that the measuring device has ***no systematic bias*** , then it is natural to assume that $E(e_{Mij}|\boldsymbol{u}_i) = 0$ for all (i, j).

From above, $E(e_{Pij}|\boldsymbol{u}_i) = 0$, so that $E(e_{ij}|\boldsymbol{u}_i) = 0$ for all (i, j). Thus, e_{ij} is the ***overall within-individual deviation*** reflecting the net effects of deviation of the actual realization from the smooth trend and deviation of the observed response from the realization due to measurement error at time t_{ij}.

The representation (2.3) provides a convenient framework for thinking about correlation arising at the ***individual level*** among the observed responses Y_{ij}, $j = 1, \ldots, n_i$. Viewing the smooth inherent trend $\mu_i(t, \boldsymbol{u}_i)$, as ***fixed*** for individual i (so continuing to ***condition*** on individual i), the ***correlation*** between two observations Y_{ij} and $Y_{ij'}$, say, is dictated in part by the properties of the deviation process $e_{Pi}(t, \boldsymbol{u}_i)$ and the measurement error deviations e_{Mij} and $e_{Mij'}$.

- Consider the ***realization deviation*** process $e_{Pi}(t, \boldsymbol{u}_i)$. If we could observe the ***error-free response process*** (2.3) at two points in time very close together, it is very likely that the two observations would tend to be on the same side, i.e., fluctuate on the same side, of the ***inherent mean response trajectory*** ; e.g., deviating it from it positively or negatively together. However, if the time points were very far apart, the two observations would be just as likely to deviate negatively or positively from the inherent trend. That is, responses close together in time would tend to be more "alike" or "***related*** " in this sense than responses far apart in time.

  This suggests that the ***correlation*** between $e_{Pi}(t, \boldsymbol{u}_i)$ and $e_{Pi}(s, \boldsymbol{u}_i)$ for times t and s is expected to be ***positive*** for t and s close together and to "damp out" to zero as t and s are farther apart. As in classical ***time series analysis*** , it might be reasonable to suppose that the correlation between $e_{Pi}(t, \boldsymbol{u}_i)$ and $e_{Pi}(s, \boldsymbol{u}_i)$ depends on the ***time distance*** $|t - s|$ (and not on the actual values of $e_{Pi}(t, \boldsymbol{u}_i)$ and $e_{Pi}(s, \boldsymbol{u}_i)$ themselves); i.e., that the process $e_{Pi}(t, \boldsymbol{u}_i)$ is ***stationary***. In Section 2.5, we discuss ***models*** that might be plausible representations of such correlation.

  Accordingly, we expect e_{Pij} and $e_{Pij'}$ associated with Y_{ij} and $Y_{ij'}$ to be (conditionally) ***positively correlated*** in a way that is related to the time distance $|t_{ij} - t_{ij'}|$.

- It is generally accepted that most measuring devices make ***haphazard*** errors, so that it is plausible that deviations due to error in the device at different time points are ***not related*** no matter how close together in time they occur. Accordingly, it is usually reasonable to assume that e_{Mij} and $e_{Mij'}$ associated with Y_{ij} and $Y_{ij'}$ are ***independent***.

- It is often also assumed that the ***magnitude*** of errors in measurement is ***unrelated*** to the size of the thing being measured, given the haphazard nature of errors. Under this assumption, it is reasonable to assume that the process $e_{Pi}(t, \boldsymbol{u}_i)$ is ***independent*** of any measurement error deviation, which implies that e_{Pij} and e_{Mij} are ***independent*** for all $j = 1, \ldots, n_i$ and in fact e_{Pij} and $e_{Mij'}$ are ***independent*** for $j, j' = 1, \ldots, n_i$.

There are measuring devices for which the magnitude of errors ***is related*** to the size of the thing being measured, in which this assumption might not hold; we discuss these considerations in later chapters.

- It is generally assumed in ***time series analysis*** and ***spatial statistics*** (which can be viewed as an extension of time series analysis to more than one dimension), that the ***magnitude*** of measurement error is ***negligible*** relative to that of the realization deviation process and can be ignored. For example, in financial applications, the outcome may be a stock price, which is observed exactly. In the spatial statistics literature, the measurement error deviation e_{Mij} is analogous to the so-called "***nugget effect***."

  For the types of longitudinal outcomes with which we are concerned, e.g., arising in health science or agricultural applications, measurement error ***may or may not*** be negligible.

Under these assumptions, conditional on each individual i, Y_{ij} and $Y_{ij'}$ are correlated, and the magnitude of the correlation is likely ***nonnegligible*** if t_{ij} and $t_{ij'}$ are close together in time.

- This correlation comes about because of the ***time-ordered*** data collection on individual i, so is a ***within-individual*** phenomenon. Under the conceptual model and our assumptions, this within-individual correlation involves the correlation between e_{Pij} and $e_{Pij'}$.

- In fact, this correlation is important even if the focus of inference is on individual i ***only***. As we remarked previously, if the goal is to make inference on $\mu_i(t, \boldsymbol{u}_i)$, the ***inherent mean response trend*** for individual i only, this correlation is ***relevant***.

  In particular, suppose we believe that $\mu_i(t, \boldsymbol{u}_i)$ is of the form

  $$\mu_i(t, \boldsymbol{u}_i) = f(t, \boldsymbol{u}_i, \beta_i), \tag{2.5}$$

  for some function f depending on an ***individual-specific*** parameter β_i. Then inference on $\mu_i(t, \boldsymbol{u}_i)$ boils down to inference on β_i. These developments demonstrate why, as we discussed for the PK study in (2.1), using standard regression methods to do this might be suspect, as standard methods assume that observations are ***uncorrelated*** or ***independent***. Thus, their use must be critically examined when the data are collected over time.

- A justification that is often given for using standard regression methods to fit a model for $\mu_i(t, \boldsymbol{u}_i)$ in ***longitudinal*** situations such as the PK study is that the intermittent response observations are ***sufficiently far apart*** in time to render this correlation, which arises because of fluctuations that are "***local***" in nature, ***practically negligible***.

In the foregoing, we treated $\mu_i(t, \boldsymbol{u}_i)$ as fixed, because the focus was on individual i and within-individual phenomena only. We now step back and consider the ***population***. For simplicity, assume that for the remainder of this discussion that there are no within-individual covariates $\boldsymbol{u}_i$, and write (2.3) as

$$\mathcal{Y}_i(t) = \mu_i(t) + e_{Pi}(t),$$

so that the ***inherent trajectory*** for individual i is $\mu_i(t)$. We return to the more general formulation including such covariates in later chapters.

AMONG-INDIVIDUAL, POPULATION-LEVEL SOURCES OF CORRELATION: Consider the population of ***all individuals*** of interest. For example, in the dental study, this might be the population of all girls only or the population of all children of either gender.

Each individual has his/her own ***inherent trend*** about which realizations of his/her stochastic process fluctuate, observations on which might be subject to measurement error.

- The bold line in Figure 2.2(b) represents the ***overall population mean response*** in ***continuous time***, averaged across all possible responses that could be observed on ***all individuals*** in the population at each time. Denote the overall mean as $\mu(t)$.
- The ***inherent trend*** for any individual i, $\mu_i(t)$, "***places***" that individual in the population of all individuals ***relative*** to the overall population mean response trajectory $\mu(t)$. Thus, from a population perspective, observed responses on, say, the uppermost individual in Figure 2.2(b) tend to be "***high***" because they are realizations (possibly subject to measurement error) fluctuating about this individual's trend, which is "***high***" relative to $\mu(t)$. Similarly, observed responses on the bottom individual tend to be "***low***" then "***high***" because of the steepness of his/her trend relative to that of the overall population mean.
- In general, responses observed on the same individual will tend to be "***alike***" because they are on the same individual and thus vary about a ***shared inherent trend***. Consequently, ***correlation*** can be thought to arise among responses Y_{ij} and $Y_{ij'}$ on individual i because of this common dependence on the shared, individual-specific trend.

This is an ***among-individual*** , ***population-level*** phenomenon. We can ***formalize*** the foregoing observations as follows. We ***decompose*** the ***inherent trajectory*** for any individual i as

$$\mu_i(t) = \mu(t) + \mathcal{B}_i(t). \tag{2.6}$$

In (2.6), $\mathcal{B}_i(t)$ represents the ***deviation*** of i's inherent trajectory from the overall population mean trend, leading to $\mu_i(t)$ begin "high" or "low" at any time t or "steeper" or "shallower" relative to $\mu(t)$.

If the population is ***heterogeneous*** ; e.g., children of both genders in the dental study, we can generalize (2.6) to allow the overall population mean to be ***different*** depending on the value of ***among-individual covariates*** $\boldsymbol{a}_i$, where each individual i's inherent trajectory deviates from the overall mean corresponding to his/her $\boldsymbol{a}_i$. For example, in the dental study, we could have a separate population mean for each gender. We represent this as

$$\mu_i(t) = \mu(t, \boldsymbol{a}_i) + \mathcal{B}_i(t), \tag{2.7}$$

where now $\mu(t, \boldsymbol{a}_i)$ is the overall population mean relevant to individual i; i.e., determined by covariate value $\boldsymbol{a}_i$. Clearly, (2.7) subsumes (2.6).

In (2.6) and (2.7), then, $\mathcal{B}_i(t)$ is the ***deviation*** from the population mean trend at any time t that dictates where the inherent trend for individual i "***sits***" in the population relative to the population mean trend. Intuitively, $\mathcal{B}_i(t)$, should have ***mean zero*** , so that the inherent trajectories over the entire population average out to yield the overall population mean. We can formalize this by assuming that

$$E\{\mathcal{B}_i(t)|\boldsymbol{a}_i\} = 0 \quad \text{for all } t,$$

where conditioning on $\boldsymbol{a}_i$ ensures that this applies for each value of $\boldsymbol{a}_i$ when the overall mean depends on among-individual covariates. Thus, $\mathcal{B}_i(t)$ characterizes ***among-individual*** behavior.

Using (2.7), we can write (2.3) (suppressing $\boldsymbol{u}_i$) as

$$\mathcal{Y}_i(t) = \mu(t, \boldsymbol{a}_i) + \mathcal{B}_i(t) + e_{Pi}(t)$$

and thus write (2.4) as

$$Y_{ij} = \mu(t_{ij}, \boldsymbol{a}_i) + \mathcal{B}_i(t_{ij}) + e_{Pij} + e_{Mij}. \tag{2.8}$$

From (2.8), if the deviations $\mathcal{B}_i(t_{ij})$ are ***correlated*** across j, then Y_{ij} and $Y_{ij'}$ will be correlated. Intuitively, we expect $\mathcal{B}_i(t_{ij})$ and $\mathcal{B}_i(t_{ij'})$ at any two times t_{ij} and $t_{ij'}$ to be ***correlated*** because they ***jointly determine*** where individual i's ***smooth*** inherent trajectory "***sits***" relative to the overall mean. This correlation is thus an ***among-individual*** , ***population-level*** phenomenon.

In our discussion of ***subject-specific*** modeling in the next section, we demonstrate explicitly in a particular example how $\mathcal{B}_i(t)$ can be characterized and how this correlation then arises.

If inference focuses on a specific individual i, it should be clear that such ***population-level*** sources of correlation are ***irrelevant***. For instance, in the PK study of ***EXAMPLE 4*** in Section 1.2, if interest focuses on the PK parameters for a ***particular*** subject, where that subject's concentration-time profile "***sits***" in the population of subjects, and thus this type of correlation, is of ***no importance***.

However, if interest focuses on the ***population*** from which the subject was drawn, then this correlation ***is relevant***.

OVERALL PATTERN OF VARIANCE AND CORRELATION: We merge these developments to obtain a representation that makes explicit how ***within-*** and ***among-individual*** sources of correlation ***combine*** to produce an ***overall pattern*** of variance and correlation.

Consider $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$, with elements recorded at times $t_{i1}, \ldots, t_{in_i}$. Define

$$\boldsymbol{\mu}_i = \begin{pmatrix} \mu(t_{i1}, \boldsymbol{a}_i) \\ \vdots \\ \mu(t_{in_i}, \boldsymbol{a}_i) \end{pmatrix}, \quad \boldsymbol{\mathcal{B}}_i = \begin{pmatrix} \mathcal{B}_i(t_{i1}) \\ \vdots \\ \mathcal{B}_i(t_{in_i}) \end{pmatrix}, \quad \boldsymbol{e}_i = \boldsymbol{e}_{Pi} + \boldsymbol{e}_{Mi} = \begin{pmatrix} e_{i1} \\ \vdots \\ e_{in_i} \end{pmatrix} = \begin{pmatrix} e_{Pi1} \\ \vdots \\ e_{Pin_i} \end{pmatrix} + \begin{pmatrix} e_{Mi1} \\ \vdots \\ e_{Min_i} \end{pmatrix},$$

where $\boldsymbol{\mu}_i$ is the fixed population mean response vector whose elements are the population mean responses at the time points t_{ij} at which i is observed, possibly depending on among-individual covariates $\boldsymbol{a}_i$. Then, from (2.8), we have

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\mathcal{B}}_i + \boldsymbol{e}_i = \boldsymbol{\mu}_i + \boldsymbol{\mathcal{B}}_i + \boldsymbol{e}_{Pi} + \boldsymbol{e}_{Mi}. \tag{2.9}$$

NOTATION: We use var$(\cdot)$ to denote both ***variance*** of a scalar random variable and ***covariance matrix*** of a random vector. The meaning should be clear from the context.

From (2.9), we can calculate the ***covariance matrix*** of $\boldsymbol{Y}_i$. We assume that $\boldsymbol{\mathcal{B}}_i$ and $\boldsymbol{e}_i$ are ***independent*** for now; we discuss situations in which this would or would not be a reasonable assumption in subsequent chapters. Recall that we assume that the process $e_{Pi}(t)$ and the measurement error deviations e_{Mij} are ***independent***, so that the random vectors $\boldsymbol{e}_{Pi}$ and $\boldsymbol{e}_{Mi}$ are independent.

With these assumptions, we have

$$\text{var}(\boldsymbol{Y}_i|\boldsymbol{a}_i) = \text{var}(\mathcal{B}_i|\boldsymbol{a}_i) + \text{var}(\boldsymbol{e}_i|\boldsymbol{a}_i) = \text{var}(\mathcal{B}_i|\boldsymbol{a}_i) + \{\text{var}(\boldsymbol{e}_{Pi}|\boldsymbol{a}_i) + \text{var}(\boldsymbol{e}_{Mi}|\boldsymbol{a}_i)\}. \quad (2.10)$$

- From the previous discussion, it is reasonable to expect that the elements of $\mathcal{B}_i$, $\mathcal{B}_i(t_{ij})$, are ***correlated*** across j, as they involve features of the ***shared inherent trajectory***. Thus, var($\mathcal{B}_i|\boldsymbol{a}_i$), is a ***nondiagonal*** matrix. Its diagonal elements reflect variability in the elements of $\boldsymbol{Y}_i$ and its off-diagonal elements reflect ***covariance***, and thus ***correlation***, due to ***among-individual*** phenomena.

- Also from above, the elements e_{Pij} of $\boldsymbol{e}_{Pi}$ are expected to be ***positively correlated*** across j, where the correlation "damps out" as the time points become farther apart. Thus, var($\boldsymbol{e}_{Pi}|\boldsymbol{a}_i$) is also a ***nondiagonal*** matrix.

- By the nature of measurement error discussed earlier, the $\boldsymbol{e}_{Mij}$ are reasonably assumed ***independent*** across j, so that var($\boldsymbol{e}_{Mi}|\boldsymbol{a}_i$) is a ***diagonal matrix***.

- The sum $\{\text{var}(\boldsymbol{e}_{Pi}|\boldsymbol{a}_i) + \text{var}(\boldsymbol{e}_{Mi}|\boldsymbol{a}_i)\}$ in braces in (2.10) reflects variability and correlation due to ***within-individual*** sources. The diagonal elements of the sum reflect variation in the elements of $\boldsymbol{Y}_i$ arising from the combined effects of fluctuations about individual i's inherent trend and measurement error. The off-diagonal elements reflect correlation due to the ***time-ordered*** nature of ***within-individual*** data collection.

RESULT: From (2.10) and with these observations, it is clear that the covariance matrix var($\boldsymbol{Y}_i|\boldsymbol{a}_i$) of the observed responses on an individual exhibits an overall pattern of variance and correlation that reflects the contributions of ***both*** within- and among-individual components. If there were ***within-individual covariates*** $\boldsymbol{u}_i$, similar arguments would apply, where the conditioning would be on $\boldsymbol{x}_i$, which incorporates $\boldsymbol{u}_i$ and $\boldsymbol{a}_i$.

The same interpretation holds in ***fancier*** versions of the above framework, which we discuss in subsequent chapters.

We now consider the two main approaches to modeling longitudinal data, which take different perspectives on acknowledging correlation.

2.4 Population-averaged versus subject-specific modeling

Zeger, Liang, and Albert (1988) coined the terms ***subject-specific*** and ***population-averaged*** to describe the two major approaches to statistical modeling of longitudinal data we now introduce. The terminology "***subject-specific***" reflects the focus of these authors on challenges arising in research involving ***humans*** in the health sciences; a more generic term would be ***individual-specific***.

We motivate both of these approaches by considering the dental study data, which are shown again in Figure 2.3. Dental distance was measured on each of $m = 27$ children (11 girls and 16 boys).

Here, Y_{ij} is the dental distance measurement for the ith child, $i = 1, \ldots, m = 27$, at time $j = 1 \ldots, 4$, where, for all children, $(t_{i1}, \ldots, t_{in_i})^T = (8, 10, 12, 14)^T$, so $n_i = 4$ for all children. As before, there is one ***among-individual covariate***, gender, where $g_i = 0$ if child i is a girl and $g_i = 1$ if child i is a boy, so that $\boldsymbol{a}_i = g_i$, $i = 1, \ldots, m$. There are no ***within-individual covariates*** $\boldsymbol{u}_i$, so that $\boldsymbol{z}_{ij} = t_{ij}$.

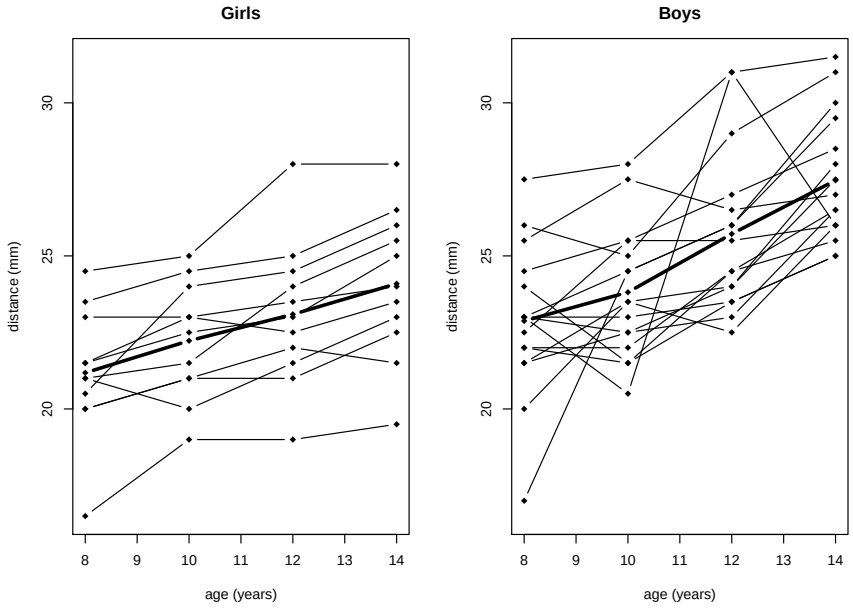

Figure 2.3: *Figure 2.1, repeated.*

The choice of modeling strategy is driven by the nature of the ***scientific questions*** of interest.

In the dental study, one goal as stated in Section 1.2 was to compare the ***rate of change*** of dental distance between boys and girls. Although this question seems straightforward on the surface, as we now demonstrate, it can be made precise in ***different ways***.

We first restate the conceptual framework developed in the previous section for convenience. We continue to assume that there are no within-individual covariates $\boldsymbol{u}_i$. Recall that in this representation, individual i has a true stochastic process

$$\mathcal{Y}_i(t) = \mu_i(t) + e_{Pi}(t) = \mu(t, \boldsymbol{a}_i) + \mathcal{B}_i(t) + e_{Pi}(t),$$

so that

$$Y_{ij} = \mu_i(t_{ij}) + e_{Pij} + e_{Mij} = \mu_i(t_{ij}) + e_{ij} \tag{2.11}$$

$$= \mu(t_{ij}, \boldsymbol{a}_i) + \mathcal{B}_i(t_{ij}) + e_{Pij} + e_{Mij} = \mu(t_{ij}, \boldsymbol{a}_i) + \mathcal{B}_i(t_{ij}) + e_{ij}, \tag{2.12}$$

where $\mu_i(t)$ is the ***inherent trend*** for individual i, $\mu(t, \boldsymbol{a}_i)$ is the ***overall population mean response*** for individuals with among-individual covariate $\boldsymbol{a}_i$, and $\mathcal{B}_i(t)$ is the ***among-individual*** deviation of i's trend from the overall mean $\mu(t, \boldsymbol{a}_i)$ at time t.

We now explicate how modeling would proceed under the two approaches and relate it back to the conceptual framework.

SUBJECT-SPECIFIC MODELING: As suggested by the name, this modeling approach is natural when questions of interest are interpreted to be about ***individual-specific behavior***. Consequently, its development mirrors the conceptual framework closely.

From Figure 2.3 , ***each child's*** distance measures appear to follow roughly a ***straight line*** trajectory, with some ***fluctuations***. This suggests adopting a ***model*** in the form of a straight line for each child, where each child has his or her own ***individual-specific*** intercept and slope. That is, represent the observed data as for child i as

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij}, \tag{2.13}$$

This model ***explicitly acknowledges*** child-specific slopes.

Taking this point of view, the question of interest of comparing ***rate of change*** between boys and girls can be ***formalized*** as comparing the "***typical***" or ***average*** individual-specific slope β_{1i} in the ***population*** of boys to that in the population of girls.

From the perspective of the conceptual framework, assuming (2.13) is tantamount to assuming that each child has his or her own ***stochastic process*** of the form

$$\mathcal{Y}_i(t) = \beta_{0i} + \beta_{1i} t + e_{Pi}(t),$$

so that i's inherent trend is $\mu_i(t) = \beta_{0i} + \beta_{1i} t$, with child-specific intercept β_{0i} and slope β_{1i}. Following the discussion in the last section, this suggests that e_{ij} in (2.13) can be decomposed into components for realization and measurement error deviations, as in (2.12).

Because interest focuses on comparing the "***typical***" or ***average*** slope between the populations of boys and girls, it is natural to conceive that intercepts and slopes ***vary*** in these populations about ***typical*** or mean values; i.e.,

$$\begin{aligned} \beta_{0i} &= \beta_{0,B} + b_{0i}, \quad \beta_{1i} = \beta_{1,B} + b_{1i}, \quad \text{if } i \text{ is a boy} \\ \beta_{0i} &= \beta_{0,G} + b_{0i}, \quad \beta_{1i} = \beta_{1,G} + b_{1i}, \quad \text{if } i \text{ is a girl} \end{aligned} \tag{2.14}$$

In (2.14), b_{0i} and b_{1i} are mean zero, ***child-specific deviations*** that acknowledge that individual child intercepts and slopes ***vary*** about the ***average intercept and slope*** $\beta_{0,B}$ and $\beta_{1,B}$ for the population of boys and $\beta_{0,G}$ and $\beta_{1,G}$ for the population of girls.

The ***question of interest*** can then be stated ***precisely*** as whether or not the average (mean) slopes $\beta_{1,B}$ and $\beta_{1,G}$ differ; that is, whether or not $\beta_{1,B} = \beta_{1,G}$.

By substitution of (2.14) into (2.13), we can rewrite (2.13) as

$$Y_{ij} = \{\beta_{0,B} g_i + \beta_{0,G}(1 - g_i)\} + \{\beta_{1,B} g_i + \beta_{1,G}(1 - g_i)\} t_{ij} + b_{0i} + b_{1i} t_{ij} + e_{ij}. \tag{2.15}$$

Relating this back to the conceptual framework, we have that the ***overall mean response*** at t_{ij} is

$$\mu(t_{ij}, \boldsymbol{a}_i) = \{\beta_{0,B} g_i + \beta_{0,G}(1 - g_i)\} + \{\beta_{1,B} g_i + \beta_{1,G}(1 - g_i)\} t_{ij}. \tag{2.16}$$

The ***individual-specific deviation*** from the overall mean at t_{ij} is

$$\mathcal{B}_i(t_{ij}) = b_{0i} + b_{1i} t_{ij}. \tag{2.17}$$

- Note from (2.17) that $\mathcal{B}_i(t_{ij})$ and $\mathcal{B}_i(t_{ij'})$ for any times t_{ij} and $t_{ij'}$ are ***correlated*** because both depend on b_{0i} and b_{1i}.
- This example thus provides an explicit demonstration of how the ***among-individual correlation*** discussed in the previous section arises.

The resulting model (2.15) thus acknowledges ***explicitly*** child-specific behavior. We can summarize the model succinctly as in (2.10) as

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\mathcal{B}}_i + \boldsymbol{e}_i, \tag{2.18}$$

where $\boldsymbol{\mu}_i$ has jth element $\{\beta_{0,B} g_i + \beta_{0,G}(1 - g_i)\} + \{\beta_{1,B} g_i + \beta_{1,G}(1 - g_i)\} t_{ij}$, and $\boldsymbol{\mathcal{B}}_i$ has jth element $b_{0i} + b_{1i} t_{ij}$.

To complete the model, we require ***assumptions*** on

- $\boldsymbol{b}_i = (b_{0i}, b_{1i})^T$, dictating individual deviations from the overall mean in (2.18) and thus ***among-individual*** variation and correlation. By definition, we have $E(\boldsymbol{b}_i|\boldsymbol{a}_i) = \mathbf{0}$; we thus require an assumption on var$(\boldsymbol{b}_i|\boldsymbol{a}_i)$ characterizing ***among-individual*** variation and correlation.
- $\boldsymbol{e}_i$ in (2.18), representing ***within-individual*** variation and correlation, which we can decompose further into the components $\boldsymbol{e}_{Pi}$ and $\boldsymbol{e}_{Mi}$ representing contributions due to realization and measurement error processes. Again, we have $E(\boldsymbol{e}_{Pi}|\boldsymbol{a}_i) = \mathbf{0}$ and $E(\boldsymbol{e}_{Mi}|\boldsymbol{a}_i) = \mathbf{0}$, so that $E(\boldsymbol{e}_i|\boldsymbol{a}_i) = \mathbf{0}$; thus, we require an assumption on var$(\boldsymbol{e}_i|\boldsymbol{a}_i)$ characterizing ***within-individual*** variation and correlation.

The key message is that, interpreting the question of interest to be one about the ***typical*** or ***average*** behavior of ***individual-specific*** features, one is led naturally to a model that acknowledges ***both within-*** and ***among-individual*** sources of variation and correlation explicitly.

- Such a model is referred to as ***subject-specific*** for obvious reasons.
- In Chapters 6 and 9, we consider such modeling in detail.

POPULATION-AVERAGED MODELING: An ***alternative interpretation*** of the question of interest of comparing ***rate of change*** between boys and girls comes about from thinking ***directly*** about the populations of boys and girls, as follows.

Each child has a ***random response vector*** $\boldsymbol{Y}_i$. For the ith boy, $\boldsymbol{Y}_i$ can be assumed to follow a ***multivariate distribution*** with an appropriate ***mean vector*** and ***covariance matrix*** , and similarly for girls. Thus, represent the populations of boys and girls directly by these distributions.

From Figure 2.1, the ***sample mean response*** trajectory for each gender appears to follow a ***straight line***. This suggests that a reasonable ***model*** for the overall mean for each gender at time t_{ij} is

$$\begin{array}{ll} \beta_{0,B} + \beta_{1,B} t_{ij} & \text{for boys,} \\ \beta_{0,G} + \beta_{1,G} t_{ij} & \text{for girls.} \end{array} \tag{2.19}$$

We use the ***same symbols*** for the intercepts and slopes that characterize these mean response vectors as we did for the subject-specific model, but the interpretation is ostensibly ***different***. E.g., whereas $\beta_{1,B}$ in (2.14) represents the "***typical***" or ***average slope*** in the population of boys, here it represents the ***slope of the population mean response vector*** for boys, and similarly for the other parameters.

From this perspective, comparing ***rate of change*** between boys and girls can be formalized as comparing the rate of change of the ***population mean response*** for boys with that for girls; that is, comparing $\beta_{1,B}$ and $\beta_{1,G}$ in (2.19).

The model is completed by making an assumption on the way in which observations Y_{ij} ***deviate from*** the population mean (2.19) for boys and girls. To this end, write

$$Y_{ij} = \mu(t_{ij}, \boldsymbol{a}_i) + \epsilon_{ij}, \tag{2.20}$$

where $E(\epsilon_{ij}|\boldsymbol{a}_i) = 0$. Assumptions on ϵ_{ij} then lead to an assumption on the ***covariance matrices*** of $\boldsymbol{Y}_i$ for boys and girls.

In the context of the conceptual framework, (2.20) is (2.12), where the among- and within-individual deviations are collected into the single term

$$\epsilon_{ij} = \mathcal{B}_i(t_{ij}) + e_{Pij} + e_{Mij} = \mathcal{B}_i(t_{ij}) + e_{ij}. \tag{2.21}$$

In (2.20), from (2.19), the overall population mean for child i at time t_{ij} is

$$\mu(t_{ij}, \boldsymbol{a}_i) = \{\beta_{0,B} g_i + \beta_{0,G}(1 - g_i)\} + \{\beta_{1,B} g_i + \beta_{1,G}(1 - g_i)\} t_{ij}. \tag{2.22}$$

The ***overall deviation*** ϵ_{ij} in (2.21) thus represents the ***combined effect*** of ***all*** sources of variation, ***within-*** and ***among-individuals***, that contribute to the fact that Y_{ij} values vary about $\mu(t_i, \boldsymbol{a}_i)$ in the populations of boys and girls (depending on if i is a boy or girl). Taking this point of view, we ***do not acknowledge explicitly*** these different sources.

We can write (2.20) succinctly as

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i, \tag{2.23}$$

where $E(\boldsymbol{\epsilon}_i|\boldsymbol{a}_i) = \boldsymbol{0}$, and $\boldsymbol{\mu}_i$ has jth element $\{\beta_{0,B} g_i + \beta_{0,G}(1 - g_i)\} + \{\beta_{1,B} g_i + \beta_{1,G}(1 - g_i)\} t_{ij}$.

- The model is completed by making assumptions directly on $\text{var}(\boldsymbol{\epsilon}_i|\boldsymbol{a}_i)$; i.e., directly on the form of the ***overall, aggregate*** pattern of covariance/correlation.

The key message is that, interpreting the question of interest to be one about ***overall population mean*** behavior, one is led naturally to the conventional approach of modeling this mean ***directly*** and more generally modeling the ***multivariate distribution*** of response vectors, or at least the ***mean*** and ***covariance matrix*** thereof, ***directly***.

- Such a model is referred to as ***population-averaged*** for obvious reasons.
- In Chapters 5 and 8, we consider such modeling in detail.

CONTRASTING SUBJECT-SPECIFIC AND POPULATION-AVERAGED MODELING: The foregoing development illustrates the ***conceptual difference*** between these two modeling strategies:

- ***Subject-specific*** modeling is appropriate when it is feasible or of direct scientific interest to postulate a model for the ***individual-specific inherent trend*** or, equivalently, ***individual-specific mean response***, and questions can be posed as pertaining to the "***typical***" or ***average*** behavior of ***individual-specific parameters*** that describe this trend, like β_{0i} and β_{1i} in (2.13).
- ***Population-averaged*** modeling is appropriate when questions of scientific interest can be posed as pertaining to the ***overall population mean response***.

As the dental study example demonstrates, in certain circumstances, taking ***either*** approach leads to the ***same model*** for overall population mean response.

- This is the case when the models used in both approaches are ***linear***.
- In particular, in the subject-specific approach to modeling the dental data, the child-specific model (2.13),

  $$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij},$$

  is ***linear*** in the individual-specific intercept and slope parameters β_{0i} and β_{1i}, and, in (2.14), β_{0i} and β_{1i} are ***linear*** in the "***typical***" parameters $\beta_{0,B}$, $\beta_{1,B}$, $\beta_{0,G}$, and $\beta_{1,G}$ in the populations of boys and girls.

  This leads to the ***linear*** model for ***population mean response*** (2.16), namely,

  $$\mu(t_{ij}, \boldsymbol{a}_i) = \{\beta_{0,B} g_i + \beta_{0,G}(1 - g_i)\} + \{\beta_{1,B} g_i + \beta_{1,G}(1 - g_i)\} t_{ij}.$$

- In the population-averaged approach to modeling these data, we modeled the population mean response in ***directly*** (2.22) by the ***linear model***

  $$\mu(t_{ij}, \boldsymbol{a}_i) = \{\beta_{0,B} g_i + \beta_{0,G}(1 - g_i)\} + \{\beta_{1,B} g_i + \beta_{1,G}(1 - g_i)\} t_{ij}.$$

- These models for population mean response are thus ***identical***.

The upshot is that, when linear models are used, the distinction between subject-specific and population-averaged modeling is ***conceptual*** only. The model resulting from either approach and the ensuing inferences can be interpreted in ***either*** of two ways. In the case of the dental study:

- From the ***subject-specific*** perspective $\beta_{1,B}$ and $\beta_{1,G}$ are the ***mean slopes*** in the populations of boys and girls and can be interpreted as the "***typical parameter values***" in those populations.
- From the ***population-averaged*** perspective, $\beta_{1,B}$ and $\beta_{1,G}$ are slopes that characterize the "***typical response vector*** ;" i.e., the ***overall mean response*** in each of these populations.

Regardless of which strategy the analyst takes to arrive at a final model, ***both interpretations are valid***.

So ***why bother*** to distinguish between the two approaches?

NONLINEAR MODELS: This dual interpretation ***does not hold*** when ***nonlinear models*** like the PK model in (2.1), are involved. This is also the case when the models are those popular when analyzing ***discrete outcome*** as for the counts in the seizure study in ***EXAMPLE 5*** or binary wheezing status in the Six Cities study in ***EXAMPLE 6*** of Section 1.2, which are also ***nonlinear*** in parameters.

In these settings, the choice between subject-specific and population-averaged modeling is guided by the nature of the scientific questions and is ***absolutely critical*** to achieving appropriate, scientifically relevant inferences. We demonstrate this in detail in Chapters 7-9.

There is one fundmental difference that persists ***whether or not*** the modeling of the mean is linear.

COVARIANCE/CORRELATION STRUCTURE: Although in the linear case the two approaches lead to the same inferences on parameters describing ***mean response*** , they dictate ***different*** models for the covariance matrix of a response vector $\boldsymbol{Y}_i$.

- As exemplified by the representation

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \mathcal{B}_i + \boldsymbol{e}_i$$

  in (2.18) arising from the subject-specific approach, this strategy leads to a covariance model that involves ***distinct components*** var($\mathcal{B}_i|\boldsymbol{a}_i$) and var($\boldsymbol{e}_i|\boldsymbol{a}_i$), say, representing ***among-*** and ***within-individual*** variation and correlation, respectively. That is, subject-specific modeling naturally ***induces*** a ***specific structure*** for the ***overall pattern*** of variation and correlation.

Thus, subject-specific modeling requires that data analyst to posit covariance models for ***each***. The models chosen then ***induce*** a structure for the overall pattern of variation and correlation.

- In contrast, as is evident from the representation

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

in (2.23) , the population-averaged approach ***does not*** distinguish these different sources of variation and correlation. Rather, only their overall net or ***aggregate*** effect is acknowledged.

Thus, in population-averaged modeling, the data analyst posits a model for the overall pattern of variation and correlation, var($\boldsymbol{\epsilon}_i|\boldsymbol{a}_i$), ***directly***.

In subsequent chapters, we consider these similarities and differences in ***excruciating detail***.

Henceforth, we use the acronyms SS and PA to refer, respectively, to subject-specific and population-averaged approaches.

2.5 Models for correlation structure

We now review some popular models for correlation structure that are used routinely in modeling longitudinal data. Depending on their features, as we discuss in subsequent chapters, these structures are used in the ***subject-specific*** approach as models for ***separate within-*** and ***among-individual*** components of the overall pattern of correlation, or in the ***population-averaged*** approach directly as models for the ***overall*** pattern.

Write $\boldsymbol{\Gamma}_i(\alpha)$ to denote a $(n_i \times n_i)$ correlation matrix depending on a vector of correlation parameters α. For now, we suppress in this notation possible dependence of $\boldsymbol{\Gamma}_i(\alpha)$ on the times t_{ij} at which i is observed. We also suppress dependence on within- and among-individual covariates.

As we demonstrate shortly and in more detail in subsequent chapters, an associated $(n_i \times n_i)$ ***covariance matrix*** with correlation structure dictated by $\boldsymbol{\Gamma}_i(\alpha)$ can be obtained by pre- and post-multiplying $\boldsymbol{\Gamma}_i(\alpha)$ by a $(n_i \times n_i)$ ***diagonal matrix*** whose diagonal elements are the ***standard deviations*** corresponding to the n_i components of the random vector (e.g., $\boldsymbol{e}_i$) being modeled.

UNSTRUCTURED CORRELATION MODEL: The most general structure is one that makes ***no assumptions*** about the pattern of association. In particular, the matrix

$$\boldsymbol{\Gamma}_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha_{12} & \alpha_{13} & \cdots & \alpha_{1n_i} \\ \alpha_{21} & 1 & \alpha_{23} & \cdots & \alpha_{2n_i} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \alpha_{n_i1} & \alpha_{n_i2} & \cdots & \alpha_{n_i,n_i-1} & 1 \end{pmatrix}, \quad -1 \leq \alpha_{jj'} \leq 1, \tag{2.24}$$

where of course $\alpha_{jj'} = \alpha_{j'j}$ for all j, j', allows the correlation between any pair of observations to be different. Thus, this matrix depends on $n_i(n_i - 1)/2$ arbitrary correlation parameters. This model is usually referred to as ***unstructured*** for obvious reasons.

This is not a very parsimonious model and moreover does not take into account the way in which the data were collected. For example, in modeling ***within-individual*** correlation due to time-ordered data collection in a SS model, we expect that correlations between observations far apart in time might be less strong than those close together in time. The model (2.24) does not impose any such restriction, but rather allows the correlations to be "anything."

As a model for the ***overall pattern of correlation*** in the PA setting, (2.24) might be plausible, as the aggregate of correlation from ***both*** within- and among-individuals sources might well result in a "***haphazard***" rather than ***systematic*** pattern of association. Even here, however, the issue of parsimony is relevant; it may well be that a simpler model with fewer parameters can do an adequate job capturing the predominant features of the overall pattern of correlation.

Thus, it is standard in both SS and PA settings to use models that attempt to represent correlation in terms of a ***small number*** (maybe one or two) of parameters.

EXCHANGEABLE OR COMPOUND SYMMETRIC MODEL: The ***exchangeable*** or ***compound symmetric*** model, is given by

$$\boldsymbol{\Gamma}_i(\alpha) = \begin{pmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \cdots & \alpha & 1 \end{pmatrix} = (1 - \alpha)\boldsymbol{I}_{n_i} + \alpha \boldsymbol{J}_{n_i}, \tag{2.25}$$

where $\boldsymbol{I}_n$ is a $(n \times n)$ identity matrix and $\boldsymbol{J}_n$ is a $(n \times n)$ matrix with 1 in every position. In many settings where this model is used, $\alpha \geq 0$; however, α can be negative and still result in a valid covariance structure.

This model is generally not an appropriate model for ***within-individual*** correlation due to time-ordered data collection in a SS model, for which correlations that "***damp out***" over time would be expected.

As we discuss in subsequent chapters, this model is often used in the PA setting to represent the overall, aggregate pattern of correlation in the case of ***clustered*** data, where there is ***no natural ordering*** to the observations within a response vector, as would be the case with repeated observations on the pups in litter born to a pregnant rat as in Chapter 1.

As we show in Chapter 3, the compound symmetric correlation structure is ***induced*** by classical models underlying a ***repeated measures analysis of variance*** approach as a representation of the overall pattern of correlation. As we discuss in subsequent chapters, this structure ***may or may not*** be a good representation of the overall, aggregate pattern. It can be a plausible model when ***among-individual*** sources of correlation ***dominate within-individual*** sources, which is often the case in practice.

This model is certainly ***parsimonious***, as it depends on only a single, scalar parameter α.

Many of the models that are used for modeling ***both*** within-individual correlation in SS models and overall correlation in PA models have their roots in ***time series analysis***. For within-individual correlation due to ***time-ordered*** data collection in a SS model, such correlation models are a ***natural*** choice. These models may or may not be reasonable for representing the ***overall, aggregate pattern of correlation*** in a PA approach; this would be the case when ***within-individual sources*** are predominant.

We review some popular correlation models from standard time series analysis. As basic time series analysis is predicated on the observations being ***equally-spaced*** in time, the first two models we discuss are appropriate only in situations where the observation times are approximately ***equidistant***.

ONE-DEPENDENT MODEL: This model may be thought of as representing the situation where observations close in time may be correlated, but correlation among those farther apart is ***negligible***. For equally-spaced data, one could imagine that observations ***adjacent*** in time might have non-negligible correlation, while the correlation between those more than one interval apart might be reasonably thought to have "damped out."

This situation is represented by the general ***one-dependent*** model

$$\boldsymbol{\Gamma}_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha_1 & 0 & \cdots & 0 \\ \alpha_1 & 1 & \alpha_2 & \cdots & 0 \\ 0 & \alpha_2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \alpha_{n_i-1} & 1 \end{pmatrix}, \tag{2.26}$$

where $0 \leq \alpha_j \leq 1$ for $j = 1, \ldots, n_i - 1$ represent the correlations between adjacent observations at times t_{ij} and $t_{i,j+1}$, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{n_i-1})^T$. Such a matrix is also referred to as a ***banded Toeplitz*** matrix.

A special case is where $\alpha_j \equiv \alpha$ for all j, resulting in the model

$$\boldsymbol{\Gamma}_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha & 0 & \cdots & 0 \\ \alpha & 1 & \alpha & \cdots & 0 \\ 0 & \alpha & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \alpha & 1 \end{pmatrix}. \tag{2.27}$$

The models in (2.26) and (2.27) can be extended to two- and higher-order dependency in the obvious way.

AUTOREGRESSIVE MODEL OF ORDER 1: The AR(1) model assumes that the correlations among observations farther apart in time ***decay*** to zero. For ***equally-spaced*** responses, this decay happens according to the number of time intervals separating two observations. In particular, if t_{ij} and $t_{i,j+1}$ are the times at which Y_{ij} and $Y_{i,j+1}$ are observed, then we have that the time interval $|t_{i,j+1} - t_{ij}|$ is a constant for all j. The model is

$$\boldsymbol{\Gamma}_i(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{n_i-1} \\ \alpha & 1 & \alpha & \alpha^2 & \cdots \\ \alpha^2 & \alpha & 1 & \alpha & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha^{n_i-1} & \cdots & \alpha^2 & \alpha & 1 \end{pmatrix}, \tag{2.28}$$

where $0 \leq \alpha \leq 1$. Clearly, as observations become farther apart, so that the number of time intervals d increases, α^d approaches 0 rather quickly. This model depends only on the scalar correlation parameter α, so may be a parsimonious representation when this approximate pattern of decay is thought to hold.

A correlation structure such as the AR(1) (2.28) embodies the property of ***stationarity*** in that correlation depends only on the ***time distance*** between two observations and not on the actual observations times themselves (except through their difference). Stationarity may or may not be a reasonable assumption, but its appeal is obvious. If correlation depends only on distance and not on the actual time points, pairs of observations at different time points have information about the ***entire*** correlation structure.

A problem with models such as (2.28) is that it is often not the case that observations are ***equally-spaced***. In situations where observations are taken over time, whether or not the responses are equally-spaced is often dictated by the ***application***. In many prospective epidemiological studies, for example, observations are taken at regular intervals for the ***convenience*** of participants; the same is often true in clinical trials where data are collected longitudinally on participants over time.

However, in many long term studies, observations are taken ***more frequently early on*** ; e.g., in the days and weeks following administration of antiretroviral therapy in HIV infection, so that the swift effect of these drugs in suppressing the virus can be observed. Once patients have reached a "***plateau*** ," additional observations are taken at much longer intervals.

Likewise, in pharmacokinetic studies , it is traditional and sensible to make unequally-spaced observations. For drugs that are administered orally, as for the pharmacokinetic study in ***EXAMPLE 4*** of Section 1.2, it is routine to take more frequent blood samples during the "***absorption phase***" where drug concentration is increasing rather quickly in hopes of capturing the nature of the absorption pattern. In the "***elimination phase***" where concentrations are decaying more slowly, fewer samples need be taken.

A model like (2.28) would ***not*** be appropriate for representing possible within-individual correlation in these applications.

Generalizations of models like the AR(1) to the case of ***unequally-spaced*** responses are available. These models continue to embody the property of ***stationarity*** , so that the correlation between Y_{ij} and $Y_{ij'}$ depends only on the distance $|t_{ij} - t_{ij'}|$ but not on the particular values t_{ij} and $t_{ij'}$.

Such models can be described in terms of the associated ***autocorrelation function***. Generically, for a process $Y(t)$ and time points t and s, the autocorrelation function $\rho(\cdot)$ is

$$\text{corr}\{Y(t), Y(s)\} = \rho(|t - s|). \tag{2.29}$$

EXPONENTIAL CORRELATION MODEL: The ***exponential*** correlation model is represented by the autocorrelation function

$$\rho(u) = \exp(-\alpha u), \quad \alpha > 0. \tag{2.30}$$

(2.30) yields

$$\text{corr}(Y_{ij}, Y_{ij'}) = \exp(-\alpha|t_{ij} - t_{ij'}|).$$

This results in the correlation matrix

$$\mathbf{\Gamma}_i(\alpha) = \begin{pmatrix} 1 & \alpha_*^{|t_{i1}-t_{i2}|} & \alpha_*^{|t_{i1}-t_{i3}|} & \cdots & \alpha_*^{|t)i1-t_{in_i-1}|} \\ & 1 & \alpha_*^{|t_{i2}-t_{i3}|} & \cdots & \vdots \\ & & \ddots & \vdots & \vdots \\ & & & 1 & \alpha_*^{|t_{in_i-1}-t_{in_i-2}|} \\ & & & & 1 \end{pmatrix}, \tag{2.31}$$

where $\alpha_* = \exp(-\alpha)$.

In the case of equally-spaced time points, (2.31) reduces to the AR(1) correlation structure (2.28). Thus, the exponential correlation model is often viewed as a ***generalization*** of the AR(1) to unequally-spaced observation times.

GAUSSIAN CORRELATION MODEL: An alternative to (2.30) is the so-called ***Gaussian*** correlation model

$$\rho(u) = \exp(-\alpha u^2), \quad \alpha > 0, \tag{2.32}$$

which yields

$$\text{corr}(Y_{ij}, Y_{ij'}) = \exp\{-\alpha(t_{ij} - t_{ij})^2\}.$$

More extensive discussion of these models is given in Chapters 3–5 of Diggle, Heagerty, Liang, and Zeger (2002). The above account of various models is by no means exhaustive; rather, we have reviewed only some of the more popular representative models that are used to model either pure within-individual serial correlation (in the SS case) or are chosen as empirical approximations to model the overall pattern of correlation in the PA case. In both situations, the hope is that such models may do a reasonable job at capturing the ***salient features*** of associations among observations with only a low-dimensional parameter α that usually must be ***estimated***.

2.6 Exploring mean and correlation structure

Just as there are procedures available that can aid the analyst in ***assessing assumptions*** , such as that of constant variance, in ordinary regression analysis, there are methods, ad hoc and otherwise, that can be used to evaluate and suggest models for patterns of correlation from both ***PA*** and ***SS*** perspectives and for overall (PA) and inherent (SS) mean response. The methods we now discuss are most relevant in the case of ***continuous*** outcome.

These methods are particularly straightforward when the data are ***balanced*** ; that is, responses are ascertained on ***all*** m individuals at the ***same*** time points, with no departures from these times or missing values for an individuals. We demonstrate using the dental study data of ***EXAMPLE 1***. Some of these techniques can be extended to ***unbalanced*** situations.

MEAN RESPONSE: For continuous response, ***spaghetti plots*** are a natural first step toward assessing the form of mean response. How these plots are constructed is guided by the scientific questions of interest. For example, for the dental study, questions concern differences between boys and girls, so it is natural to create ***separate plots*** for each gender, as in Figure 2.3.

- Under a SS perspective, where the ***inherent mean response trend*** for individuals is modeled, inspection of spaghetti plots yields insight into possible models, which, as in

  $$Y_{ij} = \beta_{0i} + \beta_{0i} t_{ij} + e_{ij}$$

  as in (2.13) suggested by Figure 2.3, depend on ***individual-specific parameters*** (intercept and slope here).

- From a PA perspective, where interest focuses on ***overall population mean response*** , plotting the ***sample means*** at each time point as in Figure 2.3 is obvious. Because interest is in how mean response ***differs*** by genders, plotting separately by gender is natural.

  Note that this is straightforward for ***balanced*** data, but more complicated otherwise. In situations where different individuals are observed at different time points (with possibly different n_i), so that there is a large number of ***distinct observation times*** across individuals, one strategy is to overlay a ***nonparametric smooth estimate*** , e.g., a scatter plot smoother using ***locally-weighted polynomial regression*** (a ***lowess curve***), where the estimate is based on treating all $N = \sum_{i=1}^{m} n_i$ observations from all individuals as independent.

NOTATION: Before we discuss considerations for assessing variation and correlation, we define notation used here and in subsequent chapters.

- We use the symbol $\boldsymbol{V}_i$ to denote an ***overall population covariance matrix*** corresponding to individual i. More precisely, for $\boldsymbol{Y}_i$ ($n_i \times 1$), we use $\boldsymbol{V}_i$ to represent var($\boldsymbol{Y}_i|\boldsymbol{x}_i$), or equivalently var($\epsilon_i|\boldsymbol{x}_i$), so that it depends potentially on ***within-individual covariates*** $\boldsymbol{u}_i$, ***among-individual covariates*** $\boldsymbol{a}_i$, and the observation times, as well as additional parameters.
- We use the symbol $\boldsymbol{R}_i$ to denote a covariance matrix associated with ***within-individual sources*** of variation and correlation associated with individual i; i.e., that of $\boldsymbol{e}_i$. This can depend on ***within-individual covariates*** $\boldsymbol{u}_i$, the observations times, and ***individual-specific parameters*** like β_{0i}, β_{1i} in (2.13) for the dental data.
- We are ***more precise*** about the nature of $\boldsymbol{V}_i$ and $\boldsymbol{R}_i$ in subsequent chapters.
- We use the symbol $\boldsymbol{\Gamma}_i$, with appropriate dependence on covariates and parameters, to denote an associated correlation matrix of either type, which should be clear from the context.

OVERALL PATTERN OF COVARIANCE AND CORRELATION: If taking a ***PA perspective*** is appropriate (which, recall, is dictated by the questions of scientific interest), then insight into the nature of the ***overall, aggregate pattern*** of variation and correlation is relevant.

For individual i, the overall pattern is embodied in the covariance matrix of $\boldsymbol{Y}_i$, or, equivalently, ϵ_i (conditional on covariates). If the individuals are drawn from a single population of individuals and the data are ***balanced*** , then it is natural to suppose that this matrix is the ***same*** for all i.

In the dental study, we identify ***two populations*** , those of boys and girls, indicated by the ***among-individual covariate*** gender, $\boldsymbol{a}_i = g_i$, and we can allow the possibility that the covariance matrices for these two populations are ***different***; that is, var($\boldsymbol{Y}_i|\boldsymbol{a}_i$) depends on $\boldsymbol{a}_i$. For simplicity, denote the covariance matrix conditional on g_i when $g_i = 0$ (girls) as $\boldsymbol{V}_G$ and when $g_i = 1$ (boys) as $\boldsymbol{V}_B$, with associated correlation matrices $\boldsymbol{\Gamma}_G$ and $\boldsymbol{\Gamma}_B$.

To gain insight into the form of these matrices, we can estimate them from the data. With balanced data, the most basic, straightforward estimators are ***sample covariance matrix*** and its associated ***sample correlation matrix***. For m individuals from the same population, recall that this estimator is

$$\widehat{\boldsymbol{V}} = (m-1)^{-1} \sum_{i=1}^{m} (\boldsymbol{Y}_i - \overline{\boldsymbol{Y}})(\boldsymbol{Y}_i - \overline{\boldsymbol{Y}})^T, \quad \overline{\boldsymbol{Y}} = m^{-1} \sum_{i=1}^{m} \boldsymbol{Y}_i.$$

Based on the 11 girls, these are

$$\widehat{\boldsymbol{V}}_G = \begin{pmatrix} 4.514 & 3.355 & 4.332 & 4.357 \\ 3.355 & 3.618 & 4.027 & 4.077 \\ 4.332 & 4.027 & 5.591 & 5.466 \\ 4.357 & 4.077 & 5.466 & 5.941 \end{pmatrix}, \quad \widehat{\boldsymbol{\Gamma}}_G = \begin{pmatrix} 1.000 & 0.830 & 0.862 & 0.841 \\ 0.830 & 1.000 & 0.895 & 0.879 \\ 0.862 & 0.895 & 1.000 & 0.948 \\ 0.841 & 0.879 & 0.948 & 1.000 \end{pmatrix}; \qquad (2.33)$$

based on the 16 boys,

$$\widehat{\boldsymbol{V}}_B = \begin{pmatrix} 6.017 & 2.292 & 3.629 & 1.613 \\ 2.292 & 4.563 & 2.194 & 2.810 \\ 3.629 & 2.194 & 7.032 & 3.241 \\ 1.613 & 2.810 & 3.241 & 4.349 \end{pmatrix}, \quad \widehat{\boldsymbol{\Gamma}}_B = \begin{pmatrix} 1.000 & 0.437 & 0.558 & 0.315 \\ 0.437 & 1.000 & 0.387 & 0.631 \\ 0.558 & 0.387 & 1.000 & 0.586 \\ 0.315 & 0.631 & 0.586 & 1.000 \end{pmatrix}. \qquad (2.34)$$

- The ***diagonal elements*** of $\widehat{\boldsymbol{V}}_G$ and $\widehat{\boldsymbol{V}}_B$ in (2.33) and (2.34) are estimates of the ***overall population variances*** at each time point in the populations of girls and boys. These are based on small numbers of observations (11 and 16), so the estimators yielding these numerical results are rather imprecise, and the estimates should not be over-interpreted.

  The variances are in the same "ballpark" over time for each gender, so it may might be reasonable to assume that the overall variance is ***constant across time*** for each.

  The variances for boys are mostly ***larger*** than those for girls, suggesting that it might be inappropriate to assume that variance is the ***same*** for each gender. From Figure 2.3, the "large" estimated variance at age 8 may in part reflect the one very "low" dental distance at that age.

- Inspection of $\widehat{\boldsymbol{\Gamma}}_G$ in (2.33) shows that the estimated correlations are ***similar*** for all pairs of ages, with no "damping out" over time. The pattern is reminiscent of ***compound symmetry*** as in (2.25).

  The estimate for boys, $\widehat{\boldsymbol{\Gamma}}_B$ in (2.34) shows roughly a similar pattern, although the values are more disparate and in general ***smaller*** than those for girls.

  These observations suggest that assuming that the ***overall correlation structure*** is ***compound symmetric*** for each population may be a reasonable approximation. However, whether or not the correlation parameter α in (2.25) is reasonably assumed to be the ***same*** for both genders is questionable.

- Under the assumption that $\boldsymbol{V}_G$ and $\boldsymbol{V}_B$, and thus $\boldsymbol{\Gamma}_G$ and $\boldsymbol{\Gamma}_B$, are the ***same*** (which seems shaky here), the common $\boldsymbol{V}$ can be estimated by the ***pooled sample covariance matrix*** and its associated correlation matrix.

Generically, if we can identify g groups ($g = 2$ here), and there are r_ℓ individuals in the data set from group ℓ, $\ell = 1, \ldots, g$, then letting $\widehat{\boldsymbol{V}}_\ell$ be the sample covariance matrix for group ℓ, the ***pooled estimator*** for the assumed common covariance matrix $\boldsymbol{V}$ is

$$\widehat{\boldsymbol{V}}_{POOLED} = (m-g)^{-1}\{(r_1-1)\widehat{\boldsymbol{V}}_1 + \cdots + (r_g-1)\widehat{\boldsymbol{V}}_g\}. \tag{2.35}$$

Although the evidence is ***not convincing*** in favor of a common overall pattern, we show the pooled sample covariance matrix and its associated correlation matrix:

$$\widehat{\boldsymbol{V}}_{POOLED} = \begin{pmatrix} 5.415 & 2.717 & 3.910 & 2.710 \\ 2.717 & 4.185 & 2.927 & 3.317 \\ 3.910 & 2.927 & 6.456 & 4.131 \\ 2.710 & 3.317 & 4.131 & 4.986 \end{pmatrix}$$

and

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

These appear to be a "***compromise***" between the estimates for girls and boys.

SCATTERPLOT MATRICES: A useful supplement to numerical estimates is a graphical display known as a ***scatterplot matrix***, which depicts associations among responses at different time points. This plot really only makes sense when all individuals are seen at the ***same*** time points.

- To achieve a visual impression that is not distorted by differences in mean and variance at each time point, this plot is based on ***centered*** and ***scaled*** observations.

  That is, for times t_j and t_k (the same for all i), letting $\overline{Y}_j$ and $\overline{Y}_k$ be the ***sample mean responses*** over all individuals at t_j and t_k and s_j and s_k be the associated ***sample standard deviations*** (square roots of jth and kth ***diagonal elements*** of the sample covariance matrix), plot the pairs

  $$\left(\frac{Y_{ij}-\overline{Y}_j}{s_j}, \frac{Y_{ik}-\overline{Y}_k}{s_k}\right)$$

  for each pair (j,k), $j \neq k$.

- Figure 2.4 shows the scatterplot matrix for girls in the dental study and is self-explanatory. The apparent association among responses at different time points appears strong and ***positive*** for each pair of time points and is fairly ***similar regardless*** of separation in time. These observations coincide with the numerical summary in $\widehat{\boldsymbol{\Gamma}}_G$.

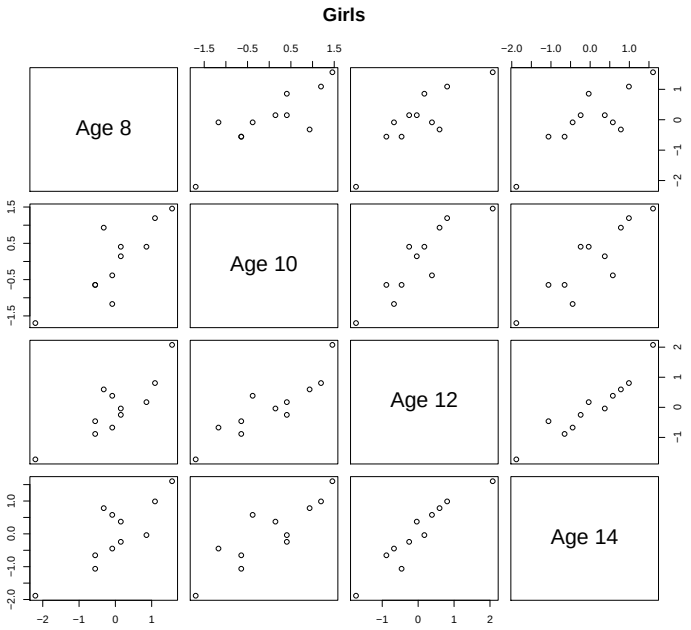

Figure 2.4: *Scatterplot matrix for the girls in the dental study.*

- Such a visual display offers the analyst further assistance in identifying ***systematic features*** in the apparent pattern of correlation that can suggest an appropriate ***correlation model***.

AUTOCORRELATION FUNCTION AND ASSOCIATED PLOTS: If it is believed that ***within-individual sources*** play a dominant role in dictating the overall, aggregate pattern of correlation, as noted in Section 2.5, the analyst may wish to consider correlation models that emphasize the ***time-ordered*** data collection. If the assumption of ***stationarity*** is reasonable, then additional diagnostic tools borrowed from the areas of ***time series analysis*** and ***spatial statistics*** are used.

For ***balanced data*** where furthermore the time points are ***equally-spaced*** , assuming stationarity, the ***autocorrelation function*** corresponding to the overall pattern of correlation can be estimated as follows. Denote the ***lag*** between two time points as u; in our situation, the lag is the number of (***equidistant***) time intervals that can separate two observations. Thus, if there are n time points, the total number of possible lags is $n - 1$. For the dental study, $n = 4$, the time interval between observations is 2 years, and there are 3 possible lags: a lag of 1 corresponds to 2 years, lag 2 to 4 years, and lag 3 to 6 years.

For the dental study, then, with no within-individual covariates, from (2.29), we can consider the ***autocorrelation function*** for each gender, which we write for all j as

$$\rho_G(u) = \text{corr}(Y_{ij}, Y_{i,j+u}|g_i = 0)$$

for girls and

$$\rho_B(u) = \text{corr}(Y_{ij}, Y_{i,j+u}|g_i = 1)$$

for boys, where, for the dental study $u = 1, 2, 3$. Because of stationarity, these functions depend only on u and is the same for all relevant j.

For given u, it is then natural to estimate $\rho_G(u)$ and $\rho_B(u)$ by the ***sample correlation*** between all pairs of observations u intervals apart for girls and for boys. To account for different means and variances at each time point, the estimator is based on ***centered*** and ***scaled*** responses. Specifically the estimator $\widehat{\rho}_G(u)$ for given u is the sample correlation coefficient among all pairs

$$\left(\frac{Y_{ij} - \overline{Y}_j}{s_j}, \frac{Y_{i,j+u} - \overline{Y}_{j+u}}{s_{j+u}} \right)$$

for all i and j for girls, treating these as if there were all independent pairs of observations on two random variables, and similarly for boys.

The estimate is often plotted against u to provide a visual impression of the ***decay*** in correlation as the time interval increases.

- For the girls in the dental study, we have

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

  where the estimates at lags $u = 1$, 2, and 3 are based on 33, 22, and 11 pairs, respectively.

  The estimates are ***relatively constant***, which is consistent with the evidence from the sample correlation matrix and scatterplot in Figure 2.4.

- As a ***visual supplement*** to these calculations, it is also customary to ***plot*** the lagged values against each other for each u; this is shown for the girls in the dental study in Figure 2.5

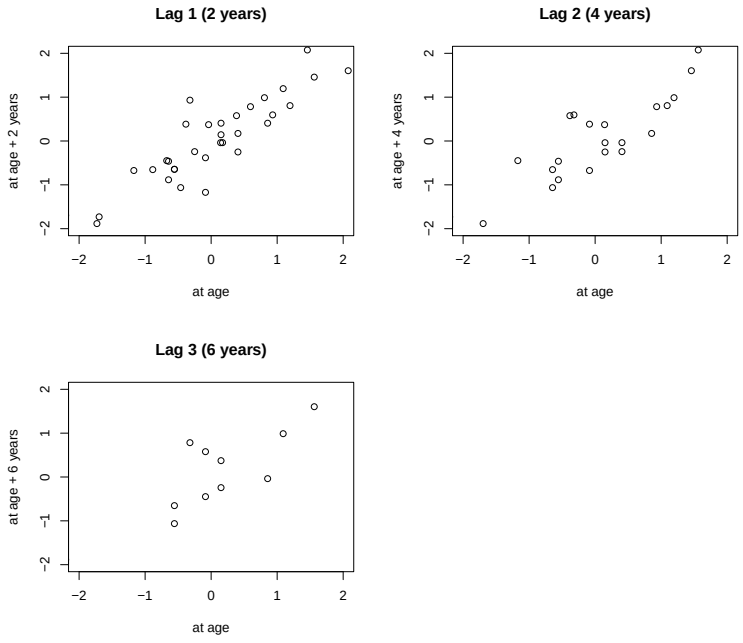

Figure 2.5: *Lag plots for the girls in the dental study.*

WITHIN-INDIVIDUAL PATTERN OF COVARIANCE AND CORRELATION: If taking a ***SS perspective*** is appropriate, then gaining insight into the nature of variation and correlation arising from ***within-individual sources*** , represented in our conceptual framework as var($\boldsymbol{e}_i|\boldsymbol{a}_i$), is important, as this will be modeled ***explicitly***. As we demonstrate in subsequent chapters, modeling the ***among-individual component*** var($\mathcal{B}_i|\boldsymbol{a}_i$) is more straightforward.

The within-individual component is potentially dictated by the ***time-ordered*** data collection; thus, it is not surprising that the ***same tools*** discussed above are relevant. Here, however, the focus is now on variation and correlation that comes about as the result of ***deviations*** from the ***inherent, individual-specific mean response***.

Accordingly, from the point of view of the conceptual representation

$$Y_{ij} = \mu_i(t_{ij}) + e_{ij},$$

we are interested in the autocorrelation function of the e_{ij}, conditional on covariates,

$$\rho(u) = \text{corr}(e_{ij}, e_{i,j+u}|\boldsymbol{a}_i).$$

Analogous to the above, the estimator $\widehat{\rho}(u)$ for given u is the ***sample correlation coefficient*** among all pairs

$$\left(\frac{Y_{ij} - \widehat{\mu}_i(t_{ij})}{\widehat{\sigma}_{ij}}, \frac{Y_{i,j+u} - \widehat{\mu}_i(t_{i,j+u})}{\widehat{\sigma}_{i,j+u}} \right)$$

for individuals i sharing a common $\boldsymbol{a}_i$ value, where $\widehat{\mu}_i(t)$ is an estimator for individual i's ***individual-specific mean response*** at time t, and $\widehat{\sigma}_{ij}$ is an estimator for the standard deviation of e_{ij}.

For the dental study, under the subject-specific model (2.13),

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij},$$

$\mu_i(t_{ij}) = \beta_{0i} + \beta_{1i} t_{ij}$, and a natural estimator for $\mu_i(t_{ij})$ is the ***predicted value*** from fitting this model to the responses on child i. We demonstrate for the boys ($g_i = 1$).

- It is natural to fit this individual-specific ***simple linear regression*** model via ***ordinary least squares*** to the data for each individual i. Figure 2.6 shows the ***residuals*** for all 16 boys; here, under the assumption that the variance of e_{ij} for boys is ***constant*** across j and the ***same*** for all boys, the residuals have been ***standardized*** by dividing by an estimate of this assumed constant variance obtained by ***pooling*** the residuals across boys. The plot suggests that the assumption of ***constant variance over time*** that is similar for all boys is reasonable.

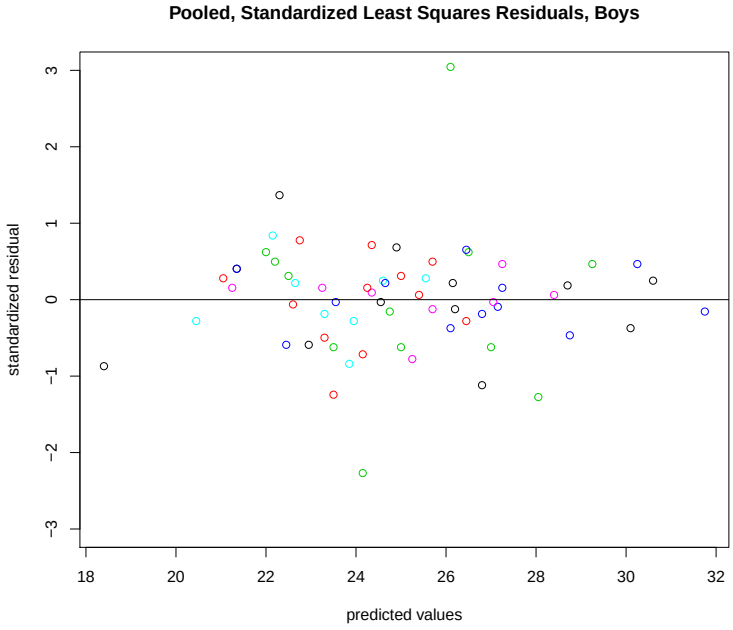

Figure 2.6: *Pooled residual plot for boys in the dental study. Each boy's residuals are displayed using a different color.*

- Figure 2.7 shows the ***lag plot*** based on the standardized residuals; the corresponding estimated autocorrelation function is

u	1	2	3
$\widehat{\rho}(u)$	-0.685	0.144	0.290

- From the plot, the rather large negative correlation at lag 1 appears to be driven strongly by one outlying pair of observations. Otherwise, at lags 2 and 3, the depicted lagged relationships appear relatively ***flat***, consistent with the numerical estimates. Of course, we do not have standard errors against which to calibrate the estimated values. However, with the exception of the outlying observation, the visual evidence and these point estimates do not offer compelling evidence of a pattern of strong correlation that decays over time.

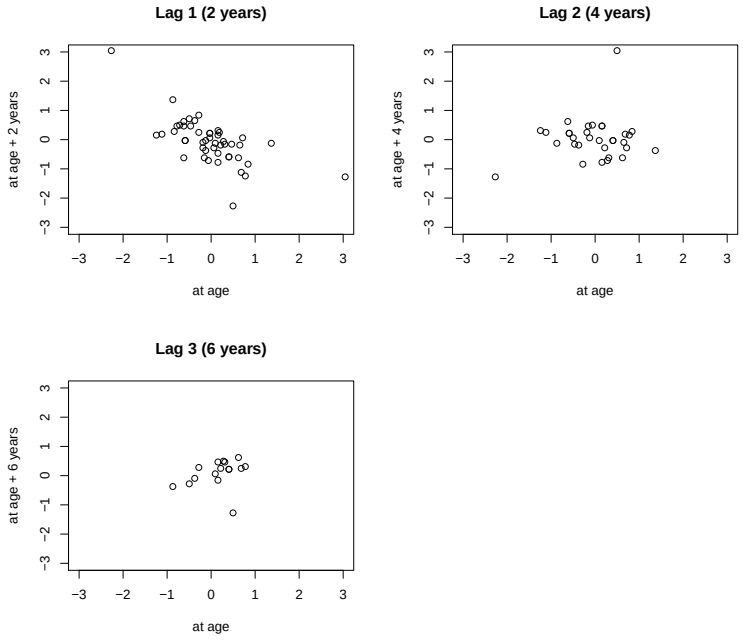

Figure 2.7: *Lag plots for within-individual correlation for the boys in the dental study.*

The foregoing calculations required that the number of responses on each individual is ***sufficiently large*** to allow the individual-specific regression models to be fitted based on each individual's data only. When some or all of the n_i are not large enough to facilitate fitting of individual-specific models, clearly the foregoing methods are not feasible. We discuss other strategies in subsequent chapters.

More generally, it is not always the case that the observation times are ***equally spaced***. In this situation, if ***stationarity*** is still a plausible assumption, the autocorrelation function can be replaced by the ***variogram***. The generic definition of the variogram for a stochastic process $Z(t)$ is

$$\gamma(u) = \frac{1}{2}E\left[\{Z(t) - Z(t-u)\}^2\right], \quad u \geq 0.$$

Under stationarity, the variogram is related to the autocorrelation function by the relationship $\gamma(u) = \text{var}\{Z(t)\}\{1 - \rho(u)\}$.

In our context, the variogram can be estimated from the standardized residuals

$$wr_{ij} = \frac{Y_{ij} - \widehat{\mu}_i(t_{ij})}{\widehat{\sigma}_{ij}}$$

by first computing $v_{ijk} = (1/2)(wr_{ij} - wr_{ik})^2$, and $u_{ijk} = t_{ij} - t_{ik}$ for all i, j, k. The (v_{ijk}, u_{ijk}) pairs for $j < k = 1, \ldots, n_i$ over all $i = 1, \ldots, m$ can be plotted and related back to the autocorrelation function.

In subsequent chapters, we discuss these and other approaches to assessing patterns of correlation in both PA and SS models.

2.7 Considerations for discrete response

Our discussion so far has been in the context of ***continuous response***. We now review general considerations for modeling continuous, repeated (multivariate) responses and then demonstrate some of the difficulties that arise when the responses are instead ***discrete*** ; e.g., binary, categorical, or in the form of counts. For simplicity of exposition, we suppress dependence on covariates.

Statistical models and associated methods for outcomes that can be viewed as continuous (or approximately so) are generally predicated on the assumption that the responses are approximately (or can be transformed to be) ***normally distributed***. In the context of modeling continuous repeated measurement data, this perspective underlies models and methods that we discuss in Chapters 3-6, which are based on the assumption that the responses vectors $\boldsymbol{Y}_i$ (conditional on covariate information) follow approximately a ***multivariate*** (n_i-variate) ***normal distribution*** when viewed from a PA perspective.

MULTIVARIATE NORMAL DISTRIBUTION: This probability distribution is the ***multivariate generalization*** of the familiar normal distribution that is widely used as a statistical model for ***scalar*** independent continuous responses. As is well known, the ***marginal*** distributions of the multivariate normal are themselves normal. Thus, this distribution is a natural framework for modeling repeated continuous responses, each of which is reasonably assumed to be approximately normally distributed.

The ***multivariate normal distribution*** has a number of properties that make it an attractive modeling framework.

- The form of the ***density*** of the multivariate normal is a ***straightforward generalization*** of that of the univariate normal. The normal distribution has the desirable property that it is ***fully characterized*** by its ***first two moments***. That is, specification of a ***mean*** and ***variance*** is sufficient to specify a normal distribution. Moreover, the mean and variance ***need not be related*** in any way.

- The multivariate normal ***shares*** this property in a generalized form: the multivariate normal is again ***fully characterized*** by its first two moments, its ***mean vector*** and ***covariance matrix***.

- Thus, for example, when the analyst is ***positing models*** for, say, ***population mean response*** and ***overall covariance structure*** in a PA modeling framework, s/he can consider each ***separately*** without concern that choice of a particular mean and particular covariance structure would ***violate*** some property of the normal distribution.

- More precisely, positing any model for population mean, so any $\boldsymbol{\mu}_i$ in the PA conceptual representation

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

  as in (2.23), and positing any model $\boldsymbol{V}_i$ for covariance structure (i.e., for var($\boldsymbol{Y}_i$) = var($\boldsymbol{\epsilon}_i$), continuing to suppress conditioning on covariates) does not violate any restrictions imposed by this distribution.

  To see this, let μ_{ij} and σ_{ij} be the mean and variance for Y_{ij} implied by these modeling choices. Then the ***correlation*** between two elements of $\boldsymbol{Y}_i$ is

  $$\text{corr}(Y_{ij}, Y_{ij'}) = \alpha_{jj'} = \frac{E(Y_{ij}Y_{ij'}) - \mu_{ij}\mu_{ij'}}{\sigma_{ij}\sigma_{ij'}},$$

  so that $E(Y_{ij}Y_{ij'}) = \alpha_{jj'}\sigma_{ij}\sigma_{ij'} + \mu_{ij}\mu_{ij'}$. It is clear that taking any $-1 \leq \alpha_{jj'} \leq 1$ would not violate any characteristic of jointly normally distributed random variables.

Thus, the analyst can reasonably contemplate models for mean and covariance structure ***separately*** without concern that the resulting covariance structure violates some distributional requirement or results in a ***pathological*** distribution.

DISCRETE RESPONSE: The situation is much different for ***discrete*** outcome.

As we noted in Chapter 1, probability distributions that are natural models for discrete responses include the ***Poisson distribution*** for responses in the form of ***counts***, the ***Bernoulli distribution*** for ***binary response*** , and the ***multinomial distribution*** for ***categorical responses***. Unfortunately, unlike for the normal distribution, ***multivariate generalizations*** of these distributions are ***not*** so straightforward.

A main issue is as follows. Because the multivariate normal distribution is fully characterized by the mean and covariance matrix, and thus the ***associated correlation matrix*** and variances, the only correlations that one need be concerned with are ***pairwise correlations*** as shown above.

In contrast , a key distinguishing feature of multivariate versions of these discrete distributions is that their densities depend in a ***complicated way*** on terms representing ***third and higher moments*** of the response vector, and thus on what have been called ***three- and higher-way associations*** among the elements. This means that it is ***not possible*** to characterize a multivariate distribution by simply specifying a mean and covariance matrix. Moreover, computation of the probability density function itself can be difficult.

These features make modeling of discrete repeated measurements a significant challenge. Without a straightforward characterization of the density of a response vector, appealing to the principles of ***maximum likelihood*** , for example, is out of the question.

As we discuss in detail in Chapter 8 , this obstacle inspired approaches to modeling and analysis of discrete, repeated outcomes that ***are*** based on positing models for only the population mean and overall covariance structure of a response vector. However, there are caveats to this approach as well, as we now demonstrate.

Unlike in the case of the multivariate normal, where there are ***no restrictions*** on the nature of ***pairwise correlations*** between two elements of a response vector, discrete multivariate responses ***do involve*** rather complicated restrictions on these.

For definiteness, consider ***binary*** responses Y_{ij} and $Y_{ij'}$, both elements of $\boldsymbol{Y}_i$. Suppose we posit a model $\boldsymbol{\mu}_i$ for the population mean response, with elements μ_{ij}. Here, of course, μ_{ij} is a ***probability*** and as such is restricted to be between 0 and 1. Because Y_{ij} is binary, the ***variance*** of Y_{ij} is dictated to be $\mu_{ij}(1-\mu_{ij})$. We now consider the ***correlation*** between Y_{ij} and $Y_{ij'}$, which is given by, suppressing dependence on covariates,

$$\text{corr}(Y_{ij}, Y_{ij'}) = \alpha_{jj'} = \frac{E(Y_{ij}Y_{ij'}) - \mu_{ij}\mu_{ij'}}{\{\mu_{ij}\mu_{ij'}(1-\mu_{ij})(1-\mu_{ij'})\}^{1/2}},$$

where $E(Y_{ij}Y_{ij'}) = \text{pr}(Y_{ij} = 1 \text{ and } Y_{ij'} = 1)$.

Clearly,

$$\text{pr}(Y_{ij} = 1 \text{ and } Y_{ij'} = 1) \leq \text{pr}(Y_{ij} = 1) = \mu_{ij} \quad \text{and} \quad \leq \text{pr}(Y_{ij'} = 1) = \mu_{ij'}.$$

Thus, it must be that $E(Y_{ij}Y_{ij'}) \leq \min(\mu_{ij}, \mu_{ij'})$. Furthermore,

$$\begin{aligned}
\text{pr}(Y_{ij} = 1 \text{ and } Y_{ij'} = 1) &= 1 - \text{pr}(Y_{ij} = 0 \text{ or } Y_{ij'} = 0) \\
&\geq 1 - \{\text{pr}(Y_{ij} = 0) + \text{pr}(Y_{ij'} = 0)\} \\
&= 1 - \{(1-\mu_{ij}) + (1-\mu_{ij'})\} = \mu_{ij} + \mu_{ij'} - 1.
\end{aligned}$$

Thus, it can be deduced that $\max(0, \mu_{ij}+\mu_{ij'}-1) \leq E(Y_{ij}Y_{ij'})$, which follows from noting that the events $(Y_{ij} = 1)$ and $(Y_{ij'} = 1)$ are either disjoint or not.

We thus have, combining the above, that

$$\max(0, \mu_{ij} + \mu_{ij'} - 1) \leq E(Y_{ij}Y_{ij'}) \leq \min(\mu_{ij}, \mu_{ij'}).$$

Arbitrarily taking $\mu_{ij} \leq \mu_{ij'}$ without loss of generality, we thus have that

$$\text{corr}(Y_{ij}, Y_{ij'}) \leq \frac{\mu_{ij} - \mu_{ij}\mu_{ij'}}{\{\mu_{ij}\mu_{ij'}(1-\mu_{ij})(1-\mu_{ij'})\}^{1/2}} = \left\{\frac{\mu_{ij}(1-\mu_{ij})}{\mu_{ij'}(1-\mu_{ij'})}\right\}^{1/2};$$

that is, the ***largest*** this correlation can be is the square root of the ***odds ratio***.

Similarly, the ***smallest*** this correlation can be is, when $\mu_{ij} + \mu_{ij'} \leq 1$,

$$\text{corr}(Y_{ij}, Y_{ij'}) \geq \frac{-\mu_{ij}\mu_{ij'}}{\{\mu_{ij}\mu_{ij'}(1-\mu_{ij})(1-\mu_{ij'})\}^{1/2}} = -\left\{\frac{\mu_{ij}\mu_{ij'}}{(1-\mu_{ij})(1-\mu_{ij'})}\right\}^{1/2}$$

or, when $\mu_{ij} + \mu_{ij'} \geq 1$,

$$\text{corr}(Y_{ij}, Y_{ij'}) \geq \frac{\mu_{ij} + \mu_{ij'} - 1 - \mu_{ij}\mu_{ij'}}{\{\mu_{ij}\mu_{ij'}(1-\mu_{ij})(1-\mu_{ij'})\}^{1/2}} = -\left\{\frac{(1-\mu_{ij})(1-\mu_{ij'})}{\mu_{ij}\mu_{ij'}}\right\}^{1/2}.$$

The result is that the fact that the data are ***binary*** imposes ***natural restrictions*** on the correlations that are possible between two binary random variables. The correlations must satisfy a constraint that depends on the means in a complicated way. Thus, in contrast to the situation of normal data, correlations cannot be "***anything***." In particular, here, assuming that the correlations are not dependent on the mean may be inappropriate. A similar phenomenon can be exhibited for other distributions, such as the Poisson.

These developments emphasize the challenges inherent in developing models and methods for analysis of ***discrete longitudinal responses***, which are not an issue for ***continuous response***. In Chapters 7-9, we discuss approaches to modeling of these data.

3 Repeated Measures Analysis of Variance

3.1 Introduction

As we have discussed, many approaches have been taken in the literature to specifying ***statistical models*** for longitudinal data. Within the framework of a specific model, the questions of scientific interest are interpreted and represented formally, and associated with the model are ***statistical methods*** that allow the questions to be addressed. Different models embodying different assumptions and taking different perspectives (e.g., SS vs. PA) lead to possibly different characterizations of the questions and different methods.

We begin our review of approaches by considering two statistical models that form the basis for what we have called ***classical methods***. These methods are appropriate for ***continuous outcome*** or outcomes that can be viewed as approximately continuous and that are reasonably thought to be approximately ***normally distributed***.

The models have several limitations relative to those underlying the more ***modern*** approaches that we discuss in subsequent chapters.

- The models are really only applicable in the case of ***balanced*** data; that is, where the responses on each individual are recorded at the ***same time points*** , with no departures from these times or missing data. Thus, in this chapter, we assume that each individual is observed at the same n time points $t_1, \ldots, t_j$, say, and each has an associated n-dimensional response vector, where the jth element corresponds to the response at time t_j.
- The models adopt a representation of the ***overall population mean*** of a response vector that is ***simplistic*** in the sense that it does not recognize the fact that the mean response might exhibit a ***systematic trajectory*** over ***continuous*** time, such as a ***straight line***. In particular, the mean ordinarily is represented using the notation that is commonplace when developing ***analysis of variance*** methods. Time and ***among-individual covariates*** are viewed as ***categorical factors*** with a small number of ***levels***. The treatment of time as a categorical factor is particularly restrictive.

- The models do not accommodate straightforwardly incorporation of ***covariate information*** beyond time and among-individual categorical factors. In our discussion, we restrict attention to a ***single*** among-individual factor such as ***group membership*** ; e.g., gender in the dental study or dose in the guinea pig diet study in ***EXAMPLE 2*** of Section 1.2. Although the models can be generalized to more than one such factor (e.g., genotype and weather pattern in the soybean growth study of ***EXAMPLE 3*** of Section 1.2), we do not consider this, deferring to the more modern approaches we discuss later that allow much greater flexibility.

- The associated analysis methods are focused almost entirely on ***hypothesis testing***. Thus, they do not readily accommodate questions regarding the nature of features of mean trajectories; e.g., in the dental study, the values of the ***slopes*** characterizing ***rate of change*** of assumed ***straight line*** population mean trajectories for boys and girls. That is, ***estimation*** of quantities like these is not straightforward within these modeling frameworks.

As we demonstrate, the models also embody assumptions on the ***overall covariance structure*** of a response vector that are possibly ***too restrictive*** or ***too general***.

- The statistical model underlying ***univariate repeated measures analysis of variance*** (ANOVA) methods is derived from a ***SS*** perspective. As we noted in Section 2.5, it induces a model for the overall covariance pattern that has a ***compound symmetric*** correlation structure, which may or may not be a plausible model.

- The statistical model underlying ***multivariate repeated measures analysis of variance*** methods arises from ***PA*** perspective. ***No specific systematic assumption*** is made on the overall covariance pattern, so that it is regarded as ***completely unstructured***. If correlation ***does*** exhibit a simpler pattern, these methods could be ***inefficient***.

We first review the ***univariate*** methods, followed by the ***multivariate*** approach. Our discussion is limited to the basic elements and thus is not meant to be ***comprehensive***. Rather, it is meant only to provide an appreciation of why statistical practice has moved toward favoring the modern methods discussed in subsequent chapters. Accordingly, we simply present results and do not offer detailed derivations or proofs.

3.2 Univariate repeated measures analysis of variance

We first discuss the basic model underlying univariate repeated measures ANOVA methods in the ***classical notation***. In particular, we present the usual "***one way***" model, where there is a ***single among-individual factor*** such as gender in the dental study of ***EXAMPLE 1*** of Section 1.2 or vitamin E dose level in the guinea pig growth study of ***EXAMPLE 2***.

BASIC SET-UP:

- The model assumes that the data arise from a study in which individuals are ***randomized*** or naturally belong to one of $g \geq 1$ groups; the group variable is often referred to as the ***between-*** or ***among-units*** factor. Thus, in the dental study, $g = 2$ genders; in the guinea pig growth study, $g = 3$ dose groups.

  From the point of view of the general notation introduced in Section 2.2, the model thus accommodates a ***single*** scalar, categorical ***among-individual covariate*** with g possible values. The model does not allow for ***within-individual covariates***.

- The response is recorded on each of n occasions or under each of n conditions. In a longitudinal study, this is usually "***time***" but could be another repeated measurement condition. E.g., if men are randomized into two groups, regular and modified diet, the response might be maximum heart rate after separate occasions during which each spent 10, 20, 30, 45, and 60 minutes walking on a treadmill. We use the generic term ***time***; this factor is often referred to in the classical literature as the ***within-units*** factor. In the dental study, this is age ($n = 4$); in the guinea pig study, weeks ($n = 6$).

  Thus, from the point of view of the ***conceptual framework*** discussed in Section 2.3, the model does not acknowledge explicitly that there is an underlying process in ***continuous time*** and that there could be values of the response at times ***other*** than these n occasions.

- As noted above, we consider only the case where there is a ***single group factor***. However, it is is straightforward to extend the development to the case where the groups are determined by a ***factorial design***; e.g. if in the guinea pig study there had been $g = 6$ groups, determined by the factorial arrangement of 3 doses and 2 genders.

NOTATION AND MODEL: We present the model first using the ***classical notation*** and then demonstrate how it can be expressed in the notation introduced in Chapter 2. Define

$$Y_{h\ell j} = \text{response on individual } h \text{ in the } \ell\text{th group at time } j.$$

- $h = 1, \ldots, r_\ell$, where r_ℓ denotes the number of units in group ℓ. Thus, in this notation, h indexes units ***within*** a particular group; $\ell = 1, \ldots, g$ indexes groups; and $j = 1, \ldots, n$ indexes the levels of time. Note then that a specific individual is ***uniquely*** identified by the indices (h, ℓ).
- The total number of individuals is $m = \sum_{\ell=1}^{g} r_\ell$. Each is observed at n time points.

The ***classical model*** for $Y_{h\ell j}$ is then given by

$$Y_{h\ell j} = \mu + \tau_\ell + b_{h\ell} + \gamma_j + (\tau\gamma)_{\ell j} + e_{h\ell j}. \tag{3.1}$$

In the usual terminology accompanying classical ANOVA methods

- μ is an "overall mean," τ_ℓ is the ***fixed deviation*** from the overall mean associated with being in group ℓ, γ_j is the ***fixed deviation*** associated with time j, and $(\tau\gamma)_{\ell j}$ is an ***additional fixed deviation*** associated with group ℓ and time j; that is, $(\tau\gamma)_{\ell j}$ is the ***interaction*** effect for group ℓ and time j.
- Thus, as we demonstrate explicitly below, the ***overall population mean response*** for the ℓth group at time j is represented as
$$\mu + \tau_\ell + \gamma_j + (\tau\gamma)_{\ell j}.$$
- $b_{h\ell}$ is a ***random effect*** assumed to be ***independent*** of the among-individual covariate group with conditional (on group) mean equal to the unconditional mean $E(b_{h\ell}) = 0$ characterizing how the "***inherent mean***" for the hth individual in group ℓ ***deviates*** from the ***overall population mean***. Thus, (3.1) represents the ***inherent (conditional) mean*** for individual (h, ℓ) as
$$\mu + \tau_\ell + \gamma_j + (\tau\gamma)_{\ell j} + b_{h\ell}, \tag{3.2}$$
and $b_{h\ell}$ characterizes ***among-individual*** behavior.
- $e_{h\ell j}$ is a ***within-individual*** deviation representing the net effect of realizations and measurement error, ***independent*** of the among-individual covariate group, with conditional (on group) mean equal to the unconditional mean $E(e_{h\ell j}) = 0$. This is often called the "***random error***," but as we have remarked previously we prefer the term ***within-individual deviation*** to reflect the fact that it embodies more than just measurement error.

Some observations are immediate.

- Model (3.1) has the ***same form*** as the statistical model for observations arising from an experiment conducted according to a ***split plot*** design. Thus, as we show shortly, the analysis is ***identical*** to that of a split plot experiment; however, the ***interpretation*** and further analyses are different.
- The ***actual values*** of the times (e.g. ages 8, 10, 12, 14 in the dental study) ***do not*** appear explicitly in the model. Rather, a separate deviation parameter γ_j and interaction parameter $(\tau\gamma)_{\ell j}$ is associated with each time. Thus, the model takes no account of where the times of observation are temporally; e.g. are they ***equally-spaced*** ?

Because (3.1) is a ***linear model*** , as discussed in Section 2.4, we can view it as a SS or PA model.

- From a SS perspective, as in (3.2),

  $$\mu + \tau_\ell + \gamma_j + (\tau\gamma)_{\ell j} + b_{h\ell}$$

  represents the inherent mean trend for the hth individual in group ℓ at time j. Note this assumes that the inherent mean for a given individual (h, ℓ) deviates from the overall population mean by the ***same amount***, $b_{h\ell}$, at each time j. Thus, this model implies that if an individual is "high" relative to the overall mean response at time j, the individual is "high" at all other times.

  This is often not a reasonable assumption. For example, consider the the conceptual representation in Figure 2.2. This assumption might be reasonable for the two uppermost individuals in panel (b), as the "inherent trends" for these are roughly parallel to the overall mean response trajectory. However, it is clearly not appropriate for the lowermost unit.

- Taking a PA perspective, write (3.1) as

  $$Y_{h\ell j} = \underbrace{\mu + \tau_\ell + \gamma_j + (\tau\gamma)_{\ell j}}_{\mu_{\ell j}} + \underbrace{b_{h\ell} + e_{h\ell j}}_{\epsilon_{h\ell j}}. \tag{3.3}$$

  In (3.3), $\epsilon_{h\ell j} = b_{h\ell} + e_{h\ell j}$ is the overall deviation reflecting aggregate deviation from this mean due to ***among-*** and ***within-individual*** sources.

Because $b_{h\ell}$ and $e_{h\ell j}$ have mean 0 (conditional on the among-individual covariate group and unconditionally), it follows that

$$E(Y_{h\ell j}) = \mu_{\ell j} = \mu + \tau_\ell + \gamma_j + (\tau\gamma)_{\ell j},$$

the overall population mean for the ℓth group at the *jth* time.

CONVENTION: Henceforth, we write $\mathcal{N}(\mu, \sigma^2)$ to denote a ***univariate normal distribution*** with mean μ and variance σ^2. We write $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{V})$ to denote a ***multivariate normal distribution*** with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$. The meaning (univariate or multivariate) is ordinarily clear from the context.

NORMALITY AND INDEPENDENCE ASSUMPTIONS: The model is completed by standard assumptions on the random deviations $b_{h\ell}$ and $e_{h\ell j}$, which lead to an assumption on the form of the overall pattern of variance and correlation.

- $b_{h\ell} \sim \mathcal{N}(0, \sigma_b^2)$ and are ***independent*** for all h and ℓ, so that where any individual "sits" in the population is unrelated to where others "sit." The fact that this normal distribution is ***identical*** for $\ell = 1, \ldots, g$ reflects the assumption that $b_{h\ell}$ is independent of the among-individual covariate group, so that ***among-individual variation*** is ***the same*** in all g populations. The ***variance component*** σ_b^2 represents the common magnitude of among-individual variation.

- $e_{h\ell j} \sim \mathcal{N}(0, \sigma_e^2)$ and are ***independent*** for all h, ℓ, and j. As for $b_{h\ell}$, that this normal distribution is the same for $\ell = 1, \ldots, 1$ follows from the assumption that the $e_{h\ell j}$ are independent of the among-individual covariate group. Moreover, it also reflects the assumption that ***within-individual variation*** is ***the same*** at all observation times. Independence across j also implies that ***within-individual correlation*** across the observation times is ***negligible***. The variance component σ_e^2 represents the magnitude of within-individual variation aggregated from all within-individual sources, namely, the realization process and measurement error.

- The $b_{h\ell}$ and $e_{h\ell j}$ are assumed to be ***mutually independent*** for all h, ℓ, and j. From the conceptual representation point of view, this says that deviations due to within-individual sources are of similar magnitude regardless of the magnitudes of the deviations $b_{h\ell}$ associated with the units on which the observations are made. This is often reasonable; however, as we will see later in the course, there are situations where it may not be reasonable.

VECTOR REPRESENTATION AND OVERALL COVARIANCE MATRIX: We can summarize the model for the responses for individual (h, ℓ) in the $(n \times 1)$ ***random vector***

$$\begin{pmatrix} Y_{h\ell 1} \\ Y_{h\ell 2} \\ \vdots \\ Y_{h\ell n} \end{pmatrix} = \begin{pmatrix} \mu + \tau_\ell + \gamma_1 + (\tau\gamma)_{\ell 1} \\ \mu + \tau_\ell + \gamma_2 + (\tau\gamma)_{\ell 2} \\ \vdots \\ \mu + \tau_\ell + \gamma_n + (\tau\gamma)_{\ell n} \end{pmatrix} + \begin{pmatrix} b_{h\ell} \\ b_{h\ell} \\ \vdots \\ b_{h\ell} \end{pmatrix} + \begin{pmatrix} e_{h\ell 1} \\ e_{h\ell 2} \\ \vdots \\ e_{h\ell n} \end{pmatrix} = \begin{pmatrix} \mu_{\ell 1} \\ \mu_{\ell 2} \\ \vdots \\ \mu_{\ell n} \end{pmatrix} + \begin{pmatrix} \epsilon_{h\ell 1} \\ \epsilon_{h\ell 2} \\ \vdots \\ \epsilon_{h\ell n} \end{pmatrix}. \tag{3.4}$$

With $\mathbf{1}$ a $(n \times 1)$ vector of 1s, we write (3.4) compactly as

$$\boldsymbol{Y}_{h\ell} = \boldsymbol{\mu}_\ell + \mathbf{1} b_{h\ell} + \boldsymbol{e}_{h\ell} = \boldsymbol{\mu}_\ell + \boldsymbol{\epsilon}_{h\ell}. \tag{3.5}$$

Under the foregoing assumptions, it is clear that each $Y_{h\ell j}$ is ***normally distributed*** with

$$E(Y_{h\ell j}) = \mu_{\ell j} = \mu + \tau_\ell + \gamma_j + (\tau\gamma)_{\ell j}, \quad \text{so that} \quad E(\boldsymbol{Y}_{h\ell}) = \boldsymbol{\mu}_\ell,$$

$$\text{var}(Y_{h\ell j}) = \text{var}(b_{h\ell}) + \text{var}(e_{h\ell j}) + 2\text{cov}(b_{h\ell}, e_{h\ell j}) = \sigma_b^2 + \sigma_e^2.$$

(conditionally and unconditionally). Moreover, it is straightforward to show (try it) that

$$\text{cov}(Y_{h\ell j}, Y_{h'\ell' j'}) = \text{cov}(\epsilon_{h\ell j}, \epsilon_{h'\ell' j'}) = 0, \quad h \neq h',$$

where $\ell \neq \ell'$ or $\ell = \ell'$ and $j \neq j'$ or $j = j'$; i.e., the covariance between observations from two different units from the same or different groups at the same or different times is zero, which implies under normality that $Y_{h\ell j}$ and $Y_{h'\ell' j'}$ are ***independent***.

Thus, under the assumptions of the model, for $\ell \neq \ell'$ or $\ell = \ell'$, the random vectors $\boldsymbol{Y}_{h\ell}$ and $\boldsymbol{Y}_{h'\ell'}$ are ***independent***, showing that the model ***automatically induces*** the usual assumption that data vectors from different individuals are independent.

It is also straightforward to derive that

$$\begin{aligned} \text{cov}(Y_{h\ell j}, Y_{h\ell j'}) = \text{cov}(\epsilon_{h\ell j}, \epsilon_{h\ell' j'}) &= E\{(Y_{h\ell j} - \mu_{\ell j})(Y_{h\ell j'} - \mu_{\ell j'})\} = E\{(b_{h\ell} + e_{h\ell j})(b_{h\ell} + e_{h\ell j'})\} \\ &= E(b_{h\ell} b_{h\ell}) + E(e_{h\ell j} b_{h\ell}) + E(b_{h\ell} e_{h\ell j'}) + E(e_{h\ell j} e_{h\ell j'}) = \sigma_b^2. \end{aligned}$$

Summarizing, we have that the m data vectors $\boldsymbol{Y}_{h\ell}$, $h = 1, \ldots, r_\ell$, $\ell = 1, \ldots, g$ are all independent and multivariate normal; that is, $\boldsymbol{Y}_{h\ell} \sim \mathcal{N}_n(\boldsymbol{\mu}_\ell, \boldsymbol{V})$, where

$$\boldsymbol{V} = \text{var}(\boldsymbol{Y}_{h\ell}) = \text{var}(\boldsymbol{\epsilon}_{h\ell}) = \begin{pmatrix} \sigma_b^2 + \sigma_e^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_e^2 & \cdots & \sigma_b^2 \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 + \sigma_e^2 \end{pmatrix}. \tag{3.6}$$

- This result follows directly from (3.5). Using the independence of $b_{h\ell}$ and $\boldsymbol{e}_{h\ell}$,

$$\text{var}(\boldsymbol{Y}_{h\ell}) = \text{var}(\boldsymbol{\epsilon}_{h\ell}) = \text{var}(\mathbf{1}b_{h\ell}) + \text{var}(\boldsymbol{e}_{h\ell}) = \text{var}(b_{h\ell})\mathbf{1}\mathbf{1}' + \text{var}(\boldsymbol{e}_{h\ell}),$$

  which, writing

$$\mathbf{1}\mathbf{1}' = \boldsymbol{J}_n = \begin{pmatrix} 1 & \cdots & 1 \\ 1 & \cdots & 1 \\ \vdots & \vdots & \vdots \\ 1 & \cdots & 1 \end{pmatrix} \text{ and } \text{var}(\boldsymbol{e}_{h\ell}) = \sigma_e^2 \boldsymbol{I}_n,$$

  yields

$$\text{var}(\boldsymbol{Y}_{h\ell}) = \text{var}(\boldsymbol{\epsilon}_{h\ell}) = \sigma_b^2 \boldsymbol{J}_n + \sigma_e^2 \boldsymbol{I}_n = \boldsymbol{V}, \tag{3.7}$$

  where (3.7) is a compact expression for (3.6).

- From (3.6),

$$\text{corr}(Y_{h\ell j}, Y_{h\ell j'}) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2}, \quad j \neq j'. \tag{3.8}$$

  Thus, the overall correlation between any two observations in $\boldsymbol{Y}_{h\ell}$ is ***the same*** and equal to (3.8). The quantity (3.8) is called the ***intraclass correlation*** in some contexts.

- These results show that the model assumes that the ***overall aggregate pattern of correlation*** is ***compound symmetric*** or ***exchangeable***. Note that, automatically, the correlation between any two observations in $\boldsymbol{Y}_{h\ell}$ is assumed to be ***positive***, as ordinarily $\sigma_b^2 > 0$ and $\sigma_e^2 > 0$.

- The model also assumes that $\text{var}(Y_{h\ell j})$ is constant for all j, so that overall variance does not ***change*** over time.

- Finally, (3.6) shows that the model implies that $\text{var}(\boldsymbol{Y}_{h\ell})$ is assumed to be ***the same*** for all groups $\ell = 1, \ldots, g$, which reflects the assumed independence of $b_{h\ell}$ and $\boldsymbol{e}_{h\ell}$ from the among-individual covariate group.

RESULT: This modeling approach and its assumptions induce a ***compound symmetric*** model for the overall aggregate pattern of correlation that is the same in each group. As we have noted previously, the compound symmetric model can be a restrictive representation of the overall pattern of correlation in the case ***within-individual*** sources of correlation are nonnegligible. The compound symmetric structure emphasizes ***among-individual*** sources of correlation, so may be reasonable when these sources are dominant.

The approach also induces the restriction that ***overall variance*** is ***constant*** across observation times. This is ***may not*** always be a realistic assumption for longitudinal data, as in many settings overall variance exhibits an ***increasing*** pattern over time.

Historically, the analysis methods we discuss shortly that are associated with this model have been used widely in agricultural, social science, and a host of other application areas, particularly before the advent of more modern methods, with little attention paid to the validity of these and other embedded assumptions. It is important that the data analyst understand the restrictions this approach involves.

ALTERNATIVE MODEL REPRESENTATION: It is of course possible to express model (3.4) in terms of the notation developed in Chapter 2. Recognizing that each (h, ℓ), $h = 1, \ldots, r_\ell$, $\ell = 1, \ldots, g$, indexes one of $m = \sum_{\ell=1}^{g} r_\ell$ unique individuals, we can reindex individuals, and thus $\boldsymbol{Y}_{h\ell}$, $b_{h\ell}$, and $\boldsymbol{e}_{h\ell}$, and $\epsilon_{h\ell}$ in (3.5), using a single index $i = 1, \ldots, m$ and reexpress the model in the form

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\mathcal{B}}_i + \boldsymbol{e}_i$$

as in (2.9), where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in})^T$ for individual i, as follows.

For illustration, take $g = 2$ and $n = 3$, and suppose individual i is in group 2. Then $E(\boldsymbol{Y}_i) = \boldsymbol{\mu}_i$ is such that $\boldsymbol{\mu}_i$ is equal to $\boldsymbol{\mu}_\ell$ in (3.5) with $\ell = 2$; that is,

$$\boldsymbol{\mu}_i = \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{pmatrix} = \begin{pmatrix} \mu + \tau_2 + \gamma_1 + (\tau\gamma)_{21} \\ \mu + \tau_2 + \gamma_2 + (\tau\gamma)_{22} \\ \mu + \tau_2 + \gamma_3 + (\tau\gamma)_{23} \end{pmatrix},$$

and the model can be written as

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \mathbf{1}b_i + \boldsymbol{e}_i = \boldsymbol{\mu}_i + \epsilon_i. \tag{3.9}$$

In fact, defining

$$\boldsymbol{\beta} = (\mu, \tau_1, \tau_2, \gamma_1, \gamma_2, \gamma_3, (\tau\gamma)_{11}, (\tau\gamma)_{12}, (\tau\gamma)_{13}, (\tau\gamma)_{21}, (\tau\gamma)_{22}, (\tau\gamma)_{23})^T,$$

and

$$\boldsymbol{X}_i = \left(\begin{array}{c|cc|ccc|cccccc} 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right),$$

(3.9) can be written as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \mathbf{1}b_i + \boldsymbol{e}_i \quad \text{or} \quad \boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, m. \tag{3.10}$$

All of this of course generalizes to any g and n.

- In this notation, information on group membership for individual *i* is incorporated in the definition of $\boldsymbol{\mu}_i$ and in particular, in (3.10), in the "design matrix" $\boldsymbol{X}_i$. All individuals with the same level of the group factor (i.e., sharing the same value of the among-individual covariate defining groups) have the same $\boldsymbol{\mu}_i$ and $\boldsymbol{X}_i$.

- As in an analysis of variance formulation, the "design matrix" $\boldsymbol{X}_i$ is ***not of full rank***, so that the "model" for the overall population mean $\boldsymbol{\mu}_i = \boldsymbol{X}_i\boldsymbol{\beta}$ for any individual is ***overparameterized***. To achieve a unique representation and ***identify*** the parameters μ, τ_ℓ, γ_j, and $(\tau\gamma)_{\ell j}$ for $\ell = 1, \ldots, 1$ and $j = 1, \ldots, n$, it is customary to impose the following ***constraints***:

$$\sum_{\ell=1}^{g} \tau_\ell = 0, \quad \sum_{j=1}^{n} \gamma_j = 0, \quad \sum_{\ell=1}^{g} (\tau\gamma)_{\ell j} = 0 = \sum_{j=1}^{n} (\tau\gamma)_{\ell j} \text{ for all } j, \ell, \tag{3.11}$$

  which is equivalent to redefining the vector of parameters $\boldsymbol{\beta}$ and the matrices $\boldsymbol{X}_i$ so that $\boldsymbol{X}_i$ is of ***full rank*** for all *i*.

QUESTIONS OF INTEREST AND STATISTICAL HYPOTHESES: As we have noted, a common objective in the analysis of longitudinal data is to assess if the way in which response changes over time is different for different populations of individuals that can be distinguished by values of among-individual covariates like gender in the dental study or dose in the guinea pig study. In ***classical*** statistical analysis, such questions are interpreted as pertaining to ***population mean response***; e.g., in the dental study, is ***pattern of change of mean response*** over age different for the populations of boys and girls?

Figure 3.8 depicts for $g = 2$ groups and $n = 3$ time points two situations in which the mean responses for each group for the three times lie on a ***straight line***. In the left panel, the ***rate of change***, represented by the ***slope*** of the two lines, is the ***same*** for both groups, so that the lines are ***parallel***, whereas in the right panel the rate of change for group 2 is steeper than for group 1, so that the lines are not parallel. Thus, in the left panel, the pattern of change is ***the same*** while in the right it is ***different***.

The left panel of Figure 3.9 shows for $g = 2$ groups and $n = 3$ time points a scenario where the ***mean response profiles*** are also ***parallel***, so that the pattern of change in each group is the same, but the means at each time ***do not*** follow an apparent straight line relationship. The right panel shows a case where the pattern is different.

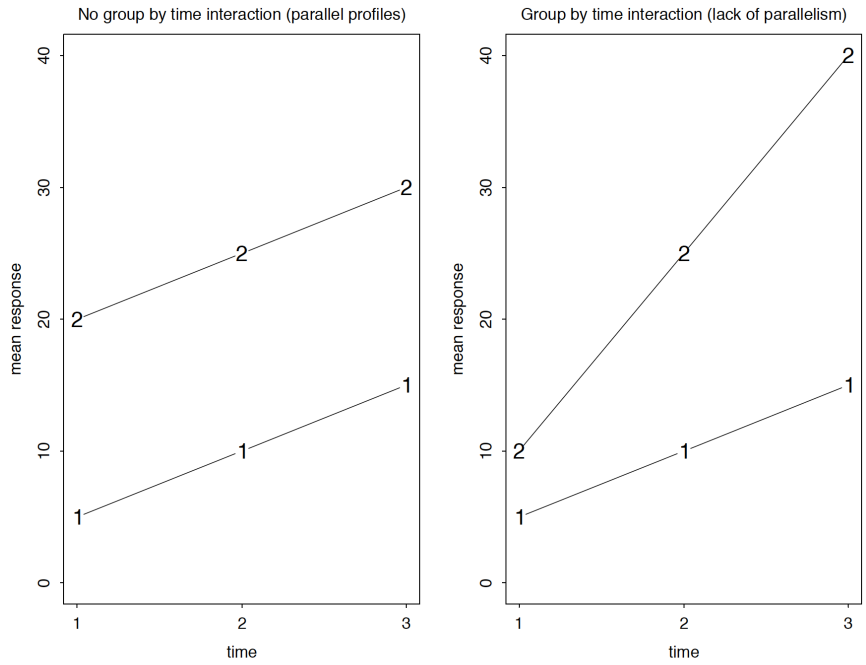

Figure 3.8: *Straight line mean profiles. Mean response for each group at each time, where the plotting symbol indicates group number. There is no interaction in the left panel; the right panel shows a quantitative interaction.*

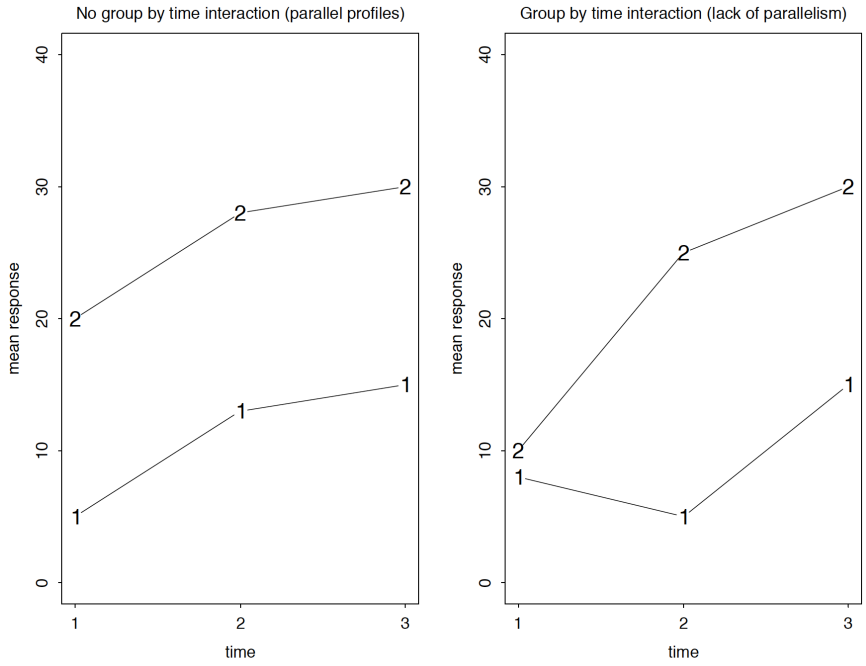

Figure 3.9: *Mean profiles not a straight line.*

GROUP BY TIME INTERACTION: In classical jargon, the situations in the right hand panels of Figures 3.8 and 3.9 depict examples of a ***group by time interaction***; in each panel, the difference in mean response between groups is ***not the same*** at all time points. In both figures, the ***direction*** of the difference in mean response between groups is ***the same*** in that the mean for group 2 is always larger. This is often referred to as a ***quantitative interaction***, particularly in health sciences research.

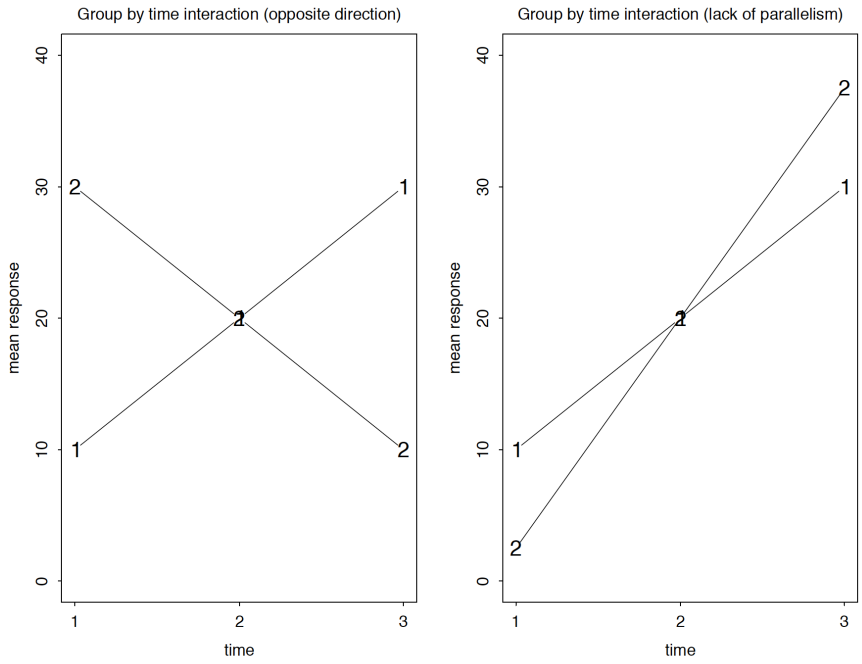

Figure 3.10: *Crossed mean profiles. Each panel represents a form of qualitative interaction.*

Figure 3.10 shows a different type of group by time interaction. In both panels, the difference in mean response between groups is again not the same at all time points, and the ***direction*** of the difference is ***not the same***, either. For example, in the left hand panel, the ***magnitude*** of the difference at times 1 and 3 is the same, but in the opposite direction. This is often referred to as a ***qualitative interaction***.

Returning to the model expressed using the classical notation as in (3.1), each mean in all of these figures is represented by

$$\mu_{\ell j} = \mu + \tau_\ell + \gamma_j + (\tau\gamma)_{\ell j}.$$

The difference between mean response in groups 1 and 2 at any time j is, under this model,

$$\mu_{1j} - \mu_{2j} = (\tau_1 - \tau_2) + \{(\tau\gamma)_{1j} - (\tau\gamma)_{2j}\}.$$

Thus, the $(\tau\gamma)_{\ell j}$ allow the difference in means between groups to be ***different*** at different times j, as in the right panels of Figures 3.8 and 3.9 and in Figure 3.10, by the amount $\{(\tau\gamma)_{1j} - (\tau\gamma)_{2j})\}$ at time j.

If the $(\tau\gamma)_{\ell j}$ were all the ***same***, the difference in means at any j reduces to

$$\mu_{1j} - \mu_{2j} = (\tau_1 - \tau_2),$$

so that the difference in mean response between groups is ***the same at all time points*** and equal to $(\tau_1 - \tau_2)$, which does not depend on j. This is the case in the left panels of Figures 3.8 and 3.9. Here, the ***pattern of change*** over time is thus ***the same***; i.e., the mean profiles for each group are ***parallel*** over time.

Under the constraints

$$\sum_{\ell=1}^{g}(\tau\gamma)_{\ell j} = 0 = \sum_{j=1}^{n}(\tau\gamma)_{\ell j} \text{ for all } \ell, j$$

in (3.11), if $(\tau\gamma)_{\ell j}$ are all ***the same*** for all ℓ, j, then it must be that

$$(\tau\gamma)_{\ell j} = 0 \text{ for all } \ell, j.$$

In general, then, if we wish to address the question of whether or not there is a common pattern of change over time, so whether or not the mean profiles are ***parallel***, we can cast this in terms of the ***null hypothesis***

$$H_0 : \text{ all } (\tau\gamma)_{\ell j} = 0, \quad \ell = 1, \ldots, g, \quad j = 1, \ldots, n, \tag{3.12}$$

with the alternative being that at least one $(\tau\gamma)_{\ell j} \neq 0$, in which case the mean difference at at least one of the time points is different from that at the others.

There are gn parameters $(\tau\gamma)_{\ell j}$; however, if the constraints above hold, then having $(g-1)(n-1)$ of the $(\tau\gamma)_{\ell j}$ equal to 0 automatically requires the remaining ones to be zero. Thus, the hypothesis (3.12) is really one about the behavior of $(g-1)(n-1)$ parameters, so there are $(g-1)(n-1)$ ***degrees of freedom*** associated with this hypothesis.

In the classical literature on analysis of variance for repeated measurements, the test of the null hypothesis (3.12) is referred to as the ***test for parallelism***. As Figure 3.9 demonstrates, parallelism does not necessarily mean that the pattern of mean response in each group follow a ***straight line***.

MAIN EFFECT OF GROUPS: If mean profiles are parallel, then the obvious next question is whether or not they are ***coincident*** ; that is, whether or not the mean response is in fact ***the same*** for each group at each time point. A little thought reveals that, if the mean profiles are ***parallel*** , if they are furthermore coincident, then the ***average*** of the mean responses over time will be the same for each group. The question of whether or not the average of mean responses is the same for each group if the profiles are ***not parallel*** may or may not be interesting or relevant.

- If in truth the situation were like those depicted in the right hand panels of Figures 3.8 and 3.9, whether or not the average of mean responses over time is different for the two groups might be interesting, as it would reflect that the mean response for group 2 is larger ***at all times***.
- On the other hand, consider the left panel of Figure 3.10. If this were the true state of affairs, this issue is ***meaningless***; the change of mean response over time is in the ***opposite*** direction for the two groups; thus, how it averages out over time is of little importance. Because the phenomenon of interest does indeed happen ***over time***, the ***average*** of what it does over time may be something that cannot be achieved – we can't make time stand still.
- Similarly, if the issue under study is something like growth, the ***average*** over time of the response may have little meaning; instead, one may be interested in, for example, how different the mean response is at the end of the time period of study. For example, in the right panel of Figure 3.10, mean response over time increases for each group at different rates, but has the same average over time. The group with the steeper rate will have a larger mean response at the end of the time period.

In general, then, the question of whether or not the average of the mean response over time is the same across groups in a longitudinal study is of most interest when the mean profiles over time are approximately parallel.

For definiteness, consider the case of $g = 2$ groups and $n = 3$ time points. For group ℓ, the average of means over time is, with $n = 3$,

$$n^{-1}(\mu_{\ell 1} + \mu_{\ell 2} + \mu_{\ell 3}) = \mu + \tau_\ell + n^{-1}(\gamma_1 + \gamma_2 + \gamma_3) + n^{-1}\{(\tau\gamma)_{\ell 1} + (\tau\gamma)_{\ell 2} + (\tau\gamma)_{\ell 3}\}.$$

The difference of the averages between $\ell = 1$ and $\ell = 2$ is then (algebra)

$$\tau_1 - \tau_2 + n^{-1}\sum_{j=1}^{n}(\tau\gamma)_{1j} - n^{-1}\sum_{j=1}^{n}(\tau\gamma)_{2j}.$$

The ***constraints*** (3.11) imposed to render the model of ***full rank*** dictate that $\sum_{j=1}^{n}(\tau\gamma)_{\ell j} = 0$ for each ℓ; thus, the two sums in this expression are 0 by assumption, so that we are left with $\tau_1 - \tau_2$.

Thus, the hypothesis may be expressed as

$$H_0 : \tau_1 - \tau_2 = 0.$$

Furthermore, under the constraint $\sum_{\ell=1}^{g}\tau_\ell = 0$, if the τ_ℓ are equal as in H_0, then they must satisfy $\tau_\ell = 0$ for each ℓ. Thus, the hypothesis may be rewritten as

$$H_0 : \tau_1 = \tau_2 = 0.$$

For general g and n, the reasoning is the same; we have

$$H_0 : \tau_1 = ... = \tau_g = 0. \tag{3.13}$$

MAIN EFFECT OF TIME: A further question of interest may be whether or not the mean response is in fact ***constant*** over time. If the profiles are parallel, then this is like asking whether the mean response averaged across groups is the ***same*** at each time. If the profiles are not parallel, then this may or may not be interesting. For example, in the left panel of Figure 3.10, the average of mean responses for groups 1 and 2 are the same at each time point. However, the mean response is certainly not constant across time for either group. If the groups represent a factor like gender, then what happens on average is something that can never be achieved.

The average of mean responses across groups for time j is

$$g^{-1}\sum_{\ell=1}^{g}\mu_{\ell j} = \gamma_j + q^{-1}\sum_{\ell=1}^{g}\tau_\ell + q^{-1}\sum_{\ell=1}^{g}(\tau\gamma)_{\ell j} = \gamma_j$$

using the constraints $\sum_{\ell=1}^{g}\tau_\ell = 0$ and $\sum_{\ell=1}^{g}(\tau\gamma)_{\ell j} = 0$ in (3.11). Thus, in the special case $g = 2$ and $n = 3$, having all these averages be the same at each time is equivalent to

$$H_0 : \gamma_1 = \gamma_2 = \gamma_3.$$

Under the constraint $\sum_{j=1}^{n}\gamma_j = 0$, then, we have $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0$. For general g and n, the hypothesis is of the form

$$H_0 : \gamma_1 = ... = \gamma_n = 0. \tag{3.14}$$

REMARK: Hypotheses (3.12), (3.13), and (3.14) are, of course, exactly the hypotheses that one tests for a ***split plot experiment*** , where, here, "time" plays the role of the "split plot" factor and "group" is the "whole plot factor." What is different is the ***interpretation*** ; because "time" has a natural ***ordering*** (longitudinal), what is interesting may be different; as noted above, of primary interest is whether or not the pattern of change in mean response over levels of time is different across groups.

ANALYSIS OF VARIANCE: Given that the statistical model and hypotheses of interest here are ***identical*** to those for a split plot, it should come as no surprise that the analysis is identical. Under the assumption that the model (3.1) is correctly specified and that the responses are normally distributed, so that

$$\boldsymbol{Y}_{h\ell} \sim \mathcal{N}_n(\boldsymbol{\mu}_\ell, \boldsymbol{V}), \quad \boldsymbol{V} = \sigma_b^2 \boldsymbol{J}_n + \sigma_e^2 \boldsymbol{I}_n. \tag{3.15}$$

as in (3.6), it can be shown the F ***ratios*** one would construct under the usual principles of analysis of variance provide the basis for valid tests of the hypotheses above. For brevity, we present the ***analysis of variance table*** and associated testing procedures without proof.

Define

- $\overline{Y}_{h\ell\cdot} = n^{-1} \sum_{j=1}^{n} Y_{h\ell j}$, the sample average over time for the hth unit in the ℓth group (over all observations on this unit)
- $\overline{Y}_{\cdot\ell j} = r_\ell^{-1} \sum_{h=1}^{r_\ell} Y_{h\ell j}$, the sample average at time j in group ℓ over all units
- $\overline{Y}_{\cdot\ell\cdot} = (r_\ell n)^{-1} \sum_{h=1}^{r_\ell} \sum_{j=1}^{n} Y_{h\ell j}$, the sample average of all observations in group ℓ
- $\overline{Y}_{\cdot\cdot j} = m^{-1} \sum_{\ell=1}^{g} \sum_{h=1}^{r_\ell} Y_{h\ell j}$, the sample average of all observations at the jth time
- $\overline{Y}_{\cdots}$ = the average of all mn observations.

Let

$$SS_G = \sum_{\ell=1}^{g} n r_\ell (\overline{Y}_{\cdot\ell\cdot} - \overline{Y}_{\cdots})^2, \quad SS_{Tot,U} = n \sum_{\ell=1}^{g} \sum_{h=1}^{r_\ell} (\overline{Y}_{h\ell\cdot} - \overline{Y}_{\cdots})^2$$

$$SS_T = m \sum_{j=1}^{n} (\overline{Y}_{\cdot\cdot j} - \overline{Y}_{\cdots})^2, \quad SS_{GT} = \sum_{j=1}^{n} \sum_{\ell=1}^{g} r_\ell (\overline{Y}_{\cdot\ell j} - \overline{Y}_{\cdots})^2 - SS_T - SS_G$$

$$SS_{Tot,all} = \sum_{\ell=1}^{g} \sum_{h=1}^{r_\ell} \sum_{j=1}^{n} (Y_{h\ell j} - \overline{Y}_{\cdots})^2.$$

Then the following analysis of variance table is constructed.

Source	SS	DF	MS	F
Among Groups	SS_G	$g-1$	MS_G	$F_G = MS_G/MS_{EU}$
Among-Unit Error	$SS_{Tot,U} - SS_G$	$m-g$	MS_{EU}	
Time	SS_T	$n-1$	MS_T	$F_T = MS_T/MS_E$
Group $\times$ Time	SS_{GT}	$(g-1)(n-1)$	MS_{GT}	$F_{GT} = MS_{GT}/MS_E$
Within-Unit Error	SS_E	$(m-g)(n-1)$	MS_E	
Total	$SS_{Tot,all}$	$nm-1$		

Here, the ***mean squares*** (MS) for each source are equal to the sum of squares (SS) divided by the degrees of freedom; e.g., $MS_G = SS_G/(g-1)$, and

$$SS_E = SS_{Tot,all} - SS_{GT} - SS_T - SS_{Tot,U}.$$

REMARK: It is traditional in the classical terminology to use the term "error;" however, it is important to recognize that the "Among Unit Error" includes variation due ***among-individual variability*** and the "Within-Unit Error" includes variation due to both ***within-individual realization*** (***fluctuations***) and ***measurement error***.

Under (3.15), the ***expectations*** of the mean squares in the table can be derived; we do not present these calculations here (they can be found in Section 3.3 of Crowder and Hand, 1990). These ***expected mean squares*** are shown in the following table; these are valid only if the model (3.15) is indeed correctly specified, so that the true overall, aggregate pattern of correlation is ***compound symmetric***.

Source	MS	Expected mean square
Among Groups	MS_G	$\sigma_e^2 + n\sigma_b^2 + n\sum_{\ell=1}^{g} r_\ell \tau_\ell^2/(g-1)$
Among-Unit error	MS_{EU}	$\sigma_e^2 + n\sigma_b^2$
Time	MS_T	$\sigma_e^2 + m\sum_{j=1}^{n} \gamma_j^2/(n-1)$
Group $\times$ Time	MS_{GT}	$\sigma_e^2 + \sum_{\ell=1}^{g} r_\ell \sum_{j=1}^{n} (\tau\gamma)_{\ell j}^2/(g-1)(n-1)$
Within-Unit Error	MS_E	σ_e^2

Inspection of the expected mean squares shows informally that we expect the F ratios in the analysis of variance table to test the appropriate issues. For example, we would expect F_{GT} to be large if the $(\tau\gamma)_{\ell j}$ are not all equal to zero, and F_G to be large is the τ_ℓ are not all equal to zero.

Note that F_G uses the appropriate denominator; intuitively, we wish to compare the mean square for groups against an "error term" that takes into account ***all*** sources of variation ***among*** (σ_b^2) and ***within*** (σ_e^2) individuals. The other two tests are on features that occur ***within individuals***; thus, the denominator takes account of the relevant source of variation, that within individuals (σ_e^2).

It can be shown formally that, as long as (3.15) is correctly specified, under the null hypotheses (3.12), (3.13), and (3.14), the ***sampling distributions*** of the F ratios in the analysis of variance table are F distributions with the degrees of freedom specified below.

TEST PROCEDURES: We now summarize the procedures for testing each of the hypotheses. Here, $\mathcal{F}_{a,b,\alpha}$ is the critical value corresponding to level of significance α for an F distribution with a numerator and b denominator degrees of freedom.

- ***Group by time interaction (parallelism)***, (3.12).

  $$H_0 : (\tau\gamma)_{\ell j} = 0 \text{ for all } j, \ell \text{ vs. } H_1 : \text{ at least one } (\tau\gamma)_{\ell j} \neq 0.$$

  A valid test rejects H_0 at level of significance α if

  $$F_{GT} > \mathcal{F}_{(g-1)(n-1),(n-1)(m-g),\alpha}$$

  or, equivalently, if the probability is less than α that one would see a value of the test statistic as large or larger than F_{GT} if H_0 were true (that is, the p-value is less than α).

- ***Main effect of group (coincidence)***, (3.13).

  $$H_0 : \tau_\ell = 0 \text{ for all } \ell \text{ vs. } H_1 : \text{ at least one } \tau_\ell \neq 0.$$

  A valid test rejects H_0 at level of significance α if

  $$F_G > \mathcal{F}_{g-1,m-g,\alpha}.$$

- ***Main effect of time (constancy)***, (3.14).

  $$H_0 : \gamma_j = 0 \text{ for all } j \text{ vs. } H_1 : \text{ at least one } \gamma_j \neq 0.$$

  A valid test rejects H_0 at level α if

  $$F_T > \mathcal{F}_{n-1,(n-1)(m-g),\alpha}.$$

VIOLATION OF COVARIANCE MATRIX ASSUMPTION: These test procedures are valid under model (3.15), which embodies the assumption of ***compound symmetry*** of the overall correlation matrix of a data vector. In fact, it can be shown that they are also valid under slightly ***more general conditions*** that include compound symmetry as a special case. However, validity of the tests is predicated on the covariance matrix being of the special form we discuss next; if not, then F ratios F_T and F_{GT} ***no longer*** have exactly an F distribution, and the associated tests are not valid and can lead to erroneous conclusions.

A $(n \times n)$ matrix $\boldsymbol{V}$ is said to be of ***Type H*** if it can be written in the form

$$\boldsymbol{V} = \begin{pmatrix} \lambda + 2\alpha_1 & \alpha_1 + \alpha_2 & \cdots & \alpha_1 + \alpha_n \\ \alpha_2 + \alpha_1 & \lambda + 2\alpha_2 & \cdots & \alpha_2 + \alpha_n \\ \vdots & \vdots & \vdots & \vdots \\ \alpha_n + \alpha_1 & \alpha_n + \alpha_2 & \cdots & \lambda + 2\alpha_n \end{pmatrix}. \tag{3.16}$$

It is straightforward to deduce that a covariance matrix with correlation structure that is compound symmetric is of Type H.

It can be shown that, as long as the data vectors $\boldsymbol{Y}_{h\ell}$ are multivariate normal with common covariance matrix $\boldsymbol{V}$ of the form (3.16), the F tests discussed above ***will be valid***. Thus, because (3.16) includes compound symmetry, the tests are valid if model 3.15) holds. If the overall covariance matrix is ***not*** of Type H, but these F tests are conducted nonetheless, they will be too ***liberal***; that is, they will reject the null hypothesis more often than they should, so that, for example, the analyst might conclude that there is sufficient evidence supporting a group by time interaction when there really is not.

It is possible to construct tests of whether or not the true overall covariance matrix is of Type H. One such test is ***Mauchly's test for sphericity***. We do not present the form and derivation of this test here; description of the test is given by Vonesh and Chinchilli (1997, p. 85), for example. The test statistic for testing the null hypothesis

$$H_0 : \boldsymbol{V} \text{ is of Type H,}$$

where $\boldsymbol{V}$ is the true covariance matrix of a response vector, has approximately a χ^2 (chi-square) distribution when the number of individuals m is "large," with degrees of freedom $(n-2)(n+1)/2$. Thus, the test is performed at level of significance α by comparing the value of the test statistic to the χ^2_α critical value with $(n-2)(n+1)/2$ degrees of freedom.

All such tests have ***limitations*** : they are not very powerful with the numbers of individuals r_ℓ in each group is not large, and they can be misleading if the true distribution of the response vectors is not multivariate normal. Accordingly, we do not discuss it further. We return to the issue of approaches when the analyst lacks confidence in the validity of the assumption of Type H covariance strutcure in the next section.

3.3 Specialized within-individual hypotheses and tests

The hypotheses of group by time interaction (parallelism) and main effect of time (constancy) have to do with questions about the ***pattern of change*** over time. However, they address these issues in an "overall" sense; e.g., the test of the group by time interaction asks only if the pattern of mean responses over time is different for different groups, but it does ***not provide insight*** into the nature of the pattern of change and how it differs.

We now review methods to carry out a ***more detailed study*** of specific aspects of how the mean response changes over time. As we demonstrate, these methods do this through testing of specialized hypotheses.

It is conventional to present the relevant null hypotheses, and, indeed, the three main null hypotheses (3.12), (3.13), and (3.14) using the following ***unified notation***. Let $\boldsymbol{\mathcal{M}}$ denote the matrix of all means $\mu_{\ell j}$ implied by the model (3.1), i.e.

$$\boldsymbol{\mathcal{M}} = \begin{pmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{g1} & \mu_{g2} & \cdots & \mu_{gn} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_1^T \\ \vdots \\ \boldsymbol{\mu}_g^T \end{pmatrix}, \tag{3.17}$$

so that the ℓth row of $\boldsymbol{\mathcal{M}}$ in (3.17) is $\boldsymbol{\mu}_\ell^T$. Let

- $\boldsymbol{C}$ be a $(c \times g)$ matrix with $c \leq g$ of full rank.
- $\boldsymbol{U}$ be a $(n \times u)$ matrix with $u \leq n$ of full rank.

Then it is possible to express null hypotheses of interest in the ***general form***

$$H_0 : \boldsymbol{C}\boldsymbol{\mathcal{M}}\boldsymbol{U} = \boldsymbol{0}. \tag{3.18}$$

In this formulation

- the matrix $\boldsymbol{C}$ specifies ***differences among*** or ***averages across*** groups
- the matrix $\boldsymbol{U}$ specifies ***differences over*** or ***averages across*** levels of time.

Depending on the choices of the matrices $\boldsymbol{C}$ and $\boldsymbol{U}$ in (3.18), the resulting ***linear function*** $\boldsymbol{C\mathcal{M}U}$ of the elements of $\mathcal{M}$ (the individual means for different groups at different time points) can be made to address ***specialized questions*** regarding differences in mean response among groups and in patterns of change over time.

To see this, first consider the null hypothesis for the ***group by time interaction*** (parallelism) (3.12), H_o : all $(\tau\gamma)_{\ell j} = 0$, with $g = 2$ groups and $n = 3$ time points. Take

$$\boldsymbol{C} = \begin{pmatrix} 1, & -1 \end{pmatrix}, \tag{3.19}$$

so that $c = 1 = g - 1$. Note that

$$\begin{aligned} \boldsymbol{C\mathcal{M}} &= \begin{pmatrix} 1, & -1 \end{pmatrix} \begin{pmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ \mu_{21} & \mu_{22} & \mu_{23} \end{pmatrix} = \begin{pmatrix} \mu_{11} - \mu_{21}, & \mu_{12} - \mu_{22}, & \mu_{13} - \mu_{23} \end{pmatrix} \\ &= \begin{pmatrix} \tau_1 - \tau_2 + (\tau\gamma)_{11} - (\tau\gamma)_{21}, & \tau_1 - \tau_2 + (\tau\gamma)_{12} - (\tau\gamma)_{22}, & \tau_1 - \tau_2 + (\tau\gamma)_{13} - (\tau\gamma)_{23} \end{pmatrix} \end{aligned}$$

Thus, $\boldsymbol{C}$ yields differences in means among groups at each time point.

Take

$$\boldsymbol{U} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix}, \tag{3.20}$$

so that $u = 2 = n - 1$. Thus, $\boldsymbol{U}$ involves differences for pairs of time points.

It is straightforward (try it) to show that

$$\begin{aligned} \boldsymbol{C\mathcal{M}U} &= \begin{pmatrix} \mu_{11} - \mu_{21} - \mu_{12} + \mu_{22}, & \mu_{12} - \mu_{22} - \mu_{13} + \mu_{23} \end{pmatrix} \\ &= \begin{pmatrix} (\tau\gamma)_{11} - (\tau\gamma)_{21} - (\tau\gamma)_{12} + (\tau\gamma)_{22}, & (\tau\gamma)_{12} - (\tau\gamma)_{22} - (\tau\gamma)_{13} + (\tau\gamma)_{23} \end{pmatrix}. \end{aligned}$$

It is an exercise in algebra to verify that, under the constraints in (3.11), if each of these elements equals zero, then H_0 follows.

Similarly, for the null hypothesis for the ***main effect of groups*** (coincidence), $H_0 : \tau_1 = \tau_2 = 0$, taking $\boldsymbol{C}$ to be as in (3.19) and

$$\boldsymbol{U} = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix},$$

it is straightforward to see that, with $n = 3$,

$$\boldsymbol{CMU} = \tau_1 - \tau_2 + n^{-1}\sum_{j=1}^{n}(\tau\gamma)_{1j} - n^{-1}\sum_{j=1}^{n}(\tau\gamma)_{2j}.$$

That is, this choice of $\boldsymbol{U}$ dictates an ***averaging*** operation across time. Imposing the constraints as above, we can express H_0 in the form $H_0 : \boldsymbol{CMU} = 0$.

To express the null hypothesis for the ***main effect of time*** (constancy), $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0$, take

$$\boldsymbol{U} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix}, \quad \boldsymbol{C} = \begin{pmatrix} 1/2, & 1/2 \end{pmatrix}.$$

Here, $\boldsymbol{C}$ involves ***averaging across groups*** while $\boldsymbol{U}$ involves differences for pairs of time points as above. Then

$$\begin{aligned} \boldsymbol{MU} &= \begin{pmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ \mu_{21} & \mu_{22} & \mu_{23} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} \mu_{11} - \mu_{12} & \mu_{12} - \mu_{13} \\ \mu_{21} - \mu_{22} & \mu_{22} - \mu_{23} \end{pmatrix} \qquad (3.21) \\ &= \begin{pmatrix} \gamma_1 - \gamma_2 + (\tau\gamma)_{11} - (\tau\gamma)_{12}, & \gamma_2 - \gamma_3 + (\tau\gamma)_{12} - (\tau\gamma)_{13} \\ \gamma_1 - \gamma_2 + (\tau\gamma)_{21} - (\tau\gamma)_{22}, & \gamma_2 - \gamma_3 + (\tau\gamma)_{22} - (\tau\gamma)_{23} \end{pmatrix}. \end{aligned}$$

from whence it is straightforward to derive, imposing the constraints in (3.11), that

$$\boldsymbol{CMU} = \begin{pmatrix} \gamma_1 - \gamma_2, & \gamma_2 - \gamma_3 \end{pmatrix}.$$

Setting this equal to zero with the constraint $\sum_{j=1}^{n}\gamma_j = 0$ yields H_0.

Clearly, the principles involved in specifying the matrices $\boldsymbol{C}$ and $\boldsymbol{U}$ to yield the form of the null hypotheses corresponding to the group by time interaction, main effect of groups, and main effect of time generalize to any g and n.

Other choices of $\boldsymbol{C}$ and $\boldsymbol{U}$ can be made to examine ***components*** making up these overall hypotheses and to isolate ***specific features*** of the pattern of change. Recall the following definition.

CONTRAST: If if $\boldsymbol{c}$ is a $(n \times 1)$ vector and $\boldsymbol{\mu}$ is a $(n \times 1)$ vector of means, then the ***linear combination*** $\boldsymbol{c}^T\boldsymbol{\mu} = \boldsymbol{\mu}^T\boldsymbol{c}$ is a ***contrast*** if $\boldsymbol{c}$ is such that its elements sum to zero; i.e., $\boldsymbol{c}^T\mathbf{1} = 0$.

Thus, for example, with $g = 2$ and $n = 3$, the columns of the matrix

$$\boldsymbol{U} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix}$$

in (3.20) define contrasts of elements of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, the mean vectors for groups 1 and 2, in that

$$\mathcal{M}\boldsymbol{U} = \begin{pmatrix} \mu_{11} - \mu_{12} & \mu_{12} - \mu_{13} \\ \mu_{21} - \mu_{22} & \mu_{22} - \mu_{23} \end{pmatrix}. \tag{3.22}$$

In (3.22), each entry is a ***contrast*** involving differences in mean response between pairs of times in each group. Specialized questions of interest can be posed by considering these contrasts.

- The ***difference*** of the contrasts in the first column of (3.22) focuses on whether or not the way in which the mean response differs from time 1 to time 2 is different in groups 1 and 2. This feature is clearly a ***component*** of the overall group by time interaction, focusing in particular on times 1 and 2. Likewise, the difference of the contrasts in the second columns of (3.22) focuses on the same for times 2 and 3, and is also part of the group by time interaction.

  Indeed, taken together, the differences of constrasts in both columns of (3.22) ***fully characterize*** the overall group by time interaction.

- Similarly, the ***average*** of the contrasts in the first column of (3.22) focuses on the difference in mean response between times 1 and 2, averaged across groups. This is clearly a component of the main effect of time. Similarly, the average of contrasts in the second column reflects the same for times 2 and 3. Again, taken together, the averages of contrasts in both rows of (3.22) ***fully characterize*** the overall main effect of time.

Thus, considering these contrasts and their differences among or averages across groups serves to "***pick apart***" how the overall group by time interaction effect and main effect of time occur and can provide insight into ***specific features*** of the pattern of change over time.

PROFILE TRANSFORMATION: For general number of groups g and number of time points n, the extension of (3.20) is the $(n \times n-1)$ matrix

$$\boldsymbol{U} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 1 \\ 0 & \cdots & 0 & -1 \end{pmatrix}. \tag{3.23}$$

Postmultiplication of $\mathcal{M}$ by $\boldsymbol{U}$ in (3.23) results in contrasts comparing means at successive pairs of time points and is often called the ***profile transformation*** of the means. Examining individually the differences among or averages across the contrasts resulting from each column provides insight on the contribution to the overall pattern of change over time.

Other $\boldsymbol{U}$ matrices allow other ways of "parsing" the pattern of change over time. For example, instead of focusing on changes from one time to the next, one might consider how the mean at a specific time point differs from what happens at ***all subsequent time points***. This might highlight at what point in time changes in the pattern begin to emerge.

We demonstrate with $g = 2$ and $n = 4$ and consider the contrast

$$\mu_{11} - (\mu_{12} + \mu_{13} + \mu_{14})/3,$$

which compares, for group 1, the mean at time 1 to the ***average*** of means at all other times. Similarly,

$$\mu_{12} - (\mu_{13} + \mu_{14})/2$$

compares for group 1 the mean at time 2 to the average of those at subsequent times. The final contrast of this type for group 1 is

$$\mu_{13} - \mu_{14},$$

which compares what happens at time 3 to the "average" of what comes next, the single mean at time 4. We may similarly specify such contrasts for the other group.

These contrasts can be obtained by postmultiplying $\mathcal{M}$ by

$$\boldsymbol{U} = \begin{pmatrix} 1 & 0 & 0 \\ -1/3 & 1 & 0 \\ -1/3 & -1/2 & 1 \\ -1/3 & -1/2 & -1 \end{pmatrix}. \tag{3.24}$$

In particular, with $g = 2$,

$$\mathcal{M}\boldsymbol{U} = \begin{pmatrix} \mu_{11} - \mu_{12}/3 - \mu_{13}/3 - \mu_{14}/3, & \mu_{12} - \mu_{13}/2 - \mu_{14}/2, & \mu_{13} - \mu_{14} \\ \mu_{21} - \mu_{22}/3 - \mu_{23}/3 - \mu_{24}/3, & \mu_{22} - \mu_{23}/2 - \mu_{24}/2, & \mu_{23} - \mu_{24} \end{pmatrix}. \tag{3.25}$$

HELMERT TRANSFORMATION: For general n, the $(n \times n-1)$ matrix whose columns define contrasts of this type is the so-called ***Helmert transformation*** matrix of the form

$$\boldsymbol{U} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1/(n-1) & 1 & 0 & \cdots & 0 \\ -1/(n-1) & -1/(n-2) & 1 & \cdots & 0 \\ \vdots & \vdots & -1/(n-3) & \vdots & \vdots \\ -1/(n-1) & -1/(n-2) & \vdots & \cdots & 1 \\ -1/(n-1) & -1/(n-2) & -1/(n-3) & \cdots & -1 \end{pmatrix}. \tag{3.26}$$

Postmultiplication of $\mathcal{M}$ by a matrix of the form (3.26) yields contrasts representing comparisons of each mean against the ***average*** of means at all subsequent times.

It is in fact the case that any $\boldsymbol{U}$ matrix with $n-1$ columns, so involving $n-1$ constrasts that "pick apart" all possible differences in means over time, as do the ***profile*** and ***Helmert*** transformation matrices (3.23) and (3.26), lead to the overall hypotheses for group by time interaction and main effect of time when paired with the appropriate $\boldsymbol{C}$ matrix.

For example, in the case $g = 2$, $n = 3$, if we premultiply ***either*** of (3.21) or (3.25) by $\boldsymbol{C} = (1/2, 1/2)$ and impose the constraints (3.11), we are led to the null hypothesis $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0$ for the ***main effect of time***.

In particular, with (3.21),

$$\boldsymbol{C}\boldsymbol{\mathcal{M}}\begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} \gamma_1 - \gamma_2, & \gamma_2 - \gamma_3 \end{pmatrix} = \mathbf{0},$$

while with (3.25),

$$\boldsymbol{C}\boldsymbol{\mathcal{M}}\begin{pmatrix} 1 & 0 \\ -1/2 & 1 \\ -1/2 & -1 \end{pmatrix} = \begin{pmatrix} \gamma_1 - 0.5\gamma_2 - 0.5\gamma_3, & \gamma_2 - \gamma_3 \end{pmatrix} = \mathbf{0},$$

both of which can be shown to imply $\gamma_1 = \gamma_2 = \gamma_3$. The diligent student can verify a similar result for the group by time interaction with $\boldsymbol{C} = (1, -1)$.

In general, for a $(n \times n-1)$ $\boldsymbol{U}$ matrix involving $n-1$ contrasts that characterizes all possible differences in means over time in a particular way,

- premultiplying $\boldsymbol{\mathcal{M}}\boldsymbol{U}$ by the $(g-1) \times g$ matrix

$$\boldsymbol{C} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{pmatrix},$$

  considers how each contrast defined by the columns of $\boldsymbol{U}$ differs across groups, which is a component of the ***group by time interaction*** (how the difference in means across groups is different at different times).

- premultiplying by $\boldsymbol{C} = (1/g, 1/g, \ldots, 1/g)$, each of the $n-1$ elements of the resulting $1 \times (n-1)$ matrix correspond to the ***average*** of each of these contrasts over groups, which all together constitute the ***main effect of time***.

SEPARATE COMPONENT TESTS: We can examine each of these components ***separately*** to explore particular aspects of the pattern of mean response over time. Formally, we can carry out separate hypothesis tests corresponding to each component. This can be accomplished as follows.

Consider the kth column of $\boldsymbol{U}$, denoted as $\boldsymbol{c}_k$, $k = 1, \ldots, n-1$.

- For each individual indexed by (h, ℓ), obtain

  $$\boldsymbol{c}_k' \boldsymbol{Y}_{h\ell},$$

  which distills the repeated measurements on the individual to a single quantity representing the value of the component contrast for that individual. If $\text{var}(\boldsymbol{Y}_{h\ell}) = \boldsymbol{V}$, so is the same for all individuals, then these values have the ***same variance*** for all individuals (see below).

- Carry out analyses on the resulting "data;" e.g., to test if the component differs across groups, conduct a usual one-way analysis of variance on these "data" and carry out the F-test for the Group effect.

- To test if the contrast is zero averaged across groups, test whether the overall mean of the "data" is equal to zero using using a standard t test (or equivalently, the F test based on the square of the t statistic).

- These tests will be valid ***regardless*** of whether or not ***compound symmetry*** holds; all that matters is that $\boldsymbol{V}$, whatever it is, is ***the same*** for all units, in which case

  $$\text{var}(\boldsymbol{c}_k' \boldsymbol{Y}_{h\ell}) = \boldsymbol{c}_k' \boldsymbol{V} \boldsymbol{c}_k,$$

  which is a constant for all h and ℓ, so that the usual assumption of constant variance necessary for the above analyses holds for the "data" corresponding to each contrast.

ORTHOGONAL CONTRASTS: Recall that if $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ are any two columns of $\boldsymbol{U}$, then if $\boldsymbol{c}_1' \boldsymbol{c}_2 = 0$, $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ are said to be ***orthogonal***. The ***contrasts*** corresponding to these vectors are ***orthogonal contrasts***. The contrasts corresponding to the columns of the ***profile transformation*** matrix (3.23) ***are not*** orthogonal, while those of the ***Helmert transformation*** matrix (3.26) ***are*** (try it).

There is an advantage to a transformation whose columns and thus embedded contrasts are orthogonal. As intuition might suggest, it can be shown that a set of $n-1$ orthogonal contrasts ***partitions*** the total Group $\times$ Time and Within-Unit Error sums of squares into $n-1$ ***distinct*** or "nonoverlapping" components. Informally, this implies that the outcome of one of the component hypothesis tests addressing a particular contrast can be considered separately, regardless of the outcome of the tests for the others. If in fact a $\boldsymbol{U}$ matrix with orthogonal columns is ***normalized*** , then, furthermore, the sums of squares for the Group effect from each of the $k = 1, \ldots, n-1$ analyses of variance above will ***sum*** to SS_{GT}, and, similarly, the error sums of squares from each of these will sum to SS_E.

To see this in a special case, consider the Helmert matrix in (3.24) with $n = 4$,

$$\boldsymbol{U} = \begin{pmatrix} 1 & 0 & 0 \\ -1/3 & 1 & 0 \\ -1/3 & -1/2 & 1 \\ -1/3 & -1/2 & -1 \end{pmatrix}.$$

Each of the columns c_k, $k = 1, \ldots, n-1 = 3$, yields a function of a data vector $\boldsymbol{c}_k' \boldsymbol{Y}_{h\ell}$ that is on a different scale, so that the sums of squares from the individual analyses of variance corresponding to each are not comparable. It is possible to modify each contrast without affecting orthogonality yield a common scale by ***normalizing*** each column; i.e., divide each column by the square root of the sums of squares of its elements so that the sum of the squares of the modified elements is equal to one. E.g., for $\boldsymbol{c}_1$, the sum of squared elements is $1^2 + (-1/3)^2 + (-1/3)^2 + (-1/3)^2 = 4/3$, yielding the normalized version $\sqrt{3/4}\boldsymbol{c}_1$; similarly, the normalized versions of $\boldsymbol{c}_2$ and $\boldsymbol{c}_3$ are $\sqrt{2/3}\boldsymbol{c}_2$ and $\sqrt{1/2}\boldsymbol{c}_3$. When all contrasts in an orthogonal transformation are scaled in this way, then they are said to be ***orthonormal.*** If the orthonormal contrasts are used in the individual analyses above to form the "data," then the sums of squares from each ***do*** sum to the overall SS_{GT} and SS_E.

It is ***not necessary*** to use normalized contrasts to obtain the correct test statistics for each component contrast. The ***same*** test statistics will result; clearly, although each analysis is on a different scale, the F ratios will be the same, as the normalization factor will ***cancel*** from numerator and denominator. The orthonormal version of the transformation is often used simply because it leads to the nice, intuitive additive property.

If the component contrasts in the chosen ***U*** matrix are ***not orthogonal*** , interpretation of the separate tests is more difficult, as the sums of squares are no longer partitioned as above, so that the outcome of one test is related to that of another.

ORTHOGONAL POLYNOMIAL CONTRASTS: As we noted at the outset, the ***statistical model*** (3.5) does not ***acknowledge explicitly*** that the response on a given individual likely evolves over ***continuous time***. In particular, in the context of the ***conceptual framework*** in Chapter 2, the model does not incorporate formally an acknowledgment of a ***smooth*** underlying trajectory, as is apparent in ***EXAMPLES 1 – 4*** in that chapter.

Accordingly, there is a need to be able to evaluate behavior of the mean response over (continuous) time in the context of the statistical model that acknowledges possible ***smooth patterns*** of change. For example, in the dental study, we might wish to evaluate whether or not there is a ***linear*** or in fact ***quadratic*** trend over time, averaged across genders and whether or not the linear or quadratic trend differs between genders.

This is facilitated by, for n time points, the set of $n-1$ ***orthogonal polynomial contrasts***. These contrasts are based on the premise that, with data at the same n time points, it is possible to fit up to a $(n-1)$th degree polynomial in time. Thus, just as the profile and Helmert transformations ***decompose*** the overall time effect into $n-1$ contrasts addressing specific differences over time, these contrasts decompose this into ***orthogonal components*** reflecting the strength of linear, quadratic, cubic, and so on contributions to the saturated $(n-1)$th degree polynomial. This is possible for times that are ***equally*** or ***unequally*** spaced; we do not present derivations here. For equally-spaced time points, the coefficients of the $n-1$ orthogonal polynomials are available in many classical statistics texts; for unequally-spaced times, the coefficients depend on the times themselves.

For example, for $n = 4$, there are $n - 1 = 3$ possible (orthogonal) contrasts corresponding to ***linear*** , ***quadratic*** , and ***cubi***c components of the overall smooth trend, which are characterized in the columns of following ***U*** matrix:

$$\boldsymbol{U} = \begin{pmatrix} -3 & 1 & -1 \\ -1 & -1 & 3 \\ 1 & -1 & -3 \\ 3 & 1 & 1 \end{pmatrix}.$$

It can be verified that the columns of ***U*** are ***orthogonal***.

With the appropriate set of orthogonal polynomial contrasts, one can proceed as above to conduct separate hypothesis tests addressing the strength of the linear, quadratic, and so on components of the mean response trajectory over time. The orthogonal polynomial transformation can also be "normalized" as discussed above.

ADJUSTED TESTS: We conclude this section by returning to the assumption embodied in the model (3.5) that the overall aggregate correlation structure is that of ***compound symmetry*** or at least of ***Type H***. As noted previously, if the assumption of Type H does ***not*** hold, then the usual F tests of the group by time interaction and main effect of time are ***invalid*** in that they will be ***too liberal***.

If the analyst doubts the relevance of this assumption, methods are available to "***adjust***" the usual *F* tests. We sketch how this approach works without providing technical justification or details.

Define

$$\epsilon = \frac{\text{tr}^2(\boldsymbol{U}'\boldsymbol{V}\boldsymbol{U})}{(n-1)\text{tr}(\boldsymbol{U}'\boldsymbol{V}\boldsymbol{U}\boldsymbol{U}'\boldsymbol{V}\boldsymbol{U})},$$

where $\boldsymbol{U}$ is any $(n \times n-1)$ matrix whose columns are ***normalized orthogonal contrasts***. It can be shown that the constant ϵ defined in this way must satisfy

$$1/(n-1) \leq \epsilon \leq 1$$

and that

$$\epsilon = 1$$

if, and only if , $\boldsymbol{V}$ is of Type H.

Because the usual *F* tests are ***too liberal*** if $\boldsymbol{V}$ is not of Type H, one suggestion is, rather than compare the *F* ratios to the usual *F* critical values with *a* and *b* numerator and denominator degrees of freedom, say, compare them instead to *F* critical values with ϵa and ϵb numerator and denominator degrees of freedom instead. This will make the degrees of freedom ***smaller*** than usual. It can be verified that, as the numerator and denominator degrees of freedom get ***smaller*** , the value of the critical value gets ***larger***. Thus, the effect of this "adjustment" is to compare *F* ratios to larger critical values, making it harder to reject the null hypothesis and thus making the test less ***liberal***.

- Of course, ϵ is not known, because it depends on the unknown $\boldsymbol{V}$ matrix. Thus, different adjustments are based on different approaches to ***estimating*** $\boldsymbol{V}$ and using the result to estimate ϵ.
- Two such approaches are the ***Greenhouse-Geisser*** and ***Huynh-Feldt*** adjustments. Each estimates ϵ in a different way; the Huynh-Feldt estimate is such that the adjustment to the degrees of freedom is not as severe as that of the Greenhouse-Geisser adjustment. These adjustments are based on ***asymptotic approximations*** , so that it is not necessarily the case that they will lead to valid tests when the numbers of individuals are small

SUMMARY: The spirit of the methods discussed above can be summarized as follows. One adopts a ***statistical model*** that makes the very specific assumption of ***compound symmetry*** of the aggregate correlation structure among responses on the same individual.

If this assumption is correct, then familiar analysis of variance methods are available to ***test hypotheses*** regarding the ***pattern of change*** over time and mean response averaged across groups. However, the model does not lend itself readily to ***estimation*** of features of the pattern of change, and the procedures to construct tests to study different features of the pattern are rather ***unwieldy***. It is possible to carry out a test of whether or not the compound symmetry assumption is supported by the data; however, the testing procedures are not reliable. Approximate, "adjusted" versions of the tests are available, but these are not necessarily reliable, either.

The bottom line is that a ***better approach*** might be to start with a more realistic and flexible ***statistical model*** within which to characterize and evaluate features of the pattern of change. This is the basis for the more ***modern*** methods we study in later chapters.

3.4 Multivariate repeated measures analysis of variance

We conclude our discussion of classical approaches with a brief overview of ***multivariate repeated measures analysis of variance*** methods.

MULTIVARIATE MODEL: The set-up and notation are identical to those introduced in Section 3.2; i.e., individuals belong to one of $g \geq 1$ groups, and the response is ascertained on each individual at n time points. There are r_ℓ individuals in each group, indexed by $h = 1 \ldots, r_\ell$, for a total of $m = \sum_{\ell=1}^{g} r_\ell$.

The representation of the ***overall population mean response*** is the same as for the univariate approach; namely, for the hth individual in the ℓth group at time j,

$$E(Y_{h\ell j}) = \mu_{\ell j} = \mu + \tau_\ell + \gamma_j + (\tau\gamma)_{\ell j}.$$

However, as we noted at the beginning of this chapter, these methods are based on a ***PA perspective*** , so do not acknowledge ***among-*** and ***within-individual*** sources of correlation separately and explicitly, and they make ***no specific assumption*** on the form of the overall pattern of covariance, taking it to be completely ***unstructured***.

In particular, the assumed model is

$$Y_{h\ell j} = \underbrace{\mu + \tau_\ell + \gamma_j + (\tau\gamma)_{\ell j}}_{\mu_{\ell j}} + \epsilon_{h\ell j} \quad \text{or} \quad \boldsymbol{Y}_{h\ell} = \boldsymbol{\mu}_\ell + \boldsymbol{\epsilon}_{h\ell}, \quad \boldsymbol{\epsilon}_{h\ell} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{V}),$$

where $\boldsymbol{V}$ is an ***arbitrary*** covariance matrix with no particular structure; that is, $\boldsymbol{V}$ is an ***unstructured*** covariance matrix.

Thus, the model can be summarized as

$$\boldsymbol{Y}_{h\ell} \sim \mathcal{N}(\boldsymbol{\mu}_\ell, \boldsymbol{V}), \quad h = 1, \ldots, r_\ell, \ \ell = 1, \ldots, q, \tag{3.27}$$

where ***V*** is ***completely unstructured***, depending on $n(n+1)/2$ ***distinct parameters*** (compared to the ***two*** parameters, σ_b^2 and σ_e^2, characterizing the compound symmetric structure underlying the univariate methods).

Under model (3.27), it is of course possible to conceive the ***same hypotheses*** of parallelism, coincidence, and constancy as under the univariate model.

- However, because of the unstructured covariance assumption, it is ***no longer possible*** to derive straightforward test statistics involving ratios of simple mean squares based on individual response observations.
- Instead, one must view the problem from a ***multivariate*** perspective and develop testing procedures based on classical ***multivariate analysis of variance*** (***MANOVA***) techniques.
- Just as the univariate methods make an assumption, that of ***compound symmetry***, for the overall aggregate covariance structure that may be ***too restrictive*** for many longitudinal data problems, the multivariate methods make an assumption that is usually ***too general***.
- Because this assumption is so general, involving $n(n+1)/2$ ***covariance parameters***, these procedures are not very ***powerful*** for detecting departures from null hypotheses of interest.
- Thus, these methods are of ***limited practical utility*** and are ***rarely used*** anymore for longitudinal data analysis. Our presentation is meant only to provide an introduction to the basic ideas.

GENERAL MULTIVARIATE PROBLEM: In the longitudinal data setting, the elements of a response vector $\boldsymbol{Y}_{h\ell}$ are observations on the ***same*** response over time. In the ***general multivariate problem***, the components of $\boldsymbol{Y}_{h\ell}$ can be, but are ***not necessarily***, observations on the same response. Instead, they can be observations on n ***different variables***. For example, in a health sciences study, $Y_{h\ell 1}$ might be systolic blood pressure, $Y_{h\ell 2}$ might be diastolic blood pressure, $Y_{h\ell 3}$ might be total cholesterol level, and so on.

In the most general case, then, hypotheses of parallelism or involving ***averaging*** over the n components of $\boldsymbol{\mu}_\ell$ are ***nonsensical***. Instead, the focus is on comparing the means of each of the components ***simultaneously*** across groups.

That is, the ***null hypothesis*** of central interest is

$$H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_g \tag{3.28}$$

versus the alternative that at least one of the g population mean vectors differs from the others in at least one component.

We first review the standard approach to testing (3.28) and then discuss how ***specialized tests*** relevant to the ***longitudinal*** situation can be developed.

HOTELLING'S T^2***:*** When $g = 2$, the test statistic for testing H_0 in (3.28) can be viewed as a ***generalization*** to multivariate response of the usual two-sample t test for scalar response. Here, (3.28) is

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2. \tag{3.29}$$

Collecting the sample averages $\overline{Y}_{\cdot \ell j}$ for each component $j = 1, \ldots, n$ for group $\ell = 1, \ldots, g$ as

$$\overline{\boldsymbol{Y}}_{\cdot \ell} = \begin{pmatrix} \overline{Y}_{\cdot \ell 1} \\ \vdots \\ \overline{Y}_{\cdot \ell n} \end{pmatrix},$$

the ***sample covariance matrix*** for group ℓ is

$$\widehat{\boldsymbol{V}}_\ell = (r_\ell - 1)^{-1} \sum_{h=1}^{r_\ell} (\boldsymbol{Y}_{h\ell} - \overline{\boldsymbol{Y}}_{\cdot \ell})(\boldsymbol{Y}_{h\ell} - \overline{\boldsymbol{Y}}_{\cdot \ell})^T.$$

The sum in this expression is referred to in the multivariate literature as a ***sum of squares and cross-products*** (***SS&CP***) matrix. Then the (assumed ***common*** across groups) ***overall covariance matrix*** $\boldsymbol{V}$ is estimated by the ***pooled*** estimator as in (2.35); with $g = 2$,

$$\widehat{\boldsymbol{V}} = (r_1 + r_2 - 2)^{-1} \{(r_1 - 1)\widehat{\boldsymbol{V}}_1 + (r_2 - 1)\widehat{\boldsymbol{V}}_2\}.$$

Analogous to the square of the usual t statistic, the ***Hotelling's*** T^2 statistic is

$$T^2 = (r_1^{-1} + r_2^{-1})^{-1} (\overline{\boldsymbol{Y}}_{\cdot 1} - \overline{\boldsymbol{Y}}_{\cdot 2})^T \widehat{\boldsymbol{V}}^{-1} (\overline{\boldsymbol{Y}}_{\cdot 1} - \overline{\boldsymbol{Y}}_{\cdot 2}). \tag{3.30}$$

It can be shown that, under model (3.27),

$$\frac{r_1 + r_2 - n - 1}{(r_1 + r_2 - 2)n} T^2 \sim F_{n, r_1 + r_2 - n - 1}.$$

Thus, the test of H_0 may be carried out at level α by comparing this version of T^2 to $\mathcal{F}_{n, r_1 + r_2 - n - 1, \alpha}$. If $n = 1$, the test reduces to the usual two-sample t test.

As an example, consider the dental study, for which $r_1 = 11$ (girls), $r_2 = 16$ (boys), $n = 4$, and

$$\overline{\boldsymbol{Y}}_{\cdot 1} = (21.182, 22.227, 23.091, 24.091)^T,$$

$$\overline{\boldsymbol{Y}}_{\cdot 2} = (22.875, 23.813, 25.719, 27.469)^T.$$

Using the estimated sample covariance matrices for each group in (2.33) and (2.34) and the resulting pooled estimate, it is straightforward to obtain

$$\frac{r_1 + r_2 - n - 1}{(r_1 + r_2 - 2)n} T^2 = 3.63,$$

which under (3.27) has an F distribution with 4 and 22 degrees of freedom; $\mathcal{F}_{4,22,0.05} = 2.816$, leading to rejection of H_0 at level $\alpha = 0.05$.

Of course, as noted in Section 2.6, the assumption of a common overall pattern of covariance for boys and girls, embodied in this procedure, does not seem to be supported by the data. Moreover, the data support a pattern for each group that, although different for each group, is approximately ***compound symmetric***, suggesting that a more powerful test could be developed.

Regardless, this hypothesis test does not address the questions of interest here. Although the result suggests there is evidence that the overall population means differ between genders, this test offers no insight into ***how*** nor into how the ***pattern of change*** differs, so is relatively ***useless***.

As discussed earlier, defining as in (3.17)

$$\mathcal{M} = \begin{pmatrix} \mu_{11} & \cdots & \mu_{1n} \\ \mu_{21} & \cdots & \mu_{2n} \end{pmatrix}$$

and letting $\boldsymbol{C} = (1, -1)$ and $\boldsymbol{U} = \boldsymbol{I}_n$, H_0 in (3.29) can be expressed as $H_0 : \boldsymbol{C}\mathcal{M}\boldsymbol{U} = \boldsymbol{0}$. We will return to this representation shortly.

ONE-WAY MANOVA: When $g > 2$, a ***multivariate version*** of the usual analysis of variance for a ***one-way layout*** can be constructed and test statistics derived for testing (3.28) as follows. The usual analysis of variance where the response is scalar involves sums of squares and mean squares for "Among Groups" and "Among-Unit Error," and the ratio of the latter yields the usual test statistic. Here, with ***multivariate*** response these are replaced by analogous ***SS&CP matrices*** as follows.

Again let $\overline{Y}_{\cdot\cdot j}$ be the sample average of all observations across all individuals and groups of the jth component of $Y_{h\ell j}$, and define the ***overall sample mean vector***

$$\overline{\boldsymbol{Y}}_{\cdot\cdot} = \begin{pmatrix} \overline{Y}_{\cdot\cdot 1} \\ \vdots \\ \overline{Y}_{\cdot\cdot n} \end{pmatrix}.$$

Then construct the ***MANOVA table*** as

Source	SS&CP	DF
Among Groups	$\boldsymbol{Q}_H = \sum_{\ell=1}^{g} r_\ell(\overline{\boldsymbol{Y}}_{\cdot\ell} - \overline{\boldsymbol{Y}}_{\cdot\cdot})(\overline{\boldsymbol{Y}}_{\cdot\ell} - \overline{\boldsymbol{Y}}_{\cdot\cdot})^T$	$g-1$
Among-unit Error	$\boldsymbol{Q}_E = \sum_{\ell=1}^{g}\sum_{h=1}^{r_\ell}(\boldsymbol{Y}_{h\ell} - \overline{\boldsymbol{Y}}_{\cdot\ell})(\boldsymbol{Y}_{h\ell} - \overline{\boldsymbol{Y}}_{\cdot\ell})^T$	$m-g$
Total	$\boldsymbol{Q}_H + \boldsymbol{Q}_E = \sum_{\ell=1}^{g}\sum_{h=1}^{r_\ell}(\boldsymbol{Y}_{h\ell} - \overline{\boldsymbol{Y}}_{\cdot\cdot})(\boldsymbol{Y}_{h\ell} - \overline{\boldsymbol{Y}}_{\cdot\cdot})^T$	$m-1$

It can be verified that

$$\boldsymbol{Q}_E = (r_1 - 1)\widehat{\boldsymbol{V}}_1 + \cdots + (r_g - 1)\widehat{\boldsymbol{V}}_g,$$

so that

$$\widehat{\boldsymbol{V}} = \boldsymbol{Q}_E/(m-g).$$

Because the entries in the MANOVA table are ***matrices***, it is not straightforward to construct a ***unique generalization*** of the usual analysis of variance F ratio that can be used to test H_0 in (3.28). Clearly, one would like to compare the "***magnitudes***" of the SS&CP matrices $\boldsymbol{Q}_H$ and $\boldsymbol{Q}_E$, but there is no one way to do this. Several statistics have been proposed.

- The most commonly discussed statistic is ***Wilks' lambda***, which can be motivated ***informally*** as follows. Letting SS_G and SS_E be the usual analysis of variance Among-Groups and Among-Unit Error sums of squares, the familiar F ratio is
$$\frac{SS_G/(g-1)}{SS_E/(m-g)}.$$
Thus, in the scalar case, H_0 is rejected when SS_G/SS_E is "large." This is equivalent to rejecting for large values of $1 + SS_G/SS_E$ or small values of
$$\frac{1}{1 + SS_G/SS_E} = \frac{SS_E}{SS_G + SS_E}.$$
For the multivariate problem, the Wilks' lambda statistic is the ***analog*** of this quantity,
$$T_W = \frac{|\boldsymbol{Q}_E|}{|\boldsymbol{Q}_H + \boldsymbol{Q}_E|}.$$
One rejects H_0 for "small" values of T_W.

- The ***Lawley-Hotelling trace*** rejects H_0 for large values of
$$T_{LH} = \text{tr}(\boldsymbol{Q}_H\boldsymbol{Q}_E^{-1}).$$

- Other statistics are ***Pillai's trace*** and ***Roy's greatest root***, which we do not present.

- None of these approaches has been shown to be superior to the others in general. All are ***equivalen***t to using the Hotelling T^2 statistic in the case $g = 2$.

A full discussion of these methods is beyond our scope. For general g and n, the sampling distributions of (functions of) these test statistics may or may not be derived ***exactly***. Thus, except in certain special cases where this is possible (see Johnson and Wichern, 2002), the sampling distributions are ***approximated*** by the F or other distributions.

As in the case of $g = 2$, the hypothesis (3.28) can be expressed in the form $H_0 : \boldsymbol{C}\mathcal{M}\boldsymbol{U} = \boldsymbol{0}$ for appropriate choice of $\boldsymbol{C}$ and $\boldsymbol{U} = \boldsymbol{I}_n$. For example, for $g = 3$,

$$\mathcal{M} = \begin{pmatrix} \mu_{11} & \cdots & \mu_{1n} \\ \mu_{21} & \cdots & \mu_{2n} \\ \mu_{31} & \cdots & \mu_{3n} \end{pmatrix}, \quad \boldsymbol{C} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}.$$

Again, as for $g = 2$, in the ***longitudinal data*** situation, testing (3.28) does not really address the questions of scientific interest, which usually focus on the ***pattern of change***. We now consider how tests of ***parallelism*** (group by time interaction) and ***constancy*** (main effect of time) are developed under model (3.27), so under the assumption of an ***unstructured covariance matrix*** common across groups.

PROFILE ANALYSIS: In the context of ***repeated measurement data*** where the n components of a data vector are repeated observations on the same response, conducting appropriate such ***multivariate tests*** for parallelism and constancy is referred to as ***profile analysis***.

Consider first the hypothesis of ***parallelism*** or ***group by time interaction***. As in the univariate case, under the usual constraints (3.11), this hypothesis is

$$H_0 : \text{ all } (\tau\gamma)_{\ell j} = 0.$$

For $g = 2$ and $n = 3$, we expressed this hypothesis in the form

$$H_0 : \boldsymbol{C}\mathcal{M}\boldsymbol{U} = \boldsymbol{0},$$

with $\boldsymbol{C}$ and $\boldsymbol{U}$ given in (3.19) and (3.20). For general g and n, with $\mathcal{M}$ as in (3.17), it follows that, if $\boldsymbol{1}_p$ denotes a column vector of 1s of length p, then choosing

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{1}_{g-1} & -\boldsymbol{I}_{g-1} \end{pmatrix} \ (g-1 \times g), \quad \boldsymbol{U} = \begin{pmatrix} \boldsymbol{1}_{n-1}^T \\ -\boldsymbol{I}_{n-1} \end{pmatrix} \ (n \times n-1) \tag{3.31}$$

yields the null hypothesis of parallelism.

We now explain informally the reason for writing hypotheses in the form $H_0 : \boldsymbol{C}\mathcal{M}\boldsymbol{U} = \boldsymbol{0}$, which provides the basis for deriving test statistics for the hypotheses of parallelism and constancy. Consider again the hypothesis (3.28), $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_g$, which can be written for general g and n in this form, with $\boldsymbol{C}$ as in (3.31) and $\boldsymbol{U} = \boldsymbol{I}_n$. It can be shown that the SS&CP matrices $\boldsymbol{Q}_H$ and $\boldsymbol{Q}_E$ on which the various test statistics are based can be expressed in a general form in terms of $\boldsymbol{C}$, $\mathcal{M}$, and $\boldsymbol{U}$, as follows.

Let $\boldsymbol{A}$ be the $(m \times q)$ matrix whose ***rows*** each correspond to one of the m individuals in the data set as follows. For individual (h, ℓ), the corresponding row is a $(1 \times q)$ vector of all 0s except for a 1 in the ℓth position; for example, with $g = 3$, for an individual in group 2, the corresponding row is

$$\boldsymbol{a}_2 = (0,\ 1,\ 0),$$

say. Premultiplying the $(g \times n) = (3 \times 4)$ matrix

$$\mathcal{M} = \begin{pmatrix} \mu_{11} & \mu_{12} & \mu_{13} & \mu_{14} \\ \mu_{21} & \mu_{22} & \mu_{23} & \mu_{24} \\ \mu_{31} & \mu_{32} & \mu_{33} & \mu_{34} \end{pmatrix}$$

then yields

$$\boldsymbol{a}_2\mathcal{M} = (\mu_{21}, \mu_{22}, \mu_{23}, \mu_{24}) = \boldsymbol{\mu}_\ell^T.$$

That is, the matrix $\boldsymbol{A}$ "picks off" the mean vector (row of $\mathcal{M}$) corresponding to the group to which each individual belongs; a little thought reveals that, in general, $\boldsymbol{A}$ will have r_1 rows $\boldsymbol{a}_1 = (1, 0, \ldots, 0)$, r_2 rows $\boldsymbol{a}_2 = (0, 1, 0, \ldots, 0)$, $\ldots$, r_g rows $\boldsymbol{a}_g = (0, 0, \ldots, 1)$. It can be deduced that the model $\boldsymbol{Y}_{h\ell} = \boldsymbol{\mu}_\ell + \boldsymbol{\epsilon}_{h\ell}$ can be written succinctly as

$$\mathcal{Y} = \boldsymbol{A}\mathcal{M} + \boldsymbol{\epsilon}, \tag{3.32}$$

where $\mathcal{Y}$ is the $(m \times n)$ matrix with rows $\boldsymbol{Y}_{h\ell}^T$, and similarly for $\boldsymbol{\epsilon}$.

Then it is an exercise in matrix algebra to show that, with $\boldsymbol{C}$ as in (3.31) and $\boldsymbol{U} = \boldsymbol{I}_n$, the SS&CP matrices $\boldsymbol{Q}_H$ and $\boldsymbol{Q}_E$ in the MANOVA table can be written as

$$\boldsymbol{Q}_H = (\boldsymbol{C}\widehat{\mathcal{M}}\boldsymbol{U})'\{\boldsymbol{C}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{C}'\}^{-1}(\boldsymbol{C}\widehat{\mathcal{M}}\boldsymbol{U}), \qquad \boldsymbol{Q}_E = \boldsymbol{U}'\mathcal{Y}'\{\boldsymbol{I}_n - \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\}\mathcal{Y}\boldsymbol{U}, \tag{3.33}$$

where $\widehat{\mathcal{M}} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\mathcal{Y}$.

A technical justification of (3.33) can be found in Vonesh and Chinchilli (1997, p. 50), who show that this representation and the form of the Wilks' lambda statistic T_W can be derived via ***maximum likelihood*** under model (3.32) and the normality assumption (3.27).

It can be shown that test statistics for ***other hypotheses***, such as parallelism and constancy, can be derived by substituting the relevant $\boldsymbol{C}$ and $\boldsymbol{U}$, such as those in (3.31), into (3.33) to yield versions of the SS&CP matrices $\boldsymbol{Q}_H$ and $\boldsymbol{Q}_E$ that can be used to construct versions of any of test statistics above, such as Wilks' lambda T_W, that address the corresponding hypotheses. Depending on g and n, these tests may be exact or approximate.

Classically, ***profile analysis*** has been carried out in practice as follows:

- The test of primary interest in longitudinal settings, that of ***parallelism*** or ***group by time interaction***, is carried out first by taking $\boldsymbol{C}$ and $\mathcal{M}$ to be as in (3.31) and constructing the desired test statistic.
- If the hypothesis of parallelism is ***not rejected***, the test of ***coincidence*** is carried out; this is the usual MANOVA test described above, with $\boldsymbol{C}$ as in (3.31) and $\boldsymbol{U} = \boldsymbol{I}_n$. If the mean response profiles are ***not parallel***, then this test seems difficult to justify. If the profiles ***are parallel***, then this test can be ***refined***. With the additional assumption of parallelism, this is equivalent to a test with $\boldsymbol{C}$ as in (3.31) and $\boldsymbol{U} = \mathbf{1}_n/n$, which leads to exactly the hypothesis corresponding to the ***main effect of group*** in the univariate analyses discussed in Section 3.2. In fact, it can be shown that any of the test statistics discussed above constructed using this $\boldsymbol{C}$ and $\boldsymbol{U}$ reduces to the F statistic for the main effect of group in the univariate analysis.
- If the hypothesis of parallelism is ***not rejected***, the test of ***constancy*** of the mean profiles over time is carried out. It can be shown that this test corresponds to taking $\boldsymbol{U}$ as in (3.31) and $\boldsymbol{C} = \boldsymbol{I}_g$. As with the test for coincidence, if the profiles are ***not parallel***, then testing whether they are ***constant*** over time seems inappropriate. Under the additional assumption of parallelism, a refined test can be constructed corresponding to $\boldsymbol{U}$ as in (3.31) and $\boldsymbol{C} = \mathbf{1}_g^T/g$. Unlike for the refined test of coincidence, the resulting multivariate tests are ***different*** from each other and from the univariate test in Section 3.2.

SUMMARY: Our discussion of multivariate analysis of variance methods has been ***deliberately brief*** because, as hopefully is evident from their formulation, these methods are clearly of ***limited utility*** in longitudinal data analysis. The focus on ***hypothesis testing*** and the assumption of a ***common but unstructured*** overall covariance matrix are key drawbacks. It is important, however, to have a basic understanding of these methods so that the appeal of the more modern methods we discuss henceforth can be fully appreciated.

4 Modern Methods: Preliminaries

4.1 Introduction

In this chapter, we set the stage for study of modern statistical models and methods for longitudinal data analysis that are ***favored*** over the classical repeated measures analysis of variance methods we reviewed in Chapter 3.

First, we recount the ***limitations*** of the classical methods that make them less attractive in practice than modern methods. We then review basic principles of large sample theory that we will use in subsequent chapters to deduce approximate approaches to inference. In particular, we will be interested in the properties of the methods under general conditions. For example, although the methods of Chapter 5 and 6 are derived based on the assumption of ***multivariate normality*** of response vectors, we will deduce the properties of the methods even if normality does not hold. In Chapter 8, we will consider methods that do not make a specific distributional assumption. Accordingly, we briefly review the concept of ***estimating equations*** that define ***estimators*** for parameters in a model of interest and the standard general arguments that lead to approximate ***large sample properties*** of such estimators. We will take this point of view and invoke this type of argument in subsequent chapters to establish approximate large sample properties of estimators for parameters in various longitudinal data models.

4.2 Drawbacks of classical methods

The following is a summary of the major drawbacks of classical methods.

1. BALANCE: Ideally, ***univariate*** repeated measures analysis of variance methods are based on the assumption that each individual is observed at the ***same*** n time points. The multivariate repeated measures analysis of variance methods seem to depend critically on this assumption.

In experiments in settings such as agriculture and manufacturing, this may not be much of a restriction, as investigators can plan and execute experiments carefully and have a good deal of ***control***. Even in this situation, unforeseen mishaps can lead to unobserved or unrecorded observations.

In most application areas, this can be a serious limitation, particularly when the individuals are ***human subjects***. For example, in many health sciences studies, subjects are asked to return to the clinic at specific time points so that the response and other information can be ascertained. However, all subjects do not always return at precisely the time instructed, and some miss visits or, more ominously ***drop out*** of the study altogether. Even if subjects do show up as required, mishaps can also occur in processing lab samples or recording information.

It is thus more often than not ***unrealistic*** to expect the final data set to be ***balanced*** in this sense. "Fixes" such as treating all responses ***within some interval*** of an intended time point as if they all observed at that time point are possible, but are ***ad hoc***, with unknown implications for inference. "Adjusted" approximate F tests that account for imbalance have been proposed. However, it seems more productive to adopt a model framework that ***does not require*** balance, under which principled methods can be developed.

2. FORM OF OVERALL COVARIANCE MATRIX: As we have noted, the univariate methods are predicated on the induced assumption that the overall, aggregate covariance structure of a data vector is ***compound symmetric***. This may be ***too restrictive*** if ***within-individual*** sources of correlation are nonnegligible. Likewise, the multivariate methods assume ***no particular structure*** for the covariance matrix, so allow for overall patterns of covariance that may be ***very unlikely*** to arise in longitudinal data. Models and methods that offer some "middle ground" or allow among- and within-individual sources of correlation to be acknowledged and modeled faithfully would offer more ***flexibility*** to the data analyst.

3. COMMON OVERALL COVARIANCE MATRIX: Both the univariate and multivariate approaches are predicated on the assumption that the overall covariance matrix of a data vector is ***the same*** for all individuals, ***regardless*** of group or any other factor. This may or may not be a reasonable assumption, just as the assumption of ***constant variance*** in classical regression analysis is sometimes violated.

For instance, it is often the case for ***biological phenomena*** that variance ***increases*** as the magnitude of the response increases. This is the case of ***pharmacokinetics***, as in the theophylline pharmacokinetics study in ***EXAMPLE 4*** of Section 1.2. In this setting, ***within-subject*** variance is well known to ***increase*** as with the magnitude of drug concentration, often in a way that appears ***proportional*** to the ***square*** of the within-subject ***inherent trend***. In later chapters, we will characterize this phenomenon formally. Under these conditions, we would expect the ***diagonal elements*** of the overall covariance matrix of a response vector to ***change over time***.

We also saw evidence of violation of this assumption in the dental study data in Chapter 2, where the sample covariance matrices calculated separately by gender and their associated correlation matrices in (2.33) and (2.34) suggested that overall variance and the magnitude of correlations might be ***different*** by gender.

4. INCORPORATION OF COVARIATE INFORMATION: In the classical set-up, an ***among-individual covariate*** is treated as the categorical ***group*** factor, with g levels; as noted in Chapter 3, two or more covariates can be included this way by considering a factorial arrangement. However, it may be of interest to view such a covariate as ***continuous***; for example, in the guinea pig diet study, the dose groups "zero," "low," and "high" might correspond to numerical doses, 0, 100, and 500 μg, say, and interest might be in how mean response changes ***smoothly with dose***. Of course, ***time*** is also treated as categorical, involving the same limitation.

Moreover, it may be relevant to incorporate other (***among-individual***) covariate information. For example, it might be believed that a subject's ***age*** at the start of the study is implicated in his/her later response to treatment (group), suggesting that a relevant model for population mean response should depend on age as well as treatment group. Although univariate and multivariate repeated measures analysis of variance methods can be extended to accommodate this (see Sections 2.4 and 3.4 of Vonesh and Chinchilli, 1997), the way in which such covariates can be incorporated in the statistical model is ***limited***.

Incorporation of covariates that ***change over time***, such as maternal smoking status in the Six Cities study, is in principle also possible in the univariate methods; however, there are ***conceptual*** issues in dealing with such covariates, not limited to these methods, and we discuss these in later chapters.

5. QUESTIONS OF INTEREST: The classical methods emphasize ***hypothesis testing***. However, it is more often than not the case that scientific questions are ***not addressed*** by carrying out a hypothesis test. Investigators often wish to obtain ***estimates*** of meaningful quantities, such as ***rates of change*** and ***differences*** among them, along with appropriate ***measures of uncertainty*** (standard errors and confidence intervals). They might also want to evaluate the extent to which rate of change of mean response over time itself ***changes*** with an individual characteristic like age. Moreover, investigators sometimes wish to make inference about mean responses at times ***other than*** those included in a study.

Clearly, the classical methods are ***too restrictive*** to accommodate these objectives, and a ***more flexible model framework*** is required.

6. NORMALITY: The classical methods are based on the assumption that a data vector has a ***multivariate normal distribution*** with a specific mean structure and the relevant (assumed ***common*** to all individuals) covariance matrix. In the case of the univariate methods, this must hold ***exactly*** to ensure that the test statistics of interest have an F distribution with appropriate degrees of freedom, so that reliable inferences can be drawn. As we have discussed, for outcomes that are ***discrete***, this assumption is clearly inappropriate. Even for ***continuous response***, normality may not be a reasonable representation.

For outcomes that are ***not continuous***, an alternative modeling framework is needed, and we discuss such frameworks in Chapters 7 - 9. For ***continuous*** response, we would like to use methods that yield reliable inferences even if the true distribution of the response is not exactly normal.

The rest of the course is devoted to study of models and associated methods that address these limitations.

4.3 Large sample theory and estimating equations

LARGE SAMPLE THEORY: As is the case with most modern statistical models, the statistical models and methods we discuss in the remainder of the course are sufficiently ***complex*** that it is not possible to derive ***exact results***. In particular, it may not be possible to express ***estimators*** for ***parameters*** involved in the models in a ***closed form***. Rather, the estimators are defined ***implicitly*** as maximizing an ***objective function*** or a solving a set of ***equations***, as discussed momentarily.

Accordingly, it is customary to appeal to ***large sample theory*** to derive ***approximate*** results. Namely, one shows that that estimators are ***consistent*** and ***asymptotically normal***. Consistency provides assurance that, for "large" sample size, an estimator "approaches" the quantity of interest. Asymptotic normality is the basis establishing an approximate ***sampling distribution*** that can be used to derive approximate standard errors, confidence intervals, test procedures, and so on.

Appendix C presents a generic review of concepts and principles of large sample theory.

DATA STRUCTURE, RESTATED: As discussed in Section 2.2, in their most general form, the data are

$$(\boldsymbol{Y}_i, \boldsymbol{z}_i, \boldsymbol{a}_i) = (\boldsymbol{Y}_i, \boldsymbol{x}_i), \quad , i = 1, \ldots, m,$$

where these are ***independent*** across i.

- The response vectors $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$ are of possibly different dimension, with Y_{ij} recorded at time t_{ij}, $j = 1, \ldots, n_i$, on individual i.
- $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}^T, \ldots, \boldsymbol{z}_{in_i}^T)^T$ comprises for individual i a vector of ***within-individual covariates*** $\boldsymbol{u}_i$ describing conditions under which the Y_{ij} were collected on i along with the times t_{ij}.
- $\boldsymbol{a}_i$ is a vector of ***among-individual covariates*** for individual i.
- $\boldsymbol{x}_i = (\boldsymbol{z}_i^T, \boldsymbol{a}_i)^T$ is the full set of covariates associated with individual i.

In the study of methods for longitudinal analysis based on these data, "***large sample***" ordinarily refers to the ***number of individuals*** m being "large," while the numbers of observations per individual, n_i, remain fixed. Thus, large sample approximations are relevant to the situation where individuals are observed at intermittent, possibly prespecified time points, and the sample of individuals available is ***sufficiently large*** to provide "good" information on the population of individuals.

ESTIMATING EQUATIONS: Most of the estimators for parameters in the longitudinal data models we study in subsequent chapters can be expressed as ***solutions*** to sets of equations commonly referred to as ***estimating equations***. Such estimators are cases of a general class of estimators known as ***M-estimators***. Viewing estimators we discuss in subsequent chapters as M-estimators, we are able to deduce properties of estimators even when some model assumptions are ***not correct***.

As background for these developments, we now present a ***generic and nonrigorous*** overview of ***M-estimation*** and ***estimating equations***, to which we refer in later chapters.

M-ESTIMATOR: Let $\mathcal{U}_i$, $i = 1, \ldots, m$, be ***independent*** random vectors with cdf F_i (we may or may not know F_i). Let η ($k \times 1$) be a parameter in a statistical model for the $\mathcal{U}_i$. E.g., if F_i has density p_i, one might specify a model for p_i depending on η. Alternatively, one might specify a model only for some features of F_i, such as a mean and covariance matrix, in terms of a parameter η.

A ***M-estimator*** for η, $\widehat{\eta}$, can be defined two ways:

(1) $\widehat{\eta}$ minimizes $\sum_{i=1}^m \rho_i(\mathcal{U}_i, \eta)$, where $\rho_i(\cdot\,, \cdot)$ are real-valued functions.

(2) $\widehat{\eta}$ is the root of a ($k \times 1$) set of estimating equations such that

$$\sum_{i=1}^{m} \boldsymbol{\Psi}_i(\mathcal{U}_i, \widehat{\eta}) = \boldsymbol{0}, \tag{4.1}$$

where $\boldsymbol{\Psi}_i(\cdot\,, \cdot)$ are vector-valued functions taking values in k-dimensional space.

Ordinarily, $\boldsymbol{\Psi}_i$ satisfies

$$E_\eta\{\boldsymbol{\Psi}_i(\mathcal{U}_i, \eta)\} = \mathbf{0}, \tag{4.2}$$

where E_η refers to expectation under the assumption that the parameter is equal to η.

The subscript *i* may or may not be relevant, depending on the situation:

- If we view $\mathcal{U}_i$ as ***independent and identically distributed*** (***iid***) draws from some joint distribution, then $\rho_i \equiv \rho$ and $\boldsymbol{\Psi}_i \equiv \boldsymbol{\Psi}$. In our context, letting $\mathcal{U}_i = (\boldsymbol{Y}_i^T, \boldsymbol{x}_i^T)^T$, $i = 1, \ldots, m$, this would correspond to the situation where we sample *m* individuals from a population of interest and record all of $\boldsymbol{Y}_i$, $\boldsymbol{u}_i$, and $\boldsymbol{a}_i$.
- If we view parts of $\boldsymbol{x}_i$ as ***fixed constants*** or view the problem ***conditional*** on the $\boldsymbol{x}_i$, then the subscript *i* on ρ_i and $\boldsymbol{\Psi}_i$ is meant to emphasize dependence on such *i*-dependent quantities.

We present the generic argument under the latter condition, but the same ideas apply in the iid case.

- If ρ_i is differentiable with respect to η, then a problem of type (1) implies one of type (2), where $\partial/\partial\boldsymbol{\eta}\, \rho_i(\mathcal{U}_i, \eta) = \boldsymbol{\Psi}_i(\mathcal{U}_i, \eta)$.
- However, a problem of type (2) can be posed without a corresponding problem of type (1), so that these problems are ***more general***.
- If p_i is the assumed density of $\mathcal{U}_i$, then choosing $\rho_i(\cdot, \eta) = \log p_i(\cdot, \eta)$ yields ***maximum likelihood estimation*** under the assumption that p_i is the true density of $\mathcal{U}_i$.

Although we consider longitudinal methods that are motivated by the principles of ***maximum likelihood***, we view all methods from the perspective of a type (2) problem, so that estimators of interest solve ***estimating equations***. This will allow us to evaluate properties even when the assumptions leading to the maximum likelihood formulation do not hold.

UNBIASED ESTIMATING EQUATIONS: Suppose that η_0 is the ***true value*** of η; that is, the value of η such that the assumed model evaluated at η_0 yields the true density (or features of the density) generating the data. In general, if an assumed model depends on a parameter η, and there is a value η_0 such that the model evaluated at η_0 corresponds to (features of) the true distribution generating the data, the model is said to be ***correctly specified***.

Formal arguments regarding ***consistency*** of M-estimators are quite involved. However, it is well known that, if (4.2) holds, under suitable ***regularity conditions***, the estimator $\widehat{\boldsymbol{\eta}}$ solving (4.1) is a ***consistent estimator*** for $\boldsymbol{\eta}_0$ if

$$E\{\boldsymbol{\Psi}_i(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_0)\} = \mathbf{0}, \tag{4.3}$$

where expectation is with respect to the true distribution of the $\boldsymbol{\mathcal{U}}_i$, and $\boldsymbol{\eta}_0$ is the unique value satisfying this requirement.

Estimating equations that satisfy these conditions are referred to as ***unbiased estimating equations***. For our purposes in this course, to ***show consistency*** of an estimator solving a set of estimating equations, it will suffice to note that the estimating equations defining it are ***unbiased***.

APPROXIMATE LARGE SAMPLE DISTRIBUTION OF M-ESTIMATOR: An approximate sampling distribution for $\widehat{\boldsymbol{\eta}}$ can be found via a standard ***Taylor series argument***. We give a ***heuristic sketch***, recognizing that technical conditions are required to validate many of the steps. See Appendix B for a review of Taylor series and notation used here.

Multiplying (4.1) by $m^{-1/2}$, we have

$$\begin{aligned}
\mathbf{0} &= m^{-1/2} \sum_{i=1}^{m} \boldsymbol{\Psi}_i(\boldsymbol{\mathcal{U}}_i, \widehat{\boldsymbol{\eta}}) \\
&= m^{-1/2} \sum_{i=1}^{m} \boldsymbol{\Psi}_i(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_0) + \left\{ m^{-1} \sum_{i=1}^{m} \partial/\partial \boldsymbol{\eta}^T \, \boldsymbol{\Psi}_i(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_*) \right\} m^{1/2}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0),
\end{aligned}$$

where $\boldsymbol{\eta}_*$ is a value between $\widehat{\boldsymbol{\eta}}$ and $\boldsymbol{\eta}_0$, and $\partial/\partial \boldsymbol{\eta}^T \, \boldsymbol{\Psi}_i(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_*)$ is the $(k \times k)$ matrix whose ℓth row is

$$\{\partial/\partial \eta_1 \boldsymbol{\Psi}_{i\ell}(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_*), \ldots, \partial/\partial \eta_k, \boldsymbol{\Psi}_{i\ell}(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_*)\},$$

and $\boldsymbol{\Psi}_{i\ell}(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_*)$ is the ℓth element of $\boldsymbol{\Psi}_i$, $\ell = 1, \ldots, k$.

As $\widehat{\boldsymbol{\eta}}$ is consistent, it is possible to define technical conditions such that $\boldsymbol{\eta}_*$ may be replaced by $\boldsymbol{\eta}_0$ in the partial derivative matrix, and the ***weak law of large numbers*** yields

$$m^{-1} \sum_{i=1}^{m} \partial/\partial \boldsymbol{\eta}^T \boldsymbol{\Psi}_i(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_0) - m^{-1} \sum_{i=1}^{m} E\left\{ \partial/\partial \boldsymbol{\eta}^T \boldsymbol{\Psi}_i(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_0) \right\} \xrightarrow{p} \mathbf{0},$$

where the expectation is with respect to the true distribution of $\boldsymbol{\mathcal{U}}_i$. Thus,

$$\mathbf{0} \approx m^{-1/2} \sum_{i=1}^{m} \boldsymbol{\Psi}_i(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_0) + \left[m^{-1} \sum_{i=1}^{m} E\left\{ \partial/\partial \boldsymbol{\eta}^T \, \boldsymbol{\Psi}_i(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_0) \right\} \right] m^{1/2}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0). \tag{4.4}$$

Assuming that the inverse exists, (4.4) can be rearranged as

$$m^{1/2}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \approx -\left[m^{-1}\sum_{i=1}^{m} E\left\{\partial/\partial\boldsymbol{\eta}^T\, \boldsymbol{\Psi}_i(\mathcal{U}_i, \boldsymbol{\eta}_0)\right\}\right]^{-1} m^{-1/2}\sum_{i=1}^{m}\boldsymbol{\Psi}_i(\mathcal{U}_i, \boldsymbol{\eta}_0). \tag{4.5}$$

Assume that as $m \to \infty$,

$$m^{-1}\sum_{i=1}^{m} E\left\{\partial/\partial\boldsymbol{\eta}^T\, \boldsymbol{\Psi}_i(\mathcal{U}_i, \boldsymbol{\eta}_0)\right\} \to \boldsymbol{A},$$

say, a nonsingular constant matrix. Now apply the ***central limit theorem*** to the rightmost term on the right hand side of (4.5). As each summand has mean **0**, this term ***converges in distribution*** to a ***multivariate normal random vector*** with mean **0** and covariance matrix

$$\lim_{m\to\infty} m^{-1}\sum_{i=1}^{m} E\left\{\boldsymbol{\Psi}_i(\mathcal{U}_i, \boldsymbol{\eta}_0)\boldsymbol{\Psi}_i^T(\mathcal{U}_i, \boldsymbol{\eta}_0)\right\} = \boldsymbol{B},$$

say. ***Slutsky's theorem*** then yields

$$m^{1/2}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \xrightarrow{\mathcal{L}} \mathcal{N}\{\mathbf{0}, \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^T\}. \tag{4.6}$$

The notation in (4.6) is ***shorthand*** for the fact that the expression on the left hand side ***converges in distribution*** to a normal random vector with mean zero and covariance matrix as shown on the right hand side.

We can use the result (4.6) to deduce an ***approximate sampling distribution*** for $\widehat{\boldsymbol{\eta}}$. Define

$$\boldsymbol{A}_m = m^{-1}\sum_{i=1}^{m}\partial/\partial\boldsymbol{\eta}^T\, \boldsymbol{\Psi}_i(\mathcal{U}_i, \boldsymbol{\eta}_0), \quad \boldsymbol{B}_m = m^{-1}\sum_{i=1}^{m}\boldsymbol{\Psi}_i(\mathcal{U}_i, \boldsymbol{\eta}_0)\boldsymbol{\Psi}_i^T(\mathcal{U}_i, \boldsymbol{\eta}_0).$$

Then from (4.6) and using the weak law of large numbers, we have the approximate result

$$\widehat{\boldsymbol{\eta}} \,\dot{\sim}\, \mathcal{N}\{\boldsymbol{\eta}_0, m^{-1}\boldsymbol{A}_m^{-1}\boldsymbol{B}_m(\boldsymbol{A}_m^{-1})^T\}, \tag{4.7}$$

where $\dot{\sim}$ denotes "approximately distributed as." Note that in (4.7), because we have rescaled (4.6), the m^{-1} terms cancel, and the covariance matrix $m^{-1}\boldsymbol{A}_m^{-1}\boldsymbol{B}_m(\boldsymbol{A}_m^{-1})^T$ can be written as

$$\left\{\sum_{i=1}^{m}\partial/\partial\boldsymbol{\eta}^T\, \boldsymbol{\Psi}_i(\mathcal{U}_i, \boldsymbol{\eta}_0)\right\}^{-1}\left\{\sum_{i=1}^{m}\boldsymbol{\Psi}_i(\mathcal{U}_i, \boldsymbol{\eta}_0)\boldsymbol{\Psi}_i^T(\mathcal{U}_i, \boldsymbol{\eta}_0)\right\}\left\{\sum_{i=1}^{m}\partial/\partial\boldsymbol{\eta}^T\, \boldsymbol{\Psi}_i(\mathcal{U}_i, \boldsymbol{\eta}_0)\right\}^{-1\,T}. \tag{4.8}$$

Substituting $\widehat{\boldsymbol{\eta}}$ into the expressions in (4.8), we arrive at the so-called ***sandwich estimator*** or ***robust estimator*** for the covariance matrix of $\widehat{\boldsymbol{\eta}}$, namely

$$\left\{\sum_{i=1}^{m}\partial/\partial\boldsymbol{\eta}^T\, \boldsymbol{\Psi}_i(\mathcal{U}_i, \widehat{\boldsymbol{\eta}})\right\}^{-1}\left\{\sum_{i=1}^{m}\boldsymbol{\Psi}_i(\mathcal{U}_i, \widehat{\boldsymbol{\eta}})\boldsymbol{\Psi}_i^T(\mathcal{U}_i, \widehat{\boldsymbol{\eta}})\right\}\left\{\sum_{i=1}^{m}\partial/\partial\boldsymbol{\eta}^T\, \boldsymbol{\Psi}_i(\mathcal{U}_i, \widehat{\boldsymbol{\eta}})\right\}^{-1\,T}. \tag{4.9}$$

These results can be used to derive approximate (large-m) standard errors and confidence intervals for the components of $\widehat{\boldsymbol{\eta}}$ in the usual way.

REMARKS:

- Note that this argument ***does not require*** a full assumption about the distribution of the $\mathcal{U}_i$. The argument depends only on ***consistency*** of the estimator $\widehat{\boldsymbol{\eta}}$ solving the estimating equations, which holds (under regularity conditions) if the estimating equations are ***unbiased*** as in (4.3) and suitable conditions hold to allow application of the ***weak law of large numbers***, ***Slutsky's theorem*** and the ***central limit theorem***. Such conditions ordinarily involve the existence of ***higher moments*** and ***differentiability*** of functions of the $\mathcal{U}_i$.
- Thus, the results that the sampling distribution of $\widehat{\boldsymbol{\eta}}$ can be approximated for "large" samples by (4.7) and that the covariance matrix of this approximate sampling distribution can be estimated by the ***sandwich estimator*** (4.9) are not predicated on a full distributional assumption for the data. These results hold only if certain ***regularity conditions*** like those above are satisfied.
- Moreover, even if an estimator is derived under a ***full distributional assumption***; e.g., a ***maximum likelihood estimator*** under a full ***parametric distributional assumption*** (like ***normality***), if it can be expressed as the solution to a set of estimating equations, its properties can be evaluated even if that distributional assumption ***does not hold*** exactly.

DEMONSTRATION: In subsequent chapters, we deduce the properties of estimators in various ***statistical models for longitudinal data*** under general conditions by recognizing that the estimators can be formulated as ***solutions to estimating equations***. As a prelude to this development, we demonstrate how the foregoing considerations apply in the familiar situation of ***linear regression***.

To place this in the context of our notation for longitudinal data, suppose we have m individuals, on each of whom a ***single*** scalar response Y_i is recorded, so that $n_i = 1$ for $i = 1, \ldots, m$, along with a vector $(p \times 1)$ of covariates $\boldsymbol{x}_i$, where for convenience in the following argument we take the first element of $\boldsymbol{x}_i$ to be identically equal to 1 for all i; and $m > p$. It is reasonable to assume that the pairs $(Y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, m$, are ***independent*** across i. Here, then, identify $\mathcal{U}_i = (Y_i, \boldsymbol{x}_i)$.

Suppose we ***assume*** that, for each $i = 1, \ldots, m$,

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad E(\epsilon_i | \boldsymbol{x}_i) = 0, \ \text{var}(\epsilon_i | \boldsymbol{x}_i) = \sigma^2, \tag{4.10}$$

so that we equivalently assume that

$$E(Y_i | \boldsymbol{x}_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}, \quad \text{var}(Y_i | \boldsymbol{x}_i) = \sigma^2, \quad i = 1, \ldots, m. \tag{4.11}$$

In (4.10) and (4.11), we ***do not*** make a full ***distributional assumption***. The ***classical assumption*** for model (4.10) is of course that ϵ_i, $i =, \ldots, m$, are ***iid*** , so that they are independent of $\boldsymbol{x}_i$ for each i, and ***furthermore*** that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, so that $Y_i \sim \mathcal{N}(\boldsymbol{x}_i^T\beta, \sigma^2)$, .

Thus, in postulating the model in (4.10) and (4.11), while we are willing to assume ***constant variance*** for all i, we are not willing to assume ***normality***. Instead, we are willing only to make the assumption about the ***first two moments*** of the conditional distribution of Y_i given $\boldsymbol{x}_i$ given in (4.11).

Now suppose that, ***in truth*** , the distribution of Y_i given $\boldsymbol{x}_i$ has first two moments of the form

$$E(Y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i^T\beta, \quad \text{var}(Y_i|\boldsymbol{x}_i) = \sigma^2 g^2(\boldsymbol{x}_i), \quad i = 1, \ldots, m, \tag{4.12}$$

for some function $g(\boldsymbol{x}) > 0$ for all $\boldsymbol{x}$. Thus, ***in truth*** , instead of being as in (4.10), the actual model generating the data is

$$Y_i = \boldsymbol{x}_i^T\beta + \epsilon_i, \quad E(\epsilon_i|\boldsymbol{x}_i) = 0, \ \text{var}(\epsilon_i|\boldsymbol{x}_i) = \sigma^2 g^2(\boldsymbol{x}_i); \tag{4.13}$$

note that the ϵ_i are ***not*** iid. Suppose that β_0 and σ_0^2 are the ***true values*** of β and σ^2 in (4.12) and (4.13) generating the observed data.

Under the ***assumed model*** in (4.10) and (4.11), it is natural to estimate β and σ by the ***ordinary least squares*** (OLS) estimator and the usual ***residual mean square***

$$\widehat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \left(\sum_{i=1}^m \boldsymbol{x}_i\boldsymbol{x}_i^T\right)^{-1} \sum_{i=1}^m \boldsymbol{x}_i Y_i, \quad \widehat{\sigma}^2 = (m-p)^{-1}\sum_{i=1}^m (Y_i - \boldsymbol{x}_i^T\widehat{\beta})^2, \tag{4.14}$$

where $\boldsymbol{X}$ is the (***full rank***) $(m \times p)$ matrix with rows $\boldsymbol{x}_i^T$, $i = 1, \ldots, m$. Of course, even if the true conditional moments of Y_i given $\boldsymbol{x}_i$ are as in (4.12) and (4.13), we can deduce immediately that the OLS estimator $\widehat{\beta}$ is ***consistent***. If we view this situation as ***conditional*** on the $\boldsymbol{x}_i$, then if

$$m^{-1}\sum_{i=1}^m \boldsymbol{x}_i\boldsymbol{x}_i^T = m^{-1}\boldsymbol{X}^T\boldsymbol{X} \ \to \boldsymbol{A}, \tag{4.15}$$

say, the ***weak law of large numbers*** implies

$$m^{-1}\sum_{i=1}^m \boldsymbol{x}_i Y_i - m^{-1}\sum_{i=1}^m \boldsymbol{x}_i(\boldsymbol{x}_i^T\beta_0) \xrightarrow{p} 0,$$

so that

$$m^{-1}\sum_{i=1}^m \boldsymbol{x}_i Y_i \xrightarrow{p} \boldsymbol{A}\beta_0$$

and thus

$$\widehat{\beta} \xrightarrow{p} \boldsymbol{A}^{-1}\boldsymbol{A}\beta_0 = \beta_0.$$

Thus, even though the constant variance assumption ***does not hold*** , the OLS estimator is ***nonetheless*** a ***consistent estimator*** for the true value β_0.

Alternatively, from (4.14), we can write $\widehat{\beta}$ as the solution to the ***estimating equations***

$$\sum_{i=1}^{m}(Y_i - \boldsymbol{x}_i^T\beta)\boldsymbol{x}_i = \boldsymbol{0}, \tag{4.16}$$

the usual ***normal equations***, so that, identifying $\eta = (\beta^T, \sigma^2)^T$,

$$\boldsymbol{\Psi}_i(Y_i, \boldsymbol{x}_i, \eta) = (Y_i - \boldsymbol{x}_i^T\beta)\boldsymbol{x}_i.$$

It is straightforward to observe that, under (4.12) and (4.13), $E_\eta\{\boldsymbol{\Psi}_i(Y_i, \boldsymbol{x}_i, \eta)|\boldsymbol{x}_i\} = \boldsymbol{0}$ so that $E_\eta\{\boldsymbol{\Psi}_i(Y_i, \boldsymbol{x}_i, \eta)\} = \boldsymbol{0}$ for all η, and thus $E\{\boldsymbol{\Psi}_i(Y_i, \boldsymbol{x}_i, \eta_0)\} = \boldsymbol{0}$, so that (4.16) is indeed an ***unbiased estimating equation*** , and consistency of $\widehat{\beta}$ is expected.

We now demonstrate how the large sample distribution of $m^{1/2}(\widehat{\beta} - \beta_0)$ can be derived using the generic estimating equation argument. From (4.16), we have

$$\begin{aligned}\boldsymbol{0} &= m^{-1/2}\sum_{i=1}^{m}(Y_i - \boldsymbol{x}_i^T\widehat{\beta})\boldsymbol{x}_i \\ &= m^{-1/2}\sum_{i=1}^{m}(Y_i - \boldsymbol{x}_i^T\beta_0)\boldsymbol{x}_i - \left(m^{-1}\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^T\right) m^{1/2}(\widehat{\beta} - \beta_0).\end{aligned} \tag{4.17}$$

(It is in fact the case that (4.17) is an ***exact*** equality.) From (4.13), define

$$\delta_i = \frac{Y_i - \boldsymbol{x}_i^T\beta_0}{\sigma_0 g(\boldsymbol{x}_i)}, \quad \text{var}(\delta_i|\boldsymbol{x}_i) = 1.$$

Thus, rearranging (4.17),

$$m^{1/2}(\widehat{\beta} - \beta_0) = \sigma_0\left(m^{-1}\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)^{-1} m^{-1/2}\sum_{i=1}^{m} g(\boldsymbol{x}_i)\boldsymbol{x}_i\delta_i. \tag{4.18}$$

By the ***central limit theorem***, letting $w_i = 1/g(\boldsymbol{x}_i)^2$ and noting that

$$\text{var}\{g(\boldsymbol{x}_i)\boldsymbol{x}_i\delta_i|\boldsymbol{x}_i\} = E\{g^2(\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T\delta_i^2|\boldsymbol{x}_i\} = w_i^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T,$$

$$m^{-1/2}\sum_{i=1}^{m} g(\boldsymbol{x}_i)\boldsymbol{x}_i\delta_i \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \sigma_0^2\boldsymbol{B}), \quad \boldsymbol{B} = \lim_{m\to\infty} m^{-1}\sum_{i=1}^{m} w_i^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T.$$

Using (4.15), we conclude by ***Slutsky's theorem*** from (4.18) that

$$m^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \sigma_0^2\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}). \tag{4.19}$$

Thus, if we let

$$\boldsymbol{W} = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & \cdots & 0 & w_m \end{pmatrix} = \text{diag}(w_1, \ldots, w_m)$$

then

$$m^{-1} \sum_{i=1}^{m} w_i^{-1} \boldsymbol{x}_i \boldsymbol{x}_i^T = m^{-1} \boldsymbol{X}^T \boldsymbol{W}^{-1} \boldsymbol{X},$$

and (4.19) yields

$$\widehat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}\{\boldsymbol{\beta}_0, \sigma_0^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1} (\boldsymbol{X}^T \boldsymbol{W}^{-1} \boldsymbol{X}) (\boldsymbol{X}^T \boldsymbol{X})^{-1}\}, \tag{4.20}$$

where the m^{-1} all cancel when $\widehat{\boldsymbol{\beta}}$ is placed on this scale.

REMARKS:

- Note that, if ***in truth*** $g(\boldsymbol{x}) \equiv 1$ for all $\boldsymbol{x}$, so that $w_i \equiv 1$ for all i, then the assumption of ***constant variance*** is ***correct***, and $\boldsymbol{W} = \boldsymbol{I}_m$, an identity matrix. In this case, (4.20) reduces to

  $$\widehat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}\{\boldsymbol{\beta}_0, \sigma_0^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}\},$$

  which is identical to the familiar exact, classical result when Y_i given $\boldsymbol{x}_i$ is ***normally distributed***. Thus, the exact, ***classical linear model theory*** normal sampling distribution result for the OLS estimator holds ***approximately*** in large samples even if normality of the response does not hold.

- If instead, ***in truth*** $g(\boldsymbol{x})$ depends on $\boldsymbol{x}$, so that w_i are different for each i depending on $\boldsymbol{x}_i$, then (4.20) is an approximate version of the exact linear model theory result when the assumption of ***constant variance does not hold***. Note that this result is ***not immediately useful*** in practice; if we do not know $g(\cdot)$, we cannot calculate the matrix $\boldsymbol{W}$, and, as discussed next, $\widehat{\sigma}^2$ will not be a consistent estimator for σ_0^2. We take up the analogous issue to this in the general longitudinal data setting in the next chapter.

- We can also consider the OLS estimator $\widehat{\sigma}^2$ for σ^2 in (4.14) from the point of view of ***estimating equations***. Recall from classical linear model theory that division here is by $(m-p)$ rather than m to achieve an exactly ***unbiased*** estimator with finite sample size m in the case of ***constant variance***. Because $m/(m-p) \to 1$ as $m \to \infty$, consider instead the estimator

  $$m^{-1} \sum_{i=1}^{m} (Y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}})^2.$$

This estimator can be written as the solution to the ***estimating equation***

$$\sum_{i=1}^{m}\{(Y_i - \boldsymbol{x}_i^T\beta)^2 - \sigma^2\} = 0, \tag{4.21}$$

jointly with the equation (4.16). If, ***in truth***, constant variance does hold, so that $g(\boldsymbol{x}) \equiv 1$ for all $\boldsymbol{x}$, then it is clear that (4.21) is an ***unbiased estimating equation***, so we expect that $\widehat{\sigma}^2$ is a consistent estimator for σ_0^2 in this case.

However, if this assumption is ***incorrect***, and $\text{var}(Y_i|\boldsymbol{x}_i) = \sigma^2 g^2(\boldsymbol{x}_i)$ as in (4.12) and (4.13), then it is straightforward that

$$E_\eta\{(Y_i - \boldsymbol{x}_i^T\beta)^2|\boldsymbol{x}_i\} = \sigma^2 g^2(\boldsymbol{x}_i),$$

and thus the summand in the estimating equation (4.21) ***does not*** have conditional (on $\boldsymbol{x}_i$) expectation 0, so that the estimating equation is not ***unbiased***. Intuitively, the meaning of the parameter σ^2 when the true variance is not constant is ***different from*** that when true variance is constant, so this is not surprising.

SUMMARY: In subsequent chapters, we derive large sample results using generalizations of this argument, without making ***full distributional assumptions*** such as normality on the conditional distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ but rather making assumptions only on conditional moments. Thus, the results will be broadly applicable as long as the number of individuals m is not too small.

5 Population-Averaged Linear Models for Continuous Response

5.1 Introduction

We begin our discussion of modern models and methods for longitudinal data analysis by considering a general class of models and associated methods for ***continuous response*** that arises from taking a ***population-averaged*** perspective. This class of models addresses all of the ***drawbacks*** of classical models and methods summarized in Section 4.2.

Namely, models of this type do not require the data set to be ***balanced***; i.e., the elements of the response vectors $\boldsymbol{Y}_i$ ***do not*** need to be observations taken at the the same n time points. In addition, the model framework allows a very general specification for the form of the ***overall aggregate covariance matrix*** of a data vector and allows it to differ depending on, for instance, the values of ***covariates***.

The ***population mean response*** is represented by a ***linear model*** that allows ***among-*** and ***within-individual*** covariates to be incorporated straightforwardly and involves ***parameters*** that characterize features of the population mean response, such as ***patterns of change*** exhibited over time, and how these features might be associated with ***among-individual covariates***.

Finally, although the model incorporates an assumption of ***multivariate normality*** of a response vector ***conditional on covariates***, using ***large sample*** (large m) arguments, as long as m is large enough and the model for the population mean response is ***correctly specified*** , it is possible to show that ***estimators*** of parameters in the models are ***consistent*** for the true values and to deduce an approximate ***normal sampling distribution*** for them, even if the true distribution of the response is ***not normal***. The approximate sampling distribution then forms the basis for ***inferential goals*** such as assessments of uncertainty and hypothesis testing procedures.

Moreover, as we demonstrate, ***even if*** the representation for the overall pattern of covariance is ***not correctly specified*** , estimators for parameters in a correctly specified population mean response model are ***still consistent*** , and an ***approximate sampling distribution*** can be derived.

5.2 Model specification

BASIC MODEL: Recall again that the observed data are

$$(\boldsymbol{Y}_i, \boldsymbol{z}_i, \boldsymbol{a}_i) = (\boldsymbol{Y}_i, \boldsymbol{x}_i), \quad , i = 1, \ldots, m,$$

independent across i, where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$, with Y_{ij} recorded at time t_{ij}, $j = 1, \ldots, n_i$ (possibly different times for different individuals); $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}^T, \ldots, \boldsymbol{z}_{in_i}^T)^T$ comprising ***within-individual*** covariate information $\boldsymbol{u}_i$ and the t_{ij}; $\boldsymbol{a}_i$ is a vector of ***among-individual*** covariates; and $\boldsymbol{x}_i = (\boldsymbol{z}_i^T, \boldsymbol{a}_i^T)^T$.

The ***population-averaged linear model*** we study in this chapter is most relevant when the responses Y_{ij} are ***continuous***. The model is written as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \epsilon_i, \quad i = 1, \ldots, m. \tag{5.1}$$

- In (5.1), $\boldsymbol{X}_i$ is a ***design matrix*** for individual i depending on individual i's ***covariates*** $\boldsymbol{x}_i$, examples of which we present momentarily.
- The ***deviation*** $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^T$ is such that

  $$E(\epsilon_i|\boldsymbol{x}_i) = \boldsymbol{0}, \quad \text{var}(\epsilon_i|\boldsymbol{x}_i) = \boldsymbol{V}_i = \boldsymbol{V}_i(\xi, \boldsymbol{x}_i), \tag{5.2}$$

  where $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ $(n_i \times n_i)$ can depend on the covariates $\boldsymbol{x}_i$ and on a vector of ***covariance parameters*** ξ, which includes ***correlation parameters*** α $(s \times 1)$ and ***variance parameters*** θ $(r \times 1)$. We discuss examples shortly. We sometimes suppress this dependence for brevity and simply write $\boldsymbol{V}_i$.
- The form of $\boldsymbol{V}_i$ is specified ***by the data analyst*** in accordance with the features of the given situation. Because of the dependence of $\boldsymbol{V}_i$ on covariates, there is no requirement, for example, that the form of the covariance matrix be the same for all individuals. We elaborate on this point in the examples below.
- Ordinarily, it is assumed that the ***conditional distribution*** of ϵ_i given $\boldsymbol{x}_i$ is ***multivariate normal*** ,

  $$\epsilon_i|\boldsymbol{x}_i \sim \mathcal{N}\{\boldsymbol{0}, \boldsymbol{V}_i(\xi, \boldsymbol{x}_i)\}, \tag{5.3}$$

  sometimes written more briefly as $\epsilon_i|\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}_i)$.

- $\boldsymbol{\beta}$ is a vector of ***parameters*** characterizing the population mean response; that is, with the assumption on ϵ_i in (5.2), we have that

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{X}_i\boldsymbol{\beta} \quad (n_i \times 1), \tag{5.4}$$

  representing the population mean response for individual *i*, or indeed any individual in the population with covariates $\boldsymbol{x}_i$.

- From (5.2), it follows that

$$\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_i = \boldsymbol{V}_i(\boldsymbol{\xi}, \boldsymbol{x}_i) \quad (n_i \times n_i), \tag{5.5}$$

  the overall population covariance matrix for an individual with covariates $\boldsymbol{x}_i$, characterizing the ***aggregate pattern of covariance*** combining among- and within-individual sources for such an individual.

- With the normality assumption (5.3), the model can be written succinctly as

$$\boldsymbol{Y}_i|\boldsymbol{x}_i \sim \mathcal{N}\{\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i(\boldsymbol{\xi}, \boldsymbol{x}_i)\}, \quad i = 1, \ldots, m, \tag{5.6}$$

  which we often abbreviate as

$$\boldsymbol{Y}_i|\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i), \quad i = 1, \ldots, m.$$

REPRESENTATION OF COVARIANCE MATRIX: To facilitate thinking about models $\boldsymbol{V}_i(\boldsymbol{\xi}, \boldsymbol{x}_i)$, it is sometimes convenient to represent this covariance matrix as a ***product*** of "***standard deviation matrices***" and a ***correlation matrix***. Let $\boldsymbol{T}_i(\theta, \boldsymbol{x}_i)$ be the $(n_i \times n_i)$ ***diagonal matrix*** whose diagonal elements are models for $\text{var}(Y_{ij}|\boldsymbol{x}_i)$, depending on a parameter θ as above. Let $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$ be a $(n_i \times n_i)$ correlation matrix, depending on a parameter α. Then it is straightforward to deduce (try it) that a model for the overall covariance structure can be obtained as

$$\boldsymbol{V}_i(\boldsymbol{\xi}, \boldsymbol{x}_i) = \boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i)\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)\boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i), \quad \boldsymbol{\xi} = (\theta^T, \alpha^T)^T, \tag{5.7}$$

where $\boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i)$ is the matrix whose diagonal elements are the models for the ***standard deviations*** $\{\text{var}(Y_{ij}|\boldsymbol{x}_i)\}^{1/2}$. Clearly, $\boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i)\boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i) = \boldsymbol{T}_i(\theta, \boldsymbol{x}_i)$. We sometimes write $\boldsymbol{T}_i$ and $\boldsymbol{\Gamma}_i$ for brevity, suppressing dependence on θ, α, and $\boldsymbol{x}_i$.

The representation (5.7) allow features of overall variance and the overall pattern of correlation to be thought of ***separately***. That is, one can entertain models for correlation structure and beliefs about variance separately to arrive at an overall specification. We demonstrate in examples below.

MODEL SUMMARY: It is often convenient to summarize the model as follows. Recall that the total number of observations Y_{ij} is $N = \sum_{i=1}^{m} n_i$. Define

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_m \end{pmatrix} \quad (N \times 1), \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_m \end{pmatrix} \quad (N \times p), \quad \boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_m \end{pmatrix} \quad (N \times 1). \tag{5.8}$$

Then (5.1) can be expressed compactly as (try it)

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}. \tag{5.9}$$

It follows from (5.2) that

$$E(\boldsymbol{\epsilon}|\tilde{\boldsymbol{x}}) = \boldsymbol{0},$$

where $\tilde{\boldsymbol{x}}$ is the collection of all covariates $\boldsymbol{x}_i$, $i = 1 \ldots, m$, for all m individuals, so that, from (5.4),

$$E(\boldsymbol{Y}|\tilde{\boldsymbol{x}}) = \boldsymbol{X}\beta.$$

Define the ***block diagonal*** matrix

$$\boldsymbol{V}(\xi, \tilde{\boldsymbol{x}}) = \begin{pmatrix} \boldsymbol{V}_1(\xi, \boldsymbol{x}_1) & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_2(\xi, \boldsymbol{x}_2) & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{V}_m(\xi, \boldsymbol{x}_m) \end{pmatrix} \quad (N \times N). \tag{5.10}$$

We often write (5.10) for brevity as

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{V}_m \end{pmatrix} \quad (N \times N). \tag{5.11}$$

Then, from (5.2) and (5.5), defining similarly $\boldsymbol{T}(\theta, \tilde{\boldsymbol{x}}) = \boldsymbol{T}$ and $\boldsymbol{\Gamma}(\alpha, \tilde{\boldsymbol{x}}) = \boldsymbol{\Gamma}$ $(N \times N)$,

$$\text{var}(\boldsymbol{\epsilon}|\tilde{\boldsymbol{x}}) = \text{var}(\boldsymbol{Y}|\tilde{\boldsymbol{x}}) = \boldsymbol{V}(\xi, \tilde{\boldsymbol{x}}) = \boldsymbol{V} = \boldsymbol{T}^{1/2}\boldsymbol{\Gamma}\boldsymbol{T}^{1/2},$$

which follows by the ***independence*** of $\boldsymbol{\epsilon}_i$ (and $\boldsymbol{Y}_i$) for $i = 1, \ldots, m$.

Note that $\boldsymbol{V}$ in (5.11) has a different definition from that in Chapter 3. Henceforth, we use the symbol $\boldsymbol{V}$ in this way to represent the covariance matrix of the "***stacked***" random vectors $\boldsymbol{\epsilon}$ and $\boldsymbol{Y}$ (conditional on the $\boldsymbol{x}_i$).

The model (5.6) can then be summarized by

$$\boldsymbol{Y}|\tilde{\boldsymbol{x}} \sim \mathcal{N}\{\boldsymbol{X}\beta, \boldsymbol{V}(\xi, \tilde{\boldsymbol{x}})\}, \tag{5.12}$$

or, briefly,

$$\boldsymbol{Y}|\tilde{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{X}\beta, \boldsymbol{V}). \tag{5.13}$$

- In the literature on longitudinal data analysis and in ***software documentation***, it is common to write the model incorporating the normality assumption using this "stacked" notation, suppressing dependence of the overall covariance matrix on parameters and covariate information; that is, as in (5.13).
- We use the more detailed notation (5.12) when we wish to emphasize ***explicitly*** the dependence of the covariance matrix on parameters and covariates.

REMARK: Recall that $\boldsymbol{x}_i$ for individual i includes the ***times*** t_{ij}, $= 1, \ldots, n_i$, at which i was observed, which, technically, are not "***covariates***" in the strict sense, although they often play the role of "***covariates***" as far as implementation is concerned. Thus, conditioning on $\boldsymbol{x}_i$ is really meant to imply conditioning on all ***among-*** and ***within-individual covariates***.

We now demonstrate features of the ***population-averaged linear model*** and its interpretation by considering its specification in several examples.

EXAMPLE 1, DENTAL STUDY: We have already considered a population-averaged model for these data in Section 2.4. Recall that there is one among-individual covariate, gender, which we represented for child i as $g_i = 0$ if i is a girl and $g_i = 1$ if i is a boy, so that $\boldsymbol{a}_i = g_i$; there are no within-individual covariates $\boldsymbol{u}_i$. The response was measured for all $m = 27$ children at ages $(t_1, \ldots, t_4) = (8, 10, 12, 14)$. Thus, $\boldsymbol{z}_{ij} = t_j$ for all i, and $\boldsymbol{x}_i$ contains g_i (and the four time points). Thus, conditioning on covariates $\boldsymbol{x}_i$ corresponds to conditioning on gender.

From a ***population-averaged*** perspective, the primary question of interest is whether or not the ***rate of change*** of the population mean response profile for boys differs from that for girls. In (2.22), we specified a model for the population mean at time t_{ij} for a child of gender g_i as

$$E(Y_{ij}|\boldsymbol{x}_i) = \{\beta_{0,B} g_i + \beta_{0,G}(1 - g_i)\} + \{\beta_{1,B} g_i + \beta_{1,G}(1 - g_i)\} t_{ij}, \tag{5.14}$$

so that $\beta_{1,G}$ and $\beta_{1,B}$ are the ***slopes*** of the ***assumed straight line*** population mean profiles for girls and boys, respectively. Interest is in comparing $\beta_{1,G}$ and $\beta_{1,B}$.

Thus,

$$\boldsymbol{\beta} = (\beta_{0,G}, \beta_{1,G}, \beta_{0,B}, \beta_{1,B})^T,$$

$p = 4$, and, for child i,

$$\boldsymbol{X}_i = \begin{pmatrix} (1-g_i) & (1-g_i)t_1 & g_i & g_i t_1 \\ \vdots & \vdots & \vdots & \vdots \\ (1-g_i) & (1-g_i)t_4 & g_i & g_i t_4 \end{pmatrix}, \tag{5.15}$$

so that

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & t_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_4 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{X}_i = \begin{pmatrix} 0 & 0 & 1 & t_1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & t_4 \end{pmatrix}, \tag{5.16}$$

if i is a girl or boy, respectively.

Clearly, $\boldsymbol{X}_i$ in (5.15) is ***not of full rank*** for any i. Intuitively, this reflects that fact that a boy does not provide information on parameters describing the population mean for girls, and vice versa. However, it is straightforward to observe that the "***stacked design matrix***" $\boldsymbol{X}$ in (5.8), has ***full column rank*** $p = 4$, as it comprises 11 matrices $\boldsymbol{X}_i$ like that on the left hand side of (5.16) stacked on top of 16 like that on the right hand side; the $p = 4$ columns of $\boldsymbol{X}$ are clearly ***linearly independent*** (check). This demonstrates that the problem of making inference on β is feasible from data like those in the study, involving children of both genders.

To complete the model, we specify a model $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ for the ***overall pattern of covariance*** $\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i)$. Because these data are ***balanced***, it was straightforward to calculate the sample overall covariance matrices and their associated correlation matrices for each gender in Section 2.6. Recall that the numerical estimates in (2.33) and (2.34) suggest the following:

- Overall variance is likely ***constant over time*** for each gender, but the variance estimates are ***larger*** for boys than for girls. Formally, the data suggest that $\text{var}(Y_{ij}|\boldsymbol{x}_i)$ for boys and girls are ***the same*** for all j but that for boys is larger. Thus, recognizing that conditioning on $\boldsymbol{x}_i$ is really conditioning on g_i, a reasonable model is

  $$\text{var}(Y_{ij}|g_i = 0) = \sigma_G^2, \quad \text{var}(Y_{ij}|g_i = 1) = \sigma_B^2. \tag{5.17}$$

  The specification in (5.17) can be represented by taking $\boldsymbol{T}_i(\theta, \boldsymbol{x}_i)$ to be the diagonal matrix with diagonal elements all equal to

  $$\sigma_G^2(1-g_i) + g_i\sigma_B^2,$$

  or, equivalently,

  $$\boldsymbol{T}_i(\theta, \boldsymbol{x}_i) = \{\sigma_G^2(1-g_i) + g_i\sigma_B^2\}\boldsymbol{I}_4, \quad \theta = (\sigma_G^2, \sigma_B^2)^T.$$

- The ages are ***equally-spaced*** in time, so any model that is reasonable under this condition is possible. The empirical evidence suggests that, for each gender, the overall pattern of correlation is approximately ***compound symmetric*** with a ***different*** correlation parameter α in (2.25) for each gender. That is,

$$\boldsymbol{\Gamma}_i(\boldsymbol{\alpha}, \boldsymbol{x}_i) = [1 - \{(1 - g_i)\alpha_G + g_i\alpha_B\}]\boldsymbol{I}_4 + \{(1 - g_i)\alpha_G + g_i\alpha_B\}\boldsymbol{J}_4,$$

  where thus $\boldsymbol{\alpha} = (\alpha_G, \alpha_B)^T$.

Combining the above, the suggested covariance model is

$$\sigma^2_G \begin{pmatrix} 1 & \alpha_G & \cdots & \alpha_G \\ \alpha_G & 1 & \cdots & \alpha_G \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_G & \cdots & \alpha_G & 1 \end{pmatrix} = \sigma^2_G\{(1 - \alpha_G)\boldsymbol{I}_4 + \alpha_G\boldsymbol{J}_4\} \tag{5.18}$$

for girls, and

$$\sigma^2_B\{(1 - \alpha_B)\boldsymbol{I}_4 + \alpha_B\boldsymbol{J}_4\} \tag{5.19}$$

for boys. The ***covariance parameter*** $\boldsymbol{\xi}$ characterizing $\boldsymbol{V}_i$ is then $\boldsymbol{\xi} = (\sigma^2_G, \sigma^2_B, \alpha_G, \alpha_B)^T$.

ALTERNATIVE PARAMETERIZATION: As with any ***linear model*** , it is possible to represent the population mean response model (5.14) using a ***different parameterization***. Because interest focuses on the ***difference in slopes*** characterizing the rates of change of population mean dental distance for boys and girls, it is natural to express the population mean ***directly*** in terms of a parameter representing this difference. Thus, an equivalent alternative to (5.14) is

$$E(Y_{ij}|\boldsymbol{x}_i) = \{\beta_{0,G} + \beta_{0,B-G}g_i\} + \{\beta_{1,G} + \beta_{1,B-G}g_i\}t_{ij}. \tag{5.20}$$

In (5.20), $\beta_{0,B-G}$ and $\beta_{1,B-G}$ represent the ***difference*** in intercept and slope between boys and girls and will be positive if that for boys exceeds that for girls. Moreover, for example, the slope of the population mean response for boys is then $\beta_{1,B-G} + \beta_G$, and similarly for intercept.

REMARK: The population mean response model in (5.14) or (5.20) in no way requires the time points t_{ij} to be the same for each child. Even if these data were ***not balanced*** , there would be ***no problem*** specifying such a model. Specification of the covariance model when data are not balanced does require some special consideration; we discuss this shortly.

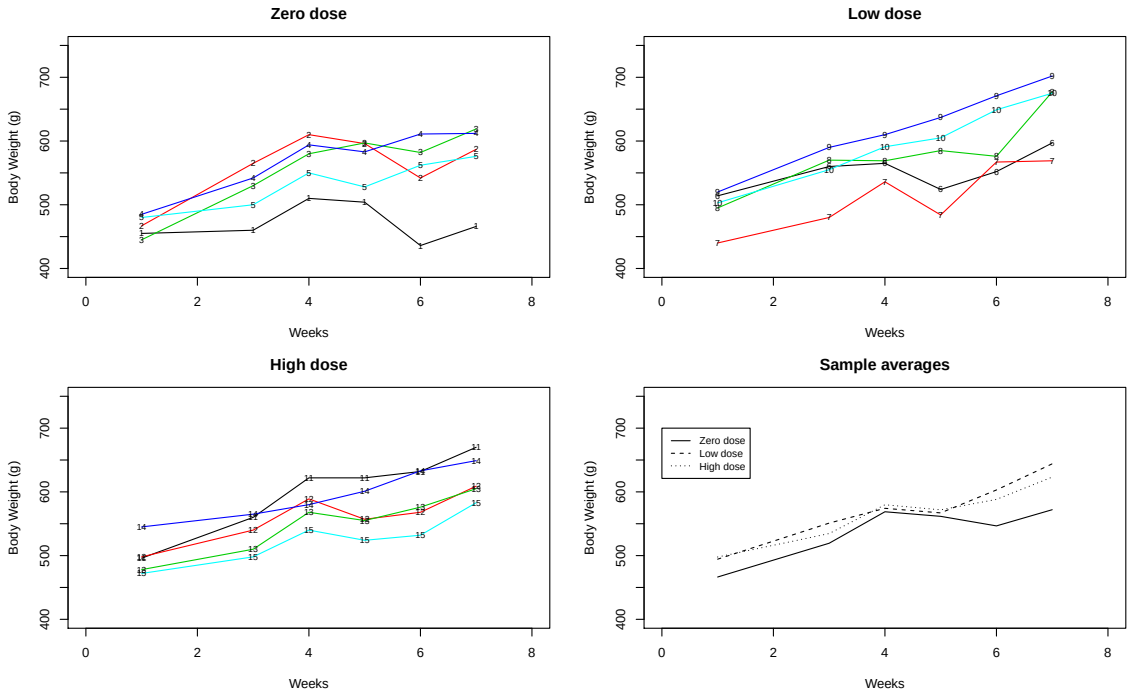

Figure 5.1: *Growth of guinea pigs receiving different doses of vitamin E diet supplement.*

EXAMPLE 2, GUINEA PIG DIET STUDY: The same considerations apply to specification of a population-averaged model for these data, which are also ***balanced***. We discuss specification of a model for population mean response, which illustrates some key issues.

Recall from Section 1.2 that 15 guinea pigs were given a growth-inhibiting substance at ***baseline*** (time 0, beginning of the first week). At weeks 1, 3, and 4, body weight we measured. Immediately after the week 4 measurement (so at the start of week 5), the pigs were ***randomized*** to receive zero, low, or high dose of vitamin E, 5 pigs per group, and body weight was subsequently recorded subsequently at weeks 5, 6, and 7. Thus all $m = 15$ pigs were observed at times $(t_1, \ldots, t_6) = (1, 3, 4, 5, 6, 7)$, so $\boldsymbol{z}_{ij} = t_j$ for all i, and $\boldsymbol{a}_i$ is the ***among-individual covariate*** dose group, with three possible values, which can be represented as $\boldsymbol{a}_i = (d_{i1}, d_{i2}, d_{i3})^T$, where $d_{i\ell} = 1$ if pig i was randomized to dose group ℓ and $= 0$ otherwise, where $\ell = 1, 2, 3$ correspond to zero, low, and high dose.

We reproduce Figure 1.3 from Chapter 1 for convenience as Figure 5.1.

Because the pigs were treated ***identically*** until the end of week 4, a reasonable model for population mean response takes it to be ***identical*** for pigs in all three groups through week 4. Because pigs were then ***randomized*** at this time to receive one of the three doses, a model should allow the population mean response profile to be potentially ***different*** for each dose group henceforth.

That is, a plausible population mean model has two "***phases***," before and after introduction of vitamin E, where the second "***phase***" is different for each group.

From the plot of ***sample averages*** over time in Figure 5.1, a model that takes each of these phases to be a ***straight line*** is reasonable, where the intercept and slope of the first phase is the same for all groups. A model that incorporates these features is the ***linear spline*** model

$$E(Y_{ij}|\boldsymbol{x}_i) = \beta_0 + \beta_1 t_{ij} + \sum_{\ell=1}^{3} \beta_{2\ell} d_{i\ell}(t_{ij} - 4)_+ \tag{5.21}$$

with a ***knot*** at week 4, where

$$\begin{aligned} x_+ &= x \quad \text{if } x \geq 0, \\ &= 0 \quad \text{if } x < 0. \end{aligned}$$

From (5.21), for any pig, population mean response follows the straight line

$$\beta_0 + \beta_1 t$$

through week $t = 4$. For $t \geq 4$, for a pig in group ℓ, population mean response is represented as

$$\beta_0 + \beta_1 t_{ij} + \beta_{2\ell}(t - 4) = \{\beta_0 + \beta_1(4)\} + (\beta_1 + \beta_{2\ell})(t - 4),$$

so that, with $t = 4$ as the "***origin***," population mean weight follows a straight line for $t \geq 4$ with "***intercept***" (value at $t = 4$ when the dose was administered) $\beta_0 + \beta_1(4)$ and ***slope*** $\beta_1 + \beta_{2\ell}$.

Differences in population mean response trajectory are reflected in (5.21) by differences among the $\beta_{2\ell}$, $\ell = 1, 2, 3$. The model (5.21) could of course be parameterized in ***alternative*** ways. The model allows the possibility that the population mean profile for the zero dose group ***changes*** after week 4, even though the pigs in this group did not receive vitamin E. If there were reason to believe that the population mean trajectory for pigs not receiving vitamin E before week 4 should ***continue*** after week 4, a modification of the model would be to take $\beta_{21} = 0$ in (5.21); however, the visual evidence in Figure 1.3 does not support this. Perhaps the effect of the growth-inhibiting substance begins to manifest at week 4, leading to a downward trend, but the addition of vitamin E mitigates this effect.

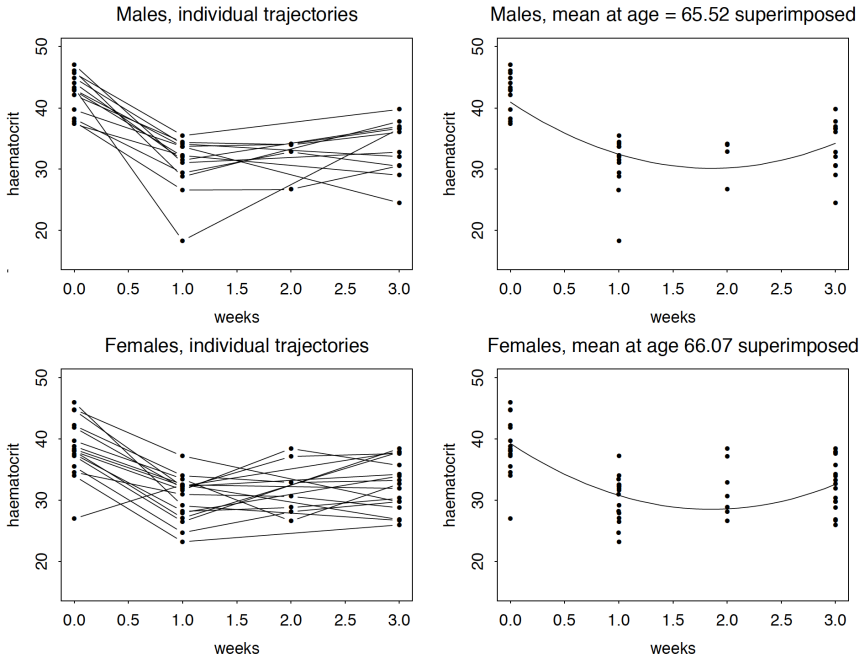

Figure 5.2: *Hæmatocrit trajectories for hip replacement patients. The left hand panels show individual profiles by gender; the right hand panels show a fitted quadratic model for the mean superimposed.*

HIP REPLACEMENT STUDY: These data are adapted from Crowder and Hand (1990, Section 5.2). Thirty patients underwent hip replacement surgery, 13 males and 15 females. Hæmatocrit, the ratio of volume packed red blood cells relative to volume of whole blood recorded as a percentage, was planned to be measured on each patient at ***baseline***, week 0, prior to surgery, and then at weeks 1, 2, and 3 post-surgery. In addition to gender, ***age*** of each patient was also recorded. The data are shown in Figure 5.2.

The primary objectives are to determine if there are differences in the ***population mean pattern of change*** of hæmatocrit following surgery between genders and to characterize the patterns of change.

It is evident in the left hand panels of Figure 5.2 that several patients of both genders are ***missing*** the measurement at week 2; there is also female who is missing both this and the baseline measurement. Crowder and Hand do not offer an explanation; because this is so ***systematic***, occurring for about half of the male and half of the female patients, it is plausible that these observations are missing for reasons having ***nothing*** to do with the health status of the patients but rather might reflect, for example, failure of the equipment used ascertain hæmatocrit values during week 2. We downplay this complication for now and return to the issue of ***missing responses*** later in this chapter.

These data exemplify the common situation where, although it was ***planned*** to record the response at $n = 4$ prespecified times (0,1,2,3 weeks), not all individuals have all responses recorded, so that n_i varies with i, although those that are available are at the prespecified times. That is, $n_i = 4$ for some patients, for whom $t_{ij} = 0, 1, 2, 3$ for $j = 1, \ldots, 4$; $n_i = 3$ for those missing the week 2 measurement, so that $t_{ij} = 0, 1, 3$, $j = 1, \ldots, 3$; and $n_i = 2$ for the female patient missing the baseline and week 2 responses, so that $t_{ij} = 1, 3$, $j = 1, 2$. For patient i, $\boldsymbol{z}_{ij} = t_{ij}$, $j = 1, \ldots, n_i$, and $\boldsymbol{a}_i = (g_i, a_i)^T$, where gender $g_i = 0$ for females and $g_i = 1$ for males; and a_i is the age of the patient (years), ranging from 47 to 79 for females (sample average 66.07) and 44 to 74 for males (65.52).

For both genders, Figure 5.2 shows that hæmatocrit ***drops*** from baseline after surgery and then begins to ***rebound*** over the 3 weeks post-surgery. This suggests that the following ***quadratic*** model for population mean is reasonable, which allows the pattern to ***differ*** between genders:

$$E(Y_{ij}|\boldsymbol{x}_i) = \{\beta_{0,F}(1 - g_i) + \beta_{0,M}g_i\} + \{\beta_{1,F}(1 - g_i) + \beta_{1,M}g_i\}t_{ij} + \{\beta_{2,F}(1 - g_i) + \beta_{2,M}g_i\}t_{ij}^2. \quad (5.22)$$

The basic model (5.23) can be modified to incorporate the possibility that features of the mean response are ***age-dependent*** ; for example,

$$\begin{aligned} E(Y_{ij}|\boldsymbol{x}_i) &= \{\beta_{0,F}(1 - g_i) + \beta_{0,M}g_i\} + \{\beta_{3,F}(1 - g_i) + \beta_{3,M}g_i\}a_i \\ &\quad + \{\beta_{1,F}(1 - g_i) + \beta_{1,M}g_i\}t_{ij} + \{\beta_{2,F}(1 - g_i) + \beta_{2,M}g_i\}t_{ij}^2. \end{aligned} \quad (5.23)$$

allows ***mean hæmatocrit at baseline*** depend on patient age of patient in a way that is different for each gender. The ***linear*** and ***quadratic*** effects that govern the pattern of change post-baseline could be modified similarly, and any of these models could be***reparameterized*** in terms of parameters representing the ***differences*** in intercept and linear and quadratic effects between genders.

Plausible models $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ for the ***overall pattern of covariance*** include those that are suited to what are ideally ***balanced*** data with ***equally-spaced*** time points; however, fitting of such models requires that the ***missing values*** for some patients be taken into account appropriately. We discuss this shortly.

HIV CLINICAL TRIAL: These data are reported in Fitzmaurice, Laird, and Ware (2011) and are from a randomized, double-blind clinical trial, AIDS Clinical Trials Group (ACTG) Study 193A, in patients infected with ***human immunodeficiency virus*** (HIV) exhibiting advanced immune suppression; i.e., CD4 T-cell counts $\leq$ 50 cells/mm^3. CD4 count is a standard measure reflecting the status of the ***immune system*** , which is compromised in patients with HIV infection.

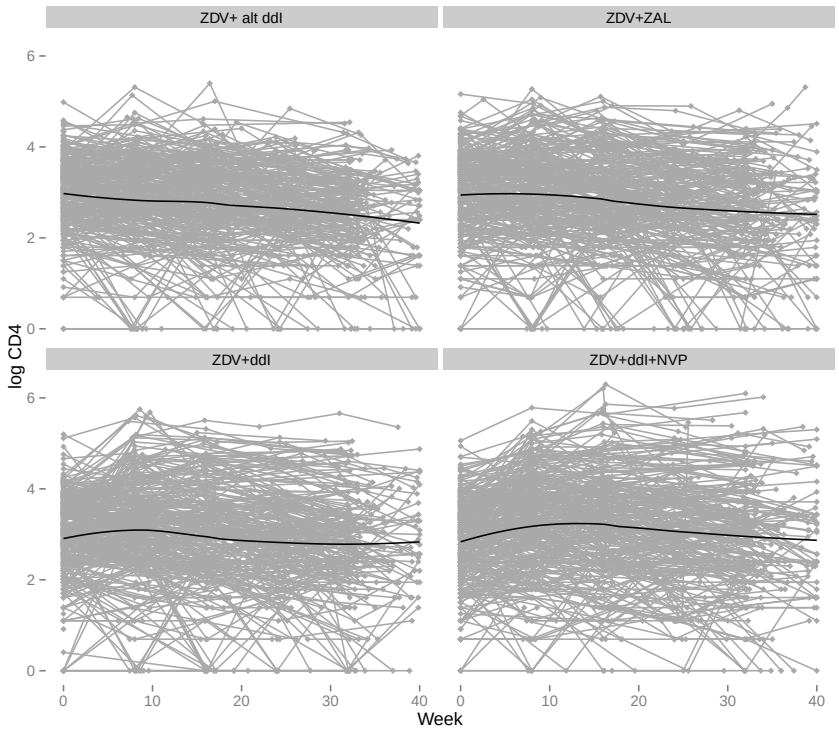

Figure 5.3: *log(CD4+1) profiles for subjects in ACTG Study 193A. A loess smoother fitted to all the data for each treatment is superimposed on the individual profiles in each panel.*

1313 subjects were ***randomized*** to one of four daily treatment regimens consisting of dual or triple combinations of drugs in the class of HIV-1 reverse transcriptase inhibitors: (1) 600 mg of zidovudine (ZDV) alternating monthly with 400 mg of didanosine (ddI); (2) 600 mg ZDV plus 2.25 mg zalcitabine (ZAL); (3) 600 mg ZDV plus 400 mg ddI; or (4) 600 mg ZDV plus 400 mg ddI and 400 mg nevirapine (NVP) (triple therapy).

CD4 measurements were planned at baseline (week 0) and then at 8-week intervals during follow-up, at weeks 8, 16, 24, 32, and 40. Figure 5.3 shows the individual ***log-transformed*** CD4 profiles for subjects randomized to each treatment regimen; because CD4 count of zero is possible, it is customary to take the response variable to be log(CD4+1) (transformed CD4 counts appear ***approximately normally distributed***). As can be seen from the plots, ***actual visits*** did not necessarily take place at ***exactly*** these times; moreover, some subjects ***skipped visits*** altogether or ***dropped out*** of the study before 40 weeks.

For example, visit times for the first subject in the ZDV+ZAL group were t_{ij} = 0, 7.6, 15.6, 23.6, 32.6, and 40 weeks; the first subject in the ZDV+ddI group had actual visits at t_{ij} = 0, 7.1, 16.1, and 32.4 weeks. The number of CD4 measurements per subject ranged from 1 to 9, with a median of 4.

An approximation to addressing this issue would be to "***bin***" actual visit times to correspond to the intended times, so that, for example, 7.6 and 7.1 weeks would be rounded to 8 weeks. However, as discussed in Section 4.2, treating all responses within some interval of an intended visit time as if they were all observed at that time is ***ad hoc***, with unknown effects on inference. If the actual visit times are available, clearly it is ***preferable*** to incorporate them in an analysis.

In addition to treatment regimen, also recorded for each subject is age (years) and gender; thus, the ***among-individual covariates*** are $\boldsymbol{a}_i = (g_i, a_i, \delta_{i1}, \ldots, \delta_{i4})^T$, where $g_i = 0$ (1) for a female (male) subject; a_i is age; and $\delta_{i\ell} = 1$ if subject i was randomized to treatment regimen ℓ and 0 otherwise, $\ell = 1, \ldots, 4$.

A ***local polynomial regression*** (***loess***) curve naïvely fitted to all the data for each treatment is superimposed on each panel in Figure 5.3 as suggested in Section 2.6 to give a rough idea of the overall population mean trend. The visual evidence suggests that a ***straight line*** might provide a reasonable representation of the overall population mean response in each group, although the triple therapy group shows a subtle rise followed by a decay, which might be better captured by a quadratic. Downplaying this for now, a simple model that allows a separate, straight line mean trajectory for each treatment is

$$E(Y_{ij}|\boldsymbol{x}_i) = \beta_0 + \{\beta_{14} + \beta_{11}\delta_{i1} + \beta_{12}\delta_{i2} + \beta_{13}\delta_{i3}\}t_{ij}. \tag{5.24}$$

In (5.24), the intercept is taken to be ***the same*** for all regimens; because subjects were ***randomized*** to the four regimens, the mean response at ***baseline*** (week 0), prior to the start of treatment, should be ***identical*** for all regimens, assuming that the randomization was carried out faithfully. Indeed, the ***sample averages*** of log-transformed CD4 at baseline are 2.98, 2.93, 2.91, and 2.84 for subjects randomized to regimens $1 - 4$.

We have parameterized the slope term in braces so that the triple therapy regimen 4 is the ***reference*** regimen. That is, β_{14} is the slope for the mean CD4 profile for regimen 4, and $\beta_{14} + \beta_{1\ell}$ is the slope for regimen $\ell = 1, 2, 3$, so that $\beta_{1\ell}$, $\ell = 1, 2, 3$ represents the difference in slope relative to triple therapy. Of course, an ***alternative parameterization*** in terms of separate slopes for each regimen is possible; likewise, allowing for ***separate intercepts*** would allow investigation of the integrity of the randomization. Model (5.24) could also be modified to incorporate dependence of intercept and slope on age and gender or to allow quadratic effects.

Specification of a covariance model $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ requires some care. Because all individuals were seen at potentially different times, with different numbers of visits, models for ***balanced*** and ***equally-spaced*** data might not be suitable.

CONSIDERATIONS FOR COVARIANCE MODELS, BALANCED DATA: When the data are ***balanced*** , as for the dental study, as discussed in Section 2.6, inspection of sample covariance and correlation matrices, scatterplot matrices, autocorrelation functions, and lag plots can assist the analyst in identifying plausible models.

In fact, these approaches can be refined to take into account a postulated population mean model so as to take advantage of the belief that the mean follows a ***smooth trajectory***. Instead of basing these diagnostic aids on ***sample means*** at each time point, one can instead estimate those means by a ***preliminary fit*** of the mean model using ***ordinary least squares*** , treating the observations from all individuals as if they are all ***mutually independent***. Although this sounds suspect, as we discuss in Section 5.5, if the mean model is ***correctly specified*** in the sense defined in Section 4.3, then the OLS estimator for β in the overall population mean model $\boldsymbol{X}\beta$ is ***consistent*** for the true value β_0. Thus, at least for m "large," using the ***predicted values*** from the OLS fit to estimate the population means should be reasonable.

CONSIDERATIONS FOR COVARIANCE MODELS, DATA NOT BALANCED: When a longitudinal data set is ***not balanced*** , not only is it more difficult to think about plausible covariance models, ***more ominously*** , if the intention was to record the response at the same ***prespecified times*** for all individuals, but some observations are ***missing*** for some individuals, then things become more complicated. In Section 5.6, we discuss the challenges associated with such ***missing data*** and the assumptions that must be fulfilled to enable ***valid inferences*** to be drawn using the models and methods in this and the next chapter.

For now, we limit our discussion to ***operational issues*** associated with specifying a covariance structure in this situation. Consider the hip replacement study, where the times of observation are the ***same*** for all individuals except that some individuals are ***missing*** the response at some of these times.

Recall that, as discussed in Section 1.3, our notational convention is that $\boldsymbol{Y}_i$ is the $(n_i \times 1)$ vector of responses ***actually observed and recorded*** at times $t_{i1}, \ldots, t_{in_i}$ on individual i. Let

$$\mathcal{Z}_i = (Z_{i1}, \ldots, Z_{in})^T \tag{5.25}$$

be the $(n \times 1)$ vector of ***intended*** responses to be collected at times $t_1, \ldots, t_n$, where $n \geq n_i$ for all $i = 1, \ldots, m$. In the literature on ***missing data*** methods, $\mathcal{Z}_i$ is referred to as the ***full data*** for subject i; see Section 5.6.

- Clearly, for an individual for whom all intended responses are observed, $n_i = n$ and $\boldsymbol{Y}_i = \mathcal{Z}_i$. Thus, $\boldsymbol{V}_i$ for such a individual is a model for $\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \text{var}(\mathcal{Z}_i|\boldsymbol{x}_i)$.

- For an individual with some components of $\mathcal{Z}_i$ ***not observed*** (missing), we can make a correspondence as follows. Consider the hip replacement study, where $n = 4$,

$$\mathcal{Z}_i = (Z_{i1}, \ldots, Z_{i4})^T, \quad (t_1, \ldots, t_4) = (0, 1, 2, 3).$$

  Consider an individual who is missing the intended observation at $t_3 = 2$ weeks. Then

$$\boldsymbol{Y}_i = (Z_{i1}, Z_{i2}, Z_{i4})^T \quad \text{at times} \quad (t_{i1}, t_{i2}, t_{i3}) = (t_1, t_2, t_4) = (0, 1, 3). \tag{5.26}$$

- Here, $\boldsymbol{V}_i$ is a model for the covariance matrix of $\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ as in (5.26), namely, for

$$\begin{pmatrix} \text{var}(Z_{i1}|\boldsymbol{x}_i) & \text{cov}(Z_{i1}, Z_{i2}|\boldsymbol{x}_i) & \text{cov}(Z_{i1}, Z_{i4}|\boldsymbol{x}_i) \\ \text{cov}(Z_{i2}, Z_{i1}|\boldsymbol{x}_i) & \text{var}(Z_{i2}|\boldsymbol{x}_i) & \text{cov}(Z_{i2}, Z_{i4}|\boldsymbol{x}_i) \\ \text{cov}(Z_{i4}, Z_{i1}|\boldsymbol{x}_i) & \text{cov}(Z_{i4}, Z_{i2}|\boldsymbol{x}_i) & \text{var}(Z_{i4}|\boldsymbol{x}_i) \end{pmatrix}, \tag{5.27}$$

  which we can write equivalently as in (5.7) as

$$\boldsymbol{V}_i = \boldsymbol{T}_i^{1/2} \boldsymbol{\Gamma}_i \boldsymbol{T}_i^{1/2}, \tag{5.28}$$

  where $\boldsymbol{T}_i = \text{diag}\{\text{var}(Z_{i1}|\boldsymbol{x}_i), \text{var}(Z_{i2}|\boldsymbol{x}_i), \text{var}(Z_{i4}|\boldsymbol{x}_i)\}$, and

$$\boldsymbol{\Gamma}_i = \begin{pmatrix} 1 & \text{corr}(Z_{i1}, Z_{i2}|\boldsymbol{x}_i) & \text{corr}(Z_{i1}, Z_{i4}|\boldsymbol{x}_i) \\ \text{corr}(Z_{i2}, Z_{i1}|\boldsymbol{x}_i) & 1 & \text{corr}(Z_{i2}, Z_{i4}|\boldsymbol{x}_i) \\ \text{corr}(Z_{i4}, Z_{i1}|\boldsymbol{x}_i) & \text{corr}(Z_{i4}, Z_{i2}|\boldsymbol{x}_i) & 1 \end{pmatrix}.$$

It should be clear from (5.27) that there is no conceptual problem in positing an ***unstructured*** covariance matrix under these circumstances; the only caveat is that some ***bookkeeping*** is necessary to establish the correspondence between observed and intended time points.

Similarly, specification of a ***compound symmetric*** correlation structure is not problematic, as correlation between any two elements of $\mathcal{Z}_i$, and thus $\boldsymbol{Y}_i$ (given $\boldsymbol{x}_i$) is ***the same*** under this model.

Here, the ***intended time points*** are equally-spaced, so that the ***one-dependent*** model in (2.27) and the ***AR(1)*** model in (2.28) are also candidates.

It is straightforward to see that the one-dependent model for the situation in (5.27) takes $\mathbf{\Gamma}_i$ in (5.28) to be

$$\begin{pmatrix} 1 & \alpha & 0 \\ \alpha & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix};$$

and the corresponding AR(1) model is

$$\begin{pmatrix} 1 & \alpha & \alpha^3 \\ \alpha & 1 & \alpha^2 \\ \alpha^3 & \alpha^2 & 1 \end{pmatrix}$$

(check). ***Software packages*** for fitting population-averaged linear models using the methods discussed in the next section incorporate the appropriate bookkeeping for this situation.

In a situation like the HIV clinical trial in ACTG Study 193A, things are more complex. In this study, we can still conceive of the ***full data*** that were intended to be collected on each subject; that is, the vector of ***intended responses*** $\mathcal{Z}_i$ as in (5.25) at prespecified times $(t_1, \ldots, t_n)$.

Strictly speaking, however, each individual i is seen at potentially ***different time points***, so that, ***operationally***, the covariance models that can be feasibly entertained are ***limited***. For example, it is not possible to take the covariance matrix to be completely ***unstructured***, as individuals seen at different time points cannot share the same covariance parameters, so that the vector ξ could be potentially different for each i (and thus infeasible to ***estimate***).

Recognizing that the actual time points for most individuals ***target*** the intended, equally-spaced time points, models such as the compound symmetric, one-dependent, AR(1) might be ***reasonable approximations*** to the true covariance structure. Alternatively, if ***within-individual*** sources of correlation are pronounced, correlation models such as the ***exponential*** (2.31) or ***Gaussian*** (2.32), which depend on the distances between ***actual*** time points, are also feasible.

In Chapter 6, we discuss ***subject-specific*** linear models, for which a model for $\boldsymbol{V}_i$ is induced through specification of ***separate*** models for contributions to the overall covariance structure from ***within-*** and ***among-individual*** sources. This structure "***automatically***" addresses complications arising because of imbalance.

5.3 Maximum likelihood estimation under normality

Given a model specification

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{X}_i\beta, \quad \text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_i = \boldsymbol{V}_i(\xi, \boldsymbol{x}_i) = \boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i)\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)\boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i),$$

as in (5.4), (5.5), and (5.7), and using the ***independence*** of $(\boldsymbol{Y}_i, \boldsymbol{x}_i)$, $i = 1, \ldots, m$, it is possible to formulate ***estimating equations*** that can be solved to yield estimators for the mean parameters β ($p \times 1$) and covariance parameters $\xi = (\theta^T, \alpha^T)^T$ ($r + s \times 1$).

LOGLIKELIHOOD: Specifically, under the ***additional assumption*** that the conditional distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ is ***multivariate normal*** as in (5.6) and using the independence across i, we can appeal to the principle of ***maximum likelihood*** to derive estimators for β and ξ as follows.

Writing the model succinctly as in (5.12), the ***joint density*** for $\boldsymbol{Y}$ conditional on $\tilde{\boldsymbol{x}}$ is

$$\begin{aligned} p(\boldsymbol{y}|\tilde{\boldsymbol{x}}; \beta, \xi) &= (2\pi)^{N/2}|\boldsymbol{V}(\xi, \tilde{\boldsymbol{x}})|^{-1/2} \exp\{-(\boldsymbol{y} - \boldsymbol{X}\beta)^T \boldsymbol{V}^{-1}(\xi, \tilde{\boldsymbol{x}})(\boldsymbol{y} - \boldsymbol{X}\beta)/2\} \\ &= \prod_{i=1}^{m}(2\pi)^{-n_i/2}|\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)|^{-1/2} \exp\{-(\boldsymbol{y}_i - \boldsymbol{X}_i\beta)^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)(\boldsymbol{y}_i - \boldsymbol{X}_i\beta)/2\}. \quad (5.29) \end{aligned}$$

It follows from (5.29) that the ***loglikelihood*** has the form, ignoring constants,

$$\begin{aligned} l(\beta, \xi) &= (-1/2)\Big\{ \log|\boldsymbol{V}(\xi, \tilde{\boldsymbol{x}})| + (\boldsymbol{Y} - \boldsymbol{X}\beta)^T \boldsymbol{V}^{-1}(\xi, \tilde{\boldsymbol{x}})(\boldsymbol{Y} - \boldsymbol{X}\beta) \Big\} \quad (5.30) \\ &= (-1/2)\sum_{i=1}^{m} \Big\{ \log|\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)| + (\boldsymbol{Y}_i - \boldsymbol{X}_i\beta)^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta) \Big\} \quad (5.31) \end{aligned}$$

ESTIMATING EQUATIONS: We appeal to standard matrix differentiation results summarized in Appendix A to derive the ***estimating equations*** (***score equations***) whose joint solution in β and ξ leads to the ***maximum likelihood estimators*** for these parameters ***under the assumption of multivariate normality***.

Differentiating (5.30) and equivalently (5.31) with respect to β ($p \times 1$) yields the estimating equation

$$\boldsymbol{X}^T \boldsymbol{V}^{-1}(\xi, \tilde{\boldsymbol{x}})(\boldsymbol{Y} - \boldsymbol{X}\beta) = \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta) = \boldsymbol{0}, \quad (5.32)$$

which follows (verify) using the following results in Appendix A:

- For $\boldsymbol{x}$ ($n \times 1$), symmetric ($n \times n$) matrix $\boldsymbol{A}$, and ***quadratic form*** $Q = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$, $\partial Q/\partial \boldsymbol{x} = 2\boldsymbol{A}\boldsymbol{x}$ ($n \times 1$).
- If $\boldsymbol{x}$ depends on β ($p \times 1$), the ***chain rule*** then gives $\partial Q/\partial \beta = (\partial \boldsymbol{x}/\partial \beta)(\partial Q/\partial \boldsymbol{x})$, where $(\partial \boldsymbol{x}/\partial \beta)$ is a ($p \times n$) matrix.

It is straightforward to observe the following.

- The estimating equation (5.32) can be rewritten as

$$\beta = \left\{ \boldsymbol{X}^T \boldsymbol{V}^{-1}(\xi, \tilde{\boldsymbol{x}}) \boldsymbol{X} \right\}^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1}(\xi, \tilde{\boldsymbol{x}}) \boldsymbol{Y} = \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i) \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i) \boldsymbol{Y}_i. \tag{5.33}$$

  Thus, if $\boldsymbol{V}$ (equivalently $\boldsymbol{V}_i$, $i = 1, \ldots, m$) were ***known***; i.e., if ξ were ***known***, then (5.33) defines explicitly an ***estimator*** for β.

  Of course, the covariance parameter ξ is ordinarily ***not known*** and must be ***estimated***, which can be accomplished by solving another ***estimating equation*** discussed below, ***jointly*** with (5.32).

- If the model $E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{X}_i\beta$ is ***correctly specified***, then it is straightforward to observe that (5.32) is an ***unbiased estimating equation***.

  In fact, even if the model $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ is ***not*** a correct specification for $\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i)$, the estimating equation is ***still unbiased***. This suggests that, ***even if*** we have specified the covariance structure incorrectly, the maximum likelihood estimator for β under normality will be ***consistent*** for the true value β_0 as long as the mean model is ***correctly specified***.

- Moreover, the foregoing observations hold ***whether or not*** the distribution of $\boldsymbol{Y}_i|\boldsymbol{x}_i$ is actually ***multivariate normal***.

Now consider ***differentiation*** of the loglikelihood (5.30) and equivalently (5.31) with respect to ξ ($r+s\times 1$). This is again straightforward using the following matrix differentiation results from Appendix A. Let $\boldsymbol{V}(\xi)$ be a $(n \times n)$ nonsingular matrix depending on a vector ξ.

- If ξ_k is the kth element of ξ, then $\partial/\partial\xi_k \, \boldsymbol{V}(\xi)$ is the $(n \times n)$ matrix whose (ℓ, p) element is the partial derivative of the (ℓ, p) element of $\boldsymbol{V}(\xi)$ with respect to ξ_k.
- $\partial/\partial\xi_k \{\log|\boldsymbol{V}(\xi)|\} = \text{tr}\left[\boldsymbol{V}^{-1}(\xi)\{\partial/\partial\xi_k \boldsymbol{V}(\xi)\}\right]$, where $\text{tr}(\boldsymbol{A})$ is the ***trace*** of square matrix $\boldsymbol{A}$.
- $\partial/\partial\xi_k \boldsymbol{V}^{-1}(\xi) = -\boldsymbol{V}^{-1}(\xi)\{\partial/\partial\xi_k \boldsymbol{V}(\xi)\}\boldsymbol{V}^{-1}(\xi)$.
- For quadratic form $Q = \boldsymbol{x}^T \boldsymbol{V}(\xi)\boldsymbol{x}$, $\partial Q/\partial\xi_k = \boldsymbol{x}^T\{\partial/\partial\xi_k \boldsymbol{V}(\xi)\}\boldsymbol{x}$. Thus, from the previous result,

$$\partial/\partial\xi_k \{\boldsymbol{x}^T \boldsymbol{V}^{-1}(\xi)\boldsymbol{x}\} = -\boldsymbol{x}^T \boldsymbol{V}^{-1}(\xi)\{\partial/\partial\xi_k \boldsymbol{V}(\xi)\}\boldsymbol{V}^{-1}(\xi)\boldsymbol{x}.$$

Let ξ_k, $k = 1, \ldots, r+s$, be the kth (scalar) component of $\boldsymbol{\xi}$. Applying the foregoing results to differentiation of the loglikelihood (5.30) and equivalently (5.31) with respect to ξ_k, it can be verified (try it) that the result is the following set of $(r+s)$ ***estimating equations***:

$$(1/2)\left((\boldsymbol{Y}-\boldsymbol{X}\beta)^T\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\{\partial/\partial\xi_k\,\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\}\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})(\boldsymbol{Y}-\boldsymbol{X}\beta) - \text{tr}\left[\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\{\partial/\partial\xi_k\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\}\right]\right)=0,\quad k=1,\ldots,r+s. \tag{5.34}$$

or, equivalently,

$$(1/2)\sum_{i=1}^{m}\left((\boldsymbol{Y}_i-\boldsymbol{X}_i\beta)^T\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)\{\partial/\partial\xi_k\,\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)(\boldsymbol{Y}_i-\boldsymbol{X}_i\beta) - \text{tr}\left[\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)\{\partial/\partial\xi_k\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)\}\right]\right)=0,\quad k=1,\ldots,r+s. \tag{5.35}$$

Stacked, the $(r+s)$ estimating equations in (5.34) or (5.35) define ***implicitly*** the maximum likelihood estimator for the covariance parameter $\boldsymbol{\xi}$ ***under the assumption of normality***. In particular, the estimator is obtained by solving these equations ***jointly*** with the equations in (5.32).

We now demonstrate that these estimating equations are ***unbiased*** if the mean model $\boldsymbol{X}_i\beta$ ***and*** the covariance model $\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)$ are ***correctly specified***. Consider form of the equations in (5.35), where they are written as a sum over i of ***independent*** quantities. It can be shown that the conditional (on $\boldsymbol{x}_i$) expectation of a summand in (5.35) is equal to zero by appealing to the following result:

- If $\boldsymbol{U}$ is a random vector with mean zero and covariance matrix $\boldsymbol{V}$, and $\boldsymbol{A}$ is a square matrix, then $E(\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U}) = \text{tr}\{E(\boldsymbol{U}\boldsymbol{U}^T)\boldsymbol{A}\} = \text{tr}(\boldsymbol{V}\boldsymbol{A}) = \text{tr}(\boldsymbol{A}\boldsymbol{V})$ (this is a special case of a more general result in Appendix A).

Using this, we have (assuming expectation is under the parameter values $\boldsymbol{\eta} = (\beta^T, \xi^T)^T$,

$$\begin{aligned} &E_\eta\left[(\boldsymbol{Y}_i-\boldsymbol{X}_i\beta)^T\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)\{\partial/\partial\xi_k\,\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)(\boldsymbol{Y}_i-\boldsymbol{X}_i\beta)\mid\boldsymbol{x}_i\right] \\ &\quad=\text{tr}\left[\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)\{\partial/\partial\xi_k\,\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)\right] \\ &\quad=\text{tr}\left[\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)\partial/\partial\xi_k\{\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)\}\right], \end{aligned} \tag{5.36}$$

from whence unbiasedness of (5.35) follows. Of course, if $\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)$ were ***incorrectly specified***, the equation is ***not*** necessarily unbiased.

As above, the argument to show that these estimating equations are unbiased ***does not require*** multivariate normality to hold; all that is necessary is that the ***first two moments*** of the distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ are correctly specified.

SUMMARY: The estimators for β and ξ in a model of the form

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{X}_i\beta, \quad \text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_i = \boldsymbol{V}_i(\xi, \boldsymbol{x}_i) = \boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i)\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)\boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i)$$

under the ***assumption*** that the conditional distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ is ***multivariate normal*** with these moments are defined as the joint solution to the estimating equations

$$\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta) = \boldsymbol{0}, \tag{5.37}$$

$$(1/2)\sum_{i=1}^{m} \Big((\boldsymbol{Y}_i - \boldsymbol{X}_i\beta)^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)\{\partial/\partial\xi_k \, \boldsymbol{V}_i(\xi, \boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i, \beta) - \text{tr}\left[\boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)\{\partial/\partial\xi_k \boldsymbol{V}_i(\xi, \boldsymbol{x}_i)\}\right]\Big) = 0, \quad k = 1, \ldots, r+s, \tag{5.38}$$

where (5.37) implies

$$\beta = \left\{\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)\boldsymbol{X}_i\right\}^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)\boldsymbol{Y}_i. \tag{5.39}$$

SPECIAL CASE: With $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i) = \boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i)\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)\boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i)$ as in (5.7), a common assumption is that

$$\text{var}(Y_{ij}|\boldsymbol{x}_i) = \sigma^2 \quad \text{for all } i,$$

so that $\boldsymbol{T}_i(\theta, \boldsymbol{x}_i) = \sigma^2\boldsymbol{I}_{n_i}$, $r = 1$, and thus

$$\boldsymbol{V}_i = \sigma^2\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i), \quad \xi = (\sigma^2, \alpha^T)^T. \tag{5.40}$$

It can be verified (do it) under these conditions that the estimating equation of form (5.38) corresponding to σ^2 ($k = 1$) reduces to

$$\sigma^2 = N^{-1}\sum_{i=1}^{m} (\boldsymbol{Y}_i - \boldsymbol{X}_i\beta)^T \boldsymbol{\Gamma}_i^{-1}(\alpha, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta). \tag{5.41}$$

We refer to this case further shortly.

IMPLEMENTATION: Solution of the estimating equations to obtain the ***maximum likelihood estimators*** (***MLEs***) for β and ξ, which we denote as $\widehat{\beta}$ and $\widehat{\xi}$, is of course equivalent to ***maximizing*** the loglikelihood (5.31) in β and ξ. This is the way it is usually implemented in software packages, using ***standard optimization techniques*** such as a ***Newton-Raphson algorithm***.

The usual implementation takes advantage of the fact that (5.37) leads to the ***expression*** for β in (5.39) in terms of ξ and, when $\boldsymbol{V}_i$ is of the form in (5.40), with a multiplicative ***scale parameter*** σ^2, (5.38) yields the expression for σ^2 in (5.41) in terms of β and α. Any β and σ^2 solving the estimating equations, or, equivalently, maximizing the loglikelihood, ***must satisfy*** these expressions.

Thus, if the expressions for β in (5.39) and, in the case of (5.40), σ^2 in (5.41) are substituted into the loglikelihood, the result is a function ***solely*** of the covariance or correlation parameters. This practice is referred to as ***profiling***. The result is that the objective function so obtained can be maximized in the covariance or correlation parameters, which is an optimization problem of ***lower dimension*** so hopefully ***more tractable*** than maximizing the loglikelihood in ***all*** parameters as-is, not taking advantage of these expressions. Once the estimates of the covariance/correlation parameters are obtained, the estimates for β and, if relevant, σ^2, maximizing the objective function can be obtained by substitution of $\widehat{\xi}$ in their expressions.

It is also possible to specify an ***iterative algorithm*** to solve the estimating equations that proceeds by cycling between solving the equation for β holding ξ fixed at the current estimate and solving that for ξ holding β fixed. This is more interesting and useful in the general ***nonlinear*** models we consider in later chapters, so we defer discussion until then.

5.4 Restricted maximum likelihood

BIASED ESTIMATION IN FINITE SAMPLES: We have already observed that the MLEs for β and ξ under the assumption of normality should be ***consistent estimators*** for their true values β_0 and ξ_0, provided that the models $E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{X}_i\beta$ and $\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ are ***correctly specified*** , under general conditions, as they solve ***unbiased estimating equations***. However, in ***finite samples*** , the estimator for ξ can be subject to ***bias*** due to a phenomenon similar to that encountered in estimation of ***variance*** of a scalar outcome Y from an iid sample or in ordinary linear regression.

In particular, if we have an iid sample $Y_1, \ldots, Y_m$ from some distribution with mean μ and variance σ^2, it is well known that the MLE for σ^2 under the assumption of ***normality*** ,

$$m^{-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2,$$

is a (downwardly) ***biased*** estimator for σ^2 for fixed m, as its expectation is $\sigma^2(m-1)/m$.

Accordingly, the usual ***sample variance*** estimator

$$s^2 = (m-1)^{-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2$$

is ***unbiased*** and thus preferred. Evidently, this bias is a consequence of the ***need to estimate*** μ rather than knowing it.

ELIMINATING THE EFFECT OF ESTIMATION OF MEAN PARAMETERS: Thus, although the rationale for s^2 is immediate from this calculation, s^2 can also be deduced by viewing it as the result of an approach that does not rely on estimation of μ. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_m)^T$, with $\mathbf{1}$ a $(m \times 1)$ vector of 1s, and let $\boldsymbol{A}$ be a $(m \times m-1)$ matrix of column rank $m-1$ such that $\boldsymbol{A}^T\mathbf{1} = \mathbf{0}$. Defining the so-called vector of $m-1$ ***error constrasts***

$$\boldsymbol{U} = \boldsymbol{A}^T\boldsymbol{Y},$$

if we assume that the Y_i are $\mathcal{N}(\mu, \sigma^2)$, so that $\boldsymbol{Y} \sim \mathcal{N}(\mu\mathbf{1}, \sigma^2\boldsymbol{I}_m)$, then it is straightforward to deduce that $\boldsymbol{U} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{A}^T\boldsymbol{A})$ and that maximizing the corresponding loglikelihood in σ^2 yields the estimator

$$\widehat{\sigma}^2 = (m-1)^{-1}\boldsymbol{Y}^T\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} = s^2$$

(try it). That is, the sample variance can be derived from effectively eliminating μ from consideration.

A similar result holds for linear regression. Here, with independent pairs $(Y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, m$, and model $Y_i = \boldsymbol{x}_i^T\beta + \epsilon_i$ with $E(\epsilon_i|\boldsymbol{x}_i) = 0$, $\text{var}(\epsilon_i|\boldsymbol{x}_i) = \sigma^2$, if $\boldsymbol{X}$ is the $(m \times p)$ design matrix with rows $\boldsymbol{x}_i$, if $\widehat{\beta}$ is the OLS estimator, it is well known that the MLE for σ^2 under the assumption that ϵ_i given $\boldsymbol{x}_i$ is ***normally distributed*** is $m^{-1}\sum_{i=1}^{m}(Y_i - \boldsymbol{x}_i^T\widehat{\beta})^2$, which can be shown to be biased. Dividing instead by $(m-p)$ yields the usual residual mean square, which is unbiased; a similar argument to that above based on suitably defined "***error contrasts***" can be made to justify this estimator, which is a simpler version of one we give shortly in the context of longitudinal data.

DEMONSTRATION: Given these observations, it is natural to be concerned that normal-theory maximum likelihood estimation of the ***covariance parameters*** ξ in our setting might be subject to similar bias. Clearly, it is not possible for general covariance model $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ to carry out a similar explicit argument. To get a sense, however, consider the special case where

$$\boldsymbol{V}_i(\xi, \boldsymbol{x}_i) = \sigma^2\boldsymbol{\Gamma}_i(\boldsymbol{x}_i), \tag{5.42}$$

where the correlation matrix $\boldsymbol{\Gamma}_i(\boldsymbol{x}_i)$ is a ***known*** function of covariates (so there is no parameter α).

Writing $\boldsymbol{\Gamma}_i$ and $\boldsymbol{\Gamma}$ for brevity, the MLEs for β and σ^2 in this case are (check)

$$\widehat{\beta} = (\boldsymbol{X}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{Y}, \qquad \widehat{\sigma}^2 = N^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta})^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}). \tag{5.43}$$

It is straightforward (try it) to show that the quadratic form in $\widehat{\sigma}^2$ in (5.43) can be written as

$$\boldsymbol{Y}^T\{\boldsymbol{\Gamma}^{-1} - \boldsymbol{\Gamma}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Gamma}^{-1}\}\boldsymbol{Y},$$

which, letting $\boldsymbol{Y}_* = \boldsymbol{\Gamma}^{-1/2}\boldsymbol{Y}$ and $\boldsymbol{X}_* = \boldsymbol{\Gamma}^{-1/2}\boldsymbol{X}$ for $\boldsymbol{\Gamma}^{-1} = \boldsymbol{\Gamma}^{-1/2}\boldsymbol{\Gamma}^{-1/2}$, can be reexpressed as

$$\boldsymbol{Y}_*^T\{\boldsymbol{I}_N - \boldsymbol{X}_*(\boldsymbol{X}_*^T\boldsymbol{X}_*)^{-1}\boldsymbol{X}_*^T\}\boldsymbol{Y}_* = \boldsymbol{Y}_*^T(\boldsymbol{I}_N - \boldsymbol{P}_*)\boldsymbol{Y}_*.$$

Here, $E(\boldsymbol{Y}_*|\tilde{\boldsymbol{x}}) = \boldsymbol{X}_*\beta$, $\text{var}(\boldsymbol{Y}_*|\tilde{\boldsymbol{x}}) = \sigma^2\boldsymbol{I}_N$ (verify), and $\boldsymbol{P}_*$ is a ***symmetric***, ***idempotent*** matrix. By the result for the ***expectation of a quadratic form*** in Appendix A,

$$E\{\boldsymbol{Y}_*^T(\boldsymbol{I}_N - \boldsymbol{P}_*)\boldsymbol{Y}_*|\tilde{\boldsymbol{x}}\} = \text{tr}\{\sigma^2\boldsymbol{I}_N(\boldsymbol{I}_N - \boldsymbol{P}_*)\} + \beta^T\boldsymbol{X}_*^T(\boldsymbol{I}_N - \boldsymbol{P}_*)\boldsymbol{X}_*\beta = \sigma^2\{N - \text{tr}(\boldsymbol{P}_*)\} + \boldsymbol{0} = \sigma^2(N - p),$$

using the fact that (see Appendix A) that the ***trace*** of a symmetric, idempotent matrix is equal to its ***rank***, and the rank of $\boldsymbol{X}$ ($N \times p$) and thus $\boldsymbol{X}_*$ and $\boldsymbol{P}_*$ is p, and $\boldsymbol{X}_*^T(\boldsymbol{I}_N - \boldsymbol{P}_*)\boldsymbol{X}_* = \boldsymbol{0}$ (check).

It follows that

$$E(\widehat{\sigma}^2|\tilde{\boldsymbol{x}}) = \frac{N-p}{N}\sigma^2,$$

demonstrating that the MLE is biased in finite samples (m individuals, N total observations) and that the alternative estimator

$$\widehat{\sigma}^2 = (N-p)^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta})^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}) \tag{5.44}$$

is preferred. Again, it is evident that the bias is a consequence of the needing to estimate β rather than knowing it. We now consider a ***generalization*** of the approach involving ***error contrasts*** above to estimation of ξ in a covariance model $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ that ***eliminates*** estimation of β from the calculation.

RESTRICTED MAXIMUM LIKELIHOOD: Analogous to the previous argument, let $\boldsymbol{A}$ be a $(N \times N-p)$ matrix of column rank $N-p$ such that $\boldsymbol{A}^T\boldsymbol{X} = \boldsymbol{0}$, where of course $\boldsymbol{X}$ is the $(N \times p)$ "stacked" design matrix for all m individuals. Define the vector of $N-p$ ***error contrasts*** to be

$$\boldsymbol{U} = \boldsymbol{A}^T\boldsymbol{Y}.$$

Then if $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}\beta, \boldsymbol{V})$, where we suppress dependence on ξ and $\tilde{\boldsymbol{x}}$ for brevity, we can write

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}),$$

and it is straightforward that

$$\boldsymbol{U} = \boldsymbol{A}^T\boldsymbol{X}\beta + \boldsymbol{A}^T\boldsymbol{\epsilon} = \boldsymbol{A}^T\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^T\boldsymbol{V}\boldsymbol{A}). \tag{5.45}$$

The ***loglikelihood*** corresponding to (5.45) is easily found to be, ignoring constants,

$$l_R(\boldsymbol{\xi}) = (-1/2)\left[\log|\boldsymbol{A}^T\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\boldsymbol{A}| + \boldsymbol{Y}^T\boldsymbol{A}\{\boldsymbol{A}^T\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\boldsymbol{A}\}^{-1}\boldsymbol{A}^T\boldsymbol{Y}\right], \tag{5.46}$$

which does not depend on β. The claim is that maximizing $l_R(\boldsymbol{\xi})$ in $\boldsymbol{\xi}$ leads to an estimator that "***corrects***" for the finite-sample bias in the spirit of (5.44).

We first rewrite (5.46) in a form that makes it ***directly comparable*** to the usual normal loglikelihood (5.30). Note that an $\boldsymbol{A}$ that satisfies $\boldsymbol{A}^T\boldsymbol{X} = \boldsymbol{0}$ is, for $(N \times N - p)$ matrix $\boldsymbol{C}$,

$$\boldsymbol{A} = \{\boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\}\boldsymbol{C}.$$

First, we show that the second term in (5.46) can be rewritten as

$$\boldsymbol{Y}^T\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{V}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta})^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}), \tag{5.47}$$

where $\widehat{\beta} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}$ as in (5.33).

- We first demonstrate that

$$\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{V}\boldsymbol{A})^{-1}\boldsymbol{A}^T = \boldsymbol{P} \quad \text{where } \boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}. \tag{5.48}$$

  Defining

$$\boldsymbol{T} = \boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T - \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T,$$

  it is straightforward to observe that $\boldsymbol{T}$ is ***symmetric and idempotent*** , where idempotency can be verified by direct multiplication to show $\boldsymbol{T}\boldsymbol{T} = \boldsymbol{T}$, using $\boldsymbol{A}^T\boldsymbol{X} = \boldsymbol{0}$. Thus,

$$\text{tr}(\boldsymbol{T}\boldsymbol{T}^T) = \text{tr}(\boldsymbol{T}) = \text{tr}(\boldsymbol{I}_N) - \text{tr}\{\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\} - \text{tr}\{\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\} = N - p - (N - p) = 0,$$

  and $\text{tr}(\boldsymbol{T}\boldsymbol{T}^T) = 0$ implies $\boldsymbol{T} = \boldsymbol{0}$ (check), from whence it follows that

$$\boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T.$$

  Because

$$\boldsymbol{A}^T\boldsymbol{X} = \boldsymbol{A}^T\boldsymbol{V}^{1/2}\boldsymbol{V}^{-1/2}\boldsymbol{X} = (\boldsymbol{V}^{1/2}\boldsymbol{A})^T(\boldsymbol{V}^{-1/2}\boldsymbol{X}) = \boldsymbol{0},$$

  the same result above holds with $\boldsymbol{A}$ replaced by $\boldsymbol{V}^{1/2}\boldsymbol{A}$ and $\boldsymbol{X}$ replaced by $\boldsymbol{V}^{-1/2}\boldsymbol{X}$, yielding

$$\boldsymbol{I}_N - \boldsymbol{V}^{-1/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1/2} = \boldsymbol{V}^{1/2}\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{V}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{V}^{1/2}. \tag{5.49}$$

  Pre- and post-multiplying (5.49) by $\boldsymbol{V}^{-1/2}$ then gives (5.48)

- It can then be shown by ***brute-force multiplication*** (try it) that

$$\boldsymbol{P} = \boldsymbol{PVP}. \tag{5.50}$$

Using (5.48) and (5.50), with $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}$, we have

$$\begin{aligned}\boldsymbol{Y}^T\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{V}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} &= \boldsymbol{Y}^T\boldsymbol{P}\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{P}\boldsymbol{V}\boldsymbol{P}\boldsymbol{Y} \\ &= \{\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}\}^T(\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{V}^{-1})\{\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}\} \\ &= (\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta})^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}),\end{aligned}$$

demonstrating (5.47).

We can thus rewrite the loglikelihood (5.46) as

$$l_R(\boldsymbol{\xi}) = (-1/2)\left\{\log|\boldsymbol{A}^T\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\boldsymbol{A}| + (\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta})^T\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta})\right\}. \tag{5.51}$$

We now argue that the first term can be expressed as

$$\log|\boldsymbol{A}^T\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\boldsymbol{A}| = \log|\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})| + \log|\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\boldsymbol{X}|. \tag{5.52}$$

Differentiate (5.51) with respect to the kth component of $\boldsymbol{\xi}$ to obtain

$$\begin{aligned}(1/2)\Big((\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta})^T\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\{\partial/\partial\xi_k\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\}\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}) \\ - \text{tr}[\{\boldsymbol{A}^T\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\boldsymbol{A}\}^{-1}\boldsymbol{A}^T\{\partial/\partial\xi_k\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\}\boldsymbol{A}]\Big).\end{aligned}$$

The second term can be written, using shorthand and letting $\boldsymbol{V}_\xi = \{\partial/\partial\xi_k\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\}$, as

$$\begin{aligned}tr\{(\boldsymbol{A}^T\boldsymbol{V}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{V}_\xi\boldsymbol{A}\} &= \text{tr}(\boldsymbol{P}\boldsymbol{V}_\xi) \\ &= \text{tr}\{(\boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1})\boldsymbol{V}_\xi\} \\ &= \text{tr}(\boldsymbol{V}^{-1}\boldsymbol{V}_\xi) - \text{tr}\{\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1})\boldsymbol{V}_\xi\} \\ &= \{\partial/\partial\xi_k\log|\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})|\} - \text{tr}\{(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{V}_\xi\boldsymbol{V}^{-1}\boldsymbol{X}\} \\ &= \{\partial/\partial\xi_k\log|\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})|\} + \{\partial/\partial\xi_k\log|\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\boldsymbol{X}|\}.\end{aligned}$$

Because this shows that the derivative of the left hand side of (5.52) is equal to the derivative of the right hand side, we conclude that the first term in (5.51) can be rewritten as $\log|\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})| + \log|\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\boldsymbol{X}|$, as required.

Substituting in (5.51) yields what is usually referred to as the ***restricted maximum likelihood*** (***REML***) objective function

$$l_R(\boldsymbol{\xi}) = (-1/2)\Big\{ \log|\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})| + (\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^T\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}) + \log|\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\boldsymbol{X}|\Big\} \tag{5.53}$$

$$= (-1/2)\left[\sum_{i=1}^{m}\Big\{ \log|\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)| + (\boldsymbol{Y}_i-\boldsymbol{X}_i\widehat{\boldsymbol{\beta}})^T\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)(\boldsymbol{Y}_i-\boldsymbol{X}_i\widehat{\boldsymbol{\beta}})\Big\} + \log\left|\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)\boldsymbol{X}_i\right|\right] \tag{5.54}$$

where

$$\widehat{\boldsymbol{\beta}} = \{\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\boldsymbol{X}\}^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\boldsymbol{Y} = \left\{\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)\boldsymbol{X}_i\right\}^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)\boldsymbol{Y}_i.$$

Note that (5.53) and (5.54) are functions of $\boldsymbol{\xi}$ only, as $\widehat{\boldsymbol{\beta}}$ depends on $\boldsymbol{\xi}$. The suggestion is to maximize (5.53), or equivalently (5.54), in $\boldsymbol{\xi}$ and then substitute the resulting estimator in the expression for $\widehat{\boldsymbol{\beta}}$.

Comparing (5.53) and (5.54) to (5.30) and (5.31) with the expression (5.33) for $\widehat{\boldsymbol{\beta}}$ substituted for $\boldsymbol{\beta}$ shows that they have the ***same*** form except for the third term on the right hand side of (5.53) and (5.54). It is this term that effects the "***correction***" for finite sample bias.

Differentiating with respect to the kth component of $\boldsymbol{\xi}$, $k = 1,\ldots,r+s$, and setting equal to zero yields

$$\begin{aligned}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^T\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\{\partial/\partial\xi_k\,\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\}\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})&\\ -\mathrm{tr}\left[\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\{\partial/\partial\xi_k\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\}\right]&\\ +\mathrm{tr}\left[\{\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\boldsymbol{X}\}^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\{\partial/\partial\xi_k\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\}\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\boldsymbol{X}\right]&=0\end{aligned} \tag{5.55}$$

or, equivalently,

$$\begin{aligned}\sum_{i=1}^{m}\Big((\boldsymbol{Y}_i-\boldsymbol{X}_i\widehat{\boldsymbol{\beta}})^T\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)\{\partial/\partial\xi_k\,\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)(\boldsymbol{Y}_i-\boldsymbol{X}_i\widehat{\boldsymbol{\beta}})&\\ -\mathrm{tr}\left[\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)\{\partial/\partial\xi_k\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)\}\right]\Big)&\\ +\mathrm{tr}\left[\left\{\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)\boldsymbol{X}_i\right\}^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)\{\partial/\partial\xi_k\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)\boldsymbol{X}_i\right]&=0.\end{aligned} \tag{5.56}$$

By the manipulations leading to (5.52), (5.55) can be rewritten, in shorthand, as

$$\boldsymbol{Y}^T\boldsymbol{P}\boldsymbol{V}_\xi\boldsymbol{P}\boldsymbol{Y} - \mathrm{tr}(\boldsymbol{P}\boldsymbol{V}_\xi),$$

and it can be shown that $E(\boldsymbol{Y}^T\boldsymbol{P}\boldsymbol{V}_\xi\boldsymbol{P}\boldsymbol{Y}) = \mathrm{tr}(\boldsymbol{P}\boldsymbol{V}_\xi)$, so that these estimating equations are ***unbiased***; the details are left as an exercise for the diligent student.

As for the MLEs, implementation is via ***maximization of the objective function*** (5.53) using standard optimization algorithms such as ***Newton-Raphson***.

DEMONSTRATION, CONTINUED: We demonstrate that estimation of ξ via REML is expected to lead to "***correction***" for bias due to estimation of β in the special case in (5.42) where $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i) = \sigma^2 \boldsymbol{\Gamma}_i(\boldsymbol{x}_i)$ and where the correlation matrix $\boldsymbol{\Gamma}_i(\boldsymbol{x}_i)$ is ***known***.

Writing $\boldsymbol{\Gamma}_i$ and $\boldsymbol{\Gamma}$ for brevity, we have as in (5.43) that

$$\widehat{\beta} = (\boldsymbol{X}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{Y},$$

and (5.55) becomes

$$(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta})^T \boldsymbol{\Gamma}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta})/\sigma^4 - \text{tr}(\boldsymbol{I}_N)/\sigma^2 + \text{tr}\{(\boldsymbol{X}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{X})^{-1} (\boldsymbol{X}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{X})\}/\sigma^2 = 0.$$

Noting that the last term is equal to $\text{tr}(\boldsymbol{I}_p) = p$, solving yields

$$\widehat{\sigma}_R^2 = (N - p)^{-1} (\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}) \boldsymbol{\Gamma}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}). \tag{5.57}$$

The REML estimator in (5.57) can be seen to be identical to that in (5.44).

REMARKS:

- Although not possible to demonstrate in general, similar "***bias correction***" is achieved for covariance parameters other than scale parameters.
- The original justification for the REML approach is attributed to Patterson and Thompson (1971). See Verbeke and Molenberghs (2000, Section 5.3) for details and other interpretations of the approach.
- It is ***not possible*** to demonstrate theoretically that one of the ML or REML approach is ***uniformly preferable*** for estimation of covariance parameters ξ in general. In the special case of balanced data collected according to a design like that in Chapter 3, with population mean model specified by the classical analysis of variance representation, it turns out that the estimators of the covariance parameters obtained using REML are ***the same*** as the ***classical ANOVA*** estimators obtained by equating mean squares to their expectations; see Verbeke and Molenberghs (2000, Section 5.3) for further references.
- ***In practice***, REML is often used ***by default*** owing to its interpretation given here as providing estimators that should exhibit less bias in finite samples. In fact, software implementing fitting of models like the ones in this and the next chapter ordinarily uses REML as the ***default method*** for estimation of covariance parameters.

5.5 Large sample inference

SAMPLING DISTRIBUTION FOR $\widehat{\beta}$: As we have seen, in the context of a particular model

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{X}_i\beta, \quad \text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_i = \boldsymbol{V}_i(\xi, \boldsymbol{x}_i) = \boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i)\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)\boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i), \quad i = 1, \ldots, m, \quad (5.58)$$

most questions of scientific interest can be represented as questions about the components of β in (5.58). To make inference on β to address the questions formally, we require an estimator for β and its ***sampling distribution***.

The obvious estimator for β is that solving the estimating equation in (5.37), namely,

$$\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta) = \boldsymbol{0}, \quad (5.59)$$

jointly with an estimating equation for ξ such as the ML equation (5.38) or the REML equation (5.56).

- An estimating equation of the general form in (5.59) is often referred to as a ***linear estimating equation*** because it depends on the response through a ***linear function*** of the response, namely $(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta)$. This will be important shortly.

Regardless of which method, ML or REML, one uses to estimate the covariance parameter ξ, even if the model in (5.58) is ***correctly specified*** and the distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ is ***exactly multivariate normal*** with these moments, it is ***not possible*** in general to derive the ***exact sampling distribution*** for the resulting estimator

$$\widehat{\beta} = \{\boldsymbol{X}^T\boldsymbol{V}^{-1}(\widehat{\xi}, \tilde{\boldsymbol{x}})\boldsymbol{X}\}^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}(\widehat{\xi}, \tilde{\boldsymbol{x}})\boldsymbol{Y} = \left\{\sum_{i=1}^{m} \boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\widehat{\xi}, \boldsymbol{x}_i)\boldsymbol{X}_i\right\}^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\widehat{\xi}, \boldsymbol{x}_i)\boldsymbol{Y}_i, \quad (5.60)$$

where $\widehat{\xi}$ in (5.60) is either of the MLE or REML estimator for the covariance parameters in the covariance model in (5.58). Clearly, (5.60) is a complicated function of the data.

LARGE SAMPLE THEORY: Accordingly, we appeal to ***large sample theory*** to derive an approximate sampling distribution for $\widehat{\beta}$ using the general approach for ***estimating equations*** discussed in Section 4.3. As we discussed there, the argument ***does not*** require that the assumption of ***normality*** of the distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ holds.

We assume that the model for $E(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ in (5.58) is ***correctly specified***. Recall that this means that there is a value β_0 such that the true expectation of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ is $\boldsymbol{X}_i\beta_0$; that is, β_0 is a parameter of the distribution that ***truly generated the data***.

- Clearly, if this is ***not*** the case, then we are in pretty serious trouble, as we are addressing the questions of interest (which are questions about population mean response) in a framework that may ***not*** be consistent with the truth.

As suggested by our development so far, specification of a model for the overall population-averaged covariance matrix is admittedly ***more difficult*** than specifying a model for the mean. Accordingly, it is reasonable to be concerned that the model we specify for var($\boldsymbol{Y}_i|\boldsymbol{x}_i$) might ***not be correctly specified***. That is, for example we might select a ***correlation model*** $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$ that does not faithfully represent the true overall correlation structure, and/or we might make incorrect assumptions about the ***overall variance***.

Accordingly, we first consider the ***ideal situation*** in which the models for ***both*** overall mean and covariance posited in (5.58) are ***correctly specified*** , and then consider the case where the latter model might be ***incorrect***.

COVARIANCE MODEL CORRECTLY SPECIFIED: If the model $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ in (5.58) is ***correctly specified***, then there is a value ξ_0 such that the ***true*** overall covariance matrix

$$\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_{0i} \tag{5.61}$$

is $\boldsymbol{V}_i(\xi_0, \boldsymbol{x}_i)$, $i = 1, \ldots, m$. That is, $\boldsymbol{V}_{0i} = \boldsymbol{V}_i(\xi_0, \boldsymbol{x}_i)$ is the covariance matrix of the conditional distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ ***actually*** generating the data.

Rather than just substituting directly into the generic argument in Section 4.3, we carry out the argument from scratch so as to demonstrate a ***fundamental*** and ***well-known result*** that persists across all types of mean-covariance models. The estimator (5.60) satisfies

$$\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\widehat{\xi}, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\beta}) = \boldsymbol{0}. \tag{5.62}$$

Collecting the parameters as $\eta = (\beta^T, \xi^T)^T$, let $\widehat{\eta} = (\widehat{\beta}^T, \widehat{\xi}^T)^T$. Because both the mean and covariance models are ***correctly specified*** , the estimating equation (5.59) and that solved to estimate ξ (ML or REML) are ***unbiased estimating equations***. Thus, we expect that $\widehat{\eta}$ is a ***consistent estimator*** for the ***true value*** $\eta_0 = (\beta_0^T, \xi_0^T)^T$.

Following the argument in Section 4.3, we multiply (5.62) by $m^{-1/2}$ and take a linear Taylor series in $\widehat{\boldsymbol{\eta}}$ about the $\boldsymbol{\eta}_0$. Here, as on the last page of Appendix B (review it), instead of writing the ***linear term*** of the series in terms of this "stacked" parameter vector, we write it as the sum of terms corresponding to each component of $\boldsymbol{\eta}$. That is,

$$\begin{aligned}
\mathbf{0} &= m^{-1/2}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\xi}},\boldsymbol{x}_i)(\boldsymbol{Y}_i-\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}) \\
&\approx m^{-1/2}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{\xi}_0,\boldsymbol{x}_i)(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta}_0)+\left\{-m^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{\xi}_0,\boldsymbol{x}_i)\boldsymbol{X}_i\right\}m^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) \\
&\quad+\left[m^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T\{\partial/\partial\boldsymbol{\xi}\boldsymbol{V}_i^{-1}(\boldsymbol{\xi}_0,\boldsymbol{x}_i)\}(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta}_0)\right]m^{1/2}(\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}_0). \qquad (5.63)
\end{aligned}$$

- In the third term on the right hand side in (5.63), we do not attempt to be more precise about the form of the partial derivative of the covariance matrix $\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)$ with respect to $\boldsymbol{\xi}$ (which this notation is meant to indicate is evaluated at $\boldsymbol{\xi}_0$). This derivative evidently is rather complicated. As we see momentarily, we needn't worry about this.
- We have used the ***consistency*** of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\xi}}$ to approximate the sums in the second and third terms as evaluated at the true value $\boldsymbol{\eta}_0$ rather than an intermediate value $\boldsymbol{\eta}_*$ as in the argument in Section 4.3.

Write the expansion compactly as

$$\mathbf{0}\approx\boldsymbol{C}_m-\boldsymbol{A}_m\,m^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+\boldsymbol{E}_m\,m^{1/2}(\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}_0), \qquad (5.64)$$

where, using $\boldsymbol{V}_i(\boldsymbol{\xi}_0,\boldsymbol{x}_i)=\boldsymbol{V}_{0i}$ as in (5.61),

$$\boldsymbol{C}_m=m^{-1/2}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_{0i}^{-1}(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta}_0),\qquad \boldsymbol{A}_m=m^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_{0i}^{-1}\boldsymbol{X}_i,$$

$$\boldsymbol{E}_m=m^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T\{\partial/\partial\boldsymbol{\xi}\boldsymbol{V}_i^{-1}(\boldsymbol{\xi}_0,\boldsymbol{x}_i)\}(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta}_0).$$

- We in fact assume that

$$m^{1/2}(\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}_0)=O_p(1); \qquad (5.65)$$

  i.e., that this quantity is ***bounded in probability*** (see Appendix C). Under regularity conditions, most estimators that are solutions to unbiased estimating equations satisfy (5.65). This says that $m^{1/2}(\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}_0)$ is "***well-behaved***" as $m\to\infty$ and describes the ***rate*** at which $\widehat{\boldsymbol{\xi}}\xrightarrow{p}\boldsymbol{\xi}_0$; i.e., (5.65) is equivalent to $\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}_0=O_p(m^{-1/2})$. This ensures that the rightmost term in (5.64) does not "***blow up***" as $m\to\infty$.

If we view the argument conditional on $\tilde{\boldsymbol{x}}$, then

$$\boldsymbol{A}_m \to \boldsymbol{A} = \lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_{0i}^{-1} \boldsymbol{X}_i.$$

By the ***central limit theorem***,

$$\boldsymbol{C}_m \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{B}),$$

where

$$\boldsymbol{B} = \lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_{0i}^{-1} \boldsymbol{V}_{0i} \boldsymbol{V}_{0i}^{-1} \boldsymbol{X}_i = \lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_{0i}^{-1} \boldsymbol{X}_i = \boldsymbol{A}.$$

By the ***weak law of large numbers***, using $E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{X}_i\beta_0$, it is straightforward that

$$\boldsymbol{E}_m \xrightarrow{p} \boldsymbol{0}. \tag{5.66}$$

Thus, rearranging and applying these results along with ***Slutsky's theorem***, we are left with

$$m^{1/2}(\widehat{\beta} - \beta_0) \approx \boldsymbol{A}^{-1}\boldsymbol{C}_m \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^{-1}). \tag{5.67}$$

- Note that (5.66) effectively ***eliminates*** any effect of having to ***estimate*** ξ. That is, if ξ_0 were ***known*** and substituted in (5.62), we could have immediately concluded (5.67).
- This reflects the ***fundamental result*** that if we obtain a estimator $\widehat{\beta}$ for a parameter β in a model for a population mean by solving a ***linear estimating equation***, with an estimated "***weight matrix***," the large sample (normal) distribution of $m^{1/2}(\widehat{\beta} - \beta_0)$ is ***the same*** as that for the (***ideal***) estimator for β we could have obtained if the "***weight matrix***" were ***known***.
- This says that there is ***no loss of precision*** suffered by the estimator for β due to having had to ***estimate*** covariance parameters versus ***knowing*** them. Intuitively, this seems like a pretty ***optimistic*** result.
- Indeed, in ***small samples*** (small number of individuals m), inference based on the result in (5.67) can be optimistic in the sense that, for example, ***standard errors*** for the components of $\widehat{\beta}$ derived from (5.67) as we discuss momentarily will be ***too small*** and thus fail to reflect the true uncertainty associated with estimating β (which includes uncertainty due to estimating ξ). In "larger" samples, inferences are often fairly reliable. Of course, what comprises "***large enough***" in any particular setting is not known.

To use the result (5.67) in practice, we approximate $\boldsymbol{A}$ by $\widehat{\boldsymbol{A}}_m$, where $\widehat{\boldsymbol{A}}_m$ is $\boldsymbol{A}_m$ with $\widehat{\xi}$ substituted for ξ_0 in $\boldsymbol{V}_{i0} = \boldsymbol{V}_i(\xi_0, \boldsymbol{x}_i)$, exploiting the fact that $\widehat{\xi}$ is a consistent estimator for ξ_0 under the conditions here. This yields the ***approximate sampling distribution*** for $\widehat{\beta}$ given by

$$\widehat{\beta} \dot{\sim} \mathcal{N}(\beta_0, m^{-1}\widehat{\boldsymbol{A}}_m^{-1}) = \mathcal{N}(\beta_0, \widehat{\boldsymbol{\Sigma}}_M), \tag{5.68}$$

where

$$\widehat{\boldsymbol{\Sigma}}_M = \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\widehat{\xi}, \boldsymbol{x}_i)\boldsymbol{X}_i\right)^{-1} = \{\boldsymbol{X}^T \boldsymbol{V}^{-1}(\widehat{\xi}, \tilde{\boldsymbol{x}})\boldsymbol{X}\}^{-1}. \tag{5.69}$$

(Note that the m^{-1} on the left hand side of (5.68) "cancels" with that in $\widehat{\boldsymbol{A}}_m$.)

- In practice, ***standard errors*** for the estimators for the components of β and associated ***confidence intervals*** for and ***test statistics*** concerning the corresponding components of the true parameter β_0 can be constructed in the usual way based on (5.68) and (5.69).

COVARIANCE MODEL POSSIBLY INCORRECTLY SPECIFIED: We can generalize the above argument to the case where the posited model $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ in (5.58) is ***not necessarily correctly specified***. That is, there is ***no value*** ξ_0 such that $\boldsymbol{V}_i(\xi_0, \boldsymbol{x}_i) = \boldsymbol{V}_{0i}$, where, again, $\boldsymbol{V}_{0i}$ is the ***true covariance matrix*** generating the data.

Of course, in practice, we would proceed unknowingly as if the model $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ ***is*** correct and solve an estimating equation of the form (5.38) (ML) or (5.56) (REML) to obtain an estimator $\widehat{\xi}$. Because the model is incorrect, it is not even clear that ξ has meaning, as it does not represent a quantity relevant to the ***true mechanism*** generating the data. Accordingly, it is not clear exactly what $\widehat{\xi}$ is "***estimating***."

In the generic argument in Section 4.3, we started from the premise that the model underlying the estimating equations being solved for the parameter η is ***correctly specified*** , so that the estimating equations $\sum_{i=1}^{m} \boldsymbol{\Psi}_i(\mathcal{U}_i, \eta) = \boldsymbol{0}$ are ***unbiased*** ; and

$$E\{\boldsymbol{\Psi}_i(\mathcal{U}_i, \eta_0)\} = \boldsymbol{0},$$

where η_0 is the true value. Inspection of (5.38) or (5.56) makes clear that, in our problem, if the model $\boldsymbol{V}_i$ is not correct, then a summand of the estimating equations ***does not*** have expectation zero necessarily, so that the estimating equations are ***not unbiased***. In this situation, we can say something about the behavior of $\widehat{\xi}$ in our problem, as follows.

In the generic case of a ***correct*** model, under regularity conditions, it is possible to ***weaken*** the argument in Section 4.3. If instead we have only that

$$\sum_{i=1}^{m} E\{\boldsymbol{\Psi}_i(\mathcal{U}_i, \eta_0)\} = \mathbf{0} \tag{5.70}$$

(so that each summand does not necessarily have mean zero, but their sum does), then it still holds in general that $\widehat{\eta} \xrightarrow{p} \eta_0$, and the argument leading to the asymptotic normality of the estimator for η goes through unchanged, except that the covariance matrix of $\boldsymbol{\Psi}_i(\mathcal{U}_i, \eta_0)$ is no longer equal to $E\{\boldsymbol{\Psi}_i(\mathcal{U}_i, \eta_0)\boldsymbol{\Psi}_i^T(\mathcal{U}_i, \eta_0)\}$, so that the definitions of the matrices $\boldsymbol{B}_m$ and $\boldsymbol{B}$ in the argument must be changed; e.g., $\boldsymbol{B}_m = m^{-1}\sum_{i=1}^{m} \text{var}\{\boldsymbol{\Psi}_i j(\mathcal{U}_i, \eta_0)\}$ instead.

If the model on which the estimating equations are based is ***incorrect*** under regularity conditions, it is usually the case that there exists η^* such that

$$\sum_{i=1}^{m} E\{\boldsymbol{\Psi}_i(\mathcal{U}_i, \eta^*)\} = \mathbf{0}, \tag{5.71}$$

where this expectation is still with respect to the ***true distribution*** of $\mathcal{U}_i$.

It turns out that, by analogy to (5.70), if (5.71) holds, solving the "***incorrect***" estimating equation will yield an "estimator" such that

$$\widehat{\eta} \xrightarrow{p} \eta^*. \tag{5.72}$$

Although η^* does not have any meaning with respect to the ***true distribution*** generating the data, it is a fixed value dictated by (5.71). A value like η^* can be thought of as the value that "tries to get closest" to the representing the truth within the confines of an incorrect model, and consequently has been referred to as the ***least false parameter***.

The key point is that, ***even with an incorrectly specified model***, we can still deduce the behavior of an "estimator" for a parameter in that model, even if the parameter has no real meaning.

Returning to our problem, we thus assume that, for ***incorrectly specified*** model $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$, if we solve estimating equations like those in (5.38) or (5.56), the solution $\widehat{\xi}$ satisfies (5.72) for some ξ^*; namely,

$$\widehat{\xi} \xrightarrow{p} \xi^*,$$

and, under regularity conditions and analogous to (5.65), $m^{1/2}(\widehat{\xi} - \xi^*) = O_p(1)$.

Suppose then that $\boldsymbol{V}_i$ is incorrectly specified, let

$$\boldsymbol{V}_i^* = \boldsymbol{V}_i(\xi^*, \boldsymbol{x}_i)$$

denote the "***incorrect covariance matrix***" implied by the choice of this incorrect model, and consider again solving (5.62), namely,

$$\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\widehat{\xi}, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\beta}) = \boldsymbol{0}.$$

First, note that for the estimating equation (5.59),

$$\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta) = \boldsymbol{0},$$

we have

$$E\{\boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\xi^*, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta_0)|\boldsymbol{x}_i\} = E\{\boldsymbol{X}_i^T \boldsymbol{V}_i^{*\,-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta_0)|\boldsymbol{x}_i\} = \boldsymbol{0},$$

$i = 1, \ldots, m$, so that the expectation of ***each summand*** is zero, ***even though*** the covariance model is ***incorrectly specified***, so that the estimating equation is still ***unbiased***, analogous to the demonstration for univariate OLS in Section 4.3. We thus conclude that $\widehat{\beta}$ is a ***consistent estimator*** for β_0, despite the fact that the "***weight matrix***" used in the linear estimating equation is not the inverse of the true covariance matrix. In fact, this holds even if we take $\boldsymbol{V}_i = \boldsymbol{I}_{n_i}$, $i = 1, \ldots, m$; that is, assume all N observations across all m individuals are ***mutually uncorrelated***. The resulting estimator for β is effectively ***OLS***, treating all N observations as if they were independent.

Expanding about $(\widehat{\beta}^T, \widehat{\xi}^T)^T = (\beta_0^T, \xi^{*\,T})^T$, analogous to (5.63),

$$\begin{aligned}
\boldsymbol{0} &= m^{-1/2} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\widehat{\xi}, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\beta}) \\
&\approx m^{-1/2} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\xi^*, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta_0) + \left\{ -m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\xi^*, \boldsymbol{x}_i)\boldsymbol{X}_i \right\} m^{1/2}(\widehat{\beta} - \beta_0) \\
&\quad + \left[m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \{\partial/\partial\xi \boldsymbol{V}_i^{-1}(\xi^*, \boldsymbol{x}_i)\}(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta_0) \right] m^{1/2}(\widehat{\xi} - \xi^*) \\
&= \boldsymbol{C}_m^* - \boldsymbol{A}_m^* m^{1/2}(\widehat{\beta} - \beta_0) + \boldsymbol{E}_m^* m^{1/2}(\widehat{\xi} - \xi^*). \qquad (5.73)
\end{aligned}$$

With $\boldsymbol{V}_i^* = \boldsymbol{V}_i(\xi^* \boldsymbol{x}_i)$ and $\boldsymbol{V}_{0i} = \mathrm{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i)$, the ***true covariance matrix***, it is clear that

$$\boldsymbol{E}_m^* \xrightarrow{p} \boldsymbol{0}, \qquad \boldsymbol{A}_m^* \to \boldsymbol{A}^* = \lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i \boldsymbol{V}_i^{*\,-1} \boldsymbol{X}_i,$$

$$\boldsymbol{C}_m^* = m^{-1/2} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{*\,-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{B}^*)$$

by the ***central limit theorem***, where

$$\boldsymbol{B}^* = \lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{*\,-1} \boldsymbol{V}_{0i} \boldsymbol{V}_i^{*\,-1} \boldsymbol{X}_i.$$

Thus, rearranging and using ***Slutsky's theorem*** as before, we have

$$m^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{A}^{*-1}\boldsymbol{B}^{*}\boldsymbol{A}^{*-1}). \tag{5.74}$$

- As in the case where the covariance model is ***correctly specified*** , because $\boldsymbol{E}_m \xrightarrow{p} \mathbf{0}$, there is ***no effect*** of estimating ξ in the incorrect model $\boldsymbol{V}_i$. If the matrix $\boldsymbol{V}_i^*$ had been ***known*** , it is straightforward to observe that (5.74) would still follow.

  This reflects a generalization of the result we saw in the case of a ***correctly specified*** covariance model, namely that the large sample distribution of $m^{1/2}(\widehat{\beta} - \beta_0)$ is the ***same*** if the "***weight matrix*** " used in the linear estimating equation for β is ***fixed*** or ***estimated***.

- In fact, the argument leading to the result (5.67) in the case of a ***correctly specified*** model is a ***special case*** of this result, where the covariance model $\boldsymbol{V}_i$ is ***correct*** after all, so that the $\xi^* = \xi_0$, the value such that $\boldsymbol{V}_{0i} = \boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$.

Note that (5.74), while informative about the behavior of the estimator for β when the posited covariance model is ***incorrect*** , cannot be used as-is in practice, as $\boldsymbol{V}_{0i}$ is of course unknown. We return to this point shortly.

OPTIMAL LINEAR ESTIMATING EQUATION: From (5.67), when the covariance model is ***correctly specified*** , the estimator solving the linear estimating equation satisfies

$$\widehat{\beta}_C \dot{\sim} \mathcal{N}\{\beta_0, (\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X})^{-1}\}, \tag{5.75}$$

where the subscript C indicates "correct," and $\boldsymbol{V}_0 = \text{block diag}(\boldsymbol{V}_{01}, \ldots, \boldsymbol{V}_{0m})$. Likewise, when the covariance model is ***incorrectly specified*** , from (5.74), the estimator solving the linear estimating equation satisfies

$$\widehat{\beta}_{IC} \dot{\sim} \mathcal{N}\{\beta_0, (\boldsymbol{X}^T\boldsymbol{V}^{*-1}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{V}^{*-1}\boldsymbol{V}_0\boldsymbol{V}^{*-1}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{V}^{*-1}\boldsymbol{X})^{-1}\}, \tag{5.76}$$

where the subscript IC indicates "incorrect," and $\boldsymbol{V}^* = \text{block diag}(\boldsymbol{V}_1^*, \ldots, \boldsymbol{V}_m^*)$.

The covariance matrices of the approximate sampling distributions in (5.75) and (5.76) reflect, at least for m large, the ***precision*** with which β can be estimated by solving the linear estimating equation for β under correct and incorrect covariance models. Both $\widehat{\beta}_C$ and $\widehat{\beta}_{IC}$ are ***consistent*** estimators for β_0; thus, we can ***compare*** the covariance matrices of their approximate sampling distributions to examine the ***relative efficiency*** of $\widehat{\beta}_{IC}$ to $\widehat{\beta}_C$.

To this end, consider the difference

$$(\boldsymbol{X}^T\boldsymbol{V}^{*-1}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{V}^{*-1}\boldsymbol{V}_0\boldsymbol{V}^{*-1}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{V}^{*-1}\boldsymbol{X})^{-1} - (\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X})^{-1}. \tag{5.77}$$

We now argue that the difference (5.77) is a ***nonnegative definite*** matrix; that is,

$$\lambda^T\{(\boldsymbol{X}^T\boldsymbol{V}^{*-1}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{V}^{*-1}\boldsymbol{V}_0\boldsymbol{V}^{*-1}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{V}^{*-1}\boldsymbol{X})^{-1} - (\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X})^{-1}\}\lambda \geq \boldsymbol{0} \tag{5.78}$$

for all λ. It follows that, if (5.78) holds, the ***diagonal elements*** (5.77) are all ≥ 0, so that the difference in the approximate ***sampling variances*** of the estimators for each component of β is ≥ 0 (check), implying that the components of $\widehat{\beta}_C$ are ***more efficient*** than those of $\widehat{\beta}_{IC}$.

Letting

$$\boldsymbol{X}_* = \boldsymbol{V}^{*-1/2}\boldsymbol{X}, \quad \boldsymbol{V}^{*-1/2}\boldsymbol{V}^{*-1/2} = \boldsymbol{V}^{*-1}, \quad \boldsymbol{W} = \boldsymbol{V}^{*1/2}\boldsymbol{V}_0^{-1}\boldsymbol{V}^{*1/2},$$

$$\boldsymbol{c} = \boldsymbol{W}^{-1/2}\boldsymbol{X}_*(\boldsymbol{X}_*^T\boldsymbol{X}_*)^{-1}\lambda, \quad \boldsymbol{W} = \boldsymbol{W}^{1/2}\boldsymbol{W}^{1/2},$$

rewrite (5.78) as (check)

$$\boldsymbol{c}^T\{\boldsymbol{I}_N - \boldsymbol{W}^{1/2}\boldsymbol{X}_*(\boldsymbol{X}_*^T\boldsymbol{W}\boldsymbol{X}_*)^{-1}\boldsymbol{X}_*^T\boldsymbol{W}^{1/2}\}\boldsymbol{c}. \tag{5.79}$$

It is straightforward to verify (try it) that

$$\boldsymbol{I}_N - \boldsymbol{W}^{1/2}\boldsymbol{X}_*(\boldsymbol{X}_*^T\boldsymbol{W}\boldsymbol{X}_*)^{-1}\boldsymbol{X}_*^T\boldsymbol{W}^{1/2} = \boldsymbol{I}_N - \boldsymbol{P}_*$$

is ***symmetric and idempotent*** , so that (5.79) can be written as

$$\boldsymbol{c}^T(\boldsymbol{I}_N - \boldsymbol{P}_*)\boldsymbol{c} = \boldsymbol{c}^T(\boldsymbol{I}_N - \boldsymbol{P}_*)(\boldsymbol{I}_N - \boldsymbol{P}_*)\boldsymbol{c} = \boldsymbol{d}^T\boldsymbol{d} \geq \boldsymbol{0},$$

demonstrating (5.78).

The result (5.78) shows that, at least approximately (for "large" m), the components of $\widehat{\beta}_C$ are more precise estimators than those of $\widehat{\beta}_{IC}$. Formally, (5.78) demonstrates that, for a given population mean model $\boldsymbol{X}\beta$, among all ***linear estimating equations*** , that formed using a ***correct covariance model*** will yields a (asymptotically) relatively more efficient estimator than any other based on an ***incorrect covariance model***. That is, the linear estimating equation with "weight matrix" based on a correct covariance model is ***optimal*** among all linear equations in this sense. Of course, this comes as no surprise.

The result ***does not*** provide insight into ***how much more precise*** in general. Evidently, the comparison of the large sample covariance matrices will depend on the ***particular situation*** – the population mean response model and covariates $\boldsymbol{x}_i$ on which it is based (assumed correct), the true covariance matrix, and the assumed covariance model.

We demonstrate a more general optimality result in the case of a ***nonlinear model*** for population-averaged mean response in Chapter 8, which subsumes the one here.

NORMALITY NOT REQUIRED: Note that nowhere in these arguments is anything assumed about the ***true distribution*** of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$; e.g., that it is ***multivariate normal***. The only assumption on this distribution required is that it possess ***sufficient moments*** so that application of the weak law of large numbers, the central limit theorem, etc, is justified. Accordingly, even though we derived the ***estimating equations*** for β and ξ in the assumed covariance model by starting with the ***normal loglikelihood*** , the resulting estimator for β has desirable properties that hold much more generally.

"ROBUST" COVARIANCE MATRIX: In practice, it is natural to be concerned that a posited covariance model is ***not correctly specified***. Identifying an appropriate model is admittedly ***challenging*** ; the structure adopted must faithfully represent the ***aggregate effects*** of both ***among-*** and ***within-individual*** variance and correlation.

Accordingly, rather than carry out inference on β_0 based on the approximate sampling distribution in (5.68), which is based on the covariance model being ***correct*** , it is conventional to base it on the foregoing argument under the condition that the posited covariance model ***may not be correct*** and the result in (5.74), which we repeat here for convenience, dropping the subscript *IC*:

$$m^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^{*-1}\boldsymbol{B}^*\boldsymbol{A}^{*-1}), \tag{5.80}$$

where

$$\boldsymbol{A}^* = \lim_{m\to\infty} m^{-1}\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{*-1}\boldsymbol{X}_i, \quad \boldsymbol{B}^* = \lim_{m\to\infty} m^{-1}\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{*-1}\boldsymbol{V}_{0i}\boldsymbol{V}_i^{*-1}\boldsymbol{X}_i.$$

Of course, $\boldsymbol{A}^*$ can be approximated by

$$m^{-1}\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\widehat{\xi}, \boldsymbol{x}_i)\boldsymbol{X}_i;$$

the difficulty is that $\boldsymbol{B}^*$ depends on the ***true covariance matrix*** $\boldsymbol{V}_{0i}$, which is not known.

However, from the argument in Section 4.3, $\boldsymbol{B}^*$ can be approximated by

$$m^{-1}\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\widehat{\xi}, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\beta})(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\beta})^T \boldsymbol{V}_i^{-1}(\widehat{\xi}, \boldsymbol{x}_i)\boldsymbol{X}_i.$$

The diligent student can verify that, using the consistency of $\widehat{\beta}$ and weak law of large numbers, this expression converges in probability to $\boldsymbol{B}^*$.

Combining, it is thus common to base inference on the approximate sampling distribution

$$\widehat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\Sigma}}_R), \tag{5.81}$$

$$\widehat{\boldsymbol{\Sigma}}_R = \tag{5.82}$$
$$\left\{\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\xi}}, \boldsymbol{x}_i)\boldsymbol{X}_i\right\}^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\xi}}, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}})(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}})^T \boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\xi}}, \boldsymbol{x}_i)\boldsymbol{X}_i \left\{\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\xi}}, \boldsymbol{x}_i)\boldsymbol{X}_i\right\}^{-1}.$$

$\widehat{\boldsymbol{\Sigma}}_R$ is often referred to as the ***robust sandwich*** or ***empirical*** (sampling) covariance matrix, in contrast to $\widehat{\boldsymbol{\Sigma}}_M$ in (5.69), which is often called the ***model-based*** covariance matrix, being based on the assumption that the model for ***overall covariance structure*** is ***correctly specified***. "***Robust***" refers to the fact that $m\times$ (5.82) is a ***consistent estimator*** for the true sampling covariance matrix of $m^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ when the covariance model is ***incorrectly specified*** (and even if it ***is correct***). It is thus ***robust*** to possible misspecification of the covariance model $\boldsymbol{V}_i$.

- It is conventional in practice to base inference on the ***robust covariance matrix*** $\widehat{\boldsymbol{\Sigma}}_R$ rather than the ***model-based*** version $\widehat{\boldsymbol{\Sigma}}_M$ to protect against the possibility of an incorrect covariance model.
- ***Software packages*** implementing these methods and those in the next chapter usually use $\widehat{\boldsymbol{\Sigma}}_R$ ***by default*** to compute approximate standard errors, confidence intervals, and so on.
- By the argument leading to (5.77), using $\widehat{\boldsymbol{\Sigma}}_R$ should result in ***less optimistic assessment of the precision*** with which the components of $\boldsymbol{\beta}$ are estimated.

QUESTIONS OF INTEREST: As discussed in the context of the examples in Section 5.2, questions of scientific interest are usually expressed in terms of ***linear functions*** of the components of $\boldsymbol{\beta}$.

For instance, in the population mean response model (5.14) for the dental study given by

$$E(Y_{ij}|\boldsymbol{x}_i) = \{\beta_{0,B}g_i + \beta_{0,G}(1 - g_i)\} + \{\beta_{1,B}g_i + \beta_{1,G}(1 - g_i)\}t_{ij},$$

$$\boldsymbol{\beta} = (\beta_{0,G}, \beta_{1,G}, \beta_{0,B}, \beta_{1,B})^T,$$

interest focuses on the ***difference in slopes*** between the genders, $\beta_{1,B} - \beta_{1,G}$, so that

$$\boldsymbol{L} = (0, -1, 0, 1). \tag{5.83}$$

If interest is in ***estimating*** the population mean response for boys at age $t_0 = 11$, then we focus on

$$\boldsymbol{L}\boldsymbol{\beta} = \beta_{0,B} + \beta_{1,B}t_0, \qquad \boldsymbol{L} = (0, 1, 0, , t_0).$$

Questions of interest can also involve more than one ***contrast*** of the components of $\boldsymbol{\beta}$; for example, continuing with the dental study, whether or not the (assumed straight line) population mean response trajectories for boys and girls in fact ***coincide*** involves the two contrasts $\beta_{0,B} - \beta_{0,G}$ and $\beta_{1,B} - \beta_{1,G}$ (equal intercepts and slopes). The null hypothesis that ***both intercepts and slopes*** for boys and girls are the same, so that the lines coincide, can be expressed as as $\boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{0}$, where

$$\boldsymbol{L} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}. \tag{5.84}$$

In general, questions can be expressed in terms of $\boldsymbol{L}$ $(c \times p)$ (of full rank), $c \geq 1$, corresponding to a set of contrasts of interest, where ordinarily rank($\boldsymbol{L}$) = c.

INFERENCE: Using the approximate sampling distribution for $\widehat{\boldsymbol{\beta}}$, an ***estimator*** for $\boldsymbol{L}\boldsymbol{\beta}$ is then $\boldsymbol{L}\widehat{\boldsymbol{\beta}}$, and, with $\widehat{\boldsymbol{\Sigma}}$ either of $\widehat{\boldsymbol{\Sigma}}_M$ or $\widehat{\boldsymbol{\Sigma}}_R$, $\boldsymbol{L}\widehat{\boldsymbol{\beta}}$ has approximate sampling distribution, from (5.68) and (5.81),

$$\boldsymbol{L}\widehat{\boldsymbol{\beta}} \stackrel{\cdot}{\sim} \mathcal{N}(\boldsymbol{L}\boldsymbol{\beta}_0, \boldsymbol{L}\widehat{\boldsymbol{\Sigma}}\boldsymbol{L}^T). \tag{5.85}$$

Thus, for example, if $\boldsymbol{L}\boldsymbol{\beta}$ represents the difference in slopes in (5.83) (c = 1), a ***standard error*** for $\boldsymbol{L}\widehat{\boldsymbol{\beta}}$ is $(\boldsymbol{L}\widehat{\boldsymbol{\Sigma}}\boldsymbol{L}^T)^{1/2}$, and a conventional ***Wald*** type $100(1-\alpha)\%$ ***confidence interval*** for $\boldsymbol{L}\boldsymbol{\beta}_0$ is

$$\boldsymbol{L}\widehat{\boldsymbol{\beta}} \pm c_{\alpha/2}(\boldsymbol{L}\widehat{\boldsymbol{\Sigma}}\boldsymbol{L}^T)^{1/2},$$

where $c_{\alpha/2}$ is an appropriate critical value, such as the the $1 - \alpha/2$ quantile of the standard normal or t distribution with some degrees of freedom, discussed further below. A test of $H_0 : \beta_{1,B} - \beta_{1,G} = 0$ versus $H_1 : \beta_{1,B} - \beta_{1,G} \neq 0$ would be based on comparing the test statistic $\boldsymbol{L}\widehat{\boldsymbol{\beta}}/(\boldsymbol{L}\widehat{\boldsymbol{\Sigma}}\boldsymbol{L}^T)^{1/2}$ to the appropriate critical value from a normal or t distribution.

More generally, approximate test statistics for the hypotheses

$$H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{h} \;\; vs. \;\; H_1 : \boldsymbol{L}\boldsymbol{\beta} \neq \boldsymbol{h},$$

where $\boldsymbol{L}$ is $(c \times p)$ with (usually) rank($\boldsymbol{L}$) = c, and $\boldsymbol{h}$ is a specified $(c \times 1)$ vector (***almost always*** $\boldsymbol{h} = \boldsymbol{0}$), can be constructed based on what is now the c-variate approximate sampling distribution (5.85).

- An approximate ***Wald test statistic*** is

$$T_L = (\boldsymbol{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{h})^T(\boldsymbol{L}\widehat{\boldsymbol{\Sigma}}\boldsymbol{L}^T)^{-1}(\boldsymbol{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{h}), \tag{5.86}$$

  which has approximately a ***chi-squared*** distribution with rank($\boldsymbol{L}$) degrees of freedom, so that the test is carried out by comparing T_L to the appropriate χ^2 critical value. If $\boldsymbol{L}$ is a row vector ($c = 1$), then this test is equivalent to the usual "Z test" based on using a standard normal critical value.

- Wald-type tests can be ***optimistic*** in practice and reject H_0 more often than they should because either of the large sample approximate sampling distributions (5.68) and (5.81) ***do not take into account*** variability associated with estimating ξ, so that the test statistic is ***too large***. In finite samples (finite m), this is often addressed by instead using a statistic of the form

$$F_L = \frac{(\boldsymbol{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{h})^T(\boldsymbol{L}\widehat{\boldsymbol{\Sigma}}\boldsymbol{L}^T)^{-1}(\boldsymbol{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{h})}{\text{rank}(\boldsymbol{L})}, \tag{5.87}$$

  which is compared to an F distribution with rank($\boldsymbol{L}$) numerator degrees of freedom and denominator degrees of freedom ***estimated*** from the data. When $c = 1$, this reduces to a t test, with degrees of freedom estimated similarly.

  Several methods have been proposed to ***estimate the denominator degrees of freedom*** for the test statistic (5.87), one of which is based on the so-called ***Satterthwaite approximation***. These are implemented in available software. We do not discuss these here; see Verbeke and Molenberghs (1997, Section 3.5.2 and Appendix A) and the documentation for SAS `proc mixed` for details. These methods usually lead to ***different results***; however, with large m, all yield degrees of freedom that are sufficiently large that the associated p-values are very similar.

When the null hypothesis corresponds to a comparison of ***nested models***, as for $\boldsymbol{L}$ in (5.83) with equal slopes or in (5.84) where the straight lines for boys and girls ***coincide*** under the null, an alternative approach is to carry out a classical ***likelihood ratio test*** based on the ***normal likelihood***

$$L_{ML}(\beta, \xi) = \prod_{i=1}^{m}(2\pi)^{-n_i/2}|\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)|^{-1/2}\exp\{-(\boldsymbol{y}_i - \boldsymbol{X}_i\beta)^T\boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)(\boldsymbol{y}_i - \boldsymbol{X}_i\beta)/2\}. \tag{5.88}$$

Here, one fits the "***full model***" of interest (5.58) first by solving the estimating equations for β and ξ dictated by (5.88) [equivalently, ***maximizing*** (5.88)] to obtain $\widehat{\beta}$ and $\widehat{\xi}$. One then imposes the condition dictated by the null hypothesis $\boldsymbol{L}\beta = \boldsymbol{0}$ and fits the resulting "***reduced model***" by maximizing the corresponding (5.88) (solving the estimating equations) to obtain $\widehat{\beta}^{(0)}$ and $\widehat{\xi}^{(0)}$, say.

The ***likelihood ratio test statistic*** is then

$$T_{LRT} = -2\{\log L_{ML}(\widehat{\beta}^{(0)}, \widehat{\xi}^{(0)}) - \log L_{ML}(\widehat{\beta}, \widehat{\xi})\}. \quad (5.89)$$

Under regularity conditions, the test statistic T_{LRT} in (5.89) has an approximate ***chi-squared distribution*** with degrees of freedom equal to the ***difference in the dimensions*** p of β in the "full" model and that for the "reduced" model; this difference is typically equal to c.

- Although the test statistic (5.89) comes about from assuming that the distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ is ***normal*** , large sample (large m) arguments show that the result that T_{LRT} has an approximate χ^2 distribution holds ***even if*** this distribution is ***not normal***.
- If one uses the ***REML objective function*** in place of the normal likelihood (5.88), a valid test is ***not obtained***. This is because the population mean parameter β is ***eliminated from consideration*** through the "error contrasts," and this parameter is ***different*** under the "full" and "reduced" models, so that the REML objective function is effectively based on ***different*** (mean zero) responses under each model and thus the two REML "loglikelihoods" are not comparable.

Inference on ***components of*** ξ is also sometimes of interest. We defer discussion of this to Chapter 6. For now, we describe the use of so-called ***information criteria*** as a way of ***informally*** comparing competing models, and in particular competing ***covariance models***.

INFORMATION CRITERIA: Although scientific questions are typically framed in terms of β in a model of the form $E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{X}_i\beta$ and can sometimes be cast as a comparison between ***nested models*** for the population mean response of this form, other questions arise where the models to be compared ***cannot*** be viewed as nested.

For example, in ***building a model*** of the form (5.58), while we have in mind a specific model $E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{X}_i\beta$ in which to frame the scientific questions, we may wish to compare the support in the data for several different models $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ for the ***overall covariance structure*** var$(\boldsymbol{Y}_i|\boldsymbol{x}_i)$. Ordinarily, competing models, e.g., compound symmetric versus AR(1), for example, are not nested.

Alternatively, we may wish to compare two competing models for $E(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ that involve different combinations of covariates and consequently are ***not nested***.

Information criteria provide an ***informal approach*** to these challenges. As is well known, the ***more parameters*** that are incorporated in a model, the ***larger the loglikelihood*** becomes; thus, if we wish to compare competing models that are not nested based on the maximized loglikelihoods for these models, we must take this into account. Simply comparing the maximized loglikelihoods directly favors "larger" models. Accordingly, the idea behind information criteria is to incorporate a ***penalty*** for using more parameters and compare instead ***penalized versions*** of the maximized loglikelihoods.

Let $\log \widehat{L}_{ML}$ denote generically the maximized loglikelihood for a specific mean-covariance model, and let P be the total number of parameters (mean and covariance) in the model ($= p + r + s$ for us). Some popular information criteria are as follows; the definitions are such that ***smaller*** values are preferred.

- ***Akaike Information Criterion (AIC):***

$$AIC = -2 \log \widehat{L}_{ML} + 2P. \tag{5.90}$$

- ***Schwarz's Bayesian Information Criterion (BIC):*** With N the total number of observations,

$$BIC = -2 \log \widehat{L}_{ML} + (P \log N). \tag{5.91}$$

- ***Hannan-Quinn Information Criterion (HQ):***

$$HQ = -2 \log \widehat{L}_{ML} + \{P \log(\log N)\}. \tag{5.92}$$

All but AIC involve ***penalties*** depending on ***both*** the ***number of model parameters*** P and the ***total number of observations*** N, so that differences in loglikelihood are calibrated relative to both of these factors.

Analogous criteria can be defined based on the logarithm of the ***REML*** objective function. ***However*** , as noted above, REML "loglikelihoods" are comparable only if they involve the ***same mean model***; thus, information criteria based on REML should be used only to compare ***covariance models*** paired with the ***same*** population mean response model. Some advocate here setting P equal to the number of covariance parameters ($P = r + s$). In addition, because the REML objective function is formulated based on $N - p$ ***error contrasts***, N in (5.91) and (5.92) should be replaced by $N - p$.

Inspection of information criteria ***should not*** be used to draw formal inferences; rather, they should be viewed only as ad hoc ***rules of thumb***. It is entirely possible in practice that ***different criteria*** will prefer ***different models***. AIC often prefers "larger" models relative to BIC, with HQ intermediate. It is beyond our scope to offer a rigorous justification for the use of information criteria for this purpose.

5.6 Missing data

Longitudinal data analysis often involves dealing with ***missing data***, most prominently because of ***attrition*** of individuals over time; that is, ***dropout***. This is, of course, a recurrent challenge when the individuals are ***human subjects***.

Here, although it is ***intended*** to ascertain the outcome of interest at specific time points, as in many of the examples we have examined, some individuals ***fail to present*** for the outcome to be recorded ***after a certain time point***, leading to what is often called a ***monotone pattern of missingness***. More generally, it is the case in many longitudinal studies that individuals do not show up at the intended times in a ***haphazard fashion***, so that the pattern of missingness for any individual can be ***nonmonotone***.

We have already discussed in Section 5.2 the hip replacement study, in which several patients exhibit a ***nonmonotone*** missingness pattern in which they are missing the intended response measurement at week 2 (with one patient also missing the baseline measurement). Recall that, because this phenomenon seems ***systematic*** and occurs for about half of the patients of each gender, it is reasonable to speculate that the fact that these observations are missing has ***nothing to do*** with the health status of patients or their genders. We return to this point shortly.

EXAMPLE: AGE-RELATED MACULAR DEGENERATION CLINICAL TRIAL: Figure 5.4 shows data reported by Molenberghs and Kenward (2007) from a multicenter clinical trial comparing an experimental (active) treatment, interferon-α, to placebo in $m = 240$ patients with age-related macular degeneration (AMD), a leading cause of vision loss among people aged 50 and older. AMD causes damage to the macula, a spot near the center of the retina and the part of the eye needed for sharp, central vision. Patients with AMD progressively lose vision at varying rates. The response, ***visual acuity***, was assessed at baseline (week 0) and then at weeks 4, 12, 24, and 52, and measured the total number of letters a patient read correctly on a standardized vision chart with lines of letters of decreasing size.

Patients were randomized to the two treatments, and all have baseline responses observed; however, ***only 188 of the 240 patients*** have observed responses at ***all five time points***. Of those remaining, 24 dropped out before the final clinic visit at 52 weeks, 8 before the 24 week visit, 6 before the 12 week visit, and 6 before the 4 week visit, with the remaining 8 missing visits intermittently. These data exemplify the very common situation in longitudinal studies in humans in which missingness is almost entirely due to ***dropout***.

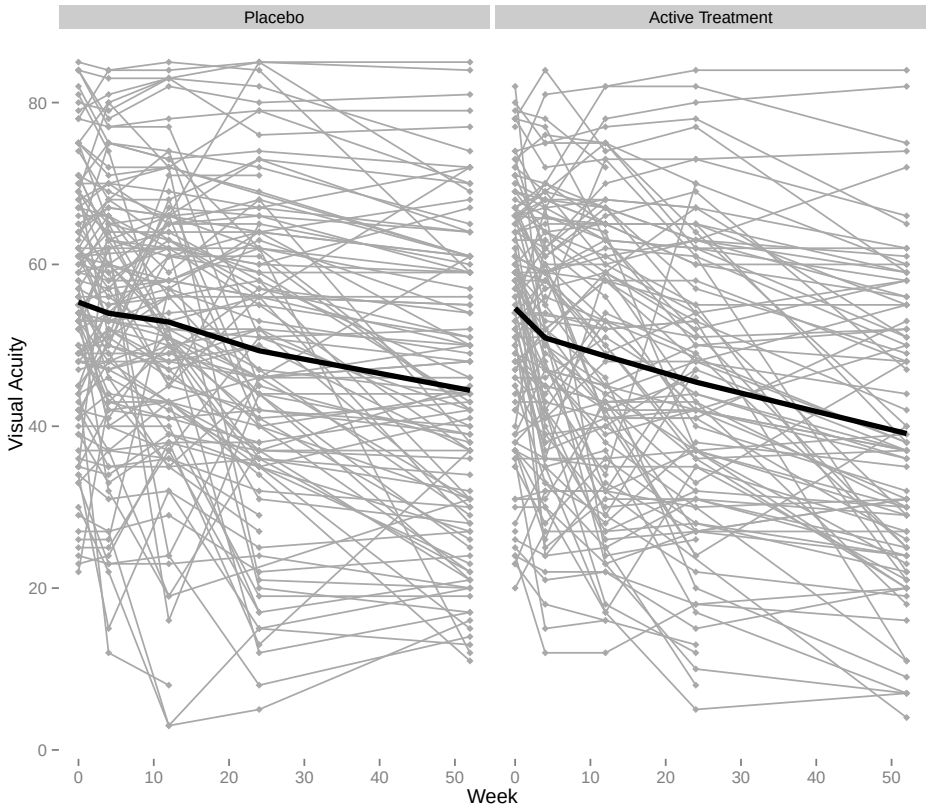

Figure 5.4: *Visual acuity profiles for subjects in the age-related macular degeneration trial. Averages of observed responses at each time point are superimposed on the individual profiles in each panel.*

A full account of the implications of such ***missing data*** on inference and of methods for valid analysis in their presence is the subject of an ***entire course***. Accordingly, we restrict our attention here to implications of missingness for the analysis methods we have discussed; these will also be relevant to those in the next chapter. Whether or not proceeding with an analysis using the ***observed data*** as if they were the ***intended data*** leads to ***valid inferences*** on questions of interest depends on the underlying ***mechanism*** responsible for the missingness, as we now discuss.

NOTATION: We first introduce notation in the context of our longitudinal data framework useful for ***formalizing*** study of missingness. We defined in (5.25) the ***full data***

$$\mathcal{Z}_i = (Z_{i1}, \ldots, Z_{in})^T; \tag{5.93}$$

that is, the responses ***intended*** to be collected on individual i at prespecified times $t_1, \ldots, t_n$. We focus on the situation where the responses ***actually observed*** , which we denote as $\boldsymbol{Y}_i$, have components that are a ***subset*** of those of $\mathcal{Z}_i$, as in (5.26) for the hip replacement data and evidently for the AMD data.

Assume that the ***covariates*** planned to be recorded, $\boldsymbol{x}_i$, are observed for ***all individuals*** $i = 1, \ldots, m$. Of course, in practice, this is ***also not always*** the case, but this is beyond the scope of our discussion here. As is customary in this context, we consider the problem ***conditional*** on $\boldsymbol{x}_i$.

From this point of view, if we ***intend*** to collect the responses (5.93), then it is clear that the ***questions of interest*** pertain to the population mean response for $\mathcal{Z}_i$ given $\boldsymbol{x}_i$. Thus, when we adopt a ***model*** for the population mean response for $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ as we have discussed up to now, implicitly, we are ordinarily ***actually*** specifying a model for the population mean response for $\mathcal{Z}_i$ given $\boldsymbol{x}_i$.

Accordingly, the questions of interest pertain to the ***distribution of the full data***. When data are missing, our objective is thus to address those questions based on the ***observed data***.

Define the ***missing data indicators*** corresponding to the n components of $\mathcal{Z}_i$ as

$$R_{ij} = \begin{cases} 1 & \text{if } Z_{ij} \text{ is observed,} \\ 0 & \text{otherwise,} \end{cases}$$

$j = 1, \ldots, n$; and let

$$\mathcal{R}_i = (R_{i1}, \ldots, R_{in})^T, \tag{5.94}$$

so that $\mathcal{R}_i$ records whether or not Z_{ij}, $j = 1, \ldots, n$, is observed. Then the ***ideal full data*** are

$$(\mathcal{Z}_i, \mathcal{R}_i),$$

which, unless $\mathcal{R}_i = \mathbf{1}$, can never be fully observed (convince yourself).

REMARK: Some authors refer to $\mathcal{Z}_i$ as the ***complete data*** and $(\mathcal{Z}_i, \mathcal{R}_i)$ as the ***full data***.

Let ***r*** denote a possible ***missingness pattern*** ; that is, a vector of zeroes and ones that is a possible value of $\mathcal{R}_i$ in (5.94). In general, there are 2^n ***possible missingness patterns***. If the only missingness patterns observed are those corresponding to ***dropout*** , and all individuals are observed at ***baseline*** (time t_1), then there are n possible patterns

$$(1, 0, \ldots, 0), \quad (1, 1, 0, \ldots, 0), \quad \ldots, \quad (1, 1, \ldots, 1).$$

For a specific pattern of missingness $\boldsymbol{r}$, write $\mathcal{Z}_{(\boldsymbol{r})i}$ to denote the part of $\mathcal{Z}_i$ that is observed, and $\mathcal{Z}_{(\bar{\boldsymbol{r}})i}$ to denote the part that is missing. Then (convince yourself), we can represent the data that we ***actually get to see*** as

$$(\mathcal{Z}_{(\mathcal{R}_i)i}, \mathcal{R}_i), \quad i = 1, \ldots, m. \tag{5.95}$$

We have been referring to the ***observed data*** as $\boldsymbol{Y}_i$, which we now identify with $\mathcal{Z}_{(\boldsymbol{r}_i)i}$ when $\mathcal{R}_i = \boldsymbol{r}_i$; however, strictly speaking, the missing indicators are ***also part of the observable information***. In the missing data literature, (5.95) is referred to as the ***observed data***.

Write the density of $\mathcal{R}_i$ given $\mathcal{Z}_i$ and $\boldsymbol{x}_i$ as

$$p(\boldsymbol{r}_i|\boldsymbol{Z}_i, \boldsymbol{x}_i) = \text{pr}(\mathcal{R}_i = \boldsymbol{r}_i|\mathcal{Z}_i = \boldsymbol{Z}_i, \boldsymbol{x}_i) \tag{5.96}$$

MISSING DATA MECHANISMS: Rubin (1976) pioneered a ***hierarchical taxonomy of missing data mechanisms***, which has become standard:

- ***Missing Completely at Random (MCAR):*** The data are said to be ***MCAR*** if

  $$\text{pr}(\mathcal{R}_i = \boldsymbol{r}_i|\mathcal{Z}_i, \boldsymbol{x}_i) \quad \text{does not depend on } \mathcal{Z}_i; \tag{5.97}$$

  that is, $\mathcal{R}_i \perp\!\!\!\perp \mathcal{Z}_i$ conditional on covariates $\boldsymbol{x}_i$. Then

  $$p(\boldsymbol{r}_i|\boldsymbol{Z}_i, \boldsymbol{x}_i) = p(\boldsymbol{r}_i|\boldsymbol{x}_i). \tag{5.98}$$

  The MCAR mechanism is ***plausible*** in situations where it is clear that missingness has ***nothing*** to do with the issues under study; for example, human subjects ***drop out*** of a study because they move away for work or family reasons. In the hip replacement study, if the missing values at week 2 are due to faulty equipment, for example, it may be reasonable to assume that the mechanism is MCAR.

  Intuitively, under a MCAR mechanism, it ***should be possible*** to make ***valid inferences*** on the questions of interest. The observed data are ***still representative*** of the information ***intended*** to be collected; there are just ***fewer observations*** than originally planned. Thus, the main consequence of proceeding with an analysis of the observed data will be ***loss of efficiency***.

- ***Missing at Random (MAR):*** The data are said to be ***MAR*** if

  $$\text{pr}(\mathcal{R}_i = \boldsymbol{r}_i|\mathcal{Z}_i, \boldsymbol{x}_i) = \text{pr}(\mathcal{R}_i = \boldsymbol{r}_i|\mathcal{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{x}_i); \tag{5.99}$$

  that is, the probability of missingness pattern $\boldsymbol{r}_i$ as a function of $\mathcal{Z}_i$ depends ***only*** on the components of $\mathcal{Z}_i$ that are ***observed*** under $\boldsymbol{r}_i$. Then

  $$p(\boldsymbol{r}_i|\boldsymbol{Z}_i, \boldsymbol{x}_i) = p(\boldsymbol{r}_i|\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{x}_i). \tag{5.100}$$

  If subjects base their decisions to ***drop out*** on their observed response values to that point, and these values ***are available to the data analyst***, then the MAR mechanism is plausible.

Intuitively, if the missingness of intended data is ***associated with*** evolving responses, and those responses reflect, for example, evolving health status, if sicker patients are ***more likely*** to drop out, then the observed data are probably ***not representative*** of the information intended to be collected. Patients who remain in the study may be the ones who are doing better on their assigned treatments; accordingly, proceeding with an analysis to address the questions of interest without taking this into account is likely to lead to ***misleading inferences***.

If ***all data*** implicated in dropout decisions are ***available*** to the data analyst, as they are if the mechanism is MAR, it should be possible to do something to "***adjust***" for the missingness on their basis in an analysis of the observed data.

- ***Missing Not at Random (MNAR):*** The data are said to be ***MNAR*** if pr($\mathcal{R}_i = \boldsymbol{r}_i | \mathcal{Z}_i, \boldsymbol{x}_i$) depends on components of $\mathcal{Z}_i$ that are ***not observed*** when $\mathcal{R}_i = \boldsymbol{r}_i$.

  Intuitively, if a MNAR mechanism governs the missingness, again, the observed data are ***not representative*** of what was intended. However, because the data that are implicated in dropout decisions are ***not available***, "***adjusting***" the analysis for missingness seems ***hopeless***.

AMD EXAMPLE, CONTINUED: In the AMD study, as in most clinical or observational studies of humans with ***dropout***, it is ***unlikely*** that the missingness has ***nothing to do*** with the health status of the subjects. For example, it may well be that patients whose vision ***continues to deteriorate*** might decide to leave the study on the advice of their physicians over concerns that they are ***achieving no benefit***. Here, MCAR is clearly ***implausible***.

If these decisions are based ***solely*** on inspection of the visual acuity measures up to that point, assuming a ***MAR*** mechanism would be reasonable. On the other hand, if the decisions are made based on ***other, unrecorded factors*** that might be associated with patients' ***future prognosis*** and that would be reflected in ***future visual acuity measures***, which are ***not observed***, the mechanism is ***MNAR***.

FUNDAMENTAL CHALLENGE: Of course, we ***cannot determine from the available data*** which of these two explanations reflects the true state of affairs. This conundrum exemplifies the ***fundamental challenge*** of inference with missing data – the ***true missingness mechanism*** is ***not identifiable from the observed data***. Accordingly, whether or not it is plausible to assume that the mechanism is MCAR or MAR, under which methods for achieving ***valid inferences*** on questions of interest based on the observed data are fairly straightforward, ***cannot be determined*** from the data.

Ordinarily, subject-matter expertise and knowledge is incorporated to justify the assumption of MCAR or MAR; however, it remains an ***unverifiable assumption***.

The upshot is that applying the longitudinal analysis methods we have discussed so far and will discuss in the remainder of the course to the observed data when there is missingness ***without acknowledgment*** of this complication can lead to ***misleading inferences***.

A ***full course*** on analysis in the presence of missing data examines this issue in ***excruciating detail***. Here, we focus on one key result that speaks directly to the validity of carrying out an analysis of the observed data using the methods in this and the next chapter under the ***assumption of MAR***.

OBSERVED DATA LIKELIHOOD: Consider the ***joint density*** of the ***ideal full data*** $(\mathcal{Z}_i, \mathcal{R}_i)$, which we write as

$$p(\boldsymbol{Z}_i, \boldsymbol{r}_i|\boldsymbol{x}_i) = p(\boldsymbol{r}_i|\boldsymbol{Z}_i, \boldsymbol{x}_i)\, p(\boldsymbol{Z}_i|\boldsymbol{x}_i). \tag{5.101}$$

In (5.101), we have ***factorized*** the density in to the product of two terms.

- The first term on the right hand side of (5.101), $p(\boldsymbol{r}_i|\boldsymbol{Z}_i, \boldsymbol{x}_i)$, is the density of the ***missingness indicator*** $\mathcal{R}_i$ given the full data $\mathcal{Z}_i$ and covariates $\boldsymbol{x}_i$. As above, depending on the missingness mechanism, this density might ***simplify*** ; we discuss this momentarily.
- The second term on the right hand side, $p(\boldsymbol{Z}_i|\boldsymbol{x}_i)$, is the density of the ***intended, full data*** given covariates. As discussed above, we now see that, from the perspective of the missing data framework, the models we have written for $E(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ and $\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ in (5.58), and indeed for the density $p(\boldsymbol{y}_i|\boldsymbol{x}_i)$ under the assumption of ***normality*** in (5.29), are really models ***implicitly*** reflecting our beliefs about the density of the ***full data*** $\mathcal{Z}_i$ given $\boldsymbol{x}_i$.

Thus, the ***ML methods*** derived in Section 5.3 correspond to assuming that $p(\boldsymbol{Z}_i|\boldsymbol{x}_i)$ in (5.101) is the *n*-***variate normal density***, depending on population mean and covariance parameters $\eta = (\beta^T, \xi^T)^T$.

In principle, we could also adopt a model for the ***density of the missingness mechanism*** , involving a parameter ψ, say. Thus write the assumed model for (5.101) as

$$p(\boldsymbol{Z}_i, \boldsymbol{r}_i|\boldsymbol{x}_i; \psi, \eta) = p(\boldsymbol{r}_i|\boldsymbol{Z}_i, \boldsymbol{x}_i; \psi)\, p(\boldsymbol{Z}_i|\boldsymbol{x}_i; \eta). \tag{5.102}$$

For $\mathcal{R}_i = \boldsymbol{r}_i$, we can ***partition*** $\mathcal{Z}_i$ as above into ***observed and missing components*** as $(\mathcal{Z}_{(\boldsymbol{r}_i)i}, \mathcal{Z}_{(\bar{\boldsymbol{r}}_i)i})$. Accordingly, we can write (5.102) as

$$p(\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i}, \boldsymbol{r}_i|\boldsymbol{x}_i; \psi, \eta) = p(\boldsymbol{r}_i|\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i}, \boldsymbol{x}_i; \psi)\, p(\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i}|\boldsymbol{x}_i; \eta).$$

It follows that we can obtain the joint density of the ***observed*** component and $\mathcal{R}_i$ as

$$\begin{aligned} p(\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{r}_i|\boldsymbol{x}_i;\psi,\eta) &= \int p(\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i}, \boldsymbol{r}_i|\boldsymbol{x}_i;\psi,\eta)\, d\boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i} \\ &= \int p(\boldsymbol{r}_i|\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i}, \boldsymbol{x}_i;\psi)\, p(\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i}|\boldsymbol{x}_i;\eta)\, d\boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i}. \end{aligned} \tag{5.103}$$

This is the density of the ***observed data*** $(\mathcal{Z}_{(\mathcal{R}_i)i}, \mathcal{R}_i)$ in (5.95) as discussed above.

Now ***under MAR***, from (5.100), the first term in the integrand of (5.103) satisfies

$$p(\boldsymbol{r}_i|\boldsymbol{Z}_i, \boldsymbol{x}_i;\psi) = p(\boldsymbol{r}_i|\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i}, \boldsymbol{x}_i;\psi) = p(\boldsymbol{r}_i|\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{x}_i;\psi).$$

Substituting in (5.103), we obtain

$$\begin{aligned} p(\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{r}_i|\boldsymbol{x}_i;\psi,\eta) &= \int p(\boldsymbol{r}_i|\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{x}_i;\psi)\, p(\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i}|\boldsymbol{x}_i;\eta)\, d\boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i} \\ &= p(\boldsymbol{r}_i|\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{x}_i;\psi) \int p(\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i}|\boldsymbol{x}_i;\eta)\, d\boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i} \\ &= p(\boldsymbol{r}_i|\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{x}_i;\psi)\, p(\boldsymbol{Z}_{(\boldsymbol{r}_i)i}|\boldsymbol{x}_i;\eta) \end{aligned} \tag{5.104}$$

Suppose now that we have a sample of ***observed data*** from m individuals, $(\mathcal{Z}_{(\mathcal{R}_i)i}, \mathcal{R}_i)$, $i = 1, \ldots, m$, as in (5.95). Consider the form of the ***likelihood*** for the parameters $(\psi^T, \eta^T)^T$ based on the observed data, often called the ***observed data likelihood***. From (5.104), the ***contribution to the likelihood*** for an individual i with $\mathcal{R}_i = \boldsymbol{r}$ is

$$p(\mathcal{Z}_{(\boldsymbol{r})i}, \boldsymbol{r}|\boldsymbol{x}_i;\psi,\eta) = p(\boldsymbol{r}|\mathcal{Z}_{(\boldsymbol{r})i}, \boldsymbol{x}_i;\psi)\, p(\mathcal{Z}_{(\boldsymbol{r})i}|\boldsymbol{x}_i;\eta); \tag{5.105}$$

when $\boldsymbol{r} = \boldsymbol{1}$, in fact $\mathcal{Z}_{(\boldsymbol{r})i} = \mathcal{Z}_i$. It follows that the contribution to the likelihood for the ith individual can be written

$$\prod_{\boldsymbol{r}} p(\mathcal{Z}_{(\boldsymbol{r})i}, \boldsymbol{r}|\boldsymbol{x}_i;\psi,\eta)^{I(\mathcal{R}_i=\boldsymbol{r})} = \prod_{\boldsymbol{r}} p(\boldsymbol{r}|\mathcal{Z}_{(\boldsymbol{r})i}, \boldsymbol{x}_i;\psi)^{I(\mathcal{R}_i=\boldsymbol{r})} p(\mathcal{Z}_{(\boldsymbol{r})i}|\boldsymbol{x}_i;\eta)^{I(\mathcal{R}_i=\boldsymbol{r})}, \tag{5.106}$$

where the product is over ***all possible missingness patterns r***. The ***observed data likelihood*** is then the product over $i = 1, \ldots, m$ of terms (5.106).

IGNORABILITY: Assume that the parameters ψ and η are ***variation independent*** in the sense that their possible values lie in a ***rectangle***, so that the range of η is ***the same*** for all possible values of ψ, and vice versa. This is often called the ***separability condition***. ***Similar assumptions*** are often made in statistical modeling more generally without comment.

Under the separability condition, there is ***no information*** about the ***parameter of interest***, η, in the first term on the right hand side of (5.106). Thus, for the purpose of maximizing the likelihood to make inference on η, we can ***ignore*** this term. Accordingly, we need only maximize in η

$$\prod_{i=1}^{m} \prod_{\boldsymbol{r}} p(\mathcal{Z}_{(\boldsymbol{r})i}|\boldsymbol{x}_i;\eta)^{I(\mathcal{R}_i=\boldsymbol{r})} \quad \text{or equivalently} \quad \sum_{i=1}^{m} \sum_{\boldsymbol{r}} I(\mathcal{R}_i = \boldsymbol{r}) \log p(\mathcal{Z}_{(\boldsymbol{r})i}|\boldsymbol{x}_i;\eta). \tag{5.107}$$

In fact, under ***ignorability*** and ***separability*** , it is common to refer to (5.107) as the ***observed data likelihood (loglikelihood)***.

Now consider (5.107) from the perspective of the ML approach in Section 5.3. As we have noted, in the context of intending to collect ***full data*** at prespecified time points, the spirit of the model for the observed response vector $\boldsymbol{Y}_i$ conditional on $\boldsymbol{x}_i$ we have discussed is that it ***really reflects*** a model $\mathcal{Z}_i$ given $\boldsymbol{x}_i$. That is, the ***questions of interest*** are formulated within a model for the data we ***intend*** to collect.

From this perspective, as noted above, we are thus assuming that the distribution of $\mathcal{Z}_i$ given $\boldsymbol{x}_i$ is ***n-variate normal***. If ***no data were missing*** , the likelihood for η would be the product of the individual n-variate normal densities dictated by our assumptions on the conditional (on $\boldsymbol{x}_i$) population mean and covariance structure.

When some of the intended observations ***are missing*** , and the missingness mechanism is assumed to be ***MAR*** (which, of course, we ***cannot verify*** from the data), the contribution

$$p(\mathcal{Z}_{(\boldsymbol{r})i}|\boldsymbol{x}_i;\eta)$$

for individual i with $\mathcal{R}_i = \boldsymbol{r}$ is the density of the ***corresponding subvector*** of $\mathcal{Z}_i$. As is ***well-known*** , any ***subvector*** of a ***multivariate normal*** random vector is ***itself multivariate normal*** with mean vector and covariance matrix corresponding to the components contained in the subvector; an example of the latter was demonstrated in (5.27) for the hip replacement study.

Accordingly, when some responses are ***missing*** , and we are willing to believe the assumption of ***MAR*** and the ***separability condition*** , we can ***ignore*** the first term on the right hand side of (5.106), and thus we can regard the likelihood (5.30) and (5.31) as the ***observed data likelihood***. That is, the ***usual analysis*** we carry out to estimate β and ξ under the assumption the response vectors $\boldsymbol{Y}_i$ are n_i-variate normal conditional on $\boldsymbol{x}_i$ corresponds to the likelihood analysis we would perform both if the ***full data*** were observed (i.e., $\boldsymbol{r} = \boldsymbol{1}$ for all m individuals) and with ***missing data*** (i.e., some individuals missing some components of $\mathcal{Z}_i$) ***under MAR***.

The first term on the right hand side of (5.106) represents the ***missing data mechanism*** under MAR. Under these conditions, then, if interest is ***solely*** in the parameters β and ξ, there is ***no need to model and fit*** the missing data mechanism.

KEY RESULT: The usual conclusion from these developments is thus that, ***under MAR***, we expect the usual analysis to yield ***valid inferences*** on β and ξ. However, we must be careful to qualify what we mean by "***valid inferences***."

- We emphasize that, for the usual analysis to yield valid inference, ***both*** (i) the assumption that the distribution of $\mathcal{Z}_i$ given $\boldsymbol{x}_i$ is ***multivariate normal*** with mean and covariance structure ***correctly specified and*** (ii) the assumption of ***MAR*** must hold. If either of these assumptions is not true, then it is ***no longer the case*** that the inferences are necessarily valid.
- The estimators for β and ξ obtained by maximizing (5.107) in η are ***identical*** to those obtained by maximizing (5.106) (under ***separability***). The estimators so obtained will be ***consistent*** for the true values of these parameters assuming, of course, that the ***full data model*** is ***correctly specified***.
- ***Likelihood ratio tests*** comparing ***nested models*** for the ***full data*** based on the statistic in (5.89) will also be ***valid***, as, under ***separability***, the missingness mechanism in (5.106) would have been estimated ***identically*** under the "full" and "reduced" full data models and thus ***cancels*** in the final test statistic.

WRINKLE: Although these results are pleasing, there is a ***catch***: obtaining an appropriate ***approximate sampling distribution*** to use as the basis for ***standard errors*** and ***Wald confidence intervals and tests*** is ***not straightforward***,, as discussed in detail by Verbeke and Molenberghs (2000, Section 17.3 and Chapter 21) and Molenberghs and Kenward (2007, Chapter 12). We focus here as we have previously on inference on β.

- Recall the Taylor series argument in (5.63) and (5.64) to derive the approximate sampling distribution for the ML estimator $\widehat{\beta}$ when the models for mean and covariance matrix are ***correctly specified***. From the vantage point of missing data, this argument was made and accordingly ***expectations*** of the quantities involved were taken acting as if the lengths n_i of the $\boldsymbol{Y}_i$ were ***fixed by design***. If, as here, ***normality holds*** and the mean and covariance models are ***correct*** this argument yields the ***same*** large-m approximate sampling distribution for $\widehat{\beta}$ as does finding the ***expected information matrix*** for $(\widehat{\beta}^T, \widehat{\xi}^T)^T$ and inverting it, where the expectation is taken from this perspective.

- Specifically, recall the definitions

$$\boldsymbol{A}_m = m^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_{0i}^{-1}\boldsymbol{X}_i, \quad \boldsymbol{E}_m = m^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T\{\partial/\partial\boldsymbol{\xi}\boldsymbol{V}_i^{-1}(\xi_0,\boldsymbol{x}_i)\}(\boldsymbol{Y}_i-\boldsymbol{X}_i\beta_0).$$

With ξ $(r+s\times 1)$, using the results for ***matrix differentiation*** in Appendix A, $\boldsymbol{E}_m$ is in fact the $(p\times r+s)$ matrix with kth column

$$-m^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_{0i}^{-1}\{\partial/\partial\xi_k\boldsymbol{V}_i^{-1}(\xi_0,\boldsymbol{x}_i)\}\boldsymbol{V}_{0i}^{-1}(\boldsymbol{Y}_i-\boldsymbol{X}_i\beta_0), \tag{5.108}$$

Thus, the ***observed information matrix***; that is, the negative of the matrix of second partial derivatives of the loglikelihood (5.31), is

$$\begin{pmatrix} \boldsymbol{A}_m & -\boldsymbol{E}_m \\ -\boldsymbol{E}_m^T & -\boldsymbol{G}_m \end{pmatrix}, \tag{5.109}$$

where $\boldsymbol{G}_m$ is the $(r+s\times r+s)$ matrix of second partial derivatives of the loglikelihood with respect to elements of ξ. If we find the ***expected information matrix*** by taking the expectation of (5.109) (conditional on $\tilde{\boldsymbol{x}}$), acting as if the lengths n_i of the $\boldsymbol{Y}_i$ were ***fixed by design***, then

$$E(\boldsymbol{E}_m|\tilde{\boldsymbol{x}}) = \boldsymbol{0}. \tag{5.110}$$

Then the expected information matrix is ***block diagonal*** so that, taking the inverse of the conditional expectation of (5.109), by standard likelihood theory, we are led to the result in (5.67),

$$m^{1/2}(\widehat{\beta}-\beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0},\boldsymbol{A}^{-1}), \quad \boldsymbol{A} = \lim_{m\to\infty} m^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_{0i}^{-1}\boldsymbol{X}_i. \tag{5.111}$$

- The foregoing argument ***hinges critically*** on the fact that ***expectation*** was taken as if the lengths n_i of the $\boldsymbol{Y}_i$ were ***fixed by design***, leading to (5.111). ***However***, from the perspective of missing data, these lengths are ***not fixed in advance***; rather, they are a consequence of the ***realized pattern of missingness***. Accordingly, calculation of the expected information must ***acknowledge*** this by placing this problem in the missing framework we have just described.

- Calculation of $E(\boldsymbol{E}_m|\tilde{\boldsymbol{x}})$ or more precisely, from (5.108), the expectation of a summand in the kth column of $\boldsymbol{E}_m$,

$$\boldsymbol{X}_i^T\boldsymbol{V}_{0i}^{-1}\{\partial/\partial\xi_k\boldsymbol{V}_i^{-1}(\xi_0,\boldsymbol{x}_i)\}\boldsymbol{V}_{0i}^{-1}(\boldsymbol{Y}_i-\boldsymbol{X}_i\beta_0), \tag{5.112}$$

from this point of view can be accomplished via a conditioning argument where the conditioning set involves ***missingness pattern*** $\mathcal{R}_i$. The details of this formulation and argument are beyond our scope here.

- It can then be shown that the expectation of (5.112) (conditional on $\tilde{\boldsymbol{x}}$) is ***not equal to*** $\mathbf{0}$ in general, so that $E(\boldsymbol{E}_m|\tilde{\boldsymbol{x}}) \neq \mathbf{0}$ and

$$\boldsymbol{E}_m \xrightarrow{p} \boldsymbol{E},$$

for some $\boldsymbol{E} \neq \mathbf{0}$ in general. Thus, the expected information is ***not block diagonal***.

- Consequently, if we appeal to standard likelihood theory, using the formula for the ***inverse of a partitioned matrix*** in Appendix A, we obtain that, acknowledging that the n_i are the result of missingness,

$$m^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}[\mathbf{0}, \{\boldsymbol{A} - \boldsymbol{E}(-\boldsymbol{G})^{-1}\boldsymbol{E}^T\}^{-1}], \quad \boldsymbol{G}_m \xrightarrow{p} \boldsymbol{G} \tag{5.113}$$

for $\boldsymbol{G}$ ($r + s \times r + s$) positive definite.

Comparing (5.113) to (5.111) shows that using the ***usual*** large sample approximation the sampling distribution of $\widehat{\beta}$ when the $\boldsymbol{Y}_i$ are ($n_i \times 1$) as a result of a MAR mechanism leads to standard errors that are ***too small*** and Wald test statistics and confidence intervals that are thus ***too optimistic*** when the differing n_i are the consequence of MAR.

- As a way around this, it has thus been advocated that, instead of basing the approximate sampling distribution on the ***usual expected information matrix*** , one should base it on the ***observed information matrix*** (5.109) and obtain standard errors and other inferences based on the inversion of this matrix. This preserves the nonzero off-diagonal, providing an empirical approximation to (5.113). A practical difficulty is that most ***software packages do not*** offer this option and ***do not output*** this matrix as a by-product of the optimization of the loglikelihood.

- It has become common practice (which of course does not make it correct) to ***disregard*** this issue and to use the ***usual*** approximate sampling distribution for inference as if it were valid. Although this has the potential to yield ***misleadingly optimistic*** inferences, there are empirical examples where it does not seem to be too terrible. ***However*** it is important to be aware of this problem. Ideally, calculation of the inverse of the full observed information matrix is ***strongly preferred***.

- It goes without saying that one should ***not use*** the ***robust*** or ***empirical*** covariance matrix (5.82) in this situation. Not only does it suffer from the same drawback, it allows the possibility that the covariance matrix is ***incorrectly specified***.

REMARK: Although in the particular case of a ***correctly specified model*** , ***MAR*** mechanism, and ***likelihood-based*** analysis it is possible to obtain valid inferences on the questions of interest (regarding aspects of the ***full data distribution***), it is important to recognize that this is ***not*** the case in general. Proceeding with a standard analysis in the presence of missing data can lead to substantially ***biased*** results.

Accordingly, it is essential that the data analyst think critically and realistically about possible reasons for missingness. An enormous body of literature exists on methods for achieving valid inferences in the presence of missing data. Verbeke and Molenberghs (2000, Chapters 14-21) and Molenberghs and Kenward (2007) offer extensive discussion of methods for handling missing data in longitudinal data analysis, including alternative approaches under MAR and methods when it is not possible to assume MAR (so that the mechanism is assumed to be MNAR).

REMARK: Contrary to widespread belief, analyses based on so-called ***Big Data*** are ***not*** somehow ***exempt*** from the issues that arise because of missing data. For example, if we have data from ***electronic health records*** on millions of subjects, the fact that some subjects have ***more observations*** on the outcome of interest might reflect that they are having encounters with the health system more frequently because of ***poorer health status***. Thus, subjects with ***fewer observations*** and thus "missing data" by comparison might be healthier, so that inferences on the effects of treatments in the population of all subjects will be compromised if this is not taken into account. With such large m, ***bias*** (***inconsistency***) of standard estimators for population quantities of interest will swamp variance. The result will be estimators that are ***very precise*** but that are ***very far off*** from representing the true quantities of interest.

6 Linear Mixed Effects Models

6.1 Introduction

In the last chapter, we discussed a general class of ***linear models*** for ***continuous response*** arising from a ***population-averaged*** point of view. Here, ***population mean response*** is represented directly by a ***linear model*** that incorporates ***among-*** and ***within-individual*** covariate information. In keeping with the population-averaged perspective, the ***overall aggregate covariance matrix*** of a response vector is also ***modeled directly***. These models are appropriate when the ***questions of scientific interest*** are questions about features of ***population mean response profiles***.

As we observed, selecting among candidate covariance models to represent the overall covariance structure is an inherent challenge. The ***aggregate pattern*** of variance and correlation may be ***sufficiently complex*** that, for example, standard correlation models like those reviewed in Section 2.5 cannot faithfully represent it.

Moreover, when the number of observations per individual n_i differs across individual and/or the observations are at different time points for different individuals, simple ***exploratory*** approaches like those in Section 2.6 are not possible, and some correlation models may not be feasible. Also, care must be taken in implementation. Of course, as discussed in Section 5.6, the reasons for ***imbalance*** must be carefully considered from the point of view of ***missing data mechanisms***.

In this chapter, we instead take a ***subject-specific perspective*** , which leads to the so-called ***linear mixed effects model*** , the ***most popular*** framework for longitudinal data analysis in practice. Here, ***individual inherent response trajectories*** are represented by a ***linear model*** incorporating covariates, and, as in Chapter 2, ***within-*** and ***among-individual*** sources of correlation are ***explicitly acknowledged and modeled separately***. Following the conceptual point of view in Chapter 2, it is natural to acknowledge individual response profiles in this way, and many scientific questions can be interpreted as pertaining to the "***typical***" features of individual trajectories; e.g., the "typical slope."

As discussed in Section 2.4, because of the use of ***linear models*** , this approach ***implies*** a linear model for ***overall population mean response*** and ***induces*** a model for the ***overall aggregate covariance matrix*** , so that a ***linear population-averaged*** model is a byproduct. Thus, as we noted there, the linear mixed effects model is a relevant framework for addressing questions of ***either*** a subject-specific or population-averaged nature.

Moreover, as we observe shortly, the induced covariance structure ameliorates the problems associated with direct specification of the overall pattern and implementation with ***unbalanced*** data discussed in Section 5.2 when a population-averaged model is adopted directly and offers the analyst ***great flexibility*** for modeling variance and correlation structure.

It follows that the ***same*** methods, namely, ***maximum likelihood*** under the assumption of ***normality*** and ***REML***, can be used to fit a linear mixed effects model, and the same large sample theory results deduced in Section 5.5 hold and are used for the basis approximate inference. Likewise, the same concerns discussed in Section 5.6 regarding ***missing data*** continue to apply.

Unlike the population-averaged approach in Chapter 5, however, because the ***subject-specific*** perspective here represents explicitly ***individual behavior***, it is possible to characterize features of individual behavior and to develop an alternative approach to implementation via maximum likelihood, which we discuss later in this chapter.

6.2 Model specification

BASIC MODEL: For convenience, we restate that the observed data are

$$(\boldsymbol{Y}_i, \boldsymbol{z}_i, \boldsymbol{a}_i) = (\boldsymbol{Y}_i, \boldsymbol{x}_i), \quad , i = 1, \ldots, m,$$

independent across i, where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$, with Y_{ij} recorded at time t_{ij}, $j = 1, \ldots, n_i$ (possibly different times for different individuals); $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}^T, \ldots, \boldsymbol{z}_{in_i}^T)^T$ comprising ***within-individual*** covariate information $\boldsymbol{u}_i$ and the t_{ij}; $\boldsymbol{a}_i$ is a vector of ***among-individual*** covariates; and $\boldsymbol{x}_i = (\boldsymbol{z}_i^T, \boldsymbol{a}_i^T)^T$.

We introduce the basic form of the ***linear mixed effects model*** and then present examples that demonstrate how it provides a general framework in which various ***subject-specific*** models can be placed. The model is

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i, \quad i = 1, \ldots, m. \tag{6.1}$$

- In (6.1), $\boldsymbol{X}_i$ ($n_i \times p$) and $\boldsymbol{Z}_i$ ($n_i \times q$) are ***design matrices*** for individual i that depend on individual i's ***covariates*** $\boldsymbol{x}_i$ and time; we present examples of how $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ arise from a subject-specific perspective momentarily.
- The vector β ($p \times 1$) in (6.1) is referred to as the ***fixed effects*** parameter.

- $\boldsymbol{b}_i$ is a $(q \times 1)$ vector of ***random effects*** characterizing ***among-individual*** behavior; i.e., where individual i "sits" in the population. The ***standard*** and most basic assumption is that the $\boldsymbol{b}_i$ are ***independent*** of the covariates $\boldsymbol{x}_i$ and satisfy, for $(q \times q)$ ***covariance matrix*** $\boldsymbol{D}$,

$$E(\boldsymbol{b}_i|\boldsymbol{x}_i) = E(\boldsymbol{b}_i) = \boldsymbol{0}, \quad \text{var}(\boldsymbol{b}_i|\boldsymbol{x}_i) = \text{var}(\boldsymbol{b}_i) = \boldsymbol{D}, \tag{6.2}$$

$$\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}). \tag{6.3}$$

  As we demonstrate, $\boldsymbol{D}$ characterizes variance and correlation due to ***among-individual*** sources. The specifications (6.2) and (6.3) can be relaxed to allow the distribution to differ depending on the values of ***among-individual covariates*** $\boldsymbol{a}_i$, as we discuss shortly, so that

$$E(\boldsymbol{b}_i|\boldsymbol{x}_i) = \boldsymbol{0}, \quad \text{var}(\boldsymbol{b}_i|\boldsymbol{x}_i) = \text{var}(\boldsymbol{b}_i|\boldsymbol{a}_i) = \boldsymbol{D}(\boldsymbol{a}_i), \quad \boldsymbol{b}_i|\boldsymbol{x}_i \sim \mathcal{N}\{\boldsymbol{0}, \boldsymbol{D}(\boldsymbol{a}_i)\}. \tag{6.4}$$

- The ***within-individual deviation*** $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{in_i})^T$ represents the ***aggregate effects*** of the ***within-individual realization*** and ***measurement error*** processes operating at the level of the individual. The ***standard*** and most basic assumption is that the $\boldsymbol{e}_i$ are ***independent*** of the random effects $\boldsymbol{b}_i$ and the covariates $\boldsymbol{x}_i$ and satisfy

$$E(\boldsymbol{e}_i|\boldsymbol{x}_i, \boldsymbol{b}_i) = E(\boldsymbol{e}_i) = \boldsymbol{0}, \quad var(\boldsymbol{e}_i|\boldsymbol{x}_i, \boldsymbol{b}_i) = \text{var}(\boldsymbol{e}_i) = \boldsymbol{R}_i(\gamma). \tag{6.5}$$

  for some $(n_i \times n_i)$ covariance matrix $\boldsymbol{R}_i(\gamma)$ depending on parameters γ. The most common assumption, often adopted ***by default*** without adequate thought, is that

$$\boldsymbol{R}_i(\gamma) = \sigma^2 \boldsymbol{I}_{n_i}, \quad \gamma = \sigma^2, \quad \text{for all } i = 1, \ldots, m; \tag{6.6}$$

  we discuss considerations for specification of $\boldsymbol{R}_i(\gamma)$ shortly. Ordinarily, it is further assumed that

$$\boldsymbol{e}_i \sim \mathcal{N}\{\boldsymbol{0}, \boldsymbol{R}_i(\gamma)\}. \tag{6.7}$$

  (6.5) and (6.7) can be relaxed to allow dependence of $\boldsymbol{e}_i$ on $\boldsymbol{x}_i$ and $\boldsymbol{b}_i$. We consider dependence of $\boldsymbol{e}_i$ on $\boldsymbol{a}_i$ here and defer discussion of more general specifications to Chapter 9.

INTERPRETATION: From the perspective of the ***conceptual model*** (2.9) in Section 2.3,

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \mathcal{B}_i + \boldsymbol{e}_i = \boldsymbol{\mu}_i + \mathcal{B}_i + \boldsymbol{e}_{Pi} + \boldsymbol{e}_{Mi}, \tag{6.8}$$

inspection of (6.1) shows that we can identify $\boldsymbol{\mu}_i = \boldsymbol{X}_i\beta$ as the $(n_i \times 1)$ overall population mean response vector, $\mathcal{B}_i = \boldsymbol{Z}_i\boldsymbol{b}_i$ as the $(n_i \times 1)$ vector of ***deviations*** from the population mean characterizing where individual i "sits" in the population and thus ***among-individual*** variation, and $\boldsymbol{e}_i$ as the $(n_i \times 1)$ vector of ***within-individual*** deviations due to the realization process and measurement error.

Thus, in the linear mixed effects model (6.1),

$$\boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i,$$

characterizes the ***individual-specific trajectory*** for individual i. As we demonstrate in examples shortly, this general form offers great latitude for representing individual profiles.

IMPLIED POPULATION-AVERAGED MODEL: It follows from (6.1) and (6.2) – (6.7) that, ***conditional on*** $\boldsymbol{b}_i$ and $\boldsymbol{x}_i$, $\boldsymbol{Y}_i$ is n_i-***variate normal*** with mean vector $\boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i$ and covariance matrix $\boldsymbol{R}_i(\gamma)$; i.e.,

$$\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i \sim \mathcal{N}\{\boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i, \boldsymbol{R}_i(\gamma)\}.$$

Thus, this conditional distribution characterizes how response observations for individual i vary and covary about the inherent trajectory $\boldsymbol{X}_i\beta+\boldsymbol{Z}_i\boldsymbol{b}_i$ for i due to the ***realization process*** and ***measurement error***.

Letting $p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)$ denote the corresponding normal density, and, from (6.3), letting $p(\boldsymbol{b}_i; \boldsymbol{D})$ be the q-variate normal density corresponding to (6.3), the density of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ is then given by

$$p(\boldsymbol{y}_i|\boldsymbol{x}_i; \beta, \gamma, \boldsymbol{D}) = \int p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)\, p(\boldsymbol{b}_i; \boldsymbol{D})\, d\boldsymbol{b}_i, \tag{6.9}$$

which is easily shown (try it) to be the density of a n_i-variate normal with mean vector $\boldsymbol{X}_i\beta$ and covariance matrix

$$\boldsymbol{V}_i(\gamma, \boldsymbol{D}, \boldsymbol{x}_i) = \boldsymbol{V}_i(\xi, \boldsymbol{x}_i) = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \boldsymbol{R}_i(\gamma), \quad \xi = \{\gamma^T, \text{vech}(\boldsymbol{D})^T\}^T, \tag{6.10}$$

where vech($\boldsymbol{D}$) is the vector of ***distinct*** elements of $\boldsymbol{D}$ (see Appendix A).

Summarizing, the linear mixed effects model framework above implies that

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{X}_i\beta, \quad \text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_i = \boldsymbol{V}_i(\xi, \boldsymbol{x}_i), \quad \boldsymbol{Y}_i|\boldsymbol{x}_i \sim \mathcal{N}\{\boldsymbol{X}_i\beta, \boldsymbol{V}_i(\xi, \boldsymbol{x}_i)\}, \quad i = 1, \ldots, m, \tag{6.11}$$

where $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ is defined in (6.10).

- As in Chapter 5, we will sometimes write $\boldsymbol{V}_i$ and $\boldsymbol{R}_i$ for ***brevity*** , suppressing dependence on parameters for brevity.
- As (6.11) shows, consistent with the discussion above and that in Section 2.4, the subject-specific linear mixed effects model implies a population-averaged model with ***overall population mean*** of the ***same form*** as in (5.4) and ***overall aggregate covariance matrix*** of the particular form (6.10).

- The specific form of the overall covariance matrix (6.10) is ***induced*** by specific choices of $\boldsymbol{R}_i(\gamma)$, reflecting the belief about the nature of the ***within-individual realization and measurement error processes***, and of the covariance matrix $\boldsymbol{D}$ of the random effects, which characterizes ***among-individual variability*** in individual trajectories $\boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i$.

- This development generalizes in the obvious way when the covariance matrix $\text{var}(\boldsymbol{b}_i|\boldsymbol{x}_i) = \text{var}(\boldsymbol{b}_i|\boldsymbol{a}_i) = \boldsymbol{D}(\boldsymbol{a}_i)$ depends on among-individual covariates $\boldsymbol{a}_i$.

- From the point of view of the conceptual model (6.8), the overall covariance matrix (6.10) is, using the assumptions on independence above,

$$\boldsymbol{V}_i(\xi, \boldsymbol{x}_i) = \text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \text{var}(\mathcal{B}_i|\boldsymbol{x}_i) + \text{var}(\boldsymbol{e}_i) = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \boldsymbol{R}_i(\gamma). \tag{6.12}$$

  The correspondence in (6.12) emphasizes that the first term represents the contribution to the induced model for the overall covariance pattern due to ***among-individual*** sources of variance and correlation, and the second term represents the contribution due to ***within-individual*** sources.

  Thus, the induced model allows the data analyst great latitude to think about and incorporate beliefs about these sources ***explicitly***.

MODEL SUMMARY: As in the case of the population-averaged model in Chapter 5, it is convenient to summarize the linear mixed effects model for all $i = 1, \ldots, m$ individuals as follows.

Define

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_m \end{pmatrix} \quad (N \times 1), \quad \boldsymbol{b} = \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \vdots \\ \boldsymbol{b}_m \end{pmatrix} \quad (mq \times 1), \quad \boldsymbol{e} = \begin{pmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \vdots \\ \boldsymbol{e}_m \end{pmatrix} \quad (N \times 1), \tag{6.13}$$

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_m \end{pmatrix} \quad (N \times p), \quad \boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{Z}_m \end{pmatrix} \quad (N \times mq), \tag{6.14}$$

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{R}_m \end{pmatrix} \quad (N \times N), \quad \widetilde{\boldsymbol{D}} = \begin{pmatrix} \boldsymbol{D} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{D} \end{pmatrix} \quad (mq \times mq). \tag{6.15}$$

In (6.13) – (6.15), we suppress dependence of $\boldsymbol{R}_i$ and thus $\boldsymbol{R}$ on γ for brevity.

Using (6.13) – (6.15), we can write the model ***succinctly*** as (verify)

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{e}, \quad E(\boldsymbol{Y}|\tilde{\boldsymbol{x}}) = \boldsymbol{X}\beta, \quad \text{var}(\boldsymbol{Y}|\tilde{\boldsymbol{x}}) = \boldsymbol{V}(\xi, \tilde{\boldsymbol{x}}) = \boldsymbol{V} = \boldsymbol{Z}\widetilde{\boldsymbol{D}}\boldsymbol{Z}^T + \boldsymbol{R}. \tag{6.16}$$

In the literature and most ***software documentation*** , the model is routinely written in the form (6.16).

We now consider several examples that highlight the features of the ***subject-specific linear mixed effects model*** and the considerations involved in model specification.

- As we demonstrated informally in Chapter 2, specification of the model is according to a ***two stage hierarchy*** in which we first represent the form of ***individual inherent trajectories*** in terms of ***individual-specific parameters*** and then "step back" and characterize how these individual-specific parameters ***vary among individuals*** in the population.
- The framework subsumes that of so-called ***random coefficient models***.

SPECIFICATION OF THE WITHIN-INDIVIDUAL COVARIANCE MATRIX $\boldsymbol{R}_i$: As noted above, the ***within-individual covariance matrix***

$$\boldsymbol{R}_i(\gamma) = \text{var}(\boldsymbol{e}_i|\boldsymbol{b}_i, \boldsymbol{x}_i)$$

represents the ***aggregate effects*** of the ***within-individual realization process*** and the ***measurement error process***. Following the conceptual representation in Chapter 2 as in (2.9), as in (6.8),

$$\boldsymbol{e}_i = \boldsymbol{e}_{Pi} + \boldsymbol{e}_{Mi},$$

where, as we noted in that chapter, we would expect that $\text{var}(\boldsymbol{e}_{Mi}|\boldsymbol{b}_i, \boldsymbol{x}_i)$, the contribution to $\boldsymbol{R}_i$ due to ***measurement error*** , to be a ***diagonal matrix*** while $\text{var}(\boldsymbol{e}_{Pi}|\boldsymbol{b}_i, \boldsymbol{x}_i)$, the contribution due to the ***realization process*** , may well exhibit ***correlation*** due to the time-ordered nature of the data collection.

Thus, when considering specification of $\boldsymbol{R}_i(\gamma)$, it is fruitful to decompose it as, in obvious notation,

$$\boldsymbol{R}_i(\gamma) = \boldsymbol{R}_{Pi}(\gamma_P) + \boldsymbol{R}_{Mi}(\gamma_M), \qquad \gamma = (\gamma_P^T, \gamma_M^T)^T, \tag{6.17}$$

where $\boldsymbol{R}_{Pi}(\gamma_P)$ is the covariance model for $\text{var}(\boldsymbol{e}_{Pi}|\boldsymbol{b}_i, \boldsymbol{x}_i)$, and $\boldsymbol{R}_{Mi}(\gamma_P)$ is the ***diagonal*** covariance model for $\text{var}(\boldsymbol{e}_{Mi}|\boldsymbol{b}_i, \boldsymbol{x}_i)$.

We now review the considerations involved from the perspective of the representation (6.17).

- First consider the common, often ***default*** specification

$$\boldsymbol{R}_i(\gamma) = \sigma^2 \boldsymbol{I}_{n_i}$$

in (6.6). From the perspective of (6.17), this can be viewed as

$$\boldsymbol{R}_i(\gamma) = \sigma_P^2 \boldsymbol{I}_{n_i} + \sigma_M^2 \boldsymbol{I}_{n_i}, \qquad \sigma^2 = \sigma_P^2 + \sigma_M^2. \tag{6.18}$$

Thus, this specification incorporates the belief that ***serial correlation*** associated with the realization process is ***negligible***, which might be a reasonable assumption if the observation times are ***sufficiently intermittent*** so that such correlation can reasonably be assumed to have "died out." Of course, this assumption should be critically examined.

From (6.18), the default specification also implies the belief noted above and in Chapter 2 that measurement errors are committed ***haphazardly*** and with variance that is ***the same*** regardless of the magnitude of the true realization of the response being measured. We discuss the practical relevance of this latter assumption in later chapters.

Thus, in (6.6),

$$\sigma^2 = \sigma_P^2 + \sigma_M^2$$

and represents variance due to the ***combined effects*** of the realization process and measurement error.

- In general, it is ***commonplace*** to make the assumption that measurement error, if it is thought to exist, occurs ***haphazardly*** with ***constant variance***, and to take

$$\boldsymbol{R}_{Mi}(\gamma_M) = \sigma_M^2 \boldsymbol{I}_{n_i}. \tag{6.19}$$

Thus, it is routine to write (6.17) without comment as

$$\boldsymbol{R}_i(\gamma) = \boldsymbol{R}_{Pi}(\gamma_P) + \sigma_M^2 \boldsymbol{I}_{n_i}, \qquad \gamma = (\gamma_P^T, \sigma_M^2)^T. \tag{6.20}$$

In applications where the response is ascertained using a ***device*** or ***analytical procedure***, as in the dental study (distance), the hip replacement study (hæmatocrit), or ACTG 193A (CD4 count), it is natural to expect the observed responses to reflect a component of measurement error as in (6.19) and thus to contemplate a model of the form (6.20).

- In some settings, it may be plausible to assume that the response is ascertained ***without measurement error***. For example, in the age-related macular degeneration trial in Section 5.6, we considered the response ***visual acuity*** , which is a count of the number of letters a patient read correctly from a vision chart. Here, it is natural to believe that it is possible to obtain this count ***exactly*** , with no or negligible error.

  In such a situation, the representation of $\boldsymbol{R}_i(\gamma)$ in (6.17) and (6.20) simplifies to

$$\boldsymbol{R}_i(\gamma) = \boldsymbol{R}_{Pi}(\gamma_P), \qquad \gamma = \gamma_P, \tag{6.21}$$

  so that the within-individual covariance matrix model ***reflects entirely*** variation and correlation due to the ***within-individual realization process***.

  Here, plausible models for $\boldsymbol{R}_i(\gamma)$ would naturally be of the form

$$\boldsymbol{R}_i(\gamma) = \boldsymbol{T}_i^{1/2}(\theta)\boldsymbol{\Gamma}_i(\alpha)\boldsymbol{T}_i^{1/2}(\theta), \qquad \gamma = (\theta^T, \alpha^T)^T, \tag{6.22}$$

  where $\boldsymbol{T}_i(\theta)$ is a ***diagonal matrix*** whose diagonal elements reflect the belief about the nature of the ***realization process variance***. For example, assuming that this variance is ***constant over time*** , so that

$$\boldsymbol{T}_i(\theta) = \sigma^2 \boldsymbol{I}_{n_i}, \qquad \theta = \sigma^2,$$

  (6.22) reduces to

$$\boldsymbol{R}_i(\gamma) = \sigma^2 \boldsymbol{\Gamma}_i(\alpha), \qquad \gamma = (\sigma^2, \alpha^T)^T, \tag{6.23}$$

  where now σ^2 is the ***assumed constant*** realization variance, and $\boldsymbol{\Gamma}_i(\alpha)$ is a $(n_i \times n_i)$ ***correlation matrix***.

  The specification (6.23) is often assumed by ***default*** , but it is prudent to consider the possibility that, if $n = \max_i(n_i)$ is the largest number of observations across all individuals, which would be the total number of ***intended times*** in a ***prospectively planned*** study, for individual i with n observations,

$$\boldsymbol{T}_i(\theta) = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2),$$

  which allows realization variance to exhibit ***heterogeneity*** over time.

- It is commonplace for users who are not well-versed in the underpinnings of the linear mixed model to assume without comment either the default specification (6.6) or possibly (6.23), failing to appreciate the implications of the foregoing discussion and the need to ***distinguish*** the contributions of the realization and measurement error processes to the overall pattern of within-individual variance and correlation.

- Moreover, in much of the literature, these considerations are often not made explicit. When they are, the default specification is usually taken to be

$$\boldsymbol{R}_i(\gamma) = \sigma_P^2 \boldsymbol{\Gamma}_i(\boldsymbol{\alpha}) + \sigma_M^2 \boldsymbol{I}_{n_i}, \qquad \gamma = (\sigma_P^2, \boldsymbol{\alpha}^T, \sigma_M^2)^T. \tag{6.24}$$

EXAMPLE 1, DENTAL STUDY: We considered a subject-specific model for these data in Section 2.4, which we recast now in the context of the linear mixed effects model. Recall that there are no within-individual covariates and one among-individual covariate, gender, $g_i = 0$ if i is a girl and $g_i = 1$ if i is a boy, so that $\boldsymbol{x}_i$ contains g_i and the four time points $(t_1, \ldots, t_4) = (8, 10, 12, 14)$.

From a ***subject-specific*** perspective, the primary question of interest is whether or not the ***typical*** or ***average rate of change*** of dental distance for boys differs from that for girls. In (2.13), we adopted a model for the ***individual trajectory*** for any child that represents it as a ***straight line*** with child-specific intercept and slope, namely

$$Y_{ij} = \beta_{0i} + \beta_{0i} t_{ij} + e_{ij}, \quad i = 1, \ldots, n_i = n = 4, \tag{6.25}$$

so that the question involves the difference in the ***typical*** or ***average slope***.

Define the child-specific "***regression parameter***" for i's straight line trajectory in (6.25) as

$$\boldsymbol{\beta}_i = \begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix}.$$

We can then summarize (6.25) as

$$\boldsymbol{Y}_i = \boldsymbol{C}_i \boldsymbol{\beta}_i + \boldsymbol{e}_i, \quad \boldsymbol{C}_i = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix} = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \\ 1 & t_4 \end{pmatrix}, \quad i = 1, \ldots, m, \tag{6.26}$$

where, because of the ***balance***, $\boldsymbol{C}_i$ is the ***same*** (4×2) matrix for all i.

As in (2.14), we allow individual-specific intercepts and slopes to vary about ***typical*** or mean values for ***each gender*** according to ***random effects*** with

$$\begin{aligned} \beta_{0i} &= \beta_{0,B} g_i + \beta_{0,G}(1 - g_i) + b_{0i}, \\ \beta_{1i} &= \beta_{1,B} g_i + \beta_{1,G}(1 - g_i) + b_{1i}. \end{aligned} \qquad \boldsymbol{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix}. \tag{6.27}$$

REMARK: In the early longitudinal data literature, a model of the form (6.26) along with a representation for $\boldsymbol{\beta}_i$ as in (6.27) is referred to as a ***random coefficient model***.

We can write (6.27) concisely as (verify)

$$\boldsymbol{\beta}_i = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i, \tag{6.28}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_{0,G} \\ \beta_{1,G} \\ \beta_{0,B} \\ \beta_{1,B} \end{pmatrix}, \quad \boldsymbol{A}_i = \begin{pmatrix} (1-g_i) & 0 & g_i & 0 \\ 0 & (1-g_i) & 0 & g_i \end{pmatrix}, \quad \boldsymbol{B}_i = \boldsymbol{I}_2.$$

Substituting (6.28) in (6.26) and rearranging, we have

$$\boldsymbol{Y}_i = \boldsymbol{C}_i\boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{C}_i\boldsymbol{B}_i\boldsymbol{b}_i + \boldsymbol{e}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i, \tag{6.29}$$

where

$$\boldsymbol{X}_i = \boldsymbol{C}_i\boldsymbol{A}_i, \quad \boldsymbol{Z}_i = \boldsymbol{C}_i\boldsymbol{B}_i.$$

Thus, it is straightforward to deduce that

$$\boldsymbol{X}_i = \begin{pmatrix} (1-g_i) & (1-g_i)t_1 & g_i & g_it_1 \\ \vdots & \vdots & \vdots & \vdots \\ (1-g_i) & (1-g_i)t_4 & g_i & g_it_4 \end{pmatrix}, \quad \boldsymbol{Z}_i = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \\ 1 & t_4 \end{pmatrix}. \tag{6.30}$$

Here, $\boldsymbol{X}_i$ is the ***same*** as the design matrix (5.15) in the ***population-averaged model*** in Chapter 5.

To complete the specification , we posit models for ***among-individual covariance matrix*** var($\boldsymbol{b}_i|\boldsymbol{a}_i$) and the ***within-individual covariance matrix*** $\boldsymbol{R}_i(\gamma)$.

- In Section 2.6, empirical exploration of the ***overall aggregate pattern of covariance*** shows evidence that ***overall correlation*** is ***different*** for boys and girls with ***overall variance constant across time*** but possibly ***larger*** for boys than for girls.
- Examination of the ***within-individual residuals*** from ***individual-specific*** fits of model (6.25) to each child ***does not*** show strong evidence of ***within-individual correlation***; we showed this for boys, and the same observation applies to girls.

- Moreover, these residuals suggest for each gender that ***within-child variance*** due to the combined effects of realization and measurement error is ***constant*** over time. Estimates of within-child variance based on pooling the residuals across children of each gender are 2.59 for boys and 0.45 for girls; the ***much larger value for boys*** is likely due in part to the very large fluctuation of distance values within one boy.

- Combining these observations, it may be reasonable to assume that the ***within-child covariance matrix*** is of the general form (6.24) with the correlation matrix $\boldsymbol{\Gamma}_i(\alpha)$ approximately equal to an identity matrix as in (6.18), so that $\boldsymbol{R}_i(\gamma)$ for any child is ***diagonal***.

  However, because the estimates of ***within-child aggregate variance*** are so different, we might consider initially a form of (6.18) that is ***different*** for each gender. That is, relaxing (6.5), so that $\boldsymbol{e}_i$ and $\boldsymbol{a}_i$ are ***not necessarily independent***, a plausible model is, in obvious notation,

$$\begin{aligned} \text{var}(\boldsymbol{e}_i|\boldsymbol{a}_i) = \boldsymbol{R}_i(\gamma) &= \sigma^2_{PG}\boldsymbol{I}_4 + \sigma^2_{MG}\boldsymbol{I}_4 \quad \text{if } i \text{ is a girl,} \\ &= \sigma^2_{PB}\boldsymbol{I}_4 + \sigma^2_{MB}\boldsymbol{I}_4 \quad \text{if } i \text{ is a boy,} \end{aligned}$$

  say. This leads to the final specification

$$\text{var}(\boldsymbol{e}_i|\boldsymbol{a}_i) = \boldsymbol{R}_i(\gamma, \boldsymbol{a}_i) = \{\sigma^2_G I(g_i = 0) + \sigma^2_B I(g_i = 1)\}\boldsymbol{I}_4, \tag{6.31}$$

  where now $\sigma^2_G = \sigma^2_{PG} + \sigma^2_{MG}$ and $\sigma^2_B = \sigma^2_{PB} + \sigma^2_{MG}$ in (6.31) represent ***within-child variance*** due to ***both*** the realization and measurement error processes (rather than overall variance as in Section 5.2).

  If the much larger estimated within-child variance for boys is mainly an ***artifact*** of the unusual pattern for one boy, an alternative model is the ***default*** (6.6), $\boldsymbol{R}_i(\gamma) = \sigma^2\boldsymbol{I}_4$. Here, one might want to examine sensitivity of fitted models to the data from the "unusual" boy by, for example, deleting him from the analysis.

- Because there is not strong evidence of within-child correlation, it is natural to attribute the overall pattern of correlation mainly to ***among-child sources***. We can examine the ***induced representation*** of the component of overall covariance structure due to among-child sources as follows. For illustration, take for each i

$$\text{var}(\boldsymbol{b}_i|\boldsymbol{a}_i) = \boldsymbol{D} = \begin{pmatrix} D_{11} & D_{12} \\ D_{12} & D_{22} \end{pmatrix}.$$

It is then straightforward to show that (try it), with $\boldsymbol{Z}_i$ as in (6.30), $\boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^T$ has ***diagonal elements***

$$D_{11} + D_{22} t_j^2 + 2 D_{12} t_j, \quad j = 1, \ldots, 4, \tag{6.32}$$

and (j, j') off-diagonal element

$$D_{11} + D_{22} t_j t_{j'} + D_{12}(t_j + t_{j'}) \quad j, j' = 1, \ldots, 4. \tag{6.33}$$

(6.32) shows that this component of the induced overall covariance structure allows for among-individual variance that possibly changes with time, and (6.33) imposes a rather complicated pattern of ***among-individual covariance and correlation*** that is clearly ***nonstationary***. Thus, this component of the model is sufficiently flexible to capture complex covariance patterns.

Because the evidence is suggestive of an overall pattern that may be ***different by gender***, one possibility is to take $\text{var}(\boldsymbol{b}_i | \boldsymbol{x}_i)$ to depend on $\boldsymbol{a}_i$ (gender) as in (6.4) and

$$\text{var}(\boldsymbol{b}_i | \boldsymbol{a}_i) = \boldsymbol{D}(\boldsymbol{a}_i) = \boldsymbol{D}_G I(g_i = 0) + \boldsymbol{D}_B I(g_i = 1). \tag{6.34}$$

However, it is hard to judge to what extent the empirical evidence reflects a ***real difference***. The simpler common model $\text{var}(\boldsymbol{b}_i | \boldsymbol{a}_i) = \boldsymbol{D}$ may well be sufficient.

HIP REPLACEMENT STUDY: Recall from Section 5.2 that, for $m = 30$ subjects undergoing hip replacement (13 male, 15 female), hæmatocrit was measured at week 0, prior to surgery, and then ideally at weeks 1 2, and 3 thereafter, where some subjects are ***missing*** the week 2 and possibly baseline measure. Also available is patient age, so that $\boldsymbol{a}_i = (g_i, a_i)^T$, where gender $g_i = 0$ for females and $g_i = 1$ for males; and a_i is the age of the patient (years).

We can interpret the primary question of interest from a SS perspective to determine if there are differences between genders in ***individual-specific features*** of the pattern of change of hæmatocrit following hip replacement. As we demonstrate, we can also investigate associations between these features and age.

Taking this point of view, from Figure 5.2, a natural model for the individual subject trajectories is

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \beta_{2i} t_{ij}^2 + e_{ij}, \tag{6.35}$$

which allows each subject to have his/her own specific ***quadratic*** profile. The model (6.35) can be written succinctly as

$$\boldsymbol{Y}_i = \boldsymbol{C}_i \boldsymbol{\beta}_i + \boldsymbol{e}_i, \quad \boldsymbol{\beta}_i = \begin{pmatrix} \beta_{0i} \\ \beta_{1i} \\ \beta_{2i} \end{pmatrix}, \quad \boldsymbol{C}_i = \begin{pmatrix} 1 & t_{i1} & t_{i1}^2 \\ \vdots & \vdots & \vdots \\ 1 & t_{in_i} & t_{in_i}^2 \end{pmatrix}. \tag{6.36}$$

As for the dental study, we can allow individual-specific intercepts, linear terms, and quadratic terms to vary about ***typical*** or mean values for ***each gender***, and we can further allow ***typical*** or mean hæmatocrit at ***baseline*** to depend on ***age*** through the model specification

$$\begin{aligned} \beta_{0i} &= \{\beta_{0,M}(1-g_i) + \beta_{0,F}g_i\} + \{\beta_{3,M}(1-g_i) + \beta_{3,F}g_i\}a_i + b_{0i} \\ \beta_{1i} &= \beta_{1,M}(1-g_i) + \beta_{1,F}g_i + b_{1i} \\ \beta_{2i} &= \beta_{2,M}(1-g_i) + \beta_{2,F}g_i + b_{2i}, \end{aligned} \tag{6.37}$$

where $\boldsymbol{b}_i = (b_{0i}, b_{1i}, b_{2i})^T$ is a vector of ***random effects***. The models for the individual-specific linear (β_{1i}) and quadratic (β_{2i}) terms could be modified to also depend on age. Letting

$$\boldsymbol{\beta}_i = (\beta_{0i}\,,\ \beta_{1i}\,,\ \beta_{2i})^T,$$

the model (6.37) can be represented as

$$\boldsymbol{\beta}_i = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i,$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_{0,M} \\ \beta_{0,F} \\ \beta_{1,M} \\ \beta_{1,F} \\ \beta_{2,M} \\ \beta_{2,F} \\ \beta_{3,M} \\ \beta_{3,F} \end{pmatrix}, \quad \boldsymbol{A}_i = \begin{pmatrix} (1-g_i) & g_i & 0 & 0 & 0 & 0 & (1-g_i)a_i & g_i a_i \\ 0 & 0 & (1-g_i) & g_i & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & (1-g_i) & g_i & 0 & 0 \end{pmatrix}, \quad \boldsymbol{B}_i = \boldsymbol{I}_3. \tag{6.38}$$

Upon substitution into (6.36), we have (verify) that $\boldsymbol{Z}_i = \boldsymbol{C}_i$ and $\boldsymbol{X}_i$ is the $(n_i \times 8)$ matrix

$$\boldsymbol{X}_i = \begin{pmatrix} (1-g_i) & g_i & (1-g_i)t_{i1} & g_i t_{i1} & (1-g_i)t_{i1}^2 & g_i t_{i1}^2 & (1-g_i)a_i & g_i a_i \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ (1-g_i) & g_i & (1-g_i)t_{in_i} & g_i t_{in_i} & (1-g_i)t_{in_i}^2 & g_i t_{in_i}^2 & (1-g_i)a_i & g_i a_i \end{pmatrix}.$$

Specification of the ***within-individual*** covariance matrix $\boldsymbol{R}_i(\gamma)$ and the covariance matrix $\text{var}(\boldsymbol{b}_i|\boldsymbol{a}_i)$ proceeds according to the same considerations as above.

Assuming for illustration that we take $\text{var}(\boldsymbol{b}_i|\boldsymbol{a}_i) = \boldsymbol{D}$ for all i, $\boldsymbol{D}$ is a (3×3) matrix, and the component $\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T$ of $\boldsymbol{V}_i$ corresponding to ***among-individual sources*** is a $(n_i \times n_i)$ matrix whose elements have a rather ***complicated*** form (try it), depending on the six distinct elements of $\boldsymbol{D}$ as well as functions of time.

In many applications that, although ***in principle*** we expect that ***all of*** individual-specific intercepts, linear terms, and quadratic terms ***vary*** in the population, practically speaking, the ***induced overall covariance model*** $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \boldsymbol{R}_i(\gamma)$ depends on a rather large vector ξ of covariance parameters. Accordingly, the induced overall covariance structure is ***highly parameterized*** and is capable of representing complex true patterns of overall variance and correlation.

It may well be that, even though quadratic terms β_{2i} ***do vary*** in the population, ***relative*** to the extent of variation in intercepts and linear terms β_{0i} and β_{1i}, this variation is ***practically negligible***. Accordingly, it is not uncommon under ***quadratic*** and ***higher-order polynomial*** individual-specific models to ***simplify*** the model for β_i by ***eliminating*** random effects associated with quadratic and higher terms. This entails redefining $\boldsymbol{Z}_i$ and $\boldsymbol{D}$ accordingly.

The resulting $\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T$ may still be ***sufficiently rich*** to approximate the true component of among-individual covariance, and the induced overall structure may still be sufficiently parametrized to capture the true overall pattern. In addition, from a ***computational perspective***, the resulting model is likely to be less burdensome and problematic to fit; see below.

We demonstrate by ***eliminating*** the random effect b_{2i} in the specification for β_{2i} in (6.37), replacing it by

$$\beta_{2i} = \beta_{2,M}(1 - g_i) + \beta_{2,F}g_i. \tag{6.39}$$

Strictly speaking, (6.39) implies that the quadratic term in (6.35) is the ***same*** for all males and for all females. While this is likely an oversimplification, as an approximation it enjoys the advantages noted above. Under (6.39),

$$\boldsymbol{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix}, \quad \boldsymbol{B}_i = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \text{so that} \quad \boldsymbol{B}_i\boldsymbol{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \\ 0 \end{pmatrix}. \tag{6.40}$$

Of course, from a ***subject-specific*** perspective, this is strictly an approximation of ***convenience***, as we most certainly do not really believe that individuals of each gender have ***individual-specific trajectories*** characterized by ***exactly the same*** quadratic component.

RELATIVE MAGNITUDES OF AMONG-INDIVIDUAL VARIATION: This foregoing demonstration with the hip replacement study scenario exemplifies an important ***general consideration*** when specifying linear mixed effects models. Although, ***conceptually***, from a SS point of view, all individual-specific parameters are expected to exhibit variation in the population, it is their ***relative magnitudes of variation*** that are practically important.

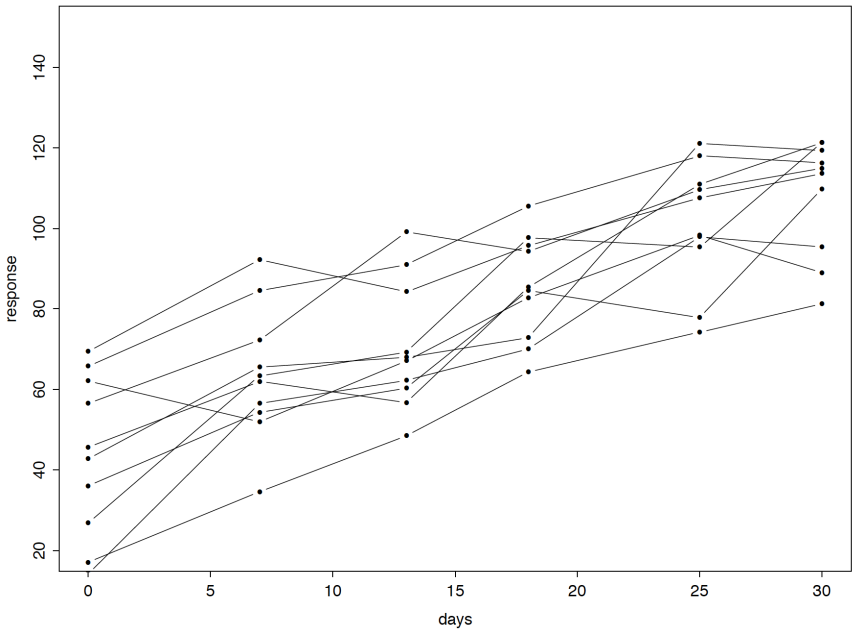

Figure 6.1: *Longitudinal data where variation in slope may be negligible.*

Consider the situation in Figure 6.1, depicting trajectories for 10 individuals for which a ***straight line inherent trend*** is a reasonable characterization. The ***individual-specific intercepts*** clearly vary substantially, but the assumed underlying lines appear to have ***very similar slopes***. Although scientifically it is reasonable to expect that individual rates of change ***should vary*** , e.g., as would be expected with patterns of growth across individual subjects or plots, ***relative*** to the variation in intercepts, the variation in slopes may well be ***orders of magnitude*** smaller.

For simplicity, assume there are no covariates. Letting β_{0i} and β_{1i} be the intercept and slope for individual i, if we assume

$$\beta_{0i} = \beta_0 + b_{0i}, \qquad \beta_{1i} = \beta_1 + b_{1i},$$

$\boldsymbol{b}_i = (b_{0i}, b_{1i})^T$, if we take var($\boldsymbol{b}_i$) = $\boldsymbol{D}$, D_{11} represents the variance of intercepts and D_{22} that of slopes. If D_{11} is ***nonnegligible*** relative to the mean intercept β_0, then intercepts vary perceptibly, but if D_{22} is ***virtually negligible*** relative to the size of the mean slope β_1, then variation in slopes is almost undetectable.

In such a situation, optimization algorithms involved in the implementation of inference by ML or REML, as discussed in the next section, can fail, as D_{22} and in fact the covariance D_{12} are not ***practically identifiable*** under these circumstances.

It is commonplace under these conditions to invoke an ***approximation*** analogous to that in (6.39) to achieve numerical stability, namely,

$$\beta_{0i} = \beta_0 + b_{0i}, \qquad \beta_{1i} = \beta_1. \tag{6.41}$$

This does not mean that we "***believe***" slopes do not vary ***at all*** in the population; rather, this is an ***approximation*** recognizing that their magnitude of variation is inconsequential ***relative*** to that of other phenomena, which allows implementation of the model to be feasible. The inclusion of the design matrix $\boldsymbol{B}_i$ in the general model specification accommodates this possibility.

In a model like (6.41), it is popular to distinguish between individual-specific features being "***fixed***" or "***random***;" in (6.41), β_{0i} would be said to be "random" while β_{1i} would be referred to as "fixed."

In Section 6.6, we discuss this and related issues further.

HIV CLINICAL TRIAL: It should be clear that the ***hierarchical framework*** of the model offers great latitude for thinking about and representing individual-specific and population-level phenomena. As a final brief example, consider ACTG Study 193A, introduced in Section 5.2. Here, subjects were ***randomized*** to four treatment regimens, with age and gender recorded at baseline, so that $\boldsymbol{a}_i = (g_i, a_i, \delta_{i1}, \ldots, \delta_{i4})^T$, where $g_i = 0$ (1) for a female (male) subject; a_i is age; and $\delta_{i\ell} = 1$ if subject i was randomized to treatment regimen ℓ and 0 otherwise, $\ell = 1, \ldots, 4$.

From Figure 5.3, a reasonable approximation is to assume that each subject has his/her own inherent underlying ***straight line*** log(CD4+1) trajectory,

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij},$$

where now β_{0i} represents individual i's inherent mean log(CD4+1) immediately prior to initiation of therapy. Although we thus would not expect the β_{0i} to be associated with randomized treatment, they may be associated with individual characteristics such as gender and age.

If interest focuses on comparing the ***patterns of change*** of log(CD4 +1) among the four regiimens, from a SS point of view, this can be cast as comparing the ***typical*** or mean slopes under the four regimens. A model that incorporates baseline associations with covariates and allows typical slopes to differ across treatments is

$$\begin{aligned} \beta_{0i} &= \beta_{00} + \beta_{01} a_i + \beta_{02} g_i + b_{0i}, \\ \beta_{1i} &= \beta_{10} + \beta_{11}\delta_{i1} + \beta_{12}\delta_{i2} + \beta_{13}\delta_{i3} + b_{1i}. \end{aligned}$$

This model could be further modified to allow way in which slopes differ across treatments to be different for each gender or to depend on age.

HIERARCHICAL MODEL SUMMARY: The ***linear mixed effects*** model is often presented formally as a ***two-stage hierarchy*** as follows. In its usual general form , for each $i = 1, \ldots, m$,

Stage 1 - Individual model.

$$\boldsymbol{Y}_i = \boldsymbol{C}_i\beta_i + \boldsymbol{e}_i \quad (n_i \times 1), \qquad \boldsymbol{e}_i|\boldsymbol{x}_i \sim \mathcal{N}\{\boldsymbol{0}, \boldsymbol{R}_i(\gamma)\}, \tag{6.42}$$

where $\boldsymbol{C}_i$ is a $(n_i \times k)$ ***design matrix*** ordinarily depending on the ***times*** $t_{i1}, \ldots, t_{in_i}$, and β_i is a $(k \times 1)$ vector of ***individual-specific regression parameters***. The regression parameter β_i can be viewed as determining individual i's ***inherent trajectory***.

The default is that $\boldsymbol{e}_i$ is independent of $\boldsymbol{x}_i$ and β_i and thus $\boldsymbol{b}_i$, , although, as we have observed, (6.42) is often generalized to allow dependence on $\boldsymbol{a}_i$, so that $\boldsymbol{R}_i(\gamma)$ depends on $\boldsymbol{a}_i$. In more general versions of this hierarchy discussed in Chapter 9, dependence on β_i and thus $\boldsymbol{b}_i$ is also allowed.

In addition, $\boldsymbol{R}_i(\gamma)$ can be decomposed into components due to the ***within-individual realization*** and ***measurement error processes***, as in (6.17.

Stage 2- Population model.

$$\beta_i = \boldsymbol{A}_i\beta + \boldsymbol{B}_i\boldsymbol{b}_i \quad (k \times 1), \qquad \boldsymbol{b}_i|\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}) \quad (q \times 1), \tag{6.43}$$

where β $(p \times 1)$ is a vector of ***fixed effects***; $\boldsymbol{A}_i$ $(k \times p)$ and $\boldsymbol{B}_i$ $(k \times q)$ are ***design matrices***; and $k = q$ in many cases, although models with $k > q$ are sometimes specified when some components of β_i are thought to vary ***negligibly*** among individuals. Typically, $\boldsymbol{A}_i$ incorporates ***among-individual covariates*** , while $\boldsymbol{B}_i$ is comprised of 0s and 1s and serves to indicate which elements of β_i are treated as "***random***" and which are treated as "***fixed***."

The default is that $\boldsymbol{b}_i$ and $\boldsymbol{x}_i$ are independent, but, as we have seen, this can be relaxed to allow dependence on $\boldsymbol{a}_i$, so that $\boldsymbol{D}(\boldsymbol{a}_i)$ depends on $\boldsymbol{a}_i$.

Substituting the ***population model*** (6.43) in the ***individual model*** (6.42) yields the ***linear mixed effects model***

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i, \quad \boldsymbol{b}_i|\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}), \quad \boldsymbol{e}_i|\boldsymbol{x}_i \sim \mathcal{N}\{\boldsymbol{0}, \boldsymbol{R}_i(\gamma)\}, \tag{6.44}$$

where $\boldsymbol{X}_i = \boldsymbol{C}_i\boldsymbol{A}_i$ $(n_i \times p)$ is the ***fixed effects design matrix*** , $\boldsymbol{Z}_i = \boldsymbol{C}_i\boldsymbol{B}_i$ $(n_i \times q)$ is the ***random effects design matrix*** , and the usual assumptions on the conditional distributions of $\boldsymbol{e}_i$ and $\boldsymbol{b}_i$ can be relaxed if need be.

6.3 Inference and considerations for missing data

IMPLIED POPULATION-AVERAGED MODEL: As shown in (6.10) and (6.11), given a particular specification of the ***two-stage hierarchy*** in (6.42) and (6.43) leading to a ***linear mixed effects model*** as in (6.44), we are led to a ***population-averaged*** model of the form

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{X}_i\beta, \quad \text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_i = \boldsymbol{V}_i(\xi, \boldsymbol{x}_i), \quad \boldsymbol{V}_i(\xi, \boldsymbol{x}_i) = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \boldsymbol{R}_i(\gamma), \quad \xi = \{\gamma^T, \text{vech}(\boldsymbol{D})^T\}^T, \tag{6.45}$$

where $(\boldsymbol{Y}_i, \boldsymbol{x}_i)$, $i = 1, \ldots, m$, are ***independent***. The model (6.45) is of course of the same form as the models considered in Chapter 5, where the covariance matrix $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ is of the ***particular form*** in (6.45). The model (6.45) can be expressed succinctly as in (6.16) as

$$E(\boldsymbol{Y}|\tilde{\boldsymbol{x}}) = \boldsymbol{X}\beta, \quad \text{var}(\boldsymbol{Y}|\tilde{\boldsymbol{x}}) = \boldsymbol{V}(\xi, \tilde{\boldsymbol{x}}) = \boldsymbol{V} = \boldsymbol{Z}\widetilde{\boldsymbol{D}}\boldsymbol{Z}^T + \boldsymbol{R}. \tag{6.46}$$

The specifications in (6.45) and (6.46) can of course be ***generalized*** to allow a more general ***among-individual covariance matrix*** of the form $\text{var}(\boldsymbol{b}_i|\boldsymbol{z}_i) = \boldsymbol{D}(\boldsymbol{a}_i)$.

ESTIMATION OF β AND ξ: From (6.45) and (6.46), it should be clear that, under the ***normality*** assumptions at each stage of the hierarchy (6.42) and (6.43), it follows that the distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ is assumed to be n_i-variate normal, i.e.,

$$\boldsymbol{Y}_i|\boldsymbol{x}_i \sim \mathcal{N}\{\boldsymbol{X}_i\beta, \boldsymbol{V}_i(\xi, \boldsymbol{x}_i)\}.$$

It follows that estimators for β and ξ can be obtained by appealing to the developments in Sections 5.3 and 5.4. That is, β and ξ can be estimated by solving the estimating equations corresponding to ***maximum likelihood*** or ***REML***.

LARGE SAMPLE INFERENCE: Moreover, the ***large sample*** results in Section 5.5 go through unchanged. Thus, the ***approximate*** sampling distributions for the estimator $\widehat{\beta}$ obtained using either ML or REML can be used as described in that section. Namely, the ***model-based*** result in (5.68),

$$\widehat{\beta} \dot{\sim} \mathcal{N}(\beta_0, \widehat{\boldsymbol{\Sigma}}_M), \qquad \widehat{\boldsymbol{\Sigma}}_M = \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\widehat{\xi}, \boldsymbol{x}_i)\boldsymbol{X}_i\right)^{-1} = \{\boldsymbol{X}^T\boldsymbol{V}^{-1}(\widehat{\xi}, \tilde{\boldsymbol{x}})\boldsymbol{X}\}^{-1}, \tag{6.47}$$

can be used as the basis for inference on β.

Likewise, the ***robust*** or ***empirical*** result in (5.81) and (5.82),

$$\widehat{\beta} \dot{\sim} \mathcal{N}(\beta_0, \widehat{\boldsymbol{\Sigma}}_R), \tag{6.48}$$

$$\widehat{\boldsymbol{\Sigma}}_R = \tag{6.49}$$
$$\left\{\sum_{i=1}^{m} \boldsymbol{X}_i \boldsymbol{V}_i^{-1}(\widehat{\xi}, \boldsymbol{x}_i)\boldsymbol{X}_i\right\}^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}\widehat{\xi}, \boldsymbol{x}_i)(Y_i - \boldsymbol{X}_i\widehat{\beta})(Y_i - \boldsymbol{X}_i\widehat{\beta})^T \boldsymbol{V}_i^{-1}\widehat{\xi}, \boldsymbol{x}_i)\boldsymbol{X}_i \left\{\sum_{i=1}^{m} \boldsymbol{X}_i \boldsymbol{V}_i^{-1}\widehat{\xi}, \boldsymbol{x}_i)\boldsymbol{X}_i\right\}^{-1}.$$

can also be used. Both (6.47) and (6.48) and (6.49) can be used for inference on ***linear*** functions $\boldsymbol{L}\beta$ as in (5.85). This inference can be from a SS or PA perspective in accordance with the scientific questions. The analyst should be careful to be clear about this.

As with the models in Chapter 5, the true distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ need not be ***normal*** for these approximations to be valid (except when there are missing data; see below).

INFORMATION CRITERIA: The ***information criteria*** (5.90) – (5.92) discussed in Section 5.5 can also be used to compare models that are not nested and in particular to compare different specifications of the ***overall covariance structure*** that are ***induced*** by combinations of choices of models for, say var($\boldsymbol{e}_i|\boldsymbol{x}_i$) and var($\boldsymbol{b}_i|\boldsymbol{a}_i$).

- For example, for the dental study, one could compare the specifications of a ***common among-individual*** covariance matrix, var($\boldsymbol{b}_i|\boldsymbol{a}_i$) = var($\boldsymbol{b}_i$) = $\boldsymbol{D}$ to taking

  $$\text{var}(\boldsymbol{b}_i|\boldsymbol{a}_i) = \boldsymbol{D}(\boldsymbol{a}_i) = \boldsymbol{D}_G I(g_i = 0) + \boldsymbol{D}_B I(g_i = 1)$$

  as in (6.34).

- Likewise, one could compare taking var($\boldsymbol{e}_i|\boldsymbol{x}_i$) = $\sigma^2\boldsymbol{I}_4$ for all children versus allowing a separate ***within-child variance*** for each gender,

  $$\text{var}(\boldsymbol{e}_i|\boldsymbol{a}_i) = \boldsymbol{R}_i(\gamma, \boldsymbol{a}_i) = \{\sigma_G^2 I(g_i = 0) + \sigma_B^2 I(g_i = 1)\}\boldsymbol{I}_4$$

  as in (6.31).

MISSING DATA: The ***same*** implications of ***missing data*** discussed in Section 5.6 apply to the linear mixed effects model. In particular, ***under the assumptions of a MAR mechanism*** and ***normality*** , the estimators for β and ξ are ***consistent*** , and the large sample approximation to the sampling distribution of $\widehat{\beta}$ as in (6.47) can be used, but with, ideally, $\widehat{\boldsymbol{\Sigma}}_M$ replaced by the appropriate element of the inverse of the ***observed information matrix***. The approximation in (6.48) and (6.49) should ***not*** be used, as discussed in Section 5.6.

BALANCED DATA: There is an interesting curiosity in the case of ***balanced*** data, so that $\boldsymbol{Y}_i$ is $(n \times 1)$ for all $i = 1, \ldots, m$, with components observed at the ***same*** n time points. In this case, $\boldsymbol{Z}_i = \boldsymbol{Z}^*$, say, is the same for all i (verify). If the linear mixed effects model specification is such that

$$\boldsymbol{R}_i(\gamma) = \sigma^2 \boldsymbol{I}_n,$$

the induced overall covariance matrix for each i is

$$\boldsymbol{V}_i(\xi, \boldsymbol{x}_i) = \boldsymbol{V}^* = \boldsymbol{Z}^* \boldsymbol{D} \boldsymbol{Z}^{*T} + \sigma^2 \boldsymbol{I}_n, \tag{6.50}$$

say. Then, under certain conditions, letting

$$\widehat{\boldsymbol{V}}^* = \boldsymbol{Z}^* \widehat{\boldsymbol{D}} \boldsymbol{Z}^{*T} + \widehat{\sigma}^2 \boldsymbol{I}_n,$$

where $\widehat{\boldsymbol{D}}$ and $\widehat{\sigma}^2$ are the estimators for $\boldsymbol{D}$ and σ^2 obtained by ML or REML, the estimator

$$\widehat{\beta} = \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \widehat{\boldsymbol{V}}^{*-1} \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \widehat{\boldsymbol{V}}^{*-1} \boldsymbol{Y}_i$$

and the ***ordinary least squares*** estimator

$$\widehat{\beta}_{OLS} = \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{Y}_i$$

are ***numerically identical***.

- This follows because it can be shown by cleverly applying ***matrix inversion results*** given in Appendix A that, with overall covariance structure $\boldsymbol{V}$ as in (6.50), the expressions

  $$\left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}^{*-1} \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}^{*-1} \boldsymbol{Y}_i \quad \text{and} \quad \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{Y}_i$$

  are ***equivalent***.

- This continues to hold even if σ^2 and $\boldsymbol{D}$ in (6.50) take on different values corresponding to different levels of an ***among-individual covariate***, as for the dental study, where these are taken to ***differ by gender***.

- Demonstration of this equivalence is left as an exercise for the ***diligent student***.

- Note that, although the ML/REML estimator and the OLS estimator are numerically equivalent, this does ***not*** mean that one can ***disregard*** the need to characterize covariance structure and just take all N observations to be ***mutually independent***.

  Correct characterization of the ***sampling distribution*** of the estimator requires that the overall covariance be ***acknowledged and modeled***, and the large sample approximate sampling distribution depends on this assumed structure.

POPULATION-AVERAGED VERSUS SUBJECT-SPECIFIC PERSPECTIVE: As we have observed, the linear mixed effects model can be viewed in different ways.

- We motivated the model from a ***subject-specific perspective***, which dovetails naturally with the ***conceptual framework*** for longitudinal data we introduced in Section 2.3. This perspective underlies the view of the model as a ***two-stage hierarchy***, as presented in Section 6.2, which involves an ***individual-level model*** expressed in terms of ***individual-specific regression parameters*** and a ***population-level model*** that characterizes how these parameters ***vary*** in the population of individuals due to (i) ***systematic associations*** with among-individual covariates and (ii) "***unexplained***" or "***natural***" sources, such as biological differences or unobserved covariates.

  This view is natural when the questions of scientific interest involve ***subject-specific phenomena***.

- This formulation also ***implies*** a ***population-averaged model***, where the form of the ***overall covariance structure*** incorporating components due to ***among-*** and ***within-individual sources*** is ***induced***. Thus, an alternative perspective on the linear mixed effects model is as a population-averaged model for which specification of a form for the overall covariance structure is facilitated "automatically" rather than chosen explicitly by the data analyst. This relieves the analyst from the often ***challenging task*** of specifying a suitable overall structure. Moreover, the induced form for the overall covariance structure dictated by the linear mixed model is ***sufficiently rich***, involving a number of parameters, that it is likely able to represent well very ***complicated, nonstationary*** patterns of overall variance and correlation, as exemplified by (6.32) and (6.33).

  Thus, it is common to adopt a linear mixed effects model even when the questions of scientific interest involve ***population-averaged phenomena***.

- As we have already emphasized, the ***fixed effects*** β and questions posed in terms of them can be interpreted from either perspective.

- However, the perspective under which the model is adopted has ***implications for inference***, in particular in regard to the interpretation and fitting of the ***overall covariance structure***. Clearly, from either perspective, we desire a model that captures the ***salient features*** of covariance so that inferences on β will be reliable. At the same time, the model should not involve more parameters to be estimated than necessary, which in finite samples can ***degrade precision of estimation*** of β (despite the optimistic first-order asymptotic theory).

- As noted above, from a ***population-averaged*** perspective, the ***induced*** form of the overall covariance structure is a convenient and flexible way of represented what might possibly be a complex structure. From this point of view, $\xi = \{\gamma^T, \text{vech}(\boldsymbol{D})^T\}^T$ is simply a vector of parameters that characterizes the structure, and thus there are ***no restrictions*** on possible values of ξ. In particular, $\boldsymbol{D}$ need not be restricted to be a legitimate covariance matrix, with non-negative diagonal elements. Likewise, γ need not be restricted to take on values that render $\boldsymbol{R}_i(\gamma)$ a legitimate covariance matrix. What matters is that the parameterization in terms of ξ can represent a legitimate ***overall covariance structure***.

- From a subject-specific perspective, however, the separate components $\boldsymbol{D}$ and $\boldsymbol{R}_i(\gamma)$ ***are interpreted*** as covariance matrices corresponding to ***among-*** and ***within-individual*** sources of variation and correlation. Thus, from this point of view, there ***are restrictions*** on the parameter space of $\xi = \{\gamma^T, \text{vech}(\boldsymbol{D})^T\}^T$ that ensure that these are legitimate covariance matrices, that is, positive (semi-) definite matrices. Thus, for example, the diagonal elements of $\boldsymbol{D}$ are restricted to be nonnegative.

- Accordingly, which perspective is relevant ***will dictate*** how assessment of and inference on the assumed covariance structure takes place. We discuss this in more detail in Section 6.6.

6.4 Best linear unbiased prediction and empirical Bayes

RANDOM EFFECTS: Ordinarily, the primary objective of an analysis is to address questions of scientific interest expressed in terms of the ***fixed effects*** β, which may have ***either*** a population-averaged or subject-specific interpretation.

When a ***subject-specific*** perspective is adopted, the ***two-stage hierarchical interpretation*** of the linear mixed effects model reflects the belief that each individual has specific regression parameters β_i characterizing his/her ***inherent trajectory***. The β_i are then represented in the population model as depending on individual-specific ***random effects*** $\boldsymbol{b}_i$ that reflect how *i*'s regression parameters deviate from the "***typical***" values and likewise how *i*'s inherent trajectory deviates from the overall population mean profile. The $\boldsymbol{b}_i$ are ***random vectors*** assumed to arise from a probability distribution(s) that characterizes the extent of variation in these features in the population.

In the ***standard version*** of the linear mixed effects model we discuss in this chapter, the distribution of the $\boldsymbol{b}_i$ is taken to be ***q-variate normal***, with mean zero with covariance matrix $\boldsymbol{D}$ (which of course can be relaxed to allow ***separate*** distributions for each level of an among-individual covariate). For the discussion here, we take

$$\boldsymbol{b}_i|\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}), \quad i = 1, \ldots, m; \tag{6.51}$$

the developments below of course can be generalized.

Thus, from a subject-specific point of view, it is often of interest to "***estimate***" $\boldsymbol{b}_i$ for each individual. These estimates can be used for ***diagnostic purposes***, e.g., to identify individuals or groups of individuals whose profiles over time may be ***outlying*** relative to the bulk of the population. They can also be used to characterize ***individual-specific trajectories***.

Because the $\boldsymbol{b}_i$ are random vectors, each corresponding to a ***randomly chosen individual*** from the population, characterizing $\boldsymbol{b}_i$ is akin to ***predicting*** the value taken on by a random vector corresponding to a randomly chosen individual. Thus, inference on $\boldsymbol{b}_i$ is often regarded as a ***prediction problem***. Because $\boldsymbol{Y}_i$ contains information about $\boldsymbol{b}_i$, it is natural to view this prediction problem as characterizing $\boldsymbol{b}_i$ ***given*** that we have observed $\boldsymbol{Y}_i = \boldsymbol{y}_i$. The usual approach is to use as a predictor the value that is "***most likely***" given that we have observed $\boldsymbol{Y}_i = \boldsymbol{y}_i$.

BAYESIAN PERSPECTIVE: It is thus natural to consider this problem based on a ***Bayesian*** formulation and to "***estimate***" $\boldsymbol{b}_i$ by the value that ***maximizes*** the ***posterior distribution*** of $\boldsymbol{b}_i$ given $\boldsymbol{Y}_i$ evaluated at $\boldsymbol{y}_i$; that is, finding the ***posterior mode***.

- In the Bayesian view of the linear mixed effect model, the $\boldsymbol{b}_i$ are regarded as ***parameters***, and the probability distribution (6.51) is referred to as the ***prior distribution*** for them.
- For the discussion here, we do not consider the parameters β and ξ from the classical Bayesian perspective as random quantities with suitable prior distributions, but treat them as ***fixed and known***; more on this momentarily.

Taking this point of view, let as in (6.9)

$$p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma) \tag{6.52}$$

be the density of the assumed conditional normal distribution

$$\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i \sim \mathcal{N}\{\boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i, \boldsymbol{R}_i(\gamma)\}.$$

Let

$$p(\boldsymbol{b}_i; \boldsymbol{D})$$

be the density corresponding to (6.51). Then, identifying this as the "***prior***" and (6.52) as the "***likelihood***," by Bayes' theorem, the ***posterior density*** of $\boldsymbol{b}_i$ conditional on observing $\boldsymbol{Y}_i = \boldsymbol{y}_i$ is given by

$$p(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{x}_i; \beta, \gamma, \boldsymbol{D}) = \frac{p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)\, p(\boldsymbol{b}_i; \boldsymbol{D})}{p(\boldsymbol{y}_i|\boldsymbol{x}_i; \beta, \gamma, \boldsymbol{D})}, \tag{6.53}$$

where, from (6.9),

$$p(\boldsymbol{y}_i|\boldsymbol{x}_i; \beta, \gamma, \boldsymbol{D}) = \int p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)\, p(\boldsymbol{b}_i; \boldsymbol{D})\, d\boldsymbol{b}_i.$$

It is straightforward to verify (do it) that, under the normal specifications (6.51) and (6.52), the ***posterior distribution*** with density (6.53) is ***also normal*** with ***mean***

$$\boldsymbol{D}\boldsymbol{Z}_i^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)\{\boldsymbol{y}_i - \boldsymbol{X}_i\beta\}. \tag{6.54}$$

Because the ***mean*** of a normal distribution is also the ***mode*** of the density, the expression (6.54) also satisfies the requirement that it ***maximizes*** the posterior density.

EMPIRICAL BAYES: From (6.54), it is natural to substitute estimators $\widehat{\beta}$ and $\widehat{\xi}$ for β and ξ, which yields the so-called ***empirical Bayes*** "***estimator***" for $\boldsymbol{b}_i$ given by

$$\widehat{\boldsymbol{b}}_i = \widehat{\boldsymbol{D}}\boldsymbol{Z}_i^T \boldsymbol{V}_i^{-1}(\widehat{\xi}, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\beta}\}, \tag{6.55}$$

where we have written (6.55) as depending on the response vector $\boldsymbol{Y}_i$ with the understanding that the actual observed value of $\boldsymbol{Y}_i$ is substituted in forming the "***estimate***."

If ξ were ***known***, so that (6.55) becomes

$$\widehat{\boldsymbol{b}}_i = \boldsymbol{D}\boldsymbol{Z}_i^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\beta}\}, \tag{6.56}$$

it is straightforward (try it) to show that (conditional on $\tilde{\boldsymbol{x}}$) $\widehat{\boldsymbol{b}}_i$ in (6.56) has mean zero and covariance matrix

$$\text{var}(\widehat{\boldsymbol{b}}_i|\tilde{\boldsymbol{x}}) = \boldsymbol{D}\boldsymbol{Z}_i^T \left\{ \boldsymbol{V}_i^{-1} - \boldsymbol{V}_i^{-1}\boldsymbol{X}_i \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right)^{-1} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \right\} \boldsymbol{Z}_i \boldsymbol{D}, \tag{6.57}$$

where we have used the streamlined notation for $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$.

Because what we really are doing is ***prediction*** of the "***moving target***" $\boldsymbol{b}_i$, which is a random rather than fixed quantity, (6.57) is known to ***understate*** the variability in $\widehat{\boldsymbol{b}}_i$.

Accordingly, it is recommended to instead use

$$\text{var}(\widehat{\boldsymbol{b}}_i - \boldsymbol{b}_i|\tilde{\boldsymbol{x}}) = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{Z}_i^T \left\{ \boldsymbol{V}_i^{-1} - \boldsymbol{V}_i^{-1}\boldsymbol{X}_i \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right)^{-1} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \right\} \boldsymbol{Z}_i \boldsymbol{D} \tag{6.58}$$

(verify). Of course, in practice, ξ is replaced by its estimator (ML or REML), in which case (6.57) and (6.58) both understate the variability in $\widehat{\boldsymbol{b}}_i$ as a ***predictor*** of $\boldsymbol{b}_i$.

Laird and Ware (1982) and Davidian and Giltinan (1995, Section 3.3) offer more discussion.

REMARK: It is possible to arrive at (6.55) directly by an argument similar to that above using Bayes theorem, where ξ is treated as known but β is viewed instead as a ***random vector independent*** of $\boldsymbol{b}_i$ with ***prior density*** $p(\beta|\beta^*, \boldsymbol{H})$ depending on ***hyperparameters*** β^* and $\boldsymbol{H}$ corresponding to the $\mathcal{N}(\beta^*, \boldsymbol{H})$ distribution.

Under these conditions, the posterior densities of $\boldsymbol{b}_i$ and β can be derived. If one assumes ***vague*** prior information on β by setting $\boldsymbol{H}^{-1} = \boldsymbol{0}$, it can be shown that mean of the posterior density for β is $\widehat{\beta}$ and that for $\boldsymbol{b}_i$ is (6.55). The details are presented in Davidian and Giltinan (1995, Section 3.3).

BEST LINEAR UNBIASED PREDICTION (BLUP): Putting the Bayesian interpretation aside, we consider another perspective on (6.55). A standard principle in statistics is that a "***best***" predictor is one that ***minimizes mean squared error***. Namely, here, $\boldsymbol{c}(\boldsymbol{Y}_i)$ is the best predictor if it minimizes

$$E[\{\boldsymbol{c}(\boldsymbol{Y}_i) - \boldsymbol{b}_i\}^T \boldsymbol{A} \{\boldsymbol{c}(\boldsymbol{Y}_i) - \boldsymbol{b}_i\}], \tag{6.59}$$

where this expectation is with respect to the joint distribution of $\boldsymbol{Y}_i$ and $\boldsymbol{b}_i$, and $\boldsymbol{A}$ is any positive definite symmetric matrix. It is a fundamental result that the ***best predictor*** in the sense of minimizing (6.59) is

$$E(\boldsymbol{b}_i|\boldsymbol{Y}_i), \tag{6.60}$$

which does not depend on $\boldsymbol{A}$. The argument is straightforward and proceeds by ***adding and subtracting*** (6.60) to each of the terms in braces in (6.59) and rearranging to show that $\boldsymbol{c}(\boldsymbol{Y}_i) = E(\boldsymbol{b}_i|\boldsymbol{Y}_i)$ (the diligent student will be sure to try this).

Thus, ***under the usual normality assumptions*** for the linear mixed model, the developments above show that (6.55) with β replacing $\widehat{\beta}$ and ξ ***known*** is "***best***" in this sense. Because (6.55) is also ***linear*** in $\boldsymbol{Y}_i$, it is the best ***linear function*** of $\boldsymbol{Y}_i$ to use as a predictor under normality.

In general, the best predictor (6.60) ***need not*** be linear. However, if attention is restricted to predictors $\boldsymbol{c}(\boldsymbol{Y}_i)$ that are ***linear*** functions of $\boldsymbol{Y}_i$, it is can be shown that, ***without any normality assumptions*** (and ξ known), (6.55) is the ***best linear unbiased predictor*** for $\boldsymbol{b}_i$ in the sense that it minimizes the mean squared error, is a ***linear*** function of $\boldsymbol{Y}_i$, and is such that $E(\widehat{\boldsymbol{b}}_i) = E(\boldsymbol{b}_i) = \boldsymbol{0}$.

We do not provide the argument here; Searle, Casella, and McCulloch (2006, Chapter 7) and Robinson (1991) offer detailed derivations.

In practice, ξ is replaced by the ML or REML estimator $\widehat{\xi}$, in which case some authors have referred to the resulting predictor as an ***estimated best linear unbiased predictor*** or ***EBLUP***.

In the linear mixed effects model literature, the term ***BLUP*** , ***empirical Bayes estimator*** , and ***EBLUP*** are often used ***interchangeably***.

HENDERSON'S MIXED MODEL EQUATIONS: Yet another approach to deducing a ***predictor*** for $\boldsymbol{b}_i$ is due to Henderson (1984). It is customary to present this using the "stacked" notation in (6.13) – (6.15). Here, we treat ξ and thus γ, $\boldsymbol{R}$, and $\boldsymbol{D}$ as ***known***.

For known ξ, Henderson proposes to "***estimate***" the $\boldsymbol{b}_i$, $i = 1, \ldots, m$, which are stacked in the vector $\boldsymbol{b}$, jointly with β, by minimizing in β and $\boldsymbol{b}$ the ***objective function***

$$\log|\widetilde{\boldsymbol{D}}| + \boldsymbol{b}^T\widetilde{\boldsymbol{D}}^{-1}\boldsymbol{b} + \log|\boldsymbol{R}| + (\boldsymbol{Y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b})^T\boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b}), \tag{6.61}$$

which under normality is twice the negative log of the ***posterior density*** of $\boldsymbol{b}$ for fixed β and twice the negative loglikelihood for β holding $\boldsymbol{b}$ fixed.

Differentiating (6.61) with respect to β and $\boldsymbol{b}$ using the matrix differentiation rules in Appendix A and setting equal to zero yields

$$\begin{aligned} \boldsymbol{X}^T\boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b}) &= \boldsymbol{0} \\ \widetilde{\boldsymbol{D}}\boldsymbol{b} - \boldsymbol{Z}^T\boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b}) &= \boldsymbol{0}, \end{aligned}$$

which can be rearranged to yield (verify) the so-called ***mixed model equations***

$$\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \widetilde{\boldsymbol{D}}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\beta} \\ \widehat{\boldsymbol{b}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Y} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Y} \end{pmatrix}. \tag{6.62}$$

It can be shown by demonstrating that

$$\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \widetilde{\boldsymbol{D}}^{-1})\boldsymbol{Z}^T\boldsymbol{R}^{-1} = (\boldsymbol{R} + \boldsymbol{Z}\widetilde{\boldsymbol{D}}\boldsymbol{Z}^T)^{-1} = \boldsymbol{V}^{-1},$$

which can be derived using matrix inversion results in Appendix A, that the solutions to (6.62) are

$$\widehat{\beta} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}, \qquad \widehat{\boldsymbol{b}} = \widetilde{\boldsymbol{D}}\boldsymbol{Z}^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}),$$

from whence the expression (6.55) for $\widehat{\boldsymbol{b}}_i$ follows.

SHRINKAGE: We demonstrate that empirical Bayes estimators (BLUPs) have the well-known property of "***shrinking***" predictions toward the mean in the sense we now describe. Consider (6.55) with ξ ***known***, that is

$$\widehat{\boldsymbol{b}}_i = \boldsymbol{D}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\beta}). \tag{6.63}$$

First consider the simplest special case of the linear mixed model, where $\boldsymbol{X}_i = \boldsymbol{1}_{n_i}$ for all i and $p = 1$ and $\boldsymbol{Z}_i = \boldsymbol{1}_{n_i}$ for all i and $q = 1$, so that the Y_{ij} have common scalar mean β for $j = 1, \ldots, n_i$, and the random effect is a ***scalar***; that is,

$$\boldsymbol{Y}_i = \boldsymbol{1}_{n_i}\beta + \boldsymbol{1}_{n_i}b_i + \boldsymbol{e}_i, \tag{6.64}$$

where $\text{var}(\boldsymbol{e}_i|\boldsymbol{x}_i) = \text{var}(\boldsymbol{e}_i) = \sigma^2\boldsymbol{I}_{n_i}$ and $\text{var}(b_i|\boldsymbol{x}_i) = \text{var}(b_i) = D$, a scalar. Then $\boldsymbol{V}_i = D\boldsymbol{J}_{n_i} + \sigma^2\boldsymbol{I}_{n_i}$, which of course has ***compound symmetric*** correlation structure. It can be shown that

$$\boldsymbol{V}_i^{-1} = \sigma^{-2}\left(\boldsymbol{I}_{n_i} - \frac{D}{\sigma^2 + n_iD}\boldsymbol{J}_{n_i}\right)$$

(verify). Then, defining $\overline{Y}_i = n_i^{-1}\sum_{j=1}^{n_i} Y_{ij}$ to be the simple average of the elements of $\boldsymbol{Y}_i$ and noting that $\widehat{\beta}$ is a weighted average of the $\overline{Y}_i$ (verify), it is straightforward to show that the BLUP (6.63) is

$$\widehat{b}_i = \frac{n_iD}{\sigma^2 + n_iD}(\overline{Y}_i - \widehat{\beta}). \tag{6.65}$$

Several insights follow from (6.65):

- First, note that we can write (6.65) as

  $$\widehat{b}_i = w_i(\overline{Y}_i - \widehat{\beta}) + (1 - w_i)0, \quad w_i = \frac{n_iD}{\sigma^2 + n_iD} < 1,$$

  so that $\widehat{b}_i$ can be interpreted as a ***weighted average*** of the estimated overall deviation $(\overline{Y}_i - \widehat{\beta})$, which is our ***best guess*** for where i "sits" in the population relative to the overall mean β based solely on the data, and 0, the mean of b_i.

  The "weight" $w_i < 0$ thus moves $\widehat{b}_i$ away from being solely based on the data and toward the mean of b_i (0).

The more data we have on i, reflected by larger n_i, the closer w_i is to 1, and the ***more weight*** is put on $(\overline{Y}_i - \widehat{\beta})$ as being a reflection of where i "sits." Likewise, if ***among-individual variation*** is large ***relative*** to ***within-individual variation***, so that D/σ^2 is large, again, $\widehat{b}_i$ puts ***more weight*** on the data from i in predicting where i sits. If, on the other hand, n_i is small and/or among-individual variation is small relative to within-individual variation, the information in the data about where i "sits" is ***not of high quality***, so $\widehat{b}_i$ puts more weight toward 0.

- In (6.64), i's individual-specific mean at any time point is $\beta + b_i$. If we were to ***predict*** this individual-specific mean from (6.64), we would naturally use $\widehat{\beta} + \widehat{b}_i$, which, from (6.65), can be written as (verify)

$$\widehat{\beta} + \widehat{b}_i = w_i \overline{Y}_i + (1 - w_i)\widehat{\beta} = \frac{n_i D}{\sigma^2 + n_i D}\overline{Y}_i + \frac{\sigma^2}{\sigma^2 + n_i D}\widehat{\beta}. \tag{6.66}$$

  In (6.66), if w_i is close to 1, then the prediction is based mainly on the data from i, $\overline{Y}_i$. This will be the case if n_i is large and/or D is large relative to σ^2, in which case the quality of information from i is high and/or there is ***little to be learned about a specific individual*** from the population. If w_i is close to 0, then the prediction is based mainly on the estimated overall population mean $\widehat{\beta}$. This will be the case if n_i is small and/or if among-individual variation, as reflected by D, is small relative to within-individual variation, reflected by σ^2, in which case the poor quality of information on i and the fact that individuals in the population do not vary much suggest that there is ***little to be learned*** about i from the data.

- The foregoing phenomena are usually referred to a ***shrinkage*** in the sense that, in predicting where an individual "sits" in the population and thus his/her individual-specific trajectory, the information from the data is "***shrunk***" toward the overall population mean.

These observations of course extend to the ***general form*** of the linear mixed model. In particular, the obvious predictor of the individual-specific trajectory $\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i$ is

$$\begin{aligned}
\boldsymbol{X}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\widehat{\boldsymbol{b}}_i &= \boldsymbol{X}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}) \\
&= (\boldsymbol{I}_{n_i} - \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1})\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{Y}_i \\
&= \boldsymbol{R}_i\boldsymbol{V}_i^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}} + (\boldsymbol{I}_{n_i} - \boldsymbol{R}_i\boldsymbol{V}_i^{-1})\boldsymbol{Y}_i.
\end{aligned} \tag{6.67}$$

Analogous to (6.66), (6.67) can be interpreted as a ***weighted average*** of the estimated overall population mean profile $\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}$ and the data $\boldsymbol{Y}_i$ on i. If $\boldsymbol{R}_i$, which reflects ***within-individual variation***, is ***large*** relative to among-individual variation, (6.67) puts more weight on the ***population mean profile***; the opposite will be true if ***among-individual variation*** is relatively large.

Similarly, viewing the model as a ***two-stage hierarchy*** , with ***stage 2 population model*** $\beta_i = \boldsymbol{A}_i\beta + \boldsymbol{B}_i\boldsymbol{b}_i$ as in (6.43), by similar reasoning, if we form "estimates" of the individual-specific parameters β_i as

$$\widehat{\beta}_i = \boldsymbol{A}_i\widehat{\beta} + \boldsymbol{B}_i\widehat{\boldsymbol{b}}_i,$$

we would expect analogous "shrinkage" in the sense that the $\widehat{\beta}_i$ will tend to be "shrunk" toward $\boldsymbol{A}_i\widehat{\beta}$.

CAVEATS ON DIAGNOSTICS USING EMPIRICAL BAYES ESTIMATES: It is tempting, and indeed popular, in practice to use the $\widehat{\boldsymbol{b}}_i$ for ***diagnostic purposes***.

- It is common to construct ***histograms and scatterplots*** of the $\widehat{\boldsymbol{b}}_i$ to identify individuals who may be regarded as ***unusual*** relative to the rest of the individuals from the relevant populations from which they arise. For example, such individuals may have individual-specific trajectories that ***evolve differently*** from those for the bulk of the other individuals in the population.
- It is also common to use the $\widehat{\boldsymbol{b}}_i$ to evaluate the relevance of the ***normality assumption*** on the random effects $\boldsymbol{b}_i$ by plotting histograms and scatterplots as well as ***normal quantile plots*** of the components of the $\widehat{\boldsymbol{b}}_i$.

There are several caveats one must bear in mind when inspecting such graphical diagnostics.

- The $\widehat{\boldsymbol{b}}_i$ have ***different distributions*** for each i unless the design matrices $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are ***the same*** for all individuals. Thus, for unbalanced data, graphics based on the raw $\widehat{\boldsymbol{b}}_i$ may be ***uninterpretable***. One approach to addressing this is to ***standardize*** the $\widehat{\boldsymbol{b}}_i$ using (6.58).
- An even more ominous concern that persists even if the $\widehat{\boldsymbol{b}}_i$ all have the same distribution is ***shrinkage***. Histograms and other graphics of the $\widehat{\boldsymbol{b}}_i$ will reflect ***less variability*** than is ***actually present*** in the distribution of the true $\boldsymbol{b}_i$. In particular, the $\boldsymbol{b}_i$ have true covariance matrix $\boldsymbol{D}$, but as in (6.58),

  $$\text{var}(\widehat{\boldsymbol{b}}_i - \boldsymbol{b}_i|\tilde{\boldsymbol{x}}) = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{Z}_i^T \left\{ \boldsymbol{V}_i^{-1} - \boldsymbol{V}_i^{-1}\boldsymbol{X}_i \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right)^{-1} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \right\} \boldsymbol{Z}_i\boldsymbol{D}.$$

  Thus, such graphical displays will ***not necessarily*** reflect the true random effects distribution. In particular, the $\widehat{\boldsymbol{b}}_i$ will tend to be "pulled in" toward the center, so that the usefulness of such plots for, for example, detecting ***departures from normality*** is suspect.

As we have demonstrated, the $\widehat{\boldsymbol{b}}_i$ can be viewed as minimizing the mean square error (6.59), which involves the ***squared-error loss function***, which also follows from normality. Louis (1984) and Shen and Louis (1991) discuss developing alternatives to the usual empirical Bayes estimators that are based on other loss functions.

The bottom line is that, while it is not entirely useless to inspect diagnostics based on $\widehat{\boldsymbol{b}}_i$, these ***potential drawbacks*** need to be kept in mind.

6.5 Implementation via the EM algorithm

With today's computational power, obtaining the ML and REML estimates of the model parameters is ***straightforward*** using standard optimization techniques such as Newton-Raphson and variants to maximize the ML and REML objective functions. However, an alternative computational approach that was popular before the advent of modern computing was to use the ***Expectation-Maximization (EM) algorithm***, as demonstrated by Laird and Ware (1982).

The EM algorithm is a ***computational technique*** to maximize an objective function and can be motivated generically from a missing data perspective in a MAR context, starting from the ***observed data likelihood*** as in (5.107); the details are presented, for example, in Section 3.4 of the instructor's notes for the course "Statistical Methods for Analysis With Missing Data." If the optimization problem can be cast cleverly as a "missing data" or "latent unobserved variable" problem, then the EM algorithm mechanics can be applied to derive an iterative scheme that, under reasonable conditions, should ***converge*** to the values of the model parameters maximizing the objective function and is ***guaranteed*** to increase toward the maximum at each iteration.

We do not attempt to derive the implementation of the EM algorithm for maximizing the ML and REML objective functions for a linear mixed effects models here from these first principles. Rather, we simply ***sketch heuristically*** the rationale for and form of the algorithm in the case of maximizing the ML objective function.

For definiteness, consider the form of the linear mixed model given by

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i, \quad \boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}), \quad \boldsymbol{e}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{n_i}), \quad , i = 1, \ldots, m,$$

so with $\boldsymbol{e}_i$ and $\boldsymbol{b}_i$ independent of $\boldsymbol{x}_i$ and $\boldsymbol{R}_i(\gamma) = \sigma^2\boldsymbol{I}_{n_i}$, the usual default specification.

In this situation, the algorithm follows by analogy to a missing data problem from viewing the ***full data*** as $(\boldsymbol{Y}_i, \boldsymbol{x}_i, \boldsymbol{b}_i)$, $i = 1, \ldots, m$, and the ***observed data*** as $(\boldsymbol{Y}_i, \boldsymbol{x}_i)$, $i = 1, \ldots, m$, so that the $\boldsymbol{b}_i$, $i = 1, \ldots, m$, are "***missing***" for all i. As we have all along, we condition on the $\boldsymbol{x}_i$.

The joint density of $(\boldsymbol{Y}_i, \boldsymbol{b}_i)$ conditional on $\boldsymbol{x}_i$, $i = 1, \ldots, m$, under the above conditions is easily seen to be ***proportional to*** (check)

$$\prod_{i=1}^{m} \sigma^{-1} \exp\{-(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta - \boldsymbol{Z}_i\boldsymbol{b}_i)^T(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta - \boldsymbol{Z}_i\boldsymbol{b}_i)/(2\sigma^2)\}|\boldsymbol{D}|^{-1/2} \exp(-\boldsymbol{b}_i^T\boldsymbol{D}^{-1}\boldsymbol{b}_i/2). \qquad (6.68)$$

If β were known, the unknown parameters in ξ are σ^2 and $\boldsymbol{D}$, and it is straightforward to observe from (6.68) that ***sufficient statistics*** for σ^2 and $\boldsymbol{D}$ are then

$$T_1 = \sum_{i=1}^{m} \boldsymbol{e}_i^T\boldsymbol{e}_i, \qquad \boldsymbol{e}_i = \boldsymbol{Y}_i - \boldsymbol{X}_i\beta - \boldsymbol{Z}_i\boldsymbol{b}_i, \qquad \boldsymbol{T}_2 = \sum_{i=1}^{m} \boldsymbol{b}_i\boldsymbol{b}_i^T. \qquad (6.69)$$

Note that the quantities in (6.69) for β known would be calculable if we had the "full data" available; that is, if we could observe $\boldsymbol{b}_i$ and $\boldsymbol{Y}_i$ and thus $\boldsymbol{e}_i$ for $i = 1, \ldots, m$. In this case, the estimators for $\boldsymbol{D}$ and σ^2 would be

$$\widehat{\sigma}^2 = T_1/N, \qquad \widehat{\boldsymbol{D}} = \boldsymbol{T}_2/m. \qquad (6.70)$$

As can be seen in Section 3.4 of the above-mentioned notes, under these conditions, the EM algorithm is based on repeated evaluation of the ***conditional expectations*** of the "full data" sufficient statistics in (6.69) given the "observed data" $\boldsymbol{Y}_i$, $i = 1, \ldots, m$ (also conditional on $\boldsymbol{x}_i$). Thus, we must derive these conditional expectations.

One way to do this is to write down the (degenerate) joint distribution of $(\boldsymbol{Y}_i^T, \boldsymbol{b}_i^T, \boldsymbol{e}_i^T)^T$, conditional on $\boldsymbol{x}_i$, and then deduce the required quantities by appealing to standard formulæ for the ***conditional moments*** of components of a multivariate normal. This joint distribution is

$$\left(\begin{array}{c} \boldsymbol{Y}_i \\ \boldsymbol{b}_i \\ \boldsymbol{e}_i \end{array}\middle|\, \boldsymbol{x}_i \right) \sim \mathcal{N}\left\{ \begin{pmatrix} \boldsymbol{X}_i\beta \\ \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \sigma^2\boldsymbol{I}_{n_i} & \boldsymbol{Z}_i\boldsymbol{D} & \sigma^2\boldsymbol{I}_{n_i} \\ \boldsymbol{D}\boldsymbol{Z}_i^T & \boldsymbol{D} & \boldsymbol{0} \\ \sigma^2\boldsymbol{I} & \boldsymbol{0} & \sigma^2\boldsymbol{I}_{n_i} \end{pmatrix} \right\}. \qquad (6.71)$$

The marginal joint distributions of $(\boldsymbol{Y}_i, \boldsymbol{b}_i)$ and $(\boldsymbol{Y}_i, \boldsymbol{e}_i)$ given $\boldsymbol{x}_i$ are embedded in (6.71). We have already seen that

$$E(\boldsymbol{b}_i|\boldsymbol{Y}_i, \boldsymbol{x}_i) = \boldsymbol{D}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta),$$

and it follows from standard calculations for conditional moments (verify) that

$$\text{var}(\boldsymbol{b}_i|\boldsymbol{Y}_i, \boldsymbol{x}_i) = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i\boldsymbol{D}.$$

Of course

$$E(\boldsymbol{b}_i\boldsymbol{b}_i^T|\boldsymbol{Y}_i,\boldsymbol{x}_i) = E(\boldsymbol{b}_i|\boldsymbol{Y}_i,\boldsymbol{x}_i)E(\boldsymbol{b}_i|\boldsymbol{Y}_i,\boldsymbol{x}_i)^T + \text{var}(\boldsymbol{b}_i|\boldsymbol{Y}_i,\boldsymbol{x}_i). \tag{6.72}$$

Similarly, it can be verified that

$$E(\boldsymbol{e}_i|\boldsymbol{Y}_i,\boldsymbol{x}_i) = \sigma^2\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta) = \boldsymbol{Y}_i - \boldsymbol{X}_i\beta - \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta)$$

$$\text{var}(\boldsymbol{e}_i|\boldsymbol{Y}_i,\boldsymbol{x}_i) = \sigma^2(\boldsymbol{I}_{n_i} - \sigma^2\boldsymbol{V}_i^{-1}),$$

and, from standard results for quadratic forms,

$$\begin{aligned} E(\boldsymbol{e}_i^T\boldsymbol{e}_i|\boldsymbol{Y}_i,\boldsymbol{x}_i) &= \text{tr}\{E(\boldsymbol{e}_i\boldsymbol{e}_i^T|\boldsymbol{Y}_i,\boldsymbol{x}_i)\} \\ E(\boldsymbol{e}_i\boldsymbol{e}_i^T|\boldsymbol{Y}_i,\boldsymbol{x}_i) &= E(\boldsymbol{e}_i|\boldsymbol{Y}_i,\boldsymbol{x}_i)E(\boldsymbol{e}_i|\boldsymbol{Y}_i,\boldsymbol{x}_i)^T + \text{var}(\boldsymbol{e}_i|\boldsymbol{Y}_i,\boldsymbol{x}_i). \end{aligned} \tag{6.73}$$

Based on (6.72) and (6.73) and some algebra, the algorithm proceeds as follows. Given starting values $\sigma^{2(0)}$ and $\boldsymbol{D}^{(0)}$, at the ℓth iteration, with $\sigma^{2(\ell)}$ and $\boldsymbol{D}^{(\ell)}$ the current iterates and $\boldsymbol{V}_i^{(\ell)} = \sigma^{2(\ell)}\boldsymbol{I}_{n_i} + \boldsymbol{Z}_i\boldsymbol{D}^{(\ell)}\boldsymbol{Z}_i^T$, carry out the following two steps:

1. Calculate
$$\beta^{(\ell)} = \left(\sum_{i=1}^m \boldsymbol{X}_i^T\boldsymbol{V}_i^{(\ell)-1}\boldsymbol{X}_i\right)^{-1}\sum_{i=1}^m \boldsymbol{X}_i^T\boldsymbol{V}_i^{(\ell)-1}\boldsymbol{Y}_i.$$

2. Define
$$\boldsymbol{r}_i^{(\ell)} = \boldsymbol{Y}_i - \boldsymbol{X}_i\beta^{(\ell)}, \qquad \boldsymbol{b}_i^{(\ell)} = \boldsymbol{D}^{(\ell)}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{(\ell)-1}\boldsymbol{r}_i^{(\ell)}, \qquad i = 1,\ldots,m.$$

   Then update $\sigma^{2(\ell)}$ and $\boldsymbol{D}^{(\ell)}$ as

$$\sigma^{2(\ell+1)} = N^{-1}\sum_{i=1}^m\{(\boldsymbol{r}_i^{(\ell)} - \boldsymbol{Z}_i\boldsymbol{b}_i^{(\ell)})^T(\boldsymbol{r}_i^{(\ell)} - \boldsymbol{Z}_i\boldsymbol{b}_i^{(\ell)}) + \sigma^{2(\ell)}\text{tr}(\boldsymbol{I}_{n_i} - \sigma^{2(\ell)}\boldsymbol{V}_i^{(\ell)-1})\},$$

$$\boldsymbol{D}^{(\ell+1)} = m^{-1}\sum_{i=1}^m(\boldsymbol{b}_i^{(\ell)}\boldsymbol{b}_i^{(\ell)\,T} + \boldsymbol{D}^{(\ell)} - \boldsymbol{D}^{(\ell)}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{(\ell)-1}\boldsymbol{Z}_i\boldsymbol{D}^{(\ell)}).$$

Iterate between steps 1 and 2 until convergence. See Laird, Lange, and Stram (1987) for details of implementation; these authors also present an algorithm for maximizing the REML objective function.

As is well known, this algorithm can be ***very slow*** to reach convergence; however, a purported advantage relative to direct maximization is that the value of the objective function is guaranteed to increase at every iteration. Frankly, the implementations of direct optimization in SAS and R have been optimized to the point that it is unusual to encounter computational difficulties; however, in this event, the EM algorithm is an alternative approach.

6.6 Testing variance components

As we discussed at the end of Section 6.3, it is possible to take ***either*** a ***population-averaged*** or a ***subject-specific*** perspective on the linear mixed effects model, which we reiterate briefly.

- Under a ***subject-specific perspective***, we explicitly adopt the ***hierarchical*** interpretation of the model, where individuals are acknowledged to have their own individual-specific trajectories governed by individual-specific parameters β_i. Questions of scientific interest have to do with the properties of the ***distributions*** of β_i. Thus, the fixed effects β represent features relevant to the ***mean*** or "***typical***" value of β_i (possibly for different among-individual covariate values). The covariance matrix $\boldsymbol{D}$ (and generalizations thereof) represents the acknowledged variation of these features in the populations of interest. Accordingly, the diagonal elements of $\boldsymbol{D}$ are interpreted as explicitly reflecting the variances of these features, while the off-diagonal elements reflect how these features co-vary in populations of interest. From this point of view, $\boldsymbol{D}$ is a ***legitimate covariance matrix*** in the sense that, at the very least, it is ***nonnegative definite*** (positive semidefinite).

  Likewise, the matrices $\boldsymbol{R}_i$ are acknowledged to also be ***legitimate covariance matrices*** reflecting within-individual variance and correlation. Thus, for example, σ^2 in the simplest specification $\sigma^2 \boldsymbol{I}_{n_i}$ is the total within-individual variance (assumed constant over time) dictating how responses on an individual vary about his/her individual-specific trajectory due to the realization process and measurement error, and it is natural that we believe that $\sigma^2 \geq 0$.

- Under a ***population-averaged perspective***, questions of scientific interest have to do with features of ***overall mean response profiles***. Here, we view the hierarchical formulation as not necessarily representing phenomena of interest but rather as a convenient mechanism to ***induce*** a rich and flexible ***overall covariance structure*** that can handle ***unbalanced data*** where responses are ascertained at possibly different time points for different individuals and that accommodates possibly ***nonstationary*** patterns of overall correlation. Thus, as we noted at the end of Section 6.3, the matrices $\boldsymbol{D}$ and $\boldsymbol{R}_i$ are simply building blocks of an ***overall*** legitimate covariance structure, and thus ***need not*** be legitimate covariance matrices themselves.

These considerations emphasize that it is ***imperative*** that the analyst acknowledge the modeling perspective taken when it comes to making ***inferences*** about covariance structure or, more precisely, inferences on the covariance parameters $\xi = (\gamma^T, \text{vech}(\boldsymbol{D})^T\}^T$, as we now describe.

EXAMPLE: For definiteness, consider the situation of the hip replacement data in Section 6.2. Suppose that we assume as in (6.35) that

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \beta_{2i} t_{ij}^2 + e_{ij}, \quad \boldsymbol{\beta}_i^T = (\beta_{0i}, \beta_{1i}, \beta_{2i})^T,$$

and then take $\boldsymbol{\beta}_i$ to be as in (6.37), so that we can write as in (6.38)

$$\boldsymbol{\beta}_i = \boldsymbol{A}_i \boldsymbol{\beta} + \boldsymbol{B}_i \boldsymbol{b}_i,$$

where

$$\boldsymbol{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \end{pmatrix}, \quad \boldsymbol{B}_i = \boldsymbol{I}_3. \tag{6.74}$$

If we take the $\boldsymbol{b}_i$ to be independent of $\boldsymbol{a}_i$ with $\text{var}(\boldsymbol{b}_i) = \boldsymbol{D}$, then $\boldsymbol{D}$ is a (3×3) matrix, which involves ***six*** distinct parameters. If we further assume that $\text{var}(\boldsymbol{e}_i) = \sigma^2 \boldsymbol{I}_{n_i}$, which involves an additional parameter, then this of course induces an overall covariance structure of the form

$$\boldsymbol{V}_i = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^T + \sigma^2 \boldsymbol{I}_{n_i},$$

involving seven parameters, so that ξ is (7×1).

- From a ***PA*** perspective, this model is a way to ***induce*** a quadratic PA population mean model and an overall covariance structure depending on ξ. Because under the PA perspective $\boldsymbol{D}$ is ***not required*** to be nonnegative definite and σ^2 is ***not required*** to be ≥ 0, there are no restrictions on ξ.
- From a ***SS*** perspective, this model embodies the belief that each individual in the population has his/her own individual-specific quadratic trajectory and that individual-specific intercepts, linear components and quadratic components ***vary*** and ***co-vary*** in the population according to the ***covariance*** matrix $\boldsymbol{D}$; in addition, individual-specific responses vary about individual-specific trajectories with ***variance*** σ^2. Here, $\boldsymbol{D}$ is ***required*** to be nonnegative definite and σ^2 is ***required*** to be ≥ 0 for this perspective to be reasonable.

Now consider ***eliminating*** b_{2i} from the model as in (6.40) and taking instead

$$\boldsymbol{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix}, \quad \boldsymbol{B}_i = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \tag{6.75}$$

so that $\text{var}(\boldsymbol{b}_i) = \boldsymbol{D}_2$ is now a (2×2) matrix with three distinct parameters and ξ is then (4×1).

- From a ***PA*** perspective, the specification (6.75) is a way to ***induce*** a ***more parsimonious*** overall covariance structure with ***fewer parameters***.

- From a ***SS*** perspective, (6.75) embodies the assumption that, while individual-specific intercepts and linear components ***vary nonegligibly*** in the population of individuals, individual-specific ***quadratic*** components either do not vary at all or, relative to the variation in intercepts and linear components, exhibit ***negligible variation*** among individuals.

  Thus, as we discussed in Section 6.2, it is popular to view this as asking whether the individual-specific quadratic components are "***fixed***" or "***random***."

Thus, from either perspective, it is of interest to evaluate whether or not (6.75) is adequate to represent the true state of affairs or if (6.74) is required.

- From a ***PA*** perspective, this corresponds to asking whether or not a ***simpler representation*** of the ***overall covariance structure*** based on fewer parameters is adequate or if the richer induced structure involving more parameters is required.

- From a ***SS*** perspective, this corresponds to what we believe about the ***relative magnitude of variation*** in individual-specific quadratic components.

To address this ***formally*** from either perspective, we might want to carry out a ***hypothesis test*** of whether or not (6.75) is sufficient to represent the situation relative to (6.74).

It is straightforward to show that (6.75) can be ***equivalently represented*** by taking the (3×3) matrix $\boldsymbol{D}$ corresponding to $\text{var}(\boldsymbol{b}_i)$ in (6.74) to be of the form

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_2 & 0 \\ 0 & 0 \end{pmatrix}, \tag{6.76}$$

say. (The diligent student will want to verify this.)

Thus, we can address this issue by testing the null hypothesis that in fact

$$H_0 : \boldsymbol{D} = \begin{pmatrix} D_{11} & D_{12} & D_{13} \\ D_{12} & D_{22} & D_{23} \\ D_{13} & D_{23} & D_{33} \end{pmatrix} = \begin{pmatrix} D_{11} & D_{12} & 0 \\ D_{12} & D_{22} & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \boldsymbol{D}_2 & 0 \\ 0 & 0 \end{pmatrix} \tag{6.77}$$

against an appropriate alternative.

As the model (6.75) is ***nested*** within the model (6.74), it is natural to consider using a ***likelihood ratio test*** for this purpose, constructed from the loglikelihoods fitting the "***full***" model under specification (6.74) and the "***reduced***" model under specification (6.75).

VALIDITY OF TEST PROCEDURES: The key issue is whether or not this likelihood ratio test is a ***valid test*** of H_0. One of the ***regularity conditions*** required for usual ***large sample theory approximations*** to hold is that the true value of a parameter is not on the ***boundary of its parameter space*** but rather lies in its ***interior***. In the context of ***hypothesis testing*** , the value of the parameter ***under the null hypothesis*** cannot be on the boundary of the parameter space but must be in the ***interior of the parameter space*** for usually asymptotic arguments leading to tests to be valid.

In particular, the ***normal approximation*** to the sampling distribution of an estimator, which is used to form Wald and F-type tests, and the ***chi-square approximation*** to the sampling distribution of the likelihood ratio test statistic ***rely critically*** on this condition.

- If we regard (3×3) matrix $\boldsymbol{D}$ in (6.77) as a symmetric matrix, whose parameters simply serve to characterize an overall covariance structure, as we do from a ***PA*** perspective, then there is ***no restriction*** on the values taken on by D_{33} (or any of the parameters, for that matter). Under this perspective, the value of D_{33} in (6.77) under H_0 (0) is in the ***interior*** of the parameter space.
- If we regard the (3×3) matrix $\boldsymbol{D}$ in (6.77) as a ***legitimate covariance matrix*** , as we do from a ***SS*** perspective, then D_{33} is a ***variance*** and, for $\boldsymbol{D}$ to nonnegative definite, it must be that $D_{33} \geq 0$. Under this perspective, the value of D_{33} under H_0 is thus on the ***boundary*** of the parameter space.

We are thus led to the following.

POPULATION-AVERAGED PERSPECTIVE: In the example, comparing the usual ***likelihood ratio test statistic*** described above to the appropriate ***chi-square critical value*** will yield a ***valid test*** of H_0 in (6.77), whose interpretation is as above.

In general, if a PA perspective is taken, the matrices $\boldsymbol{D}$ and $\boldsymbol{R}_i(\gamma)$ are ***not required*** to be nonnegative definite, so that there are no restrictions on ξ. Thus, the values of ξ under a null hypothesis representing simpler structure will not lie on the boundary of the parameter space, and testing whether or not there is evidence a more complex ***induced overall covariance structure*** is preferred over a simpler one can be conducted in the usual way.

SUBJECT-SPECIFIC PERSPECTIVE: In the example, as noted above, H_0 places at least one parameter on the boundary of the parameter space. Thus, carrying out the likelihood ratio test in the usual way ***will not*** lead to a valid test.

To achieve a valid test, one must appeal to specialized theoretical results for ***nonstandard testing situations*** in a classic paper by Self and Liang (1987). Stram and Lee (1994) used this theory to demonstrate that, when $\boldsymbol{R}_i = \sigma^2 \boldsymbol{I}_{n_i}$, the large sample distribution of the likelihood ratio test statistic is, under reasonable conditions, a ***mixture of chi-squared distributions***.

For $\boldsymbol{D}$ $(q+1 \times q+1)$, for testing a general null hypothesis of the form

$$H_0 : \boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_q & \boldsymbol{0} \\ \boldsymbol{0} & 0 \end{pmatrix},$$

where $\boldsymbol{D}_q$ is a $(q \times q)$ postitive definite matrix versus the alternative that $\boldsymbol{D}$ is a general $(q+1 \times q+1)$ nonnegative definite matrix, the large sample distribution of the likelihood ratio test statistic under H_0 is a mixture of a χ^2_{q+1} distribution and a χ^2_q distribution with equal weights of 0.5.

- Our example is the special case of $q = 2$.
- The effect is to ***reduce*** the p-value that results relative to that that would be obtained if one (incorrectly) used the likelihood ratio testing procedure in the usual way. Thus, ignoring the "***boundary problem***" will lead in general to rejection of H_0 of less often and to possibly adopting models that are too parsimonious.

RESULT: We do not discuss this further here; details can be found in Stram and Lee (1994) and Section 6.3 of Verbeke and Molenberghs (2000); see also Verbeke and Molenberghs (2003).

The takeaway message is that ***faithfully acknowledging*** the perspective to be taken (PA vs. SS) on the scientific questions is critical to achieving reliable inferences. The data analyst must probe his or her scientific collaborators to ensure that the appropriate perspective is taken.

STANDARD ERRORS FOR COVARIANCE PARAMETERS: Testing as discussed above is usually carried out to refine the model with the goal of improving inferences on the overall population mean structure from a PA perspective or to assist interpretation from a SS perspective. From a PA perspective, reducing the number of covariance parameters and thereby achieving a ***more parsimonious*** representation of overall covariance structure will hopefully lead to inferences on β and the population mean that are ***more efficient in finite samples***. Here, the covariance parameters are ordinarily ***not*** of scientific interest in their own right.

From a SS perspective, this testing provides insight on the ***relative magnitudes of variation*** of features of ***individual inherent trajectories*** (e.g., individual-specific intercepts and slopes) in the population of individuals. In this case, the diagonal elements of the matrix $\boldsymbol{D}$ represent the magnitudes of variation of these features, and the off-diagonal elements represent how these co-vary in the population of individuals. Thus, ***scientific questions*** may involve characterizing these magnitudes of variation and thus may be stated formally in terms of the diagonal elements of $\boldsymbol{D}$.

When a model specification is adopted such that the diagonal elements of $\boldsymbol{D}$ are assumed to be non-zero, ***estimates*** of these elements characterize the variation in the features to which they correspond on the individual trajectory model (that is, the β_i). Thus, it is of interest to report these estimates accompanied by ***appropriate standard errors***.

In principle , calculation of standard errors for these elements and more generally for all components of the covariance parameter ξ can be based on a large sample approximation to the sampling distribution of the estimator $\widehat{\xi}$. Such an approximation can be derived by an estimating equation argument similar to that for $\widehat{\beta}$ if one is willing to assume that the n_i are ***fixed*** (so no missing data as discussed in Section 5.6). A key issue is that the covariance matrix of the asymptotic distribution of the estimator $\widehat{\xi}$ depends on the ***third*** and ***fourth moments*** of the true distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$. If one is willing to assume ***normality*** of the response, this covariance matrix can be derived from the information matrix in (5.109) and depends on the ***fourth moment*** of a normal distribution. If the true distribution of the response is ***not normal*** , then the approximate sampling distribution for ξ so obtained and thus approximate standard errors derived from it can be ***very unreliable***.

We do not present details here. This discussion underscores the general issue that inference on second moment properties is more problematic than inference on first moment properties.

7 Generalized and Nonlinear Models for Univariate Response

7.1 Introduction

The models for longitudinal data we have discussed so far are suitable for responses that are or can be viewed as approximately ***continuous***. Moreover, the models incorporate the assumption that the overall population mean (PA perspective) and inherent individual trajectories (SS perspective) can be approximated by representations that are ***linear*** in parameters.

Such models are clearly ***unsuitable*** for ***discrete responses***, such as binary or categorical outcomes or responses whose values are small counts, for which standard models are not linear. They are also not appropriate for continuous outcomes when population or individual trajectories ***cannot*** be well-approximated by linear functions of parameters.

For instance, in ***EXAMPLE 4*** of Chapter 1 on the pharmacokinetics of theophylline, the ***mechanistic model*** for (continuous) drug concentration at time t within an ***individual subject*** in (1.3) and (2.1), derived from the one-compartment representation of the body in Figure 1.6, is a natural way to represent the ***inherent individual trajectory*** of drug concentrations over time. As we review shortly, this model is ***nonlinear*** in individual-specific parameters k_a, Cl, and V reflecting absorption rate; drug clearance, which has to do with how the drug is eliminated from the body; and volume of distribution, which is related to the extent to which the drug distributes through the body, respectively. These individual-specific parameters thus have ***meaningful scientific interpretations***, so an appropriate analysis should incorporate the mechanistic model.

Likewise, in ***EXAMPLE 6*** of Chapter 1, the Six Cities Study, the wheezing response is ***binary***. Thus, if $Y_{ij} = 0$ if the ith child is not wheezing at time (age) j and 1 if s/he is, the "typical" or ***population mean*** response at age j given covariates is $\text{pr}(Y_{ij} = 1|\boldsymbol{x}_i)$. Popular regression models for probabilities, such as ***logistic*** or ***probit*** regression models, are ***nonlinear*** in parameters, as we demonstrate in the next section.

Clearly, ***population-averaged*** and ***subject-specific*** models for longitudinal data in these situations are required. In this chapter, as a prelude to discussing these longitudinal models and associated inferential methods, we review classical ***nonlinear regression models*** for ***univariate response***.

7.2 Nonlinear mean-variance models

GENERAL NONLINEAR MODEL: We consider the following situation and notation. Let Y denote a scalar response of interest and $\boldsymbol{x}$ denote a vector of covariates, and suppose we observe $(Y_j, \boldsymbol{x}_j)$, $j = 1, \ldots, n$, ***independent*** across j. Here, we use j as the index in anticipation of our discussion of SS nonlinear models; see below. In this chapter, we focus on models of the general form

$$E(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \beta), \qquad \text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2 g^2(\beta, \delta, \boldsymbol{x}_j), \quad j = 1, \ldots, n. \tag{7.1}$$

where $\theta = (\sigma^2, \delta^T)^T$ is $(r \times 1)$, and β is $(p \times 1)$.

- In (7.1), $f(\boldsymbol{x}, \beta)$ is a ***nonlinear*** function of parameters β depending on the covariates $\boldsymbol{x}_j$.
- $g^2(\beta, \delta, \boldsymbol{x}_j)$ is the ***variance function***, which allows variance to be ***nonconstant*** over j in a systematic fashion depending on $\boldsymbol{x}_j$ and which is also possibly ***nonlinear*** in β and possibly additional ***variance parameters*** δ. Here, σ^2 is a ***scale parameter***.

EXAMPLES: The model (7.1) is used to represent a variety of situations, depending on the context.

- As noted above, when Y_j is ***binary*** taking values 0 or 1, $E(Y_j|\boldsymbol{x}_j) = \text{pr}(Y_j = 1|\boldsymbol{x}_j)$, and a natural model is the classical ***logistic regression model***

  $$f(\boldsymbol{x}_j, \beta) = \frac{\exp(\boldsymbol{x}_j^T\beta)}{1 + \exp(\boldsymbol{x}_j^T\beta)}, \quad \text{or equivalently} \quad \text{logit}\{f(\boldsymbol{x}_j, \beta)\} = \log\left\{\frac{f(\boldsymbol{x}_j, \beta)}{1 - f(\boldsymbol{x}_j, \beta)}\right\} = \boldsymbol{x}_j^T\beta, \tag{7.2}$$

  where $\text{logit}(u) = \log\{u/(1-u)\}$. Here, then, $f(\boldsymbol{x}_j, \beta)$ represents a ***probability***.

  For binary response with mean $f(\boldsymbol{x}_j, \beta)$, it is immediate that we ***must have***

  $$\text{var}(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \beta)\{1 - f(\boldsymbol{x}_j, \beta)\}, \tag{7.3}$$

  so that $\sigma^2 \equiv 1$, and there is no unknown parameter δ. Implicit is the assumption that the binary response can be ascertained ***perfectly***, with no potential ***misclassification error***, which is analogous to measurement error in the case of binary response.

  This situation might arise in a study where the jth of n participants has baseline covariates $\boldsymbol{x}_j$, and the single binary response Y_j is ascertained on each individual j at some follow-up time. Here, $\boldsymbol{x}_j$ is an ***among-individual*** covariate, and interest focuses on the probability of positive response in the population as a function of these covariates.

Thus, the ***scope of inference*** is the ***entire population*** from which the sample of n individuals was drawn, and the parameter β has a ***PA interpretation***. The extension to a ***longitudinal study*** is if the response were ascertained ***repeatedly*** over time on each individual j.

Alternatively, the n binary responses might all be on the ***same individual*** after s/he was given different doses x_j of a drug on occasions $j = 1, \ldots, n$, where these responses are assumed to be ascertained ***sufficiently far apart*** in time to be ***approximately independent***. In this case, interest focuses on the dose-response relationship for this individual, so that the ***scope of inference*** is this ***single individual***, and x_j is a ***within-individual*** covariate. In this case, the parameter β characterizes the probability of positive response for ***this individual only*** as a function of dose.

- As discussed in Section 2.2, a model of the form (7.1) is often used to describe ***individual pharmacokinetics***. For example, from (2.1), if our focus is on a ***given individual*** who received dose D of theophylline at time 0, and Y_j is drug concentration measured on this individual at time t_j, then $\boldsymbol{x}_j = (D, t_j)$, $j = 1, \ldots, n$, and

$$f(\boldsymbol{x}_j, \boldsymbol{\beta}) = \frac{\beta_1}{\beta_3(\beta_1 - \beta_2/\beta_3)}\{\exp(-\beta_2 t_j/\beta_3) - \exp(-\beta_1 t_j)\}, \quad \boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T. \tag{7.4}$$

In (7.4), $\boldsymbol{x}_j$ has the interpretation as what we have referred to as a ***within-individual covariate*** (appended by time); we have used the notation $\boldsymbol{z}_{ij}$ for the jth such covariate on individual i in a longitudinal data context.

As noted previously, it is often further assumed that the sampling times t_j are ***sufficiently intermittent*** that ***serial correlation*** among the Y_j is ***negligible***, so that the assumption of ***independence*** of the $(Y_j, \boldsymbol{x}_j)$ over j is taken to hold ***approximately***.

Here, $\text{var}(Y_j|\boldsymbol{x}_j)$ in (7.4) reflects the ***aggregate*** variation due to the ***within-individual realization process*** and ***measurement error*** in ascertaining drug concentrations. As noted in Section 2.2, in pharmacokinetics this aggregate variance typically exhibits ***constant coefficient of variation***, so a popular ***empirical model*** for aggregate within-individual variance in practice is

$$\text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2 f^2(\boldsymbol{x}_j, \beta), \tag{7.5}$$

which is of the form in (7.1) with $g^2(\beta, \delta, \boldsymbol{x}_j) = f^2(\boldsymbol{x}_j, \beta)$, so that σ is the coefficient of variation (CV). In (7.5), there is no unknown variance parameter δ.

A common generalization of (7.5) is the so-called "***power of the mean***" variance model

$$\text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2 f^{2\delta}(\boldsymbol{x}_j, \beta), \quad \delta > 0, \tag{7.6}$$

so $g^{2\delta}(\beta, \delta, \boldsymbol{x}_j) = f^{2\delta}(\boldsymbol{x}_j, \beta)$, which represents aggregate variance as proportional to an ***arbitrary power*** δ of the mean response. This is a popular model when the combined effects of effect of realization and measurement error appear to yield ***more profound*** pattern of variance than dictated by the constant CV model.

From the point of view of the ***conceptual representation*** in Chapter 2 , models like (7.5) and (7.6) are indeed ***approximations*** to a potentially more complex mechanism. To see this, write the *j*th drug concentration as

$$Y_j = f(\boldsymbol{x}_j, \beta) + e_{Pj} + e_{Mj}, \tag{7.7}$$

where as before e_{Pj} represents the within-individual deviation due to the ***realization process*** and e_{Mj} represents the ***measurement error*** deviation at time t_j, with $E(e_{Pj}|\boldsymbol{x}_j) = 0$ and $E(e_{Mj}|\boldsymbol{x}_j) = 0$. Then (7.7) of course implies $E(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \beta)$ as in (7.1) and allows us to contemplate the ***contributions*** of each to the aggregate within-individual variance $\text{var}(Y_j|\boldsymbol{x}_j)$ as follows.

Many biological processes exhibit approximate ***constant CV*** or other dependence of the variance of the process on the ***level of mean response***. Here, this implies that an appropriate model for the ***variance of the realization process deviation*** is

$$\text{var}(e_{Pj}|\boldsymbol{x}_j) = \sigma_P^2 f(\boldsymbol{x}_j, \beta)^{2\delta_P}, \tag{7.8}$$

say, where δ_P might indeed be equal to 1.

As we have discussed, some measuring techniques commit errors such that the magnitude of the error is ***related*** to the size of the thing being measured. This is sometimes the case for ***assays*** used to ascertain levels of drug or other agents in blood, plasma, or other samples. In (7.7), the thing being measured at time t_j is the ***actual realized drug concentration***

$$f(\boldsymbol{x}_j, \beta) + e_{Pj}.$$

Thus, ideally, this suggests that e_{Mj} and e_{Pj} are ***correlated***, so an overall model for $\text{var}(Y_j|\boldsymbol{x}_j)$ should reflect this. However, it is well accepted in pharmacokinetics that the aggregate variance of drug concentrations is ***dominated by measurement error*** in that the deviations from the inherent drug concentration trajectory $f(\boldsymbol{x}_j, \beta)$ are "negligible" compared to those for measurement error.

From this point of view, at the level of the individual, for whom β is ***fixed***, it is common to view e_{Pj} and e_{Mj} as approximately independent and to approximate $\text{var}(e_{Mj}|\boldsymbol{x}_j)$ as depending on $f(\boldsymbol{x}_j, \beta)$, in which case a model for ***measurement error variance*** might be of the form

$$\text{var}(e_{Mj}|bx_j) = \sigma_M^2 f(\boldsymbol{x}_j, \beta)^{2\delta_M}, \tag{7.9}$$

Following these considerations and combining (7.8) and (7.9), we are led to the representation

$$\text{var}(Y_j|\boldsymbol{x}_j) = \sigma_P^2 f(\boldsymbol{x}_j, \beta)^{2\delta_P} + \sigma_M^2 f(\boldsymbol{x}_j, \beta)^{2\delta_M} \tag{7.10}$$

A further approximation reflecting the belief that measurement error dominates the realization process would be to ***disregard*** e_{Pi} and thus the first term in (7.10) entirely, in which case the common models (7.5) and (7.6) can be viewed as ***representing primarily*** measurement error variance. Alternatively, these models can be viewed as a "***compromise***" approximation to (7.10).

If it is in fact believed that measurement errors are of ***similar magnitude*** regardless of the size of the thing being measured, so that e_{Mj} and e_{Pj} are reasonably taken as ***independent***, an aggregate variance model representing this is a simplification of (7.10), usually parameterized as

$$\text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2\{\delta_1 + f^{2\delta_2}(\boldsymbol{x}_j, \beta)\}, \quad \boldsymbol{\delta} = (\delta_1, \delta_2)^T, \tag{7.11}$$

so that $\sigma_P^2 = \sigma^2$ and $\sigma_M^2 = \sigma^2\delta_1$.

In this example, the ***scope of inference*** is confined to the ***single individual*** on whom the drug concentrations over time were ascertained. Here, then, β pertains to this individual only. The ***same*** modeling considerations would of course apply to ***each individual*** in a sample of m individuals on whom concentration-time data are available, as in the SS longitudinal data model framework we discuss in Chapter 9.

- Although in (7.1) we allow the dependence of the variance function on β and $\boldsymbol{x}_j$ to be arbitrary, as in the foregoing examples, it is almost always the case that if it is taken to depend on ***both*** β and $\boldsymbol{x}_j$, this dependence is solely through the ***mean response*** $f(\boldsymbol{x}_j, \beta)$.

- Note that in (7.1) it could be that the variance function depends only on covariates $\boldsymbol{x}_j$ and variance parameters δ and ***not on*** β or the mean response. For example,

$$\text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2 \exp(\boldsymbol{x}_j^T \boldsymbol{\delta})$$

is a popular ***empirical model*** that allows variance to change directly with the values of covariates. Such models are widely used in econometrics.

In the most general case of model (7.1), we make no further assumptions on the distribution of Y_j given $\boldsymbol{x}_j$ beyond the first two moments. For binary response Y_j, of course, given a model $f(\boldsymbol{x}_j, \beta)$ for $E(Y_j|\boldsymbol{x}_j)$, the entire (Bernoulli) distribution of Y_j given $\boldsymbol{x}_j$ is ***fully specified***. Likewise, if we take the distribution of $Y_j|\boldsymbol{x}_j$ to be ***normal***, then given a model (7.1) the distribution is fully specified.

SCALED EXPONENTIAL FAMILY: A special case of the general model (7.1) is obtained by making the assumption that the distribution of Y_j given $\boldsymbol{x}_j$ is a member of a particular class of distributions that includes the Bernoulli/binomial and the normal with ***constant variance*** for all j. A random variable *Ylt* is said to have distribution belonging to the ***scaled exponential family*** if it has density or probability mass function

$$p(y\,;\zeta,\sigma) = \exp\left\{\frac{y\zeta - b(\zeta)}{\sigma^2} + c(y,\sigma)\right\}, \tag{7.12}$$

where ζ and σ are real-valued parameters characterizing the density, and $b(\zeta)$ and $c(y,\sigma)$ are real-valued functions.

- If σ is ***known*** (often $\sigma = 1$ in this case), then (7.12) is exactly the density of a ***one-parameter exponential family*** with ***canonical parameter*** ζ.
- It is straightforward to derive (try it) that

  $$E(Y) = b_\zeta(\zeta) = d/d\zeta\, b(\zeta), \qquad \text{var}(Y) = \sigma^2 b_{\zeta\zeta}(\zeta) = \sigma^2 d^2/d\zeta^2\, b(\zeta),$$

  so that if $E(Y) = \mu$ and $b_\zeta(\,\cdot\,)$ is a one-to-one function, ζ can be regarded as a function of μ, namely, $\zeta = b_\zeta^{-1}(\mu)$,, and thus $\text{var}(Y) = \sigma^2 b_{\zeta\zeta}\{(b_\zeta^{-1}(\mu)\} = \sigma^2 g^2(\mu)$. This demonstrates that the density (7.12) induces a ***specific relationship between mean and variance***.
- Common distributions that are members of the class (7.12) are as follows:

Distribution	$b(\zeta)$	$\zeta(\mu)$	$g^2(\mu)$
Normal, constant variance	$\zeta^2/2$	μ	1
Poisson	$\exp(\zeta)$	$\log\mu$	μ
Gamma	$-\log(-\zeta)$	$-1/\mu$	μ^2
Inverse Gaussian	$-(-2\zeta)^{1/2}$	$1/\mu^2$	μ^3
Binomial	$\log(1+e^\zeta)$	$\log\{\mu/(1-\mu)\}$	$\mu(1-\mu)$

For the Poisson and binomial distributions, $\sigma = 1$. For the others, σ is a free parameter characterizing the density.

GENERALIZED (NON)LINEAR MODEL: If the distribution of $Y_j|\boldsymbol{x}_j$ has density (7.12) with $b_\zeta(\zeta_j) = f(\boldsymbol{x}_j, \beta)$, then it follows that

$$E(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \beta), \quad \text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2 g^2\{f(\boldsymbol{x}_j, \beta)\}, \tag{7.13}$$

for function $g^2(\cdot,)$ dictated by $b(\cdot)$, and (7.13) with this density is referred to as a ***generalized (non)linear model***.

- In (7.13), we emphasize that the implied variance function is a ***known function of the mean***.
- Model (7.13) is a slight extension of the ***generalized linear model*** , for which $\boldsymbol{x}_j$ and β enter the mean model ***only*** through the ***linear combination*** $\boldsymbol{x}_j^T\beta$, in which case we write $f(\boldsymbol{x}_j^T\beta)$.
- For a generalized linear model with $E(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j^T\beta)$ and $f(\cdot)$ ***monotone*** in its single argument, its ***inverse*** $f^{-1}(\cdot)$ is called the ***link function*** , and $\boldsymbol{x}_j^T\beta$ is called the ***linear predictor***. If furthermore the link function satisfies $f^{-1}(\mu) = \zeta$ for ζ as in (7.12), then it is called the ***canonical link***. There is ***no special significance*** to the canonical link as far as data analysis is concerned; e.g., there is ***no reason*** it should provide a better fitting model than some other f.
- The usual ***logistic regression model*** in (7.2) and (7.3) is a special case of a generalized linear model, arising from the simplest ***binomial distribution***, the Bernoulli. This model uses the ***canonical link*** ; the classical ***probit*** model, which instead takes

  $$f(\boldsymbol{x}_j, \beta) = \Phi(\boldsymbol{x}_j^T\beta),$$

  where $\Phi(\cdot)$ is the cdf of the standard normal distribution, is also a generalized linear model that ***does not*** use the canonical link.
- For responses in the form of (nonnegative integer) ***counts*** , as in ***EXAMPLE 5*** of the epileptic seizure study in Chapter 1, the ***Poisson distribution*** is a standard model, and the classical model for $E(Y_j|\boldsymbol{x}_j)$ is the ***loglinear model***

  $$f(\boldsymbol{x}_j, \beta) = \exp(\boldsymbol{x}_j^T\beta),$$

  with $\text{var}(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \beta)$. This is also a generalized linear model with canonical link.
- The classical ***linear regression model*** $f(\boldsymbol{x}_j, \beta) = f(\boldsymbol{x}_j^T\beta) = \boldsymbol{x}_j^T\beta$ where $Y_j|\boldsymbol{x}_j$ is assumed ***normal*** with with ***constant variance*** is also a special case of a generalized linear model, where $f(\cdot)$ is the so-called ***identity link***.

- Despite widespread usage, there is ***no reason*** that dependence on the covariates must be through the ***linear combination*** $\boldsymbol{x}_j^T\beta$ the case except convention. For example, in dose-toxicity modeling, where the response Y_j is binary and x_j is dose given to the jth laboratory rat, modifying the usual logistic model to be

$$E(Y_j|x_j) = \frac{\exp(\beta_0 + \beta_1 x_j^{\beta_2})}{1 + \exp(\beta_0 + \beta_1 x_j^{\beta_2})}$$

  often provides a ***better fit***.

- Model (7.13) can be ***extended*** without altering the foregoing results. For example, if Y_j is the number of "successes" observed in a fixed number r_j trials with success probability $\pi(\boldsymbol{c}_j, \beta)$, say, then letting $\boldsymbol{x}_j = (r_j, \boldsymbol{c}_j^T)^T$,

$$E(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \beta) = r_j\pi(\boldsymbol{c}_j, \beta), \quad \text{var}(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \beta)\{r_j - f(\boldsymbol{x}_j, \beta)\}/r_j = g^2\{f(\boldsymbol{x}_j, \beta), \boldsymbol{x}_j\}.$$

  We suppress this additional dependence of the variance function on $\boldsymbol{x}_j$ in generalized (non)linear models henceforth and continue to write the variance function as in (7.13), but all developments apply to this more general formulation.

- For distributions like the Poisson for counts or binomial for numbers of "successes," the ***scale parameter*** $\sigma^2 = 1$. However, in some circumstances the mean-variance relationship in (7.13) with $\sigma^2 = 1$ may be ***insufficient*** to represent the ***true magnitude*** of the aggregate variation in the data. ***Overdispersion*** refers to the phenomenon in which the variance of the response exceeds the nominal variance dictated by the distributional model. This can be because of ***measurement error*** or due to ***clustering***.

  For example, if r rats are placed in each of n cages, the rats in cage j are given a dose x_j of a toxic agent, and Y_j is the number of rats in cage j having an adverse reaction, then Y_j is the sum of r binary responses, one for each rat. If all rats have the ***same probability*** π_j of having an adverse reaction to the dose x_j, then Y_j is binomial with parameters r and π_j. However, if rats are heterogeneous, so that the kth rat in the cage j has probability p_{jk} of having an advese reaction, where the p_{jk} are such that $E(p_{jk}|x_j) = \pi_j$ and $\text{var}(p_{jk}|x_j) = \tau^2\pi_j(1-\pi_j)$, it can be shown (try it) that $Y_j|x_j$ is such that

$$E(Y_j|x_j) = r\pi_j, \qquad \text{var}(Y_j|x_j) = \sigma^2 r\pi_j(1-\pi_j), \tag{7.14}$$

  where σ^2 is a function of τ^2 and r.

The mean-variance model in (7.14) resembles that of the usual binomial ***except*** for the ***scale factor*** σ^2. Because there is ***additional among-rat variation*** in that all rats do not have the same probability of an adverse reaction, we might expect $\sigma^2 > 1$, which would make the variability ***more profound*** than that dictated by the binomial.

It is thus commonplace to allow a scale factor σ^2 in (7.13) to accommodate potential such ***overdispersion***.

As we discuss next, it turns out that maximum likelihood estimation of $\boldsymbol{\beta}$ in a generalized (non)linear model (7.13) under density (7.12) is equivalent to solving the same ***linear estimating equation*** that one is led to more generally from a variety of viewpoints.

7.3 Estimation of mean and variance parameters

We assume henceforth that the model for the mean $E(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \boldsymbol{\beta})$ in (7.1) is ***correctly specified***.

MAXIMUM LIKELIHOOD FOR THE SCALED EXPONENTIAL FAMILY: Taking the derivative of the logarithm of (7.12) with respect to $\boldsymbol{\beta}$ with ζ represented as a function of the mean (and thus of $\boldsymbol{\beta}$), using the chain rule, it is straightforward to show (verify) that the ***maximum likelihood estimator*** for $\boldsymbol{\beta}$ is the solution to the ***estimating equation***

$$\sum_{j=1}^{n} f_{\beta}(\boldsymbol{x}_j, \boldsymbol{\beta}) g^{-2}\{f(\boldsymbol{x}_j, \boldsymbol{\beta})\}\{Y_j - f(\boldsymbol{x}_j, \boldsymbol{\beta})\} = \mathbf{0}, \tag{7.15}$$

where $f_{\beta}(\boldsymbol{x}_j, \boldsymbol{\beta}) = \partial/\partial\boldsymbol{\beta} f(\boldsymbol{x}_j, \boldsymbol{\beta})$ is the $(p \times 1)$ vector of partial derivatives of $f(\boldsymbol{x}_j, \boldsymbol{\beta})$ with respect to the elements of $\boldsymbol{\beta}$. Clearly, this is an ***unbiased estimating equation*** (verify).

- In the special case of (7.12) corresponding to the ***normal distribution with constant variance*** , (7.15) reduces to

$$\sum_{j=1}^{n} f_{\beta}(\boldsymbol{x}_j, \boldsymbol{\beta})\{Y_j - f(\boldsymbol{x}_j, \boldsymbol{\beta})\} = \mathbf{0}, \tag{7.16}$$

  which is the estimating equation corresponding to ***ordinary nonlinear least squares***. If in fact $f(\boldsymbol{x}_j, \boldsymbol{\beta}) = \boldsymbol{x}_j^T\boldsymbol{\beta}$, a ***linear*** model, then $f_{\beta}(\boldsymbol{x}_j, \boldsymbol{\beta}) = \boldsymbol{x}_j$, and (7.16) are the usual ordinary least squares ***normal equations*** , as expected.

LINEAR ESTIMATING EQUATION FOR β: For the ***general mean-variance model*** (7.1), with ***no distributional assumptions*** beyond the two specified moments and possibly unknown variance parameters θ, the standard approach to estimation of β is by solving an obvious generalization of the ***linear estimating equation*** (7.15), given by

$$\sum_{j=1}^{n} f_{\beta}(\boldsymbol{x}_j, \beta) g^{-2}(\beta, \delta, \boldsymbol{x}_j)\{Y_j - f(\boldsymbol{x}_j, \beta)\} = \mathbf{0}, \tag{7.17}$$

jointly with estimating equations for the variance parameters θ, discussed momentarily. Obviously, (7.17) is an ***unbiased estimating equation***.

It is common to justify solving (7.17) under these conditions as follows. If the value of the ***variance function*** $g^2(\beta, \delta, \boldsymbol{x}_j)$ were ***known*** for each j, then the reciprocal of the variance function specifies a set of ***fixed weights*** $w_j = g^{-2}(\beta, \delta, \boldsymbol{x}_j)$, $j = 1, \ldots, n$, say. If one were to make the assumption that the distribution of $Y_j|\boldsymbol{x}_j$ is ***normal*** for each j, then the ***maximum likelihood estimator*** for β is the ***weighted least squares estimator*** , which solves

$$\sum_{j=1}^{n} f_{\beta}(\boldsymbol{x}_j, \beta) w_j\{Y_j - f(\boldsymbol{x}_j, \beta)\} = \mathbf{0}, \tag{7.18}$$

(Weighted) least squares estimation is often justified more generally, ***without*** the normality assumption, as minimizing an ***intuitively appealing objective function***, here, the ***weighted least squares*** criterion

$$\sum_{j=1}^{n} w_j\{Y_j - f(\boldsymbol{x}_j, \beta)\}^2. \tag{7.19}$$

Of course, as the variance function ***depends*** on β and δ, which are unknown, the suggestion is effectively to replace the unknown weights w_j in (7.18) and (7.19) by ***estimated weights***, formed by substituting estimators for β and δ, as we demonstrate momentarily.

QUADRATIC ESTIMATING EQUATION FOR θ: Analogous to the approach to estimation of the covariance parameters ξ in the linear longitudinal data models we discussed in Chapters 5 and 6, an appealing estimating equation to be solved to obtain an estimator for $\theta = (\sigma^2, \delta^T)^T$ can be derived by differentiating the ***loglikelihood*** corresponding to taking the distribution of $Y_j|\boldsymbol{x}_j$ to be ***normal*** with mean and variance as in (7.1).

This loglikelihood is given by (ignoring constants)

$$-(n/2)\log\sigma^2 - (1/2)\sum_{j=1}^{n}\log g^2(\beta,\delta,\boldsymbol{x}_j) - (1/2)\sum_{j=1}^{n}\frac{\{Y_j - f(\boldsymbol{x}_j,\beta)\}^2}{\sigma^2 g^2(\beta,\delta,\boldsymbol{x}_j)}. \tag{7.20}$$

This ***does not mean*** that we necessarily ***believe*** normality; we simply use this approach to derive an estimating equation. Differentiating (7.20) yields the $(r \times 1)$ estimating equation (verify)

$$\sum_{j=1}^{n}\left[\frac{\{Y_j - f(\boldsymbol{x}_j,\beta)\}^2}{\sigma^2 g^2(\beta,\delta,\boldsymbol{x}_j)} - 1\right]\begin{pmatrix}1\\ \nu_\delta(\beta,\delta,\boldsymbol{x}_j)\end{pmatrix} = \mathbf{0}, \tag{7.21}$$

where

$$\nu_\delta(\beta,\delta,\boldsymbol{x}_j) = \partial/\partial\delta\ \log g(\beta,\delta,\boldsymbol{x}_j) = \frac{\partial/\partial\delta\, g(\beta,\delta,\boldsymbol{x}_j)}{g(\beta,\delta,\boldsymbol{x}_j)}.$$

The diligent student will be sure to make the ***analogy*** to equation (5.35) for estimation of covariance parameters ξ in the linear longitudinal data models in Chapters 5 and 6.

It is straightforward to observe (verify) that if the variance model $\text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2 g^2(\beta,\delta,\boldsymbol{x}_j)$ in (7.1) is ***correctly specified*** , then (7.21) is an ***unbiased estimating equation***.

In the nonlinear modeling literature, this approach to estimation of θ, and thus δ in the "weights," in a mean-variance model (7.1) has been referred to as ***pseudolikelihood***. A ***REML*** version of (7.21) has also been proposed. Other estimating equations for θ based on ***alternatives*** to a quadratic functions of the ***deviations*** , $\{Y_j - f(\boldsymbol{x}_j,\beta)\}^2$, such as the ***absolute deviations*** $|Y_j - f(\boldsymbol{x}_j,\beta)|$, have also been proposed as a way to offer robustness to ***outliers*** ; see Carroll and Ruppert (1988, Chapter 3), Davidian and Carroll (1987), and Pinheiro and Bates (2000, Section 5.2)

GENERALIZED LEAST SQUARES: Of course, the estimating equation (7.21) must be solved ***jointly*** with the equation for β in (7.17); that is, we solve jointly in β and θ the estimating equations

$$\sum_{j=1}^{n} f_\beta(\boldsymbol{x}_j,\beta)g^{-2}(\beta,\delta,\boldsymbol{x}_j)\{Y_j - f(\boldsymbol{x}_j,\beta)\} = \mathbf{0}, \tag{7.22}$$

$$\sum_{j=1}^{n}\left[\frac{\{Y_j - f(\boldsymbol{x}_j,\beta)\}^2}{\sigma^2 g^2(\beta,\delta,\boldsymbol{x}_j)} - 1\right]\begin{pmatrix}1\\ \nu_\delta(\beta,\delta,\boldsymbol{x}_j)\end{pmatrix} = \mathbf{0}. \tag{7.23}$$

This can be implemented by an ***iterative algorithm***, starting from an initial estimate $\widehat{\beta}^{(0)}$, such as the ***nonlinear OLS estimator*** solving (7.16). At iteration ℓ,

1. Holding β fixed at $\widehat{\beta}^{(\ell)}$, solve the quadratic estimating equation (7.23) for θ to obtain $\widehat{\theta}^{(\ell)} = (\widehat{\sigma}^{2(\ell)}, \widehat{\delta}^{(\ell)T})^T$.
2. Holding δ fixed at $\widehat{\delta}^{(\ell)}$, solve the linear estimating equations (7.22) in β to obtain $\widehat{\beta}^{(\ell+1)}$. Set $\ell = \ell + 1$ and return to step 1.

A variation on step 2 is to substitute $\widehat{\beta}^{(\ell)}$ in $g^{-2}(\beta, \delta, \boldsymbol{x}_j)$ in (7.22) along with $\widehat{\delta}^{(\ell)}$, so that the "weights" are held fixed.

This procedure and variations on it is often referred to as (***estimated***) ***generalized least squares*** (GLS). One would ordinarily iterate between steps 1 and 2 to "***convergence***."

- It is important to recognize that, for arbitrary variance function $g^2(\beta, \delta, \boldsymbol{x})$, it is not necessarily the case that solving the system (7.22)-(7.23) corresponds to ***maximizing*** some ***objective function***. That is, in general, we view the resulting final estimators $(\widehat{\beta}^T, \widehat{\theta}^T)^T$ as ***M-estimators*** of the ***second type*** as in (4.2).
- Thus, there is no reason to expect that there is a ***unique solution*** to (7.22)-(7.23) or that the above algorithm should ***converge*** to a solution. ***Luckily***, in practice, it almost always does.
- Operationally, in this case it is not possible to obtain the solution $(\widehat{\beta}^T, \widehat{\theta}^T)^T$ directly by ***standard optimization techniques*** applied to an ***overall objective function*** as was the case for the longitudinal data methods in Chapters 5 and 6. Instead, an iterative algorithm like that above must be used.

  For fixed $\widehat{\beta}^{(\ell)}$, step 1 of the algorithm can in fact be carried out by maximizing the ***normal likelihood*** corresponding to general model (7.1) in θ. Then, for fixed $\widehat{\theta}^{(\ell)}$, step 2 can be carried out by so-called ***iteratively reweighted least squares*** (IRWLS), which is ***itself an iterative process*** that can be derived by taking a ***linear Taylor series*** of (7.22) in β about some β^*.

Defining $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$,

$$\boldsymbol{X}(\beta) = \begin{pmatrix} f_\beta^T(\boldsymbol{x}_1, \beta) \\ \vdots \\ f_\beta^T(\boldsymbol{x}_n, \beta) \end{pmatrix} \quad (n \times p), \qquad \boldsymbol{W}(\beta) = \text{diag}\{g^{-2}(\beta, \delta, \boldsymbol{x}_1), \ldots, g^{-2}(\beta, \delta, \boldsymbol{x}_n\}$$

for ***fixed*** δ, the ath iteration of IRWLS is

$$\beta_{(a+1)} = \beta_{(a)} + \{\boldsymbol{X}_{(a)}^T \boldsymbol{W}_{(a)} \boldsymbol{X}_{(a)}\}^{-1} \boldsymbol{X}_{(a)}^T \boldsymbol{W}_{(a)} (\boldsymbol{Y} - \boldsymbol{f}_{(a)}), \quad \boldsymbol{W}_{(a)} = \boldsymbol{W}(\beta_{(a)}), \quad \boldsymbol{X}_{(a)} = \boldsymbol{X}(\beta_{(a)}). \quad (7.24)$$

Iteration continues until some convergence criterion is met.

The ***diligent student*** will look up or verify him/herself the derivation of (7.24).

- When the mean-variance model is of the form for a generalized (non)linear model, so that there is ***no unknown*** δ in the variance function, the estimating equation (7.22) for β is in fact the ***score equation*** (7.15), and its solution corresponds to maximizing the loglikelihood, which is carried out by an IRWLS approach. Thus, IRWLS is the standard way to implement ***maximum likelihood*** in the class of generalized (non)linear models.

For future reference, we can write the system of estimating equations (7.22)-(7.23) ***compactly*** in obvious streamlined notation as (check)

$$\sum_{j=1}^{n} \begin{pmatrix} f_{\beta j} & \boldsymbol{0} \\ \boldsymbol{0} & 2\sigma^2 g_j^2 \begin{pmatrix} 1/\sigma \\ \nu_{\delta j} \end{pmatrix} \end{pmatrix} \begin{pmatrix} \sigma^2 g_j^2 & 0 \\ 0 & 2\sigma^4 g_j^4 \end{pmatrix}^{-1} \begin{pmatrix} Y_j - f_j \\ (Y_j - f_j)^2 - \sigma^2 g_j^2 \end{pmatrix} = \boldsymbol{0}. \quad (7.25)$$

QUADRATIC ESTIMATING EQUATION FOR β***:*** There is a common ***misconception*** that solving (7.22)-(7.23) corresponds to ***maximizing*** the ***normal loglikelihood*** in (7.20). Of course, (7.23) does arise from differentiating (7.20) with respect to $\boldsymbol{\theta}$.

However, it is straightforward to derive (do it) that differentiating (7.20) with respect to β yields the alternative estimating equation

$$\sum_{j=1}^{n} f_\beta(\boldsymbol{x}_j, \beta) g^{-2}(\beta, \delta, \boldsymbol{x}_j)\{Y_j - f(\boldsymbol{x}_j, \beta)\} + \sigma^2 \sum_{j=1}^{n} \left[\frac{\{Y_j - f(\boldsymbol{x}_j, \beta)\}^2}{\sigma^2 g^2(\beta, \delta, \boldsymbol{x}_j)} - 1 \right] \nu_\beta(\beta, \delta, \boldsymbol{x}_j) = \boldsymbol{0}, \quad (7.26)$$

where

$$\nu_\beta(\beta, \delta, \boldsymbol{x}_j) = \partial/\partial\beta \log g(\beta, \delta, \boldsymbol{x}_j) = \frac{\partial/\partial\beta\, g(\beta, \delta, \boldsymbol{x}_j)}{g(\beta, \delta, \boldsymbol{x}_j)}.$$

The second term in (7.26) is a result of the fact that the variance function $g^2(\beta, \delta, \boldsymbol{x}_j)$ ***depends on*** β.

Note that the first term in the estimating equation (7.26) is identical to the ***linear estimating equations*** (7.22). The second term thus demonstrates that, when the variance is believed to depend on β (usually through the ***mean response***), there is ***additional information*** about β in the ***squared deviations*** $\{Y_j - f(\boldsymbol{x}_j, \beta)\}^2$ above and beyond that in the mean itself.

- This is a consequence of the fact that the normal distribution places ***no restrictions*** on the form of the mean and variance. Intuitively, then, when the variance depends on the parameter β that describes the mean, it stands to reason that ***more*** can be learned about it from the quadratic function $\{Y_j - f(\boldsymbol{x}_j, \beta)\}^2$, which obviously reflects the nature of ***variance***.
- This suggests that, under the assumption of normality, it is possible to obtain an estimator for β that is ***more efficient*** than that obtained from the linear GLS equation.
- Of course, if the variance function ***does not depend*** on β, then (7.26) reduces to the linear equation (7.22), in which case the ***maximum likelihood estimators under normality*** for β and θ do jointly solve (7.22)-(7.23).
- In contrast, the ***scaled exponential family*** distributions with density (7.12) are such that the variance is a ***specific function of the mean*** dictated by the particular distribution. Intuitively, this suggests that, under these distributions, there is ***no additional information*** to be gained about β from the variance, reflected in the fact that the resulting estimating equation (7.15) does ***not*** involve a quadratic function of the deviations.

REMARK: A critical feature of the estimating equation (7.26) is that it is ***not enough*** for $f(\boldsymbol{x}_j, \beta)$ to be ***correctly specified*** for this to be an ***unbiased estimating equation***.

- With $f(\boldsymbol{x}_j, \beta)$ correctly specified, (7.26) is an unbiased estimating equation if the ***variance model*** $\sigma^2 g^2(\beta, \delta, \boldsymbol{x})$ is ***also correctly specified***. Thus, in general, for (7.26) to yield a ***consistent estimator*** for the true value β_0, it is necessary to specify ***both*** the mean and variance models correctly.
- Thus, there is a ***trade-off*** between ***gaining information*** about β to obtain a ***more efficient*** estimator and ending up with an ***inconsistent estimator*** for β due to ***misspecification*** of the variance model.
- Intuitively, as it is ***more difficult*** to model ***variances*** than it is to model means, this is a ***non-trivial*** concern.

In summary, under the assumption that the distribution of $Y_j|\boldsymbol{x}_j$ is ***normal*** with first two moments as in (7.1), the maximum likehood estimators for β and θ jointly solve (7.26) and (7.23). For future reference, we write this system of estimating equations ***compactly*** in streamlined form as (verify)

$$\sum_{j=1}^{n} \begin{pmatrix} f_{\beta j} & 2\sigma g_j^2 \nu_{\beta j} \\ \mathbf{0} & 2\sigma^2 g_j^2 \begin{pmatrix} 1/\sigma \\ \nu_{\delta j} \end{pmatrix} \end{pmatrix} \begin{pmatrix} \sigma^2 g_j^2 & 0 \\ 0 & 2\sigma^4 g_j^4 \end{pmatrix}^{-1} \begin{pmatrix} Y_j - f_j \\ (Y_j - f_j)^2 - \sigma^2 g_j^2 \end{pmatrix} = \mathbf{0}. \tag{7.27}$$

Of course, this system of estimating equations differs from the GLS equations in (7.25) only by the presence of the non-zero off-diagonal entry in the leftmost matrix, which serves to introduce the ***quadratic dependence*** of the equation for β and which equals zero when $g^2(\beta, \delta, \boldsymbol{x}_j)$ does not depend on β.

7.4 Large sample results

It is possible via large sample theory arguments to derive approximate sampling distributions for the estimators for β obtained by solving the ***linear estimating equation*** (7.22) jointly with (7.23), i.e., (7.25) ; or the ***quadratic estimating equation*** (7.26) jointly with (7.23), (i.e., (7.27). Here, "***large sample***" implies $n \to \infty$.

The calculations are ***simpler versions*** of those required to deduce the large sample (large m) properties of the estimators for ***general PA longitudinal data models for mean and covariance matrix*** we discuss in Chapter 8). We thus provide a brief sketch of these results for (7.25) and (7.27), whose implications carry over to the longitudinal setting.

LINEAR ESTIMATING EQUATION: Analogous to the situation of a ***possibly incorrectly specified covariance model*** in the case of the linear PA models in Section 5.5, we can carry out a similar M-estimation argument under a misspecified variance model.

Assume that we posit a ***correct*** mean model $E(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \beta)$, and suppose that the ***true*** variance ***actually generating the data*** is given by

$$\text{var}(Y_j|\boldsymbol{x}_j) = v_{0j}. \tag{7.28}$$

Suppose, however, that we posit a variance model

$$\text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2 g^2(\beta, \delta, \boldsymbol{x}_j) = v(\beta, \theta, \boldsymbol{x}_j)$$

such that ***there is not necessarily*** a $\theta_0 = (\sigma_0^2, \delta_0^T)^T$ such that $v(\beta_0, \theta_0, \boldsymbol{x}_j) = v_{0j}$.

Suppose further that we estimate $\boldsymbol{\theta}$ by solving the estimating equation (7.23) jointly with (7.22). The equation (7.23) is ***not an unbiased estimating equation*** if the variance model is ***incorrect*** ; however, assume that, under this incorrect variance model, the resulting "estimator" $\widehat{\boldsymbol{\theta}}_* = (\widehat{\sigma}^2, \widehat{\boldsymbol{\delta}}^T)^T \xrightarrow{p} (\sigma^{2*}, \boldsymbol{\delta}^{*T})^T = \boldsymbol{\theta}^*$ for some $\boldsymbol{\theta}^*$. Note that for the linear estimating equation (7.22), we then have

$$E[f_\beta(\boldsymbol{x}_j, \beta_0) g^{-2}(\beta_0, \boldsymbol{\delta}^*, \boldsymbol{x}_j)\{Y_j - f(\boldsymbol{x}_j, \beta_0)\} \mid \boldsymbol{x}_j] = \mathbf{0}, \quad j = 1, \ldots, m,$$

so that (7.22) is still an ***unbiased estimating equation***, and thus $\widehat{\beta}$ is a ***consistent estimator*** for β_0 nonetheless.

Define $v_j^* = \sigma^{2*} g^{-2}(\beta_0, \boldsymbol{\delta}^*, \boldsymbol{x}_j)$, $j = 1, \ldots, n$, and let $f_{\beta\beta}(\boldsymbol{x}, \beta) = \partial^2 / \partial\beta \partial\beta^T f(\boldsymbol{x}, \beta)$, the $(p \times p)$ matrix of second partial derivatives of $f(\boldsymbol{x}, \beta)$. Let

$$\boldsymbol{V}_0 = \text{diag}(v_{01}, \ldots, v_{0n}), \quad \boldsymbol{V}^* = \text{diag}(v_1^*, \ldots, v_n^*).$$

Note that we use $\boldsymbol{V}^*$ here ***differently*** from its definition in Chapters 5 and 6. Assume also that $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = O_p(1)$ (bounded in probability).

Expanding the right hand side of

$$\mathbf{0} = \sigma^{-2*} n^{-1/2} \sum_{j=1}^{n} f_\beta(\boldsymbol{x}_j, \widehat{\beta}) g^{-2}(\widehat{\beta}, \widehat{\boldsymbol{\delta}}, \boldsymbol{x}_j)\{Y_j - f(\boldsymbol{x}_j, \widehat{\beta})\}$$

in a Taylor series about $(\widehat{\beta}^T, \widehat{\boldsymbol{\delta}}^T)^T = (\beta_0^T, \boldsymbol{\delta}^{*T})^T$, analogous to (5.73), we obtain

$$\mathbf{0} \approx \boldsymbol{C}_n^* + (\boldsymbol{A}_{n1}^* + \boldsymbol{A}_{n2}^* + \boldsymbol{A}_{n3}^*) n^{1/2}(\widehat{\beta} - \beta_0) + \boldsymbol{E}_n^* n^{1/2}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*), \tag{7.29}$$

where (check)

$$\boldsymbol{A}_{n1}^* = n^{-1} \sum_{j=1}^{n} v_j^{*-1} f_{\beta\beta}(\boldsymbol{x}, \beta_0)\{Y_j - f(\boldsymbol{x}_j \beta_0)\} \xrightarrow{p} \mathbf{0},$$

$$\boldsymbol{A}_{n2}^* = -n^{-1} \sum_{j=1}^{n} v_j^{*-1} f_\beta(\boldsymbol{x}_j, \beta_0) f_\beta^T(\boldsymbol{x}_j, \beta_0) \xrightarrow{p} \boldsymbol{A}^* = lim_{n\to\infty} n^{-1} \boldsymbol{X}^T \boldsymbol{V}^{*-1} \boldsymbol{X}, \qquad \boldsymbol{X} = \boldsymbol{X}(\beta_0)$$

$$\boldsymbol{A}_{n3}^* = -2n^{-1} \sum_{j=1}^{n} v_j^{*-1} f_\beta(\boldsymbol{x}_j, \beta_0) \nu_\beta^T(\beta_0, \boldsymbol{\delta}^*, \boldsymbol{x}_j)\{Y_j - f(\boldsymbol{x}_j, \beta_0)\} \xrightarrow{p} \mathbf{0},$$

$$\boldsymbol{E}_n^* = -2n^{-1} \sum_{j=1}^{n} v_j^{*-1} f_\beta(\boldsymbol{x}_j, \beta_0) \nu_\delta^T(\beta_0, \boldsymbol{\delta}^*, \boldsymbol{x}_j)\{Y_j - f(\boldsymbol{x}_j, \beta_0)\} \xrightarrow{p} \mathbf{0},$$

$$\boldsymbol{C}_n^* = n^{-1/2} \sum_{j=1}^{n} v_j^{*-1} f_\beta(\boldsymbol{x}_j, \beta_0)\{Y_j - f(\boldsymbol{x}_j, \beta_0)\} \xrightarrow{\mathcal{L}} \mathcal{N}\{\mathbf{0}, \boldsymbol{B}^*\}, \qquad \boldsymbol{B}^* = \lim_{n\to\infty} n^{-1} \boldsymbol{X}^T \boldsymbol{V}^{*-1} \boldsymbol{V}_0 \boldsymbol{V}^{*-1} \boldsymbol{X}.$$

It follows that

$$n^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{A}^{*-1}\boldsymbol{B}^{*}\boldsymbol{A}^{*-1}). \tag{7.30}$$

Moreover, if in fact the variance model is ***correctly specified*** after all, so that $\delta^* = \delta_0$ for which $v_{0j} = g^2(\beta_0, \delta_0, \boldsymbol{x}_j)$, then $v_j^* = v_{0j}$, $j = 1, \ldots, n$, and (7.30) reduces to

$$n^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{A}^{-1}), \quad \boldsymbol{A} = \lim_{n\to\infty} n^{-1}\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X}. \tag{7.31}$$

- The results in (7.30) and (7.31) are of course entirely ***analogous*** to those we obtained for the ***PA linear model*** in Section 5.5, with the ***exception*** that the matrix $\boldsymbol{X} = \boldsymbol{X}(\beta_0)$ here is a ***nonlinear function*** of the true value β_0 and the covariates rather than a ***fixed design matrix***.
- In the case of a ***generalized (non)linear model*** , so that there is ***no unknown parameter*** δ, $\widehat{\beta}$ is in fact the ***MLE*** and thus (7.31) is the large sample result for maximum likelihood under a scaled exponential family distribution
- These results are used to specify ***approximate sampling distributions*** in the usual way; e.g., under the assumption the ***variance model is correctly specified*** , one would derive ***model-based standard errors*** by substituting the estimates into $\boldsymbol{X}$ and $\boldsymbol{V}_0$ to obtain, in obvious notation,

$$\widehat{\beta} \dot{\sim} \mathcal{N}[\beta_0, \{\boldsymbol{X}^T(\widehat{\beta})\boldsymbol{V}^{-1}(\widehat{\beta}, \widehat{\theta})\boldsymbol{X}(\widehat{\beta})\}^{-1}], \quad \boldsymbol{V}(\beta, \theta) = \sigma^2\text{diag}\{g^{-2}(\beta, \delta, \boldsymbol{x}_j), \ldots, g^{-2}(\beta, \delta, \boldsymbol{x}_j)\}. \tag{7.32}$$

- Likewise, ***robust*** or ***empirical standard errors*** can be derived from (7.30).

In the next chapter, we will see that analogous results hold for a ***general nonlinear population-averaged mean-covariance model***.

QUADRATIC ESTIMATING EQUATION: It is likewise possible to derive the large sample distribution of the estimator for β solving the system in (7.27) jointly in θ; that is, solving the ***quadratic estimating equation*** (7.26). Because this equation is ***not unbiased*** unless the variance function is ***correctly specified*** , the argument proceeds under the assumption that the variance function model is ***correct***. We thus assume that there are ***true values*** β_0 and θ_0 such that the posited mean and variance models yield the true mean and variance relationships.

The resulting approximate sampling distribution can be compared to that we just derived for the estimator for β solving (7.25) to gain insight into the ***potential gains in efficiency*** for estimating β achieved when the variance model is indeed correctly specified by using the ***quadratic*** rather than the ***linear equation*** under different conditions.

The argument entails expanding $n^{-1/2}\times$ (7.26) in $(\widehat{\beta}^T, \widehat{\theta}^T)^T$ about $(\beta_o^T, \theta_0^T)^T$ to find an approximate expression for

$$n^{1/2}\begin{pmatrix} \widehat{\beta} - \beta_0 \\ \widehat{\theta} - \theta_0 \end{pmatrix}$$

and then ***isolating the implied distribution*** of $n^{1/2}(\widehat{\beta} - \beta_0)$ by appealing to formulæ for the ***inverse of a partitoned matrix*** (see Appendix A).

It is ***not possible*** to expand the estimating equation (7.26) alone to arrive at this ***directly*** as we did for the ***linear estimating equation*** because it turns out that the dependence of the distribution of $\widehat{\beta}$ on that of $\widehat{\theta}$ ***does not vanish*** as it does for (7.22) above.

The argument is thus ***tedious***; accordingly we do not give it here but only present the result. The argument assumes that, although the equations (7.25) are derived under the assumption of ***normality***, the ***true distribution*** of $\boldsymbol{Y}_j|\boldsymbol{x}_j$ is ***not necessarily normal***.

HIGHER MOMENT PROPERTIES: Letting

$$\epsilon_j = \frac{Y_j - f(\boldsymbol{x}_j, \beta_0)}{\sigma_0 g(\beta_0, \delta_0, \boldsymbol{x}_j)},$$

$E(\epsilon_j^3|\boldsymbol{x}_j) = \zeta$ is the ***coefficient of skewness*** of the distribution of $Y_j|\boldsymbol{x}_j$ (***third*** moment property) and, with $\text{var}(\epsilon_j^2|\boldsymbol{x}_j) = 2 + \kappa$, κ is the ***coefficient of excess kurtosis*** (***fourth*** moment property). For the ***normal distribution***, $\zeta = \kappa = 0$.

Define $\tau_\theta(\beta, \delta, \boldsymbol{x}_j) = \{1, \nu_\delta^T(\beta, \delta, \boldsymbol{x}_j)\}^T$. Using streamlined notation where a "0" subscript indicates evaluation at the ***true values*** of the parameters, let

$$\boldsymbol{R} = \begin{pmatrix} \nu_{\beta 01}^T \\ \vdots \\ \nu_{\beta 0n}^T \end{pmatrix} \quad (n \times p), \quad \boldsymbol{Q} = \begin{pmatrix} \tau_{\delta 01}^T \\ \vdots \\ \tau_{\delta 0n}^T \end{pmatrix} \quad (n \times r), \quad \boldsymbol{P} = \boldsymbol{I} - \boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{Q}^T.$$

Then it can be shown that, if the ***skewness and excess kurtosis*** of the ***true distribution*** of $Y_j|\boldsymbol{x}_j$ are ζ and κ,

$$n^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Lambda}^{-1}\boldsymbol{\Delta}\boldsymbol{\Lambda}^{-1}), \tag{7.33}$$

$$\boldsymbol{\Lambda} = \lim_{n\to\infty} n^{-1}(\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X} + 2\boldsymbol{R}^T\boldsymbol{P}\boldsymbol{R}),$$

$$\boldsymbol{\Delta} = \lim_{n\to\infty} n^{-1}\left\{\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X} + (2+\kappa)\boldsymbol{R}^T\boldsymbol{P}\boldsymbol{R} + \zeta(\boldsymbol{X}^T\boldsymbol{V}_0^{-1/2}\boldsymbol{P}\boldsymbol{R} + \boldsymbol{R}^T\boldsymbol{P}\boldsymbol{V}_0^{-1/2}\boldsymbol{X})\right\}.$$

- The dependence of $\boldsymbol{\Delta}$ on third and fourth moment properties of the true distribution of $Y_j|\boldsymbol{x}_j$ is a consequence of the fact that the summand of the estimating equation (7.26) involves both ***linear and quadratic terms*** in $\{Y_j - f(\boldsymbol{x}_j, \beta)\}$, so that ζ and κ show up in the ***variance*** of the summand when the ***central limit theorem*** is applied.

- Both components of the covariance matrix in (7.33) depend on the covariance matrix of the linear estimator, $(\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X})^{-1}$ ***plus*** additional terms arise because of the ***quadratic component*** of the estimating equation (7.26) for β ($\boldsymbol{R}$) and the need to estimate θ ($\boldsymbol{Q}$). Thus, inclusion of the quadratic term in the estimating equation for β has the effect of making the properties of $\widehat{\beta}$ depend on those of $\widehat{\theta}$.

- When $\zeta = 0$ and $\kappa = 0$, corresponding to the ***third and fourth moments of the normal distribution***, so that the true distribution of $Y_j|\boldsymbol{x}_j$ is ***really normal*** ,

$$\boldsymbol{\Delta} = \boldsymbol{\Lambda}.$$

  Then (7.33) implies approximately that

$$\widehat{\beta} \dot{\sim} \mathcal{N}(\beta_0, n^{-1}\boldsymbol{\Lambda}^{-1}), \quad n^{-1}\boldsymbol{\Lambda}^{-1} \approx (\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X} + 2\boldsymbol{R}^T\boldsymbol{P}\boldsymbol{R})^{-1}, \tag{7.34}$$

  whereas, for the ***linear estimating equation*** when the variance function is ***correctly specified*** as we assume here, (7.31) implies approximately that

$$\widehat{\beta} \dot{\sim} \mathcal{N}(\beta_0, n^{-1}\boldsymbol{A}^{-1}), \quad n^{-1}\boldsymbol{A}^{-1} \approx (\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X})^{-1}, \tag{7.35}$$

  It is straightforward to observe that the difference

$$(\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X})^{-1} - (\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X} + 2\boldsymbol{R}^T\boldsymbol{P}\boldsymbol{R})^{-1}$$

  is ***nonnegative definite*** (check); thus, (7.34) and (7.35) imply that, when the true distribution really is normal, the quadratic estimator for β is ***more efficient*** than the linear estimator.

- However, if the true distribution is ***not normal*** and instead has ***arbitrary*** coefficients of skewness and kurtosis ζ and κ, ***relative efficiency*** of the two estimators is less clear. Approximately for large n, analogous to (7.34) and (7.35), this involves comparing $n^{-1}\boldsymbol{A}^{-1}$ in (7.35) to

$$\begin{aligned}(\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X} + 2\boldsymbol{R}^T\boldsymbol{P}\boldsymbol{R})^{-1} &\left\{\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X} + (2+\kappa)\boldsymbol{R}^T\boldsymbol{P}\boldsymbol{R} + \zeta(\boldsymbol{X}^T\boldsymbol{V}_0^{-1/2}\boldsymbol{P}\boldsymbol{R} + \boldsymbol{R}^T\boldsymbol{P}\boldsymbol{V}_0^{-1/2}\boldsymbol{X})\right\} \\ &\times(\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X} + 2\boldsymbol{R}^T\boldsymbol{P}\boldsymbol{R})^{-1}.\end{aligned}$$

  Evidently, whether or not the difference of these two covariance matrices is nonnegative definite depends in a complicated way on ζ, κ, and the matrices $\boldsymbol{R}$ and $\boldsymbol{Q}$.

The takeaway message is that, although estimation of β via the ***quadratic estimating equation*** (7.26), jointly with that of θ via (7.23), will be ***more efficient*** than using the ***linear equation*** (7.22), if $Y_j|\boldsymbol{x}_j$ is ***exactly normal*** , if it is ***not*** , it is not clear that the extra trouble is worthwhile.

Indeed, use of the quadratic equation ***requires*** that the ***variance model*** is ***correctly specified*** to achieve ***consistent estimation*** of β, so that the potential efficiency gain must be weighed against the possibility of ***misspecification*** of this model.

LARGE SAMPLE THEORY FOR VARIANCE PARAMETER ESTIMATORS: It is also possible to derive an approximate sampling distribution for the estimator for the ***variance parameter*** θ in either case. We do not pursue this here.

- From the results for the quadratic estimator for β above, because the estimating equation (7.23) depends on $\{Y_j - f(\boldsymbol{x}_j, \beta)\}^2$, we expect that properties of $\widehat{\theta}$ are sensitive to whether or not the true distribution of $Y_j|\boldsymbol{x}_j$ is ***really normal*** and thus depend on the ***coefficients of skewness and excess kurtosis*** of the true distribution.
- This reflects a ***more general*** phenomenon. The properties of estimators of ***second moment*** properties like ***variance and covariance*** depend on the ***third and fourth moment*** properties of the true distribution of the data. Thus, obtaining ***realistic assessments of uncertainty*** of such estimators is ***inherently challenging***. In particular, unless the true distribution is really ***exactly normal*** , assessments based on the assumption of normality will be ***unreliable***.

GENERALIZATION: All of these results ***generalize*** to the longitudinal data setting. We discuss some of these in Chapter 8.

CURIOSITY: We end this chapter by noting an interesting feature of the linear estimating equations (7.17) for β, namely, in shorthand,

$$\sum_{j=1}^{n} f_{\beta j}\sigma^{-2}g_j^{-2}(Y_j - f_j) = \mathbf{0}, \tag{7.36}$$

and the system of joint estimating equations (7.27) for β and θ,

$$\sum_{j=1}^{n} \begin{pmatrix} f_{\beta j} & 2\sigma g_j^2 \nu_{\beta j} \\ \mathbf{0} & 2\sigma^2 g_j^2 \begin{pmatrix} 1/\sigma \\ \nu_{\delta j} \end{pmatrix} \end{pmatrix} \begin{pmatrix} \sigma^2 g_j^2 & 0 \\ 0 & 2\sigma^4 g_j^4 \end{pmatrix}^{-1} \begin{pmatrix} Y_j - f_j \\ (Y_j - f_j)^2 - \sigma^2 g_j^2 \end{pmatrix} = \mathbf{0}. \tag{7.37}$$

It is straightforward to see or show (verify) that (7.36) and (7.37) are of the general form

$$\sum_{j=1}^{n} \mathcal{D}_j^T(\eta)\mathcal{V}_j^{-1}(\eta)\{\boldsymbol{s}_j(\eta) - \boldsymbol{m}_j(\eta)\} = \boldsymbol{0}, \tag{7.38}$$

where η is a $(k \times 1)$ vector of parameters; $\boldsymbol{s}_j(\eta)$ is a $(v \times 1)$ vector of functions of Y_j, $\boldsymbol{x}_j$, and η;

$$\boldsymbol{m}_j(\eta) = E\{\boldsymbol{s}_j(\eta)|\boldsymbol{x}_j\} \quad (v \times 1), \qquad \mathcal{V}_j(\eta) = \text{var}\{\boldsymbol{s}_j(\eta)|\boldsymbol{x}_j\} \quad (v \times v), \qquad \mathcal{D}_j(\eta) = \partial/\partial\eta^T \boldsymbol{m}_j(\eta) \quad (v \times k).$$

- The ***linear estimating equation*** for β in (7.36) with θ treated as ***fixed*** is trivially of this form, with $\eta = \beta$, $v = 1$, and

$$\boldsymbol{s}_j(\eta) = Y_j, \qquad \boldsymbol{m}_j(\eta) = f(\boldsymbol{x}_j, \beta), \qquad \mathcal{V}_j(\eta) = \sigma^2 g^2(\beta, \delta, \boldsymbol{x}_j), \qquad \mathcal{D}_j^T(\eta) = f_\beta(\boldsymbol{x}_j, \beta).$$

- The joint ***quadratic estimating equations*** in (7.37) are also of this form, with $\eta = (\beta^T, \theta^T)^T$, $v = 2$, and, in shorthand,

$$\boldsymbol{s}_j(\eta) = \begin{pmatrix} Y_j \\ (Y_j - f_j)^2 \end{pmatrix}, \qquad \boldsymbol{m}_j(\eta) = \begin{pmatrix} f_j \\ \sigma^2 g_j^2 \end{pmatrix}, \qquad \mathcal{V}_j(\eta) = \begin{pmatrix} \sigma^2 g_j^2 & 0 \\ 0 & 2\sigma^4 g_j^4 \end{pmatrix}, \tag{7.39}$$

$$\mathcal{D}_j^T(\eta) = \begin{pmatrix} f_{\beta j} & 2\sigma g_j^2 \nu_{\beta j} \\ \boldsymbol{0} & 2\sigma^2 g_j^2 \begin{pmatrix} 1/\sigma \\ \nu_{\delta j} \end{pmatrix} \end{pmatrix}.$$

  Note that $\mathcal{V}_j(\eta)$ in (7.39) is $\text{var}(\boldsymbol{s}_j|\boldsymbol{x}_j)$ ***under the assumption of normality*** , so that $\text{cov}\{Y_j, (Y_j - f_j)^2|\boldsymbol{x}_j\} = 0$ and $\text{var}\{(Y_j - f_j)^2|\boldsymbol{x}_j\} = 2\sigma^4 g_j^4$, which of course correspond to the normal, which has coefficients of skewness and excess kurtosis $\zeta = \kappa = 0$.

- This suggests that, if we instead believe that the true distribution of $Y_j|\boldsymbol{x}_j$ has skewness and kurtosis $\zeta \neq 0$, $\kappa > 0$ for some ζ and κ, the "***covariance matrix***" $\mathcal{V}_j(\eta)$ in (7.39) is ***incorrectly specified***.

- To gain insight into the consequences of this, we can make an ***analogy*** to the argument we made in Chapter 5 comparing the covariance matrices (5.75) and (5.76) that resulted from using ***correct and incorrect specifications*** for the overall covariance matrix of a response vector in the ***linear estimating equation*** for β in the linear PA models of that chapter. This argument showed that using an incorrect model for the covariance matrix $\boldsymbol{V}_i$ leads to an estimator for β that is ***inefficient*** relative to that obtained using a ***correct model*** , which corresponds to using the ***optimal linear estimating equation***.

It is straightforward to see (verify) that, if we identify $\mathcal{D}_j^T$ with $\boldsymbol{X}_i^T$, $\mathcal{V}_j$ with $\boldsymbol{V}_i$, $\boldsymbol{s}_j$ with $\boldsymbol{Y}_i$, and $\boldsymbol{m}_j$ with $\boldsymbol{X}_i\beta$ in the estimating equation (5.59), the equation (7.38), namely,

$$\sum_{j=1}^{n} \mathcal{D}_j^T(\eta)\mathcal{V}_j^{-1}(\eta)\{\boldsymbol{s}_j(\eta) - \boldsymbol{m}_j(\eta)\} = \boldsymbol{0},$$

is of the ***same form*** and can be viewed as a ***linear estimating equation*** in the "***response***" $\boldsymbol{s}_j$. Thus, the same (large sample) argument regarding inefficiency applies here with these correspondences, and thus suggests that using $\mathcal{V}_j(\eta)$ in (7.39) should result in ***inefficiency*** of the resulting estimators for β and θ ***relative to*** instead taking

$$\mathcal{V}_j(\eta) = \begin{pmatrix} \sigma^2 g_j^2 & \zeta\sigma^3 g_j^3 \\ \zeta\sigma^3 g_j^3 & (2+\kappa)\sigma^4 g_j^4 \end{pmatrix},$$

which is the "***correct covariance matrix***" and should thus result in the "***optimal linear estimating equation***" of the form (7.38).

- Of course, it is ***extremely unlikely*** we would ever know the true ζ and κ in practice. However, this shows that, by assuming normality, we are ***effectively*** making the assumption that the ***first four moments*** of the distribution of $Y_j|\boldsymbol{x}_j$ are the ***same*** as those of the normal distribution with mean and variance given by the posited mean-variance model (7.1).

These considerations will arise in a ***multivariate*** context in the overview of ***generalized estimating equations*** in the next chapter.

8 Population-Averaged Models and Generalized Estimating Equations

8.1 Introduction

In this chapter, we consider ***population-averaged*** models for longitudinal data where

(i) the responses may be ***discrete*** ,

(ii) an appropriate model for the ***overall population mean response*** trajectory may be ***nonlinear*** in parameters; and/or

(iii) the aggregate ***variance*** of the response given covariates is possibly a ***function*** of the parameters in the mean model ***and*** additional, unknown ***variance parameters***.

These models extend those of Chapter 5 to incorporate these features and can also be viewed as a ***multivariate extension*** of the mean-variance models for univariate response in Chapter 7. The models are of course all applicable to ***continuous*** responses under conditions (ii) and (iii).

As we discussed in Section 2.7 , when the response is ***discrete*** , specification of a full ***multivariate distribution*** for $\boldsymbol{Y}_i|\boldsymbol{x}_i$ is problematic and thus an infeasible basis for modeling and inference. As we noted in that section, unlike for the normal distribution, where the extension from univariate to multivariate is immediate, multivariate versions of discrete distributions have densities that depend in a complicated way on so-called ***higher-way associations*** among the elements of a data vector. Moreover, in contrast to the multivariate normal distribution, for which there are no ***restrictions*** on the nature of ***pairwise correlations*** among elements of a response vector, discrete multivariate responses do involve ***complicated restrictions*** on these. We demonstrated this in the particular case of ***binary response*** in Section 2.7.

These challenges led to a ***classic paper*** , Liang and Zeger (1986), in which the authors proposed a framework in which one specifies models only for the ***first two moments*** of the distribution of $\boldsymbol{Y}_i|\boldsymbol{x}_i$. In particular, one posits models for the ***mean response*** and ***aggregate variance of the response*** along with a "***working model*** " for the ***aggregate correlation*** among elements of a response vector. The working model is most certainly ***misspecified*** , but the hope is that it can capture the salient features of pairwise correlation structure and lead to a more efficient analysis than, say, erroneously assuming all observations are ***independent***.

Liang and Zeger (1986) popularized this approach, which is now considered fundamental. This original paper restricts to considering responses such as binary data, counts, and so on that from a univariate point of view could be modeled by the ***scaled exponential family*** , and to models for the mean and variance that are thus of the "***generalized linear model type***," with ***no unknown parameters*** in the variance model. However, this restriction is ***unnecessary*** ; accordingly, in this chapter, we consider more general nonlinear mean models and variance functions depending on possibly unknown parameters.

Liang and Zeger (1986) proposed that the model be fitted by solving a ***linear estimating equation*** for the mean parameters along with a suitable method for estimating the parameters in the ***working correlation model***. Subsequently, other authors proposed refinements and extensions of this approach. There are numerous references that cover the types of estimating equations we are about to discuss. Some of the key references are Prentice (1988), Zhao and Prentice (1990), Prentice and Zhao (1991), and Liang, Zeger, and Qaqish (1992). See also Chapter 9 of Vonesh and Chinchilli (1997) and Chapter 3 of Fitzmaurice et al. (2009) and the references therein.

In this chapter, we discuss this modeling framework and key inferential approaches. We also discuss an issue that we have so far downplayed, that of the difficulties that arise in modeling and interpretation when covariate information is ***time-dependent***.

8.2 Model specification

DATA, RESTATED: We first restate the form of the observed data for convenience. These data are

$$(\boldsymbol{Y}_i, \boldsymbol{z}_i, \boldsymbol{a}_i) = (\boldsymbol{Y}_i, \boldsymbol{x}_i), \quad , i = 1, \ldots, m,$$

independent across i, where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$, with Y_{ij} recorded at time t_{ij}, $j = 1, \ldots, n_i$ (possibly different times for different individuals); $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}^T, \ldots, \boldsymbol{z}_{in_i}^T)^T$ comprising ***within-individual*** covariate information $\boldsymbol{u}_i$ and the t_{ij}; $\boldsymbol{a}_i$ is a vector of ***among-individual*** covariates; and $\boldsymbol{x}_i = (\boldsymbol{z}_i^T, \boldsymbol{a}_i^T)^T$.

Before we state the basic model, we note an important consideration that we discuss in detail in Section 8.6. Up to now, we have assumed that the among-individual covariates $\boldsymbol{a}_i$ ***do not change over time*** , as is the case with individual characteristics such as gender, treatment group in a conventional randomized study, or characteristics such as age or weight ascertained once at ***baseline***.

If there are any within-individual covariates $\boldsymbol{u}_i$, we have assumed that these are either ***time-independent***, as in the case of a one-time drug dose D_i administered to subject i in a pharmacokinetic study, or are determined according to a ***fixed design***, as would be the case in a pharmacokinetic study where each participant is to receive repeated doses over several fixed ***dosing intervals***.

Of course, this ***need not always*** be the case. To fix ideas, recall the following example.

EXAMPLE 6: Maternal smoking and child respiratory health, continued. Recall from Chapter 1 the example of the Six Cities Study. In this part of the study, $m = 300$ mother-child pairs were to be examined once a year at the child's ages 9–12, so that the ***intended number*** of examinations for each pair i is $n = 4$, although some pairs are missing data at some examinations, so that $n_i \leq 4$ in general. At each examination, the outcome of interest is the ***binary response*** "***wheezing status***," coded as 0 (no) or 1 (yes), where 1 corresponds to respiratory problems. ***Maternal smoking status*** was recorded at each examination as a categorical variable indicating the current level of the mother's smoking: 0=none, 1=moderate, 2=heavy. The ***city*** in which each pair lived was also recorded.

The goals of this portion of the study were to determine how the ***wheezing response pattern*** in the population changes with age and how it might be associated with ***level of maternal smoking***. These are questions regarding ***population-averaged*** phenomena; for example, as is usually the case in studies of ***public health***, the second question is focused on the association between maternal smoking and child respiratory status overall in the ***population***.

As we discussed in Section 2.2, from this point of view, the ***covariate*** maternal smoking status would ordinarily be interpreted as an ***among-individual*** covariate, as it reflects how a child was ***treated*** over the period of the study and thus is relevant to the overall ***population-level*** question.

Thus, in this study, if mother-child pair i from city c_i (= 0 for Portage, = 1 for Kingston) was examined at ages t_{ij}, $j = 1, \ldots, n_i$, we have $\boldsymbol{a}_i = (c_i, s_{i1}, \ldots, s_{in_i})^T$, where s_{ij}, $j = 1, \ldots, n_i$, is such that $s_{ij} = 0, 1, 2$ according to mother i's smoking status at t_{ij}. There are no within-individual level covariates. We can thus write $\boldsymbol{x}_i$, the collection of all covariates on pair i, as

$$\boldsymbol{x}_i = (\boldsymbol{x}_{i1}^T, \ldots, \boldsymbol{x}_{in_i}^T)^T = \{(t_{i1}, c_i, s_{i1})^T, \ldots, (t_{in_i}, c_i, s_{in_i})^T\}^T, \tag{8.1}$$

where $\boldsymbol{x}_{ij} = (t_{ij}, c_i, s_{ij})^T$ records the age, city, and smoking status at the jth examination for pair i.

We contrast this situation with one in which ***among-individual*** covariates do not change over time.

EXAMPLE 5: Epileptic seizures and chemotherapy, continued. Recall from Chapter 1 the clinical trial conducted in $m = 59$ subjects suffering from simple or partial seizures, who were randomized to the anti-epileptic drug progabide or placebo. On each, a ***baseline measure*** of each subject's propensity for seizures was recorded, namely, the number of seizures suffered in the 8 weeks leading up to the start of the study. Also recorded was each subject's ***age*** at the start of assigned treatment. After initiation of assigned treatment, the ***number of seizures*** experienced by each subject in $n = 4$ consecutive two-week periods was recorded, so that the response is a ***count***. Seizure counts were recorded for all m subjects at all n periods, so there are no missing data.

Thus, in this study, the ***among-individual covariates*** are the assigned treatment $\delta_i = 0$ for placebo and 1 for progabide, baseline seizure count c_i, and age a_i at the start of the study, both of which are ***time-independent*** , and there are no within-individual covariates. Thus, letting $\boldsymbol{a}_i = (\delta_i, c_i, a_i)^T$, $i = 1, \ldots, m$, and letting $t_j = 1, 2, 3, 4$ indicate the observation period, $j = 1, \ldots, 4$. The we can write $\boldsymbol{x}_i$, the collection of all covariates on subject i, as

$$\boldsymbol{x}_i = (\boldsymbol{x}_{i1}^T, \ldots, \boldsymbol{x}_{in_i}^T)^T = \{(t_{i1}, \boldsymbol{a}_i^T)^T, \ldots, (t_{in_i}, \boldsymbol{a}_i^T)^T\}^T = \{(t_1, \boldsymbol{a}_i^T)^T, \ldots, (t_4, \boldsymbol{a}_i^T)^T\}^T. \tag{8.2}$$

In (8.2), $\boldsymbol{x}_{ij}$, the component of $\boldsymbol{x}_i$ associated with time j, involves among-individual covariates that ***do not change*** over time. Thus, the subscript j thus corresponds only to the time of the jth response measure. In constrast, in (8.1), $\boldsymbol{x}_{ij}$ involves among-individual covariate information that ***does change*** over time.

Although it is conventional to write the model for mean response in terms of $\boldsymbol{x}_{ij}$ as we do below, it is it is ***critical*** to appreciate exactly what one ***implicitly assumes*** when specifying such a model in a situation like (8.1), as we discuss momentarily and in detail in Section 8.6.

BASIC MODEL: We focus on the general PA mean-covariance model of the form

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{f}_i(\boldsymbol{x}_i, \beta), \qquad \text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i). \tag{8.3}$$

- In model (8.3), with $\boldsymbol{x}_i$ partitioned as $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}^T, \ldots, \boldsymbol{x}_{in_i}^T)^T$ as in (8.1) and (8.2), $\boldsymbol{f}_i(\boldsymbol{x}_i, \beta)$ is ordinarily taken to be the $(n_i \times 1)$ vector

$$\boldsymbol{f}_i(\boldsymbol{x}_i, \beta) = \begin{pmatrix} f(\boldsymbol{x}_{i1}, \beta) \\ \vdots \\ f(\boldsymbol{x}_{in_i}, \beta) \end{pmatrix} \quad (n_i \times 1) \tag{8.4}$$

  for some function $f(\boldsymbol{x}, \beta)$ that may be ***nonlinear*** in β $(p \times 1)$. We say more about this model below.

- The covariance model $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ $(n_i \times n_i)$ is taken to have the form

$$\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i) = \boldsymbol{T}_i^{1/2}(\beta, \theta, \boldsymbol{x}_i)\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)\boldsymbol{T}_i^{1/2}(\beta, \theta, \boldsymbol{x}_i). \quad (8.5)$$

- In (8.5), $\boldsymbol{T}_i(\beta, \theta, \boldsymbol{x}_i)$ is the $(n_i \times n_i)$ diagonal matrix whose diagonal elements are the models for $\text{var}(Y_{ij}|\boldsymbol{x}_i)$, e.g., involving a variance function

$$\text{var}(Y_{ij}|\boldsymbol{x}_i) = \sigma^2 g^2(\beta, \delta, \boldsymbol{x}_{ij}),$$

  so depending on possibly unknown variance parameters $\theta = (\sigma^2, \delta^T)^T$ $(r \times 1)$. In the case of popular models for responses that would from a univariate point of view follow a ***scaled exponential family*** distribution, the variance function g^2 ***does not depend*** on unknown parameters δ. Moreover, for some of these distributions, such as the Bernoulli for binary response and the Poisson for responses in the form of counts, the ***scale parameter*** σ^2 might also be taken to be known and $\sigma^2 = 1$. In the case of possible ***overdispersion***, as discussed in Section 7.2, an unknown scale parameter σ^2 would be incorporated.

  Considerations for specifying an appropriate ***variance function model*** are as in Section 7.2.

- In (8.5), the $(n_i \times n_i)$ matrix $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$ is a correlation matrix that generally depends on $\boldsymbol{x}_i$ only through the within-individual times or other conditions t_{ij} at which observations in $\boldsymbol{Y}_i$ are taken. Here, α $(s \times 1)$ is a vector of unknown correlation parameters. Considerations for specifying $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$ are discussed momentarily.

- The vector of variance and correlation parameters $\xi = (\theta^T, \alpha^T)^T$ $(r + s \times 1)$ may be ***entirely unknown***. Alternatively, it may be that only α is unknown in models where the form of the variance function is ***entirely specified***, as discussed above.

Model (8.3) can be viewed as a ***multivariate analog*** to the univariate mean-variance models discussed in Chapter 7. Thus, it should come as no surprise that inferential strategies for (8.3) exploit some of the same ideas as in the univariate case.

In particular, estimation of β and ξ is typically carried out by solution of ***linear and quadratic estimating equations*** that are similar in spirit to those discussed in Chapters 5, 6, and 7. Of necessity, these equations are ***more complicated***, as we will see in Sections 8.3 and 8.4, although the basic principles are the same. The term ***generalized estimating equations*** (GEEs), first coined by Liang and Zeger (1986), has come to refer broadly to the body of techniques for inference for β and ξ based on solution of such estimating equations.

- As suggested by the title of Liang and Zeger (1986), "Longitudinal data analysis using generalized linear models," and noted above, the original formulation was restricted to responses and models of the "***generalized linear model***" type. This explains why in much of the classical literature on GEEs there are ***no unknown variance parameters*** δ, and interest focuses exclusively on estimating ***correlation parameters*** α in a "working" correlation model as discussed below and possibly a scale parameter σ^2 in the case of ***overdispersion***.

- Our development allows the model $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ to involve ***both*** unknown variance parameters θ and unknown correlation parameters α. We note simplifications that occur in the case where the variance function does not depend on any unknown parameters θ.

WORKING CORRELATION MODEL: In accordance with the considerations given at the beginning of this chapter, $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$ is often a ***working correlation model*** that is acknowledged in many circumstances to be ***incorrect*** but that is specified as a way to accommodate the expected ***aggregate pairwise correlations*** among elements of $\boldsymbol{Y}_i$.

- As for the linear population-averaged models in Chapter 5, the modeling framework (8.3) is feasible in situations where there are *n* ***intended times*** at which all individuals are to be observed. (We discuss implications of ***missing responses*** in this context in Section 8.7.)

- In this case, $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$ is might taken to be ***completely unstructured***, so that $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ involves $n(n-1)/2$ unknown correlation parameters in addition to possibly unknown variance parameters θ. Although the large sample theory we discuss in Section 8.5 suggests that, analogous to that in Sections 5.5 and 7.4, the properties of the estimator for β ***do not depend*** on whether or not ξ is ***estimated or known***, in finite samples (m), estimation of a large number of correlation and variance parameters can lead to ***inefficient*** estimation of β.

- This framework can also be used in more general situations where the observation times t_{ij} are ***different for different individuals***, although the working correlation models that are practically feasible in this situation are ***limited***.

- Accordingly, it is popular to use working correlation models that involve a ***small number of correlation parameters*** α. If it is thought that ***within-individual sources of correlation*** due to time-ordered data collection are dominant, AR(1), exponential, or Gaussian correlation models might be used. Conversely, if ***among-individual sources of correlation*** dominate, a ***compound symmetric*** correlation model might be selected. This model is in principle feasible even if the observation times are ***different*** for different individuals.

- Recognizing that such working models are ***almost certainly misspecified*** , it is commonplace to use the appropriate form of the ***robust sandwich*** or ***empirical*** estimator for the covariance matrix of the approximate sampling distribution of the estimator for β to assess uncertainty, as we demonstrate in Section 8.5.

SPECIFICATION OF THE MEAN MODEL: Models for the $E(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ are specified in accordance with the particular type of response and the nature of the questions of interest. For example, if Y_{ij} is ***binary*** , then a natural model is a general ***logistic*** model as in (7.2), i.e.,

$$f(\boldsymbol{x}_{ij},\beta) = \frac{\exp\{h(\boldsymbol{x}_{ij})^T\beta\}}{1+\exp\{h(\boldsymbol{x}_{ij})^T\beta\}}, \quad \text{or equivalently} \quad \text{logit}\{f(\boldsymbol{x}_{ij},\beta)\} = \log\left\{\frac{f(\boldsymbol{x}_{ij},\beta)}{1-f(\boldsymbol{x}_{ij},\beta)}\right\} = h(\boldsymbol{x}_{ij})^T\beta, \tag{8.6}$$

where $h(\cdot)$ is a vector of functions of $\boldsymbol{x}_{ij}$. Similarly, for Y_{ij} in the form of counts, a ***loglinear model***

$$f(\boldsymbol{x}_{ij},\beta) = \exp\{h(\boldsymbol{x}_{ij})^T\beta\) \quad \text{or equivalently} \quad \log\{f(\boldsymbol{x}_{ij},\beta)\} = h(\boldsymbol{x}_{ij})^T\beta \tag{8.7}$$

would be an obvious choice. Other models, linear or nonlinear, can of course be posited as appropriate. E.g., for a ***continuous response*** for which the population mean trajectory approaches an ***asymptote*** as $t \to \infty$, a possible model is

$$f(\boldsymbol{x}_{ij},\beta) = \beta_1 + (\beta_2 - \beta_1)\exp\{-\exp(-\beta_3)t_{ij}\},$$

where the rate constant is parameterized to enforce positivity.

As noted in (8.4), with $\boldsymbol{x}_i$ partitioned as $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}^T,\ldots,\boldsymbol{x}_{in_i}^T)^T$, the jth element of $\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)$ is ordinarily taken to depend on $\boldsymbol{x}_i$ ***through*** $\boldsymbol{x}_{ij}$ ***only***. It is ***critical*** that the data analyst understand the implications of this, as we now discuss.

Specifically, as $\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)$ is a model for $E(\boldsymbol{Y}_i|\boldsymbol{x}_i)$, its jth component is a model for $E(Y_{ij}|\boldsymbol{x}_i)$. Thus, restricting the jth component of $\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)$, $f(\boldsymbol{x}_{ij},\beta)$ to depend on $\boldsymbol{x}_i$ ***only through*** $\boldsymbol{x}_{ij}$ implicitly embodies the assumption that

$$E(Y_{ij}|\boldsymbol{x}_i) = E(Y_{ij}|\boldsymbol{x}_{i1},\ldots,\boldsymbol{x}_{in_i}) = E(Y_{ij}|\boldsymbol{x}_{ij}). \tag{8.8}$$

It is ***critical*** to recognize that (8.8) is an ***assumption*** that ***may or may not*** hold.

- In a situation like that of ***EXAMPLE 5*** , the seizure trial, $\boldsymbol{x}_{ij}$ involves only the ***time-independent covariate*** $\boldsymbol{a}_i$, which includes treatment, baseline seizure count, and age at study entry. Here, then, $\boldsymbol{x}_{ij}$ varies with j ***only*** through the ***planned*** time periods t_j, so, effectively,

  $$E(Y_{ij}|\boldsymbol{x}_i) = E(Y_{ij}|\boldsymbol{a}_i)$$

  for all j, in which case (8.8) with $\boldsymbol{x}_{ij} = (t_{ij},\boldsymbol{a}_i^T)^T$ ***necessarily holds***.

- Similarly, consider a study where each individual is assigned to receive n different doses of a drug d_j on n separate occasions $t_1, \ldots, t_n$, and the response Y_{ij} is ascertained at each occasion for each individual i. Interest focuses on the ***relationship*** between population mean response and dose. Assuming ***no missing data***, $\boldsymbol{x}_{ij} = (t_j, d_j)$ for all i, and thus $\boldsymbol{x}_i = \{(t_1, d_1)^T, \ldots, (t_n, d_n)^T\}^T$. Here, then, the $\boldsymbol{x}_{ij}$, are ***fixed by design***, specified ***a priori*** in a way that is ***unrelated to*** the responses that they might elicit, rather than ***observed***.

  Thus, the relationship between dose and response is ***clear cut***. E.g., if the goal is to evaluate the relationship between mean response and ***current dose level***, assuming that the occasions are ***sufficiently far apart*** that effects of previous doses on the current response have "***washed out***," (8.8), so that the response at occasion j depends on all doses ***only*** through the dose given at occasion j, is reasonable. If we redefine $\boldsymbol{x}_{ij} = (t_j, d_1 + \cdots + d_j)$ and wish to evaluate the relationship between mean response and ***cumulative dose***, the assumption in (8.8) is also reasonable. The main issue is that, because the $\boldsymbol{x}_{ij}$ are ***fixed in advance***, their values are ***not impacted*** by the response. Contrast this situation with the following.

- Consider the setting of the Six Cities study. Recall that $\boldsymbol{x}_{ij} = (t_{ij}, c_i, s_{ij})^T$, where t_{ij} is the time variable (examination occasion corresponding to the age of the child), c_i is the (fixed) city, and s_{ij} is the mother's smoking status status at the jth examination for pair i. Here, smoking status is ***observed*** rather than dictated by design.

  Suppose that, at examination j for pair i, the child's wheezing status $Y_{ij} = 1$, and the mother is currently engaging in heavy smoking, $s_{ij} = 2$. After seeing her child's wheezing result, the mother decides to ***cut back*** on her ***future smoking***. In this case, her smoking status $s_{i,j+1}$ at the $(j+1)$th examination is ***associated with*** and thus ***not independent of***, the child's wheezing status Y_{ij} at the jth examination. Clearly, then, it ***cannot be true*** that

$$E(Y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{x}_{i,j+1}) = E(Y_{ij}|\boldsymbol{x}_{ij}), \tag{8.9}$$

  because, given the smoking status s_{ij} in $\boldsymbol{x}_{ij}$ at examination j, the mother's smoking status $s_{i,j+1}$ in $\boldsymbol{x}_{i,j+1}$ is ***not independent*** of Y_{ij}. Applying this reasoning more generally, it should be clear that (8.8) ***cannot hold*** under these circumstances.

- From a ***causal perspective***, under which we wish to attribute mother's smoking status s_{ij} at examination j as "***causing***" the child's wheezing status Y_{ij} at j, smoking status $s_{i,j+1}$ ***confounds*** the relationship between Y_{ij} and s_{ij}.

That is, if (8.9) holds, we ***cannot hope*** to isolate the effect of smoking status at examination j on wheezing at j. If we adopt a model in that implicitly assumes (8.8), the parameter β in that model ***does not characterize*** the ***causal mechanism*** of interest, and thus ***misleading inferences*** could result.

- In general, for ***time-dependent among-individual covariates*** that are ***not fixed by design***, as in the dose-response study above, ***great care*** must be taken in modeling relationships between response and covariates and interpretating the model. ***Blindly adopting*** a model for which conditional mean at time j depends only on covariates at time j as in (8.4) can lead to ***complex difficulties*** with ***interpretation***. We discuss this further in Section 8.6.

8.3 Linear estimating equations

LINEAR ESTIMATING EQUATION: Analogous to developments in Chapter 5 for linear PA models and Chapter 7 for the case of univariate response, we can derive an estimating equation for β by considering the situation where the covariance matrix $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ is ***known***. Writing $\boldsymbol{V}_i$ to denote this known matrix and adopting a working assumption of ***normality*** for $\boldsymbol{Y}_i|\boldsymbol{x}_i$, we can differentiate the loglikelihood

$$\log L = -(1/2)\sum_{i=1}^{m}\left[\log|\boldsymbol{V}_i| + \{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}^T\boldsymbol{V}_i^{-1}\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}\right]$$

with respect to β to obtain the estimating equation

$$\sum_{i=1}^{m}\boldsymbol{X}_i^T(\beta)\boldsymbol{V}_i^{-1}\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\} = \boldsymbol{0}, \qquad \boldsymbol{X}_i(\beta) = \begin{pmatrix} f_\beta^T(\boldsymbol{x}_{i1},\beta) \\ \vdots \\ f_\beta^T(\boldsymbol{x}_{in_i},\beta) \end{pmatrix} \quad (n_i \times p). \tag{8.10}$$

This follows from the same matrix differentiation results used in Section 5.3. Alternatively, we can take a "***weighted least squares***" point of view to arrive at (8.10), analogous to (7.18).

When $\boldsymbol{V}_i$ is taken to depend on β but not on covariance parameters ξ, we can make a multivariate analogy to the ***maximum likelihood estimating equation*** for the ***scaled exponential family*** in (7.15) to arrive at an estimating equation of the form in (8.10), where now $\boldsymbol{V}_i$ also depends on β.

These considerations suggest that, when the model $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ depends on ***both*** β and covariance parameters ξ, we estimate β by solving the ***linear estimating equation***

$$\sum_{i=1}^{m}\boldsymbol{X}_i^T(\beta)\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\} = \boldsymbol{0}, \tag{8.11}$$

where an estimator for ξ is substituted.

Analogous to the streamlined notation used in Chapter 5, define $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, \ldots, \boldsymbol{Y}_m^T)^T$,

$$\boldsymbol{f}(\beta) = \{\boldsymbol{f}_1^T(\boldsymbol{x}_1, \beta), \ldots, \boldsymbol{f}_m^T(\boldsymbol{x}_1, \beta)\}^T \quad (N \times 1), \quad \boldsymbol{X}(\beta) = \begin{pmatrix} \boldsymbol{X}_1(\beta) \\ \vdots \\ \boldsymbol{X}_m(\beta) \end{pmatrix} \quad (N \times p), \tag{8.12}$$

$$\boldsymbol{V}(\beta, \xi) = \text{block diag}\{\boldsymbol{V}_1(\beta, \xi, \boldsymbol{x}_1), \ldots, \boldsymbol{V}_m(\beta, \xi, \boldsymbol{x}_m)\}, \quad (N \times N).$$

Then we can write (8.11) compactly as

$$\boldsymbol{X}^T(\beta)\boldsymbol{V}^{-1}(\beta, \xi)\{\boldsymbol{Y} - \boldsymbol{f}(\beta)\} = \boldsymbol{0}. \tag{8.13}$$

- ***Even if*** the covariance model $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ is ***misspecified***, the estimating equation in (8.11) and equivalently (8.13) is ***unbiased*** as long as the model for the mean is correctly specified ***and*** satisfies conditions we discuss in Section 8.6.
- Given that the correlation model incorporated in $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ is ***quite possibly misspecified***, as discussed above, this is ***critically important***.

Henceforth in this chapter, except when we discuss modeling considerations in Section 8.6, we assume that the model $\boldsymbol{f}_i(\boldsymbol{x}_i, \beta)$ for $E(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ is ***correctly specified***.

IMPLEMENTATION: By analogy to the univariate case, an obvious strategy for solving (8.11) is to use a multivariate version of the ***generalized least squares algorithm*** for solving (7.22) in Section 7.3. Specifically, start with an ***initial estimate*** $\widehat{\beta}^{(0)}$. A natural choice for $\widehat{\beta}^{(0)}$ is the ***OLS estimator*** treating all N elements of $\boldsymbol{Y}$ as if they were ***mutually independent*** with the ***same*** conditional variance; i.e., replacing $\boldsymbol{V}$ in (8.13) by a $(N \times N)$ identity matrix. That the OLS estimator is ***consistent*** for the true value of β follows from arguments in Section 8.5, as we discuss shortly. Then at iteration ℓ:

1. Holding β fixed at $\widehat{\beta}^{(\ell)}$, estimate ξ by $\widehat{\xi}^{(\ell)}$; we discuss approaches to doing this momentarily.
2. Substitute $\widehat{\xi}^{(\ell)}$ for ξ in $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ in (8.11). Then holding ξ fixed at $\widehat{\xi}^{(\ell)}$ solve the linear estimating equations (8.11) in β to obtain $\widehat{\beta}^{(\ell+1)}$. Set $\ell = \ell + 1$ and return to step 1.

A variation on step 2 is to substitute $\widehat{\beta}^{(\ell)}$ in $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ in (8.11) along with $\widehat{\xi}^{(\ell)}$, so that the "***weights***" are held fixed.

As in the univariate case, one would iterate between steps 1 and 2 until "***convergence***." As in that case, it is not necessarily true that this algorithm should ***converge***, as the procedure does not correspond to maximizing some ***objective function***.

ESTIMATION OF COVARIANCE PARAMETERS: By analogy to the univariate case, a natural approach to estimating ξ would be to solve a suitable ***quadratic estimating equation***, as we discuss shortly.

In the early papers on GEEs in the biostatistical literature by Liang and Zeger, estimation of ξ was advocated based on ***simple, moment-based approaches***. Actually, as this early work was in the context of "***generalized linear model-type***" problems, this involved estimation only of correlation parameters α, as in this setting the variance function ***does not depend*** on unknown parameters (except perhaps a scale parameter σ^2 in the case of overdispersion or distributions like the gamma).

For example, for the ***exponential correlation model*** , reparameterized here as

$$\text{corr}(Y_{ij}, Y_{ij'}|\boldsymbol{x}_i) = \alpha^{|t_{ij}-t_{ij'}|},$$

where Y_{ij} and $Y_{ij'}$ are observed at times t_{ij} and $t_{ij'}$, and $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$ depends on the scalar parameter α. Assume that $\text{var}(Y_{ij}|\boldsymbol{x}_{ij}) = \sigma^2 g^2(\beta, \delta, \boldsymbol{x}_{ij})$ with δ ***known***. Given an estimate $\widehat{\beta}^{(\ell)}$ at the ℓth iteration of the above algorithm, the weighted residuals

$$wr_{ij} = \{Y_{ij} - f(\boldsymbol{x}_{ij}, \widehat{\beta}^{(\ell)})\}/g(\widehat{\beta}^{(\ell)}, \delta, \boldsymbol{x}_{ij})$$

have ***approximate*** mean 0 and satisfy

$$E(wr_{ij}wr_{ij'}|\boldsymbol{x}_i) \approx \sigma^2\alpha^{|t_{ij}-t_{ij'}|}.$$

Taking logarithms of both sides of this expression yields the approximate relationship

$$\log(wr_{ij}wr_{ij'}) \approx \log\sigma^2 + |t_{ij} - t_{ij'}|\log\alpha;$$

thus, the suggestion was to form ***all pairs of lagged residuals*** for each i, ***pool*** them together, and estimate $\log\alpha$ by ***simple linear regression*** of the $\log(wr_{ij}wr_{ij'})$ on the $|t_{ij} - t_{ij'}|$. The resulting estimator may be exponentiated to yield an estimator for α.

Similar ***moment-based methods*** were proposed by Liang and Zeger for other correlation models.

In subsequent publications, it was proposed ***instead*** to estimate α, and more generally ξ, by solving a ***suitable estimating equation*** derived analogously to those used in Chapters 5 and 7 by starting with the loglikelihood under the assumption $\boldsymbol{Y}_i|\boldsymbol{x}_i$ is normally distributed, namely

$$\log L = -(1/2)\sum_{i=1}^{m}\left[\log|\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)| + \{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}^T\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}\right]. \tag{8.14}$$

Using ***matrix differentiation results*** in Appendix A as we did in Section 5.3, treating β as fixed and taking the partial derivatives of (8.14) with respect to each element ξ_k, $k = 1,\ldots,r+s$, of ξ, we obtain the $(r+s\times 1)$ estimating equations

$$(1/2)\sum_{i=1}^{m}\Big(\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}^T\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\partial/\partial\xi_k\,\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}$$
$$-\operatorname{tr}\left[\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\partial/\partial\xi_k\{\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\right]\Big) = 0,\quad k = 1,\ldots,r+s. \tag{8.15}$$

By analogy to the linear and univariate cases, if $\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)$ is ***correctly specified***, then using the result for the expectation of a quadratic form in Appendix A, it is straightforward (verify) that (assuming expectation is under the parameter values β and ξ)

$$\begin{aligned}E\Big[\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}^T\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\partial/\partial\xi_k\,\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}\mid\boldsymbol{x}_i\Big]&\\ = \operatorname{tr}\left[\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\partial/\partial\xi_k\,\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\right]&\\ = \operatorname{tr}\left[\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\partial/\partial\xi_k\{\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\right],&\end{aligned}$$

from whence it follows that (8.15) is an ***unbiased estimating equation***.

These considerations lead to the $(p+r+s\times 1)$ system of estimating equations

$$\sum_{i=1}^{m}\boldsymbol{X}_i^T(\beta)\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\} = \boldsymbol{0}, \tag{8.16}$$

$$(1/2)\sum_{i=1}^{m}\Big(\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}^T\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\partial/\partial\xi_k\,\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}$$
$$-\operatorname{tr}\left[\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\partial/\partial\xi_k\{\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\right]\Big) = 0,\quad k = 1,\ldots,r+s. \tag{8.17}$$

The multiplicative factor of (1/2) in the equation for ξ in (8.17) could be disregarded, but we maintain it for now, as it proves important for the developments of Section 8.9, the motivation for which we now discuss.

ALTERNATIVE FORM OF THE QUADRATIC ESTIMATING EQUATION FOR ξ: In a series of papers in the late 1980s/early 1990s, Prentice (1988), Zhao and Prentice (1990), and Prentice and Zhao (1991), an ***alternative way*** of representing estimating equations such as (8.17) was popularized and became the standard way to write such equations.

Recall from (7.38) that we recognized that the ***system of estimating equations*** (7.37) could be seen to be of the ***generic form*** (7.38). In the current context, this generic form is

$$\sum_{i=1}^{m} \mathcal{D}_i^T(\eta)\mathcal{V}_i^{-1}(\eta)\{\boldsymbol{s}_i(\eta) - \boldsymbol{m}_i(\eta)\} = \boldsymbol{0}, \tag{8.18}$$

where η is a $(k \times 1)$ vector of parameters; $\boldsymbol{s}_i(\eta)$ is a $(v \times 1)$ vector of functions of Y_i, $\boldsymbol{x}_i$, and η;

$$\boldsymbol{m}_i(\eta) = E\{\boldsymbol{s}_i(\eta)|\boldsymbol{x}_i\} \quad (v \times 1), \qquad \mathcal{V}_i(\eta) = \text{var}\{\boldsymbol{s}_i(\eta)|\boldsymbol{x}_i\} \quad (v \times v), \qquad \mathcal{D}_i(\eta) = \partial/\partial\eta^T \boldsymbol{m}_i(\eta) \quad (v \times k).$$

It is possible to express (8.17) in the form (8.18). To demonstrate this directly analytically is ***quite involved*** and is presented in Section 8.9.

Instead, the approach was to develop directly a set of equations of the form (8.18) for estimating ξ. This formulation leads to an estimating equation for ξ of the form discussed in the papers cited above.

We consider β fixed for now and consider estimation of ξ only; we combine the equations for ξ we now derive with the ***linear equation*** (8.11) for β at the end of the following argument, and we consider combining with ***quadratic estimating equations*** for β in the next section.

If we are interested in estimating the elements of ξ, which describe an ***entire covariance structure*** (variances and correlations, so variances and covariances), we must consider the ***variances*** of each element of a response vector and ***all pairwise associations*** among elements of a response vector. Of course, if there are ***no unknown variance parameters*** θ, we need only consider associations.

Let $v_{ijk}(\beta, \xi)$ be the (j, k) element of $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ (which is of course equal to the (k, j) element by symmetry). Then if $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ is specified correctly,

$$E\left[\{Y_{ij} - f(\boldsymbol{x}_{ij}, \beta)\}^2|\boldsymbol{x}_i\right] = v_{ijj}(\beta, \xi),$$

$$E\left[\{Y_{ij} - f(\boldsymbol{x}_{ij}, \beta)\}\{Y_{ik} - f(\boldsymbol{x}_{ik}, \beta)\}|\boldsymbol{x}_i\right] = v_{ijk}(\beta, \xi).$$

The idea is to identify $\boldsymbol{s}_i(\eta)$ in (8.18) as containing ***all quadratic and distinct cross-product terms*** of the form $\{Y_{ij} - f(\boldsymbol{x}_{ij}, \beta)\}^2$ and $\{Y_{ij} - f(\boldsymbol{x}_{ij}, \beta)\}\{Y_{ik} - f(\boldsymbol{x}_{ik}, \beta)\}$ for $j, k = 1, \ldots, n_i$, so that $\boldsymbol{m}_i(\eta)$ contains the expectations of these given above, which comprise the distinct elements of $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$.

- By ***symmetry***, for $\boldsymbol{Y}_i$ ($n_i \times 1$), there are n_i ***quadratic terms*** (one for each entry of $\boldsymbol{Y}_i$) and $n_i(n_i-1)/2$ ***distinct crossproducts*** (i.e., number of covariances), for a total of $n_i(n_i+1)/2$ distinct terms.

- Explicitly, the $n_i(n_i+1)/2$ distinct terms are the n_i ***squared deviations*** $\{Y_{i1}-f(\boldsymbol{x}_{i1},\beta)\}^2, \ldots, \{Y_{in_i}-f(\boldsymbol{x}_{in_i},\beta)\}^2$ and the $n_i(n_i-1)/2$ ***crossproduct terms***

$$\{Y_{i1}-f(\boldsymbol{x}_{i1},\beta)\}\{Y_{i2}-f(\boldsymbol{x}_{i2},\beta)\},\ \{Y_{i1}-f(\boldsymbol{x}_{i1},\beta)\}\{Y_{i3}-f(\boldsymbol{x}_{i3},\beta)\}, \ldots,$$

$$\{Y_{i,n_i-1}-f(\boldsymbol{x}_{i,n_i-1},\beta)\}\{Y_{in_i}-f(\boldsymbol{x}_{in_i},\beta)\}.$$

  Thus, $\boldsymbol{s}_i(\eta)$ in (8.18) should be of length $n_i(n_i+1)/2$, assuming that the model contains unknown variance parameters. If it does not, then only the $n_i(n_i-1)/2$ crossproduct terms are required.

- Of course, as the quadratic estimating equation (8.17) depends on the quadratic form in $\{\boldsymbol{Y}_i-\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}$, it ***also depends*** on these squared deviations and crossproducts.

To formalize, define

$$u_{ijk}(\beta)=\{Y_{ij}-f(\boldsymbol{x}_{ij},\beta)\}\{Y_{ik}-f(\boldsymbol{x}_{ik},\beta)\}. \tag{8.19}$$

Then we can collect the ***distinct*** $u_{ijk}(\beta)$ defined in (8.19) in a vector of length $n_i(n_i+1)/2$ (with unknown variance parameters θ) or $n_i(n_i-1)/2$ (with no unknown variance parameters) in some order. In the former case, suppressing the dependence of the u_{ijk} on β for brevity, let

$$\boldsymbol{u}_i(\beta)=(u_{i11},u_{i12},u_{i13},\ldots,u_{i22},u_{i23},\ldots,u_{in_i-1n_i-1},u_{in_i-1n_i-1},u_{in_in_i})^T.$$

- Here, we have used the ordering imposed by defining

$$\boldsymbol{u}_i(\beta)=\text{vech}\left[\{\boldsymbol{Y}_i-\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}\{\boldsymbol{Y}_i-\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}^T\right],$$

  where vech($\cdot$) is defined in Appendix A. If there are ***no unknown variance parameters*** in ξ, $\boldsymbol{u}_i(\beta)$ would be defined by ***deleting*** the squared components.

We can define a ***corresponding vector***

$$\boldsymbol{v}_i(\beta,\xi)=\{v_{i11}(\beta,\xi),v_{i12}(\beta,\xi),v_{i13}(\beta,\xi),\ldots,v_{i22}(\beta,\xi),v_{i23}(\beta,\xi),\ldots,v_{in_in_i}(\beta,\xi)\}^T;$$

i.e., $\boldsymbol{v}_i(\beta,\xi)=\text{vech}\{\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}$. Clearly,

$$E\{\boldsymbol{u}_i(\beta)|\boldsymbol{x}_i\}=\boldsymbol{v}_i(\beta,\xi).$$

If $\boldsymbol{u}_i(\beta)$ contains no squared components (i.e., no unknown variance parameters), then the corresponding elements of $\boldsymbol{v}_i$ would be deleted.

Thus, identifying $\boldsymbol{s}_i(\eta) = \boldsymbol{u}_i(\beta)$, we have $\boldsymbol{m}_i(\eta) = \boldsymbol{v}_i(\beta, \xi)$.

It is important to recognize in reading the literature that there are ***variations*** on the construction we describe here. For example, some authors instead base the equations on $\boldsymbol{s}_i(\eta) = \text{vech}(\boldsymbol{Y}_i \boldsymbol{Y}_i^T)$, and/or may "***stack***" things in a different order.

With $\boldsymbol{m}_i(\eta) = \boldsymbol{v}_i(\beta, \xi)$, to identify $\mathcal{D}_i(\eta)$ in (8.18), define

$$\boldsymbol{E}_i(\beta, \xi) = \partial / \partial \xi \; \boldsymbol{v}_i(\beta, \xi).$$

$\boldsymbol{E}_i(\beta, \xi)$ has $n_i(n_i+1)/2$ or $n_i(n_i-1)/2$ rows, depending on the form of $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$, and $(r+s)$ columns.

To identify $\mathcal{V}_i(\eta)$ in (8.18), we must find, suppressing dependence of $\boldsymbol{u}_i(\beta)$ and its elements on β for brevity,

$$\text{var}(\boldsymbol{u}_i | \boldsymbol{x}_i) = \boldsymbol{Z}_i(\beta, \xi),$$

say. To specify this matrix, we must be ***willing to make assumptions*** about quantities of the general form

$$\text{cov}(u_{ijk}, u_{i\ell p} | \boldsymbol{x}_i) = E(u_{ijk} u_{i\ell p} | \boldsymbol{x}_i) - E(u_{ijk} | \boldsymbol{x}_i) E(u_{i\ell p} | \boldsymbol{x}_i). \tag{8.20}$$

Clearly, this is ***quite complex***. To demonstrate, consider the particular model for $\text{var}(\boldsymbol{Y}_i | \boldsymbol{x}_i)$ such that

$$\text{var}(Y_{ij} | \boldsymbol{x}_i) = \sigma^2 g^2(\beta, \delta, \boldsymbol{x}_{ij}),$$

with correlation matrix $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$. Define

$$\epsilon_{ij} = \{ Y_{ij} - f(\boldsymbol{x}_{ij}, \beta) \} / \{ \sigma g(\beta, \delta, \boldsymbol{x}_{ij}) \}.$$

Then, of course $E(\epsilon_{ij} | \boldsymbol{x}_i) = 0$, and $\text{var}(\epsilon_{ij} | \boldsymbol{x}_i) = 1$, where expectation here and subsequently is under the parameter values β and ξ. Clearly, the elements of $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^T$ are ***correlated***, with correlation matrix equal to $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$ (assuming as we are that this matrix is correctly specified for the purposes of formulating this approach).

Under this model, for $j, k = 1, \ldots, n_i$,

$$u_{ijj} = \sigma^2 g^2(\beta, \delta, \boldsymbol{x}_{ij}) \epsilon_{ij}^2, \quad u_{ijk} = \sigma^2 g(\beta, \delta, \boldsymbol{x}_{ij}) g(\beta, \theta, \boldsymbol{x}_{ik}) \epsilon_{ij} \epsilon_{ik}.$$

Thus,

$$v_{ijj} = \sigma^2 g^2(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{x}_{ij}) E(\epsilon_{ij}^2 | \boldsymbol{x}_i) = \sigma^2 g^2(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{x}_{ij}),$$

$$v_{ijk} = \sigma^2 g(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{x}_{ij}) g(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{x}_{ik}) E(\epsilon_{ij}\epsilon_{ik} | \boldsymbol{x}_i) = \sigma^2 g(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{x}_{ij}) g(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{x}_{ik}) \text{corr}(\epsilon_{ij}, \epsilon_{ik} | \boldsymbol{x}_i),$$

where $\text{corr}(\epsilon_{ij}, \epsilon_{ik})$ is the (j, k) element of the correlation matrix $\boldsymbol{\Gamma}_i(\boldsymbol{\alpha}, \boldsymbol{x}_i)$.

In general, using the shorthand notation

$$g_{ij} = g(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{x}_{ij}),$$

we may rewrite (8.20) in terms of the ϵ_{ij} as

$$\text{cov}(u_{ijk}, u_{i\ell p} | \boldsymbol{x}_i) = \sigma^4 g_{ij} g_{ik} g_{i\ell} g_{ip} \{ E(\epsilon_{ij}\epsilon_{ik}\epsilon_{i\ell}\epsilon_{ip} | \boldsymbol{x}_i) - E(\epsilon_{ij}\epsilon_{ik} | \boldsymbol{x}_i) E(\epsilon_{i\ell}\epsilon_{ip} | \boldsymbol{x}_i) \}. \tag{8.21}$$

The representation (8.21) highlights how ***complex*** specification of $\boldsymbol{Z}_i(\boldsymbol{\beta}, \boldsymbol{\xi})$ is; in particular, some special cases of (8.21) are

$$\text{cov}(u_{ijj}, u_{ijj} | \boldsymbol{x}_i) = \text{var}(u_{ijj} | \boldsymbol{x}_i) = \sigma^4 g_{ij}^4 \text{var}(\epsilon_{ij}^2 | \boldsymbol{x}_i),$$

$$\text{cov}(u_{ijk}, u_{ijk} | \boldsymbol{x}_i) = \text{var}(u_{ijk}) = \sigma^4 g_{ij}^2 g_{ik}^2 [\, E(\epsilon_{ij}^2 \epsilon_{ik}^2 | \boldsymbol{x}_i) - \{ E(\epsilon_{ij}\epsilon_{ik} | \boldsymbol{x}_i) \}^2 \,],$$

$$\text{cov}(u_{ijj}, u_{i\ell\ell} | \boldsymbol{x}_i) = \sigma^2 g_{ij}^2 g_{i\ell}^2 \{ E(\epsilon_{ij}^2 \epsilon_{i\ell}^2 | \boldsymbol{x}_i) - 1 \},$$

$$\text{cov}(u_{ijk}, u_{ijp} | \boldsymbol{x}_i) = \sigma^4 g_{ij}^2 g_{ik} g_{ip} \{ E(\epsilon_{ij}^2 \epsilon_{ik}\epsilon_{ip} | \boldsymbol{x}_i) - E(\epsilon_{ij}\epsilon_{ik} | \boldsymbol{x}_i) E(\epsilon_{ij}\epsilon_{ip} | \boldsymbol{x}_i) \},$$

$$\text{cov}(u_{ijj}, u_{ijp} | \boldsymbol{x}_i) = \sigma^2 g_{ij}^3 g_{ip} \{ E(\epsilon_{ij}^3 \epsilon_{ip} | \boldsymbol{x}_i) - E(\epsilon_{ij}\epsilon_{ip} | \boldsymbol{x}_i) \}.$$

Of course, (8.21) represents the general case for $j \neq k \neq \ell \neq p$.

The result is that, to specify the "***covariance matrix***" $\boldsymbol{\mathcal{V}}_i(\boldsymbol{\eta}) = \boldsymbol{Z}_i(\boldsymbol{\beta}, \boldsymbol{\xi})$ in (8.18), we must be prepared to make assumptions about ***numerous higher moments*** involving the elements of $\boldsymbol{Y}_i$, or equivalently, $\boldsymbol{\epsilon}_i$, up to ***four-way associations***, i.e., $E(\epsilon_{ij}\epsilon_{ik}\epsilon_{i\ell}\epsilon_{ip} | \boldsymbol{x}_i)$.

The diligent student will verify that, if $n_i = 1$, so that $\boldsymbol{Y}_i$ is a ***scalar***, this reduces to ***quadratic estimating equation*** (7.21) for θ in the univariate case, where all we have are ***variance parameters***.

Putting aside the troublesome issue of specifying $\boldsymbol{Z}_i(\boldsymbol{\beta}, \boldsymbol{\xi})$, the preceding developments suggest the following approach. Given some assumption on $\boldsymbol{Z}_i(\boldsymbol{\beta}, \boldsymbol{\xi})$ (thus, some assumption on ***higher moments*** of the ϵ_{ij}), to estimate $\boldsymbol{\xi}$ (treating $\boldsymbol{\beta}$ fixed for now), one would solve an ***estimating equation*** of the form

$$\sum_{i=1}^{m} \boldsymbol{E}_i^T(\boldsymbol{\beta}, \boldsymbol{\xi}) \boldsymbol{Z}_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\xi}) \{ \boldsymbol{u}_i(\boldsymbol{\beta}) - \boldsymbol{v}_i(\boldsymbol{\beta}, \boldsymbol{\xi}) \} = \boldsymbol{0}. \tag{8.22}$$

- The estimating equation (8.22) is clearly ***unbiased*** , ***regardless*** of the choice of $\boldsymbol{Z}_i(\beta, \xi)$.
- The discussion at the end of Chapter 7 suggests that the ***optimal estimating equation*** for ξ of form (8.22) is that found by specifying $\boldsymbol{Z}_i(\beta, \xi)$ ***correctly*** , so that, in truth, $\text{var}\{\boldsymbol{u}_i(\beta)|\boldsymbol{x}_i\} = \boldsymbol{Z}_i(\beta, \xi)$.
- In step 1 of the iterative algorithm, at the ℓth iteration, ξ could be estimated by replacing β in (8.22) by $\widehat{\beta}^{(\ell)}$ everywhere, including in $\boldsymbol{u}_i(\beta)$, and solving in ξ.

Two issues must be resolved:

- Clearly, specification of $\boldsymbol{Z}_i(\beta, \xi)$ is ***challenging*** , and the chance one could correctly specify all the relevant moments is ***slim to none***. Thus, a ***realistic strategy*** for specifying $\boldsymbol{Z}_i(\beta, \xi)$ in practice is required.
- Both (8.22) and (8.17) depend on $\boldsymbol{Y}_i$ through the elements of $\boldsymbol{u}_i(\beta)$, the latter equation through the ***quadratic form*** in $\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta)\}$. Because the second equation was derived from the loglikelihood assuming that the distribution of $\boldsymbol{Y}_i|\boldsymbol{x}_i$ is ***normally distributed*** , intuition suggests that choosing $\boldsymbol{Z}_i(\beta, \xi)$ to be the "covariance matrix" for $\boldsymbol{u}_i(\beta)$ that would be obtained under this condition should lead to an equation (8.22) that is ***equivalent*** to (8.17).

We consider each of these issues in turn.

SPECIFICATION OF $\boldsymbol{Z}_i(\beta, \xi)$: Because correct specification is a considerable challenge, the suggestion is to make a "***working assumption***" for $\boldsymbol{Z}_i(\beta, \xi)$, similar in spirit to that made for the correlation matrix of $\boldsymbol{Y}_i|\boldsymbol{x}_i$ in the ***linear estimating equation*** for β. Popular working assumptions are

- ***Independence working assumption.*** Take $\boldsymbol{Z}_i(\beta, \xi)$ to be the covariance matrix for $\boldsymbol{u}_i|\boldsymbol{x}_i$ that would be obtained if the elements Y_{ij} of $\boldsymbol{Y}_i$ were assumed to be ***mutually independent*** across j.
- ***Gaussian working assumption.*** Take $\boldsymbol{Z}_i(\beta, \xi)$ to be the covariance matrix for $\boldsymbol{u}_i|\boldsymbol{x}_i$ that would be obtained by assuming that the distribution of $\boldsymbol{Y}_i|\boldsymbol{x}_i$ is ***normal*** with the first two moments ***correctly specified*** according to the assumed mean-covariance model (8.3). Equivalently, assume that $\epsilon_i|\boldsymbol{x}_i$ is normal with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$.

It can be shown under this condition that

$$\text{cov}(u_{ijk}, u_{i\ell p}|\boldsymbol{x}_i) = v_{ij\ell}v_{ikp} + v_{ijp}v_{ik\ell} = \sigma^4 g_{ij}g_{ik}g_{i\ell}g_{ip}\{E(\epsilon_{ij}\epsilon_{i\ell}|\boldsymbol{x}_i)E(\epsilon_{ik}\epsilon_{ip}|\boldsymbol{x}_i) + E(\epsilon_{ij}\epsilon_{ip}|\boldsymbol{x}_i)E(\epsilon_{ik}\epsilon_{i\ell}|\boldsymbol{x}_i)\}. \tag{8.23}$$

The entries of $\boldsymbol{Z}_i(\beta, \xi)$ can then be determined from the ***simplified*** relationship (8.23).

Note that, as $E(\epsilon_{ij}\epsilon_{ik}|\boldsymbol{x}_i)$ is equal to the conditional correlation between ϵ_{ij} and ϵ_{ik}, from (8.23) all of the needed entries of $\boldsymbol{Z}_i(\beta, \xi)$ depend only on the assumed correlation model $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$. Moreover, (8.23) also yields

$$\text{var}(u_{ijj}|\boldsymbol{x}_i) = 2\sigma^4 g_{ij}^4,$$

where the "2" agrees with the fourth moment properties of the normal.

Although these choices may indeed be misspecifications, the hope is that they will produce estimators "***closer to***" being "***optimal***" than simply ***ignoring*** the pattern of association among the elements of $\boldsymbol{u}_i$ altogether.

RELATIONSHIP BETWEEN (8.22) ***AND*** (8.17): The estimating equation in (8.17), derived ***directly*** from the assumed normal loglikelihood for $\boldsymbol{Y}_i|\boldsymbol{x}_i$, namely,

$$(1/2)\sum_{i=1}^{m}\Big(\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta)\}^T \boldsymbol{V}_i^{-1}(\beta, \xi, \boldsymbol{x}_i)\{\partial/\partial\xi_k\, \boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\beta, \xi, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta)\}$$
$$- \text{tr}\left[\boldsymbol{V}_i^{-1}(\beta, \xi, \boldsymbol{x}_i)\partial/\partial\xi_k\{\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)\}\right]\Big) = 0, \quad k = 1, \ldots, r+s, \tag{8.24}$$

is quadratic, as it depends on a quadratic form in $\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta)\}$. A quadratic form may of course be written as a ***linear combination*** of squared deviations and crossproducts; e.g., for square matrix $\boldsymbol{A}_i$,

$$\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta)\}^T \boldsymbol{A}_i^{-1}\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta)\} = \sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\{Y_{ij} - f(\boldsymbol{x}_{ij}, \beta)\}\{Y_{ik} - f(\boldsymbol{x}_{ik}, \beta)\}a^{ijk},$$

where a^{ijk} is the (j, k) element of $\boldsymbol{A}_i^{-1}$. The quadratic form in the summand of (8.24) is thus a ***linear combination*** of the elements of $\boldsymbol{u}_i(\beta)$.

- Of course, the summands of estimating equation in (8.22),

$$\sum_{i=1}^{m}\boldsymbol{E}_i^T(\beta, \xi)\boldsymbol{Z}_i^{-1}(\beta, \xi)\{\boldsymbol{u}_i(\beta) - \boldsymbol{v}_i(\beta, \xi)\} = \boldsymbol{0}, \tag{8.25}$$

  are ***also linear combinations*** of the elements of $\boldsymbol{u}_i(\beta)$.

- If the distribution of $\boldsymbol{Y}_i|\boldsymbol{x}_i$ ***really is*** normal, and we choose $\boldsymbol{Z}_i(\beta, \xi)$ according to the ***Gaussian working assumption***, then, as discussed at the end of Chapter 7, (8.25) must be the (asymptotically) "***optimal***" estimating equation of this (quadratic) form, as the "weight matrix" $\boldsymbol{Z}_i^{-1}(\beta, \xi)$ is correctly specified.

- The quadratic equation (8.24) is also the ***normal theory ML estimating equation*** for ξ. Thus, it is also (asymptotically) "optimal" if the data ***really are*** normally distributed.
- Both equations ***cannot*** be the "optimal" quadratic equation and be different; thus, intuition suggests that estimating equations (8.24) and (8.25) must be ***the same***.

It is possible to show this equivalence analytically; the argument is carried out in detail in Section 8.9.

STACKED EQUATIONS: Using a (quadratic) equation of the form (8.25) to estimate ξ, along with the linear equation for β, it is clear that the iterative two-step scheme given earlier solves the $p + r + s$-dimensional system of equations

$$\sum_{i=1}^{m} \begin{pmatrix} \boldsymbol{X}_i^T(\beta) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{E}_i^T(\beta,\xi) \end{pmatrix} \begin{pmatrix} \boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z}_i(\beta,\xi) \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta) \\ \boldsymbol{u}_i(\beta) - \boldsymbol{v}_i(\beta,\xi) \end{pmatrix} = \mathbf{0}. \tag{8.26}$$

Compare these equations to the univariate equations (7.25) (which can be seen to incorporate the ***univariate version*** of the "Gaussian working assumption" to yield the "$2\sigma^4 g_j^4$" term). It is straightforward to show that (verify), with $n_i = 1$ for all i and the Gaussian working assumption, (8.26) reduces to (7.25).

The equations (8.26) can be written as

$$\sum_{i=1}^{m} \mathcal{D}_i^T(\eta)\mathcal{V}_i^{-1}(\eta)\{\boldsymbol{s}_i(\eta) - \boldsymbol{m}_i(\eta)\} = \mathbf{0}, \tag{8.27}$$

where

$$\mathcal{D}_i(\eta) = \begin{pmatrix} \boldsymbol{X}_i(\beta) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{E}_i(\beta,\xi) \end{pmatrix} \quad \{n_i + n_i(n_i+1)/2 \times p + r + s\},$$

$$\mathcal{V}_i(\eta) = \begin{pmatrix} \boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z}_i(\beta,\xi) \end{pmatrix} \quad \{n_i + n_i(n_i+1)/2 \times n_i + n_i(n_i+1)/2\},$$

$$\boldsymbol{s}_i(\eta) = \begin{pmatrix} \boldsymbol{Y}_i \\ \boldsymbol{u}_i(\beta) \end{pmatrix}, \quad \boldsymbol{m}_i(\eta) = \begin{pmatrix} \boldsymbol{f}_i(\boldsymbol{x}_i,\beta) \\ \boldsymbol{v}_i(\beta,\xi) \end{pmatrix} \quad \{n_i + n_i(n_i+1)/2 \times 1\}.$$

With the equations written in the form (8.27), it is clear that the two-step algorithm for solving them is natural, and that each of steps 1 and 2 can be implemented in principle via a ***Gauss-Newton*** updating scheme, which is common in practice. In particular

- Step 1, with β held fixed [last $r + s$ rows of (8.27)], can be implemented by a Gauss-Newton algorithm iterated to convergence to obtain the next iterate of ξ.
- Step 2, with ξ held fixed [first p rows of (8.27)], can be implemented by a Gauss-Newton algorithm iterated to convergence to obtain the next iterate of β.

8.4 Quadratic estimating equations

As in the univariate case, it is natural in the longitudinal context to consider the potential ***increase efficiency*** for estimation of β by extracting information about β from the covariance matrix $\boldsymbol{V}_i(\beta, \xi)$. As in that case, we can deduce a ***quadratic estimating equation*** for β by appealing to the ***normal loglikelihood***.

Differentiating the normal loglikelihood (8.14) with respect to β, it is straightforward using the same matrix differentiation operations as before to obtain the resulting equation

$$\begin{aligned}\sum_{i=1}^{m} \Big\{ & \boldsymbol{X}_i^T(\beta)\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\} \\ & + \Big(\Big(\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}^T \boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\partial/\partial\beta_\ell \boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\} \\ & - \text{tr}\left[\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\partial/\partial\beta_\ell\{\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\right]\Big)\Big) \Big\} = \boldsymbol{0} \quad (p\times 1), \end{aligned} \tag{8.28}$$

where the double parentheses indicate p terms of the form inside them stacked for $\ell = 1, \ldots, p$. Of course, underlying estimating equation (8.28) is the assumption of normality. This equation would be solved ***jointly*** with the quadratic equation (8.24) for ξ to obtain the MLEs for β and ξ under the assumption of normality.

Alternatively, following the ***same reasoning*** as in the previous section for estimation of ξ, we can formulate a general quadratic equation. Defining $\boldsymbol{v}_i(\beta,\xi)$ and $\boldsymbol{u}_i(\beta)$ as before, and letting

$$\boldsymbol{B}_i(\beta,\xi) = \partial/\partial\beta\, \boldsymbol{v}_i(\beta,\xi) \quad \{n_i(n_i+1)/2 \times p\},$$

so that $\boldsymbol{B}_i(\beta,\xi)$ is the "gradient matrix" of $E\{\boldsymbol{u}_i(\beta)|\boldsymbol{x}_i\} = \boldsymbol{v}_i(\beta,\xi)$, the obvious joint estimating equations for β and ξ are

$$\sum_{i=1}^{m} \begin{pmatrix} \boldsymbol{X}_i^T(\beta) & \boldsymbol{B}_i^T(\beta,\xi) \\ \boldsymbol{0} & \boldsymbol{E}_i^T(\beta,\xi) \end{pmatrix} \begin{pmatrix} \boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}_i(\beta,\xi) \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta) \\ \boldsymbol{u}_i(\beta) - \boldsymbol{v}_i(\beta,\xi) \end{pmatrix} = \boldsymbol{0}. \tag{8.29}$$

- Because of the presence of the "gradient matrix" $\boldsymbol{B}_i(\beta,\xi)$ in the first p rows of (8.29), this leads to an estimating equation for β that is ***quadratic***.
- If $\boldsymbol{Z}_i(\beta,\xi)$ were chosen according to the ***Gaussian working assumption***, then, by the same reasoning as in the previous section, the equation corresponding to the first p rows of (8.29) should be the "optimal" such equation if ***normality really holds***, and, thus, intuitively, should be ***identical*** to the normal theory ML equation (8.28). This can be shown by arguments analogous to those in Section 8.9.

In (8.28), the ***off-block-diagonal elements*** of the "covariance matrix"

$$\begin{pmatrix} \boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}_i(\beta, \xi) \end{pmatrix}$$

are all equal to zero. ***Under normality*** of $\boldsymbol{Y}_i|\boldsymbol{x}_i$, it is indeed the case that

$$\text{cov}\{\boldsymbol{Y}_i, \boldsymbol{u}_i(\beta)|\boldsymbol{x}_i\} = \boldsymbol{0}.,$$

which is consistent with the view that (8.29) with the ***Gaussian working assumption*** is the optimal joint equation if ***normality really holds***, as in this case $\mathcal{V}_i(\beta, \xi)$ is exactly the covariance matrix of the "response" $\boldsymbol{s}_i(\eta) = \{\boldsymbol{Y}_i^T, \boldsymbol{u}_i^T(\beta)\}^T$.

By analogy to the univariate case discussed at the end of Chapter 7, if the distribution of $\boldsymbol{Y}_i|\boldsymbol{x}_i$ is not believed to be normal, we could ***in principle*** arrive at a more general equation by specifying the covariance matrix of the "response" $\boldsymbol{s}_i(\eta) = \{\boldsymbol{Y}_i^T, \boldsymbol{u}_i^T(\beta)\}^T$ to embody corresponding assumptions about the moments of $\boldsymbol{s}_i(\eta)$. If we specify

$$\text{var}\{\boldsymbol{u}_i(\beta)|\boldsymbol{x}_i\} = \boldsymbol{Z}_i(\beta, \xi) \quad \text{and} \quad \text{cov}\{\boldsymbol{Y}_i, \boldsymbol{u}_i(\beta)|\boldsymbol{x}_i\} = \boldsymbol{C}_i(\beta, \xi),$$

say, we obtain

$$\sum_{i=1}^{m} \begin{pmatrix} \boldsymbol{X}_i^T(\beta) & \boldsymbol{B}_i^T(\beta, \xi) \\ \boldsymbol{0} & \boldsymbol{E}_i^T(\beta, \xi) \end{pmatrix} \begin{pmatrix} \boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i) & \boldsymbol{C}_i(\beta, \xi) \\ \boldsymbol{C}_i^T(\beta, \xi) & \boldsymbol{Z}_i(\beta, \xi) \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta) \\ \boldsymbol{u}_i(\beta) - \boldsymbol{v}_i(\beta, \xi) \end{pmatrix} = \boldsymbol{0}. \tag{8.30}$$

- In the ***unlikely*** event that the assumptions for $\boldsymbol{Z}_i(\beta, \xi)$ and $\boldsymbol{C}_i(\beta, \xi)$ are ***correct***, we would expect to have constructed the "***optimal***" such quadratic equation.

- This entails making the moment assumptions on $\text{var}\{\boldsymbol{u}_i(\beta)|\boldsymbol{x}_i\}$ in $\boldsymbol{Z}_i(\beta, \boldsymbol{x}_i)$, but ***also*** for $\boldsymbol{C}_i(\beta, \xi)$, which involves specifying quantities of the form

  $$\text{cov}(Y_{ij}, u_{ik\ell}|\boldsymbol{x}_i) = \sigma^3 g_{ij} g_{ik} g_{i\ell} E(\epsilon_{ij}\epsilon_{ik}\epsilon_{i\ell}|\boldsymbol{x}_i).$$

  That is, we need to be willing to specify ***not only*** the ***skewness*** $E(\epsilon_{ij}^3|\boldsymbol{x}_i)$ as in the univariate case, but ***also*** the "three-way associations."

- The chance that we can specify the matrices $\boldsymbol{Z}_i(\beta, \xi)$ and $\boldsymbol{C}_i(\beta, \xi)$ ***completely correctly*** in practice is ***slim***. Indeed, specifying $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ correctly in itself is ***difficult enough***.

Putting this issue aside for the moment, clearly, solution of estimating equations of the general form (8.30) can be ***implemented*** by a Gauss-Newton updating scheme, redefining $\mathcal{D}_i(\eta)$ and $\mathcal{V}_i(\eta)$ in (8.27) in the obvious way as

$$\mathcal{V}_i(\eta) = \begin{pmatrix} \boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i) & \boldsymbol{C}_i(\beta, \xi) \\ \boldsymbol{C}_i^T(\beta, \xi) & \boldsymbol{Z}_i(\beta, \xi) \end{pmatrix}. \tag{8.31}$$

This must be carried out via a ***single*** such updating algorithm of dimension $p+r+s$, as it is no longer possible to ***decouple*** the equations for β (first p rows) and ξ (last $r+s$ rows) and to use separate, lower-dimensional updating. Thus, solution of (8.30) by this approach is ***more complex*** numerically.

WORKING ASSUMPTIONS: As above, in practice, one would make a "***working assumption***" about the entire matrix $\mathcal{V}_i(\eta)$ in (8.31). This of course involves an assumption on $\boldsymbol{Z}_i(\beta, \xi)$ as before, along with one on $\boldsymbol{C}_i(\beta, \xi)$. Popular working assumptions for $\mathcal{V}_i(\eta)$ are as follows, analogous to the previous discussion.

- ***Independence working assumption.*** Taking the elements of $\boldsymbol{Y}_i$ to be ***mutually independent*** leads to the choice for $\boldsymbol{Z}_i(\beta, \xi)$ discussed previously and $\boldsymbol{C}_i(\beta, \xi) = \boldsymbol{0}$.
- ***Gaussian working assumption.*** Taking of $\boldsymbol{Y}_i|\boldsymbol{x}_i$ to be normal leads to the choice for $\boldsymbol{Z}_i(\beta, \xi)$ given in (8.23) and $\boldsymbol{C}_i(\beta, \xi) = \boldsymbol{0}$.

TERMINOLOGY: Equations like those in (8.26) and (8.30) have been referred to as ***generalized estimating equations*** of specific types:

- Equations of the form (8.26), which involve solving a ***linear*** estimating equation for β jointly with a quadratic one for ξ, have been called ***GEE-1***.
- Equations of the form (8.30), which involve solving a ***quadratic*** estimating equation for β jointly with a quadratic one for ξ, have been called ***GEE-2***.
- This terminology was evidently coined in a paper by Liang, Zeger, and Qaqish (1992).
- The unqualified acronym "GEE" is used popularly both to refer to the general approach of specifying estimating equations for mean-covariance models of the form (8.3) ***and*** to the particular case where the ***linear*** equation for β in (8.11) is solved with the elements of ξ estimated by simple ***moment-based*** estimators.

- "GEE" is often also taken to imply that "***working assumptions***" are involved, including those on the correlation matrix of $\boldsymbol{Y}_i|\boldsymbol{x}_i$, that are likely to be ***incorrect*** , so that the ***robust sandwich covariance matrix*** , discussed in the next section, should be used by default when approximating the sampling distribution of the estimator for β.

REMARKS: The ***same*** issues involving ***trade-offs*** between linear and quadratic estimating equations for β that we discussed for the univariate case extend to the longitudinal, multivariate setting.

- The linear equation is clearly unbiased ***even if*** the covariance model $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ is ***misspecified***. Such misspecification is ***more likely*** in the multivariate case, as the analyst must model ***not only*** variance but also correlation structure. The latter is more difficult, so the focus is on possible ***incorrect modeling*** of the correlation matrix $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$.
- The quadratic equation for β will be unbiased ***as long as*** the covariance model $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ is correctly specified. Thus, one must be confident that this is the case to reap the benefits of ***possible increased efficiency***. Even in this case, the ***optimal*** quadratic equation ***also*** requires correct specification of $\boldsymbol{Z}_i(\beta, \xi)$ and $\boldsymbol{C}_i(\beta, \xi)$. This is almost certainly ***not the case*** in practice, so even though the estimating equation is unbiased with $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ correct, whether or not it leads to a more efficient estimator for β is ***no longer clear*** , analogous to the results at the end of Chapter 7.
- The quadratic equations will ***not be unbiased*** if $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ is not correctly specified so could result in ***inconsistent*** estimation of β. Thus, in practice, it is generally agreed that "GEE-1" estimation based on (8.26) is the "safer" choice for routine use in fitting population-averaged models. Accordingly, we discuss large sample inference ***only*** for the estimator for β obtained by solving the ***linear estimating equation***.

8.5 Large sample inference

As we did for the ***linear population-averaged models*** discussed in Chapter 5, we derive an approximate sampling distribution for $\widehat{\beta}$ solving (8.11), ***linear estimating equation***

$$\sum_{i=1}^{m} \boldsymbol{X}_i^T(\beta)\boldsymbol{V}_i^{-1}(\beta, \xi, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta)\} = \boldsymbol{0} \tag{8.32}$$

with an estimator $\widehat{\xi}$ ***substituted***.

The estimator $\widehat{\xi}$ can be a ***moment-based*** estimator or that found by solving the ***quadratic estimating equation*** for ξ under some ***working assumption***, all of which satisfy the conditions below.

As in those arguments, we adopt a large sample framework in which the number of individuals $m \to \infty$ with the n_i treated as ***fixed***. As we noted in Section 5.6, it is ***not appropriate*** to regard the n_i as fixed when data vectors of intended length n are of different lengths as the result of some ***missingness mechanism***. We discuss the implications of ***missing data*** for inference using GEEs in Section 8.7.

COVARIANCE MODEL POSSIBLY INCORRECTLY SPECIFIED: As in Section 5.5 for a linear PA model with covariance matrix not depending on β and in the univariate setting in Section 7.4, consider the general situation in which, although the mean model $\boldsymbol{f}_i(\boldsymbol{x}_i, \beta)$ is ***correctly specified*** , so that the ***true mean*** is

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{f}_i(\boldsymbol{x}_i, \beta_0),$$

where β_0 is the true value of β, the covariance model $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ ***need not be*** correctly specified. Thus, analogous to these previous arguments, letting the ***true covariance matrix*** be

$$\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_{i0},$$

there is ***not necessarily*** a value ξ_0 such that $\boldsymbol{V}_i(\beta_0, \xi_0, \boldsymbol{x}_i)$.

Assume that the estimator $\widehat{\xi}$ is such that $\widehat{\xi} \xrightarrow{p} \xi^*$ for some ξ^* and $m^{1/2}(\widehat{\xi} - \xi^*) = O_p(1)$, and let

$$\boldsymbol{V}_i^* = \boldsymbol{V}_i(\beta_0, \xi^*, \boldsymbol{x}_i).$$

As in Section 5.5, even with the ***covariance model misspecified*** , (8.32) is still an ***unbiased estimating equation*** , as clearly

$$E\Big[\boldsymbol{X}_i^T(\beta_0)\boldsymbol{V}_i^{-1}(\beta_0, \xi^*, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta_0)\}|\boldsymbol{x}_i\Big] = \boldsymbol{0}.$$

The argument we now present is a ***generalization*** of those in Sections 5.5 and 7.4. Expanding (8.32) in a Taylor series in $(\widehat{\beta}^T, \widehat{\xi}^T)^T$ about $(\beta_0^T, \xi^{*T})^T$, we have

$$\begin{aligned} \boldsymbol{0} &= m^{-1/2}\sum_{i=1}^{m}\boldsymbol{X}_i^T(\widehat{\beta})\boldsymbol{V}_i^{-1}(\widehat{\beta}, \widehat{\xi}, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \widehat{\beta})\} \\ &\approx \boldsymbol{C}_m^* + (\boldsymbol{A}_{m1}^* + \boldsymbol{A}_{m2}^* + \boldsymbol{A}_{m3}^*)\, m^{1/2}(\widehat{\beta} - \beta_0) + \boldsymbol{E}_m^*\, m^{1/2}(\widehat{\xi} - \xi^*). \end{aligned} \tag{8.33}$$

Let $\boldsymbol{X}_i = \boldsymbol{X}_i(\beta_0)$, and, as in (8.12), $\boldsymbol{X} = \boldsymbol{X}(\beta_0)$. Also define

$$\boldsymbol{V}_0 = \text{block diag}(\boldsymbol{V}_{01}, \ldots, \boldsymbol{V}_{0m}), \qquad \boldsymbol{V}^* = \text{block diag}(\boldsymbol{V}_1^*, \ldots, \boldsymbol{V}_m^*).$$

It is straightforward that

$$\boldsymbol{C}_m = m^{-1/2} \sum_{i=1}^{m} \boldsymbol{X}_i^T(\beta_0) \boldsymbol{V}_i^{*-1} \{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta_0)\} \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{B}^*),$$

$$\boldsymbol{B}^* = \lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{*-1} \boldsymbol{V}_{0i} \boldsymbol{V}^{*-1} \boldsymbol{X}_i = \lim_{m\to\infty} m^{-1} \boldsymbol{X}^T \boldsymbol{V}^{*-1} \boldsymbol{V}_0 \boldsymbol{V}^{*-1} \boldsymbol{X},$$

$$\boldsymbol{A}_{m2}^* = -m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T(\beta_0) \boldsymbol{V}_i^{*-1} \boldsymbol{X}_i(\beta_0) \xrightarrow{p} \boldsymbol{A}^* = \lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{*-1} \boldsymbol{X}_i = \lim_{m\to\infty} m^{-1} \boldsymbol{X}^T \boldsymbol{V}^{*-1} \boldsymbol{X},$$

$$\boldsymbol{A}_{m1}^* = m^{-1} \sum_{i=1}^{m} \{\partial/\partial\beta \, \boldsymbol{X}_i^T(\beta_0)\} \boldsymbol{V}_i^{*-1} \{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta_0)\} \xrightarrow{p} \boldsymbol{0},$$

$$\boldsymbol{A}_{m3}^* = m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T(\beta_0) \{\partial/\partial\beta \boldsymbol{V}_i^{-1}(\beta_0, \xi^*, \boldsymbol{x}_i)\} \boldsymbol{V}_i^{*-1} \{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta_0)\} \xrightarrow{p} \boldsymbol{0},$$

$$\boldsymbol{E}_m^* = m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T(\beta_0) \{\partial/\partial\xi \boldsymbol{V}_i^{-1}(\beta_0, \xi^*, \boldsymbol{x}_i)\} \{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta_0)\} \xrightarrow{p} \boldsymbol{0}.$$

Combining, we obtain the (***not surprising***) result

$$m^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^{*-1} \boldsymbol{B}^* \boldsymbol{A}^{*-1}), \tag{8.34}$$

which is ***identical*** in form to the result (5.74) in the linear PA model case, but with $\boldsymbol{X}$ and $\boldsymbol{V}^{*-1}$ depending on β_0.

- From (8.33), because $\boldsymbol{A}_{m3}^*$ and $\boldsymbol{E}_m^* \xrightarrow{p} \boldsymbol{0}$, there is ***no effect*** of estimating ξ in the incorrect model $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$, ***nor*** is there an effect of the fact that this model ***depends on*** β, which must be estimated to form "weights."

- As before, when the model $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ is ***correctly specified***, $\xi^* = \xi_0$, the value such that $\boldsymbol{V}_{0i} = \boldsymbol{V}_i(\beta_0, \xi_0, \boldsymbol{x}_i)$, and $\boldsymbol{V}_i^* = \boldsymbol{V}_{0i}$, so that (8.34) reduces to

$$m^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^{-1}), \quad \boldsymbol{A} = \lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_{0i}^{-1} \boldsymbol{X}_i = \lim_{m\to\infty} m^{-1} \boldsymbol{X}^T \boldsymbol{V}_0^{-1} \boldsymbol{X}. \tag{8.35}$$

OPTIMAL LINEAR ESTIMATING EQUATION: Analogous to (5.75) and (5.76), (8.35) and (8.34) yield the approximate sampling distributions

$$\widehat{\beta}_C \dot{\sim} \mathcal{N}\{\beta_0, (\boldsymbol{X}^T \boldsymbol{V}_0^{-1} \boldsymbol{X})^{-1}\} \tag{8.36}$$

when $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ is ***correctly specified***, where *C* indicates "correct;" and, when $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ is ***incorrectly specified***,

$$\widehat{\beta}_{IC} \dot{\sim} \mathcal{N}\{\beta_0, (\boldsymbol{X}^T \boldsymbol{V}^{*-1} \boldsymbol{X})^{-1} (\boldsymbol{X}^T \boldsymbol{V}^{*-1} \boldsymbol{V}_0 \boldsymbol{V}^{*-1} \boldsymbol{X}) (\boldsymbol{X}^T \boldsymbol{V}^{*-1} \boldsymbol{X})^{-1}\}, \tag{8.37}$$

where *IC* indicates "incorrect."

- The comparison of the approximate covariance matrices in (8.36) and (8.37) is ***identical*** to that in Chapter 5. Thus, we conclude immediately that, for "large" m, the components of $\widehat{\boldsymbol{\beta}}_{IC}$ are ***inefficient relative to*** the corresponding components of $\widehat{\boldsymbol{\beta}}_{C}$.

- It follows that using a ***correct covariance model*** leads to the ***optimal linear estimating equation*** of this type, extending the result in Chapter 5 to ***nonlinear models*** and covariance models that ***depend on*** $\boldsymbol{\beta}$.

- In fact, it is possible to obtain an ***even more general*** result. Consider the linear estimating equation of the form

$$\sum_{i=1}^{m} \boldsymbol{\mathcal{A}}_i^T(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \boldsymbol{\beta}) = \boldsymbol{0}, \tag{8.38}$$

  where $\boldsymbol{\mathcal{A}}_i(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{x}_i)$, $i = 1, \ldots, m$, are arbitrary $(n_i \times p)$ matrices depending on $\boldsymbol{\beta}$, some additional parameter $\boldsymbol{\gamma}$, and possibly $\boldsymbol{x}_i$. Clearly, (8.38) is an ***unbiased estimating equation*** for $\boldsymbol{\beta}$. Let $\widehat{\boldsymbol{\beta}}_G \xrightarrow{p} \boldsymbol{\beta}_0$ be the solution to (8.38), where $\boldsymbol{\gamma}$ has been replaced by some $\widehat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}^*$ that is bounded in probability. It is straightforward, expanding (8.38) evaluated at $(\widehat{\boldsymbol{\beta}}_G^T, \widehat{\boldsymbol{\gamma}}^T)^T$ about $(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}^{*T})^T$ and letting $\boldsymbol{\mathcal{A}}_i = \boldsymbol{\mathcal{A}}_i(\boldsymbol{\beta}_0, \boldsymbol{\gamma}^*, \boldsymbol{x}_i)$ and $\boldsymbol{\mathcal{A}} = (\boldsymbol{\mathcal{A}}_1^T, \ldots, \boldsymbol{\mathcal{A}}_m^T)^T$ $(N \times p)$, to deduce that

$$\widehat{\boldsymbol{\beta}}_G \dot{\sim} \mathcal{N}\left\{\boldsymbol{\beta}_0, (\boldsymbol{\mathcal{A}}^T\boldsymbol{X})^{-1}(\boldsymbol{\mathcal{A}}^T\boldsymbol{V}_0\boldsymbol{\mathcal{A}})(\boldsymbol{X}^T\boldsymbol{\mathcal{A}})^{-1}\right\}. \tag{8.39}$$

  By an argument similar to that in Section 5.5, it can be shown that the matrix difference

$$(\boldsymbol{\mathcal{A}}^T\boldsymbol{X})^{-1}(\boldsymbol{\mathcal{A}}^T\boldsymbol{V}_0\boldsymbol{\mathcal{A}})(\boldsymbol{X}^T\boldsymbol{\mathcal{A}})^{-1} - (\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X})^{-1}$$

  is ***nonnegative definite*** (try it).

- It follows from (8.36), (8.39), and this result that, among ***all linear estimating equations*** for $\boldsymbol{\beta}$ of ***arbitrary form*** (8.38), the equation (8.11) with covariance model ***correctly specified*** is ***optimal***. This is an ***even more general*** result than that above.

- This result can be derived formally from a ***geometric perspective*** by appealing to ***semiparametric theory*** ; the gory details are presented in Chapter 4 of Tsiatis (2006).

- In fact, this result suggests that general estimating equations of the form (8.27), namely,

$$\sum_{i=1}^{m} \boldsymbol{\mathcal{D}}_i^T(\boldsymbol{\eta})\boldsymbol{\mathcal{V}}_i^{-1}(\boldsymbol{\eta})\{\boldsymbol{s}_i(\boldsymbol{\eta}) - \boldsymbol{m}_i(\boldsymbol{\eta})\} = \boldsymbol{0},$$

  with the "covariance matrix" $\boldsymbol{\mathcal{V}}_i^{-1}(\boldsymbol{\eta})$ ***correctly specified*** are ***optimal*** among all equations linear in $\{\boldsymbol{s}_i(\boldsymbol{\eta}) - \boldsymbol{m}_i(\boldsymbol{\eta})\}$.

ROBUST COVARIANCE MATRIX: As we have discussed, because the covariance model $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ and in particular the correlation model $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$ are likely to be ***misspecified***, the latter being a "***working model***," it is standard to use (8.37) as the basis for the approximate sampling distribution for the estimator $\widehat{\beta}$.

In particular, analogous to the argument in Section 5.5,

$$\widehat{\beta} \dot{\sim} \mathcal{N}(\beta_0, \widehat{\boldsymbol{\Sigma}}_R), \qquad \widehat{\boldsymbol{\Sigma}}_R = \widehat{\boldsymbol{A}}_m^{*-1} \widehat{\boldsymbol{B}}_m^* \widehat{\boldsymbol{A}}_m^{*-1} \tag{8.40}$$

$$\widehat{\boldsymbol{A}}_m^* = \sum_{i=1}^{m} \boldsymbol{X}_i^T(\widehat{\beta}) \boldsymbol{V}^{-1}(\widehat{\beta}, \widehat{\xi}, \boldsymbol{x}_i) \boldsymbol{X}_i(\widehat{\beta}),$$

$$\widehat{\boldsymbol{B}}_m^* = \sum_{i=1}^{m} \boldsymbol{X}_i^T(\widehat{\beta}) \boldsymbol{V}^{-1}(\widehat{\beta}, \widehat{\xi}, \boldsymbol{x}_i) \{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \widehat{\beta}\} \{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \widehat{\beta}\}^T \boldsymbol{V}^{-1}(\widehat{\beta}, \widehat{\xi}, \boldsymbol{x}_i) \boldsymbol{X}_i(\widehat{\beta}).$$

In (8.40), $\widehat{\boldsymbol{\Sigma}}_R$ is the ***robust sandwich*** or ***empirical*** covariance matrix, and it is straightforward to demonstrate that $m^{-1}\widehat{\boldsymbol{\Sigma}}_R$ is a consistent estimator for the true sampling covariance matrix in (8.34).

Under the assumption that the covariance model is ***correctly specified***, one would instead use (8.36) as the basis for the approximate sampling distribution of $\widehat{\beta}$, namely,

$$\widehat{\beta} \dot{\sim} \mathcal{N}(\beta_0, \widehat{\boldsymbol{\Sigma}}_M), \qquad \widehat{\boldsymbol{\Sigma}}_M = \left\{ \sum_{i=1}^{m} \boldsymbol{X}_i^T(\widehat{\beta}) \boldsymbol{V}^{-1}(\widehat{\beta}, \widehat{\xi}, \boldsymbol{x}_i) \boldsymbol{X}_i(\widehat{\beta}) \right\}^{-1} = \widehat{\boldsymbol{A}}_m^{*-1}, \tag{8.41}$$

where $\widehat{\boldsymbol{\Sigma}}_M$ is the so-called ***model-based*** covariance matrix.

- Not surprisingly, software for solving GEEs typically uses the robust covariance matrix in (8.40) ***by default***, and the user must request explicitly the model-based analysis.

8.6 Modeling issues

As the preceding sections demonstrate, the mechanics of specifying and fitting a general population-averaged mean-covariance model of the form (8.3) and carrying out approximate large-sample inference for β defining the assumed mean response model are ***messy but straightforward***. ***However***, there are ***more abstract issues*** that can render inferences and practical interpretation suspect. In this and the next section, we discuss these in some detail.

TIME-DEPENDENT AMONG-INDIVIDUAL COVARIATES: As we indicated in Section 8.2, specification of a sensible mean model when some covariates change over time can involve critical ***conceptual challenges***. We now take a more formal look at this issue.

Recall from our initial discussion of the basic mean model that, ***ordinarily*** , the model $\boldsymbol{f}_i(\boldsymbol{x}_i, \beta)$ for $E(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ satisfies

$$\boldsymbol{f}_i(\boldsymbol{x}_i, \beta) = \begin{pmatrix} f(\boldsymbol{x}_{i1}, \beta) \\ \vdots \\ f(\boldsymbol{x}_{in_i}, \beta) \end{pmatrix} \quad (n_i \times 1). \tag{8.42}$$

As in (8.8), (8.42) ***implicitly incorporates*** a rather ***strong assumption*** , namely, that

$$E(Y_{ij}|\boldsymbol{x}_i) = E(Y_{ij}|\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}) = E(Y_{ij}|\boldsymbol{x}_{ij}). \tag{8.43}$$

We noted earlier that, when the $\boldsymbol{x}_{ij}$ are ***time-independent*** , or when they are time-dependent but ***fixed by design*** or depend on j only through t_{ij}, there is no conceptual difficulty with (8.43) because the values of the $\boldsymbol{x}_{ij}$ are determined in a way that is ***unrelated*** to the longitudinal response. In contrast, if the $\boldsymbol{x}_{ij}$ are ***observed*** and not under the control of the investigator, there is the possibility that their values are ***impacted*** by those of the longitudinal response in ways that can ***distort*** the relationship between mean response and covariates and lead to difficulties in ***interpretation***.

This can be formalized as follows.

CONVENTION: For this discussion, assume that there are n intended observation times t_j, $j = 1, \ldots, n$, for each of which $\boldsymbol{x}_{ij}$ comprises ***time-independent*** covariates, such as gender, age at study entry, and so on, that ***do not vary*** with time, along with other ***time-dependent*** covariates that ***do vary*** over the observation times, such as dose in a designed experiment or smoking status in the Six Cities study. We do not make this explicit in the notation; however, it is worth noting that it is implicit in the following developments that the presence of the time-independent covariates in each of $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in}$ means that all expressions are ***conditional*** on the time-independent covariates.

Recall from Chapter 2 that we conceptualize that each individual i has an associated ***stochastic response process*** $\mathcal{Y}_i(t)$ in ***continuous time*** , where we suppress dependence on within-individual covariates $\boldsymbol{u}_i$ for brevity. With time-dependent covariates, we can also conceptualize an individual ***covariate process*** $\boldsymbol{x}_i(t)$, say.

As in previous chapters, we assume that time-dependent covariates are observed ***without measurement error*** , so that in principle it is possible to observe $\boldsymbol{x}_i(t)$ at any t ***without error***.

For the following developments, we adopt the ***convention*** that, for observation time j, Y_{ij} is the response ascertained at t_j, and we let $\boldsymbol{x}_{ij}$ involve values of $\boldsymbol{x}_i(t)$ at $t < t_j$, so up to a time ***immediately prior to*** t_j. That, is we adopt the convention that the ***temporal ordering*** of the data is

$$\boldsymbol{x}_{i1}, Y_{i1}, \boldsymbol{x}_{i2}, Y_{i2}, \ldots, \boldsymbol{x}_{i,n-1}, Y_{i,n-1}, \boldsymbol{x}_{in}, Y_{in}. \tag{8.44}$$

- In the dosing study above, where the doses d_j are ***fixed by design***, this convention coincides with the expectation that the response at t_j is ascertained ***after*** the dose d_j is administered.

- In the Six Cities study, where $\boldsymbol{x}_{ij}$ includes mother i's smoking status s_{ij} at age t_j, smoking status s_{ij} is naturally regarded as being ***already established*** at age t_j and thus ***preceding*** the child's wheezing response at t_j.

- In studies where t_1 corresponds to ***baseline***, (8.44) implies that $\boldsymbol{x}_{i1}$ comprises covariates whose values are ascertained ***immediately prior*** to the first measure of the response, Y_{i1}, including the ***randomized treatment***. Previously, in a ***randomized study***, we have used the index $j = 1$ to refer to baseline and taken Y_{i1} to be the response ascertained ***prior to initiation treatment***.

  For purposes of the discussion here, we adhere to the temporal ordering (8.44), so that, with randomized treatment included in $\boldsymbol{x}_{i1}$, Y_{i1} represents the first measure of the response ***after*** initiation of treatment. If a value of the response is ascertained ***prior to treatment***, we take this to be at the same time $\boldsymbol{x}_{i1}$ is recorded and can include this in $\boldsymbol{x}_{i1}$ in the developments below.

  In an ***observational longitudinal study*** such as the Six Cities study, the temporal ordering (8.44) is natural.

  We make further remarks on handling of such ***baseline responses*** in Section 8.8.

With these conventions, we can formalize the issues raised above.

EXOGENOUS COVARIATE: A covariate process is said to be ***exogenous*** with respect to a response/outcome process if, given all previous values of the covariate and response, the covariate at time j is ***independent*** of all preceding responses. Formally, in our context and obvious notation, for all individuals i and $j = 2, \ldots, n$, an exogenous covariate process satisfies

$$p(\boldsymbol{x}_{ij} | y_{i1}, \ldots, y_{i,j-1}, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,j-1}) = p(\boldsymbol{x}_{ij} | \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,j-1}). \tag{8.45}$$

- Such covariates have also been referred to as ***external*** in the survival analysis literature.

- Practically speaking, (8.45) states that the values taken on by the covariate at time j depend only on previous covariate values and ***not*** on values taken on by the response prior to j.

- Clearly, $\boldsymbol{x}_{ij} = (t_j, d_j)$ in the fixed dose design study is ***exogenous***, as the doses are fixed in advance so are determined ***completely independently*** of responses they might elicit.

- In the Six Cities study, the covariate $\boldsymbol{x}_{ij} = (t_{ij}, c_i, s_{ij})^T$ is obviously ***not exogenous***. As discussed earlier, if a mother who has past smoking history embodied in $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,j-1}$ alters her future smoking behavior as a result of observing the current wheezing status of her child, then $\boldsymbol{x}_{ij}$ cannot be independent of $Y_{i,j-1}$. More generally, she might decide to alter her future smoking as a result of knowing her past smoking history and observing part or all of the ***entire past trajectory*** of wheezing of her child, so that $\boldsymbol{x}_{ij}$ is not independent of $Y_{i1}, \ldots, Y_{i,j-1}$.

- A covariate process that is not exogenous is referred to as ***endogenous*** (or ***internal***).

- In some settings, ***care*** must be taken in evaluating whether or not a covariate process is exogenous. For example, in studies of ***environmental health***, Y_{ij} might be some ***health outcome*** for study participant i, and $\boldsymbol{x}_{ij}$ might include ***air pollution levels*** recorded at a monitoring site near to i's residence when the health outcome is measured (and thus assumed to be in place ***prior*** to the observed health outcome). Here, future pollution levels are likely related to past pollution levels but are not impacted by, and are thus ***clearly independent*** of, the previous health outcomes of study participants.

  However, suppose ***instead*** that $\boldsymbol{x}_{ij}$ includes a measure of i's ***personal exposure*** to air pollution, e.g., a summary measure based on the level at the monitoring station and i's time spent outdoors in the past month. If i decides to ***limit*** his/her future personal exposure by spending less time outdoors because of his/her current and past trajectory of health outcomes, then the covariate process is ***not exogenous***.

- In fact, regarding ***time*** itself as a ***covariate***, it should be clear that observation times in a study that are ***predetermined in advance***, such as the visit times in the epileptic seizure study, are ***exogenous***. ***However*** consider an ***observational study***, such as one based on electronic health records. If individuals have different observation times t_{ij} at which responses and other information is recorded, it is plausible that an individual's future ***realized observation times*** might be ***related*** to his/her past responses and observation times, as s/he might require ***more frequent visits*** to his/her healthcare provider if s/he has poorer responses.

- From (8.45), a strategy to check the validity of the ***assumption of exogeneity*** in practice is to develop ***regression models*** for $j = 2, \ldots, n$ for $\boldsymbol{x}_{ij}$ as a function of $Y_{i1}, \ldots, Y_{i,j-1}$ and $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,j-1}$. If, given $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,j-1}$, there is no evidence of dependence on $Y_{i1}, \ldots, Y_{i,j-1}$, this would be taken as support for the contention that the covariate process is exogenous.

KEY RESULT: If a covariate process is exogenous, it can be shown that

$$E(Y_{ij}|\boldsymbol{x}_i) = E(Y_{ij}|\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in}) = E(Y_{ij}|\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{ij}), \qquad j = 1, \ldots, n-1, \tag{8.46}$$

where of course (8.46) is trivially true when $j = n$.

This result can be demonstrated formally by repeated application of the exogeneity condition (8.45). In particular, for $j < n$,

$$\begin{aligned} p(y_{ij}|\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in}) &= \frac{p(y_{ij}, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in})}{p(\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in})} = \frac{p(\boldsymbol{x}_{in}|y_{ij}, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,n-1})p(y_{ij}, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,n-1})}{p(\boldsymbol{x}_{in}|\boldsymbol{x}_{i1} \ldots, \boldsymbol{x}_{i,n-1})p(\boldsymbol{x}_{i1} \ldots, \boldsymbol{x}_{i,n-1})} \\ &= \frac{p(\boldsymbol{x}_{in}|\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,n-1})p(y_{ij}, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,n-1})}{p(\boldsymbol{x}_{in}|\boldsymbol{x}_{i1} \ldots, \boldsymbol{x}_{i,n-1})p(\boldsymbol{x}_{i1} \ldots, \boldsymbol{x}_{i,n-1})} = p(y_{ij}|\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,n-1}), \end{aligned} \tag{8.47}$$

where the first equality in (8.47) follows because (8.45) implies that Y_{ij} for $j < n$ is independent of $\boldsymbol{x}_{in}$ given $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,n-1}$. By applying these same steps to the result in (8.47) repeatedly, (8.46) follows.

Of course, (8.46) is ***not as strong*** a condition as (8.43), that is,

$$E(Y_{ij}|\boldsymbol{x}_i) = E(Y_{ij}|\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in}) = E(Y_{ij}|\boldsymbol{x}_{ij}). \tag{8.48}$$

However, which of the assumptions (8.48) or (8.45) is required depends on the precise questions of interest.

- If, as in our examples and implicit in the above arguments, $\boldsymbol{x}_{ij}$ is defined as involving ***only*** values of time-dependent covariates recorded at time j, then

$$E(Y_{ij}|\boldsymbol{x}_{ij}) \tag{8.49}$$

  is the ***marginal population mean*** at time j, representing the ***marginal relationship*** between ***current*** response and ***current*** covariate values. In some contexts, investigators might well be interested in simply characterizing the ***association*** between current values of response and covariates, the so-called ***cross-sectional association***, in which case they might posit a marginal model (8.49) ***directly*** as a way of ***empirically representing*** this relationship.

Here, the investigators are ***not interested*** in making a ***causal interpretation***. ***Nonetheless***, as we demonstrate momentarily, unless they are willing to assume (8.48), ***great care*** must be taken in fitting the directly specified model (8.49)

- More often, however, (whether they admit it or not) investigators ***are*** interested in making ***causal interpretations***. From this perspective, when investigators adopt a marginal model, they wish to conclude that the current value of the covariate alone "***causes***" the response. As we have discussed in the context of the Six Cities study, such an interpretation ***cannot be made*** from fitting a marginal model because of the ***confounding*** that is likely present. That is, future smoking status is not independent of current response given past responses and smoking behavior. Thus, for instance, children whose wheezing status is poor might be more likely to be exposed to less smoking in the future than those not suffering from respiratory problems. Clearly, both of the assumptions (8.48) or (8.45) would be suspect for this study.

- In some situations, investigators might be interested in the relationship between ***cumulative exposure*** and response. Consider the air pollution example above, where interest focuses on the relationship between air pollution levels at nearby monitoring sites and health outcomes. Here, it is reasonable to assume that (8.45) holds, but (8.48) as stated is probably not true, as longitudinal health outcomes are likely impacted by the ***cumulative history*** of exposure to pollution.

  However, if we define a new covariate $\boldsymbol{x}_{ij}^* = (\boldsymbol{x}_{i1}^T, \ldots, \boldsymbol{x}_{ij}^T)^T$, then under (8.45), (8.48) holds with $\boldsymbol{x}_i^*$ and $\boldsymbol{x}_{ij}^*$ replacing $\boldsymbol{x}_i$ and $\boldsymbol{x}_{ij}$. As we demonstrate momentarily, this will facilitate valid inferences based on fitting a model based on $\boldsymbol{x}_{ij}^*$.

- Of course, the ***critical issue*** is whether or not (8.45) is a realistic assumption. If it is, as in the pollution example, then it is possible to draw ***causal interpretations***.

  If it is ***not***, then it is ***not possible*** to make causal interpretations by simply fitting models of the type discussed in this course. A specialized statistical framework for ***causal inference*** in the presence of ***time-dependent confounding*** is required of the type pioneered by Robins (1994), Robins, Greenland, and Hu (1999), and Robins, Hernán, and Brumback (2000); Vansteelandt and Joffe (2014) provide a comprehensive review and cite numerous references. This is the subject of an entire course.

UNBIASEDNESS OF THE LINEAR ESTIMATING EQUATION, REVISITED: As we noted previously, ***under the assumption*** that the model in (8.4),

$$\boldsymbol{f}_i(\boldsymbol{x}_i, \beta) = \begin{pmatrix} f(\boldsymbol{x}_{i1}, \beta) \\ \vdots \\ f(\boldsymbol{x}_{in_i}, \beta) \end{pmatrix} \tag{8.50}$$

is ***correctly specified***, the ***linear estimating equation*** (8.32), namely

$$\sum_{i=1}^{m} \boldsymbol{X}_i^T(\beta) \boldsymbol{V}_i^{-1}(\beta, \xi, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta)\} = \boldsymbol{0}, \tag{8.51}$$

is an ***unbiased estimating equation***, so that we expect $\widehat{\beta}$ obtained by solving (8.51) to be a ***consistent estimator*** for the true value β_0 of β.

Implicit in (8.50) is that the assumption (8.48),

$$E(Y_{ij}|\boldsymbol{x}_i) = E(Y_{ij}|\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in}) = E(Y_{ij}|\boldsymbol{x}_{ij}),$$

holds. We now take a closer look at the implications of this.

In a ***world-famous*** paper, Pepe and Anderson (1994) made the following simple but critically important observation. Assume, as would almost always be the case, that the ***working*** covariance model (8.5) for var$(\boldsymbol{Y}_i|\boldsymbol{x}_i)$,

$$\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i) = \boldsymbol{T}_i^{1/2}(\beta, \theta, \boldsymbol{x}_i)\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)\boldsymbol{T}_i^{1/2}(\beta, \theta, \boldsymbol{x}_i),$$

is such that $\boldsymbol{T}_i(\beta, \theta, \boldsymbol{x}_i) = \sigma^2 \text{diag}\{g^2(\beta, \delta, \boldsymbol{x}_{i1}), \ldots, g^2(\beta, \delta, \boldsymbol{x}_{in_i})\}$, as when the variance function depends on the mean response; and the ***working correlation model*** $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$ possibly depends on $\boldsymbol{x}_i$, usually through the times t_{ij}.

Under these conditions, the estimating equation (8.51) can be written as, ignoring the multiplicative scale parameter σ^2, (verify)

$$\sum_{i=1}^{m} \left(\frac{f_\beta(\boldsymbol{x}_{i1}, \beta)}{g^2(\beta, \delta, \boldsymbol{x}_{i1})} \cdots \frac{f_\beta(\boldsymbol{x}_{in_i}, \beta)}{g^2(\beta, \delta, \boldsymbol{x}_{in_i})} \right) \begin{pmatrix} \Gamma^{i11} & \cdots & \Gamma^{i1n_i} \\ \vdots & \vdots & \vdots \\ \Gamma^{in_i1} & \cdots & \Gamma^{in_in_i} \end{pmatrix} \begin{pmatrix} Y_{i1} - f(\boldsymbol{x}_{i1}, \beta) \\ \vdots \\ Y_{in_i} - f(\boldsymbol{x}_{in_i}, \beta) \end{pmatrix}, \tag{8.52}$$

where Γ^{ijk} is the (j, k) element of the ***inverse*** of the ***working correlation matrix***, and dependence of this matrix on $\boldsymbol{x}_i$ through the t_{ij} is suppressed.

Using shorthand notation

$$f_{\beta ij} = f_\beta(\boldsymbol{x}_{ij}, \beta), \qquad g_{ij}^{-2} = g^{-2}(\beta, \delta, \boldsymbol{x}_{ij}),$$

it is straightforward to rewrite (8.52) as

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} f_{\beta ik} g_{ik}^{-2} \Gamma^{ikj} \{ Y_{ij} - f(\boldsymbol{x}_{ij}, \beta) \}. \tag{8.53}$$

Taking the ***conditional expectation*** of a summand in (8.53) given $\boldsymbol{x}_i$ yields

$$\begin{aligned} E\left[f_{\beta ik} g_{ik}^{-2} \Gamma^{ikj} \{ Y_{ij} - f(\boldsymbol{x}_{ij}, \beta) \} | \boldsymbol{x}_i \right] &= f_{\beta ik} g_{ik}^{-2} \Gamma^{ikj} E\left[\{ Y_{ij} - f(\boldsymbol{x}_{ij}, \beta) \} | \boldsymbol{x}_i \right] \\ &= f_{\beta ik} g_{ik}^{-2} \Gamma^{ikj} \{ E(Y_{ij} | \boldsymbol{x}_i) - f(\boldsymbol{x}_{ij}, \beta) \}. \end{aligned} \tag{8.54}$$

It follows that (8.54) will be equal to zero ***only if***

$$E(Y_{ij} | \boldsymbol{x}_i) = f(\boldsymbol{x}_{ij}, \beta);$$

that is, if $E(Y_{ij}|\boldsymbol{x}_i)$ depends on $\boldsymbol{x}_i$ ***only through*** $\boldsymbol{x}_{ij}$, as in (8.48).

This result has ***important but often unappreciated*** implications. If the analyst assumes a model of the form (8.50), s/he must be willing to assume that (8.48),

$$E(Y_{ij} | \boldsymbol{x}_i) = E(Y_{ij} | \boldsymbol{x}_{ij}),$$

holds. If this ***does not hold*** , then the estimator for β can be ***inconsistent*** for the true value β_0.

Thus, if one assumes a ***marginal population mean model*** as in (8.49) directly, one must also be willing to assume (8.48) to ensure valid inferences.

Pepe and Anderson (1994) noted an additional result. If one adopts an ***independence working assumption*** for var($\boldsymbol{Y}_i|\boldsymbol{x}_i$), so takes the ***working correlation matrix*** to be a $(n_i \times n_i)$ ***identity matrix*** , it is straightforward that (8.53) reduces to

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} f_{\beta ij} g_{ij}^{-2} \{ Y_{ij} - f(\boldsymbol{x}_{ij}, \beta) \}. \tag{8.55}$$

Note that, for a summand of (8.55),

$$E\left[f_{\beta ij} g_{ij}^{-2} \{ Y_{ij} - f(\boldsymbol{x}_{ij}, \beta) \} | \boldsymbol{x}_{ij} \right] = f_{\beta ij} g_{ij}^{-2} \{ E(Y_{ij} | \boldsymbol{x}_{ij}) - f(\boldsymbol{x}_{ij}, \beta) \}. \tag{8.56}$$

It follows from (8.56) that, as long as $f(\boldsymbol{x}_{ij}, \beta)$ is a ***correctly specified*** model for $E(Y_{ij}|\boldsymbol{x}_{ij})$, the estimating equation (8.51) is ***unbiased***.

RESULT: These results suggest that, if one is ***not willing*** to assume (8.48), the ***working correlation matrix*** should be taken to be an identity matrix; that is, a ***independence working assumption*** should be made to ensure consistent inference. This is critical if one is interested in ***marginal inference*** as discussed above.

- Unfortunately, as noted above, appreciation for this result is not widespread, and this advice is ***rarely*** followed in practice.

8.7 Missing data

As discussed in Section 5.6, a key challenge in longitudinal data analysis is ***missing data***, in particular ***dropout*** of individuals over the course of the observation period. In this section, we review briefly the implications of such dropout for the ***validity of inferences*** in population-averaged, possibly nonlinear models for general response types fitted using ***GEE methods***. A more detailed treatment of this issue is presented Chapter 5 of the instructor's notes for the course "Statistical Methods for Analysis With Missing Data."

DATA STRUCTURE: As we did previously, we consider the situation where the ***full data*** on the response ***intended*** to be collected on each individual i are at n prespecified times $t_1, \ldots, t_n$. As in (5.93), define the full data response vector as

$$\mathcal{Z}_i = (Z_{i1}, \ldots, Z_{in})^T. \tag{8.57}$$

The responses ***actually observed*** on i,, $\boldsymbol{Y}_i$, are a subset of the components of $\mathcal{Z}_i$.

The ***missing data indicators*** are, as in (5.94

$$R_{ij} = \begin{cases} 1 & \text{if } Z_{ij} \text{ is observed,} \\ 0 & \text{otherwise,} \end{cases}$$

$j = 1, \ldots, n$, and

$$\mathcal{R}_i = (R_{i1}, \ldots, R_{in})^T, \tag{8.58}$$

where we denote possible values of $\mathcal{R}_i$ (vectors of 0s and 1s) by $\boldsymbol{r}$. In the particular case of ***dropout***, assuming all individuals are observed at t_1, which we take here to be ***baseline***, there are n possible ***missingness patterns r***, given by

$$\boldsymbol{r}^{(1)} = (1, 0, \ldots, 0), \quad \boldsymbol{r}^{(2)} = (1, 1, 0, \ldots, 0), \quad \ldots, \quad = \boldsymbol{r}^{(n)} = (1, 1, \ldots, 1). \tag{8.59}$$

Dropout is of course a ***monotone missingness pattern***.

We henceforth assume that ***all individuals*** are observed at baseline, so that $R_{i1} = 1$ for all i.

- It is clear that, under a ***dropout mechanism***, if $R_{ij} = 1$, then it must be the case that $R_{i,j-1} = \cdots = R_{i2} = R_{i1} = 1$ (convince yourself).
- If $R_{ij} = 1$, then clearly $Z_{i1} = Y_{i1}, \ldots, Z_{ij} = Y_{ij}$; that is, the responses ***actually observed*** are those ***intended*** through time j.

At each time t_j, $j = 1, \ldots, n$, it is also intended to collect ***covariates*** $\boldsymbol{x}_{ij}$, where $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}^T, \ldots, \boldsymbol{x}_{in}^T)^T$. Interest focuses on the ***relationship*** between population mean response and $\boldsymbol{x}_i$.

- As discussed in the last section, ***conceptual challenges*** arise in the case of ***endogenous covariates*** $\boldsymbol{x}_{ij}$. Accordingly, the developments we present in this section are restricted to the case where the $\boldsymbol{x}_{ij}$ are ***exogenous covariates***.
- Under ***exogeneity***, the value of $\boldsymbol{x}_i$ is observed/known to the data analyst ***throughout the study period***, regardless of whether or not i drops out.
- This would be the case in a study in which all covariates are ***time-independent***, recorded only at ***baseline*** (t_1).
- This would also hold in a situation like the air pollution monitoring example in the previous section, where the covariate (pollution) process evolves ***independently*** of the responses of any study participant.
- We thus take in this section $\boldsymbol{x}_{ij} = (\boldsymbol{a}_i, t_j)$, where $\boldsymbol{a}_i$ is a vector of ***exogenous covariates*** whose values are known to the data analyst at all n intended time points, regardless of dropout. The simplest case is where $\boldsymbol{a}_i$ are ***baseline covariates***.

It may also be the case that ***additional information*** $\boldsymbol{v}_{ij}$ is recorded on each i at each t_j.

- We emphasize that there is ***no interest*** in the $\boldsymbol{v}_{ij}$ insofar as the mean response goes. As we discuss shortly, the $\boldsymbol{v}_{ij}$ are of interest only for their potential utility for describing the ***missingness/dropout mechanism***.
- The $\boldsymbol{v}_{ij}$ are ***observed*** along with $Z_{ij} = Y_{ij}$ as long as i has ***not yet dropped out***; that is, as long as $R_{ij} = 1$.

SUMMARY: Combining, the scenario we consider in this section can be summarized as follows. For each individual i, the ***full data*** on responses in (8.57) and additional information $\boldsymbol{v}_{ij}$ ***intended to be collected*** are

$$\widetilde{\mathcal{Z}}_i = \{(Z_{i1}, \boldsymbol{v}_{i1}^T)^T, \ldots, (Z_{in}, \boldsymbol{v}_{in}^T)^T\}^T. \tag{8.60}$$

along with $\boldsymbol{x}_i$, which is ***always observed***.

Here, $\mathcal{R}_i = (R_{i1}, \ldots, R_{in})^T$ in (8.58) is such that

$$R_{ij} = \begin{cases} 1 & \text{if } (Z_{ij}, \boldsymbol{v}_{ij}^T)^T \text{ is observed,} \\ 0 & \text{otherwise,} \end{cases}$$

If $R_{ij} = 1$, then $(Z_{ij}, \boldsymbol{v}_{ij}^T)^T = (Y_{ij}, \boldsymbol{v}_{ij}^T)^T$.

DROPOUT INDICATOR: It is conventional in the situation of dropout to define for each i the ***dropout indicator***

$$D_i = 1 + \sum_{j=1}^{n} R_{ij}. \tag{8.61}$$

- It is straightforward from (8.61) that $D_i = j$ implies that i was ***last seen*** at t_{j-1}, so dropped out sometime in the time interval (t_{j-1}, t_j). In this case, ***by convention***, i is said to have dropped out at t_j.
- Moreover, from (8.59), the possible values of $\mathcal{R}_i$ are $\boldsymbol{r}^{(j)}$, $j = 1, \ldots, n$, where $\boldsymbol{r}^{(j)}$ corresponds to dropout at time $j+1$; i.e., being last seen at time t_j. It is thus clear that the events

  $$\{D_i = j + 1\} \qquad \text{and} \qquad \{\mathcal{R}_i = \boldsymbol{r}^{(j)}\} \tag{8.62}$$

  are ***equivalent*** (check).
- Because $R_{i1} = 1$ always, D_i has possible values $2, \ldots, n+1$, where $D_i = n+1$ corresponds the situation where the ***full data are observed***.

DROPOUT MECHANISMS: Recall from (5.96) that the various ***missing data mechanisms*** of MCAR, MAR, and MNAR are defined in terms of the density of $\mathcal{R}_i$ given the full data, which are $\widetilde{\mathcal{Z}}_i$ here, and $\boldsymbol{x}_i$; that is,

$$p(\boldsymbol{r}_i | \boldsymbol{Z}_i, \boldsymbol{x}_i) = \text{pr}(\mathcal{R}_i = \boldsymbol{r}_i | \widetilde{\mathcal{Z}}_i = \boldsymbol{Z}_i, \boldsymbol{x}_i).$$

Because, under dropout, the possible values of $\boldsymbol{r}_i$ are ***restricted*** to be $\boldsymbol{r}^{(j)}$, $j = 1, \ldots, n$, it is straightforward from (8.61) and (8.62) that we can express these mechanisms ***equivalently*** in terms of

$$\text{pr}(D_i = j+1 | \widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(\mathcal{R}_i = \boldsymbol{r}^{(j)} | \widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i), \qquad j = 1, \ldots, n. \tag{8.63}$$

Letting $\widetilde{\mathcal{Z}}_{(\boldsymbol{r}^{(j)}),i}$ denote the part of $\widetilde{\mathcal{Z}}_i$ that is ***observed*** when $\mathcal{R}_i = \boldsymbol{r}^{(j)}$, ***equivalently*** , when $D_i = j+1$, from (8.60),

$$\widetilde{\mathcal{Z}}_{(\boldsymbol{r}^{(j)}),i} = \{(Z_{i1}, \boldsymbol{v}_{i1}^T)^T, \ldots, (Z_{ij}, \boldsymbol{v}_{ij}^T)^T\}^T = \{(Y_{i1}, \boldsymbol{v}_{i1}^T)^T, \ldots, (Y_{ij}, \boldsymbol{v}_{ij}^T)^T\}^T.$$

For convenience shortly, define the ***history*** through time t_j as

$$\mathcal{H}_{ij} = [\{(Z_{i1}, \boldsymbol{v}_{i1}^T)^T, \ldots, (Z_{ij}, \boldsymbol{v}_{ij}^T)^T\}^T, \boldsymbol{x}_i], \qquad j = 1, \ldots, n, \tag{8.64}$$

recognizing that $\boldsymbol{x}_i$ is known throughout time. If $D_i = j+1$, then note that

$$\mathcal{H}_{ik} = [\{(Y_{i1}, \boldsymbol{v}_{i1}^T)^T, \ldots, (Y_{ik}, \boldsymbol{v}_{ik}^T)^T\}^T, \boldsymbol{x}_i], \qquad k = 1, \ldots, j.$$

Using (8.63) and (8.64), the ***dropout mechanisms*** are conventionally represented as follows.

- ***Missing Completely at Random (MCAR).*** The probability of dropout at $j+1$ does not depend on $\widetilde{\mathcal{Z}}_i$; that is

  $$\text{pr}(D_i = j+1 | \widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(D_i = j+1 | \boldsymbol{x}_i) = \pi(j+1, \boldsymbol{x}_i). \tag{8.65}$$

  Analogous to (5.97), (8.65) implies that

  $$D_i \perp\!\!\!\perp \widetilde{\mathcal{Z}}_i \,|\, \boldsymbol{x}_i \tag{8.66}$$

  which is equivalent to $\mathcal{R}_i \perp\!\!\!\perp \widetilde{\mathcal{Z}}_i \,|\, \boldsymbol{x}_i$.

- ***Missing at Random (MAR).*** The probability of dropout at $j+1$ depends on $\widetilde{\mathcal{Z}}_i$ ***only*** through components of $\widetilde{\mathcal{Z}}_i$ that ***are observed*** under dropout at $j+1$,

  $$\text{pr}(D_i = j+1 | \widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(D_i = j+1 | \widetilde{\mathcal{Z}}_{(\boldsymbol{r}^{(j)}),i}, \boldsymbol{x}_i) = pr(D_i = j+1 | \mathcal{H}_{ij}) = \pi(j+1, \mathcal{H}_{ij}). \tag{8.67}$$

- ***Missing Not at Random (MNAR).*** The probability of dropout at $j+1$ depends on components of $\widetilde{\mathcal{Z}}_i$ that ***are not observed*** under dropout at $j+1$.

ADDITIONAL INFORMATION $\boldsymbol{v}_{ij}$: Although the $\boldsymbol{v}_{ij}$ are not of ***direct interest*** for modeling, the definitions above demonstrate that they may be implicated in the ***missingness mechanism***. Thus, if the $\boldsymbol{v}_{ij}$ are available to the data analyst, they are ***critical*** for justifying the ***assumption of MAR*** , as we do shortly.

OBSERVED DATA GEE: We are now in a position to consider the behavior of the estimator $\widehat{\beta}$ for β obtained by solving the usual linear estimating equation (8.51) based on the ***observed data***. It proves convenient to ***modify the notation*** as follows.

Recall from above that we restrict attention to $\boldsymbol{x}_{ij} = (\boldsymbol{a}_i, t_j)$, where $\boldsymbol{a}_i$ are ***exogenous covariates*** known throughout time and observed for all i. Suppose we have posited a model for the mean of the ***intended full response vector*** $\mathcal{Z}_i$ in (8.57), $E(\mathcal{Z}_i|\boldsymbol{x}_i)$, of the form

$$\boldsymbol{f}_i(\boldsymbol{x}_i, \beta) = \begin{pmatrix} f(\boldsymbol{x}_{i1}, \beta) \\ \vdots \\ f(\boldsymbol{x}_{in}, \beta) \end{pmatrix} = \begin{pmatrix} f(\boldsymbol{a}_i, t_1, \beta) \\ \vdots \\ f(\boldsymbol{a}_i, t_n, \beta) \end{pmatrix} = \begin{pmatrix} f_1(\boldsymbol{a}_i, \beta) \\ \vdots \\ f_n(\boldsymbol{a}_i \beta) \end{pmatrix} \qquad (n \times 1). \tag{8.68}$$

Note that conditioning on $\boldsymbol{x}_i$ here is ***equivalent*** to conditioning on $\boldsymbol{a}_i$. Assume that this model is ***correctly specified***.

With $\boldsymbol{f}_i(\boldsymbol{x}_i, \beta)$ defined as in (8.68), let

$$\mathcal{X}_i(\boldsymbol{x}_i, \beta) = \begin{pmatrix} f_\beta^T(\boldsymbol{x}_{i1}, \beta) \\ \vdots \\ f_\beta^T(\boldsymbol{x}_{in}, \beta) \end{pmatrix} = \begin{pmatrix} f_{1\beta}^T(\boldsymbol{a}_i, \beta) \\ \vdots \\ f_{n\beta}^T(\boldsymbol{a}_i, \beta) \end{pmatrix} \qquad (n \times p) \tag{8.69}$$

be the ***gradient matrix*** of the mean model for the intended full response vector in (8.68), and likewise and let $\mathcal{V}_i(\beta, \xi, \boldsymbol{x}_i)$ be a $(n \times n)$ ***working covariance model*** for $\text{var}(\mathcal{Z}_i|\boldsymbol{x}_i)$.

If $D_i = j + 1$, then the ***observed response vector*** on i is $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ij})^T$ $(n_i = j)$. Let $\boldsymbol{X}_i^{(j)}(\beta)$ $(j \times p)$ and $\boldsymbol{V}_i^{(j)}(\beta, \xi, \boldsymbol{x}_i)$ $(j \times j)$ be the corresponding submatrices of $\mathcal{X}_i(\boldsymbol{x}_i, \beta)$ and $\mathcal{V}_i(\beta, \xi, \boldsymbol{x}_i)$. These are as defined previously in this chapter for each i, where we have added the superscript (j) to ***emphasize*** that these correspond an ***observed data vector*** of length $n_i = j \leq n$.

With these definitions, the ***linear estimating equation*** (8.51) can be written as (verify)

$$\sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} I(D_i = j + 1) \boldsymbol{X}_i^{(j)T}(\beta) \{ \boldsymbol{V}_i^{(j)}(\beta, \xi, \boldsymbol{x}_i) \}^{-1} \begin{pmatrix} Y_{i1} - f_1(\boldsymbol{a}_i, \beta) \\ \vdots \\ Y_{ij} - f_j(\boldsymbol{a}_i, \beta) \end{pmatrix} \right\} = \boldsymbol{0}. \tag{8.70}$$

Consider the conditional expectation of the ith summand in (8.70) given $\boldsymbol{x}_i$ $(\boldsymbol{a}_i)$.

Assuming expectation is under the parameter values β and ξ, the estimator $\widehat{\beta}$ solving (8.70) will be ***consistent*** for β_0 if

$$E\left\{ I(D_i = j+1)\boldsymbol{X}_i^{(j)T}(\beta)\{\boldsymbol{V}_i^{(j)}(\beta,\xi,\boldsymbol{x}_i)\}^{-1} \begin{pmatrix} Y_{i1} - f_1(\boldsymbol{a}_i,\beta) \\ \vdots \\ Y_{ij} - f_1(\boldsymbol{a}_i,\beta) \end{pmatrix} \middle| \boldsymbol{x}_i \right\} = \boldsymbol{0} \qquad j = 1,\ldots,n. \tag{8.71}$$

Consider the left hand side of (8.71) under ***different dropout/missingness mechanisms***.

- ***MCAR.*** Writing this as

$$E\left[E\left\{ I(D_i = j+1)\boldsymbol{X}_i^{(j)T}(\beta)\{\boldsymbol{V}_i^{(j)}(\beta,\xi,\boldsymbol{x}_i)\}^{-1} \begin{pmatrix} Y_{i1} - f_1(\boldsymbol{a}_i,\beta) \\ \vdots \\ Y_{ij} - f_1(\boldsymbol{a}_i,\beta) \end{pmatrix} \middle| \boldsymbol{x}_i, \widetilde{\mathcal{Z}}_i \right\} \middle| \boldsymbol{x}_i \right], \tag{8.72}$$

  and using the fact that, as in (8.66), D_i is ***independent of*** $\widetilde{\mathcal{Z}}_i$ given $\boldsymbol{x}_i$, the left hand side of (8.71) becomes (verify)

$$E\left\{ \pi(j+1,\boldsymbol{x}_i)\,\boldsymbol{X}_i^{(j)T}(\beta)\{\boldsymbol{V}_i^{(j)}(\beta,\xi,\boldsymbol{x}_i)\}^{-1} \begin{pmatrix} Y_{i1} - f_1(\boldsymbol{a}_i,\beta) \\ \vdots \\ Y_{ij} - f_1(\boldsymbol{a}_i,\beta) \end{pmatrix} \middle| \boldsymbol{x}_i \right\} = \boldsymbol{0}.$$

  Thus, under ***MCAR***, the estimating equation (8.70) is ***unbiased***, as intuition would suggest. Accordingly, if ***dropout*** is completely at random, so is ***unrelated*** to the evolving response, as might be the case if study participants dropped out to move to another city, $\widehat{\beta}$ obtained by solving the usual linear GEE is ***consistent***.

- ***MAR.*** Using the equivalent form (8.72), (8.71) can be written using (8.67) as (verify)

$$\boldsymbol{X}_i^{(j)T}(\beta)\{\boldsymbol{V}_i^{(j)}(\beta,\xi,\boldsymbol{x}_i)\}^{-1} E\left\{ \pi(j+1,\mathcal{H}_{ij}) \begin{pmatrix} Y_{i1} - f_1(\boldsymbol{a}_i,\beta) \\ \vdots \\ Y_{ij} - f_1(\boldsymbol{a}_i,\beta) \end{pmatrix} \middle| \boldsymbol{x}_i \right\}. \tag{8.73}$$

  The kth element in the expectation in (8.73) is (verify), $k = 1,\ldots,j$,

$$\text{cov}\{\pi(j+1,\mathcal{H}_{ij}),\, Y_{ik}\}.$$

  From (8.67), $\pi(j+1,\mathcal{H}_{ij})$ ***depends on*** $Y_{i1},\ldots,Y_{ij}$; thus, this covariance is almost certainly ***not equal to zero*** in general. Thus, under MAR, the usual estimator $\widehat{\beta}$ ***is not consistent for*** β_0 ***in general***.

- ***MNAR.*** It should be clear that ***similar considerations*** apply here, and the estimator is ***not consistent*** in general.

RESULT: Even under the simpler condition of ***exogenous covariates***, when there are missing data due to ***dropout***, the usual GEE estimator obtained by solving (8.51) based on the ***observed data*** is ***only guaranteed to be consistent*** if the dropout mechanism is MCAR.

- Given that MCAR is often ***not realistic*** in practice, this is a rather ***scary*** result.
- The data analyst thus must take ***great care*** to understand the possible reasons for dropout.

REMARK: The foregoing developments were derived in the context of ***dropout***, which is a ***monotone missingness mechanism***. Under ***nonmonotone patterns***, the same scary considerations apply: the estimator for β based on the ***observed data*** is only guaranteed to be consistent if the missingness mechanism is MCAR.

Moreover, methods to achieve a consistent estimator for β under the assumption of ***MAR*** with ***nonmonotone missingness*** are ***very difficult*** to develop. However, with ***monotone missingness***, i.e., dropout, such methods are possible, as we now discuss.

MODIFIED ESTIMATING EQUATIONS UNDER MAR DROPOUT: When the dropout mechanism is thought to be ***MAR***, it is possible to ***modify*** the usual ***linear estimating equation*** (8.70) to obtain estimating equations that ***are unbiased***. We briefly sketch the two main approaches.

We first consider a convenient representation of $\text{pr}(D_i = j+1|\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i)$. Define the ***cause-specific hazard function*** of dropout as

$$\lambda_j(\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(D_i = j|D_i \geq j, \widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i), \qquad j = 2, \ldots, n. \tag{8.74}$$

Note that $\lambda_1(\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(D_i = 1|D_i \geq 1, \widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = 0$ because $(Z_{i1}, \boldsymbol{v}_{i1})$ is ***always observed*** for all i; and $\lambda_{n+1}(\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(D_i = n+1|D_i \geq n+1, \widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = 1$ by construction. It can then be deduced (do it) that

$$\overline{\pi}_j(\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(R_{ij} = 1|\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \prod_{\ell=1}^{j}\{1 - \lambda_\ell(\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i)\}, \qquad j = 2, \ldots, n, \tag{8.75}$$

where $\overline{\pi}_1(\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(R_{i1} = 1|\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(R_{i1} = 1) = \overline{\pi}_1 = 1$, because all individuals are observed at baseline, and thus (verify)

$$\text{pr}(D = j+1|\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \overline{\pi}_j(\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i)\lambda_{j+1}(\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i), \qquad j = 1, \ldots, n. \tag{8.76}$$

Under MAR, (8.74) can be written as

$$\lambda_j(\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(D_i = j | D_i \geq j, \widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(D_i = j | D_i \geq j, \mathcal{H}_{i,j-1}) = \lambda_j(\mathcal{H}_{i,j-1}), \qquad j = 2, \ldots, n. \quad (8.77)$$

so that the ***hazard of dropping out*** at time t_j, so last being seen at time t_{j-1}, depends only on the ***observed history*** through time t_{j-1}. Similarly, (8.75) and (8.76) become

$$\overline{\pi}_j(\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \overline{\pi}_j(\mathcal{H}_{i,j-1}) = \text{pr}(R_{ij} = 1 | \mathcal{H}_{i,j-1}) = \prod_{\ell=1}^{j} \{1 - \lambda_\ell(\mathcal{H}_{i,j-1}\}, \qquad j = 2, \ldots, n, \quad (8.78)$$

where $\overline{\pi}_1 = 1$, and

$$\text{pr}(D = j + 1 | \widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(D = j + 1 | \mathcal{H}_{ij}) = \overline{\pi}_j(\mathcal{H}_{i,j-1})\lambda_{j+1}(\mathcal{H}_{ij}), \qquad j = 1, \ldots, n. \quad (8.79)$$

Equations (8.77)–(8.79) demonstrate that, ***under the assumption of MAR***, it is possible to develop ***models*** for the ***hazard functions***

$$\lambda_j(\mathcal{H}_{i,j-1})$$

based on the ***observed data*** and thereby obtain models for

$$\text{pr}(R_{ij} = 1 | \mathcal{H}_{i,j-1}) = \overline{\pi}_j(\mathcal{H}_{i,j-1}) \quad \text{and} \quad \text{pr}(D = j + 1 | \mathcal{H}_{ij}).$$

These models can then be ***fitted and substituted*** into the ***modified estimating equations*** we now discuss.

WEIGHTED GENERALIZED ESTIMATING EQUATIONS (WGEEs): The modified estimating equations involve ***weighting*** each the summand of the usual linear estimating equation, which we wrote in the form (8.70), in a way that yields unbiasedness. The two main approaches are:

- ***Inverse probability weighting at the individual level.*** This was proposed by Fitzmaurice, Molenberghs, and Lipsitz (1995) and involves ***weighting*** the contribution to (8.70) for each individual by the ***inverse*** of the probability of that individual's observed dropout time, conditional on his/her observed history.
- ***Inverse probability weighting at the occasion level.*** This was proposed by Robins, Rotnitzky, and Zhao (1995) and involves ***weighting*** contributions to (8.70) at ***each time point*** for ***each individual*** by the inverse of the probability of having an observed response at that time point, conditional on observed history.
- These and more advanced techniques can be justified theoretically by appealing to the general theory of ***semiparametrics*** and missing data as in Tsiatis (2006).

INVERSE PROBABILITY WEIGHTING AT THE INDIVIDUAL LEVEL: Define the weight

$$w_{ij} = \frac{I(D_i = j+1)}{\overline{\pi}_j(\mathcal{H}_{i,j-1})\lambda_{j+1}(\mathcal{H}_{ij})},$$

where, from (8.79), the denominator of w_{ij} is $\text{pr}(D = j+1|\mathcal{H}_{ij})$. The ***WGEE*** is then

$$\sum_{i=1}^{m}\left\{\sum_{j=1}^{n} w_{ij}\boldsymbol{X}_i^{(j)T}(\beta)\{\boldsymbol{V}_i^{(j)}(\beta,\xi,\boldsymbol{x}_i)\}^{-1}\begin{pmatrix} Y_{i1}-f_1(\boldsymbol{a}_i,\beta)\\ \vdots \\ Y_{ij}-f_j(\boldsymbol{a}_i,\beta)\end{pmatrix}\right\} = \boldsymbol{0}. \tag{8.80}$$

Comparing (8.80) to (8.70), the only difference is the ***inverse weighting*** by $\text{pr}(D = j+1|\mathcal{H}_{ij})$.

The ***diligent student*** can verify that, if $\lambda_j(\mathcal{H}_{i,j-1})$ are the ***true hazard functions***, the expectation of a summand of (8.80) is indeed equal to zero, so that (8.80) is an ***unbiased estimating equation*** under MAR. This can be accomplished by representing the expectation of a summand as in (8.72), with inner conditioning on $\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i$.

INVERSE PROBABILITY WEIGHTING AT THE OCCASION LEVEL: Define the $(n \times n)$ diagonal weight matrix

$$\mathcal{W}_i = \text{diag}\left(\frac{R_{i1}}{\overline{\pi}_1}, \frac{R_{i2}}{\overline{\pi}_2(\mathcal{H}_{i1})}, \ldots, \frac{R_{in}}{\overline{\pi}_n(\mathcal{H}_{i,n-1})}\right).$$

The ***WGEE*** is then

$$\sum_{i=1}^{m}\mathcal{X}_i^T(\boldsymbol{x}_i,\beta)\mathcal{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\mathcal{W}_i\begin{pmatrix} Z_{i1}-f_1(\boldsymbol{a}_i,\beta)\\ \vdots \\ Z_{in}-f_n(\boldsymbol{a}_i,\beta)\end{pmatrix} = \boldsymbol{0}, \tag{8.81}$$

where $\mathcal{X}_i^T(\boldsymbol{x}_i\beta)$ and $\mathcal{V}_i(\beta,\xi,\boldsymbol{x}_i)$ are defined in and below (8.69). From the form of $\mathcal{W}_i$, when $R_{ij} = 1$, so that i has not yet dropped out, $Z_{ij} = Y_{ij}$, the observed response at t_j. If then $R_{i,j+1} = 0$, all subsequent $R_{ik} = 0$, $k > j$, and it is straightforward to observe that the summand will depend only on $Y_{i1}, \ldots, Y_{ij}$, $\boldsymbol{X}_i^{(j)}(\beta)$ $(j \times p)$, and the upper left $(j \times j)$ submatrix of $\mathcal{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)$ (verify).

As with (8.80), it can be shown by a similar conditioning argument that the conditional (on $\boldsymbol{x}_i$) expectation of a summand of (8.81) is equal to zero, so that (8.81) is an ***unbiased estimating equation*** under MAR.

RESULT: If the assumption of ***MAR dropout*** is plausible, then these methods can be used to obtain consistent estimators for β in a ***full data model*** of interest with exogenous covariates.

- However, this requires ***correct specification*** of ***models*** for the ***dropout hazard functions*** $\lambda_j(\mathcal{H}_{i,j-1})$, $j = 2, \ldots, n$. If these models are ***misspecified***, then the estimating equations ***no longer*** need be unbiased.

Considerations for such modeling are discussed in Chapter 5 of the instructor's notes for the course "Statistical Methods for Analysis With Missing Data," where it is demonstrated that one approach is to adopt ***logistic regression models*** for the hazards for each j.

- It is ***not possible*** to show that one weighting approach yields ***more efficient inferences*** than the other in general. The individual-level approach has been preferred in practice on the grounds that it is ***simpler to implement***. Specifically, one can model and fit the dropout hazard functions as above and form ***fixed, estimated weights*** w_{ij}. Many software packages for solving the linear GEE allow fixed weights for each individual, so that these estimated weights can be incorporated straightforwardly.

  The only widely available software that implements both methods, including specification of the ***hazard models***, is SAS `proc gee`. Its use is demonstrated in the instructor's notes for the course "Statistical Methods for Analysis With Missing Data."

8.8 Examples

We briefly describe how modeling might proceed in two examples.

EXAMPLE 5: Epileptic seizures and chemotherapy, continued. Recall that the ***among-individual covariates*** randomized treatment δ_i (= 0 for placebo and = 1 for progabide), baseline seizure count over the 8 weeks prior to the start of the study, c_i, and age a_i at the start of the study ***do not change*** over time. Thus, $\boldsymbol{x}_{ij} = (t_{ij}, \delta_i, c_i, a_i)^T$ are ***exogenous***. All $m = 59$ subjects are seen at all $n = 4$ visits, with no missing responses. The response Y_{ij} is a ***count***, so a natural model is a ***loglinear model*** as in (8.7),

$$f(\boldsymbol{x}_{ij}, \beta) = \exp\{h(\boldsymbol{x}_{ij})^T\beta) \quad \text{or equivalently} \quad \log\{f(\boldsymbol{x}_{ij}, \beta)\} = h(\boldsymbol{x}_{ij})^T\beta,$$

where $h(\cdot)$ is a vector of functions of $\boldsymbol{x}_{ij}$.

In the above, we treat the ***baseline seizure count*** as a ***covariate***. However, strictly speaking, baseline seizure count also a measure of the ***response*** prior to the start of treatment, albeit over an observation period of 8 weeks rather than 2 weeks as is the case for the post-treatment responses. Many authors, including Thall and Vail (1990) themselves, have regarded baseline seizure count as a ***covariate***, in part to avoid the issue of the ***different lengths*** of the observation periods. While this is convenient, it could also be ***inefficient***, as the baseline count also contains information on the ***distribution of responses***.

It is a ***simple matter*** to address the time scale issue, as we demonstrate momentarily, so in what follows, we follow Diggle et al. (2002) and treat the baseline 8-week seizure count as a ***response measure*** Y_{i1} at time $t_{ij} = 0$, and ***reindex*** the responses at periods 1–4 as $Y_{i2}, \ldots, Y_{i5}$, so that $j = 1, \ldots, n = 5$. Accordingly, we ***redefine*** $\boldsymbol{x}_{ij} = (t_{ij}, \delta_i, a_i)^T$. Note that, here, our temporal convention is ***different from*** that in Section 8.6, although this poses no difficulty.

The more ***fundamental issue*** is whether or not it is a good idea to treat a baseline response as a ***covariate*** to take into account the fact that individuals differ in their responses prior to treatment or if it is preferable to treat the baseline value as ***part of the response vector*** for each individual. In the case of ***linear models*** for the mean response, the two strategies can be ***equivalent*** ; however, when the mean response model is ***nonlinear*** as it is here, this is no longer the case.

This is a matter of ***considerable debate*** in the literature, and the choice is in part guided by the nature of the ***questions of interest*** and the ***type of study***. This instructor agrees with Fitzmaurice, Laird, and Ware (2011) that, as a general strategy, it is ***preferable*** to treat a baseline response as part of the response vector rather than as a covariate, partly on ***efficiency grounds***. A very nice, practical discusion of this issue is given in Sections 5.6 and 5.7 of Fitzmaurice et al. (2011).

Accordingly, define $o_{ij} = 8$ if $j = 1$ ($t_{ij} = 0$, baseline) and $o_{ij} = 2$, $j = 2, \ldots, 5$ ($t_{ij} > 0$, observation period/visit 1–4). Thus, o_{ij} records the time scale over which Y_{ij} was ascertained (8 or 2 weeks). Define $v_{ij} = 0$ if $t_{ij} = 0$ (baseline) and $v_{ij} = 1$ if $t_{ij} > 0$ (visit 1–4). We consider the loglinear model

$$E(Y_{ij}|\boldsymbol{x}_i) = \exp(\log o_{ij} + \beta_0 + \beta_1 v_{ij} + \beta_2 \delta_i + \beta_3 \delta_i v_{ij}). \tag{8.82}$$

In (8.82), $\log o_{ij}$ is an ***offset*** in the following sense. On the log scale, (8.82) is equivalent to

$$\log\{E(Y_{ij}|\boldsymbol{x}_i)\} - \log o_{ij} = \log\{E(Y_{ij}/o_{ij}|\boldsymbol{x}_i)\} = \beta_0 + \beta_1 v_{ij} + \beta_2 \delta_i + \beta_3 \delta_i v_{ij}. \tag{8.83}$$

Thus, (8.83) shows that (8.82) is equivalent to modeling the means of $Y_{i1}/8$ and $Y_{ij}/2$, $j = 2, \ldots, 5$; that is, the ***average number of seizures*** per week over each period.

Model (8.82) specifies that the mean of average numbers of seizures per week at baseline ($j = 1$) is

$$E(Y_{i1}/8|\boldsymbol{x}_i) = \exp(\beta_0 + \beta_2 \delta_i)$$

and for $J = 2, \ldots, 5$ is

$$E(Y_{i1}/2|\boldsymbol{x}_i) = \exp\{\beta_0 + \beta_1 + (\beta_2 + \beta_3)\delta_i\}.$$

This model is consistent with the impression given by the sample mean average numbers of seizures, as summarized below:

Visit	*Placebo*	*Progabide*
0 (baseline)	7.70	7.90
1	9.35	8.58
2	8.29	8.42
3	8.79	8.13
4	7.96	6.71
average over visits 1–4	8.60	7.96

The sample means in each group take a ***jump*** at visit 1 to a higher level that is possibly different in each group, represented in the model by the terms $\beta_1 v_{ij}$ and $\beta_3 v_{ij}\delta_i$. The remain relatively flat in each group thereafter, although there is an apparent drop at visit 4 in the progabide group (see below).

Given that this is a ***randomized study*** , a ***simplification*** of the model would be to ***remove*** the term $\beta_2\delta_i$, which allows the mean to differ at baseline between the two treatment groups. Indeed, the raw sample means of average number of seizures in the two groups are almost ***identical***. Average age at baseline (average of a_i) is 29.6 (SD 6.0) for the placebo group and 27.7 (6.6) for the progabide group, which are very similar, further supporting the contention that the randomization was carried out appropriately. We maintain the $\beta_2\delta_i$ term in analyses on the course website, but it probably could be deleted.

A modification of the model is as follows. As above, the sample means seem to suggest a ***possible drop*** in mean response at the 4th visit, we define an additional indicator variable $v4_{ij} = 0$ unless $j = 5$ for the possibility that the mean response at baseline is associated with age of the subject. The model is modified to

$$E(Y_{ij}|\boldsymbol{x}_i) = \exp(\log o_{ij} + \beta_0 + \beta_1 v_{ij} + \beta_2\delta_i + \beta_3 v_{ij}\delta_i + \beta_4 v4_{ij} + \beta_5 v4_{ij}\delta_i).$$

The parameter β_5 reflects whether or not the difference in post-baseline mean response in fact ***changes*** at the fourth visit, while β_4 allows the possibility that the mean response "shifts" at the 4th visit relative to the earlier ones. A further modification would be to incorporate age at baseline into the model to evaluate if treatment effects differ with age.

Fits of these models in SAS and R are on the course website.

EXAMPLE 6: Maternal smoking and child respiratory health, continued. The ***among-individual covariates*** are city c_i (= 0 for Portage, = 1 for Kingston), which is ***time-independent*** , and mother's smoking status at child's age (time) t_{ij}, s_{ij} (=0, 1, or 2 as the mother's smoking was none, moderate, or heavy). Define $\boldsymbol{x}_{ij} = (t_{ij}, c_i, s_{ij})^T$ for mother-child pair i.

Recall that this is an ***observational study*** ; thus, as discussed in Section 8.6, it is very likely that mother's smoking status s_{ij} is ***not exogenous***. Thus, because of the ***confounding*** that could be present, attempting to draw ***causal interpretations*** regarding the effect of mother's smoking on child respiratory status is ill-advised. Accordingly, we specify a ***marginal model*** only recognizing that we can do nothing more than evaluate the ***association*** between mother's ***current smoking status*** and and the probability her child is currently experiencing respiratory problems. A ***causal analysis*** would require the use of specialized techniques for this purpose, as discussed in Section 8.6.

Because the response is ***binary***, $E(Y_{ij}|\boldsymbol{x}_{ij}) = \text{pr}(Y_{ij} = 1|\boldsymbol{x}_{ij}) = f(\boldsymbol{x}_{ij}, \beta)$. We specify directly such a marginal model as

$$\text{logit}\{f(\boldsymbol{x}_{ij}, \beta)\} = \beta_0 + \beta_1 c_i + \beta_2 I(s_{ij} = 0) + \beta_3 I(s_{ij} = 1). \quad (8.84)$$

Thus, for example, the probability that child i is wheezing at age t_{ij} ($Y_{ij} = 1$) if that child is from Kingston and his mother is a heavy smoker at t_{ij} is

$$\frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)},$$

and thus the ***odds*** that such a child would be wheezing are

$$\exp(\beta_0 + \beta_1).$$

If follows that the ***odds ratio*** comparing the odds that a child from Kingston whose mother is a nonsmoker is wheezing at t_{ij} to the odds that a child from Kingston whose mother is a heavy smoker is then $\exp(\beta_2)$. If $\beta_2 < 0$, the odds of wheezing are smaller under a nonsmoking mother. Of course, these should be interpreted as ***purely associational*** statements.

Following the discussion regarding ***unbiasedness*** of the linear generalized estimating equation at the end of Section 8.6, fitting the ***directly-specified marginal model*** (8.84) should probably be implemented using the ***independence working assumption*** for $\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i)$.

To complicate matters ***further***, of the $m = 32$ mother-child pairs, 14 have $(Y_{ij}, \boldsymbol{x}_{ij})$ is ***missing*** at one or more of the ages (times) t_{ij}, and this missingness is ***nonmonotone*** for most of these pairs. As discussed in Section 8.7, unless the missingness mechanism is ***MCAR***, the above associational analysis could be flawed due to possible ***inconsistency*** of the estimator for β used.

Of course, without further information, it is ***difficult*** to make a subject-matter based determination if MCAR is plausible. It ***would be*** possible to investigate, under the assumption that the mechanism is MAR, whether or not it is in fact MCAR by fitting an appropriate models for the missingness (e.g., ***hazard models*** as in Section 8.7) to investigate if missingness is related to wheezing status; this is beyond our scope here. However, as discussed in Section 5.6, it is ***impossible*** to determine from the data if MAR is the true mechanism. Accordingly, analyses based on the ***observed data*** should be viewed with caution.

Even if we are willing to make the assumption of MAR, because the missingness is not ***monotone***, the ***WGEE*** methods at the end of Section 8.7 are not feasible. Other approaches that accommodate nonmonotone patterns must be used. See the instructor's notes for the course "Statistical Methods for Analysis With Missing Data."

Subject to these caveats, fits of (8.84) using SAS and R, which should be viewed as ***purely illustrative***, are on the course website.

8.9 Further results for quadratic equations

As noted in Section 8.3, it is possible to show analytically the equivalence between (8.24), namely,

$$(1/2)\sum_{i=1}^{m}\Big(\{\boldsymbol{Y}_i-\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}^T\boldsymbol{V}_i^{-1}(\beta,\boldsymbol{\xi},\boldsymbol{x}_i)\{\partial/\partial\xi_k\,\boldsymbol{V}_i(\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\beta,\boldsymbol{\xi},\boldsymbol{x}_i)\{\boldsymbol{Y}_i-\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}$$
$$-\operatorname{tr}\left[\boldsymbol{V}_i^{-1}(\beta,\boldsymbol{\xi},\boldsymbol{x}_i)\partial/\partial\xi_k\{\boldsymbol{V}_i(\beta,\boldsymbol{\xi},\boldsymbol{x}_i)\}\right]\Big)=0,\quad k=1,\ldots,r+s, \tag{8.85}$$

and the alternative form (8.25) with $\boldsymbol{Z}(\beta,\boldsymbol{\xi})$ is chosen according to the Gaussian working assumption,

$$\sum_{i=1}^{m}\boldsymbol{E}_i^T(\beta,\boldsymbol{\xi})\boldsymbol{Z}_i^{-1}(\beta,\boldsymbol{\xi})\{\boldsymbol{u}_i(\beta)-\boldsymbol{v}_i(\beta,\boldsymbol{\xi})\}=\boldsymbol{0}, \tag{8.86}$$

where

$$\boldsymbol{u}_i(\beta)=\operatorname{vech}\left[\{\boldsymbol{Y}_i-\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}\{\boldsymbol{Y}_i-\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}^T\right].$$

In practice, squared terms are deleted if the model contains no unknown variance parameters. We do not note this explicitly in the following argument.

To show that (8.85) and (8.86) are in fact the ***same estimating equation*** under the conditions above, it ***suffices*** to show that their kth rows coincide. The kth row of (8.85) may be written, using the identity for quadratic forms $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \text{tr}(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T)$, as

$$(1/2)\sum_{i=1}^{m}\text{tr}\Big[\{\partial/\partial\xi_k\,\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\boldsymbol{Y}_i-\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}\{\boldsymbol{Y}_i-\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}^T\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i) - \{\partial/\partial\xi_k\,\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\Big] = 0. \tag{8.87}$$

Noting that $\boldsymbol{E}_i(\beta,\xi)$ has kth row $\{\partial/\partial\xi_k\,\boldsymbol{v}_i(\beta,\xi)\}^T$, we can write the kth row of (8.86) as

$$\sum_{i=1}^{m}\{\partial/\partial\xi_k\,\boldsymbol{v}_i(\beta,\xi)\}^T\boldsymbol{Z}_i(\beta,\xi)\{\boldsymbol{u}_i(\beta)-\boldsymbol{v}_i(\beta,\xi)\} = 0. \tag{8.88}$$

We can thus show the result by showing that the ith summand in (8.87) is equal to that in (8.88).

We use several matrix results given in Appendix A, which we repeat here for convenience; these can be found with discussion in Chapter 16 of Harville (1997) or in Appendix 4.A of Fuller (1987). For matrices $\boldsymbol{A}$ ($a\times a$), $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$,

(i) $\text{tr}(\boldsymbol{A}\boldsymbol{B}) = \{\text{vec}(\boldsymbol{A})\}^T\{\text{vec}(\boldsymbol{B}^T)\} = \{\text{vec}(\boldsymbol{A}^T)\}^T\{\text{vec}(\boldsymbol{B})\}$.

(ii) $\text{tr}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{D}^T\boldsymbol{C}^T) = \{\text{vec}(\boldsymbol{A})\}^T(\boldsymbol{B}\otimes\boldsymbol{C})\text{vec}(\boldsymbol{D})$, where "$\otimes$" is Kronecker product.

(iii) For $\boldsymbol{A}$ symmetric, there exists a unique matrix $\boldsymbol{\Phi}$ of dimension $\{a^2\times a(a+1)/2\}$ such that

$$\text{vec}(\boldsymbol{A}) = \boldsymbol{\Phi}\text{vech}(\boldsymbol{A}).$$

Clearly, $\boldsymbol{\Phi}$ is unique and of full column rank, as there is only one way to write the distinct elements of $\boldsymbol{A}$ in a full, redundant vector.

There also exist ***many*** (not unique) linear transformations of vec($\boldsymbol{A}$) into vech($\boldsymbol{A}$). Consider a transformation matrix $\boldsymbol{\Psi}$ of dimension $\{a(a+1)/2\times a^2\}$ such that

$$\text{vech}(\boldsymbol{A}) = \boldsymbol{\Psi}\text{vec}(\boldsymbol{A}).$$

One particular choice of $\boldsymbol{\Psi}$ is the Moore-Penrose generalized inverse of $\boldsymbol{\Phi}$, $\boldsymbol{\Psi} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T$. Fuller (1987, page 383) gives the actual form of $\boldsymbol{\Phi}$.

Because we are addressing equivalency under the ***assumption of normality*** , take $\boldsymbol{Y}_i|\boldsymbol{x}_i$ to be normally distributed for the purposes of the argument.

Under these conditions, it is possible to show [see, for example, Fuller (1987, Lemma 4.A.1)] that

$$\text{var}\{\boldsymbol{u}_i(\beta)|\boldsymbol{x}_i\} = 2\boldsymbol{\Psi}\{\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\otimes\{\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\boldsymbol{\Psi}^T. \tag{8.89}$$

In fact, (8.89) is a compact way of expressing (8.23). Result 4.A.3.1 of Fuller (1987, page 385) then yields that

$$\{\text{var}\{\boldsymbol{u}_i(\beta)|\boldsymbol{x}_i\}\}^{-1} = (1/2)\boldsymbol{\Phi}^T\{\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\otimes\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\}\boldsymbol{\Phi}. \tag{8.90}$$

We are now in a position to show the desired correspondence. For brevity, we suppress the arguments of all matrices and vectors. The estimating equation in (8.87) has two parts:

$$(1/2)\text{tr}\{(\partial/\partial\xi_k\boldsymbol{V}_i)\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i-\boldsymbol{f}_i)(\boldsymbol{Y}_i-\boldsymbol{f}_i)^T\boldsymbol{V}_i^{-1}\} \tag{8.91}$$

and

$$-(1/2)\text{tr}\{(\partial/\partial\xi_k\boldsymbol{V}_i)\boldsymbol{V}_i^{-1}\}. \tag{8.92}$$

Consider (8.91). By (ii) above, identifying $\boldsymbol{A} = \partial/\partial\xi_k\boldsymbol{V}_i$, $\boldsymbol{B} = \boldsymbol{V}_i^{-1}$, $\boldsymbol{D}^T = (\boldsymbol{Y}_i-\boldsymbol{f}_i)(\boldsymbol{Y}_i-\boldsymbol{f}_i)^T$, and $\boldsymbol{C}^T = \boldsymbol{V}_i^{-1}$, and using the definition of $\boldsymbol{u}_i$,

$$\begin{aligned}
&(1/2)\text{tr}\{(\partial/\partial\xi_k\boldsymbol{V}_i)\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i-\boldsymbol{f}_i)(\boldsymbol{Y}_i-\boldsymbol{f}_i)^T\boldsymbol{V}_i^{-1}\}\\
&\qquad = (1/2)\{\text{vec}(\partial/\partial\,\xi_k\boldsymbol{V}_i)\}^T(\boldsymbol{V}_i^{-1}\otimes\boldsymbol{V}_i^{-1})\text{vec}\{(\boldsymbol{Y}_i-\boldsymbol{f}_i)(\boldsymbol{Y}_i-\boldsymbol{f}_i)^T\}\\
&\qquad = (1/2)\{\boldsymbol{\Phi}\text{vech}(\partial/\partial\,\xi_k\boldsymbol{V}_i)\}^T(\boldsymbol{V}_i^{-1}\otimes\boldsymbol{V}_i^{-1})\boldsymbol{\Phi}\boldsymbol{u}_i\\
&\qquad = \{\text{vech}(\partial/\partial\,\xi_k\boldsymbol{V}_i)\}^T\{(1/2)\boldsymbol{\Phi}^T(\boldsymbol{V}_i^{-1}\otimes\boldsymbol{V}_i^{-1})\boldsymbol{\Phi}\}\boldsymbol{u}_i,
\end{aligned} \tag{8.93}$$

From the definition of $\boldsymbol{v}_i$,

$$\text{vech}(\partial/\partial\xi_k\,\boldsymbol{V}_i) = \partial/\partial\xi_k\,\text{vech}(\boldsymbol{V}_i) = \partial/\partial\xi_k\boldsymbol{v}_i.$$

Moreover, the "middle" term in (8.93) in braces, by (8.90), equals $\boldsymbol{Z}_i^{-1}$, as we are doing these calculations under normality. Substituting these developments into (8.93) yields

$$\{\partial/\partial\xi_k\boldsymbol{v}_i\}^T\boldsymbol{Z}_i^{-1}\boldsymbol{u}_i. \tag{8.94}$$

Now consider (8.92). Applying (ii) gives

$$\begin{aligned}
&-(1/2)\{\text{vec}(\partial/\partial\,\xi_k\boldsymbol{V}_i)\}^T(\boldsymbol{V}_i^{-1}\otimes\boldsymbol{V}_i^{-1})\text{vec}(\boldsymbol{V}_i)\\
&\qquad = -(1/2)\{\boldsymbol{\Phi}\text{vech}(\partial/\partial\,\xi_k\boldsymbol{V}_i)\}^T(\boldsymbol{V}_i^{-1}\otimes\boldsymbol{V}_i^{-1})\boldsymbol{\Phi}\text{vech}(\boldsymbol{V}_i)\\
&\qquad = -\{\text{vech}(\partial/\partial\,\xi_k\boldsymbol{V}_i)\}^T\{(1/2)\boldsymbol{\Phi}^T(\boldsymbol{V}_i^{-1}\otimes\boldsymbol{V}_i^{-1})\boldsymbol{\Phi}\}\boldsymbol{v}_i\\
&\qquad = -\{\partial/\partial\xi_k\boldsymbol{v}_i\}^T\boldsymbol{Z}_i^{-1}\boldsymbol{v}_i.
\end{aligned} \tag{8.95}$$

Combining (8.94) and (8.95), we obtain that the kth row of the PL summand in (8.87) is equal to the kth row of the GEE summand in (8.88), namely

$$\{\partial/\partial\xi_k \boldsymbol{v}_i\}^T \boldsymbol{Z}_i^{-1}(\boldsymbol{u}_i - \boldsymbol{v}_i),$$

as desired. Of course, it is fact possible to carry out the argument in the reverse direction, starting from (8.88).

The same type of argument can be applied to the second term in the quadratic estimating equation for β, so that the joint normal ML equations may be written in the "GEE-2" form with the Gaussian working assumption.

9 Nonlinear and Generalized Linear Mixed Effects Models

9.1 Introduction

In Chapter 8, we considered ***population-averaged*** models for longitudinal data involving responses that are ***discrete or continuous*** and models for the overall population mean response that may be ***nonlinear*** in parameters. These models are an appropriate framework for addressing questions of scientific interest that focus on the ***overall population***.

- Interest may be in comparing the ***pattern of change*** of the population mean response over time between two treatments for a ***continuous*** response. For example, interest may be in comparing the ***rate of decrease*** of average ***viral load*** for the population HIV-infected subjects were all subjects in the population to receive two different "cocktails" of anti-retroviral therapy.
- It may be of interest to compare the ***odds*** of a positive response in the population of patients if all were to receive a new drug versus the standard treatment over the study period.

Such questions are typically of interest in the context of ***public health*** and the need to make ***public policy recommendations***. For instance, the FDA makes ***regulatory decisions*** based on how a new product performs relative to the standard of care overall in the population; e.g., does it lower the odds of an undesirable health outcome relative to the standard in the population of patients?

In many applications, however, interest naturally focuses instead on ***individual-specific behavior***. A key example we have discussed is that of ***pharmacokinetics***. Here, data on drug concentrations achieved over time are collected on each subject in a sample drawn from a population of interest. However, as discussed in Chapter 1 in ***EXAMPLE 4***, the theophylline pharmacokinetic study, interest is ***not*** in how population mean drug concentration changes over time.

Rather, interest focuses on the underlying, ***within-individual processes*** of absorption, distribution, and elimination of the drug, in particular the "***typical***" or mean values of ***individual-specific*** parameters characterizing these processes in the population and how their values ***vary*** across the population. Here, then, interest is in ***subject-specific*** inference.

Similarly, while the FDA is interested in population-averaged inference for the purpose of making ***broad public policy decisions***, in routine practice a physician may be more interested in the comparison of an ***individual patient's*** odds of having a positive response under two different treatments. Again, this is a ***subject-specific*** question.

In this chapter, we discuss a broad class of ***subject-specific*** models for longitudinal data that are an appropriate framework in which to pose and address such questions, that of ***nonlinear mixed effects models***. As the name implies, these models are appropriate when a model for individual-level behavior that is ***nonlinear*** in ***individual-specific parameters*** is available, and questions of interest can be formulated in terms of the individual-level model and its parameters. Versions of the model accommodate both continuous and discrete longitudinal responses. In particular, ***generalized linear mixed effects models*** are a ***special case*** of this class relevant when the response is of the "generalized linear model type."

9.2 Model specification

DATA, RESTATED, AGAIN: We review the form of the observed data once again. These data are

$$(\boldsymbol{Y}_i, \boldsymbol{z}_i, \boldsymbol{a}_i) = (\boldsymbol{Y}_i, \boldsymbol{x}_i), \quad , i = 1, \ldots, m, \tag{9.1}$$

independent across i, where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$, with Y_{ij} recorded at time t_{ij}, $j = 1, \ldots, n_i$ (possibly different times for different individuals); $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}^T, \ldots, \boldsymbol{z}_{in_i}^T)^T$, comprising ***within-individual*** covariate information $\boldsymbol{u}_i$ and the ***times*** t_{ij}; $\boldsymbol{a}_i$ is a vector of ***among-individual*** covariates; and $\boldsymbol{x}_i = (\boldsymbol{z}_i^T, \boldsymbol{a}_i^T)^T$.

As in Section 8.2, if there are ***within-individual covariates*** $\boldsymbol{u}_i$, these are either ***time-independent***, as in the theophylline pharmacokinetic study we discuss momentarily, where each subject i received a dose D_i at baseline; or are determined according to a ***fixed design***, as in the case where each subject received repeated doses over several ***dosing intervals***, as we discuss shortly.

For now, we take the ***among-individual*** covariates $\boldsymbol{a}_i$ to be ***time-independent***, reflecting, for example, treatment assignment in a randomized study, baseline measures, static characteristics such as gender, and so on, so that the complications associated with ***time-dependent*** covariates discussed in Section 8.6 are not an issue.

We first return to the theophylline pharmacokinetics example to motivate the basic model specification and then consider further examples.

EXAMPLE 4: Pharmacokinetics of theophylline, continued. Recall from Section 1.2 that $m = 12$ subjects were each given a dose D_i of the anti-asthmatic agent theophylline at time 0, where the dose (mg/kg) was scaled to each individual's body weight (kg). Blood samples were taken from each subject at $n_i = 10$ subsequent time points and assayed for ***theophylline concentration*** (mg/L).

Figure 9.1 shows the data on all 12 subjects.

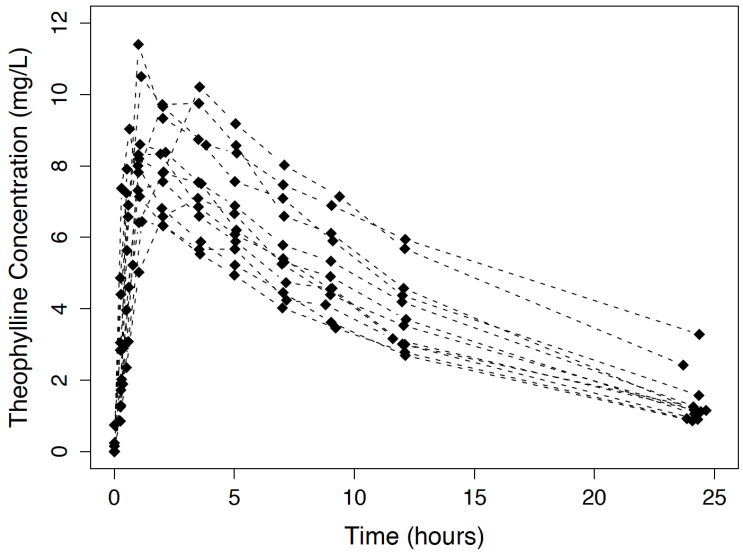

Figure 9.1: *Theophylline concentration-time profiles for $m = 12$ subjects receiving an oral dose of theophylline at time 0.*

As discussed in Sections 1.2 and 2.2 , a ***mechanistic*** , ***theoretical*** model for achieved theophylline concentration $C_i(t)$ at time t following dose D_i at $t = 0$ ***within*** individual i is the ***one-compartment model with first order absorption and elimination*** given by

$$C_i(t) = \frac{k_{ai} D_i}{V_i(k_{ai} - Cl_i/V_i)}\{\exp(-Cl_i\, t_{ij}/V_i) - \exp(-k_{ai} t_{ij})\}, \quad \beta_i = (k_{ai}, Cl_i, V_i)^T, \tag{9.2}$$

where we have taken the bioavailability $F \equiv 1$; and k_{ai}, Cl_i, and V_i are the ***fractional absorption rate*** , ***clearance*** , and ***volume of distribution*** governing individual i's pharmacokinetics (PK). In (9.2), the ***individual-specific PK parameters*** k_{ai}, Cl_i, and V_i characterize the processes of absorption, elimination, and distribution, respectively, taking place ***within*** individual i and thus have ***scientifically meaningful interpretations***.

Recall from Section 1.2 that a ***fundamental principle*** of pharmacokinetics is that these processes, and thus ***PK parameters*** in a ***theoretical model*** like (9.2), ***vary*** among individuals, and it is these differences in the PK parameters that lead to differences in ***steepness, peak, and decay*** of individual-specific concentration-time profiles like those in Figure 9.1.

As noted above, interest thus focuses on gaining insight on the "***typical***" or ***mean*** values of these parameters and the ***extent to which they vary*** in the population based on the observed data.

- More generally, interest focuses on characterizing the (multivariate) ***distribution*** of the PK parameters β_i in the population of interest and its features, including ***mean and/or median values*** of each component of β_i and how the components vary and ***covary*** in the population.

For many drugs, pharmacokinetic properties may be ***systematically associated*** with subject characteristics.

- For example, ***elimination***, in particular ***excretion*** and thus drug ***clearance***, is often associated with ***renal function*** (kidney function) and ***weight***. Subjects who suffer from ***renal impairment*** may clear drug from their bodies ***more slowly*** that those with normal kidney function. Renal impairment is often characterized by a continuous measure, ***creatinine clearance***. Creatinine is a byproduct of muscle metabolism that is removed from the body by the kidneys; if the kidneys are impaired, creatinine is cleared from the body more slowly than if the kidneys are healthy, so that creatinine clearance reflects renal function.
- Likewise, categorical characteristics such as ***gender***, ***age***, ***ethnicity/race***, and whether or not a subject is a ***smoker*** or is in a ***fed or fasting state*** when taking the drug can also be ***associated*** with PK processes.
- In fact, a ***more refined*** goal of a PK study is to evaluate the ***evidence suggesting such associations***. If drug absorption, distribution, and elimination are associated with subject characteristics, this may have implications for ***dosing recommendations***, that is, ***how often*** and in what ***amount*** the drug should be given to achieve the desired ***therapeutic effect*** while keeping the probability of ***adverse side effects*** low. If the PK parameters exhibit considerable ***variation*** across individuals, this makes "one-size-fits-all" dosing recommendations difficult.
- Accordingly, it is of interest to determine how much of the variation in PK parameters in the population can be ***attributed to such systematic associations***.

An excellent review of pharmacokinetics can be found in Giltinan (2014).

The foregoing considerations suggest that an appropriate ***statistical model*** should

- Acknowledge that pharmacokinetic processes take place at the ***individual level*** and thus incorporate the ***within-individual PK model*** (9.2)
- Allow ***individual-specific parameters*** in such a model to have a ***distribution*** that characterizes how they ***vary and covary*** in the population of individuals
- As with any longitudinal data situation, account for ***correlation*** among concentration measures on the same individual due to ***within- and among-individual*** sources.

The general ***nonlinear mixed effects model*** satisfies these requirements.

NONLINEAR MIXED EFFECTS MODEL: The model can be expressed as a ***two-stage hierarchy*** , analogous to that in (6.42) and (6.43) for the ***linear mixed effects model***. The specification here is more general to accommodate typical features that arise in the applications for which the model is relevant. Thus, this model ***subsumes*** the linear mixed effects model as viewed from the ***hierarchical perspective*** as a ***special case***.

It proves convenient, in contrast to the population-averaged models in Chapter 8, where it sufficed to collect all covariates for individual i in $\boldsymbol{x}_i$, to highlight explicitly the ***distinction*** between ***within-*** and ***among-individual*** covariates in stating the general form of each ***stage*** of the ***nonlinear mixed effects model***. Before presenting the general form of the hierarchy, we discuss considerations for each stage.

Stage 1 - Individual model. At the ***first stage*** of the hierarchy, ***individual-level behavior*** is modeled, depending on the data (9.3) on individual i. From (9.1), the available data on i are the pairs

$$(Y_{i1}, \boldsymbol{z}_{ij}), \ldots, (Y_{in_i}, \boldsymbol{z}_{in_i}), \tag{9.3}$$

where $\boldsymbol{z}_{ij}$ incorporates t_{ij} and the conditions $\boldsymbol{u}_i$ under which Y_{ij} was collected.

- For example, in a PK study involving ***multiple dosing intervals*** at which subjects are to be given repeated doses of the drug, suppose individual i received d_i doses $D_{i1}, \ldots, D_{id_i}$ at times $s_{i1}, \ldots, s_{id_i}$ over the study period.

Then his/her entire ***dosing history***, which summarizes the conditions under which his/her response data were collected and thus comprises the ***within-individual covariate*** $\boldsymbol{u}_i$, can be summarized as

$$\boldsymbol{u}_i = \{(s_{i\ell}, D_{i\ell}),\ \ell = 1, \ldots, d_i\}. \tag{9.4}$$

The dosing times $s_{i\ell}$, $\ell = 1, \ldots, d_i$, ordinarily ***do not coincide*** with the sampling times t_{ij}, $j = 1, \ldots, n_i$. As we demonstrate shortly in a specific example, a PK model for drug concentration at time t_{ij} depends on t_{ij} and the dosing history ***up to*** t_{ij},

$$\boldsymbol{u}_{ij} = \{(s_{i1}, D_{i1}), \ldots, (s_{i\ell}, D_{i\ell}),\ \ s_{i\ell} < t_{ij}\}, \tag{9.5}$$

say. Accordingly, it is reasonable to define $\boldsymbol{z}_{ij} = (t_{ij}, \boldsymbol{u}_{ij})$, where $\boldsymbol{u}_{ij}$ is as in (9.5). More generally, one could take $\boldsymbol{z}_{ij} = (t_{ij}, \boldsymbol{u}_i)$ where $\boldsymbol{u}_i$ is the entire dosing history (9.4); however, only $\boldsymbol{u}_{ij}$ in (9.5) is relevant at t_{ij}, as ***future doses*** do not affect the concentration at t_{ij}.

- As above, the ***within-individual covariates*** are summarized as $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}^T, \ldots, \boldsymbol{z}_{in_i}^T)^T$.

With $\boldsymbol{z}_{ij}$ appropriately defined, assume there is a ***model*** f that describes the ***individual-level relationship*** between Y_{ij} and $\boldsymbol{z}_{ij}$ in terms of individual-specific parameters β_i of the form

$$E(Y_{ij}|\boldsymbol{z}_{ij}, \beta_i) = f(\boldsymbol{z}_{ij}, \beta_i), \quad \text{so that} \quad E(\boldsymbol{Y}_i|\boldsymbol{z}_i, \beta_i) = \boldsymbol{f}_i(\boldsymbol{z}_i, \beta_i) = \begin{pmatrix} f(\boldsymbol{z}_{i1}, \beta_i) \\ \vdots \\ f(\boldsymbol{z}_{in_i}, \beta_i) \end{pmatrix}. \tag{9.6}$$

- In the theophylline example, $\boldsymbol{z}_{ij} = (D_i, t_{ij})^T$, and, from (9.2),

$$f(\boldsymbol{z}_{ij}, \beta_i) = \frac{k_{ai} D_i}{V_i(k_{ai} - Cl_i/V_i)}\{\exp(-Cl_i\, t_{ij}/V_i) - \exp(-k_{ai} t_{ij})\}, \quad \beta_i = (k_{ai}, Cl_i, V_i)^T. \tag{9.7}$$

  Because $\boldsymbol{u}_i$ is required ***only*** to specify fully the model (9.7) and is ***not implicated*** in the scientific questions of interest, there are no conceptual complications involved, even if it is time-dependent.

- In (9.6), we ***condition on*** β_i to acknowledge that the relationship depends on individual i's parameters β_i, which are regarded as ***fixed*** at the ***level of the individual*** but are viewed as ***random vectors*** at the population level.

- Indeed, if interest focuses ***only*** on individual i, under suitable assumptions, if n_i is ***sufficiently large***, it would be possible in to fit (9.6), i.e., estimate β_i, based on the data (9.3) on i.

Stage 2 - Population model. The ***second stage*** involves a model for ***population-level behavior*** , which relates the ***among-individual covariates*** $\boldsymbol{a}_i$ to β_i and and implies a ***distributional model*** for β_i given covariates that represents ***variation and covariation*** of the components of β_i in the population, as formalized below.

BASIC MODEL: We consider the following general SS hierarchical nonlinear mixed effects model.

Stage 1 - Individual model. Given a model f as in (9.6), the random vectors $\boldsymbol{Y}_i$, $i = 1, \ldots, m$, are assumed to satisfy

$$E(\boldsymbol{Y}_i|\boldsymbol{z}_i, \beta_i) = E(\boldsymbol{Y}_i|\boldsymbol{z}_i, \boldsymbol{a}_i, \boldsymbol{b}_i) = E(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i) = \boldsymbol{f}_i(\boldsymbol{z}_i, \beta_i) = \boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{a}_i, \beta, \boldsymbol{b}_i) = \boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i),$$

$$\text{var}(\boldsymbol{Y}_i|\boldsymbol{z}_i, \beta_i) = \text{var}(Y_i|\boldsymbol{z}_i, \boldsymbol{a}_i, \boldsymbol{b}_i) = \boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i) = \boldsymbol{R}_i(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{b}_i). \quad (9.8)$$

Here, β_i is a $(k \times 1)$ individual-specific regression parameter characterizing the model $f(\boldsymbol{z}_{ij}, \beta_i)$ for individual behavior, and γ is a vector of ***within-individual covariance parameters***. We say more about the form of the ***within-individual covariance matrix*** $\boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i)$, and more generally the ***distribution*** of $\boldsymbol{Y}_i|\boldsymbol{z}_i, \beta_i$, below.

The expressions $\boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i)$ and $\boldsymbol{R}_i(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{b}_i)$, the ***within-individual conditional mean and covariance matrix*** , as functions of $\boldsymbol{x}_i$, $\boldsymbol{b}_i$, and β, are obtained by substituting the ***Stage 2 population model*** for β_i in (9.8).

Stage 2 - Population model. The individual-specific parameter β_i is assumed to be a function of ***among-individual covariates*** $\boldsymbol{a}_i$, ***fixed effects*** β $(p \times 1)$, and ***random effects*** $\boldsymbol{b}_i$ $(q \times 1)$, namely,

$$\beta_i = \boldsymbol{d}(\boldsymbol{a}_i, \beta, \boldsymbol{b}_i), \quad (9.9)$$

where $\boldsymbol{d}$ is a k-dimensional vector of possibly ***nonlinear*** functions of $\boldsymbol{a}_i$, β, and $\boldsymbol{b}_i$. Note that the ***linear*** population model (6.43)

$$\beta_i = \boldsymbol{A}_i\beta + \boldsymbol{B}_i\boldsymbol{b}_i, \quad (9.10)$$

for design matrices $\boldsymbol{A}_i$ $(k \times p)$ and $\boldsymbol{B}_i$ $(k \times q)$ depending on $\boldsymbol{a}_i$, is a ***special case*** of (9.9). We give examples of population models (9.9) below that demonstrate the advantage of allowing ***nonlinear*** relationships.

In either of (9.9) or (9.10), the ***random effects*** $\boldsymbol{b}_i$ represent ***among-individual variation*** that ***cannot*** be attributed to ***systematic relationships*** of β_i to covariates. Ordinarily, it is implicit that the $\boldsymbol{b}_i$ are ***independent*** of the within-individual covariates $\boldsymbol{z}_i$, as otherwise β_i would be related to $\boldsymbol{z}_i$, which is nonsensical in a "regression model" like (9.6).

As in the linear mixed effects model, $\boldsymbol{b}_i$ ***need not be independent*** of the ***among-individual covariates*** $\boldsymbol{a}_i$. The usual, ***default*** assumption is that the $\boldsymbol{b}_i$ are ***iid*** , i.e.,

$$E(\boldsymbol{b}_i|\boldsymbol{x}_i) = E(\boldsymbol{b}_i|\boldsymbol{a}_i) = E(\boldsymbol{b}_i) = \boldsymbol{0}, \text{var}(\boldsymbol{b}_i|\boldsymbol{x}_i) = \text{var}(\boldsymbol{b}_i|\boldsymbol{a}_i) = \text{var}(\boldsymbol{b}_i) = \boldsymbol{D}. \quad (9.11)$$

As in Chapter 6, the covariance matrix $\boldsymbol{D}$ embodies ***among-individual variation and covariation*** in the population. The iid assumption (9.11) should ***critically evaluated*** for relevance by the data analyst. This assumption (9.11) can be relaxed to allow ***dependence on*** $\boldsymbol{a}_i$, as in the linear case, so that the covariance matrix differs according to values of elements of $\boldsymbol{a}_i$; i.e.,

$$\text{var}(\boldsymbol{b}_i|\boldsymbol{x}_i) = \text{var}(\boldsymbol{b}_i|\boldsymbol{a}_i) = \boldsymbol{D}(\boldsymbol{a}_i). \quad (9.12)$$

The population model is completed by making an assumption on the ***distribution*** of $\boldsymbol{b}_i|\boldsymbol{x}_i$, that is, $\boldsymbol{b}_i|\boldsymbol{a}_i$. The ***usual assumption*** is, as in Chapter 6, that this distribution is ***normal***. The ***popular default*** is to take the $\boldsymbol{b}_i$ to be iid as in (9.11), in which case the assumption is

$$\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}). \quad (9.13)$$

In the sequel, we restrict attention to population models in which $\boldsymbol{b}_i$ are taken to be ***iid and normal*** as in (9.11) and (9.13) in particular ***for the purpose of describing inferential methods***. All of the methods can be ***extended easily*** to accommodate dependence on covariates as in (9.12)

SPECIFICATION OF THE STAGE 2 POPULATION MODEL (9.9)***:*** Pharmacokineticists have long appreciated that the ***marginal distributions*** of PK parameters such as drug clearance and volume of distribution in the population ***do not appear to be normal*** , as is often true for ***biological characteristics***. Rather, it is more plausible that such parameters, which are of course constrained to be ***positive*** , have ***skewed distributions*** with ***positive support*** in the population of individuals.

- If $\boldsymbol{b}_i$ is taken to be ***normally distributed*** , as in (9.13), and if β_i is taken to follow a ***linear population model*** as in (9.10), the result is thus a ***potentially unrealistic*** model for the distribution of PK parameters in the population.
- Accordingly, pharmacokineticists have favored modeling the components of β_i in such a way that, if $\boldsymbol{b}_i$ were approximately normally distributed, the components of β_i would have ***skewed distributions with positive support***.

To illustrate, consider the theophylline study and the model (9.7). ***Among-individual covariates*** weight and creatinine clearance were also recorded at baseline. Letting $\boldsymbol{a}_i = (w_i, c_i)^T$, where w_i (kg) is weight and c_i (ml/min) is creatinine clearance, consider the ***population model*** for clearance Cl_i

$$Cl_i = \exp(\beta_{Cl0} + \beta_{Clw} w_i + \beta_{Clc} c_i + b_{i,Cl}); \tag{9.14}$$

a variation on (9.14) is

$$Cl_i = (\beta_{Cl0} + \beta_{Clw} w_i + \beta_{Clc} c_i) \exp(b_{i,Cl}).$$

These models accommodate the possibility of a ***systematic association*** of clearance with weight and creatinine clearance, with the remaining among-individual variation in clearance that is ***not explained*** by this association represented by the associated ***random effect*** $b_{i,Cl}$. The first model (9.14) further ***enforces positivity*** of Cl_i.

In both models, $b_{i,Cl}$ enters the model in a ***multiplicative***, and thus ***nonlinear***, fashion; if $b_{i,Cl}$ were ***normally*** distributed, then Cl_i would be ***lognormally*** distributed. The dependence on the covariates takes a different functional form in each case; in (9.14), the dependence on ***fixed effects*** is also ***nonlinear***.

An alternative to (9.14) is based on ***reparameterizing*** the PK model (9.7). Write (9.14) as

$$\log Cl_i = \beta_{Cl0} + \beta_{Clw} w_i + \beta_{Clc} c_i + b_{i,Cl}, \tag{9.15}$$

which is ***linear*** in both fixed and random effects. Consider the parameterization of (9.7)

$$f(\boldsymbol{z}_{ij}, \boldsymbol{\beta}_i) = \frac{e^{k_{ai}^*} D_i}{e^{V_i^*}(e^{k_{ai}^*} - e^{Cl_i^*}/e^{V^{**}i})}[\exp\{-(e^{Cl_i^*}/e^{V_i^*})t_{ij}\} - \exp(-e^{k_{ai}^*} t_{ij})], \quad \boldsymbol{\beta}_i = (k_{ai}^*, Cl_i^*, V_i^*)^T. \tag{9.16}$$

In (9.16), $k_{ai}^* = \log k_{ai}$, $Cl_i^* = \log Cl_i$, and $V_i^* = \log V_i$, so that the model is parameterized ***directly*** in terms of the ***logarithms*** of the PK parameters. Parameterizations like (9.16) ***enforce positivity*** of parameters that must be positive to be ***biologically plausible*** and can make model fitting ***more numerically stable***.

In the context of a ***hierarchical model***, alternative parameterizations are also introduced to accommodate the belief that the ***population distribution*** of each parameter is ***skewed with positive support***. Moreover, such a parameterization supports use of a simpler, ***linear*** stage 2 population model; e.g., analogous to (9.15),

$$Cl_i^* = \beta_{Cl0} + \beta_{Clw} w_i + \beta_{Clc} c_i + b_{i,Cl}. \tag{9.17}$$

- The model parameterized as in (9.7) in terms of the PK parameters ***directly*** , with second stage model (9.14), and the model parameterized as (9.16) with a linear second stage model of the form (9.17) are two strategies for achieving the ***same objective***.

- Pharmacokineticists tend to prefer the first approach, while statisticians usually adopt the latter, the rationale being that (9.16) is a ***more stable*** parameterization that might improve practical performance of the inferential methods we discuss in subsequent sections. The associated ***linear population model*** also results in ***simpler*** implementation for some methods.

Another issue is the ***relative magnitudes*** of ***among-individual variation*** of the elements of β_i. As in the linear case in Chapter 6, great disparity in these might dictate an ***approximate*** population model.

After ***systematic variation*** due to associations with covariates is taken into account, the ***remaining variation*** in PK parameters represented by random effects can be of ***considerably different magnitudes***. In loglinear models like (9.15) and (9.17) for $\log Cl_i$, the ***standard deviation*** of the random effect corresponds roughly to the ***coefficient of variation*** (CV) in the population of Cl_i. If, for example, the CVs of Cl_i and V_i are ***much larger*** than that of k_{ai}, ***numerical challenges*** can arise in implementation of the nonlinear mixed effects model using the methods we discuss in this chapter.

Accordingly, it is common to adopt an ***approximate model*** that treats the CV of k_{ai} as ***negligible***, accomplished by specifying a population model involving ***no associated random effect*** for k_{ai}, as in

$$\beta_i = \begin{pmatrix} \log Cl_i \\ \log V_i \\ \log k_{ai} \end{pmatrix} = \begin{pmatrix} 1 & w_i & c_i & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & w_i & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & w_i \end{pmatrix} \begin{pmatrix} \beta_{Cl0} \\ \beta_{Clw} \\ \beta_{Clc} \\ \beta_{V0} \\ \beta_{Vw} \\ \beta_{ka0} \\ \beta_{kaw} \end{pmatrix} + \begin{pmatrix} b_{i,Cl} \\ b_{i,V} \\ 0 \end{pmatrix}.$$

This clearly can be expressed in the form (9.10) (verify).

- This model implies that ***all variation*** in $\log k_{ai}$ in the population can be explained by a ***systematic relationship*** with weight. Obviously, we ***do not strictly believe this*** , but it is an ***approximation*** that facilitates implementation when variation in $\log Cl_i$ and $\log V_i$ is much larger by comparison.

SPECIFICATION OF THE STAGE 1 MODEL (9.8)***:*** Given an appropriate model f for the individual-level ***conditional mean*** $E(\boldsymbol{Y}_i|\boldsymbol{z}_i, \beta_i)$, the individual model is completed by specification of

- ***The within-individual covariance matrix*** $\boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i)$. Analogous to (6.17), assuming independence of the ***within-individual realization*** and ***measurement error processes*** $\boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i)$ can be decomposed as

$$\boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i) = \boldsymbol{R}_{Pi}(\beta_i, \gamma_P, \boldsymbol{z}_i) + \boldsymbol{R}_{Mi}(\beta_i, \gamma_M, \boldsymbol{z}_i), \tag{9.18}$$

  where $\boldsymbol{R}_{Pi}(\beta_i, \gamma_P, \boldsymbol{z}_i)$ is the component due to the ***within-individual realization process***, and $\boldsymbol{R}_{Mi}(\beta_i, \gamma_M, \boldsymbol{z}_i)$ is a ***diagonal matrix*** whose diagonal elements reflect ***within-individual measurement error variance***.

  As in Chapter 7, we allow dependence of the diagonal elements of these components, and thus the ***aggregate within-individual variance***

$$\text{var}(Y_{ij}|\boldsymbol{z}_{ij}, \beta_i),$$

  on β_i. The ***diagonal elements*** of each of $\boldsymbol{R}_{Pi}$ and $\boldsymbol{R}_{mi}$ in (9.18) can be specified by appealing to the considerations discussed in Section 7.2 for univariate response, leading to a model for $\text{var}(Y_{ij}|\boldsymbol{z}_{ij}, \beta_i)$ that takes into account variation due to both the ***within-individual realization process*** and ***measurement error***. Likewise, $\boldsymbol{R}_{Pi}$ can involve a model for possible ***within-individual serial correlation*** due to the realization process, as discussed in (6.2).

- ***The distribution of*** $\boldsymbol{Y}_i|\boldsymbol{z}_i, \beta_i$. This is dictated by the ***application*** and the ***nature of the response***.

We first discuss specification of (9.18); a more detailed account is given by Davidian and Giltinan (2003, Section 2.2.2).

Consider the theophylline study. Here, the response, drug concentration, is ***continuous***. As noted in Sections 2.2 and 7.2, it is well-established that drug concentrations on a given individual do not exhibit ***constant variance***. Rather, as in (7.5), a standard model for the ***aggregate*** within-individual variance is that of ***constant coefficient of variation***, which is assumed to be dominated by ***measurement error***. Thus, a common model for the aggregate within-individual variance is

$$\text{var}(Y_{ij}|\boldsymbol{z}_{ij}, \beta_i) = \sigma^2 f(\boldsymbol{z}_{ij}, \beta_i)^{2\delta}, \tag{9.19}$$

where δ may well be equal to 1. In (9.19), the ***variance parameters*** σ^2 and δ are not taken to depend on i, reflecting the belief that the pattern of within-individual variance should be ***similar*** for all individuals due to the use of a ***common measuring technique*** to ascertain concentrations.

More generally, depending on the application, one can posit a ***variance model*** of the form

$$\text{var}(Y_{ij}|\boldsymbol{z}_{ij}, \beta_i) = \sigma^2 g^2(\beta_i, \delta, \boldsymbol{z}_{ij}), \quad \theta = (\sigma^2, \delta^T)^T, \tag{9.20}$$

analogous to (7.1), chosen based on the considerations discussed in Section 7.2.

- As for (9.19), it is customary to take the ***within-individual variance parameters*** to be the ***same*** for all individuals i, reflecting the belief that the aggregate pattern of within-individual variance due to realization process and measurement error is ***similar*** for all i. This is certainly reasonable for the ***measurement error*** component when the same device or technique is used to ascertain the response.
- Although realization variance could conceivably manifest ***differently*** for different individuals, the assumption of common parameters may be made as an ***approximation*** to achieve ***parsimony*** , as the parameters may be "***similar enough***" across individuals. Estimation of ***individual-specific*** variance parameters could be ***challenging*** , particularly when n_i is ***not large***.
- The variance model (9.20) may be dictated by the ***nature of the response***. For example, if Y_{ij} is ***binary*** , and it is assumed that there is no ***misclassification error*** in ascertaining the values of the Y_{ij}, ***of necessity*** ,

$$\text{var}(Y_{ij}|\boldsymbol{z}_{ij}, \beta_i) = f(\boldsymbol{z}_{ij}, \beta_i)\{1 - f(\boldsymbol{z}_{ij}, \beta_i)\}. \tag{9.21}$$

A ***correlation model*** $\Gamma_i(\alpha, \boldsymbol{z}_i)$ for ***within-individual serial correlation*** in $\boldsymbol{R}_{Pi}(\beta_i, \gamma_P, \boldsymbol{z}_i)$ in (9.18) can also be specified.

- Typically, as we have noted for PK and other applications, it is assumed that such within-individual correlation is ***negligible*** due to the ***intermittent*** nature of data collection, such that responses are ascertained ***sufficiently far apart in time*** that correlation due to the realization process has ***died out***.
- This ***need not*** be the case in general. Ideally, the overall model $\boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{Z}_i)$ should embody ***whatever assumptions*** on within-individual covariance are relevant. However, complex such models can render practical implementation of the model ***computationally challenging*** , in which case such an assumption may be made as an ***approximation***.

- Specification may also be limited by the capability of ***available software***. ***At best*** , most widely available software implementing the methods for fitting these models discussed in subsequent sections allows the within-individual covariance matrix to be of the form

$$\boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i) = \boldsymbol{T}_i^{1/2}(\beta_i, \theta, \boldsymbol{z}_i)\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{z}_i)\boldsymbol{T}_i^{1/2}(\beta_i, \theta, \boldsymbol{z}_i), \tag{9.22}$$

  where $\boldsymbol{T}_i(\beta_i, \theta, \boldsymbol{z}_i)$ is a ***diagonal matrix*** depending on a built-in or user-specified ***variance model*** , and $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{z}_i)$ is one of the "***standard*** " correlation models.

  The form (9.22) with a non-diagonal correlation model $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{z}_i)$ makes the most sense when measurement error is assumed to be ***negligible*** , in which case it is a model for $\boldsymbol{R}_{Pi}$ in (9.18). Alternatively, with $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{z}_i) = \boldsymbol{I}_{n_i}$, (9.22) makes sense as a model for $\boldsymbol{R}_{Mi}$ in (9.18) when the ***realization process*** is taken as negligible ***or*** as a model for the sum $\boldsymbol{R}_{Pi} + \boldsymbol{R}_{Mi}$ when serial correlation is assumed ***negligible***.

  It is generally ***not possible*** to implement more complex models (9.18) without specialized programming.

Specification of the distribution of $\boldsymbol{Y}_i|\boldsymbol{z}_i, \beta_i$ is based on the features of the particular application.

Again consider the theophylline study, which is representative of the considerations in ***pharmacokinetics*** more generally. As drug concentrations must be ***nonnegative*** and mostly likely are ***positive*** in a typical study, a natural specification is the ***lognormal distribution***. However, it is common instead to assume

$$\boldsymbol{Y}_i|\boldsymbol{z}_i, \beta_i \sim \mathcal{N}\{\boldsymbol{f}_i(\boldsymbol{z}_i, \beta_i), \boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i)\}, \tag{9.23}$$

where $\boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i)$ need not be a diagonal matrix as above. Of course, (9.23) implies that

$$Y_{ij}|\boldsymbol{z}_{ij}, \beta_i \sim \mathcal{N}\{f(\boldsymbol{z}_{ij}, \beta_i), \sigma^2 g^2(\beta_i, \delta, \boldsymbol{z}_{ij})\}, \quad j = 1, \ldots, n_i,$$

under a general variance model (9.20).

A justification of the ***normality assumption*** (9.23) is as follows. From the plots of the data in Figures 1.5 and 9.1, it is evident that the ***noise-to-signal ratio*** is very ***small*** , reflecting the ***high quality*** of these data. This is a typical feature of PK data. If the distribution of $Y_{ij}|\boldsymbol{z}_{ij}, \beta_i$ were ***lognormal*** , then of necessity the variance follows the "power model" (9.19) with $\delta = 1$, and the ***scale parameter*** σ is the CV, reflecting "noise-to-signal." It can be shown that, under these conditions, if σ is "***small*** ," the lognormal distribution can be ***approximated*** by a normal distribution (try it).

More generally, as in ***classical univariate regression*** modeling, the normal distribution is often a reasonable model for ***continuous response*** given covariates.

REMARK: When generic reference is made to the ***nonlinear mixed effects model***, as in the linear case, it is ***implicit*** that the distribution of $\boldsymbol{Y}_i|\boldsymbol{z}_i, \beta_i$ is taken to be ***normal***, analogous to the simpler linear case.

For ***other types*** of responses, it is more appropriate to assume a different ***within-individual distributional model*** for $\boldsymbol{Y}_i|\boldsymbol{z}_i, \beta_i$. In particular, for ***certain continuous and discrete responses***, a member of the ***scaled exponential family*** class is an appropriate model for the distribution of $Y_{ij}|\boldsymbol{z}_{ij}, \beta_i$ for each j. This leads to an important ***special case*** of the general model (9.8)-(9.9).

GENERALIZED LINEAR MIXED EFFECTS MODEL: As in Chapter 7, we could consider the broader class of "***generalized (non)linear mixed effects models***;" for simplicity, we restrict presentation here to the classical "linear" case.

If the distribution of $Y_{ij}|\boldsymbol{z}_{ij}, \beta_i$, $j = 1, \ldots, n_i$, is assumed to be a member of the ***scaled exponential family class***, then, analogous to the discussion in Section 7.2 following (7.13), it is natural to posit

$$E(Y_{ij}|\boldsymbol{z}_{ij}, \beta_i) = f(\boldsymbol{z}_{ij}^T \beta_i), \qquad \text{var}(Y_{ij}|\boldsymbol{z}_{ij}, \beta_i) = \sigma^2 g^2\{f(\boldsymbol{z}_{ij}^T \beta_i)\}, \tag{9.24}$$

where $f(\cdot)$ is one of the ***usual*** models, such as the ***logistic*** or ***probit*** for ***binary*** response, ***loglinear*** model for response in the form of a ***count***, and so on; $g^2(\cdot)$ is dictated by the particular scaled exponential family distribution, which may or may not involve an unknown scale parameter σ^2, such as (9.21) for binary response; and $\boldsymbol{z}_{ij}^T \beta_i$ is the ***linear predictor***.

Ordinarily, it is assumed that there is ***no measurement or misclassification error***, so that the $\text{var}(Y_{ij}|\boldsymbol{z}_{ij}, \beta_i)$ in (9.24) is a model for the variance of the ***within-individual realization process***, which is assumed to follow the scaled exponential family at each t_{ij}.

Because there is ***no straightforward multivariate generalization*** of scaled exponential family distributions such as the Bernoulli or Poisson, it is customary to assume that the Y_{ij} given $\boldsymbol{z}_i, \beta_i$, $j = 1, \ldots, n_i$, are ***conditionally independent***. This may or may not be a ***reasonable assumption***, depending on the application, but it is standard and built in to available software, so is often made ***by default***.

The standard ***generalized linear mixed effects model***, typically abbreviated as ***GLMM***, is specified as the following ***two-stage hierarchy***.

Stage 1 - Individual model. Given a model f as in (9.24),, the Y_{ij} are assumed to be ***conditionally independent*** given $\boldsymbol{z}_i, \beta_i$, and $Y_{ij}|\boldsymbol{z}_{ij}, \beta_i$ is assumed to follow a ***scaled exponential family*** distribution with

$$E(Y_{ij}|\boldsymbol{z}_{ij}, \beta_i) = E(Y_{ij}|\boldsymbol{z}_{ij}, \boldsymbol{a}_i, \boldsymbol{b}_i) = E(Y_{ij}|\boldsymbol{x}_i, \boldsymbol{b}_i) = f(\boldsymbol{z}_{ij}^T\beta_i) = f(\boldsymbol{u}_{ij}^T\beta + \boldsymbol{v}_{ij}^T\boldsymbol{b}_i),$$

$$\text{var}(Y_{ij}|\boldsymbol{z}_{ij}, \beta_i) = \text{var}(Y_{ij}|\boldsymbol{z}_{ij}, \boldsymbol{a}_i, \boldsymbol{b}_i) = \text{var}(Y_{ij}|\boldsymbol{x}_i, \boldsymbol{b}_i) = \sigma^2 g^2\{f(\boldsymbol{z}_{ij}^T\beta_i)\} = \sigma^2 g^2\{f(\boldsymbol{u}_{ij}^T\beta + \boldsymbol{v}_{ij}^T\boldsymbol{b}_i),\}, \quad (9.25)$$

where $\boldsymbol{u}_{ij}^T = \boldsymbol{z}_{ij}^T\boldsymbol{A}_i$ and $\boldsymbol{v}_{ij}^T = \boldsymbol{z}_{ij}^T\boldsymbol{B}_i$ by ***substitution*** of the population model below, and the ***within-individual variance function*** g^2 in (9.25) is dictated by the particular scaled exponential family.

Stage 2 - Population model. The individual-specific parameter β_i is assumed to depend on ***among-individual covariates*** $\boldsymbol{a}_i$, ***fixed effects*** β ($p \times 1$), and ***random effects*** $\boldsymbol{b}_i$ ($q \times 1$) through the ***linear population model***

$$\beta_i = \boldsymbol{A}_i\beta + \boldsymbol{B}_i\boldsymbol{b}_i. \quad (9.26)$$

Ordinarily, it is further assumed that the $\boldsymbol{b}_i$ are ***iid normal*** as in (9.13); i.e.,

$$\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}) \quad (9.27)$$

We discuss the generalized linear mixed effects model further in in Sections 9.5 and 9.6.

We consider briefly several more examples that illustrate ***modeling considerations***.

EXAMPLE 3: Growth of two different soybean genotypes, continued. Recall the soybean study in Section 1.2, in which different experimental plots were randomly planted in each of three years with two different soybean genotypes, Forrest (F) and Plant Introduction #416937 (P). The response Y_{ij}, average leaf weight per plant (g), was calculated by sampling each plot approximately weekly (t_{ij}) over the growing season. The goal was to compare ***growth characteristics*** of the genotypes, and especially the ***asymptotic growth*** at the end of the growing season.

Here, there are ***no within-individual covariates*** $\boldsymbol{u}_i$, so that $\boldsymbol{z}_{ij} = t_{ij}$, and the ***among-individual covariates*** are $\delta_i = 0$ (F) or 1 (P) and $w_i = 1, 2, 3$ according to the year, where each year corresponds to a different condition (dry, normal, wet), so that $\boldsymbol{a}_i = (\delta_i, w_i)^T$.

As discussed in Section 1.2, a ***theoretical model*** for the growth process taking place ***within an individual plot*** is the ***logistic growth model*** , which involves ***scientifically meaningful parameters*** reflecting the ***growth characteristics*** of the plot.

Because the response is a ***biological characteristic***, it is natural to suppose that the ***within-individual realization process*** has ***nonconstant variance***. From Figure 1.4, the within-plot "noise-to-signal" appears fairly low, so it might be reasonable to assume that the within-individual distribution is ***normal***, with the following ***individual model***.

$$E(Y_{ij}|\boldsymbol{z}_{ij},\boldsymbol{\beta}_i) = f(\boldsymbol{z}_{ij},\boldsymbol{\beta}_i) = \frac{\beta_{1i}}{1+\beta_{2i}\exp(-\beta_{3i}t_{ij})}, \qquad \text{var}(Y_{ij}|\boldsymbol{z}_{ij},\boldsymbol{\beta}_i) = \sigma^2 f^{2\delta}(\boldsymbol{z}_{ij},\boldsymbol{\beta}_i), \tag{9.28}$$

where the model for $\text{var}(Y_{ij}|\boldsymbol{z}_{ij},\boldsymbol{\beta}_i)$ may be a "compromise" representing the ***aggregate within-individual variance***. It ***may or may not*** be reasonable to assume that ***within-individual correlation*** is negligible. This could be investigated informally by fitting (9.28) to the data on ***each plot*** using the methods in Chapter 7 (assuming within-individual ***independence***), forming ***within-individual weighted residuals***, and using the techniques in Section 2.6.

In the context of (9.28), each individual plot *i* has its ***own growth characteristics*** represented by $\boldsymbol{\beta}_i = (\beta_{1i},\beta_{2i},\beta_{3i})^T$, and β_{1i} in particular represents the ***asymptotic behavior*** achieved in the plot. Thus, it is natural to view the question of interest formally as ***subject-specific*** and ask if there is evidence that the components of $\boldsymbol{\beta}_i$, and especially β_{1i} representing ***plot-specific asymptotic growth***, are ***systematically associated*** with genotype and/or weather.

This suggests ***population models*** that explicitly represent such associations. For example, possible models for β_{1i} are

$$\beta_{1i} = \beta_1 + \beta_2\delta_i + b_{1i} \quad \text{or} \quad \beta_{1i} = \exp(\beta_1 + \beta_2\delta_i + \beta_3 I(w_i=1) + \beta_4 I(w_i=2) + \beta_5 I(w_i=3) + b_{1i}), \tag{9.29}$$

and similarly for β_{2i} and β_{3i}. If ***random effects*** $\boldsymbol{b}_i = (b_{1i},b_{2i},b_{3i})^T$ corresponding to each component of $\boldsymbol{\beta}_i$ are taken to be approximately ***normal***, the second model in (9.29), which ***enforces positivity*** and allows the population distribution of β_{1i} to be ***skewed***, might be more reasonable,

PHARMACOKINETICS OF PHENOBARBITAL IN NEONATES. This is a ***world-famous*** PK example reported by many authors; see Davidian and Giltinan (1995, Section 6.6) and Pinheiro and Bates (2000, Section 6.4). The data are from a study conducted on $m = 59$ preterm infants given phenobarbital for prevention of seizures during the first 16 days after birth. Each infant received an initial, or ***loading***, dose at baseline followed by one or more ***sustaining doses*** by ***intravenous administration***. Thus, as in (9.4), infant *i* receiving a total of d_i doses has ***dose history***

$$\boldsymbol{u}_i = \{(s_{i\ell}, D_{i\ell}),\ \ell = 1,\ldots,d_i\}.$$

Infant 9			*Infant 50*		
time (hrs)	*dose* (μg/kg)	*conc.* (μg/L)	*time* (hrs)	*dose* (μg/kg)	*conc.* (μg/L)
0.0	27.0	–	0.0	20.0	–
1.1	–	22.1	3.0	–	22.2
11.1	3.2	–	12.5	2.5	–
22.3	3.2	–	24.5	2.5	–
34.6	3.2	–	36.5	2.5	–
46.6	3.2	–	48.0	2.5	–
58.7	3.2	–	60.5	2.5	–
70.9	3.2	–	72.5	2.5	–
82.7	–	29.2	81.0	–	30.5
83.2	3.2	–	84.5	2.5	–
94.6	3.2	–	88.0	30.0	–
106.6	3.2	–	89.0	–	67.9
118.6	3.2	–	96.5	2.5	–
130.6	3.2	–	108.5	3.5	–
142.1	–	34.2	120.5	3.5	–
142.6	3.2	–	132.5	3.5	–
312.6	–	19.6	144.5	3.5	–
			157.0	3.5	–
			162.0	–	58.7
Apgar	8		Apgar	6	
weight	1.4 kg		weight	1.1 kg	

Table 9.1: *Data for two infants, pharmacokinetic study of phenobarbital.*

A total of $n_i = 1$ to 6 concentration measurements were obtained from each infant as part of ***routine monitoring*** (these are ***separated substantially in time***, as it is not feasible to draw blood from preterm infants frequently), for a total of $N = 155$ concentrations. On each infant, two ***among-individual covariates***, birthweight w_i (kg) and 5-minute Apgar score a_i, were recorded. Apgar score is an ordinal score taking values from 1 - 10 of the overall physical condition of an infant 5-minutes after birth, where higher scores are better.

Table 9.1 shows the data for two infants. As can be seen in Table 9.1, the dosing times and observation times ***do not coincide***.

The pharmacokinetics of phenobarbital here can be described by the ***one-compartment open model with intravenous administration and first-order elimination***. Following a ***single dose*** $D_{i\ell}$ given at time $s_{i\ell}$, the model states that concentration $C_i(t)$ for individual i at time $t > s_{i\ell}$ is

$$C_i(t) = \frac{D_{i\ell}}{V_i} \exp\left\{-\frac{Cl_i}{V_i}(t - s_{i\ell})\right\},$$

where Cl_i and V_i are phenobarbital ***clearance and volume of distribution*** for infant i.

An assumption that is often reasonable is that PK behavior is ***unchanged regardless*** of the number of doses given, and that achieved concentrations are governed by the ***principle of superposition***, which dictates that a new dose contributes in an ***additive fashion*** to the amount of drug ***already present*** in the system due to previous doses. Under these conditions and the repeated dosing in this study, the concentration achieved at time t following a ***series of doses***

$$(s_{i\ell}, D_{i\ell}),\ \ell : s_{i\ell} < t,$$

is given by a ***sum*** of such terms, namely

$$C_i(t) = \sum_{\ell: s_{i\ell} < t} \frac{D_{i\ell}}{V_i} \exp\left\{-\frac{Cl_i}{V_i}(t - s_{i\ell})\right\}. \tag{9.30}$$

Figure 9.2 shows the data for infant 9 with a fit of (9.30) superimposed and shows the effect of the cumulative doses.

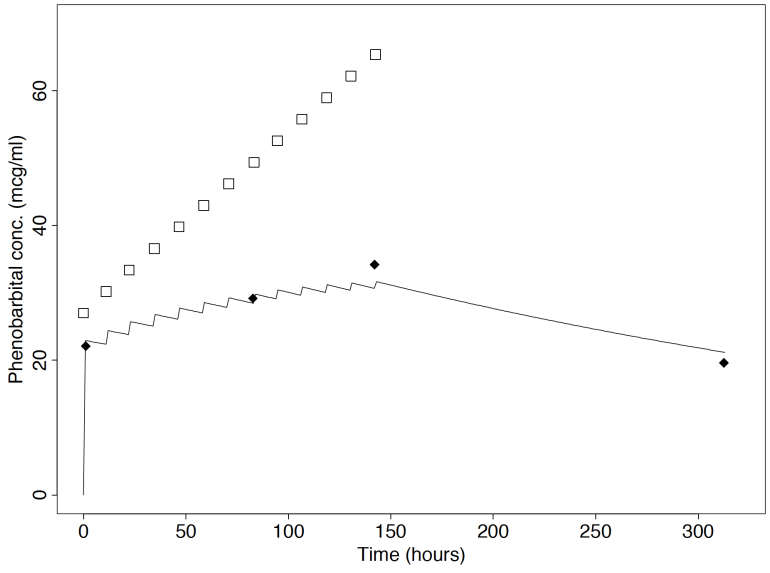

Figure 9.2: *Phenobarbital data for infant 9. The diamonds are phenobarbital concentrations, the open squares represent cumulative dose (μg/kg), and the solid line is a fit of the model* (9.30).

As in any PK study, a goal was to investigate associations between the PK parameters Cl_i and V_i and the ***among-individual covariates*** $\boldsymbol{a}_i = (w_i, a_i)^T$. As with the theophylline study, a reasonable hierarchical model for the phenobarbital study is as follows; this is a model used by Davidian and Giltinan (1995, Section 6.6).

Let $\boldsymbol{z}_{ij} = (t_{ij}, \boldsymbol{u}_{ij})$, where $\boldsymbol{u}_{ij}$ is as in (9.5). From (9.30), ***within-individual behavior*** can be represented as, with $\beta_i = (\log Cl_i, \log V_i)^T$,

$$E(Y_{ij}|\boldsymbol{z}_{ij}, \beta_i) = f(\boldsymbol{z}_{ij}, \beta_i) = \sum_{\ell: s_{i\ell} < t_{ij}} \frac{D_{i\ell}}{e^{\beta_{2i}}} \exp\left\{-\frac{e^{\beta_{1i}}}{e^{\beta_{2i}}}(t_{ij} - s_{i\ell})\right\}, \qquad \text{var}(Y_{ij}|\boldsymbol{z}_{ij}, \beta_i) = \sigma^2 f^{2\delta}(\boldsymbol{z}_{ij}, \beta_i), \quad (9.31)$$

where the distribution of $Y_{ij}|\boldsymbol{z}_{ij}, \beta_i$ might be approximated by a ***normal distribution*** with the moments in (9.31). Because the time intervals between concentration measures are ***large***, it might be reasonable to assume that the Y_{ij} are ***conditionally independent***.

A ***population model*** used by these authors is

$$\beta_{1i} = \beta_1 + \beta_3 w_i + b_{1i}, \qquad \beta_{2i} = \beta_2 + \beta_4 w_i + \beta_5 I(a_i < 5) + b_{2i}, \quad (9.32)$$

where $\boldsymbol{b}_i = (b_{1i}, b_{2i})^T$ might be taken as $\mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$.

REMARK: In the theophylline and soybean study examples, the number of observations n_i on any given individual i is likely ***sufficiently large*** that the ***within-individual model*** could be fitted to each individual ***separately***. For example, the one-compartment model for theophylline PK (9.7), parameterized in this way or as in (9.16), along with a suitable ***variance model***, could in principle be implemented for ***each individual separately*** using the methods in Chapter 7. In contrast, many individuals in the phenobarbital study have $n_i = 1$, 2, or 3, so that individual fitting of (9.31) is not possible. As we demonstrate shortly, it is ***not necessary*** for n_i to be large enough for individual fitting to fit the overall hierarchical nonlinear mixed effects model.

In the foregoing examples, taking a ***subject-specific*** perspective is ***natural***, as the questions of interest unambiguously have to do with how the "typical" values of ***scientifically meaningful*** parameters in ***theoretical models*** for within-individual behavior are associated with individual characteristics and the extent to which they vary. In other settings, whether to take a ***subject-specific*** or ***population-averaged perspective*** can be ***less clear*** and depends on how the questions of interest are ***interpreted***.

EXAMPLE 5: Epileptic seizures and chemotherapy, continued. Recall the epileptic seizure trial, for which the goal was to determine if progabide ***reduces the rate of seizures*** relative to placebo. Recall from Section 8.8 that Y_{ij} is the number of seizures experienced in period j of length o_{ij}, where $j = 1$ corresponds to baseline and $o_{ij} = 8$ and $j = 2, \ldots, 5$ are post-baseline periods with $o_{ij} = 2$. Define as before $v_{ij} = 0$ if $t_{ij} = 0$ and $v_{ij} = 1$ if $t_{ij} > 0$, and let $\delta_i = 0$ for placebo and = 1 for progabide. Thus, $\boldsymbol{z}_{ij} = (o_{ij}, t_{ij})^T$, and $\boldsymbol{x}_i = (\boldsymbol{z}_{ij}^T, \delta_i, a_i)$ (we do not consider age a_i here).

From a ***population-averaged*** point of view, this question is about comparing the ***pattern of change of the seizure rates*** in the population of patients if they were to take placebo to that if they were to take progabide.

Accordingly, in Section 8.8, we considered ***population-averaged*** models such as (8.82), a simplified version of which is

$$E(Y_{ij}|\boldsymbol{x}_i) = \exp(\log o_{ij} + \beta_0 + \beta_1 v_{ij} + \beta_2 \delta_i v_{ij}), \qquad \text{or} \qquad E(Y_{ij}/o_{ij}|\boldsymbol{x}_i) = \exp(\beta_0 + \beta_1 v_{ij} + \beta_2 \delta_i v_{ij}). \quad (9.33)$$

In (9.33), $E(Y_{ij}/o_{ij}|\boldsymbol{x}_i)$ is the ***population seizure rate*** under the conditions in $\boldsymbol{x}_i$, so that $\exp(\beta_0)$ is the baseline ($v_{ij} = 0$) ***population seizure rate***, and $\exp(\beta_2)$ represents the ***ratio*** of the population seizure rate experienced in the post-baseline period ($v_{ij} = 1$) under progabide to that under placebo (verify).

From a ***subject-specific*** perspective, this question is interpreted as having to do with ***individual-level seizure rates***. To make this precise, consider the ***individual model***

$$E(Y_{ij}|\boldsymbol{z}_{ij}, \boldsymbol{\beta}_i) = \exp(\log o_{ij} + \beta_{0i} + \beta_{1i} v_{ij}), \quad (9.34)$$

where it is natural to assume that the ***distribution*** of $Y_{ij}|\boldsymbol{z}_{ij}, \boldsymbol{\beta}_i$ is ***Poisson*** with mean (9.34). In (9.34), $\exp(\beta_{1i})$ is thus the ratio of the ***seizure rate for individual*** i in the post-baseline period to that at baseline.

Consider the ***population model***

$$\beta_{0i} = \beta_0 + b_{0i}, \qquad \beta_{1i} = \beta_1 + \beta_2 \delta_i + b_{1i}, \qquad \boldsymbol{b}_i = (b_{0i}, b_{1i})^T. \quad (9.35)$$

Substituting (9.35) into (9.34) yields

$$\begin{aligned} E(Y_{ij}|\boldsymbol{x}_i, \boldsymbol{b}_i) &= \exp\{\log o_{ij} + (\beta_0 + b_{0i}) + (\beta_1 + \beta_2 \delta_i + b_{1i}) v_{ij})\} \\ &= \exp\{\log o_{ij} + \beta_0 + (\beta_1 + \beta_2 \delta_i) v_{ij} + b_{0i} + b_{1i} v_{ij})\}. \end{aligned} \quad (9.36)$$

- In (9.35), with $E(\boldsymbol{b}_i) = \boldsymbol{0}$ as usual, β_0 is the "***typical***" value of ***log baseline seizure rate*** among individuals in the population; equivalently, the log baseline seizure rate for a "***typical individual***" in the population, defined as one with $\boldsymbol{b}_i = \boldsymbol{0}$, so having the "***typical***" value.

  Thus, $\exp(\beta_0)$ can be interpreted as "***typical***" baseline seizure rate in the population or seizure rate for a "***typical***" ***individual***.

- Similarly, exp(β_0 + β_1) is the "typical" seizure rate/seizure rate for a "typical" individual who received placebo, and exp(β_0 + β_1 + β_2) is that for a "typical" individual receiving progabide.

  Thus, exp(β_2) can be viewed as the ***ratio of seizure rates*** for a "typical" individual in the population under progabide and placebo. In fact, exp(β_2) can ***also*** be viewed as the ratio of seizure rates for two individuals with the ***same value*** of $\boldsymbol{b}_i$ if one of them received placebo and the other progabide.

Thus, exp(β_2) in (9.36) has a decidedly ***different interpretation*** from exp(β_2) in (9.33). The latter is the ratio of the rates of seizures experienced in the ***entire population*** under the two treatments. The former is the ratio of the rates of seizures experienced by a typical ***individual*** (or individuals who share the same ***propensities for seizures*** at baseline and under treatment, reflected by having the ***same*** values of b_{0i} and b_{1i}). Clearly, as noted at the beginning of this chapter, this interpretation may be ***more relevant to a clinician*** deciding how to treat an individual patient.

This leads to a more general discussion of the contrast between the ***hierarchical nonlinear mixed effects model*** (9.8)-(9.9) and the ***population-averaged model*** (8.3) considered in Chapter 8. In contrast to the case of the ***linear mixed effects model*** , which of course is ***subsumed*** by the model here, the two modeling approaches ***do not lead*** to models that coincide.

POPULATION-AVERAGED VERSUS SUBJECT-SPECIFIC PERSPECTIVE: As we have noted, the ***linear mixed effects model*** discussed in Chapter 6 is a ***special case*** of the general nonlinear mixed effects model (9.8)-(9.9). In particular, from (6.42)-(6.43), this model is

$$E(\boldsymbol{Y}_i|\boldsymbol{z}_i,\beta_i) = \boldsymbol{f}_i(\boldsymbol{z}_i,\beta_i) = \boldsymbol{C}_i\beta_i, \quad \text{var}(\boldsymbol{Y}_i|\boldsymbol{z}_i,\beta_i) = \boldsymbol{R}_i(\gamma,\boldsymbol{z}_i), \tag{9.37}$$

$$\beta_i = \boldsymbol{d}(\boldsymbol{a}_i,\beta,\boldsymbol{b}_i) = \boldsymbol{A}_i\beta + \boldsymbol{B}_i\boldsymbol{b}_i, \qquad E(\boldsymbol{b}_i) = \boldsymbol{0}, \ \text{var}(\boldsymbol{b}_i) = \boldsymbol{D}, \tag{9.38}$$

where we highlight possible dependence of $\boldsymbol{R}_i(\gamma,\boldsymbol{z}_i)$ on $\boldsymbol{z}_i$; and, in the usual formulation, $\boldsymbol{R}_i(\gamma,\boldsymbol{z}_i)$ ***does not depend*** on β_i. It follows from (9.37)-(9.38), as argued in Section 6.2 under ***normality*** , that

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i,\boldsymbol{b}_i) = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i, \quad \boldsymbol{X}_i = \boldsymbol{C}_i\boldsymbol{A}_i, \ \boldsymbol{Z}_i = \boldsymbol{C}_i\boldsymbol{B}_i, \qquad \text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i,\boldsymbol{b}_i) = \boldsymbol{R}_i(\gamma,\boldsymbol{z}_i).$$

Thus, the ***overall population-averaged mean*** is given by

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = E\{E(\boldsymbol{Y}_i|\boldsymbol{x}_i,\boldsymbol{b}_i)|\boldsymbol{x}_i\} = E(\boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i|\boldsymbol{x}_i) = \boldsymbol{X}_i\beta. \tag{9.39}$$

Moreover, using the relationship $\text{var}(\boldsymbol{Z}) = E\{\text{var}(\boldsymbol{Z}|\boldsymbol{V})\} + \text{var}\{E(\boldsymbol{Z}|\boldsymbol{V})\}$ for random vectors $\boldsymbol{Z}$ and $\boldsymbol{V}$,

$$\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = E\{\boldsymbol{R}_i(\gamma, \boldsymbol{z}_i)|\boldsymbol{x}_i\} + \text{var}(\boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i|\boldsymbol{x}_i) = \boldsymbol{R}_i(\gamma, \boldsymbol{z}_i) + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T. \tag{9.40}$$

Thus, although motivated from a SS point of view, the linear mixed effects model ***implies*** a ***linear population-averaged model*** as in (9.39), with ***induced overall covariance structure*** (9.40).

- The key result is that the fixed effects β ***can be interpreted*** from ***either*** a subject-specific or population-averaged perspective; both interpretations are ***valid***.
- Thus, a ***population-averaged perspective*** on the linear mixed effects model is that is it a mechanism by which to induce a ***rich and flexible*** model for the ***overall, population-averaged covariance structure***.

As we now demonstrate, these features ***do not*** carry over to the ***nonlinear*** case. Analogous to (9.39), for ***arbitrary nonlinear*** individual conditional mean model and population model

$$E(\boldsymbol{Y}_i|\boldsymbol{z}_i, \beta_i) = \boldsymbol{f}_i(\boldsymbol{z}_i, \beta_i), \qquad \beta_i = \boldsymbol{d}(\boldsymbol{a}_i, \beta, \boldsymbol{b}_i), \qquad E(\boldsymbol{b}_i) = \boldsymbol{0}, \;\; \text{var}(\boldsymbol{b}_i) = \boldsymbol{D},$$

substituting for β_i,

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i) = \boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i), \qquad \text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i) = \boldsymbol{R}_i(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{b}_i) \tag{9.41}$$

are possibly ***nonlinear functions*** of $\boldsymbol{b}_i$.

The ***implied overall population mean*** is thus

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = E\{E(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i)|\boldsymbol{x}_i\} = E\{f(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i)|\boldsymbol{x}_i\} = \int \boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i)\, dF_b(\boldsymbol{b}_i), \tag{9.42}$$

where F_b is the distribution function of the iid $\boldsymbol{b}_i$.

- If f is a ***nonlinear*** function of $\boldsymbol{b}_i$ (by virtue of being nonlinear in β_i), it is ***highly unlikely*** that the integral in (9.42) can be ***evaluated analytically*** , even if the distribution F_b is ***normal*** ; i.e., the integral cannot be obtained in a ***closed form***.
- Moreover, it is ***almost certainly not*** the case that

  $$\int \boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i)\, dF_b(\boldsymbol{b}_i) = \boldsymbol{f}_i(\boldsymbol{x}_i, \beta);$$

  that is, it is highly unlikely that, if not ***impossible*** for, the solution to the integral to be of the ***same form*** f if f is ***nonlinear*** in $\boldsymbol{b}_i$.

Thus, ***in contrast to*** the linear case, the nonlinear mixed effects model ***does not*** in general imply a population-averaged model of the ***same functional form***.

- Instead, as (9.42) shows, the model implies a population-averaged overall mean model that has a ***different, complex form*** that may not be possible to express analytically.

- Accordingly, if we posit ***directly*** a ***population-averaged*** model in terms of a function f,

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{f}_i(\boldsymbol{x}_i, \beta), \tag{9.43}$$

say, the ***interpretation*** of β in (9.43) is ***different from*** that of β in (9.41) using the same f.

Thus, if one uses the ***same*** function f to model the individual conditional mean response and the population mean responses, the two approaches ***do not lead to*** the same ***population-averaged model***.

- Likewise, the ***implied*** overall aggregate covariance structure is given by

$$\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = E\{\boldsymbol{R}_i(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{b}_i)|\boldsymbol{x}_i\} + \text{var}\{\boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i)|\boldsymbol{x}_i\}, \tag{9.44}$$

Both components of (9.44) involve ***integrals*** that are almost certainly ***analytically intractable***.

Thus, the ***overall covariance structure*** implied by the hierarchical model is likely ***not available*** in a ***closed form***. Nonetheless, the first term in (9.44) evidently represents the contribution from ***within-individual sources*** , while the second represents that from ***among-individual sources*** and is almost certainly ***not*** a diagonal matrix (convince yourself).

RESULT: When ***nonlinear*** models are involved, as is the case in ***numerous*** scientific areas and when the response is of the "generalized linear model type" (or more generally), ***great care*** must be taken in identifying the appropriate ***inferential perspective*** ; that is, which of a ***subject-specific*** or ***population-averaged*** point of view is more relevant to the questions of interest. As the epileptic seizure study example above illustrates, the ***interpretation*** of the model and of β in particular is ***different*** depending on which modeling approach is adopted.

The models in this chapter are appropriate when the questions of interest are best formulated and interpreted from a ***subject-specific*** perspective.

REMARK: A popular, but ***misguided***, view of nonlinear mixed effects models, and GLMMS in particular, among many (often non-statistician) practitioners, is that the introduction of random effects is ***mainly*** a way to take correlation among the components of a response vector into account "***automatically***" without having to specify a model for it directly, as in the models of Chapter 8. This is accompanied by the mistaken impression that inferences and their interpretation ***are not impacted*** by this approach.

Clearly, the foregoing discussion invalidates this naive point of view. Indeed, the entire ***interpretation*** of these models and their parameters is ***different***.

9.3 Maximum likelihood

Given a specified nonlinear or generalized linear mixed effects model (9.8)-(9.9) and the goal of ***subject-specific inference*** within its context, the main objective is to estimate β and $\boldsymbol{D}$, as these parameters characterize the "***typical***" behavior of individual-specific parameters and how these parameters ***vary and covary*** in the population, including both ***systematic*** variation due to relationships to covariates and "unexplained" ***inherent*** population variation.

The full model (9.8)-(9.9) also involves the within-individual covariance parameter γ, which includes within-individual ***variance parameters*** θ as well as possible within-individual ***correlation parameters*** α, say. As in Chapter 6, let $\xi = \{\gamma^T, \text{vech}(\boldsymbol{D})^T\}^T$ denote the collection of all variance and covariance parameters.

MAXIMUM LIKELIHOOD: The data-analytic objective is to estimate (β^T, ξ^T). Intuitively, to make progress, it seems critical from (9.42) that we need to make an assumption on the distribution F_b of the $\boldsymbol{b}_i$ (or be able to ***estimate*** this distribution somehow). In general, let

$$p(\boldsymbol{b}_i; \boldsymbol{D})$$

denote the density of the assumed distribution of $\boldsymbol{b}_i$. As noted in the previous section, the ***usual assumption*** is that

$$\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D});$$

however other models are also possible.

As suggested by the developments in the previous section, we can also make an appropriate assumption on the ***distribution*** of $\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i$; the usual assumptions are to take this to be ***normal*** for continuous responses or; assuming ***independence*** of the elements Y_{ij} of $\boldsymbol{Y}_i$, to take $Y_{ij}|\boldsymbol{x}_i, \boldsymbol{b}_i$ for $j = 1, \ldots, n_i$ to follow one of the ***scaled exponential family*** distribution for responses that are in that class. Let the density of the assumed distribution of $\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i$ be

$$p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma), \tag{9.45}$$

where, by substitution of β_i, (9.45) depends on β and γ.

Letting as usual $\boldsymbol{Y}$ denote the "stacked" vector of the $\boldsymbol{Y}_i$, $i = 1, \ldots, m$, and $\tilde{\boldsymbol{x}}$ the collection of the $\boldsymbol{x}_i$, $i = 1, \ldots, m$, the conditional density of the observed data $\boldsymbol{Y}$ given the covariates $\tilde{\boldsymbol{x}}$ is

$$p(\boldsymbol{y}|\tilde{\boldsymbol{x}}; \beta, \gamma, \boldsymbol{D}) = \prod_{i=1}^{m} p(\boldsymbol{y}_i|\boldsymbol{x}_i; \beta, \gamma, \boldsymbol{D}), \tag{9.46}$$

using the ***independence*** of $(\boldsymbol{Y}_i, \boldsymbol{x}_i)$, $i = 1, \ldots, m$. The conditional density $p(\boldsymbol{y}_i|\boldsymbol{x}_i; \beta, \gamma, \boldsymbol{D})$ of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ can be written as

$$\begin{aligned} p(\boldsymbol{y}_i|\boldsymbol{x}_i; \beta, \gamma, \boldsymbol{D}) &= \int p(\boldsymbol{y}_i, \boldsymbol{b}_i|\boldsymbol{x}_i; \beta, \gamma, \boldsymbol{D})\, d\boldsymbol{b}_i = \int p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)\, p(\boldsymbol{b}_i|\boldsymbol{x}_i; \boldsymbol{D})\, d\boldsymbol{b}_i \\ &= \int p(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)\, p(\boldsymbol{b}_i; \boldsymbol{D})\, d\boldsymbol{b}_i, \end{aligned} \tag{9.47}$$

using the iid assumption for the $\boldsymbol{b}_i$. Substituting (9.47) in (9.46),

$$p(\boldsymbol{y}|\tilde{\boldsymbol{x}}; \beta, \gamma, \boldsymbol{D}) = \prod_{i=1}^{m} \int p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)\, p(\boldsymbol{b}_i; \boldsymbol{D})\, d\boldsymbol{b}_i. \tag{9.48}$$

From (9.48), the ***loglikelihood*** for (β, ξ) is then

$$\ell(\beta, \xi) = \log \left\{ \prod_{i=1}^{m} \int p(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)\, p(\boldsymbol{b}_i; \boldsymbol{D})\, d\boldsymbol{b}_i \right\}. \tag{9.49}$$

Ideally, (β, ξ) can be estimated by maximizing (9.49) in these parameters.

- The obvious practical challenge is that, as before, with ***nonlinear*** models, the m q-dimensional integrals in (9.49) are ***analytically intractable*** in all but the simplest situations. Thus, a means of evaluating these integrals, either ***numerically*** or some other way, is required.
- For example, one approach would be to implement a ***numerical integration*** approach such as ***Gaussian quadrature*** to "do" the integrals numerically.

Quadrature rules rely on a ***deterministic approximation*** to an integral as a ***weighted sum*** of the integrand evaluated at a specially chosen set of values, or ***abscissæ***, where the weights are also specially chosen; a full description is given in Monahan (2001, Chapter 10). The approximation thus requires the integrand to be evaluated at each abscissa, and these values are weighted and summed.

The ***accuracy*** of the approximation is predicated on the number of abscissæ, L, say, which is chosen by the user. The ***more*** abscissæ, the ***better*** the approximation. For $q = 1$, it is not too difficult computationally to carry out such numerical integration, as the abscissæ need only be chosen in ***one dimension***. The approximation often works well for L as small as 5 or 10. However, for $q > 1$, abscissæ must be chosen in ***each dimension***, and the integrand must be evaluated at ***each combination***; e.g., for $q = 3$ and $L = 10$, there are $10^3 = 1000$ function evaluations to perform. Thus, for larger q, the ***computational challenge increases*** substantially.

- This might not be a big deal if (β, ξ) is ***known*** and a single evaluation of the integrals in (9.49) is all that is required. However, maximization of (9.49) via standard ***iterative optimization*** techniques such as Newton-Raphson requires ***repeated evaluations*** of (9.49) at each iteration, each of which involves evaluation of m q-dimensional integrals. Obviously, this is potentially ***computationally intensive***, and there is a ***trade-off*** between reducing L to ameliorate this and accuracy of the approximation.

- An alternative approach is ***Monte Carlo integration***, which involves a ***stochastic approximation*** of the integrals. The integral in (9.48) can be viewed as the ***expected value*** of $p(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)$ with respect to the distribution corresponding to $p(\boldsymbol{b}_i; \boldsymbol{D})$. A natural approximation to this expected value is to draw a sample $\boldsymbol{b}_i^{(\ell)}$, $\ell = 1, \ldots, L$, from $p(\boldsymbol{b}_i; \boldsymbol{D})$ and approximate the integral as

$$L^{-1} \sum_{\ell=1}^{L} p(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i^{(\ell)}; \beta, \gamma).$$

  Ordinarily, L must be fairly ***large*** to achieve acceptable accuracy. As with quadrature, this sampling scheme must be carried out for each $i = 1, \ldots, m$ ***repeatedly*** at each iteration of optimization algorithm.

  Importance sampling is another stochastic approximation method that is advantageous when it is difficult to sample from $p(\boldsymbol{b}_i; \boldsymbol{D})$; this is beyond our scope here.

- In the early 1980s when nonlinear mixed effects models were first being used widely, limited computing power rendered these and other numerical integration approaches ***too burdensome*** for routine application. This led to development of methods to ***approximate*** the loglikelihood (9.49) ***analytically*** by a ***closed form*** expression, effectively "doing" the integrals. These methods are still in widespread use today and are reviewed in the next two sections.

- With ***modern computing***, the computational burden associated with maximizing (9.49) is ***much less ominous*** than it was in the 1980s, and numerical integration methods are incorporated in standard software. In Section 9.6, we briefly discuss numerical integration and other approaches to "exact" likelihood inference.

- However, another issue, ***regardless*** of computing power, is that the loglikelihood (9.49) is often a ***highly nonlinear function*** of the parameters and is replete with ***local maxima***, making the optimization problem ***challenging***. This is true even if the loglikelihood is ***approximated*** analytically by a closed form expression. Accordingly, it is often recommended, even in the case of analytic approximations to (9.49), to ***repeat*** the optimization numerous times over a ***grid*** or ***sample*** of starting values for (β, ξ) to establish the ***true*** maximum.

EMPIRICAL BAYES ESTIMATION: As for linear mixed effects models in Section 6.4, it is of interest to "***estimate***" the $\boldsymbol{b}_i$ and more generally the β_i, particularly as the latter often have ***scientific meaning*** in the context of a theoretical model f for individual behavior. As in Section 6.4, once estimates $\widehat{\beta}$ and $\widehat{\xi}$ are available, a natural approach is to ***maximize*** in $\boldsymbol{b}_i$ the ***posterior density***

$$p(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{x}_i; \beta, \gamma, \boldsymbol{D}) = \frac{p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)\, p(\boldsymbol{b}_i; \boldsymbol{D})}{p(\boldsymbol{y}_i|\boldsymbol{x}_i; \beta, \gamma, \boldsymbol{D})}, \tag{9.50}$$

evaluated at $\widehat{\beta}$ and $\widehat{\xi}$ for each $i = 1, \ldots, m$ where from (9.47),

$$p(\boldsymbol{y}_i|\boldsymbol{x}_i; \beta, \gamma, \boldsymbol{D}) = \int p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)\, p(\boldsymbol{b}_i; \boldsymbol{D})\, d\boldsymbol{b}_i.$$

The resulting ***posterior mode*** $\widehat{\boldsymbol{b}}_i$ is the ***empirical Bayes estimator*** for $\boldsymbol{b}_i$.

In contrast to the case of the linear mixed effects model under the usual normality assumptions, it is generally ***not possible*** to obtain the posterior mode in a ***closed form*** as in (6.54). This is true in a nonlinear mixed effects model even if ***both densities*** $p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)$ and $p(\boldsymbol{b}_i; \boldsymbol{D})$ are ***normal***. However, it is typically straightforward to maximize (9.50) in $\boldsymbol{b}_i$, particularly when q is fairly small.

Once $\widehat{\boldsymbol{b}}_i$ are obtained, it is customary to "***estimate***" the β_i by substitution into the population model (9.9); i.e.,

$$\widehat{\beta}_i = \boldsymbol{d}(\boldsymbol{a}_i, \widehat{\beta}, \widehat{\boldsymbol{b}}_i). \tag{9.51}$$

- As with the linear mixed effects model, the $\widehat{\boldsymbol{b}}_i$ and $\widehat{\beta}_i$ are often used for ***diagnostic purposes*** to identify "unusual" individuals or groups of individuals or to assess the relevance of the assumption of ***normality*** of the $\boldsymbol{b}_i$.

- Likewise, the $\widehat{\beta}_i$ are used to characterize ***individual-specific*** conditional means and as estimates for individual-specific parameters β_i, such as ***PK parameters***. In this application, $\widehat{\beta}_i$ may be used to ***simulate*** subject *i*'s ***expected achieved drug concentrations*** under ***different dosing strategies***.

- A popular approach to ***building population models***, so identifying ***covariate relationships*** with the components of β_i, is based on the $\widehat{\boldsymbol{b}}_i$. An initial fit of a nonlinear mixed effects with ***no covariates*** in the population model is carried out. The $\widehat{\boldsymbol{b}}_i$ from this fit are ***plotted*** against ***among-individual*** covariates. ***Systematic patterns*** in these plots may indicate relationships that should be ***incorporated*** in a refined population model $\boldsymbol{d}(\boldsymbol{a}_i, \beta, \boldsymbol{b}_i)$ and the forms of the components of this model.

- A ***caveat*** to these uses of the the empirical Bayes estimates $\widehat{\boldsymbol{b}}_i$ and $\widehat{\beta}_i$ is the tendency for ***shrinkage*** to ***distort*** visual impressions and relationships.

The methods in Section 9.5, which invoke various ***analytic approximations*** to the integrals in the loglikelihood (9.49), take advantage of the $\widehat{\boldsymbol{b}}_i$ to improve the approximation. Likewise, the $\widehat{\boldsymbol{b}}_i$ play a role in ***quadrature methods*** discussed in Section 9.6.

9.4 Approximate inference based on individual estimates

When n_i for each individual $i = 1, \ldots, m$ is ***sufficiently large*** to support fitting the individual model (9.8) ***separately*** by individual to obtain reasonable individual-specific estimators for β_i, an ***intuitively appealing*** approach that circumvents the difficulties with the loglikelihood is to use these estimators as "data" for estimation of (β, ξ). Clearly, this requires $n_i \geq k$ for all i, where, practically speaking, n_i must be ***much larger*** than k to obtain reliable individual estimators.

The basic idea is as follows.

For each $i = 1, \ldots, m$, using the data $(Y_{i1}, \boldsymbol{z}_{i1}), \ldots, (Y_{in_i}, \boldsymbol{z}_{in_i})$, fit the within-individual model (9.8),

$$E(\boldsymbol{Y}_i|\boldsymbol{z}_i, \beta_i) = \boldsymbol{f}_i(\boldsymbol{z}_i, \beta_i), \quad \text{var}(\boldsymbol{Y}_i|\boldsymbol{z}_i, \beta_i) = \boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i)$$

using standard methods for univariate response, e.g., GLS, to obtain the ***individual estimators*** $\widehat{\beta}_i$, $i = 1 \ldots, m$. Note that we use the symbol $\widehat{\beta}_i$ differently here from its use as the empirical Bayes estimator in the previous section.

When $\boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i)$ is ***diagonal*** , this can be carried out using the methods in Chapter 7; extension to ***non-diagonal*** specifications is possible. We discuss this in more detail momentarily.

In restricting attention to each i separately, this fitting is ***conditional*** on β_i, so that β_i is viewed as a ***fixed parameter***. For n_i "large," then, the usual ***asymptotic theory*** for univariate response models dictates that $\widehat{\beta}_i$ ***conditional*** on β_i is ***approximately normal*** with mean β_i and a covariance matrix $\boldsymbol{\Sigma}_i$ that depends on β_i and other (variance and possibly correlation) parameters that can be estimated by substituting $\widehat{\beta}_i$ and estimators for the other parameters. Letting $\widehat{\boldsymbol{\Sigma}}_i$ denote this estimated covariance matrix, we can state this formally as

$$\widehat{\beta}_i|\beta_i, \boldsymbol{z}_i \dot{\sim} \mathcal{N}(\beta_i, \widehat{\boldsymbol{\Sigma}}_i).$$

This result is ***independent*** of $\boldsymbol{a}_i$, so we can write this equivalently as being conditional on $\boldsymbol{x}_i$, namely,

$$\widehat{\beta}_i|\beta_i, \boldsymbol{x}_i \dot{\sim} \mathcal{N}(\beta_i, \widehat{\boldsymbol{\Sigma}}_i). \tag{9.52}$$

In the following developments, $\widehat{\boldsymbol{\Sigma}}_i$, the estimated covariance matrix of the ***approximate sampling distribution*** (9.52), is treated as ***fixed and known*** , as $\widehat{\beta}_i$ and estimators for other unknown parameters have been substituted.

From (9.52), then,

$$E(\widehat{\beta}_i|\beta_i, \boldsymbol{x}_i) \approx \beta_i, \qquad \text{var}(\widehat{\beta}_i|\beta_i, \boldsymbol{x}_i) \approx \widehat{\boldsymbol{\Sigma}}_i. \tag{9.53}$$

Consider the ***linear*** population model (9.10),

$$\beta_i = \boldsymbol{A}_i\beta + \boldsymbol{B}_i\boldsymbol{b}_i,$$

where the $\boldsymbol{b}_i$ are iid with $E(\boldsymbol{b}_i) = \boldsymbol{0}$ and $\text{var}(\boldsymbol{b}_i) = \boldsymbol{D}$.

It follows from (9.53) by substitution of the population model that

$$E(\widehat{\beta}_i|\boldsymbol{x}_i) = E\{E(\widehat{\beta}_i|\beta_i, \boldsymbol{x}_i)|\boldsymbol{x}_i\} \approx E(\beta_i|\boldsymbol{x}_i) = E(\boldsymbol{A}_i\beta + \boldsymbol{B}_i\boldsymbol{b}_i|\boldsymbol{x}_i) = \boldsymbol{A}_i\beta, \tag{9.54}$$

$$\begin{aligned} \text{var}(\widehat{\beta}_i|\boldsymbol{x}_i) &= \text{var}\{E(\widehat{\beta}_i|\beta_i, \boldsymbol{x}_i)|\boldsymbol{x}_i\} + E\{\text{var}(\widehat{\beta}_i|\beta_i, \boldsymbol{x}_i)|\boldsymbol{x}_i\} \\ &\approx \text{var}(\boldsymbol{A}_i\beta + \boldsymbol{B}_i\boldsymbol{b}_i|\boldsymbol{x}_i) + E(\widehat{\boldsymbol{\Sigma}}_i|\boldsymbol{x}_i) = \boldsymbol{B}_i\boldsymbol{D}\boldsymbol{B}_i^T + \widehat{\boldsymbol{\Sigma}}_i. \end{aligned} \tag{9.55}$$

The approximate moments in (9.54) and (9.55) have the form of those of a ***population-averaged linear model*** for "response vectors" $\widehat{\beta}_i$ conditional on $\boldsymbol{x}_i$. This suggests that β and $\boldsymbol{D}$ can be estimated using GEEs as in Chapter 8.

Alternatively, thinking of (9.53) as an ***approximate linear stage 1 individual model*** with linear stage 2 population model as in (9.37) and (9.38) with $\boldsymbol{C}_i = \boldsymbol{I}_k$, it is natural to express (9.53) as

$$\widehat{\beta}_i \approx \beta_i + \boldsymbol{e}_i^* = \boldsymbol{A}_i\beta + \boldsymbol{B}_i\boldsymbol{b}_i + \boldsymbol{e}_i^*, \qquad \boldsymbol{e}_i^* \dot{\sim} \mathcal{N}(\boldsymbol{0}, \widehat{\boldsymbol{\Sigma}}_i), \tag{9.56}$$

where we continue to treat the $\widehat{\boldsymbol{\Sigma}}_i$, $i = 1, \ldots, m$, as known matrices. The representation (9.56) of course leads to (9.54) and (9.55). Moreover, (9.56) has the form of a ***linear mixed effects model*** where the covariance matrix of the within-individual deviation $\boldsymbol{e}_i^*$ is ***known***. This suggests that it might be possible to "fit" (9.56) to estimate β and $\boldsymbol{D}$ using ***linear mixed effects model software***.

The foregoing developments lead to several approaches to estimation of β and $\boldsymbol{D}$ that have been proposed in the nonlinear mixed effects and pharmacokinetics literature, presented next.

REMARK: Before we discuss implementation, we offer an justification of regarding this approach as following from an ***analytical approximation*** to the loglikelihood (9.49). If the $\widehat{\beta}_i$ are viewed roughly as conditional (on $\boldsymbol{x}_i$ and $\boldsymbol{b}_i$) "***sufficient statistics***" for the β_i for each i, this approach can be viewed as approximating (9.49) with a ***change of variables*** to β_i by replacing $p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)$ in the integrals (9.48) by

$$p(\widehat{\beta}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma),$$

the approximate normal density based on the ***large***-n_i asymptotic theory. Clearly, n_i must be ***sufficiently large*** for the asymptotic approximation to be justified.

ESTIMATION OF β_i: As noted previously, it is standard to assume that the within-individual covariance parameter γ is ***the same*** for all individuals i. This suggests that, rather than estimating γ ***separately*** for each individual in the course of estimating β_i, a natural approach is to "***pool***" information on γ across individuals to obtain a ***common estimator*** and then use this common estimator to estimate β_i, $i = 1, \ldots, m$.

We present this idea first in the case where $\boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i)$ is a ***diagonal*** matrix, as would be the case if we believed that correlation due to within-individual sources is ***negligible***. Extension of the method we now describe to ***non-diagonal*** $\boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i)$ is discussed below. Assuming then that the $(Y_{ij}, \boldsymbol{z}_{ij})$, $j = 1, \ldots, n_i$, are ***independent*** conditional on β_i, consider the stage 1 individual model

$$E(Y_{ij}|\boldsymbol{z}_{ij}, \beta_i) = f(\boldsymbol{z}_{ij}, \beta_i), \quad \text{var}(Y_{ij}|\boldsymbol{z}_{ij}, \beta_i) = \sigma^2 g^2(\beta_i, \delta, \boldsymbol{z}_{ij}), \tag{9.57}$$

where, as in Chapter 7, we regard the Y_{ij} as ***conditionally independent*** given $\boldsymbol{z}_{ij}$ (and β_i), and $\gamma = \theta = (\sigma^2, \delta^T)^T$, common to all i.

Note that, if $\beta_1, \ldots, \beta_m$ were all ***known***, assuming as we did in Chapter 7 for the purpose of deriving an ***estimating equation*** for variance parameters that the distributions of $\boldsymbol{Y}_i|\boldsymbol{z}_i, \beta_i$ are ***normal*** and using the ***conditional independence*** of the $\boldsymbol{Y}_i$ across i, the ***loglikelihood*** for $\theta = (\sigma^2, \delta^T)^T$ across all m individuals is the ***sum*** of the individual loglikelihoods (verify), i.e., ignoring constants,

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} \left[-\log \sigma - \log g(\beta_i, \delta, \boldsymbol{z}_{ij}) - (1/2) \frac{\{Y_{ij} - f(\boldsymbol{z}_{ij}, \beta_i)\}^2}{\sigma^2 g^2(\beta_i, \delta, \boldsymbol{z}_{ij})} \right]. \tag{9.58}$$

Differentiating (9.58) yields the estimating equation

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} \left[\frac{\{Y_{ij} - f(\boldsymbol{z}_{ij}, \beta_i)\}^2}{\sigma^2 g^2(\beta_i, \delta, \boldsymbol{z}_{ij})} - 1 \right] \begin{pmatrix} 1 \\ \nu_\delta(\beta_i, \delta, \boldsymbol{z}_{ij}) \end{pmatrix} = \boldsymbol{0}, \tag{9.59}$$

where ν_δ is defined following (7.21). Note that (9.59) is the sum over $i = 1, \ldots, m$ of ***quadratic estimating equations*** of the form (7.21). Thus, (9.59) can be interpreted as "***pooling***" across the data from all m individuals to estimate the ***common*** σ^2 and δ.

This suggests the following extension of the iterative ***GLS algorithm*** introduced in Section 7.3. For each $i = 1, \ldots, m$, estimate β_i by $\widehat{\beta}_i^{(0)}$, where $\widehat{\beta}_i^{(0)}$ is some initial estimate, e.g., OLS, based on i***'s data only***. Then, at iteration ℓ:

1. Holding β_i fixed at $\widehat{\beta}_i^{(\ell)}$, $i = 1, \ldots, m$, solve the "***pooled***" quadratic estimating equation (9.59) to obtain $\widehat{\theta}^{(\ell)} = (\widehat{\sigma}^{2(\ell)}, \widehat{\delta}^{(\ell)T})^T$.

2. Holding δ fixed at $\widehat{\delta}^{(\ell)}$, ***for each*** $i = 1, \ldots, m$, estimate β_i by solving the ***linear estimating equation*** (7.22) in β_i to obtain $\widehat{\beta}_i^{(\ell+1)}$, $i = 1, \ldots, m$. Set $\ell = \ell + 1$ and return to step 1.

As in the case of a ***single individual***, a variation is to substitute $\widehat{\beta}_i^{(\ell)}$ for β_i in $g^{-2}(\beta_i, \delta, \boldsymbol{x}_j)$ in (7.22) along with $\widehat{\delta}^{(\ell)}$, so that the "weights" are held fixed. The iteration continues to "***convergence***."

When $\boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i)$ is ***not*** a diagonal matrix and involves additional ***correlation parameters*** α, so that the individual model is

$$E(Y_{ij}|\boldsymbol{z}_{ij}, \beta_i) = f(\boldsymbol{z}_{ij}, \beta_i), \quad \text{var}(Y_i|\boldsymbol{z}_i, \beta_i) = \boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i), \tag{9.60}$$

conditioning on $\beta_1, \ldots, \beta_m$, it is clear (verify) that, assuming that distributions of $\boldsymbol{Y}_i|\boldsymbol{z}_i, \beta_i$ are ***normal*** and using the ***conditional independence*** of the $\boldsymbol{Y}_i$ across i, (9.58) is replaced by, ignoring constants,

$$(-1/2)\sum_{i=1}^{m}\Big[\log|\boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i)| + \{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{z}_i, \beta_i)\}^T\boldsymbol{R}_i^{-1}(\beta_i, \gamma, \boldsymbol{z}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{z}_i, \beta_i)\}\Big]. \tag{9.61}$$

Thus, in this case, step 1 of the iterative algorithm involves substituting the current estimates $\widehat{\beta}_i^{(\ell)}$, $i = 1, \ldots, m$, in (9.61) and maximizing in γ. Of course, (9.61) can be differentiated to yield the corresponding ***quadratic estimating equations*** ; these are of the form of those in Chapter 5.

REMARKS:

- Intuition suggests that, if it is believed that γ is ***common*** across individuals, "pooling" information from all m individuals should result in a ***more precise estimator*** for γ than the m estimators $\widehat{\gamma}_i$, say, that would be obtained from each i ***separately***. Although the (conditional on β_i) ***asymptotic theory*** for estimators $\widehat{\beta}_i$ solving the linear estimating equation (7.22) in the independence case implies that it ***does not matter*** how variance parameters are estimated, given that n_i are likely ***not large*** , the intuition is that the "pooled" algorithm above should result in ***more efficient*** individual estimators $\widehat{\beta}_i$ than those obtained separately.

  Intuition suggests further that using these more efficient estimators $\widehat{\beta}_i$ as "data" for estimating β and $\boldsymbol{D}$ should lead to ***more efficient*** estimation of these quantities.

- Unfortunately, there is ***no standard software*** that implements this algorithm. However, it is not difficult to program using ***nonlinear regression*** software such as SAS `proc nlin` or R `nls()`. At the conclusion of the iterative algorithm, approximate estimated ***covariance matrices*** $\widehat{\boldsymbol{\Sigma}}_i$ are available from the final execution of step 2.

- However, if the within-individual covariance model involves a ***scale parameter*** σ^2, an ***adjustment*** must be made. The software will ***automatically*** estimate σ^2 by $\widehat{\sigma}_i^2$, say, based on i's data only and then use this estimate to calculate $\widehat{\boldsymbol{\Sigma}}_i$. Thus, one must ***multiply*** the estimated covariance matrices for each i from the software by $\widehat{\sigma}^2/\widehat{\sigma}_i^2$, where $\widehat{\sigma}^2$ is the pooled estimator, to make sure that they are based on the more efficient estimator $\widehat{\sigma}^2$.

Given "data" $(\widehat{\beta}_i, \widehat{\boldsymbol{\Sigma}}_i)$, $i = 1, \ldots, m$, there are several ways to implement estimation of β and $\boldsymbol{D}$ based on the foregoing observations.

APPROXIMATE POPULATION-AVERAGED ALGORITHM: Consider the approximate PA model in (9.54) and (9.55). In the common special case where $\boldsymbol{B}_i = \boldsymbol{I}_q$, so that each component of β_i has an associated random effect, this becomes

$$E(\beta_i|\boldsymbol{x}_i) \approx \boldsymbol{A}_i\beta, \qquad \text{var}(\widehat{\beta}_i|\boldsymbol{x}_i) \approx \boldsymbol{D} + \widehat{\boldsymbol{\Sigma}}_i. \tag{9.62}$$

From Chapters 5 and 8, ***assuming normality*** for the purpose of deriving estimating equations and differentiating the normal loglikelihood for (9.62)

$$-(1/2)\sum_{i=1}^{m}\log|\boldsymbol{D} + \widehat{\boldsymbol{\Sigma}}_i| - (1/2)\sum_{i=1}^{m}(\widehat{\beta}_i - \boldsymbol{A}_i\beta)^T(\boldsymbol{D} + \widehat{\boldsymbol{\Sigma}}_i)^{-1}(\widehat{\beta}_i - \boldsymbol{A}_i\beta) \tag{9.63}$$

with respect to β and $\boldsymbol{D}$ yields the ***estimating equations***

$$\sum_{i=1}^{m}\boldsymbol{A}_i^T(\boldsymbol{D} + \widehat{\boldsymbol{\Sigma}}_i)^{-1}(\widehat{\beta}_i - \boldsymbol{A}_i\beta) = \boldsymbol{0} \tag{9.64}$$

$$(1/2)\sum_{i=1}^{m}\left[(\widehat{\beta}_i - \boldsymbol{A}_i\beta)^T(\boldsymbol{D} + \widehat{\boldsymbol{\Sigma}}_i)^{-1}\frac{\partial \boldsymbol{D}}{\partial \omega_k}(\boldsymbol{D} + \widehat{\boldsymbol{\Sigma}}_i)^{-1}(\widehat{\beta}_i - \boldsymbol{A}_i\beta) - \text{tr}\left\{(\boldsymbol{D} + \widehat{\boldsymbol{\Sigma}}_i)^{-1}\frac{\partial \boldsymbol{D}}{\partial \omega_k}\right\}\right] = 0, \quad k = 1, \ldots, q(q+1)/2, \tag{9.65}$$

where $\omega = (\omega_1, \ldots, \omega_{q(q+1)/2})^T$ is the vector of distinct elements of $\boldsymbol{D}$.

From (9.64), the estimator for β satisfies

$$\widehat{\beta} = \left\{\sum_{i=1}^{m}\boldsymbol{A}_i^T(\boldsymbol{D} + \widehat{\boldsymbol{\Sigma}}_i)^{-1}\boldsymbol{A}_i\right\}^{-1}\sum_{i=1}^{m}\boldsymbol{A}_i^T(\boldsymbol{D} + \widehat{\boldsymbol{\Sigma}}_i)^{-1}\widehat{\beta}_i, \tag{9.66}$$

and (9.65) are of course the usual ***quadratic estimating equations*** under the "Gaussian working assumption." The equations (9.65) can be expressed in an ***alternative form*** using the following results for symmetric matrix $\boldsymbol{\Lambda}$:

- $\partial/\partial\boldsymbol{\Lambda}\{\log|\boldsymbol{\Lambda}|\} = \boldsymbol{\Lambda}^{-1}$.
- $\partial/\partial\boldsymbol{\Lambda}\{(\boldsymbol{x} - \mu)^T\boldsymbol{\Lambda}^{-1}(\boldsymbol{x} - \mu)\} = -\boldsymbol{\Lambda}^{-1}(\boldsymbol{x} - \mu)(\boldsymbol{x} - \mu)\boldsymbol{\Lambda}^{-1}$.

With $(\boldsymbol{D} + \widehat{\boldsymbol{\Sigma}}_i)$ playing the role of $\boldsymbol{\Lambda}$, it is straightforward to show that differentiating (9.63) with respect to $\boldsymbol{D}$ in this special case and setting equal to zero yields

$$\sum_{i=1}^{m}(\boldsymbol{D} + \widehat{\boldsymbol{\Sigma}}_i)^{-1} - \sum_{i=1}^{m}(\boldsymbol{D} + \widehat{\boldsymbol{\Sigma}}_i)^{-1}(\widehat{\beta}_i - \boldsymbol{A}_i\beta)(\widehat{\beta}_i - \boldsymbol{A}_i\beta)^T(\boldsymbol{D} + \widehat{\boldsymbol{\Sigma}}_i)^{-1} = \boldsymbol{0}.$$

By using matrix inversion results from Appendix A, it can be shown (try it), that this may be ***reexpressed*** as

$$\boldsymbol{D} = m^{-1}\sum_{i=1}^{m}(\boldsymbol{D}^{-1}+\widehat{\boldsymbol{\Sigma}}_i^{-1})^{-1}+m^{-1}\sum_{i=1}^{m}(\boldsymbol{D}^{-1}+\widehat{\boldsymbol{\Sigma}}_i^{-1})^{-1}\widehat{\boldsymbol{\Sigma}}_i^{-1}(\widehat{\beta}_i-\boldsymbol{A}_i\beta)(\widehat{\beta}_i-\boldsymbol{A}_i\beta)^T\widehat{\boldsymbol{\Sigma}}_i^{-1}(\boldsymbol{D}^{-1}+\widehat{\boldsymbol{\Sigma}}_i^{-1})^{-1}. \quad (9.67)$$

The form of (9.67) suggests an ***iterative algorithm*** in which, given the current estimate $\widehat{\beta}^{(\ell)}$ for β obtained from (9.66) with the current estimate $\boldsymbol{D}^{(\ell)}$ substituted, substitute $\widehat{\beta}^{(\ell)}$ and $\boldsymbol{D}^{(\ell)}$ in the right hand side of (9.67) to obtain the update $\boldsymbol{D}^{(\ell+1)}$. This scheme is iterated to convergence.

APPROXIMATE LINEAR MIXED EFFECTS MODEL: The representation (9.56), an ***approximate linear mixed effects*** model

$$\widehat{\beta}_i \approx \boldsymbol{A}_i\beta + \boldsymbol{B}_i\boldsymbol{b}_i + \boldsymbol{e}_i^*, \qquad \boldsymbol{e}_i^* \dot{\sim} \mathcal{N}(\boldsymbol{0}, \widehat{\boldsymbol{\Sigma}}_i), \quad (9.68)$$

suggests that ***standard software*** , such as SAS `proc mixed` and R `nlme()`, can be used to estimate β and $\boldsymbol{D}$. A ***wrinkle*** is that the covariance matrices

$$\text{var}(\boldsymbol{e}_i^*|\boldsymbol{x}_i) \approx \widehat{\boldsymbol{\Sigma}}_i$$

are ***fixed, nondiagonal matrices***. The usual software ***default*** is to take $\text{var}(\boldsymbol{e}_i^*|\boldsymbol{x}_i) = \sigma_e^2\boldsymbol{I}_{n_i}$ for some σ_e^2, which is ***estimated***. This suggests "***preprocessing*** " the "data" $\widehat{\beta}_i$ as follows.

If $\widehat{\boldsymbol{\Sigma}}_i^{-1/2}$ is the Cholesky decomposition of $\widehat{\boldsymbol{\Sigma}}_i^{-1}$; i.e., an upper triangular matrix satisfying $\widehat{\boldsymbol{\Sigma}}_i^{-1/2\,T}\widehat{\boldsymbol{\Sigma}}_i^{-1/2} = \widehat{\boldsymbol{\Sigma}}_i^{-1}$, then $\widehat{\boldsymbol{\Sigma}}_i^{-1/2}\widehat{\boldsymbol{\Sigma}}_i\widehat{\boldsymbol{\Sigma}}_i^{-1/2\,T} = \boldsymbol{I}_p$, so that $\widehat{\boldsymbol{\Sigma}}_i^{-1/2}\boldsymbol{e}_i^*$ has identity covariance matrix. Premultiplying (9.68) by $\widehat{\boldsymbol{\Sigma}}_i^{-1/2}$ yields the "new" ***linear mixed effects model***

$$\widehat{\boldsymbol{\Sigma}}_i^{-1/2}\widehat{\beta}_i \approx (\widehat{\boldsymbol{\Sigma}}_i^{-1/2}\boldsymbol{A}_i)\beta + (\widehat{\boldsymbol{\Sigma}}_i^{-1/2}\boldsymbol{B}_i)\boldsymbol{b}_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p). \quad (9.69)$$

Fitting (9.69) to the "data" $\widehat{\boldsymbol{\Sigma}}_i^{-1/2}\widehat{\beta}_i$ with "design matrices"

$$\boldsymbol{X}_i = \widehat{\boldsymbol{\Sigma}}_i^{-1/2}\boldsymbol{A}_i \qquad \text{and} \qquad \boldsymbol{Z}_i = \widehat{\boldsymbol{\Sigma}}_i^{-1/2}\boldsymbol{B}_i$$

and the ***constraint*** $\sigma_e^2 = 1$ can then be carried out with standard software.

"TWO-STAGE ALGORITHMS:" In the ***special case*** $\boldsymbol{B}_i = \boldsymbol{I}_q$, ***pharmacokineticists*** have fitted the "***linear mixed effects model*** " (9.68) via an ***EM algorithm*** as follows.

(i) Obtain starting values

$$\widehat{\beta}_{(0)} = m^{-1}\sum_{i=1}^{m}\widehat{\beta}_i, \quad \widehat{\boldsymbol{D}}_{(0)} = (m-1)^{-1}\sum_{i=1}^{m}(\widehat{\beta}_i - \boldsymbol{A}_i\widehat{\beta}_{(0)})(\widehat{\beta}_i - \boldsymbol{A}_i\widehat{\beta}_{(0)})^T.$$

Set $\ell = 0$.

(ii) "***E-step.***" Produce current ***empirical Bayes*** estimates of the β_i, $i = 1, \ldots, m$, given by

$$\widetilde{\beta}_{i,(\ell+1)} = (\widehat{\boldsymbol{D}}_{(\ell)}^{-1} + \widehat{\boldsymbol{\Sigma}}_i^{-1})^{-1}(\boldsymbol{C}_i^{-1}\widehat{\beta}_i + \widehat{\boldsymbol{D}}_{(\ell)}^{-1}\boldsymbol{A}_i\widehat{\beta}_{(\ell)}).$$

It is straightforward to deduce that this expression is of the form of the "***posterior mean***" $E(\beta_i|\widehat{\beta}_i, \boldsymbol{x}_i)$ for the "model" (9.68) under normality.

(iii) "***M-step.***" Obtain updated estimates as

$$\widehat{\beta}_{(k+1)} = \sum_{i=1}^{m} \boldsymbol{W}_{i,(\ell)}\widetilde{\beta}_{i,(\ell+1)}, \quad \boldsymbol{W}_{i,(\ell)} = \left(\sum_{i=1}^{m} \boldsymbol{A}_i^T\widehat{\boldsymbol{D}}_{(\ell)}^{-1}\boldsymbol{A}_i\right)^{-1}\boldsymbol{A}_i^T\widehat{\boldsymbol{D}}_{(\ell)}^{-1},$$

$$\widehat{\boldsymbol{D}}_{(\ell+1)} = m^{-1}\sum_{i=1}^{m}(\widehat{\boldsymbol{D}}_{(\ell)}^{-1} + \widehat{\boldsymbol{\Sigma}}_i^{-1})^{-1} + m^{-1}\sum_{i=1}^{m}(\widetilde{\beta}_{i,(\ell+1)} - \boldsymbol{A}_i\widehat{\beta}_{(\ell+1)})(\widetilde{\beta}_{i,(\ell+1)} - \boldsymbol{A}_i\widehat{\beta}_{(\ell+1)})^T.$$

Set $\ell = \ell + 1$ and return to (ii).

The algorithm is iterated to ***convergence***. The results should be ***identical*** to those obtained by direct maximization (e.g., using the software mentioned above).

APPROXIMATE SAMPLING DISTRIBUTION: Regardless of which of these methods is used to estimate β and $\boldsymbol{D}$, it follows from the standard ***asymptotic theory*** in Chapters 5 and 6 that, for ***large*** m,

$$\widehat{\beta} \dot{\sim} \mathcal{N}\left[\beta, \left\{\sum_{i=1}^{m} \boldsymbol{A}_i^T(\boldsymbol{B}_i\widehat{\boldsymbol{D}}\boldsymbol{B}_i^T + \widehat{\boldsymbol{\Sigma}}_i)^{-1}\boldsymbol{A}_i\right\}^{-1}\right]. \tag{9.70}$$

An alternative ***robust sandwich*** covariance matrix can be obtained as in Chapters 5 and 6.

TERMINOLOGY:

- In the ***pharmacokinetics literature***, for unknown reasons, the foregoing EM algorithm is referred to as the ***Global Two-Stage*** (GTS) method. This may be to distinguish it from a simpler, ad hoc approach referred to as the ***Standard Two-Stage*** (STS) method, in which β and $\boldsymbol{D}$ are estimated by

$$\widehat{\beta}_{STS} = \left(\sum_{i=1}^{m} \boldsymbol{A}_i^T\boldsymbol{A}_i\right)^{-1}\sum_{i=1}^{m} \boldsymbol{A}_i^T\widehat{\beta}_i,$$

$$\widehat{\boldsymbol{D}}_{STS} = (m-1)^{-1}\sum_{i=1}^{m}(\widehat{\beta}_i - \boldsymbol{A}_i\widehat{\beta}_{STS})(\widehat{\beta}_i - \boldsymbol{A}_i\widehat{\beta}_{STS})^T.$$

Clearly, $\widehat{\beta}_{STS}$ is ***inefficient*** relative to that discussed above. More disturbingly, $\widehat{\boldsymbol{D}}_{STS}$ is a ***biased*** estimator for $\boldsymbol{D}$ (try it).

- In fact, pharmacokineticists often view the approaches above based on ***individual estimates*** $\widehat{\beta}_i$ as being somehow ***distinct*** from the ***nonlinear mixed effects model***. They mistakenly refer to ***two-stage approaches*** as a ***separate modeling approach*** and use the term ***nonlinear mixed effects model*** only in the context of the methods we discuss in the next two sections.

9.5 Approximate inference based on linearization

Historically, methods based on ***individual estimates*** have been attractive because they break down fitting of the ***nonlinear mixed effects model*** (9.8)-(9.9) into two stages, each of which can be carried out using ***standard methods*** (with some minor modifications).

A drawback of these methods is that they require n_i to be large enough for ***all*** of the m individuals so that estimation of the β_i is feasible ***and*** the ***large sample approximation*** to the distribution of $\widehat{\beta}_i|\beta_i, \boldsymbol{x}_i$ is reasonable. In practice, the n_i may not always be ***large enough*** for these conditions to be met.

Often, although many of the m individuals have large enough n_i to support individual estimation, some ***do not***. Disregarding these individuals raises the possibility for inefficient and even ***biased inference*** , as the remaining individuals may no longer be a random sample from the population. In some settings, such as for ***population pharmacokinetic studies*** like the phenobarbital study, the sampling design for all individuals involves n_i that are too small. In these situations, the methods of Section 9.4 are ***not an option***.

An alternative class of methods is motivated by returning to the ***loglikelihood*** (9.49), which involves the product of terms

$$p(\boldsymbol{Y}_i|\boldsymbol{x}_i;\beta,\gamma,\boldsymbol{D}) = \int p(\boldsymbol{Y}_i|\boldsymbol{x}_i,\boldsymbol{b}_i;\beta,\gamma)\, p(\boldsymbol{b}_i;\boldsymbol{D})\, d\boldsymbol{b}_i. \tag{9.71}$$

The methods are based on approximating the integral (9.71) by a ***closed form expression*** , thereby leading to an ***analytical*** , closed form approximation to the loglikelihood. This approximate objective function can maximized directly and/or differentiated to yield ***estimating equations*** in the spirit of those in Chapter 8.

FIRST ORDER LINEARIZATION METHODS: The ***simplest*** such methods approximate (9.71) for each i by referring directly to the stage 1 individual model

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i,\boldsymbol{b}_i) = \boldsymbol{f}_i(\boldsymbol{x}_i,\beta,\boldsymbol{b}_i), \qquad \text{var}(\boldsymbol{Y}|\boldsymbol{x}_i,\boldsymbol{b}_i) = \boldsymbol{R}_i(\beta,\gamma,\boldsymbol{x}_i,\boldsymbol{b}_i).$$

Assuming that $\boldsymbol{R}_i(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{b}_i)$ is ***positive definite***, letting $\boldsymbol{R}_i^{1/2}(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{b}_i)$ be its ***Cholesky decomposition*** (or other ***square root matrix***), and defining

$$\epsilon_i = \boldsymbol{R}_i^{-1/2}(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{b}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i)\},$$

write the model as

$$\boldsymbol{Y}_i = \boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i) + \boldsymbol{R}^{1/2}(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{b}_i)\epsilon_i. \tag{9.72}$$

Note that $\epsilon_i | \boldsymbol{x}_i, \boldsymbol{b}_i$ has mean $\boldsymbol{0}$ and ***identity*** covariance matrix.

The ***difficulty*** with the integration in (9.71) is the fact that $\boldsymbol{b}_i$ enters in a ***nonlinear*** fashion. (This same issue arises ***even*** if we are only interested in obtaining the mean and covariance matrix of $\boldsymbol{Y}_i | \boldsymbol{x}_i$.) Thus, the idea is to ***approximate*** (9.72) by a model that is ***linear*** in the $\boldsymbol{b}_i$. The simplest way to do this is to invoke a ***linear approximation*** of (9.72) about the ***mean*** of the $\boldsymbol{b}_i$, $\boldsymbol{0}$.

This approach was advocated in the ***pharmacokinetics literature*** in the early 1980s; see, for example, Beal and Sheiner (1985), where it is referred to as the ***first order method*** (FO). In the special case of ***generalized linear mixed effects models*** it is referred to by Breslow and Clayton (1993), Zeger, Liang, and Albert (1988), and others as ***marginal quasilikelihood*** (MQL).

By Taylor series of (9.72) about $\boldsymbol{b}_i = \boldsymbol{0}$,

$$\begin{aligned} \boldsymbol{Y}_i &\approx \boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{0}) + \partial/\partial \boldsymbol{b}_i\{\boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{0})\}(\boldsymbol{b}_i - \boldsymbol{0}) + \boldsymbol{R}_i^{1/2}(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{0})\epsilon_i + \partial/\partial \boldsymbol{b}_i\{\boldsymbol{R}_i^{1/2}(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{0})\}(\boldsymbol{b}_i - \boldsymbol{0})\epsilon_i \\ &\approx \boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{0}) + \boldsymbol{Z}_i(\boldsymbol{x}_i, \beta, \boldsymbol{0})\boldsymbol{b}_i + \boldsymbol{R}_i^{1/2}(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{0})\epsilon_i. \end{aligned} \tag{9.73}$$

In (9.73), $\boldsymbol{Z}_i(\boldsymbol{x}_i, \beta, \boldsymbol{0}) = \partial/\partial \boldsymbol{b}_i\{\boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i)\}|_{b_i=0}$. The ***crossproduct term*** involving $\boldsymbol{b}_i\epsilon_i$ is disregarded as "***small***" relative to the leading three terms, as both $\boldsymbol{b}_i$ and ϵ_i have mean $\boldsymbol{0}$ conditional on $\boldsymbol{x}_i$.

The ***approximate model*** in (9.73) implies the ***approximate population-averaged moments***

$$E(\boldsymbol{Y}_i | \boldsymbol{x}_i) \approx \boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{0}), \quad \text{var}(\boldsymbol{Y}_i | \boldsymbol{x}_i) \approx \boldsymbol{Z}_i(\boldsymbol{x}_i, \beta, \boldsymbol{0})\boldsymbol{D}\boldsymbol{Z}_i^T(\boldsymbol{x}_i, \beta, \boldsymbol{0}) + \boldsymbol{R}_i(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{0}). \tag{9.74}$$

The covariance matrix $\text{var}(\boldsymbol{Y}_i | \boldsymbol{x}_i)$ in (9.74) has the same form of that for a ***linear mixed effects model***; in fact, as we demonstrate shortly in the context of a more refined approximation, by a Taylor series in β, (9.73) can be ***further approximated*** so as to have a ***linear*** "***mean model***."

Note that the approximation involves evaluation of the function f at $\boldsymbol{b}_i = \boldsymbol{0}$ for all individuals. Clearly, this approximation destroys the "***individuality***" of the mean response. We say more about this momentarily.

The approximation (9.74) suggests ***several ways*** to estimate (β, ξ) as follows.

- Under the assumption of ***normality*** of the conditional distributions $\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i$ and with $\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$, it follows from (9.73) that $\boldsymbol{Y}_i|\boldsymbol{x}_i$ is ***approximately normal*** with mean and covariance matrix as in (9.74). This suggests approximating the integral (9.71) by a ***normal density*** with these moments, so that (9.49) involves a product of these densities.

  From (9.74), the approximate covariance matrix ***depends on*** β. Thus, as in Section 8.4, this results in ***quadratic estimating equations*** for both β and ξ incorporating the the***Gaussian working assumption***; that is, ***GEE-2*** equations with this working assumption.

  This is implemented in the software package `nonmem`, a suite of Fortran programs that is heavily focused on ***pharmacokinetic analysis***. This method is also available in SAS `proc nlmixed`.

- Alternatively, a ***GEE-1*** approach using the appropriate ***linear estimating equation*** for β is possible. This is implemented in the SAS macro `nlinmix`.

- Standard errors for the estimator for β are obtained by applying the ***usual theory***.

DRAWBACK: Clearly, these estimating equations cannot be expected to be ***unbiased***. The approximate population-averaged mean in (9.74) is clearly ***not equal*** to the true mean

$$\int \boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i)\, p(\boldsymbol{b}_i; \boldsymbol{D})\, d\boldsymbol{b}_i.$$

Replacing $\boldsymbol{b}_i$ by its mean, $\boldsymbol{0}$, obviously yields only a ***crude approximation*** to this integral. Thus, as with the methods of the previous section, which also rely on approximations, there is ***no reason*** to expect the estimator for β to be ***consistent*** , or even approximately so.

- ***Amazingly*** , in some applications where the ***magnitude of the inter-individual variation*** (represented by $\boldsymbol{D}$) is not too great, so that $\boldsymbol{b}_i$ are not too far from $\boldsymbol{0}$, estimators that are ***nearly unbiased*** in finite samples (finite m) can be obtained. This has been observed in ***extensive simulation studies*** in the area of pharmacokinetics. Although this is fortuitous, it is by ***no means necessary***.

MORE REFINED APPROXIMATIONS: Approximations that do not remove the "***individuality***" of the model that can be ***more accurate*** can be motivated in different ways. It is beyond our scope to give a full, detailed account of the many strategies that have been proposed. Rather, we provide a ***heuristic motivation*** for the general refined approach and refer to the literature for more details, alternative derivations, and variations on this theme.

One such approximation was advocated by Lindstrom and Bates (1990). A simple motivation follows from a ***linearization argument*** similar to that above. Instead of approximating (9.72),

$$\boldsymbol{Y}_i = \boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i) + \boldsymbol{R}^{1/2}(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{b}_i)\epsilon_i,$$

by expansion about $\boldsymbol{b}_i = \boldsymbol{0}$, Lindstrom and Bates argued that expansion about a value "***closer to***" $\boldsymbol{b}_i$ should result in a ***more accurate approximation***.

By analogy to the steps above, expanding about $\boldsymbol{b}_i^*$ "close to" $\boldsymbol{b}_i$,

$$\begin{aligned}\boldsymbol{Y}_i &\approx \boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i^*) + \partial/\partial \boldsymbol{b}_i\{\boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i^*)\}(\boldsymbol{b}_i - \boldsymbol{b}_i^*) + \boldsymbol{R}_i^{1/2}(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{b}_i^*)\epsilon_i \\ &\quad +\partial/\partial \boldsymbol{b}_i\{\boldsymbol{R}_i^{1/2}(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{b}_i^*)\}(\boldsymbol{b}_i - \boldsymbol{b}_i^*)\epsilon_i \\ &= \{\boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i^*) - \boldsymbol{Z}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i^*)\boldsymbol{b}_i^*\} + \boldsymbol{Z}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i^*)\boldsymbol{b}_i + \boldsymbol{R}_i^{1/2}(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{b}_i^*)\epsilon_i. \qquad (9.75)\end{aligned}$$

Here, $\boldsymbol{Z}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i^*) = \partial/\partial \boldsymbol{b}_i\{\boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i)\}|_{b_i = b_i^*}$, and the term involving $(\boldsymbol{b}_i - \boldsymbol{b}_i^*)\epsilon_i$ has been disregarded as "negligible."

Treating $\boldsymbol{b}_i^*$ as a ***fixed constant*** , (9.75) yields the approximate moments

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i) \approx \boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i^*) - \boldsymbol{Z}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i^*)\boldsymbol{b}_i^*, \quad \text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) \approx \boldsymbol{Z}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i^*)\boldsymbol{D}\boldsymbol{Z}_i^T(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i^*) + \boldsymbol{R}_i(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{b}_i^*). \qquad (9.76)$$

From the approximate moments given in (9.76), if $\boldsymbol{b}_i^*$ were ***known***, as for the ***cruder approximation*** about $\boldsymbol{0}$, it is possible to deduce ***estimating equations*** (e.g., GEE-1 or GEE-2) that can be solved to estimate β and ξ.

Thus, a suitable value $\boldsymbol{b}_i^*$ to substitute in (9.76) is required. Lindstrom and Bates (1990) focus on the situation where $\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$ and $p(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)$ is a normal density. Under these conditions, they suggest substituting the ***posterior mode*** for $\boldsymbol{b}_i$, $\widehat{\boldsymbol{b}}_i$. From (9.50), $\widehat{\boldsymbol{b}}_i$ maximizes, ignoring constants

$$-(1/2)\log|\boldsymbol{R}_i(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{b}_i)| - (1/2)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i)\}^T\boldsymbol{R}_i^{-1}(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{b}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i)\} - (1/2)\boldsymbol{b}_i^T\boldsymbol{D}^{-1}\boldsymbol{b}_i. \qquad (9.77)$$

- Lindstrom and Bates (1990) restricted attention to models where $\boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i)$ ***does not depend on*** β_i, and thus does not depend on $\boldsymbol{b}_i$. In this case, the first term in (9.77) is constant with respect to $\boldsymbol{b}_i$ and can be disregarded. We relax this in the following.

When $p(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)$ and $p(\boldsymbol{b}_i; \boldsymbol{D})$ are ***normal*** , these considerations suggest the following basic ***iterative strategy*** for estimation of β and ξ.

(i) Obtain ***initial estimators*** $\widehat{\beta}^{(0)}$ and $\widehat{\xi}^{(0)}$; e.g., by fitting the approximate model (9.74) obtained from expanding about $\boldsymbol{b}_i = \boldsymbol{0}$. In almost all the literature, a ***GEE-1*** approach with the ***Gaussian working assumption*** is used.

Then obtain ***initial empirical Bayes estimators*** $\widehat{\boldsymbol{b}}_i^{(0)}$, $i = 1, \ldots, m$, by substituting $\widehat{\beta}^{(0)}$ and $\widehat{\xi}^{(0)}$ in (9.77) for ***each*** i, and, holding these fixed, ***maximize*** in $\boldsymbol{b}_i$. Thus, m maximizations, one for each i, are performed. Set $\ell = 0$.

(ii) ***Substitute*** $\widehat{\boldsymbol{b}}_i^{(\ell)}$ for $\boldsymbol{b}_i^*$ in the approximate moments (9.76). Treating $\widehat{\boldsymbol{b}}_i^{(\ell)}$ as ***fixed***, solve a set of estimating equations, usually using ***GEE-1*** equations with the ***Gaussian working assumption***, to obtain the ***updated estimators*** $\widehat{\beta}^{(\ell+1)}$ and $\widehat{\xi}^{(\ell+1)}$.

(iii) ***Substituting*** $\widehat{\beta}^{(\ell+1)}$ and $\widehat{\xi}^{(\ell+1)}$ in (9.77) and holding fixed, ***maximize (***9.77) in $\boldsymbol{b}_i$ for each i in m separate maximizations to obtain $\widehat{\boldsymbol{b}}_i^{(\ell+1)}$, $i = 1, \ldots, m$. Set $\ell = \ell + 1$ and go to (ii).

Iteration between steps (ii) and (iii) proceeds until "***convergence***," where this is usually defined as relative change in successive estimates of all components of β and ξ being less than some tolerance.

Various versions of this scheme are implemented in the SAS macro `nlinmix` and the R function `nlme()`. These software packages focus specifically on the case where $p(\boldsymbol{Y}_i|\boldsymbol{x}_i\boldsymbol{b}_i;\beta,\gamma)$ is the ***normal density***. There are a few ***subtleties***.

- These implementations ***ignore*** the first term in (9.77) when performing the update in step (iii).
- These and other implementations invoke a ***further approximation*** to allow algorithms for fitting ***linear mixed effects models*** to be exploited. At step (ii), the ***approximate moments*** are

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i) \approx \boldsymbol{f}_i(\boldsymbol{x}_i,\beta,\widehat{\boldsymbol{b}}_i^{(\ell)}) - \boldsymbol{Z}_i(\boldsymbol{x}_i,\beta,\widehat{\boldsymbol{b}}_i^{(\ell)})\widehat{\boldsymbol{b}}_i^{(\ell)},$$

$$\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) \approx \boldsymbol{Z}_i(\boldsymbol{x}_i,\beta,\widehat{\boldsymbol{b}}_i^{(\ell)})\boldsymbol{D}\boldsymbol{Z}_i^T(\boldsymbol{x}_i,\beta,\widehat{\boldsymbol{b}}_i^{(\ell)}) + \boldsymbol{R}_i(\beta,\gamma,\boldsymbol{x}_i,\widehat{\boldsymbol{b}}_i^{(\ell)}).$$

Substituting the previous iterate for β, $\widehat{\beta}^{(\ell)}$, in the expression for $\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i)$, $\boldsymbol{Z}_i(\boldsymbol{x}_i,\widehat{\beta}^{(\ell)},\widehat{\boldsymbol{b}}_i^{(\ell)})$ and $\boldsymbol{R}_i(\widehat{\beta}^{(\ell)},\gamma,\boldsymbol{x}_i,\widehat{\boldsymbol{b}}_i^{(\ell)})$ are ***constant*** with respect to β, as in a ***linear mixed effects model***.

Moreover, by a ***further approximation*** to the mean, expanding $\boldsymbol{f}_i(\boldsymbol{x}_i,\beta,\widehat{\boldsymbol{b}}_i^{(\ell)})$ and $\boldsymbol{Z}_i(\boldsymbol{x}_i,\beta,\widehat{\boldsymbol{b}}_i^{(\ell)})$ to ***linear terms*** about $\widehat{\beta}^{(\ell)}$ and ***ignoring*** "negligible" terms,

$$\begin{aligned} E(\boldsymbol{Y}_i|\boldsymbol{x}_i) &\approx \boldsymbol{f}_i(\boldsymbol{x}_i,\widehat{\beta}^{(\ell)},\widehat{\boldsymbol{b}}_i^{(\ell)}) + \boldsymbol{X}_i(\boldsymbol{x}_i,\widehat{\beta}^{(\ell)},\widehat{\boldsymbol{b}}_i^{(\ell)})(\beta-\widehat{\beta}^{(\ell)}) - \boldsymbol{Z}_i(\boldsymbol{x}_i,\widehat{\beta}^{(\ell)},\widehat{\boldsymbol{b}}_i^{(\ell)})\widehat{\boldsymbol{b}}_i^{(\ell)} \\ &= \boldsymbol{f}_i(\boldsymbol{x}_i,\widehat{\beta}^{(\ell)},\widehat{\boldsymbol{b}}_i^{(\ell)}) - \boldsymbol{X}_i(\boldsymbol{x}_i,\widehat{\beta}^{(\ell)},\widehat{\boldsymbol{b}}_i^{(\ell)})\widehat{\beta}^{(\ell)} - \boldsymbol{Z}_i(\boldsymbol{x}_i,\widehat{\beta}^{(\ell)},\widehat{\boldsymbol{b}}_i^{(\ell)})\widehat{\boldsymbol{b}}_i^{(\ell)} + \boldsymbol{X}_i(\boldsymbol{x}_i,\widehat{\beta}^{(\ell)},\widehat{\boldsymbol{b}}_i^{(\ell)})\beta, \end{aligned}$$

where $\boldsymbol{X}_i(\boldsymbol{x}_i,\beta,\boldsymbol{b}_i) = \partial/\partial\beta\,\{\boldsymbol{f}_i(\boldsymbol{x}_i,\beta,\boldsymbol{b}_i)\}$.

Defining the "***pseudo-response vector***"

$$\boldsymbol{w}_i^{(\ell)} = \boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \widehat{\beta}^{(\ell)}, \widehat{\boldsymbol{b}}_i^{(\ell)}) + \boldsymbol{X}_i(\boldsymbol{x}_i, \widehat{\beta}^{(\ell)}, \widehat{\boldsymbol{b}}_i^{(\ell)})\widehat{\beta}^{(\ell)} + \boldsymbol{Z}_i(\boldsymbol{x}_i, \widehat{\beta}^{(\ell)}, \widehat{\boldsymbol{b}}_i^{(\ell)})\widehat{\boldsymbol{b}}_i^{(\ell)},$$

then, approximately, from above,

$$E(\boldsymbol{w}_i^{(\ell)}|\boldsymbol{x}_i) \approx \boldsymbol{X}_i(\boldsymbol{x}_i, \widehat{\beta}^{(\ell)}, \widehat{\boldsymbol{b}}_i^{(\ell)})\beta. \tag{9.78}$$

The approximate mean model (9.78) for $\boldsymbol{w}_i^{(\ell)}$ is ***linear*** in β; moreover, from above,

$$\text{var}(\boldsymbol{w}_i^{(\ell)}|\boldsymbol{x}_i) \approx \boldsymbol{Z}_i(\boldsymbol{x}_i, \widehat{\beta}^{(\ell)}, \widehat{\boldsymbol{b}}_i^{(\ell)})\boldsymbol{D}\boldsymbol{Z}_i^T(\boldsymbol{x}_i, \widehat{\beta}^{(\ell)}, \widehat{\boldsymbol{b}}_i^{(\ell)}) + \boldsymbol{R}_i(\widehat{\beta}^{(\ell)}, \gamma, \boldsymbol{x}_i, \widehat{\boldsymbol{b}}_i^{(\ell)}). \tag{9.79}$$

Together, (9.78) and (9.79) represent an approximate ***linear mixed effects model*** with "***constant design matrices***"

$$\boldsymbol{X}_i^{(\ell)} = \boldsymbol{X}_i(\boldsymbol{x}_i, \widehat{\beta}^{(\ell)}, \widehat{\boldsymbol{b}}_i^{(\ell)}) \quad \text{and} \quad \boldsymbol{Z}_i^{(\ell)} = \boldsymbol{Z}_i(\boldsymbol{x}_i, \widehat{\beta}^{(\ell)}, \widehat{\boldsymbol{b}}_i^{(\ell)})$$

and "constant" ***within-individual covariance matrix*** $\boldsymbol{R}_i(\widehat{\beta}^{(\ell)}, \gamma, \boldsymbol{x}_i, \widehat{\boldsymbol{b}}_i^{(\ell)})$. Thus, β and ξ can be estimated using ***standard techniques and software*** for linear mixed effects models.

- In fact, step (iii) can ***also*** be approximated. Under the ***approximate linear mixed effects model*** (9.78) and (9.79), the ***posterior mode*** for $\boldsymbol{b}_i$ can be updated using the expression for the ***posterior mean*** for a (normal) linear mixed effects model in (6.54). In particular,

$$\begin{aligned}\widehat{\boldsymbol{b}}_i^{(\ell+1)} &= \hat{\boldsymbol{D}}^{(\ell+1)}\boldsymbol{Z}_i^T(\boldsymbol{x}_i, \widehat{\beta}^{(\ell)}, \widehat{\boldsymbol{b}}_i^{(\ell)})\boldsymbol{V}_i^{-1}(\widehat{\beta}^{(\ell)}, \widehat{\xi}^{(\ell+1)}, \boldsymbol{x}_i)\{\boldsymbol{w}_i^{(\ell)} - \boldsymbol{X}_i(\boldsymbol{x}_i, \widehat{\beta}^{(\ell)}, \widehat{\boldsymbol{b}}_i^{(\ell)})\widehat{\beta}^{(\ell+1)}\} \\ &= \hat{\boldsymbol{D}}^{(\ell+1)}\boldsymbol{Z}_i^{(\ell)}\boldsymbol{V}_i^{(\ell)-1}(\boldsymbol{w}_i^{(\ell)} - \boldsymbol{X}_i^{(\ell)}\widehat{\beta}^{(\ell+1)}),\end{aligned} \tag{9.80}$$

$$\begin{aligned}\boldsymbol{V}_i^{(\ell)} &= \boldsymbol{V}_i(\widehat{\beta}^{(\ell)}, \widehat{\xi}^{(\ell+1)}, \boldsymbol{x}_i) = \boldsymbol{Z}_i(\boldsymbol{x}_i, \widehat{\beta}^{(\ell)}, \widehat{\boldsymbol{b}}_i^{(\ell)})\hat{\boldsymbol{D}}^{(\ell+1)}\boldsymbol{Z}_i^T(\boldsymbol{x}_i, \widehat{\beta}^{(\ell)}, \widehat{\boldsymbol{b}}_i^{(\ell)}) + \boldsymbol{R}_i(\widehat{\beta}^{(\ell)}, \widehat{\gamma}^{(\ell+1)}, \boldsymbol{x}_i, \widehat{\boldsymbol{b}}_i^{(\ell)}) \\ &= \boldsymbol{Z}_i^k\hat{\boldsymbol{D}}^{(\ell+1)}\boldsymbol{Z}_i^{(\ell)T} + \boldsymbol{R}_i(\widehat{\beta}^{(\ell)}, \widehat{\gamma}^{(\ell+1)}, \boldsymbol{x}_i, \widehat{\boldsymbol{b}}_i^{(\ell)})\end{aligned}$$

- The SAS macro `nlinmix` implements these further approximations. Steps (ii) and (iii) are carried out by forming the $\boldsymbol{w}_i^{(\ell)}$ and the matrices $\boldsymbol{X}_i^{(\ell)}$ and $\boldsymbol{Z}_i^{(\ell)}$ for the current iteration and calling `proc mixed` to fit the approximate linear mixed effects model. The ***approximate posterior modes*** (9.80) are a ***byproduct*** of calling `proc mixed`, as `proc mixed` calculates these based on (6.54) evaluated at the final estimates of β, γ, and $\boldsymbol{D}$ by default.

  The R function `nlme()` ***does not*** use the further approximation (9.80) but rather maximizes (9.77), ***disregarding*** the leading term.

Standard errors for the estimator for β obtained via this approach are generally computed by using the usual large sample results, with the final estimates of β and ξ and final value for $\widehat{\boldsymbol{b}}_i$ substituted.

ALTERNATIVE DERIVATION: An alternative derivation of this scheme is possible based on the ***Laplace approximation*** to an integral of the form

$$\int \exp\{n\ell(\boldsymbol{\tau})\}\, d\boldsymbol{\tau} \approx \left(\frac{2\pi}{n}\right)^{q/2} |\ell''(\widehat{\boldsymbol{\tau}})|^{-1/2} \exp\{n\ell(\widehat{\boldsymbol{\tau}})\}. \tag{9.81}$$

Here, $\boldsymbol{\tau}$ is $(q \times 1)$, $\ell(\boldsymbol{\tau})$ is a real-valued function of $\boldsymbol{\tau}$ that is maximized at $\widehat{\boldsymbol{\tau}}$, and $\ell''(\boldsymbol{\tau}) = \partial^2/\partial\boldsymbol{\tau}\partial\boldsymbol{\tau}^T\{\ell(\boldsymbol{\tau})\}$. The ***Laplace approximation*** is valid when n is "large." In particular, the approximation is $O(n^{-1})$.

Wolfinger (1993) and Vonesh (1996) discuss how the Laplace approximation (9.81) can be applied when $p(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \boldsymbol{\beta}, \boldsymbol{\gamma})$ and $p(\boldsymbol{b}_i; \boldsymbol{D})$ are ***normal densities***. This proceeds by identifying $\boldsymbol{b}_i$ with $\boldsymbol{\tau}$ and n with n_i for individual i in the integral in (9.71).

These authors consider the situation where $\boldsymbol{R}_i(\boldsymbol{\beta}_i, \boldsymbol{\gamma}, \boldsymbol{z}_i)$ ***does not*** depend on $\boldsymbol{\beta}_i$, which we write as $\boldsymbol{R}_i(\boldsymbol{\gamma}, \boldsymbol{x}_i)$. In this case, the integral (9.71) is

$$\begin{aligned}&(2\pi)^{-n_i/2}(2\pi)^{-q/2}|\boldsymbol{R}_i(\boldsymbol{\gamma}, \boldsymbol{x}_i)|^{-1/2}|\boldsymbol{D}|^{-1/2} \\ &\quad\times \int \exp[-(1/2)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)\}^T \boldsymbol{R}_i^{-1}(\boldsymbol{\gamma}, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)\} - (1/2)\boldsymbol{b}_i^T \boldsymbol{D}^{-1}\boldsymbol{b}_i]\, d\boldsymbol{b}_i.\end{aligned} \tag{9.82}$$

Consider approximating the integral in (9.82) by (9.81). Identifying

$$\ell(\boldsymbol{b}_i) = -\frac{1}{2n_i}\left[\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)\}^T \boldsymbol{R}_i^{-1}(\boldsymbol{\gamma}, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)\} + \boldsymbol{b}_i^T \boldsymbol{D}^{-1}\boldsymbol{b}_i\right],$$

and using various ***matrix differentiation results***, it is straightforward to show (try it) that $\ell'(\boldsymbol{b}_i) = \partial/\partial\boldsymbol{b}_i\{\ell(\boldsymbol{b}_i)\}$ satisfies

$$\ell'(\boldsymbol{b}_i) = -n_i^{-1}\boldsymbol{D}^{-1}\boldsymbol{b}_i + n_i^{-1}\boldsymbol{Z}_i^T(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)\boldsymbol{R}_i^{-1}(\boldsymbol{\gamma}, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)\} \tag{9.83}$$

and the matrix of second partial derivatives is

$$\begin{aligned}\ell''(\boldsymbol{b}_i) \;=\; & -n_i^{-1}\boldsymbol{D}^{-1} - n_i^{-1}\boldsymbol{Z}_i^T(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)\boldsymbol{R}_i^{-1}(\boldsymbol{\gamma}, \boldsymbol{x}_i)\boldsymbol{Z}_i(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{b}_i) \\ & + n_i^{-1}\partial/\partial\boldsymbol{b}_i^T\{\boldsymbol{Z}_i^T(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)\}\boldsymbol{R}_i^{-1}(\boldsymbol{\gamma}, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)\}.\end{aligned} \tag{9.84}$$

The third term on the right hand side of (9.84) has ***conditional expectation zero***. It is standard to ***disregard*** this term, so effectively making a ***further approximation*** to the Laplace approximation (9.81) by replacing $\ell''(\cdot)$ by its ***conditional expectation*** on the right hand side of (9.81).

Substituting the conditional expectation of (9.84) in (9.81) yields

$$\begin{aligned}p(\boldsymbol{Y}_i|\boldsymbol{x}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{D}) &\approx (2\pi)^{-n_i/2}(2\pi)^{-q/2}|\boldsymbol{R}_i(\boldsymbol{\gamma}, \boldsymbol{x}_i)|^{-1/2}|\boldsymbol{D}|^{-1/2}(2\pi)^{q/2}n_i^{-q/2} \\ &\quad\times n_i^{q/2}|\boldsymbol{D}^{-1} + \boldsymbol{Z}_i^T(\boldsymbol{x}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{b}}_i)\boldsymbol{R}_i^{-1}(\boldsymbol{\gamma}, \boldsymbol{x}_i)\boldsymbol{Z}_i(\boldsymbol{x}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{b}}_i)|^{-1/2} \\ &\quad\times \exp[-(1/2)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{b}}_i)\}^T \boldsymbol{R}_i^{-1}(\boldsymbol{\gamma}, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{b}}_i)\} - (1/2)\widehat{\boldsymbol{b}}_i^T \boldsymbol{D}^{-1}\widehat{\boldsymbol{b}}_i].\end{aligned} \tag{9.85}$$

Because the posterior mode $\widehat{\boldsymbol{b}}_i$ ***maximizes*** $\ell(\boldsymbol{b}_i)$, $\widehat{\boldsymbol{b}}_i$ must be such that $\ell'(\widehat{\boldsymbol{b}}_i) = \mathbf{0}$. From (9.83), we thus have that $\widehat{\boldsymbol{b}}_i$ must satisfy

$$\widehat{\boldsymbol{b}}_i = \boldsymbol{D}\boldsymbol{Z}_i^T(\boldsymbol{x}_i, \beta, \widehat{\boldsymbol{b}}_i)\boldsymbol{R}_i^{-1}(\gamma, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \widehat{\boldsymbol{b}}_i)\}. \tag{9.86}$$

Via ***complicated matrix algebra***, it can be shown, using the representation of $\widehat{\boldsymbol{b}}_i$ in (9.86) and defining

$$\boldsymbol{h}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i) = \boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i) - \boldsymbol{Z}_i(\boldsymbol{x}_i, \beta, \boldsymbol{b}_i)\boldsymbol{b}_i,$$

that (9.85) can be written as

$$\begin{aligned} p(\boldsymbol{Y}_i|\boldsymbol{x}_i; \beta, \gamma, \boldsymbol{D}) &\approx (2\pi)^{-n_i/2}|\boldsymbol{R}_i(\gamma, \boldsymbol{x}_i) + \boldsymbol{Z}_i(\boldsymbol{x}_i, \beta, \widehat{\boldsymbol{b}}_i)\boldsymbol{D}\boldsymbol{Z}_i^T(\boldsymbol{x}_i, \beta, \widehat{\boldsymbol{b}}_i)|^{-1/2} \\ &\times \exp[-(1/2)\{\boldsymbol{Y}_i - \boldsymbol{h}_i(\boldsymbol{x}_i, \beta, \widehat{\boldsymbol{b}}_i)\}^T\{\boldsymbol{R}_i(\gamma, \boldsymbol{x}_i) + \boldsymbol{Z}_i(\boldsymbol{x}_i, \beta, \widehat{\boldsymbol{b}}_i)\boldsymbol{D}\boldsymbol{Z}_i^T(\boldsymbol{x}_i, \beta, \widehat{\boldsymbol{b}}_i)\}^{-1}\{\boldsymbol{Y}_i - \boldsymbol{h}_i(\boldsymbol{x}_i, \beta, \widehat{\boldsymbol{b}}_i)\}] \end{aligned} \tag{9.87}$$

- Note that (9.87) has the form of a ***normal density*** with mean

  $$\boldsymbol{h}_i(\boldsymbol{x}_i, \beta, \widehat{\boldsymbol{b}}_i) = \boldsymbol{f}_i(\boldsymbol{x}_i, \beta, \widehat{\boldsymbol{b}}_i) - \boldsymbol{Z}_i(\boldsymbol{x}_i, \beta, \widehat{\boldsymbol{b}}_i)\widehat{\boldsymbol{b}}_i$$

  and covariance matrix

  $$\boldsymbol{R}_i(\gamma, \boldsymbol{x}_i) + \boldsymbol{Z}_i(\boldsymbol{x}_i, \beta, \widehat{\boldsymbol{b}}_i)\boldsymbol{D}\boldsymbol{Z}_i^T(\boldsymbol{x}_i, \beta, \widehat{\boldsymbol{b}}_i).$$

  These approximate moments are the ***same*** as those in (9.76); thus, this derivation leads to the ***same approximate moments*** obtained earlier.

- This leads ***naturally*** to replacing $\boldsymbol{b}_i$ by the ***posterior mode***. The approximation (9.87) can be substituted for the ith integral in the loglikelihood (9.49) to yield a ***closed form*** expression.

- It should be clear that maximization of this approximation to (9.49) results in ***GEE-2 estimating equations*** with the Gaussian working assumption. As noted above, and following the recommendations discussed in Chapter 8, it is standard to use the ***GEE-1*** approach instead. The form of (9.87) suggests that an ***iterative strategy*** like that described above can be used.

- When $\boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i)$ ***does depend*** on β_i, and thus on $\boldsymbol{b}_i$, the above argument ***no longer applies***, as pointed out by Vonesh (1996). It can be shown that, if $\boldsymbol{R}_i(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{b}_i)$ has the form of a ***scale parameter*** σ^2 times a matrix, then if σ is "***small***," the same approximation as in (9.87) can be obtained. Because in many applications, such as pharmacokinetics, ***within-individual variation*** is indeed "small," this further approximation is often ***relevant in practice***.

REMARKS:

- As with the approximation about $\boldsymbol{b}_i = \mathbf{0}$, it is not clear that the estimators for β and ξ obtained via the iterative strategy ***need be consistent***. For that matter, it is not clear that the procedure need even "***converge***" to a "***solution***." Luckily, in practice, it usually does.

  The Laplace approximation requires n_i to be "***large***." This suggests that if ***both*** n_i for all $i = 1, \ldots, m$ and $m \to \infty$, the estimators ***are consistent***. Vonesh (1996) discusses this in more detail.

- In fact, it should be evident that $n_i \to \infty$ ***and*** $m \to \infty$ are required to show that the estimators based on ***individual estimates*** in the previous section are consistent, as these depend on the relevance of the ***individual-level asymptotic theory approximation***. It turns out that, under these conditions, the methods we have discussed here and the "two-stage" methods are ***virtually identical***.

- In practice, the approximation discussed here often works quite well, ***even when*** n_i is ***not too large*** for all i. Simulation evidence shows that the estimator for β obtained via the iterative algorithm outlined above can be ***virtually unbiased*** for moderate m, even when n_i is not large.

- It is often assumed that the Y_{ij} are ***independent*** conditional on $\boldsymbol{x}_i, \boldsymbol{b}_i$, so that $\boldsymbol{R}_i(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{b}_i)$ is a ***diagonal matrix*** with diagonal elements equal to the assumed ***within-individual variance function***. Under these conditions, implementation of the above strategy ***simplifies*** somewhat, but the principle is the same.

GENERALIZED LINEAR MIXED EFFECTS MODELS: In the special case of a generalized linear mixed effects model as in (9.25)-(9.26), a ***similar*** approximation can be obtained. Recall here that it is assumed that the Y_{ij} are ***conditionally independent*** given β_i and $\boldsymbol{z}_i$ and are distributed according to one of the ***scaled exponential family*** distributions.

Thus, letting $p(Y_{ij}|\boldsymbol{z}_{ij}, \boldsymbol{a}_i, \boldsymbol{b}_i; \beta, \gamma)$ be the assumed ***conditional density of*** Y_{ij}, a member of the ***scaled exponential family class***, then under the conditional independence assumption,

$$p(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma) = \prod_{j=1}^{n_i} p(Y_{ij}|\boldsymbol{z}_{ij}, \boldsymbol{a}_i, \boldsymbol{b}_i; \beta, \gamma). \tag{9.88}$$

Recall from (7.12) the form of the scaled exponential family density,

$$p(y\,;\zeta,\sigma) = \exp\left\{\frac{y\zeta - b(\zeta)}{\sigma^2} + c(y,\sigma)\right\}, \tag{9.89}$$

and that $E(Y) = \mu = b_\zeta(\zeta)$, and $\text{var}(Y) = \sigma^2 b_{\zeta\zeta}\{(b_\zeta^{-1}(\mu)\} = \sigma^2 g^2(\mu)$. It can be shown (try it) by ***clever manipulations*** that

$$b(\zeta) = \int_{-\infty}^{\mu(\zeta)} \frac{u}{g^2(u)}\,du, \qquad \zeta = \int_{-\infty}^{\mu(\zeta)} \frac{1}{g^2(u)}\,du,$$

so that

$$\frac{y\zeta - b(\zeta)}{\sigma^2} = \sigma^{-2}\int_{-\infty}^{\mu} \frac{y-u}{g^2(u)}\,du.$$

It follows that (check) the quantity in the exponent in (9.89) can be written as

$$\sigma^{-2}\int_{y}^{\mu} \frac{y-u}{g^2(u)}\,du + c_*(y,\sigma), \qquad c_*(y,\sigma) = \sigma^{-2}\int_{-\infty}^{y} \frac{y-u}{g^2(u)}\,du + c(y,\sigma).$$

Thus, ignoring the term $c_*(y,\sigma)$ that depends only on the scale parameter σ^2, which is ordinarily ***known*** in popular models such as the Bernoulli/binomial and Poisson, taking $\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$ as in (9.27), the contribution for i to the likelihood, (9.47), can be written as

$$p(\boldsymbol{Y}_i|\boldsymbol{x}_i;\beta,\gamma,\boldsymbol{D}) \propto |\boldsymbol{D}|^{-1/2}\int \exp\left[\frac{1}{\sigma^2}\sum_{j=1}^{n_i}\int_{Y_{ij}}^{\mu_{ij}} \frac{Y_{ij}-u}{g^2(u)}\,du - (1/2)\boldsymbol{b}_i^T\boldsymbol{D}^{-1}\boldsymbol{b}_i\right] d\boldsymbol{b}_i. \tag{9.90}$$

In a ***famous*** paper, Breslow and Clayton (1993) suggested approximating (9.90) by using the***Laplace approximation*** (9.81). Identifying

$$\ell(\boldsymbol{b}_i) = \frac{1}{n_i\sigma^2}\sum_{j=1}^{n_i}\int_{Y_{ij}}^{\mu_{ij}} \frac{Y_{ij}-u}{g^2(u)}\,du - \frac{1}{2n_i}\boldsymbol{b}_i^T\boldsymbol{D}^{-1}\boldsymbol{b}_i,$$

$$\ell'(\boldsymbol{b}_i) = -n_i^{-1}\boldsymbol{D}^{-1}\boldsymbol{b}_i + \sigma^{-2}\sum_{j=1}^{n_i} \frac{Y_{ij} - f(\boldsymbol{z}_{ij},\boldsymbol{a}_i,\beta,\boldsymbol{b}_i)}{g^2\{f(\boldsymbol{z}_{ij},\boldsymbol{a}_i,\beta,\boldsymbol{b}_i)\}}\partial/\partial\boldsymbol{b}_i\,\{f(\boldsymbol{z}_{ij},\boldsymbol{a}_i,\beta,\boldsymbol{b}_i)\}.$$

Letting

$$\boldsymbol{R}_i(\beta,\gamma,\boldsymbol{x}_i,\boldsymbol{b}_i) = \sigma^2\text{diag}[g^2\{f(\boldsymbol{z}_{i1},\boldsymbol{a}_i,\beta,\boldsymbol{b}_i)\},\ldots,g^2\{f(\boldsymbol{z}_{in_i},\boldsymbol{a}_i,\beta,\boldsymbol{b}_i)\}]$$

and $\boldsymbol{Z}_i(\boldsymbol{x}_i,\beta,\boldsymbol{b}_i)$ be the $(n_i \times q)$ matrix with jth row $[\partial/\partial\boldsymbol{b}_i\,\{f(\boldsymbol{z}_{ij},\boldsymbol{a}_i,\beta,\boldsymbol{b}_i)\}]^T$, we can write this as

$$\ell'(\boldsymbol{b}_i) = -n_i^{-1}\boldsymbol{D}^{-1}\boldsymbol{b}_i + \boldsymbol{Z}_i^T(\boldsymbol{x}_i,\beta,\boldsymbol{b}_i)\boldsymbol{R}_i^{-1}(\beta,\gamma,\boldsymbol{x}_i,\boldsymbol{b}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta,\boldsymbol{b}_i)\}. \tag{9.91}$$

This looks identical to (9.83) in the case of conditionally normal $\boldsymbol{Y}_i$, with the exception that $\boldsymbol{R}_i$ depends on $\boldsymbol{b}_i$.

Differentiating (9.91) again with respect to $\boldsymbol{b}_i$, and, as with (9.84), ignoring the term with expectation zero, yields

$$\ell''(\boldsymbol{b}_i) \approx -n_i^{-1}\{\boldsymbol{D}^{-1} + \boldsymbol{Z}_i^T(\boldsymbol{x}_i,\beta,\boldsymbol{b}_i)\boldsymbol{R}_i^{-1}(\beta,\gamma,\boldsymbol{x}_i,\boldsymbol{b}_i)\boldsymbol{Z}_i(\boldsymbol{x}_i,\beta,\boldsymbol{b}_i)\}.$$

Substituting these expressions into (9.81), we obtain, ignoring constants,

$$p(\boldsymbol{Y}_i|\boldsymbol{x}_i;\beta,\gamma,\boldsymbol{D}) \dot{\propto} |\boldsymbol{D}|^{-1/2}|\boldsymbol{D}^{-1} + \boldsymbol{Z}_i^T(\boldsymbol{x}_i,\beta,\widehat{\boldsymbol{b}}_i)\boldsymbol{R}_i^{-1}(\beta,\gamma,\boldsymbol{x}_i,\widehat{\boldsymbol{b}}_i)\boldsymbol{Z}_i(\boldsymbol{x}_i,\beta,\widehat{\boldsymbol{b}}_i)|^{-1/2}$$
$$\times \exp\left[\frac{1}{\sigma^2}\sum_{j=1}^{n_i}\int_{Y_{ij}}^{f(z_{ij},a_i,\beta,\hat{b}_i)}\frac{Y_{ij}-u}{g^2(u)}\,du - (1/2)\widehat{\boldsymbol{b}}_i^T\boldsymbol{D}^{-1}\widehat{\boldsymbol{b}}_i\right] d\boldsymbol{b}_i. \quad (9.92)$$

Through additional manipulations that are left as an exercise for the ***diligent student***, it can be shown that (9.92) leads to a ***further approximation*** in terms of a ***linear mixed effects model*** representation, as in the case of normal conditional distribution. Thus, essentially the ***same iterative strategy*** discussed in that case is applicable here and is implemented in the SAS macro `glimmix`. Wolfinger and O'Connell (1993) discuss a slightly different derivation; see also Schall (1991).

In the context of generalized linear mixed effects models, the iterative scheme is referred to as ***penalized quasi-likelihood*** (PQL); see Breslow and Clayton (1993).

Of course, the same comments about ***consistency*** of the estimators given in the normal case apply here as well.

DANGER: Under some circumstances, the approximation underlying these developments can be ***very poor***. In particular, when the Y_{ij} are ***binary*** and n_i are ***small***, it has been observed that the resulting estimators for β, γ, and $\boldsymbol{D}$, and ***particularly the latter***, can show ***nontrivial bias*** in practice. This is discussed by Breslow and Lin (1995) and Lin and Breslow (1996), who propose analytical approaches for ***correcting*** this bias. Alternatively, a number of authors have suggested that the only way around this problem is to try to "***do***" the integral in (9.71) more directly.

9.6 "Exact" likelihood inference

The approximate methods of the last two sections often work ***remarkably well*** in practice in that the resulting estimators for β, γ, and $\boldsymbol{D}$ are ***approximately unbiased*** in finite samples (finite m and n_i). However, as noted for ***binary*** response at the end of Section 9.5, sometimes these approximations ***fail***.

Accordingly, implementing the nonlinear and generalized linear mixed effects models by maximizing the loglikelihood (9.49) ***without*** resorting to ***analytical approximation*** is desirable.

QUADRATURE: We mentioned the use of ***quadrature*** in Section 9.3. Here, we review generically the main idea of Gaussian quadrature. Let $\varphi(z)$ be the ***standard normal density*** , and let $f(z)$ be a known function. Gaussian quadrature is designed to approximate an integral of the form

$$\int f(z)\,\varphi(z)\,dz. \tag{9.93}$$

In particular, the integral (9.93) is approximated by the ***weighted sum***

$$\int f(z)\,\varphi(z)\,dz \approx \sum_{\ell=1}^{L} w_\ell f(z_\ell), \tag{9.94}$$

where w_ℓ are appropriately chosen ***weights*** , and the ***abscissæ*** z_ℓ are solutions to the Lth order ***Hermite polynomial***. Standard algorithms are available for calculating the abscissæ and weights.

This can be used to approximate the integrals in (9.47),

$$\int p(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)\, p(\boldsymbol{b}_i; \boldsymbol{D})\, d\boldsymbol{b}_i$$

by invoking a change of variables to $\boldsymbol{z}_i = \boldsymbol{D}^{-1/2}\boldsymbol{b}_i$, so that $\boldsymbol{z}_i$ is has a standard multivariate normal distribution. It is then possible to transform the problem of evaluating this integral to one of successive applications of the one-dimensional quadrature rule (9.94). This is demonstrated by Pinheiro and Bates (1995) and Davidian and Giltinan (1995, Section 7.3).

A drawback of the standard quadrature rule (9.94) is that the abscissæ are chosen based solely on $\varphi(x)$, so ***regardless*** of the form of the function $f(z)$. As a result, the z_ℓ may or may not lie in the relevant region of integration, depending on the support of $f(z)$. A modification referred to as ***adaptive Gaussian quadrature*** was proposed by Pinheiro and Bates (1995) to center the abscissæ, when transformed to the scale of the $\boldsymbol{b}_i$, around the ***posterior modes*** $\widehat{\boldsymbol{b}}_i$ rather than $\mathbf{0}$ and to scale them appropriately. The result is that the abscissæ will tend to lie in the region of interest, meaning that L can be chosen to be smaller and achieve the same accuracy as ordinary quadrature.

See Pinheiro and Bates (1995) for a detailed demonstration of how all this works in the case of the nonlinear mixed effects model with $p(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)$ a ***normal*** density. Gaussian quadrature and adaptive Gaussian quadrature are implemented in SAS `proc nlmixed`, and adaptive quadrature is available in SAS `proc glimmix` for fitting generalized linear mixed effects models.

The paper by Pinheiro and Bates (1995) reviews and compares several other ways of approximating the integrals and is the best resource for understanding the details.

EM ALGORITHMS: Another approach is to use the ***EM*** algorithm, regarding the random effects as "missing data" as for the linear mixed effects model, discussed in Section 6.5. Because of the ***nonlinearity*** of the model in $\boldsymbol{b}_i$, it is no longer possible to evaluate ***analytically in a closed form*** the ***conditional expectations*** given the observed data of "full data" sufficient statistics. Instead, evaluation of the required conditional expectations for this intractable ***E-step*** involves ***integration***.

In the context ***generalized linear mixed effects models*** , McCulloch (1997) and Booth and Hobert (1999) proposed using various versions of ***Monte Carlo simulation*** to evaluate the conditional expectations. Walker (1996) describes an EM algorithm that uses ***Monte Carlo integration*** to compute the E-step. These algorithms can achieve good accuracy is computationally intensive but can be ***slow to converge***. Alternatively, quadrature can be used to compute the E-step.

A nice reference on generalized linear mixed effects models that shows some of these implementations in detail is Rabe-Hesketh and Skrondal (2009).

MISSING DATA: We conclude this section by highlighting considerations for ***missing data***.

Following the same considerations presented in Section 5.6, under the ***assumptions*** of a ***missing at random*** (MAR) missingness mechanism and the ***separability condition*** , so that ***ignorability*** holds, likelihood-based inference for the nonlinear mixed effects model based on the ***observed data*** will be ***valid***.

- That is, ***assuming that the model is exactly, correctly specified*** , the estimators for β and ξ obtained by maximizing (9.49) using the observed data will be ***consistent*** for the true values of the parameters.
- Because methods for implementing the nonlinear mixed effects model based on "exact" calculation of the likelihood are likely to get ***closer*** to achieving maximization of the "true" loglikelihood than those based on ***analytical approximation*** , they are to be preferred in this situation.
- Although consistent estimators can be obtained via this approach under MAR, as in Section 5.6, assuming that m is ***sufficiently large*** for asymptotic theory to be relevant, ***standard errors*** calculated based on the ***expected information matrix*** will ***misstate*** the true sampling variability. Standard errors should ideally be derived based on the ***observed information matrix*** , as discused in Section 5.6.

9.7 Examples

In this section we present some examples.

PHARMACOKINETIC STUDY OF ARGATROBAN: This example is taken from Davidian and Giltinan (1995, section 9.5) and concerns a study of the pharmacokinetics and pharmacodynamics of the anti-coagulant agent argatroban conducted at the biotechnology company Genentech. We consider the pharmacokinetic data here.

In the study, m = 37 subjects each received a four hour (240 minute) ***intravenous infusion*** of one of several doses of argatroban. For each infusion rate from 1 μg/kg/min to 5 μg/kg/min in increments of 0.5 μg/kg/min, 4 subjects were randomized to receive that infusion rate; a 37th subject received a rate of 4.37 μg/kg/min. Serial blood samples were taken from each patient at several time points over the 360 minutes (6 hours) following the start of the infusion and were assayed for argatroban concentration. Figure 9.3 shows concentration-time profiles for 4 subjects at different doses, with a fit of the pharmacokinetic model given below superimposed.

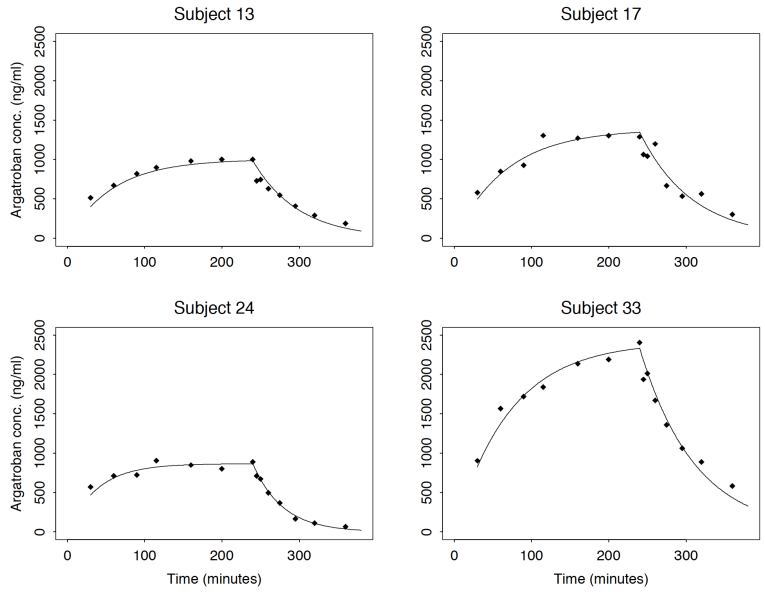

Figure 9.3: *Concentration-time data for four subjects from the argatroban pharmacokinetic study.*

For each individual i with infusion rate R_i, we thus have concentration measurements $Y_{i1}, \ldots, Y_{in_i}$ at times $t_{i1}, \ldots, t_{in_i}$, so that $\boldsymbol{z}_{ij} = (R_i, t_{ij})^T$. Here, n_i is in the range of 10 to 14 for each.

A standard model for concentration at time t_{ij} during and following administration by a ***constant-rate infusion*** of amount per unit time R_i of duration t_{inf} is the one-compartment model

$$f(\boldsymbol{z}_{ij}, \boldsymbol{\beta}_i) = \frac{R_i}{Cl_i}\left\{\exp\left(-\frac{Cl_i}{V_i}t^*_{ij}\right) - \exp\left(-\frac{Cl_i}{V_i}t_{ij}\right)\right\}, \tag{9.95}$$

where

$$t^*_{ij} = 0 \text{ for } t_{ij} \leq t_{inf}; t^*_{ij} = t_{ij} - t_{inf} \text{ for } t_{ij} > t_{inf},$$

Cl_i and V_i are the clearance rate and volume of distribution for subject i, and $\boldsymbol{\beta}_i = (\beta_{1i}, \beta_{i2})^T$ is defined so that

$$Cl_i = \exp(\beta_{1i}), \quad V_i = \exp(\beta_{2i}).$$

From Figure 9.3, this model appears to provide an adequate representation of the concentration time relationship. A fundamental assumption of the model is that Cl_i and V_i are ***not dose (infusion-rate) dependent***; they do not depend on R_i. This assumption means that an individual's clearance and volume characteristics do not change depending on the dose administered. Thus, in principle, information on the values of Cl_i and V_i for a particular individual can be obtained from concentration-time data at any dose R_i and thus information on the population distribution these parameters can be obtained from individuals receiving different doses.

The inclusion of different doses was so that subjects would achieve different concentrations, as seen in Figure 9.3, facilitating investigation of the relationship between concentration and a response, ***activated partial thromboplastin time*** (aPPT), roughly, a measure of how long it takes the blood to clot. We do not report on this ***pharmacodynamic analysis*** here.

We assume the ***stage 1 individual model*** , where (9.95) describes the concentration-time relationship for each subject, so that each individual has parameters $\boldsymbol{\beta}_i = (\beta_{1i}, \beta_{2i})^T$. As it is well known that the within-individual variance of pharmacokinetic concentration measurements is ***nonconstant*** and likely dominated by measurement error. we assume that

$$\text{var}(Y_{ij}|\boldsymbol{z}_{ij}, \boldsymbol{\beta}_i) = \sigma^2 f^{2\delta}(\boldsymbol{z}_{ij}, \boldsymbol{\beta}_i),$$

the power-of-the-mean variance model. Here, σ^2 and δ are ***common across subjects***; this is reasonable if the major source of within-individual variation is indeed measurement error due to the ***assay*** used to determine argatroban concentrations. That this variance is nonconstant is supported by plots of pooled within-individual residuals obtained from either individual OLS fits or from empirical Bayes estimates from a fit of the nonlinear mixed effects model here assuming constant within-individual variance, not shown here.

Here, there are no among-individual covariates $\boldsymbol{a}_i$; the infusion rate R_i is a within-individual covariate, as it is a condition of measurement. We also assume that the Y_{ij} are ***conditionally independent*** given $\boldsymbol{x}_i = \boldsymbol{z}_i$ and $\boldsymbol{b}_i$.

With no among-individual covariates and the model parameterized as above in terms of the ***logarithms*** of the PK parameters, we take the second stage population model to be

$$\beta_i = \beta + \boldsymbol{b}_i.$$

Assuming the default $\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$, D_{11} and D_{22} represent the variances of log *Cl* and log *V* values in the population, so are roughly the squares of the coefficients of variation of these parameters in the population.

Programs on the course webpage show fits of this nonlinear mixed effects models using several methods implemented in different software.

EXAMPLE 3: Growth of two different soybean genotypes, continued. As discussed in Section 9.2, we consider the stage 1 individual model as in (9.28),

$$E(Y_{ij}|\boldsymbol{z}_{ij}, \beta_i) = f(\boldsymbol{z}_{ij}, \beta_i) = \frac{\beta_{1i}}{1 + \exp\{-\beta_{3i}(t_{ij} - \beta_{2i})\}}, \qquad \text{var}(Y_{ij}|\boldsymbol{z}_{ij}, \beta_i) = \sigma^2 f^{2\delta}(\boldsymbol{z}_{ij}, \beta_i), \tag{9.96}$$

$\beta_i = (\beta_{1i}, \beta_{2i}, \beta_{3i})^T$, where in (9.96) we have used an ***alternative parameterization*** of the ***logistic growth model***. This model was adopted by Davidian and Giltinan (1995, Section 11.2), and Pinheiro and Bates (2000, Section 6.3) used a related parameterization. Pinheiro and Bates plotted empirical Bayes estimates of pooled within-individual residuals versus fitted values from a fit of the nonlinear mixed effects model below assuming constant within-individual variance and including no among-individual covariates; this plot showed strong evidence of non-constant within-plot variance.

These authors chose to rewrite the model in terms of the "***soybean half-life***" parameter β_{2i} in (9.96), the time at which 50% of the final asymptotic growth is achieved, based on informal diagnostic plots of empirical Bayes estimates of the parameters from a fit of a nonlinear mixed effects model using (9.28) with no stage 2 among-individual covariates suggesting that β_{2i} in the original model is not normally distributed in the population. These diagnostics for the other parameters did not suggest major departures from normality.

Because the goal of the study was to investigate whether or not there are systematic associations between the components of β_i, especially ***asymptotic growth*** β_{1i}, and the among individual covariates δ_i (genotype) and w_i (year/weather condition), we consider a general stage 2 population model allowing a separate population mean ("typical value") for each year-genotype combination for each of β_{1i}, β_{2i}, and β_{3i}. The general initial population model is thus

$$\begin{aligned}
\beta_{1i} &= \beta_{11}(1-\delta_i)I(w_i=1) + \beta_{12}\delta_i I(w_i=1) + \beta_{13}(1-\delta_i)I(w_i=2) + \beta_{14}\delta_i I(w_i=2) \\
&\quad + \beta_{15}(1-\delta_i)I(w_i=3) + \beta_{16}\delta_i I(w_i=3) + b_{1i} \\
\beta_{2i} &= \beta_{21}(1-\delta_i)I(w_i=1) + \beta_{22}\delta_i I(w_i=1) + \beta_{23}(1-\delta_i)I(w_i=2) + \beta_{24}\delta_i I(w_i=2) \\
&\quad + \beta_{25}(1-\delta_i)I(w_i=3) + \beta_{26}\delta_i I(w_i=3) + b_{2i} \qquad (9.97) \\
\beta_{3i} &= \beta_{31}(1-\delta_i)I(w_i=1) + \beta_{32}\delta_i I(w_i=1) + \beta_{33}(1-\delta_i)I(w_i=2) + \beta_{34}\delta_i I(w_i=2) \\
&\quad + \beta_{35}(1-\delta_i)I(w_i=3) + \beta_{36}\delta_i I(w_i=3) + b_{3i}
\end{aligned}$$

Fits of the nonlinear mixed effects model given by (9.96)-(9.97) and of a simpler, preliminary model with no among-individual covariates in stage 2 using several methods implemented in different software on the course webpage.

PHARMACOKINETICS OF PHENOBARBITAL IN NEONATES, continued. As discussed in Section 9.2, we consider the stage 1 individual model using the one-compartment model with repeated dosing; i.e., if infant *i* has dosing history

$$(s_{i\ell}, D_{i\ell}),\ \ell : s_{i\ell} < t,$$

the model is as in (9.31), namely,

$$E(Y_{ij}|\boldsymbol{z}_{ij},\beta_i) = f(\boldsymbol{z}_{ij},\beta_i) = \sum_{\ell: s_{i\ell}<t_{ij}} \frac{D_{i\ell}}{e^{\beta_{2i}}} \exp\left\{-\frac{e^{\beta_{1i}}}{e^{\beta_{2i}}}(t_{ij}-s_{i\ell})\right\}, \qquad \text{var}(Y_{ij}|\boldsymbol{z}_{ij},\beta_i) = \sigma^2 f^{2\delta}(\boldsymbol{z}_{ij},\beta_i), \quad (9.98)$$

where (9.98) is parameterized in terms of $\beta_i = (\log Cl_i, \log V_i)^T$. We take $\delta = 1$ in the analyses discussed below.

On the course webpage, we use `nlme()` to fit (9.98) with three different stage 2 population models.

$$(i) \quad \beta_{1i} = \beta_1 + b_{1i}, \quad \beta_{2i} = \beta_2 + b_{2i},$$

a model with no systematic associations with the among-individual covariates birthweight w_i and Apgar score a_i;

$$(ii) \quad \beta_{1i} = \beta_1 + \beta_3 w_i + b_{1i}, \quad \beta_{2i} = \beta_2 + \beta_4 w_i + b_{2i},$$

allowing systematic association of each PK parameter on birthweight; and

$$(iii) \quad \beta_{1i} = \beta_1 + \beta_3 w_i + \beta_5 I(a_i \geq 5) + b_{1i}, \quad \beta_{2i} = \beta_2 + \beta_4 w_i + \beta_6 I(a_i \geq 5) b_{2i},$$

allowing additional association with Apgar category.

Assuming the default $\boldsymbol{b}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{D})$, D_{11} and D_{22} represent the variances of log *Cl* and log *V* values in the population, so are roughly the squares of the coefficients of variation of these parameters in the population.

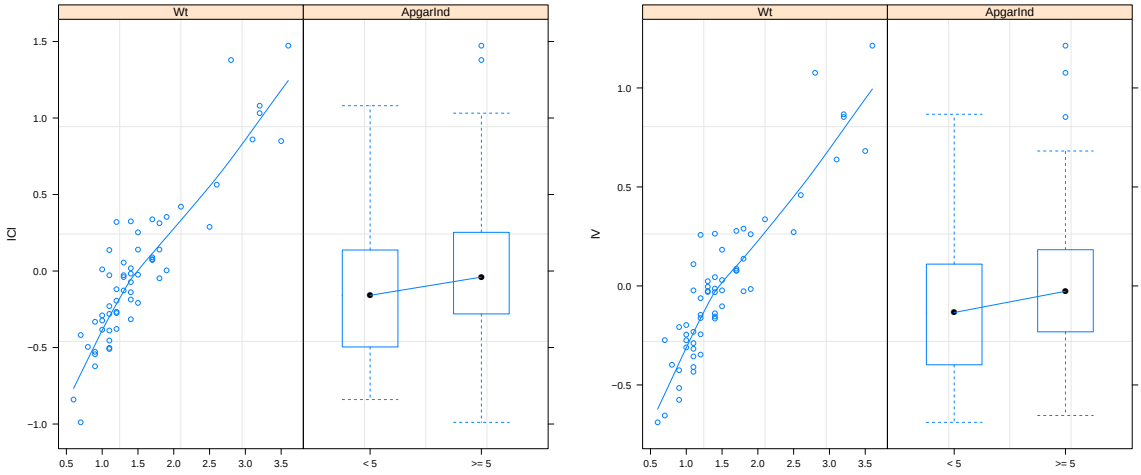

Figure 9.4: *Empirical Bayes estimates of the random effects b_{1i} and b_{2i} from a fit of model (i) plotted against the covariates birthweight and Apgar score category..*

Figure 9.4 shows plots of the empirical Bayes estimates of the random effects from a fit of model (i). The plot suggests a strong association of each PK parameter with birthweight, with systematic association with Apgar category much less clear. On the course webpage, fits of models (i) - (iii) show that, while adding birthweight the population model for each PK parameter, as in (ii), greatly improves the fit, there does not seem to be strong enough evidence to suggest the need for model (iii). See the course webpage for these fits and more plots.

10 Additional Topics

10.1 Introduction

Chapters 1-9 present an overview of the ***fundamental topics*** that ordinarily comprise a first course on longitudinal data analysis. Understanding of these fundamentals is preparation for study of ***additional topics*** connected with modeling and analysis of longitudinal and other correlated/clustered data and for reading the current literature on longitudinal data methods.

In this chapter, we provide an introduction to several of these additional topics.

DATA, RESTATED: As in previous chapters, the observed data are

$$(\boldsymbol{Y}_i, \boldsymbol{z}_i, \boldsymbol{a}_i) = (\boldsymbol{Y}_i, \boldsymbol{x}_i), \quad , i = 1, \ldots, m, \tag{10.1}$$

independent across i, where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$, with Y_{ij} recorded at time t_{ij}, $j = 1, \ldots, n_i$ (possibly different times for different individuals); $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}^T, \ldots, \boldsymbol{z}_{in_i}^T)^T$, comprising ***within-individual*** covariate information $\boldsymbol{u}_i$ and the ***times*** t_{ij}; $\boldsymbol{a}_i$ is a vector of ***among-individual*** covariates; and $\boldsymbol{x}_i = (\boldsymbol{z}_i^T, \boldsymbol{a}_i^T)^T$. We take the ***among-individual*** covariates to be ***time-independent*** unless otherwise stated.

10.2 Bayesian formulation of hierarchical models

BAYESIAN VS. FREQUENTIST INFERENCE: We have restricted attention to so-called frequentist-type inference. In particular, for most of the methods discussed, we have considered model parameters to be ***fixed quantities*** and viewed data to be a ***repeatable sample*** arising as a result of some generative data process (e.g., according to an experimental design or data gathering scheme). Inference regarding the (fixed) values of the parameters of interest is by reference to (ideally) an exact, or in most cases, an ***approximate*** (derived via ***large sample theory***), ***sampling distribution***. In this view, probability refers to the conceptual repeated sampling if the generative process were to be repeated infinitely.

Bayesian inference is based on a different point of view in which the data are viewed as ***fixed***, model parameters are viewed as ***random***, and, in the classical Bayesian paradigm, probability refers to "***degree of belief***" regarding the values of the parameters.

Given a ***prior distribution*** specifying degree of belief before the data are observed, ***Bayes theorem*** is used to obtain the ***posterior distribution***, reflecting the ***updated*** degree of belief having seen the data. "Estimation" of parameters is usually based on the ***mode*** of the posterior density, and assessments of uncertainty are based on the variance (standard deviation) of the posterior.

It can be shown that, despite the apparent divergence of these frameworks, they can lead to very ***similar results***. Thus, in modern statistics, they are often viewed as ***complementary strategies*** for framing scientific inquiry. Owing to the ***high-dimensional integration*** involved in applying Bayes theorem in complex models, before the latter half of the twentieth century, it was often ***prohibitive*** to implement Bayesian models. However, with the computational advances that began in the last quarter of the twentieth century, and in particular the use and refinement of ***Markov chain Monte Carlo*** (MCMC) techniques, formulation and implementation of complex statistical models from a Bayesian perspective is now ***commonplace***. Indeed, with ***noninformative prior*** specifications, models of both types overlap, and a Bayesian framework with fitting via MCMC techniques is sometimes viewed as a convenient way to implement frequentist analyses.

HIERARCHICAL MODELS FROM A BAYESIAN PERPSECTIVE: Hierarchical models such as the linear, generalized linear, and nonlinear mixed effects models discussed in Chapters 6 and 9 are placed ***naturally*** in a Bayesian framework, as we now demonstrate. Because the latter models subsume the former, we present this in the case of a general nonlinear mixed effects model.

From the Bayesian perspective, the model is a ***three-stage*** hierarchy, and the model parameters β, γ, and $\boldsymbol{D}$ involved in the usual two-stage hierarchy (9.8)-(9.9) are viewed as ***random vectors***. In a ***classical*** Bayesian formulation, ***full parametric distributional assumptions*** are made at each stage, although this can be relaxed, in particular for the distribution of the random effects discussed briefly later. For our discusion here, we adopt full parametric assumptions, so we write the stage 1 model differently from the general specification in (9.8), as follows.

Stage 1 - Individual model. Given a model f for $E(Y_{ij}|\boldsymbol{z}_{ij}, \beta_i$ possibly ***nonlinear*** in β_i, the random vectors $\boldsymbol{Y}_i$, $i = 1, \ldots, m$, are assumed to satisfy

$$\boldsymbol{Y}_i|\boldsymbol{z}_i, \beta_i \sim p(\boldsymbol{y}_i|\boldsymbol{z}_i, \beta_i; \gamma), \tag{10.2}$$

where $p(\boldsymbol{y}_i|\boldsymbol{z}_i, \beta_i; \gamma)$ is a parametric density such that

$$E(\boldsymbol{Y}_i|\boldsymbol{z}_i, \beta_i) = \boldsymbol{f}_i(\boldsymbol{z}_i, \beta_i), \quad \text{var}(\boldsymbol{Y}_i|\boldsymbol{z}_i, \beta_i) = \boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i).$$

For continuous responses, $p(\boldsymbol{y}_i|\boldsymbol{z}_i, \beta_i; \gamma)$ in (10.2) is ordinarily taken to be the $\mathcal{N}\{\boldsymbol{f}_i(\boldsymbol{z}_i, \beta_i), \boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i)\}$ density.

For responses of the "generalized linear model type," the Y_{ij}, $j = 1, \ldots, n_i$, are taken to be ***conditionally independent*** given $\boldsymbol{z}_i$ and β_i, so that

$$p(\boldsymbol{y}_i|\boldsymbol{z}_i, \beta_i; \gamma) = \prod_{j=1}^{n_i} p(y_{ij}|\boldsymbol{z}_{ij}, \beta_i; \gamma);$$

and $p(y_{ij}|\boldsymbol{z}_{ij}, \beta_i; \gamma)$ is a ***scaled exponential family*** density for each j appropriate to the form of the response, so that $\boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i)$ is of necessity a ***diagonal matrix*** whose diagonal elements are determined by the ***variance function*** corresponding to the particular scaled exponential family density.

Stage 2 - Population model. The individual-specific parameter β_i is assumed to be a function of ***among-individual covariates*** $\boldsymbol{a}_i$, ***fixed effects*** β ($p \times 1$), and ***random effects*** $\boldsymbol{b}_i$ ($q \times 1$), namely,

$$\beta_i = \boldsymbol{d}(\boldsymbol{a}_i, \beta, \boldsymbol{b}_i), \qquad \boldsymbol{b}_i \sim p(\boldsymbol{b}_i|\boldsymbol{D}), \tag{10.3}$$

where $\boldsymbol{d}$ is a k-dimensional vector of possibly ***nonlinear*** functions of $\boldsymbol{a}_i$, β, and $\boldsymbol{b}_i$; and $\boldsymbol{b}_i$ density $p(\boldsymbol{b}_i|\boldsymbol{D})$, usually assumed to be the $\mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$ density. As in Chapter 9, this can be relaxed to allow the distribution of $\boldsymbol{b}_i$ to depend on $\boldsymbol{a}_i$.

In the classical Bayesian literature, the distribution of $\boldsymbol{b}_i$ or that implied for β_i are sometimes confusingly referred to as the ***prior distribution***.

Stage 3 - Hyperprior distribution. $(\beta, \gamma, \boldsymbol{D})$ are assumed to have joint density

$$(\beta, \gamma, \boldsymbol{D}) \sim p(\beta, \gamma, \boldsymbol{D}|\Omega), \tag{10.4}$$

where Ω are ***known hyperparameters*** characterizing this density.

REMARKS:

- From the classical Bayesian viewpoint, the hyperprior reflects ***prior beliefs*** about the values of β, γ, and $\boldsymbol{D}$.
- Ordinarily, in practice, the joint density (10.4) is written as

  $$p(\beta, \gamma, \boldsymbol{D}|\Omega) = p(\beta|\Omega_1)\, p(\gamma|\Omega_2)\, p(\boldsymbol{D}|\Omega_3), \tag{10.5}$$

  so that β, γ, and $\boldsymbol{D}$ are taken to be independent, and hyperpriors for them are specified separately. The individual components (10.5) of the hyperprior (10.4) are usually taken to reflect ***weak knowledge*** of the values of the parameters.

- An advantage of this formulation is that the hyperprior is a natural way to incorporate ***historical*** or other information about the parameters in to the overall model. For example, if it is known from past studies that there is a ***range of plausible values*** for β, the hyperprior for β can be taken to concentrate on that range.

- This has been used advantageously in implementation of ***highly complex*** models for ***pharmacokinetics***, ***viral dynamics***, and other phenomena that can be represented as ***compartmental systems*** involving numerous parameters, as we discuss shortly.

- As an example, in the special case where $p(\boldsymbol{y}_i|\boldsymbol{z}_i, \beta_i; \gamma)$ is the $\mathcal{N}\{\boldsymbol{f}_i(\boldsymbol{z}_i, \beta_i), \boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i)\}$ distribution, where $\boldsymbol{R}_i(\beta_i, \gamma, \boldsymbol{z}_i) = \sigma^2 \text{diag}\{f^{2\delta}(\boldsymbol{z}_{i1}, \beta_i), \ldots, f^{2\delta}(\boldsymbol{z}_{in_i}, \beta_i)\}$, and the Y_{ij} are conditionally independent with $\gamma = (\sigma^2, \delta)^T$, common specifications for the components of (10.5) are

$$\beta \sim \mathcal{N}(\beta^*, \boldsymbol{H}), \qquad \sigma^{-2} \sim \text{Ga}(\nu_0/2, \nu_0\tau_0/2),$$

$$\boldsymbol{D}^{-1} \sim \text{Wi}\{(\rho\boldsymbol{D}^*)^{-1}, \rho\}, \qquad \delta \sim U(0, \delta_0), \tag{10.6}$$

  where $\text{Ga}(\cdot, \cdot)$ $\text{Wi}(\cdot, \cdot)$, and $U(\cdot, \cdot)$ denote the gamma, Wishart, and uniform distributions, respectively; and the ***hyperparameters*** β^*, $\boldsymbol{H}$, ν_0, τ_0, ρ, $\boldsymbol{D}^*$, and δ_0 are taken to be known.

- Hyperprior specifications for ***generalized linear mixed effects models*** are discussed by Rabe-Hesketh and Skrondal (2009) and references therein; those for ***more complex*** general nonlinear mixed models, and especially in the context of pharmacokinetics, are discussed by Wakefield et al. (1994), Wakefield (1996), Rosner and Müller (1994), and Müller and Rosner (1997), among many others.

POSTERIOR DISTRIBUTIONS: Given the above hierarchy, it is straightforward to deduce expressions for the ***marginal posterior densities*** of each of the parameter β, γ, and $\boldsymbol{D}$. We can express the stage 1 individual model (10.2) density equivalently by substituting the stage 2 population model (10.3) as

$$p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma). \tag{10.7}$$

Letting as usual $\boldsymbol{Y}$ and $\boldsymbol{b}$ denote the "stacked" vectors of the individual response vectors $\boldsymbol{Y}_i$ and random effects $\boldsymbol{b}_i$, by the conditional independence of $\boldsymbol{Y}_i$ given $\boldsymbol{z}_i$ and $\boldsymbol{b}_i$ and among the $\boldsymbol{b}_i$, the densities of $\boldsymbol{Y}$ and $\boldsymbol{b}$ are the products of the individual densities in (10.7) and (10.3), which we can write as

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{b}; \beta, \gamma) \quad \text{and} \quad p(\boldsymbol{b}|\boldsymbol{D}).$$

We take the entire analysis to be conditional on $\boldsymbol{x}$, as in prior chapters. Then the ***joint posterior density*** of $(\beta, \gamma, \boldsymbol{D})$, where we implicitly condition on $\boldsymbol{x}$ and recall that the hyperparameters Ω are ***known***, is seen to be

$$p(\beta, \gamma, \boldsymbol{D}|\boldsymbol{y}; \boldsymbol{x}, \Omega) = \frac{p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{b}; \beta, \gamma)\, p(\boldsymbol{b}|\boldsymbol{D})\, p(\beta, \gamma, \boldsymbol{D}|\Omega)}{\int\int\int\int p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{b}; \beta, \gamma)\, p(\boldsymbol{b}|\boldsymbol{D})\, p(\beta, \gamma, \boldsymbol{D}|\Omega)\, d\beta\, d\gamma\, d\boldsymbol{b}\, d\boldsymbol{D}}. \tag{10.8}$$

To obtain the ***marginal posterior densities*** of β, γ, and $\boldsymbol{D}$, whose ***modes*** are the Bayesian "***estimators***" it is necessary ***integrate*** (10.8) with respect to the rest of the parameter; e.g., to obtain the posterior for β, $p(\beta|\boldsymbol{y}; \boldsymbol{x}, \Omega)$,

$$p(\beta|\boldsymbol{y}; \boldsymbol{x}, \Omega) = \frac{\int\int\int p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{b}; \beta, \gamma)\, p(\boldsymbol{b}|\boldsymbol{D})\, p(\beta, \gamma, \boldsymbol{D}|\Omega)\, d\gamma\, d\boldsymbol{b}\, d\boldsymbol{D}}{\int\int\int\int p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{b}; \beta, \gamma)\, p(\boldsymbol{b}|\boldsymbol{D})\, p(\beta, \gamma, \boldsymbol{D}|\Omega)\, d\beta\, d\gamma\, d\boldsymbol{b}\, d\boldsymbol{D}} \tag{10.9}$$

- Clearly, the potentially ***high-dimensional integration*** in (10.9) and the analogous expressions for the marginal posteriors of the other parameters is ***not analytically tractable*** in general.

IMPLEMENTATION BY MARKOV CHAIN MONTE CARLO SIMULATION: In the early 1990s, the use of ***MCMC techniques*** as a way to "***do***" the required integrations ***numerically*** was popularized. These techniques employ a clever scheme that leads to ***simulated draws*** from the posterior distribution of the parameters. These simulated values can them be used to ***construct numerically*** any ***functional of the posterior distribution*** desired; e.g., posterior modes for each of the parameters and the posterior variance.

A course in Bayesian inference covers these methods in detail; here, we just remark on the simplest version of this technique, the ***Gibbs sampler***, which relies on the premise that the conditional distributions of each parameter given all the others and the data might have forms from which it is ***straightforward*** to simulate random deviates.

Generically, the basic idea is as follows. Suppose we have J random variables $(U_1, \ldots, U_J)$ with joint density $p(u_1, \ldots, u_J)$, and we would like to find the marginal distributions $p(u_j)$, $j = 1, \ldots, J$. Assume that the joint density is uniquely determined (not automatic!) by the ***full conditional densities***

$$p(u_j|u_1, \ldots, u_{j-1}, u_{j+1}, \ldots, u_J), \quad j = 1, \ldots, J,$$

from which it is "easy" to sample.

The Gibbs sampler is an ***iterative algorithm*** for obtaining a sample from the joint distribution based on the full conditional distributions that can then be used to obtain the marginals. Given a set of initial values $(u_1^{(0)}, \ldots, u_J^{(0)})$, at the ℓth iteration, one generates random variates from the full conditionals as follows:

$$\begin{aligned} U_1^{(\ell+1)} &\sim p(u_1 | u_2^{(\ell)}, u_3^{(\ell)}, \ldots, u_J^{(\ell)}) \\ U_2^{(\ell+1)} &\sim p(u_2 | u_1^{(\ell+1)}, u_3^{(\ell)}, \ldots, u_J^{(\ell)}) \\ &\vdots \\ U_J^{(\ell+1)} &\sim p(u_J | u_1^{(\ell+1)}, u_2^{(\ell+1)}, \ldots, u_{J-1}^{(\ell+1)}) \end{aligned} \tag{10.10}$$

After T iterations, we have a realization of the random vector $(U_1^{(T)}, \ldots, U_J^{(T)})$. It can be shown that the sequence generated in this way is a ***Markov chain*** with ***stationary distribution*** $p(u_1, \ldots, u_J)$. It thus follows that, as $T \to \infty$, this random vector ***tends in distribution*** to a draw from the joint distribution $p(u_1, \ldots, u_J)$ of interest.

- This suggests that one can obtain a sample of size L from $p(u_1, \ldots, u_J)$ by obtaining ***one long chain*** of length T and, after an initial "***burn-in***" period after which the chain is thought to have "***stabilized***," collecting L ***suitably spaced*** realizations from the chain (to eliminate correlation among them).
- Alternatively, one can perform L ***parallel chains***, each of length T, and take the ***final realization*** from each.
- This final sample of L realizations from the joint distribution can then be used to construct the desired functionals, such as ***marginal posterior summaries*** like the mode, mean, and variance.

This generic scheme is used in the context of ***hierarchical models*** by identifying each element of $(U_1, \ldots, U_J)$ with the parameters of the model; e.g., in the example above, β, γ, $\boldsymbol{D}$, as well as β_i or $\boldsymbol{b}_i$, $i = 1, \ldots, m$. One then derives the full conditional distributions, at least up to a ***proportionality constant***, from the assumptions embodied in the three-stage hierarchy.

In the particular case of the ***normal model with nonconstant variance*** above and ***hyperprior*** specifications in (10.6), it can be shown that the full conditional distributions (10.10) are as follows; the diligent student may want to verify this.

$$(\beta|\boldsymbol{y},\sigma^2,\delta,\boldsymbol{D},\beta_i,\ i=1,\ldots,m)\sim\mathcal{N}\{\boldsymbol{\Lambda}(m\boldsymbol{D}^{-1}\overline{\beta}+\boldsymbol{H}^{-1}\beta^*),\boldsymbol{\Lambda}\},$$

$$\boldsymbol{\Lambda}^{-1}=m\boldsymbol{D}^{-1}+\boldsymbol{H}^{-1},\qquad \overline{\beta}=m^{-1}\sum_{i=1}^{m}\beta_i,$$

$$(\boldsymbol{D}^{-1}|\boldsymbol{y},\beta,\sigma^2,\delta,\beta_i,\ i=1,\ldots,m)\sim \text{Wi}\left[\left\{\sum_{i=1}^{m}(\beta_i-\beta)\beta_i-\beta)^T+\rho\boldsymbol{D}^*\right\}^{-1},m+\rho\right],$$

$$(\sigma^{-2}|\boldsymbol{y},\beta,\delta,\boldsymbol{D},\beta_i,\ i=1,\ldots,m)\sim \text{Ga}\{(\nu_0+N)/2,A_0\},$$

$$A_0=\left[\sum_{i=1}^{m}\{\boldsymbol{y}_i-\boldsymbol{f}_i(\boldsymbol{z}_i,\beta_i)\}^T\boldsymbol{R}_i^{-1}(\beta_i,\gamma,\boldsymbol{z}_i)\{\boldsymbol{y}_i-\boldsymbol{f}_i(\boldsymbol{z}_i,\beta_i)\}+\nu_0\tau_0\right]/2.$$

The full conditional $(\beta_i|\boldsymbol{y},\beta,\sigma^2,\boldsymbol{D},\beta_k,\ k\neq i)$ is ***proportional to***

$$\exp\left[-\frac{1}{2\sigma^2}\{\boldsymbol{y}_i-\boldsymbol{f}_i(\boldsymbol{z}_i,\beta_i)\}^T\boldsymbol{R}_i^{-1}(\beta_i,\gamma,\boldsymbol{z}_i)\{\boldsymbol{y}_i-\boldsymbol{f}_i(\boldsymbol{z}_i,\beta_i)\}\right]$$

$$\times\exp\left\{-(\beta_i-\beta)^T\boldsymbol{D}^{-1}(\beta_i-\beta)/2\right\}\sigma|\boldsymbol{R}_i(\beta_i,\gamma,\boldsymbol{z}_i)|^{-1/2}.$$

The full conditional $(\delta|\boldsymbol{y},\beta,\sigma^2,\boldsymbol{D},\beta_i,\ i=1,\ldots,m)$ is proportional to

$$\prod_{i=1}^{m}\exp\left[-\frac{1}{2\sigma^2}\{\boldsymbol{y}_i-\boldsymbol{f}_i(\boldsymbol{z}_i,\beta_i)\}^T\boldsymbol{R}_i^{-1}(\beta_i,\gamma,\boldsymbol{z}_i)\right]\sigma|\boldsymbol{R}_i(\beta_i,\gamma,\boldsymbol{z}_i)|^{-1/2}.$$

Although the idea is simple, there are ***several challenges*** for implementation.

- There is a need to ***monitor convergence*** of the chain(s) to feel confident that they have stabilized, and diagnostic techniques for doing so have been developed. The required length of the ***burn-in period*** is also an issue.
- Gibbs sampling has obvious appeal when it is ***straightforward to sample*** from all the full conditional distributions. However, it is ordinarily the case in ***complex statistical models*** that this is not always possible. In the example above, sampling from the full conditionals for β, σ^2, and $\boldsymbol{D}$ is straightforward using standard procedures for ***random variate generation*** from the normal, gamma, and Wishart distributions, which are available in popular software.
- However, sampling from those for the β_i and δ is ***more challenging*** ; this is a consequence of the ***nonlinearity*** of the model f in β_i. One must resort to techniques such as ***rejection sampling*** for which the target distribution to be sampled is known only up to a proportionality constant. Other methods, such as importance sampling, can also be employed. Alternatively, one can embed embed a random variate sampling scheme based on the ***Metropolis-Hastings algorithm***. Usually, the choice and implementation of method depends on the ***particular model*** and must be tailored to the specific problem.

- Accordingly, implementation of the Bayesian formulation of the general nonlinear mixed effects model can require some ***sophistication*** on the part of the user.

- Software such as ***BUGS*** (Bayesian inference Using Gibbs Sampling) accommodates these situations. An interface to BUGS for population pharmacokinetic (and pharmacodynamic) analysis that has many popular PK models built-in and can accommodate complex individual dosing histories, ***PKBugs***, is also available. However, the user must have ***sufficient background*** in MCMC techniques, and especially appreciation for the issues of practical implementation discussed above.

Further discussion is beyond our scope here. Some classic papers discussing implementation of the Bayesian formulation of nonlinear mixed effects models are Wakefield, Smith,, and Racine-Poon (1994), Wakefield (1996), and Bennett, Racine-Poon, and Wakefield (1996). Rosner and Müller (1994), Wakefield (1996), and Müller and Rosner (1997) discuss specific pharmacokinetic applications.

10.3 Complex nonlinear models

As the quantitative and biological sciences continue to ***converge***, complex ***mathematical models of biological systems*** have become commonplace. ***Nonlinear dynamical systems*** models, of which the one-compartment PK models we have discussed in previous chapters are trivial, simple cases, are used in a number of application areas. It is natural to ***embed*** these complex ***mathematical models*** in the ***statistical*** nonlinear mixed effects model framework to address scientific questions of interest, as the following two examples demonstrate.

PHYSIOLOGICALLY-BASED PHARMACOKINETIC MODELS: Ordinary ***compartmental models*** of pharmacokinetics for the study of the disposition of drugs and biologics in humans are typically ***gross simplifications*** of the physiology involved. Although these models can be ***extraordinary useful approximations***, more sophisticated such models are required to address scientific questions in some key settings.

Toxicokinetics refers to the study of ***pharmacokinetics*** of environmental, chemical, or other agents in the context of assessment of their possible ***toxic effects***.

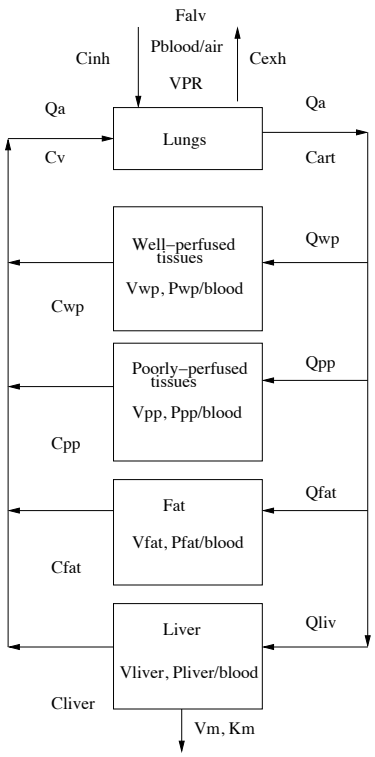

Figure 10.1: *A representative physiologically-based pharmacokinetic model.*

Intensive ***toxicokinetic studies*** are conducted in animal models and involve exposure of each animal to the agent and collection of frequent blood and other samples from which concentrations are ascertained. They are also conducted to a lesser extent in humans, from whom blood, breath, and urine samples are obtained and assayed for concentrations of the agent. Understandably, toxicokinetics is of great interest to ***environmental regulatory agencies*** such as the EPA.

Here, interest focuses on learning about key processes, such as the rate at which the agent is ***metabolized*** in the liver. This information is used in the overall toxicological assessment. Because the scientific questions involve the such organ-specific processes, it is standard to entertain ***more detailed compartmental models*** that represent the body by ***physiologically identifiable compartments*** such as fatty tissues, poorly- and well-perfused tissues, the liver, and so on. Figure 10.1 shows a prototypical such model.

These so-called ***physiologically-based pharmacokinetic*** (PBPK) models generally are complex ***systems of differential equations*** that ***do not admit closed form solutions*** for observable concentrations, so must be solved numerically.

For example, equations corresponding to the model in Figure 10.1 are given by

$$C_{\text{art}} = \frac{F_{\text{card}} C_{\text{ven}} + F_{\text{alv}} C_{\text{inh}}}{F_{\text{card}} + F_{\text{alv}}/P_{\text{blood/air}}}, \quad C_{\text{ven}} = \sum_s \frac{F_s C_s}{F_{\text{card}}}$$

$$C_{\text{exh}} = (1-\delta)\frac{C_{\text{art}}}{P_{\text{blood/air}}} + \delta C_{\text{inh}}$$

$$\frac{dC_s}{dt} = \frac{F_s}{V_s}\left(C_{\text{art}} - \frac{C_s}{P_{s/\text{blood}}}\right), \quad s = \text{wp, pp, fat}$$

$$\frac{dC_{\text{liv}}}{dt} = \frac{F_{\text{liv}}}{V_{\text{liv}}}\left(C_{\text{art}} - \frac{C_{\text{liv}}}{P_{\text{liv/blood}}}\right) - R_{\text{liv}} \ (s = \text{liv}),$$

$$R_{\text{liv}} = \frac{V_{\max} C_{\text{liv}}}{V_{\text{liv}}(K_m + C_{\text{liv}})},$$

In the figure, the Qs are amounts, which are divided by corresponding compartmental ***volume*** parameters V to yield concentrations C. The F parameters are blood flow rates, the P parameters are tissue-over-blood ***partition coefficients***, and $V_{\max}$ and K_m are ***Michaelis-Menten metabolism coefficients*** that describe ***metabolism in the liver***.

Interest thus focuses on these metabolism parameters, their ***typical values*** and how they ***vary in the population***. Thus, the nonlinear hierarchical model we have discussed is a natural framework for addressing this. However, there are some serious ***challenges***.

- The solution to the above system for the ***observable concentration*** (usually in blood or exhaled breath) provides the model f of interest, and it is clear that a closed form expression is extremely unlikely.
- In general, such PBPK models involve ***numerous parameters***, most of which are ***not identifiable*** from longitudinal concentration-time data from these studies.

A naive approach to inference on the ***typical values*** of $V_{\max}$ and K_m has been to ***fix*** all parameters except these at "***literature values***" for all individuals, estimate $V_{\max}$ and K_m for each individual, and base inference on the typical values and variability on these estimates. Clearly, this crude approximation is ***highly suspect***.

Gelman, Bois, and Jiang (1996) were the first to propose formally placing this problem within the ***nonlinear hierarchical framework***. To address the challenge of handling the numerous unidentifiable parameters in a more ***principled*** way, they proposed a ***Bayesian formulation*** as in the previous section, where the hyperprior specifications for each parameter are based on ***historical and literature information***.

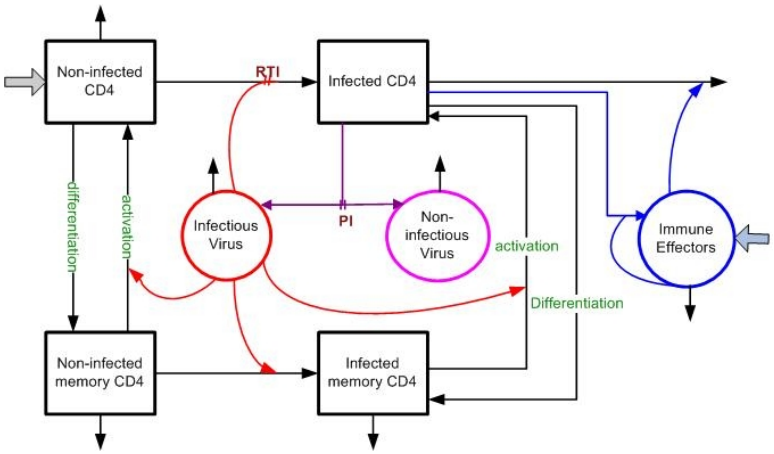

Figure 10.2: *A representative HIV dynamic model.*

These priors are "***informative***" in that they do not reflect ***vague knowledge*** but rather incorporate ***scientific information and judgment***. Implementation of the model is via MCMC techniques.

A software package, ***MCSim***, that implements this general approach, is available. Of course, its appropriate use requires a fairly deep understanding of both the theoretical model and the statistical model in which it is embedded.

Interestingly, the naive approach is still often used.

HIV DYNAMICS: Human immunodeficiency virus Type-1 (HIV) progressively destroys the body's ability to fight infection by killing or damaging cells in the immune system. Since the mid-1990s, there has been considerable interest in developing mathematical models to represent hypothesized mechanisms governing the ***interplay between HIV and the immune system***. These so-called ***HIV dynamic models*** have led to advances in understanding of plausible mechanisms underlying HIV pathogenesis and in developing ***antiretroviral*** treatment strategies for HIV-infected individuals.

As in pharmacokinetics, these models are predicated on representing processes involved in the virus-immune system ***interplay*** via ***hypothetical compartments***, where the compartments characterize different populations of virus; immune system cells targeted by the virus, namely $CD4^+$ T cells; and so on, that interact within a subject. An example is the model in Figure 10.2, which shows a typical such model.

As an example, we show a model studied by Adams et al. (2007), which involves ***compartments*** denoted by T_1, type 1 target cells, e.g., CD4$^+$ cells (cells/μl); T_2, type 2 target cells, such as macrophages (cells/μl); V_I and V_{NI}, infectious and noninfectious free virus, respectively (RNA copies/ml); and E, cytotoxic T-lymphocytes (cells/μl). With a superscript asterisk (*) denoting infected target cells; and, e.g., with $T_1(t)$ = concentration of type 1 target cells at time t, the model is

$$\begin{aligned}
\dot{T}_1 &= \lambda_1 - d_1 T_1 - \{1 - \epsilon_1 u(t)\} k_1 V_I T_1, \\
\dot{T}_2 &= \lambda_2 - d_2 T_2 - \{1 - f\epsilon_1 u(t)\} k_2 V_I T_2, \\
\dot{T}_1^* &= \{1 - \bar{\epsilon}_1(t)\} k_1 V_I T_1 - \delta T_1^* - m_2 E T_1^*, \\
\dot{T}_2^* &= \{1 - f\epsilon_1 u(t)\} k_2 V_I T_2 - \delta T_2^* - m_2 E T_2^*, \\
\dot{V}_I &= \{1 - \epsilon_2 u(t)\} 10^3 N_T \delta (T_1^* + T_2^*) - c V_I - \{1 - \epsilon_1 u(t)\} \rho_1 10^3 k_1 T_1 V_I \\
&\quad - \{1 - f\epsilon_1 u(t)\} \rho_2 10^3 k_2 T_2 V_I, \\
\dot{V}_{NI} &= \epsilon_2 u(t) 10^3 N_T \delta (T_1^* + T_2^*) - c V_{NI}, \\
\dot{E} &= \lambda_E + \frac{b_E (T_1^* + T_2^*)}{(T_1^* + T_2^*) + K_b} E - \frac{d_E (T_1^* + T_2^*)}{(T_1^* + T_2^*) + K_d} E - \delta_E E,
\end{aligned} \tag{10.11}$$

along with an initial condition vector $\{T_1(0), T_2(0), T_1^*(0), T_2^*(0), V_I(0), V_{NI}(0), E(0)\}^T$. In (10.11), most dependence on t is suppressed for brevity, and the factors of 10^3 convert between μl and ml scales. The model depends on ***numerous meaningful parameters*** β that are thought to vary across subjects; e.g., c (1/day), the natural death rate of the virus; δ (1/day), the death rate of infected target cells; and λ_k, $k = 1, 2$, production rates (cells/μl-day) of type 1 and 2 target cells. The function $u(t)$, $0 \leq u(t) \leq 1$, represents ***time-dependent input*** of antiretroviral therapy, with $u(t) = 0$ corresponding to fully off treatment and $u(t) = 1$ to fully on treatment. The parameters ϵ_k, $k = 1, 2$, $0 \leq \epsilon_k \leq 1$, are ***efficacies*** of reverse transcriptase inhibitor treatments for blocking new infections and protease inhibitors for causing infected cells to produce noninfectious virus, respectively. See Adams et al. (2007) for a complete description.

In a typical study of HIV-infected subjects, longitudinal data on ***combinations*** of one or more of these compartments are collected, and interest focuses on ascertaining the typical values of some of these parameters and how they vary across subjects to gain insight into ***viral mechanisms*** and their possible associations with subject characteristics. Usually, longitudinal measurements of total CD4$^+$ cells, $T_1 + T_1^*$, and total viral load, $V_i + V_{NI}$, are available on each subject. Needless to say, solution of (10.11) can only be carried out numerically. A further complication is that total viral load measurements may be ***left-censored*** by the ***lower limit of quantification*** of the assay. Finally, as with PBPK models, the available data ***fail to identify*** all the parameters.

Clearly, the scientific questions can be addressed within the nonlinear mixed effects model framework. There is an ongoing literature on approaches to doing this, which included ***integrating numerical solution of the system of differential equations with estimation***. Not surprisingly, many of the approaches place the problem in a Bayesian formulation.

10.4 Time-dependent covariates in nonlinear mixed effects models

In Chapter 9, we did not address the situation in which ***among-individual covariates change valu***e over the course of observation of an individual. We now remark briefly on the implications of this in the nonlinear mixed effects model context.

GENERALIZATION: When time-dependent among-individual covariates are available, a first thought is to ***modify*** the basic two-stage hierarchy as follows. For this discussion, let $\boldsymbol{a}_{ij}$ denote the values of among-individual covariates at time j for individual i.

Stage 1 - Individual model. Given a model f as in (9.6), the random vectors $\boldsymbol{Y}_i$, $i = 1, \ldots, m$, are assumed to satisfy

$$E(\boldsymbol{Y}_i|\boldsymbol{z}_i, \beta_{ij}) = \boldsymbol{f}_i(\boldsymbol{z}_i, \beta_{ij}), \qquad \text{var}(\boldsymbol{Y}_i|\boldsymbol{z}_i, \beta_{ij}) = \boldsymbol{R}_i(\beta, \gamma, \boldsymbol{x}_i, \boldsymbol{b}_i). \tag{10.12}$$

In (10.12), we allow the individual-specific parameters β_{ij} to ***change with*** j.

Stage 2 - Population model. The individual-specific parameter β_{ij} satisfies

$$\beta_{ij} = \boldsymbol{d}(\boldsymbol{a}_{ij}, \beta, \boldsymbol{b}_i); \tag{10.13}$$

a ***linear*** version of (10.13) is

$$\beta_{ij} = \boldsymbol{A}_{ij}\beta + \boldsymbol{B}_{ij}\boldsymbol{b}_i,$$

where the design matrix $\boldsymbol{A}_{ij}$ changes with changing values $\boldsymbol{a}_{ij}$. Here and in (10.13), the value of β_{ij} thus changes as $\boldsymbol{a}_{ij}$ changes.

There is no ***operational barrier*** to fitting the model in (10.12)- (10.13); e.g., the R function `nlme()` has this capability, which is discussed in Pinheiro and Bates (2000, Section 7.1). Thus, implementing such a model is ***entirely possible***.

The issue is whether or not such a model is ***appropriate or even makes sense*** , as we now discuss.

GENERALIZED LINEAR MIXED EFFECTS MODELS: With ***generalized linear mixed effects models*** , the objective is ordinarily inference on the association between the response and among-individual covariates such as treatment assignment in a randomized clinical trial and/or subject characteristics from a ***subject-specific*** perspective. The model is thus an ***empirical framework*** in which to address these questions. When these covariates are ***time-independent*** , there is ***no conceptual difficulty*** with specifying these models and making the desired inferences, as discussed in Section 8.6 in the case of population-averaged models.

When ***time-dependent*** among-individual covariates are involved, however, the ***same conceptual issues*** discussed in Section 8.6 arise. Modeling involving ***endogenous covariates*** suffers from the ***same difficulties of interpretation*** discussed in that section. Namely, ***time-dependent confounding*** complicates and frankly renders impossible the ability to draw ***causal inferences*** based on these models.

Thus, adopting the model (10.12)-(10.13) when such time-dependent covariates are involved is almost certainly a ***prescription for misleading inference*** and challenging interpretation. As noted in Section 8.6, an entirely ***different approach*** is required.

NONLINEAR MIXED EFFECTS MODELS: In the case of a nonlinear hierarchical model in which the function f representing the within-individual mean trajectory is derived from ***mechanistic***,, ***theoretical considerations***, (10.12)- (10.13) is problematic for ***different reasons***. Such theoretical models are derived under the fundamental assumption that that the ***scientifically meaningful parameters*** involved are ***constants with respect to time***.

For example, in simple compartmental models of the type we have discussed in previous chapters, PK parameters like clearance Cl_i and volume of distribution V_i of necessity are ***fixed*** for individual i, and the ***differential equations*** giving rise to the concentration-time model are predicated on this. Similarly, in the HIV dynamic model (10.11), the numerous parameters are regarded as ***fixed constants*** for each individual. In each case, a practical interpretation is that these fixed parameters are thought to be ***inherent characteristics*** of the individual that govern his/her kinetics or dynamics.

In such systems, if a ***parameter*** is thought to ***vary with time*** , then it is taken in the model to be a ***function of time*** , which might be represented for fitting purposes as a ***parametric or nonparametric*** model. Allowing a mechanistic parameter to vary with time ***fundamentally alters*** the nature of the ***system of differential equations*** and thus the form of the solution.

time (hours)	conc. (mg/L)	dose (mg)	age (years)	weight (kg)	creat. (ml/min)	glyco. (mg/dl)
0.00	–	166	75	108	> 50	69
6.00	–	166	75	108	> 50	69
11.00	–	166	75	108	> 50	69
17.00	–	166	75	108	> 50	69
23.00	–	166	75	108	> 50	69
27.67	0.7	–	75	108	> 50	69
29.00	–	166	75	108	> 50	94
35.00	–	166	75	108	> 50	94
41.00	–	166	75	108	> 50	94
47.00	–	166	75	108	> 50	94
53.00	–	166	75	108	> 50	94
65.00	–	166	75	108	> 50	94
71.00	–	166	75	108	> 50	94
77.00	0.4	–	75	108	> 50	94
161.00	–	166	75	108	> 50	88
168.75	0.6	–	75	108	> 50	88
height=72 inches, Caucasian, smoker, no ethanol abuse, no CHF						

Table 10.1: *Data for a subject in the quinidine study. conc. = quinidine concentration, glyco. =* α_1*-acid glycoprotein concentration, CHF = congestive heart failure.*

Thus, for example, in a one-compartment model, allowing clearance, to vary with the value of an among-individual covariate and writing Cl_{ij}, say, must be carefully considered for ***plausibility*** and can have implications for the ***validity*** of the model used.

In pharmacokinetics, under certain conditions, it is accepted that individual-specific parameters ***can fluctuate*** over observation periods but ***not over*** each observation time *j*. To discuss this, we consider a ***world-famous study*** of the pharmacokinetics of the anti-arrhythmic drug quinidine, which has been cited by numerous authors (e.g., Davidian and Giltinan, 1995, Sections 1.1.2 and 9.3; Wakefield, 1996; Pinheiro and Bates, 2000, Sections 3.4 and 8.2).

The $m = 136$ subjects in the study were hospitalized and undergoing routine treatment with oral quinidine for atrial fibrillation or ventricular arrhythmia. Table 10.1 shows ***abridged data*** for one subject, which typify the information collected. Demographical and physiological characteristics included age; weight; height; ethnicity/race (Caucasian/Black/Hispanic); smoking status (yes/no); ethanol abuse (no/current/previous); congestive heart failure (no or mild/moderate/severe); creatinine clearance, a measure of renal function ($\leq$ 50 ml/min indicates renal impairment); and α_1-acid glycoprotein concentration, the level of a molecule that binds quinidine. In addition, ***dosing history*** (times, amounts) was recorded along with quinidine concentrations.

A ***one-compartment model*** with first-order absorption and elimination has been used to describe the PK of quinidine. With repeated dosing, in addition to the ***principle of superposition*** discussed in Section 9.2, it is also assumed that drug accumulates in the system until a "***steady state***" is reached at which, roughly, rate of administration of drug is ***equal to*** the rate of elimination (e.g., Giltinan, 2014).

Under these conditions, the model can be written as follows. Denoting the ℓth (dose time, amount) by (s_ℓ, D_ℓ) as before, the amount of quinidine in the "***absorption depot***," $A_a(s_\ell)$, and the concentration of quinidine in the blood, $C(s_\ell)$, at dose time s_ℓ for a subject who has ***not yet achieved a steady state*** are given by

$$\begin{aligned} A_a(s_\ell) &= A_a(s_{\ell-1})\exp\{-k_a(s_\ell - s_{\ell-1})\} + D_\ell, \\ C(s_\ell) &= C(s_{\ell-1})\exp\{-k_e(s_\ell - s_{\ell-1})\} + A_a(s_{\ell-1})\frac{k_a}{V(k_a - k_e)} \\ &\quad \times\Big[\exp\{-k_e(s_\ell - s_{\ell-1})\} - \exp\{-k_a(s_\ell - s_{\ell-1})\}\Big], \end{aligned}$$

and the concentration of quinidine at time t in the next dosing interval $(s_\ell, s_{\ell+1})$ is

$$\begin{aligned} C(t) &= C(s_\ell)\exp\{-k_e(t - s_\ell)\} + A_a(s_\ell)\frac{k_a}{V(k_a - k_e)} \\ &\quad \times\Big[\exp\{-k_e(t - s_\ell)\} - \exp\{-k_a(t - s_\ell)\}\Big], \quad s_\ell < t < s_{\ell+1}, \end{aligned} \tag{10.14}$$

where $k_e = Cl/V$. Once a ***steady state*** has been reached, a ***further*** set of equations governs the values of $A_a(s_\ell)$ and $C(s_\ell)$ at dose times, which we exclude here for brevity. The model for concentration at time t that dictates f thus depends on the meaningful parameters $\beta = (k_a, V, Cl)^T$.

The quinidine study is representative of the situation where one or more subject characteristics thought to be ***associated*** with pharmacokinetic behavior ***change*** over the observation period on the subject, as is the case here for α_1-acid glycoprotein concentration. For the subject in Table 10.1, it is likely that α_1-acid glycoprotein concentration was ***measured intermittently*** at times 0, 29, and 161. In this situation, a standard modeling approach is based on the following idea.

If a subject is observed over several treatment intervals, it may be reasonable to expect that, although a basic compartment model with ***static parameters*** applies ***in any interval***, ***fluctuations*** in the values of his/her pharmacokinetic parameters may occur over time that ***show an association*** with ***other characteristics*** that also change. From this point of view, for the quinidine study, the assumption is that the pharmacokinetic parameters in (10.14) for the individual in Table 10.1 are ***constant within the intervals*** 0–29 hours, 29–77 hours, and after 161 hours, but may have ***fluctuated*** over the entire period in a way that is associated with α_1-acid glycoprotein concentration.

Such an assumption is clearly subject to ***scientific debate*** but is often invoked as a practical way to view the problem.

Denoting these intervals by I_k, $k = 1, \ldots, a$ ($a = 3$), a standard modeling approach in the pharmacokinetic literature is to modify the stage 2 population model (10.13) as

$$\beta_{ij} = \boldsymbol{d}(\boldsymbol{a}_{ik}, \beta, \boldsymbol{b}_i), \tag{10.15}$$

where $\boldsymbol{a}_{ik}$ is the value of the subject characteristics for $t_{ij} \in I_k$. In (10.15), the element of β that is the coefficient of α_1-acid glycoprotein concentration is taken to be ***constant*** over all intervals; e.g., for $t_{ij} \in I_k$,

$$\log Cl_{ij} = \beta_0 + \beta_1 a_{ik} + b_i. \tag{10.16}$$

The model (10.15) implies that, within a given individual, "***inter-interval variation***" is ***entirely*** "***explained***," by the change in covariates for that individual. The model can be extended to include ***nested random effects*** that allow for ***unexplained biological variation*** within intervals, as discussed in the next section.

10.5 Multilevel models

Longitudinal data with the structure we have focused on in this course can be viewed as having a ***single level*** of ***clustering***. In particular, responses fall into natural ***clusters*** because they are ascertained longitudinally on ***different individuals***. Responses from the same cluster are naturally viewed as ***correlated*** due to the fact that they are "***more alike***" by virtue of belonging to the same cluster, the phenomenon we have referred to as the ***among-individual source of correlation***. ***Linear and nonlinear mixed effects models*** naturally account for this correlation through the introduction of ***random effects***.

From this point of view, the longitudinal data structure we have focused on in this course is a ***special case*** of a ***general clustered data structure***, in which it is possible to identify ***multiple levels*** of such clustering. In this section, we give a brief introduction to statistical models for this data structure. The models we have discussed are thus a particular case of these models.

MULTIPLE LEVELS OF CLUSTERING: In many settings, the data structure is such that it is possible to identify ***multiple levels of clustering***:

- A classic example is an ***agricultural study*** in which ***plots*** are ***nested*** within experimental ***blocks***, and plots are randomized to treatments within blocks. A single response may be ascertained on each plot at the end of the growing season. Alternatively, in a more complicated version, longitudinal responses are collected on each plot throughout the growing season.

- Similarly, a common situation in ***clinical trials*** is that in which subjects are recruited from a sample of ***clinics***. Each subject within each clinic is randomized to a study treatment. Here, subjects are ***nested*** within clinics. The response or interest may be ascertained on each subject at the end of the study period. Alternatively, each subject may be followed ***longitudinally*** for the response at several clinic visits over the study period.

- This structure occurs frequently in studies in ***public health*** and ***education*** where it is ***not feasible*** to expose participants to treatment or conditions of interest ***individually***. For example, in a study of school-based interventions to prevent smoking, entire ***schools*** might be randomized to receive a particular intervention program; allowing different students within same school to receive different interventions opens the possibility that they could discuss and compare them, which could ***compromise*** the ability to assess their effects.

  Individual instructors might actually ***deliver the interventions*** at the ***classroom level***, so this represents a ***source of variation***, in that particular instructors might be more or less effective at delivering the same intervention. Smoking behavior might then be recorded over time on individual students. Here, students are ***nested*** within classrooms within schools.

  The response, ***score*** on a questionnaire assessing tobacco and health knowledge, might be ascertained at ***baseline***, prior to intervention, and then again at the end of the intervention on each student.

COVARIATES: In the foregoing examples, it is natural that there might be ***covariates*** that recorded at ***different levels*** of the hierarchy.

- In the smoking intervention study, there may be covariates collected at the level of the ***individual student***, such as gender, GPA, socioeconomic status, and so on.

Covariates at the ***classroom level*** , such as those recorded on the ***instructor*** , including gender, previous experience with such interventions, years teaching, etc, might also be collected. Finally, ***school-level*** covariates, such as location, racial/ethnic makeup, proportion of students receiving free/reduced price lunch, etc, might be available.

- Questions of ***scientific interest*** may involve associations between response or level-specific behavior and these covariates. In particular, the goal is often to determine the relative importance of characteristics at ***different levels*** in terms of effects on the response.

MULTILEVEL HIERARCHICAL MODELS: A natural framework in which to place such questions is that of ***multilevel hierarchical models***. An important feature of these models is that they take account of ***correlation*** due to ***clustering*** at each stage of the hierarchy, noted above, as we demonstrate shortly.

The linear and nonlinear mixed effects models we have discussed for longitudinal data are ***special cases*** of this general framework. In the more general model, there ***may or may not*** be a longitudinal aspect. If there is, it corresponds to a ***level*** in the hierarchy.

For simplicity, we consider ***linear models*** , and comment on the obvious generalization to nonlinear and generalized linear models at the end of this section.

LEVELS OF THE HIERARCHY: It is common to identify ***units*** , be they individual subjects or something else, at each ***level*** of the hierarchy. ***Level 1*** units are at the '***lowest*** " level and are the units on which the response is ascertained. ***Level 2*** units are clusters composed of level 1 units. ***Level 3*** units comprise clusters of level 2 units, and so on.

This is most easily appreciated through examples.

- Consider the clinical trial example above, suppose that a single measure of the response is recorded at the end of the study. In this case, level 1 units are the ***individual subjects*** on whom the response measure will obtained. Level 2 units are the ***clinics*** , in which subjects are ***nested***. Thus, this is a two-level structure.
- Consider the same example, but where now each subject will visit the clinic at several occasions during the study period, and the response will be ascertained at each visit. Here, level 1 units are the ***measurement occasions*** , i.e., the longitudinal times at which the response is ascertained. These are ***nested*** within level 2 units, the ***subjects***. Level 3 units are the clinics in which level 2 units, subjects, are ***nested***.

- In the smoking intervention example, if the baseline score is treated as a ***covariate*** , it is natural to view this as a hierarchy with ***three levels***. Level 1 units are the individual students, nested within level 2 units, classrooms, nested within level 3 units, schools. If we view the baseline and final scores as ***longitudinal responses*** instead, level 1 is the measurement occasion. level 2 is the student, level 3 is classroom, and level 4 is school.

- In the longitudinal data structure which we have been concerned in this course, it follows that level 1 units are the ***longitudinal occasions*** at which the response is ascertained, and level 2 units are the ***individuals***.

- This convention obviously can be applied in general.

NOTATIONAL CONVENTION: In the literature, it is ***conventional*** (although there are exceptions) to use i to index level 1 units, j to index level 2 units, and k to index level 3 units in a ***three-level model***.

The ***numbers of units*** are identified as follows. There are n_3 level 3 units, indexed by $k = 1, \ldots, n_3$. Within the kth, there are n_{2k} level 2 units indexed by $j = 1, \ldots, n_{2k}$. Within the jth level 2 unit, there are n_{1jk} level 1 units indexed by $i = 1, \ldots, n_{1jk}$.

In a ***two-level model*** , there are n_2 level 2 units indexed by $j = 1, \ldots, n_2$ and within the jth level 2 unit there are n_{1j} level 1 units indexed by $i = 1, \ldots, n_{1j}$.

- Thus, under this indexing convention, the ***longitudinal data structure*** we have considered would use i to index time points and j to index individuals, exactly ***opposite*** of the standard indexing convention in the longitudinal data literature, which we have used to this point.

  Here, the ***total number of individuals*** m in the notation we have used corresponds to n_2 and the number of observations on each individual, which we denote by n_i for individuals indexed by i, is n_{1j} for the jth individual in this indexing scheme.

- It is ***prudent*** when reading the multilevel modeling literature to pay attention to the indexing convention used.

In the following example, we demonstrate this indexing convention.

MULTICENTER LONGITUDINAL CLINICAL TRIAL: As above, suppose a ***clinical trial*** is conducted comparing two treatments, coded as 0 and 1. The trial involves n_3 clinics (level 3 units) indexed by $k = 1, \ldots, n_3$. Within each clinic k, n_{2k} subjects are recruited, indexed by $j = 1, \ldots, n_{2k}$ (level 2 units), each of whom is randomized to receive treatment 0 or 1. On subject j within clinic k, n_{1jk} responses are ascertained at several clinic visits, indexed by $i = 1 \ldots, n_{1jk}$, where $i = 1$ corresponds to ***baseline***.

The usual perspective is that the clinics represent a ***random sample*** from the ***hypothetical population*** of all possible such clinics in which subjects could be recruited and given the study agents. The subjects within clinics likewise are viewed as ***random samples*** from the hypothetical populations of all possible subjects who could attend each clinic.

Suppose that interest focuses on comparing the patterns of change in the response over the study period for the two treatments, where, for ***any subject*** , the expected longitudinal trajectory over time is expected to show a ***constant rate of change***. From this perspective, it is natural to take a ***subject-specific perspective*** to developing a model, as follows.

Let Y_{ijk} represent the ***response*** on subject j from clinic k at the ith time at which the subject is observed, $i = 1, \ldots, n_{1jk}$. Under the principles we used to develop the ***two-stage linear mixed effect model hierarchy*** in Chapter 6, the following formulation is natural.

Letting t_{ijk} denote the ***observation times*** for subject j in clinic k, represent the responses at the subject level as

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk} t_{ijk} + e_{ijk}, \quad i = 1, \ldots, n_{1jk}, \tag{10.17}$$

where, in (10.17), β_{0jk} and β_{1jk} are the ***subject-specific*** intercept and slope dictating subject (j, k)'s ***inherent longitudinal trajectory***; and e_{ijk} is a ***mean zero*** (conditional on covariates) ***within-subject*** deviation representing the effects of the realization process and measurement error.

Defining $\boldsymbol{Y}_{jk}$ and $\boldsymbol{e}_{jk}$ $(n_{1jk} \times 1)$ in the obvious way, we can write (10.17) succinctly as

$$\boldsymbol{Y}_{jk} = \boldsymbol{C}_{jk}\boldsymbol{\beta}_{jk} + \boldsymbol{e}_{jk}, \quad \boldsymbol{\beta}_{jk} = (\beta_{0jk}, \beta_{1jk})^T. \tag{10.18}$$

Assumptions on $\boldsymbol{e}_{jk}$ would be made as discussed in Chapters 2, 6, 7, and 9.

The next step in formulating the multilevel model is to represent the ***subject-specific intercept and slope*** in terms of ***fixed effects*** , covariates, and ***random effects***.

Given that this is a ***randomized study*** , so that we do not expect baseline response to be associated with treatment assignment, write

$$\beta_{0jk} = \beta_0 + b_{0k} + b_{0jk}, \tag{10.19}$$

where $E(b_{0k}) = 0$, $E(b_{0jk}) = 0$, so that $E(\beta_{0jk}) = \beta_0$, the mean or "typical" value of intercept across all clinics and subjects within them. In (10.19), b_{0k} is a ***random effect*** representing how ***subject-specific intercepts*** for subjects in the kth clinic deviate from the overall mean intercept β_0,, and b_{0jk} is a ***random effect*** representing further how the intercept for the jth subject within that clinic deviates from the ***clinic-specific*** (conditional on clinic) mean intercept (across subjects in the clinic) $\beta_0 + b_{0k}$.

Similarly, a model for the ***subject-specific*** slope is

$$\beta_{1jk} = \beta_1 + \beta_2\delta_{jk} + b_{1k} + b_{1jk}, \tag{10.20}$$

where $\delta_{jk} = 0$ if the jth subject in clinic k received drug 0, and $\delta_{jk} = 1$ if s/he received drug 1; and $E(b_{1k}|\delta_{jk}) = 0$, $E(b_{1jk}|\delta_{jk}) = 0$. Thus, in (10.20), β_1 is the mean or "typical" slope ***across all clinics and subjects*** for subjects receiving drug 0, and $\beta_1 + \beta_2$ is that for drug 1, so that β_2 represents the ***difference in mean or typical rate of change*** between the two drugs.

As for the intercept, b_{1k} is a ***random effect*** representing how subject-specific slopes for subjects in the kth clinic deviate from the overall mean slope for each treatment, and b_{1jk} is a ***random effect*** representing further how the slope for the jth subject within that clinic deviates from the ***clinic-specific*** (conditional on clinic) mean slope $\beta_1 + \beta_2\delta_{jk} + b_{1k}$, depending on the treatment assigned.

As in Chapter 6, we can ***summarize*** (10.19)-(10.20) as

$$\boldsymbol{\beta}_{jk} = \boldsymbol{A}_{jk}\boldsymbol{\beta} + \boldsymbol{B}_k^{(1)}\boldsymbol{b}_k + \boldsymbol{B}_{jk}^{(2)}\boldsymbol{b}_{jk}, \tag{10.21}$$

where

$$\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T, \quad \boldsymbol{b}_{jk} = (b_{0jk}, b_{1jk})^T, \quad \boldsymbol{b}_k = (b_{0k}, b_{1k})^T,$$

$$\boldsymbol{A}_{jk} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \delta_{jk} \end{pmatrix}, \quad \boldsymbol{B}_k^{(1)} = \boldsymbol{I}_2, \quad \boldsymbol{B}_{jk}^{(2)} = \boldsymbol{I}_2.$$

It should be clear that it is possible to modify (10.19) and (10.20) to incorporate dependence on both ***clinic-level*** and ***subject-level covariates***. For example, suppose that w_{jk} is the weight of subject j in clinic k, and s_k is the average number of suffers of the condition the drugs are supposed to treat seen at clinic k over the past five years.

A model that allows subject-specific baseline response to depend on these covariates is

$$\beta_{0jk} = \beta_{00} + \beta_{01} s_k + \beta_{02} w_{jk} + b_{0k} + b_{0jk}. \tag{10.22}$$

In (10.22), inference on β_{01} addresses the ***clinic-level*** issue of whether or not the average number of patients seen is associated with subject-specific baseline response; e.g., do patients with ***worse*** baseline response seek out clinics that treat ***larger numbers*** of patients with this condition? Likewise, β_{02} addresses the ***subject-level*** issue of whether or not weight is associated with baseline response.

The slope specification (10.20) can of course be modified analogously to incorporate dependence on such clinic- and subject-level characteristics. Under (10.22), the definitions of $\boldsymbol{A}_{jk}$ and β above would be revised in the ***obvious way***.

The model is completed by ***assumptions*** on the ***random effects*** $\boldsymbol{b}_k$ and $\boldsymbol{b}_{jk}$. The random effects corresponding to different levels of the model (clinics and subjects within clinics here) are ordinarily assumed to be ***independent*** of one another. Under the assumption that the $\boldsymbol{b}_k$ and $\boldsymbol{b}_{jk}$ are both ***iid*** , so that their distributions do not depend on covariates, the usual assumption is

$$\boldsymbol{b}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}^{(1)}), \qquad \boldsymbol{b}_{jk} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}^{(2)}), \tag{10.23}$$

where, in (10.23), $\boldsymbol{D}^{(1)}$ is a covariance matrix corresponding to the ***level 1*** (clinic) random effect and $\boldsymbol{D}^{(2)}$ is a covariance matrix corresponding to the ***level 2*** (subject) random effect.

- The ***elements*** of $\boldsymbol{D}^{(1)}$ thus characterize variation in intercepts and slope and covariation between them due to ***clinic-level*** phenomena, and those of $\boldsymbol{D}^{(2)}$ thus represent the additional variation and correlation due to ***subject-level*** sources.

GENERAL THREE-LEVEL MODEL: In general, substituting (10.21) in (10.18) yields, analogous to (6.44), a general model of the form

$$\boldsymbol{Y}_{jk} = \boldsymbol{X}_{jk}\beta + \boldsymbol{Z}^{(1)}_{jk}\boldsymbol{b}_k + \boldsymbol{Z}^{(2)}_{jk}\boldsymbol{b}_{jk} + \boldsymbol{e}_{jk}, \tag{10.24}$$

where, letting β be $(p \times 1)$, $\boldsymbol{b}_k$ be $(q_1 \times 1)$, and $\boldsymbol{b}_{jk}$ be $(q_2 \times 1)$,

$$\boldsymbol{X}_{jk} = \boldsymbol{C}_{jk}\boldsymbol{A}_{jk} \ (n_{1jk} \times p), \quad \boldsymbol{Z}^{(1)}_{jk} = \boldsymbol{C}_{jk}\boldsymbol{B}^{(1)}_{k} \ (n_{1jk} \times q_1), \quad \boldsymbol{Z}^{(2)}_{jk} = \boldsymbol{C}_{jk}\boldsymbol{B}^{(2)}_{jk} \ (n_{1jk} \times q_2);$$

and, as in (10.23), the usual assumption is that

$$\boldsymbol{b}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}^{(1)}), \qquad \boldsymbol{b}_{jk} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}^{(2)}), \qquad \boldsymbol{e}_{jk} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{n_{1jk}}). \tag{10.25}$$

As we have discussed, more general specifications would ***not*** take the random effects and $\boldsymbol{e}_{jk}$ to be ***independent*** of covariates $\boldsymbol{x}_{jk}$, where $\boldsymbol{x}_{jk}$ is the collection of all covariates on unit (j, k) at all levels; would allow ***serial correlation*** among elements of $\boldsymbol{e}_{jk}$ when level 1 units are longitudinal measurement occasions; and would allow the variance of elements of $\boldsymbol{e}_{jk}$ to be ***nonconstant*** and to depend on the random effects.

- Clearly, in principle, the general hierarchical formulation shown here for three levels can be ***extended*** to ***any number*** of levels.
- Likewise, the ***same considerations*** can be invoked to specify ***multilevel generalized linear mixed effects models*** and ***multilevel nonlinear mixed effects models***.

INFERENCE: Given the specification in (10.24)-(10.25), it is possible to write down the corresponding ***normal loglikelihood*** and to ***maximize*** it in β, σ^2, and the distinct elements of $\boldsymbol{D}^{(1)}$ and $\boldsymbol{D}^{(2)}$.

- This is routine in software such as SAS `proc mixed` and `proc glimmix` and in the R packages `nlme` and `lme4`.
- It is also possible to derive and obtain ***empirical Bayes estimates*** of ***posterior modes*** for $\boldsymbol{b}_k$ and $\boldsymbol{b}_{jk}$ and to use these to obtain empirical Bayes estimates of ***subject-specific parameters*** such as β_{jk}.

A practical introduction to multilevel models is given in Chapter 22 of Fitzmaurice, Laird, and Ware (2011). More generally, there is an enormous literature on these models, particularly in the literature on social and behavioral science.

We conclude this section by noting that, in the context of ***pharmacokinetics*** with ***time-dependent among-individual covariates*** as in (10.15), the same considerations can be invoked to specify models for individual-specific PK parameters that are taken to "***fluctuate***" over time intervals where such covariates change value. Reverting to the standard longitudinal data indexing, for example, the model (10.16) for clearance corresponding to $t_{ij} \in I_k$ could be modified as

$$\log Cl_{ij} = \beta_0 + \beta_1 a_{ik} + b_i + b_{ik}$$

to allow this fluctuation to depend on biological variation as well as systematic association with the changing covariate.

10.6 Distribution of random effects

We now return to the two-stage, subject-specific hierarchical linear and nonlinear mixed effects models for the standard longitudinal data structure discussed in Chapters 6 and 9. A standard, default assumption in these models, and one that is embedded in standard software, is that the ***random effects*** $\boldsymbol{b}_i$ are ***normally distributed***. In Chapter 9, we discussed in the context of pharmacokinetics how the individual-level model can be ***reparameterized*** so that individual-specific PK parameters can be taken to have ***skewed distributions*** in the population; however, this formulation is still predicated on normal random effects.

WHY SHOULD RANDOM EFFECTS BE NORMALLY DISTRIBUTED? In general, there is ***no fundamental principle*** that requires that random effects implicated in the distributions of ***individual-specific*** parameters need be normally distributed or, for that matter, have ***unimodal distributions***.

- It may well be that, in the population, individual-specific features such as intercepts, slopes, or PK parameters have ***heavy-tailed*** distributions relative to the normal, so that the extent of spread in the population of these values is greater than that dictated by a normal distribution.
- It may even be possible for such features to have ***multimodal*** distributions. For example, it could be that there are underlying ***subpopulations*** for which values of such features tend to "cluster" around different mean or "typical" values.
- Such ***subpopulations*** might correspond to a particular ***among-individual characteristic***. For example, suppose that in a PK setting values of drug clearance tend to be ***different*** for smokers and nonsmokers. If log clearance values cluster as above according to smoking status, the overall distribution of clearance values, if we could see it, would appear ***bimodal***.

  It could be that, ***within*** each of these subpopulations, the distribution of log clearance ***is approximately normal***. The result is that the overall distribution of log clearance is thus ***a mixture of normal*** distributions. Normal mixtures can be ***bimodal*** if the means each normal component of the mixture are sufficiently far apart.
- This suggests that, in implementation of these models, failure to include such a ***among-individual covariate*** in the population model could lead to evidence of ***nonnormality*** of random effects, perhaps through plots of empirical Bayes estimates.

RELAXING THE NORMALITY ASSUMPTION: There is a large body of work on models and methods for linear and nonlinear mixed effects models that allow the distribution of the random effects to be ***nonnormal***. In these approaches, under assumptions about the ***true random effects distribution***, the distribution itself is ***estimated*** from the data along with other model parameters.

- In one class of models and methods, ***no restrictions*** are placed on the nature of the random effects distribution. That is, the distribution is assumed to lie in the class of ***all probability distributions***; this includes ***discrete*** distributions and distributions that are a mixture of ***continuous*** and discrete components.

  In this case, the ***estimator*** for the random effects distribution is ***fully nonparametric*** and itself a ***discrete distribution***.

- When the random effects correspond to individual-specific features such as drug clearance or slope of an individual-specific trajectory that are naturally regarded as ***continuous***, distributions that are discrete do not seem ***realistic*** or ***plausible models*** for the nature of the random effects. The models and methods above thus include as possibilities distributions that are highly unlikely to be plausible representations of the true density.

  This has led to models and methods that take as a starting point the assumption that the distribution of the random effects has a ***density*** that obeys certain ***smoothness restrictions***; for example, that is has a certain number of continuous derivatives. The assumption that the random effects distribution has such a ***smooth density*** of necessity restricts the class of plausible distributions, but the tradeoffs are that the restricted class is highly likely to contain the true density and that the resulting estimate will itself lead to a smooth distribution.

In the context of mixed effects models, there have been many proposals along both of these lines; see Davidian and Giltinan (1995, Chapter 7), Verbeke and Molenberghs (2000, Chapter 12), and and Zhang and Davidian (2001) for review of these approaches. In a Bayesian formulation of mixed effects model, similar developments have been proposed in which the random effects distribution F_b is taken to be unrestricted or to lie within some restricted class; see, for example, Müller and Rosner (1997).

Here, we briefly describe two popular approaches to representing the distribution of the random effects when it is assumed to have a ***smooth density***.

MIXTURE OF NORMAL DENSITIES: A number of authors have proposed to represent the distribution of random effects by a ***mixture of normal*** distributions; see the above references. A version of this has been called the ***heterogeneity model*** by Verbeke and Molenberghs (2000, Chapter 12). Here, under the assumption of ***iid random effects*** $\boldsymbol{b}_i$ ($q \times 1$), the usual normality assumption is replaced by

$$\boldsymbol{b}_i \sim \sum_{k=1}^{K} p_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{D}_k), \qquad \sum_{k=1}^{K} p_k = 1, \tag{10.26}$$

where the contraint

$$\sum_{k=1}^{K} p_k \boldsymbol{\mu}_k = \boldsymbol{0} \tag{10.27}$$

is imposed to ensure that $E(\boldsymbol{b}_i) = \boldsymbol{0}$.

The representation (10.26) is standard and can be interpreted as saying that the population is a ***combination*** of K ***subpopulations***, where the kth population is a fraction p_k of the overall populations. Under this model, we can view each individual i as being a member of one of the underlying subpopulations.

It is straightforward to show, defining $u_{ik} = 1$ if the random effect $\boldsymbol{b}_i$ for individual i is from the kth subpopulation and = 0 otherwise and using a conditioning argument (try it), that the ***overall covariance matrix*** var($\boldsymbol{b}_i$) is

$$\boldsymbol{D} = \sum_{k=1}^{K} p_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T + \sum_{k=1}^{K} p_k \boldsymbol{D}_k. \tag{10.28}$$

IMPLEMENTATION: Under (10.26), it follows that, with an individual model involving within-individual covariance parameters γ and a population model involving fixed parameters β,

$$p(\boldsymbol{y}_i|\boldsymbol{x}_i; \beta, \gamma, \boldsymbol{D}) = \sum_{k=1}^{K} p_k \int p(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)\, p_k(\boldsymbol{b}_i; \mu_k, \boldsymbol{D}_k)\, d\boldsymbol{b}_i, \tag{10.29}$$

where $p_k(\,\cdot\,; \mu_k, \boldsymbol{D}_k)$ is the q-variate normal density with mean μ_k and covariance matrix $\boldsymbol{D}_k$.; and $\boldsymbol{D}$ depends on all of these as in (10.28).

- In this formulation, the number of mixture components K can be regarded as a ***tuning parameter*** controlling the ***flexibility*** of the representation. The more components in a mixture, the more flexible it is for representing ***complicated true densities***; however, this is at the expense of ***more parameters*** that need to be estimated; namely, more p_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{D}_k$.

- It is customary to treat K as ***known***, fit the model incorporating (10.26) subject to the constraints (10.27) by maximizing the loglikelihood corresponding to (10.29) for $i = 1, \ldots, m$. Here, $K = 0$ corresponds to the ***usual normality assumption***, and successively increasing K leads to richer parameterizations of the density. One thus fits the model with $K = 0$, and usually $K = 1, 2$, and 3.

- Testing for the number of components K in a mixture as in (10.26) is subject to ***boundary problems*** similar to those we discussed for variance parameters in Section 6.6. Thus, it is not possible to deduce K under the assumption that the mixture model (10.26) is ***correct*** via ***likelihood ratio tests*** comparing models with increasing K, for example.

- An ad hoc approach is to inspect ***information criteria*** and choose the value of K that optimizes a given criterion.

For given K, it is possible to maximize the loglikelihood corresponding to (10.29) via an ***EM algorithm***. See Verbeke and Molenberghs (2000, Chapter 12).

SEMINONPARAMETRIC (SNP) DENSITY REPRESENTATION: An ***alternative approach*** to representing the density of the random effects was proposed for use with the nonlinear mixed effects model by Davidian and Gallant (1993) and represented in an advantageous parameterization by Zhang and Davidian (2001). In this formulation, under the assumption of ***iid random effects*** $\boldsymbol{b}_i$ $(q \times 1)$, the random effects are first written as

$$\boldsymbol{b}_i = \boldsymbol{\mu} + \boldsymbol{S}\boldsymbol{U}_i \quad (q \times 1), \tag{10.30}$$

where $\boldsymbol{\mu}$ is a $(q \times 1)$ vector of parameters, $\boldsymbol{S}$ is a lower triangular matrix, and $\boldsymbol{U}_i$ is a $(q \times 1)$ random vector. If the random vector $\boldsymbol{U}_i$ in (10.30) is taken to be ***standard multivariate normal***, and $\boldsymbol{\mu} = \boldsymbol{0}$, then (10.30) reduces to the usual normality assumption with $\boldsymbol{D} = \boldsymbol{S}\boldsymbol{S}^T$.

Instead, $\boldsymbol{U}_i$ is taken to have a ***smooth density*** that falls in a class of such densities that can be represented by an ***infinite series expansion***. As proposed by Davidian and Gallant (1993), the idea is to approximate this density, and thus the density of $\boldsymbol{b}_i$ in (10.30), by a ***truncated series expansion***; this has been referred to as the ***seminonparametric*** (SNP) approximation.

That is, the density of $\boldsymbol{U}_i$ is represented as

$$h_K(\boldsymbol{u}, \boldsymbol{a}) = P_K^2(\boldsymbol{u}, \boldsymbol{a})\, \varphi(\boldsymbol{u}) = \left\{ \sum_{|\lambda| \le K} a_\lambda \boldsymbol{u}^\lambda \right\}^2 \varphi(\boldsymbol{u}), \tag{10.31}$$

where $\varphi(\boldsymbol{u})$ is the standard q-variate normal density; $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)^T$ is a vector of nonnegative integers; $\boldsymbol{u}^\lambda = u_1^{\lambda_1} \cdots u_q^{\lambda_q}$, the ***monomial*** of order $|\boldsymbol{\lambda}| = \sum_{\ell=1}^q \lambda_\ell$, and $P_K(\boldsymbol{u}, \boldsymbol{a})$ is thus a polynomial of order K with coefficients collected in the vector $\boldsymbol{a}$. For example, when $K = 2$, $q = 2$,

$$P_K(\boldsymbol{u}) = a_{00} + a_{10} u_1 + a_{01} u_2 + a_{20} u_1^2 + a_{11} u_1 u_2 + a_{02} u_2^2,$$

and $\boldsymbol{a} = (a_{00}, a_{10}, a_{01}, a_{20}, a_{11}, a_{02})^T$.

In the representation (10.31), as in the ***normal mixture*** representation (10.26), the degree of the polynomial K plays the role of a a ***tuning parameter*** controlling the ***flexibility*** of the representation. The higher the order of the polynomial, the more flexible it is for representing ***complicated true densities***; however, this is again at the expense of ***more parameters*** that need to be estimated. Vast experience shows that $K = 1$ or 2 is often sufficient to approximate complex shapes, including ***multimodality and skewness***.

For (10.31) to be a ***legitimate density***, the coefficients in the polynomial $P_K(\boldsymbol{u}, \boldsymbol{a})$ must be chosen so that

$$\int h_K(\boldsymbol{u}, \boldsymbol{a})\, d\boldsymbol{u} = 1. \tag{10.32}$$

For (10.32) to hold, then, it is necessary to ***impose a constraint*** on the coefficients in $P_K(\boldsymbol{u}, \boldsymbol{a})$. For example, when $K = 0$, it must be that $a_{00} = 1$, so that $h_K(\boldsymbol{u})$ reduces to a standard normal density, and thus from (10.30) $\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S}\boldsymbol{S}^T)$. A special case is to take $\boldsymbol{\mu} = \boldsymbol{0}$; otherwise, in ***linear population models*** for which the kth element of $\boldsymbol{\beta}_i$ is of the form

$$\beta_{ki} = \beta_{k0} + \boldsymbol{\beta}_{k1}^T h(\boldsymbol{a}_i) + b_{ki},$$

say, with "***intercept***" β_{k0}, where $h(\cdot)$ is a function of among-individual covariates $\boldsymbol{a}_i$, with arbitrary $\boldsymbol{\mu}$, one would write instead

$$\beta_{ki} = \boldsymbol{\beta}_{k1}^T h(\boldsymbol{a}_i) + b_{ki},$$

so that μ_k, the kth element of $\boldsymbol{\mu}$, plays the role of the population model "intercept." In the following, we take $\boldsymbol{\mu} = \boldsymbol{0}$.

More generally, to ensure (10.32), note from (10.31) that it must be that

$$E\{P_K^2(\boldsymbol{U}, \boldsymbol{a})\} = 1, \quad \text{where } \boldsymbol{U} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_q). \tag{10.33}$$

It is possible to show that the expectation in (10.33) can be written as a ***quadratic form*** involving the coefficients in $P_K^2(\boldsymbol{U}, \boldsymbol{a})$ and ***moments of a standard normal random variable***, as follows. For example, for $q = 2$, there are $d = (K+1)(K+2)/2$ distinct coefficients; for $K = 3$,

$$\boldsymbol{a} = (a_{00}, a_{10}, a_{01}, a_{20}, a_{11}, a_{02}, a_{30}, a_{21}, a_{12}, a_{03})^T.$$

is the $(d \times 1)$ vector of these coefficients. Writing the ℓth element of $\boldsymbol{a}$ as $a_{\ell_1 \ell_2}$, $\ell = 1, \ldots, d$, let $\boldsymbol{U}_a$ be the random vector whose ℓth element is $U_1^{\ell_1} U_2^{\ell_2}$, where U_1 and U_2 are independent, standard normal random variables, $\ell = 1, \ldots, d$. Then it can be shown that $P_K(\boldsymbol{U}, \boldsymbol{a}) = \boldsymbol{a}^T \boldsymbol{U}_a$, so that

$$E\{P_K^2(\boldsymbol{U}) = \boldsymbol{a}^T E(\boldsymbol{U}_a \boldsymbol{U}_a^T)\boldsymbol{a} = \boldsymbol{a}^T \boldsymbol{A}\boldsymbol{a}, \tag{10.34}$$

where $\boldsymbol{A} = E(\boldsymbol{U}_a \boldsymbol{U}_a^T)$ is the matrix with (ℓ, ℓ') element $E(U_1^{\ell_1+\ell'_1})E(U_2^{\ell_2+\ell'_2})$. These moments are readily available via ***standard recursive formulæ***. This formulation can be ***generalized*** to any q and K. It follows from (10.34) that (10.33) can be written as

$$\boldsymbol{a}^T \boldsymbol{A}\boldsymbol{a} = 1. \tag{10.35}$$

Thus, from (10.30) with $\boldsymbol{\mu} = \boldsymbol{0}$ and (10.31), we can represent the density of $\boldsymbol{b}_i$ for fixed K as

$$h_K(\boldsymbol{b}_i, \boldsymbol{a}) = P_K^2(\boldsymbol{S}^{-1}\boldsymbol{b}_i, \boldsymbol{a})\, p(\boldsymbol{b}_i; \boldsymbol{0}, \boldsymbol{S}\boldsymbol{S}^T), \tag{10.36}$$

where $p(\cdot\,; \boldsymbol{0}, \boldsymbol{S}\boldsymbol{S}^T)$ is the $\mathcal{N}(\boldsymbol{0}, \boldsymbol{S}\boldsymbol{S}^T)$ density, and $\boldsymbol{a}$ satisfies the ***constraint*** (10.35). It follows from (10.36) that, with an individual model involving within-individual covariance parameters γ and a population model involving fixed parameters β,

$$p(\boldsymbol{y}_i | \boldsymbol{x}_i; \beta, \gamma, \boldsymbol{D}) = \int p(\boldsymbol{Y}_i | \boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)\, P_K^2(\boldsymbol{S}^{-1}\boldsymbol{b}_i, \boldsymbol{a})\, p(\boldsymbol{b}_i; \boldsymbol{0}, \boldsymbol{S}\boldsymbol{S}^T)\, d\boldsymbol{b}_i, \tag{10.37}$$

where $\boldsymbol{D}$ is now determined by $\boldsymbol{S}$ and $\boldsymbol{a}$. Using the representation (10.37), one can write down loglikelihood for β, γ, $\boldsymbol{a}$, and $\boldsymbol{S}$ and maximize in these parameters ***subject to the constraint*** (10.35).

- Zhang and Davidian (2001) show that, for a ***linear mixed effects model***, the loglikelihood corresponding to (10.37) can be expressed in a ***closed form***.

- For general ***nonlinear*** models, the loglikeliood is ***not*** available in a closed form in general. In this case, Davidian and Gallant (1993) suggest "doing" the integrals in (10.36) using ***Gaussian quadrature*** or ***Monte Carlo integration***, taking advantage of the fact that they depend on integration against a ***normal density***.

- Zhang and Davidian (2001) show further that, regardless of the form of the model, the constraint (10.35) can be imposed "***automatically***" by a ***reparameterization*** via a ***spherical transformation*** of (10.35).

- As with the mixture of normals representation, it is suggested to choose K by fitting the model first for $K = 0$ corresponding to normal random effects, and followed by successive fits with $K = 1, 2$, and 3, and selecting K via inspection of ***information criteria***.

PHARMACOKINETICS OF PHENOBARBITAL IN NEONATES, continued. We conclude with a brief look at an analysis of the phenobarbital PK study using the SNP representation for the random effects; a full account is in Chapter 7 of Davidian and Giltinan (1997). Assuming the ***individual model*** (9.98) with repeated dosing and the ***population model***

$$(ii) \quad \beta_{1i} = \beta_1 + \beta_3 w_i + b_{1i}, \quad \beta_{1i} = \beta_2 + \beta_4 w_i + b_{2i},$$

the density of $\boldsymbol{b}_i = (b_{1i}, b_{2i})^T$ is represented by the SNP approximation (10.30) and (10.31).

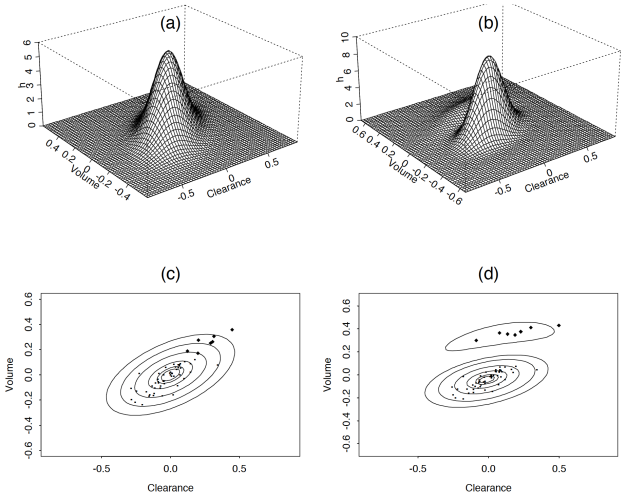

Figure 10.3: *Estimated random effects densities for $\boldsymbol{b}_i$ using the SNP approach. The left hand column shows the estimated density and contour plot with empirical Bayes estimates of random effects superimposed when $K = 0$, corresponding to normal random effects. The right hand column shows the same when $K = 2$ in the SNP representation.*

Figure 10.3 shows plots of the ***estimated random effects density*** with ***empirical Bayes estimates*** of the random effects superimposed for $K = 0$, corresponding to the usual assumption of normal random effects, and $K = 2$, the preferred fit based on inspection of ***information criteria***. The estimated density for $K = 2$ shows a ***clear second mode***.

The random effects estimates for seven infants are shown as diamonds; these seven infants appear to correspond to the second mode and thus seem to arise from a separate ***subpopulation*** relative to the rest of the infants. These infants can be seen upon inspection of the data to have low measured phenobarbital concentrations after the loading dose. Because the initial concentration measurement is highly influential for determining the estimate of log ***volume of distribution*** $\log V_i$, the observed pattern makes sense. Given that birthweight is already accounted for in the population model and Apgar score does not seem associated with either $\log Cl_i$ or $\log V_i$ from the analysis in Section 9.7, it is not possible to explain this by a possible systematic association with an infant characteristic. It may well be that a relevant, ***unmeasured attribute*** is being reflected here.

Appendix A: Fun Matrix Facts

For convenience, we summarize several useful matrix facts here.

SQUARE MATRIX RESULTS: Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be square matrices of the same dimension. Inverses below are assumed to exist.

- $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$, $(\boldsymbol{A}^{-1})^T = (\boldsymbol{A}^T)^{-1}$.
- We denote the ***determinant*** of $\boldsymbol{A}$ by $|\boldsymbol{A}|$.
- $|\boldsymbol{A}| = |\boldsymbol{A}^T|$, $|\boldsymbol{A}| = 1/|\boldsymbol{A}^{-1}|$
- $|\boldsymbol{AB}| = |\boldsymbol{A}||\boldsymbol{B}|$
- $(\boldsymbol{A}+\boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}(\boldsymbol{A}^{-1}+\boldsymbol{B}^{-1})^{-1}\boldsymbol{A}^{-1}$
- $(\boldsymbol{A}+\boldsymbol{CBD})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{C}(\boldsymbol{B}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{C})^{-1}\boldsymbol{DA}^{-1}$. Here, $\boldsymbol{A}$ and $\boldsymbol{B}$ need not be of the same dimension, and $\boldsymbol{C}$ and $\boldsymbol{D}$ are conformable matrices.
- The following are equivalent: (i) $\boldsymbol{A}$ is ***nonsingular*** , (ii) $|\boldsymbol{A}| \neq 0$, (iii) $\boldsymbol{A}^{-1}$ exists.
- We denote the ***trace*** of a square matrix $\boldsymbol{A}$ by $\text{tr}(\boldsymbol{A})$.
- $\text{tr}(\boldsymbol{A}) = \text{tr}(\boldsymbol{A}^T)$, $\text{tr}(b\boldsymbol{A}) = b\text{tr}(\boldsymbol{A})$
- $\text{tr}(\boldsymbol{A}+\boldsymbol{B}) = \text{tr}(\boldsymbol{A}) + \text{tr}(\boldsymbol{B})$, $\text{tr}(\boldsymbol{AB}) = \text{tr}(\boldsymbol{BA})$
- If $\boldsymbol{A}$ is $(n \times n)$ and $\boldsymbol{x}$ is $(n \times 1)$, then the ***quadratic form***

  $$\boldsymbol{x}^T\boldsymbol{Ax}$$

  and $\boldsymbol{A}$ are ***nonnegative definite*** if $\boldsymbol{x}^T\boldsymbol{Ax} \geq 0$. The quadratic form and $\boldsymbol{A}$ are ***positive definite*** if $\boldsymbol{x}^T\boldsymbol{Ax} > 0$. If $\boldsymbol{A}$ is positive definite, then it is symmetric and nonsingular (so its inverse exists).
- $\boldsymbol{x}^T\boldsymbol{Ax} = \text{tr}(\boldsymbol{Axx}^T)$
- If $\boldsymbol{x}$ is a ***random vector*** with mean μ and covariance matrix $\boldsymbol{V}$, then

  $$E(\boldsymbol{x}^T\boldsymbol{Ax}) = \text{tr}\{E(\boldsymbol{xx}^T)\boldsymbol{A}\} = \text{tr}(\boldsymbol{VA}) + \mu^T\boldsymbol{A}\mu = \text{tr}(\boldsymbol{AV}) + \mu^T\boldsymbol{A}\mu.$$

vec and vech NOTATION:

- For a $(n \times r)$ matrix $\boldsymbol{A}$, vec($\boldsymbol{A}$) is defined as the $(nr \times 1)$ vector consisting of the r columns of $\boldsymbol{A}$ stacked in the order $1, \ldots, r$.

- If furthermore $\boldsymbol{A}$ is $(n \times n)$ and symmetric, then vec($\boldsymbol{A}$) contains redundant entries. The vech$(\cdot)$ operator yields the column vector containing all the distinct entries of $\boldsymbol{A}$ by stacking the lower diagonal elements; e.g., for $n = 3$,

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{pmatrix} \quad \text{and} \quad \text{vech}(\boldsymbol{A}) = \begin{pmatrix} a_{11} \\ a_{12} \\ a_{13} \\ a_{22} \\ a_{23} \\ a_{33} \end{pmatrix}.$$

- For matrices $\boldsymbol{A}$ $(a \times a)$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$,

  (i) $\text{tr}(\boldsymbol{AB}) = \{\text{vec}(\boldsymbol{A})\}^T\{\text{vec}(\boldsymbol{B}^T)\} = \{\text{vec}(\boldsymbol{A}^T)\}^T\{\text{vec}(\boldsymbol{B})\}$.

  (ii) $\text{tr}(\boldsymbol{ABD}^T\boldsymbol{C}^T) = \{\text{vec}(\boldsymbol{A})\}^T(\boldsymbol{B} \otimes \boldsymbol{C})\text{vec}(\boldsymbol{D})$, where $\otimes$ represents Kronecker product.

  (iii) For $\boldsymbol{A}$ symmetric, there is a relationship between vec($\boldsymbol{A}$) and vech($\boldsymbol{A}$). In particular, there exists a unique matrix $\boldsymbol{\Phi}$ of dimension $\{a^2 \times a(a+1)/2\}$ such that

  $$\text{vec}(\boldsymbol{A}) = \boldsymbol{\Phi}\text{vech}(\boldsymbol{A}).$$

  Clearly, $\boldsymbol{\Phi}$ is unique and of full column rank, as there is only one way to write the distinct elements of $\boldsymbol{A}$ in a full, redundant vector.

INVERSE OF PARTITIONED MATRIX: Consider a generic $(k \times k)$ matrix

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{pmatrix},$$

where the $\boldsymbol{C}_{ij}$ are submatrices such that $\boldsymbol{C}_{11}$ is $(k_1 \times k_1)$ and $\boldsymbol{C}_{22}$ is $(k_2 \times k_2)$ such that $k = k_1 + k_2$, and $\boldsymbol{C}_{11}^{-1}$ and $\boldsymbol{C}_{22}^{-1}$ exist, as do all other inverses below. Then

$$\boldsymbol{C}^{-1} = \begin{pmatrix} \boldsymbol{D}_{11} & \boldsymbol{D}_{12} \\ \boldsymbol{D}_{21} & \boldsymbol{D}_{22} \end{pmatrix},$$

where

$$\begin{aligned}
\boldsymbol{D}_{11} &= (\boldsymbol{C}_{11} - \boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21})^{-1} \\
\boldsymbol{D}_{22} &= (\boldsymbol{C}_{22} - \boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}\boldsymbol{C}_{12})^{-1} = \boldsymbol{C}_{22}^{-1} + \boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21}\boldsymbol{D}_{11}\boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1} \\
\boldsymbol{D}_{12} &= -\boldsymbol{C}_{11}^{-1}\boldsymbol{C}_{12}\boldsymbol{D}_{22} = -\boldsymbol{D}_{11}\boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1} \\
\boldsymbol{D}_{21} &= -\boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21}\boldsymbol{D}_{11}.
\end{aligned}$$

MATRIX DIFFERENTIATION: Let $\boldsymbol{x}$ be a $(n \times 1)$ vector depending on a $(p \times 1)$ vector β, and let $\boldsymbol{A}$ be a $(n \times n)$ square matrix.

- For quadratic form $Q = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$, $\partial Q/\partial \boldsymbol{x} = 2\boldsymbol{A}\boldsymbol{x}$. Note that this is a $(n \times 1)$ vector.
- The chain rule then gives $\partial Q/\partial \beta = (\partial \boldsymbol{x}/\partial \beta)(\partial Q/\partial \boldsymbol{x})$. Note that $\partial \boldsymbol{x}/\partial \beta$ is a $(p \times n)$ matrix.

Let $\boldsymbol{V}(\xi)$ be a $(n \times n)$ nonsingular matrix depending on a $(q \times 1)$ (parameter) vector ξ.

- If ξ_k is the kth element of ξ, then $\partial/\partial\xi_k\, \boldsymbol{V}(\xi)$ is the $(n \times n)$ matrix whose (ℓ, p) element is the partial derivative of the (ℓ, p) element of $\boldsymbol{V}(\xi)$ with respect to ξ_k.
- $\partial/\partial\xi_k\{\log|\boldsymbol{V}(\xi)|\} = \text{tr}\left[\boldsymbol{V}^{-1}(\xi)\{\partial/\partial\xi_k \boldsymbol{V}(\xi)\}\right]$.
- $\partial/\partial\xi_k \boldsymbol{V}^{-1}(\xi) = -\boldsymbol{V}^{-1}(\xi)\left\{\partial/\partial\xi_k\, \boldsymbol{V}(\xi)\right\}\boldsymbol{V}^{-1}(\xi)$.
- For quadratic form $Q = \boldsymbol{x}^T\boldsymbol{V}(\xi)\boldsymbol{x}$, $\partial Q/\partial\xi_k = \boldsymbol{x}^T\{\partial/\partial\xi_k\, \boldsymbol{V}(\xi)\}\boldsymbol{x}$. Thus, from the previous result,

$$\partial/\partial\xi_k\left\{\boldsymbol{x}^T\boldsymbol{V}^{-1}(\xi)\boldsymbol{x}\right\} = -\boldsymbol{x}^T\boldsymbol{V}^{-1}(\xi)\left\{\partial/\partial\xi_k\, \boldsymbol{V}(\xi)\right\}\boldsymbol{V}^{-1}(\xi)\boldsymbol{x}.$$

Appendix B: Notation and Taylor Series

The following is a generic review of Taylor series and the type of notation we use in certain parts of the course.

REAL-VALUED FUNCTIONS: Let $h(\boldsymbol{x}, \boldsymbol{\alpha})$ be a real-valued function of a vector $\boldsymbol{x}$ (which is irrelevant to the developments here) and a $(r \times 1)$ vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_r)^T$. We write

$$h_\alpha(\boldsymbol{x}, \boldsymbol{\alpha}) = \partial/\partial\boldsymbol{\alpha}\, h(\boldsymbol{x}, \boldsymbol{\alpha}) = \begin{pmatrix} \partial/\partial\alpha_1 h(\boldsymbol{x}, \boldsymbol{\alpha}) \\ \vdots \\ \partial/\partial\alpha_r h(\boldsymbol{x}, \boldsymbol{\alpha}) \end{pmatrix} \quad (r \times 1). \tag{B.1}$$

The vector $h_\alpha(\boldsymbol{x}, \boldsymbol{\alpha})$ of partial derivatives of $h(\boldsymbol{x}, \boldsymbol{\alpha})$ with respect to the elements of $\boldsymbol{\alpha}$ is referred to as the ***gradient*** (vector). Of course, $h_\alpha^T(\boldsymbol{x}, \boldsymbol{\alpha})$ denotes its transpose, a $(1 \times r)$ vector.

We can extend this notation to second partial derivatives with respect to the elements of $\boldsymbol{\alpha}$. This is done by writing the $(r \times r)$ symmetric matrix of second partial derivatives of $h(\boldsymbol{x}, \boldsymbol{\alpha})$ as follows.

$$h_{\alpha\alpha}(\boldsymbol{x}, \boldsymbol{\alpha}) = \partial/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T h(\boldsymbol{x}, \boldsymbol{\alpha}) = \begin{pmatrix} \partial^2/\partial\alpha_1^2 h(\boldsymbol{x}, \boldsymbol{\alpha}) & \partial^2/\partial\alpha_1\partial\alpha_2 h(\boldsymbol{x}, \boldsymbol{\alpha}) & \cdots & \partial^2/\partial\alpha_1\partial\alpha_r h(\boldsymbol{x}, \boldsymbol{\alpha}) \\ & \partial^2/\partial\alpha_2^2 h(\boldsymbol{x}, \boldsymbol{\alpha}) & \cdots & \partial^2/\partial\alpha_2\partial\alpha_r h(\boldsymbol{x}, \boldsymbol{\alpha}) \\ & & \ddots & \vdots \\ & & & \partial^2/\partial\alpha_r^2 h(\boldsymbol{x}, \boldsymbol{\alpha}) \end{pmatrix} \tag{B.2}$$

This is often called the ***Hessian***.

More generally, suppose that $h(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\delta})$ is a real-valued function of two vectors $\boldsymbol{\alpha}$ $(r \times 1)$ and $\boldsymbol{\delta}$ $(s \times 1)$. Then define

$$h_{\alpha\delta}(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\delta}) = \partial/\partial\boldsymbol{\alpha}\partial\boldsymbol{\delta}^T h(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\delta}) = \begin{pmatrix} \partial^2/\partial\alpha_1\partial\delta_1 h & \partial^2/\partial\alpha_1\partial\delta_2 h & \cdots & \partial^2/\partial\alpha_1\partial\delta_s h \\ \partial^2/\partial\alpha_2\partial\delta_1 h & \partial^2/\partial\alpha_2\partial\delta_2 h & \cdots & \partial^2/\partial\alpha_2\partial\delta_s h \\ \vdots & \vdots & \ddots & \vdots \\ \partial^2/\partial\alpha_r\partial\delta_1 h & \cdots & \cdots & \partial^2/\partial\alpha_r\partial\delta_s h \end{pmatrix}, \tag{B.3}$$

where we have written $h = h(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\delta})$ for brevity. This is a $(r \times s)$ matrix; it follows that, by reducing the roles of $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$, $h_{\delta\alpha}$ is a $(s \times r)$ matrix, and is the transpose of $h_{\alpha\delta}$ in (B.3).

VECTOR-VALUED FUNCTIONS: Now suppose that $\boldsymbol{h}(\boldsymbol{x}, \delta)$ is a vector-valued function of dimension n of a vector $\boldsymbol{x}$ and parameter δ ($s \times 1$); i.e.

$$\boldsymbol{h}(\boldsymbol{x}, \delta) = \begin{pmatrix} h_1(\boldsymbol{x}, \delta) \\ \vdots \\ h_n(\boldsymbol{x}, \delta) \end{pmatrix}.$$

Thus, the vector-valued function $\boldsymbol{h}$ has real-valued component functions $h_1, \ldots, h_n$.

We will write

$$\boldsymbol{h}_\delta(\boldsymbol{x}, \delta) = \partial / \partial \delta^T \boldsymbol{h}(\boldsymbol{x}, \delta) = \begin{pmatrix} \partial/\partial\delta_1 h_1(\boldsymbol{x}, \delta) & \cdots & \partial/\partial\delta_s h_1(\boldsymbol{x}, \delta) \\ \vdots & \vdots & \vdots \\ \partial/\partial\delta_1 h_n(\boldsymbol{x}, \delta) & \cdots & \partial/\partial\delta_s h_n(\boldsymbol{x}, \delta) \end{pmatrix},$$

a ($n \times s$) matrix. In particular, note that if we have a real-valued function $h(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\delta})$ for α ($r \times 1$) and $\boldsymbol{\delta}$ ($s \times 1$), then

$$h_\alpha(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\delta}) = \begin{pmatrix} \partial/\partial\alpha_1 h(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\delta}) \\ \vdots \\ \partial/\partial\alpha_r h(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\delta}) \end{pmatrix} = \begin{pmatrix} h_{\alpha_1}(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\delta}) \\ \vdots \\ h_{\alpha_r}(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\delta}) \end{pmatrix} \quad (r \times 1).$$

Thus, $h_\alpha(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\delta})$ is a vector-valued function. Applying the above definition of the partial derivative of a vector-valued function with respect to a vector parameter to $h_\alpha(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\delta})$, we may conclude that the result of differentiating $h_\alpha(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\delta})$ with respect to $\boldsymbol{\delta}$ is the ($r \times s$) matrix $h_{\alpha\delta}(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\delta})$ defined in (B.3); that is,

$$\partial / \partial \boldsymbol{\delta}^T h_\alpha(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\delta}) = \partial / \partial \boldsymbol{\alpha} \partial \boldsymbol{\delta}^T h(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\delta}).$$

We now consider various forms of Taylor's theorem.

UNIVARIATE TAYLOR'S THEOREM: Assume $h(\alpha)$ has $(k-1)$ continuous derivatives in $[a, b]$ and finite kth derivative in (a, b), where α is univariate and h is a real-valued function. Let $\alpha_0 \in [a, b]$. For each $\alpha \in [a, b]$, $\alpha \neq \alpha_0$, there exists α_* interior to the interval joining α_0 and α such that

$$h(\alpha) = h(\alpha_0) + \sum_{\ell=1}^{k-1} \frac{1}{\ell!} \{\partial^\ell / \partial \alpha^\ell h(\alpha)\}_{\alpha=\alpha_0} (\alpha - \alpha_0)^\ell + \frac{1}{k!} \{\partial^k / \partial \alpha^k h(\alpha)\}_{\alpha=\alpha_*} (\alpha - \alpha_0)^k.$$

Taylor's theorem is used heavily in making large sample arguments in statistics. In such arguments, we usually deal with vector-valued functions of vector-values parameters, for which a multivariate version of Taylor's theorem is required.

MULTIVARIATE TAYLOR'S THEOREM: First consider a real-valued function $h(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is $(r \times 1)$. We I state the theorem for general k, but write the form of the representation of $h(\boldsymbol{\alpha})$ for $k = 2$ only, as things get messy pretty quickly. One should be able to deduce the form for larger k by analogy to the univariate version.

Assume that $h(\boldsymbol{\alpha})$ on $\mathcal{R}^r$ has continuous partial derivatives of order k at each point in an open set $S \subset \mathcal{R}^r$. Let $\boldsymbol{\alpha}_0 \in S$. For each $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ such that the line segment joining $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}_0$ lies in S, there exists $\boldsymbol{\alpha}_*$ in the interior of this line segment such that, in the case $k = 2$,

$$h(\boldsymbol{\alpha}) = h(\boldsymbol{\alpha}_0) + \sum_{\ell=1}^{r} \{\partial/\partial \boldsymbol{\alpha}_\ell h(\boldsymbol{\alpha})\}_{\alpha=\alpha_0} (\boldsymbol{\alpha}_\ell - \boldsymbol{\alpha}_{0,\ell}) + (1/2) \sum_{\ell=1}^{r} \sum_{t=1}^{r} \{\partial^2/\partial \boldsymbol{\alpha}_\ell \partial \boldsymbol{\alpha}_t h(\boldsymbol{\alpha})\}_{\alpha=\alpha_*} (\boldsymbol{\alpha}_\ell - \boldsymbol{\alpha}_{0,\ell})(\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_{0,t}).$$

Using the shorthand notation above, we can write this more succinctly as

$$h(\boldsymbol{\alpha}) = h(\boldsymbol{\alpha}_0) + h_\alpha^T(\boldsymbol{\alpha}_0)(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0) + (1/2)(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T h_{\alpha\alpha}(\boldsymbol{\alpha}_*)(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0).$$

Here, we have used a further shorthand

$$h_\alpha(\boldsymbol{\alpha}_0) = \{\partial/\partial \boldsymbol{\alpha} h(\boldsymbol{\alpha})\}_{\alpha=\alpha_0}.$$

We use similar notation for $h_{\alpha\alpha}(\boldsymbol{\alpha})$ and other expressions.

If h is a function of two vectors $\boldsymbol{\alpha}$ $(r \times 1)$ and $\boldsymbol{\delta}$ $(s \times 1)$, we can define the single, "stacked" vector $(\boldsymbol{\alpha}^T, \boldsymbol{\delta}^T)^T$ $(r + s \times 1)$ and apply the theorem to obtain a representation of $h(\boldsymbol{\alpha}, \boldsymbol{\delta})$ about some value $(\boldsymbol{\alpha}_0^T, \boldsymbol{\delta}_0^T)^T$. This is handy when we want to maintain the distinction between two arguments of a function and treat them separately. Using the above, it is easy to show that, for $k = 2$, we can write

$$\begin{aligned} h(\boldsymbol{\alpha}, \boldsymbol{\delta}) = {} & h(\boldsymbol{\alpha}_0, \boldsymbol{\delta}_0) + \{h_\alpha^T(\boldsymbol{\alpha}_0, \boldsymbol{\delta}_0)(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0) + h_\delta^T(\boldsymbol{\alpha}_0, \boldsymbol{\delta}_0)(\boldsymbol{\delta} - \boldsymbol{\delta}_0)\} \\ & + (1/2)\{(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T h_{\alpha\alpha}(\boldsymbol{\alpha}_*, \boldsymbol{\delta}_*)(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0) + 2(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T h_{\alpha\delta}(\boldsymbol{\alpha}_*, \boldsymbol{\delta}_*)(\boldsymbol{\delta} - \boldsymbol{\delta}_0) + \\ & (\boldsymbol{\delta} - \boldsymbol{\delta}_0)^T h_{\delta\delta}(\boldsymbol{\alpha}_*, \boldsymbol{\delta}_*)(\boldsymbol{\delta} - \boldsymbol{\delta}_0)\}. \end{aligned}$$

It is often necessary to apply Taylor's theorem to vector-valued functions. This can appear complicated, but it is mainly a matter of notation. Suppose $\boldsymbol{h}(\boldsymbol{\alpha}) = [h_1(\boldsymbol{\alpha}), \ldots, h_v(\boldsymbol{\alpha})]^T$ $(v \times 1)$, where again $\boldsymbol{\alpha}$ is $(r \times 1)$.

If we apply the multivariate Taylor theorem to each component of $\boldsymbol{h}$ (each of which is a real-valued function), it can be shown that the expansion of $\boldsymbol{h}(\boldsymbol{\alpha})$ about $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ for $k = 2$ can be written compactly as

$$\boldsymbol{h}(\boldsymbol{\alpha}) = \boldsymbol{h}(\boldsymbol{\alpha}_0) + \begin{pmatrix} h_{1\alpha}^T(\boldsymbol{\alpha}_0) \\ \vdots \\ h_{v\alpha}^T(\boldsymbol{\alpha}_0) \end{pmatrix} (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0) + (1/2)\{\boldsymbol{I}_v \otimes (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\} \boldsymbol{H}_{\alpha\alpha}^* (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0),$$

where $\boldsymbol{H}_{\alpha\alpha}^*$ is the $(rv \times r)$ matrix consisting of the $(r \times r)$ matrices $h_{1\alpha\alpha}(\boldsymbol{\alpha}_*), \ldots, h_{v\alpha\alpha}(\boldsymbol{\alpha}_*)$ stacked vertically; $\boldsymbol{I}_v$ is a $(v \times v)$ identity matrix, and $\otimes$ denotes ***Kronecker product***.

Finally, if $\boldsymbol{h}(\boldsymbol{\alpha}, \boldsymbol{\delta})$ $(v \times 1)$ is a vector-valued function depending on $\boldsymbol{\alpha}$ $(r \times 1)$ and $\boldsymbol{\delta}$ $(s \times 1)$, and we wish to separate explicitly the terms involving $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$, the above extends in the obvious way; for $k = 1$, we have, using obvious notation,

$$\boldsymbol{h}(\boldsymbol{\alpha}, \boldsymbol{\delta}) = \boldsymbol{h}(\boldsymbol{\alpha}_0, \boldsymbol{\delta}_0) + \begin{pmatrix} h_{1\alpha}^T(\boldsymbol{\alpha}_*, \boldsymbol{\delta}_*) \\ \vdots \\ h_{v\alpha}^T(\boldsymbol{\alpha}_*, \boldsymbol{\delta}_*) \end{pmatrix} (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0) + \begin{pmatrix} h_{1\delta}^T(\boldsymbol{\alpha}_*, \boldsymbol{\delta}_*) \\ \vdots \\ h_{v\delta}^T(\boldsymbol{\alpha}_*, \boldsymbol{\delta}_*) \end{pmatrix} (\boldsymbol{\delta} - \boldsymbol{\delta}_0).$$

This can, of course, be written more compactly as

$$\boldsymbol{h}(\boldsymbol{\alpha}, \boldsymbol{\delta}) = \boldsymbol{h}(\boldsymbol{\alpha}_0, \boldsymbol{\delta}_0) + \begin{pmatrix} h_{1\alpha}^T(\boldsymbol{\alpha}_*, \boldsymbol{\delta}_*) & h_{1\delta}^T(\boldsymbol{\alpha}_*, \boldsymbol{\delta}_*) \\ \vdots & \vdots \\ h_{v\alpha}^T(\boldsymbol{\alpha}_*, \boldsymbol{\delta}_*) & h_{v\delta}^T(\boldsymbol{\alpha}_*, \boldsymbol{\delta}_*) \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} - \boldsymbol{\alpha}_0 \\ \boldsymbol{\delta} - \boldsymbol{\delta}_0 \end{pmatrix}.$$

Appendix C: Review of Large Sample Theory

Because large sample theory results are fundamental to modern statistical methods, for which ***exact*** results cannot be derived, we review generically and informally the basics of large sample theory.

In particular, suppose we have an estimator for a ***parameter*** of interest in a statistical model. Recall that an ***estimator*** is a function of random variables/vectors representing data, and an ***estimate*** is the numerical value that results from evaluating the estimator at a particular realization of data.

Ideally, we would like to derive the exact ***sampling distribution*** of the estimator to deduce appropriate ***assessments of uncertainty*** , such as standard errors and confidence intervals, and to develop ***hypothesis testing*** procedures. However, in complex statistical models, this is often not possible analytically. It is thus customary to appeal to ***theory*** as the ***sample size*** approaches infinity and use the theory to ***approximate*** the behavior of the estimator.

The fundamental concepts are:

- ***Consistency***. Does the estimator "estimate the right stuff?" That is, for larger and larger sample sizes, does the estimator "approach" the true value of the parameter in some sense?
- ***Asymptotic distribution.*** Can we approximate the true, unknown sampling distribution of the estimator to use as a basis for inference and gain understanding of precision of estimation?
- ***Asymptotic relative efficiency.*** (There are different definitions of this concept; we consider a standard one.) Can we compare the performance of two or more competing estimators for the same quantity? If both are "consistent," which one is "better" in terms of precision?

We review basic concepts in probability and large sample theory relevant to the above goals.

CONSISTENCY AND ORDER IN PROBABILITY: To evaluate whether or not an estimator "approaches" the "right stuff," we must define precisely what we mean by this. Along with this concept is a convenient notation that summarizes behavior of relevant quantities in this sense.

STOCHASTIC CONVERGENCE: To discuss consistency, we need a basic understanding of ***convergence*** of random variables. The following concepts are usually introduced in a probability course, but often their practical usefulness is not elucidated.

- Estimators are functions of random variables, so that they are ***themselves*** random variables (vectors). Thus, convergence of random variables (vectors) in a probabilistic sense is directly relevant to defining ***consistency***, as we now show.

For this discussion, let Y_n be a ***generic random variable*** (scalar) or vector that depends on some index n. Let Y be another random variable or vector. Let P be a relevant probability measure.

DEFINITION C.1 Almost sure convergence. $Y_n \xrightarrow{\text{a.s}} Y$; i.e., Y_n converges to Y with probability one or almost surely, if

$$P(\lim_{n\to\infty} Y_n = Y) = 1.$$

DEFINITION C.2 Convergence in probability. $Y_n \xrightarrow{p} Y$; i.e., Y_n converges to Y in probability if

$$\lim_{n\to\infty} P(|Y_n - Y| < \delta) = 1 \quad \text{for all } \delta > 0.$$

For random vectors, the definitions extend element by element.

FACTS: The following can be derived from the above definitions.

- $Y_n \xrightarrow{\text{a.s}} Y$ implies that $Y_n \xrightarrow{p} Y$.
- If h is a ***continuous*** function in its argument, then if $Y_n \xrightarrow{\text{a.s}} Y$, it follows that $h(Y_n) \xrightarrow{\text{a.s}} h(Y)$. Similarly, if $Y_n \xrightarrow{p} Y$, then $h(Y_n) \xrightarrow{p} h(Y)$.

We will make routine use of the second fact in the sequel.

Taken alone, the definitions do not seem to be relevant to a study of practical issues in estimation. However, if we identify them with the generic estimation problem, their importance becomes clear.

- n is the ***sample size***.
- Y_n represents an ***estimator*** of some parameter of interest in a ***statistical model*** , η, say. Recall that a statistical model is a class of probability distributions assumed to have generated the data used to form the estimator. Thus, for example, a ***parametric*** statistical model would be a probability distribution depending on a finite-dimensional parameter η. A statistical model is ***correct*** if it includes the true distribution generating the data. In this case, the ***true value*** of η, η_0, say, is then the value of η such that the probability distribution evaluated at η_0 is that truly generating the data.

- The estimator is a ***function*** of n through its dependence on the n (assumed randomly sampled) observations, so we can write $\widehat{\eta}_n$. Ordinarily, we do not include a subscript "n" in standard notation for estimators, but it is important to recognize that they do depend on the sample size. If we view an estimator properly as a ***function*** of sample data, then the sample size serves as an ***index*** for a ***sequence*** of estimators, the functions of sample size n for each n.

- Y represents the thing Y_n, and thus an estimator $\widehat{\eta}_n$, "approaches." In the estimation problem, we hope that $\widehat{\eta}_n$" $\longrightarrow$ "η_0, where, assuming that the statistical model in which η appears is ***correct*** , η_0 is the ***true value*** generating the data.

- Thus, in ***DEFINITIONS*** **C.1** and **C.2** of modes of stochastic convergence, the estimator $\widehat{\eta}_n$ plays the role of Y_n while the value η_0, which in this case is a ***fixed constant*** , plays the role of Y. So in the case of applying these definitions to ***consistency*** of estimators, which we define formally momentarily, the random variable or vector Y is degenerate.

TERMINOLOGY: Special terminology is used to describe how $\widehat{\eta}_n$ approaches η_0.

- ***Strong consistency:*** $\widehat{\eta}_n \xrightarrow{\text{a.s}} \eta_0$
- ***Weak consistency:*** $\widehat{\eta}_n \xrightarrow{p} \eta_0$.

WHAT DOES THIS MEAN?

- Both types of consistency state that the estimator ***approaches*** the quantity to be estimated in a probabilistic sense.

- From the definition of almost sure convergence, the interpretation of ***strong*** consistency is that, if the sample size n is sufficiently large, the probability that $\widehat{\eta}_n$ will assume values outside an arbitrarily small "neighborhood" of η_0 is zero. This follows from the fact that the limit appears inside the probability statement in ***DEFINITION*** **C.1**. Recall that, for a ***deterministic sequence*** , a_n has ***limit*** a if, for each $\epsilon > 0$, there is a value n_ϵ such that

$$|a_n - a| < \epsilon \quad \text{for all } n > n_\epsilon.$$

  This can be applied to the probability.

- From the definition of convergence in probability, the interpretation of ***weak*** consistency is that, for n large, the probability is small that $\widehat{\eta}_n$ assumes a value outside an arbitrarily small neighborhood of η_0. This again follows from the definition of a limit; the difference between 1 and the probability that $\widehat{\eta}_n$ is within δ of η_0 is less that ϵ if n is greater than some n_ϵ.

- The names seem to imply that strong is ***better than*** weak.

PRACTICAL DIFFERENCE: Here is a popular argument in favor of strong consistency. Suppose that one were to collect data ***sequentially***, and, periodically, re-estimate η by $\widehat{\eta}_n$, where n is the number of observations collected so far. Thus, with this scheme, $n \to \infty$. A ***sequence of estimators*** indexed by n, $\widehat{\eta}_n$, is thus generated.

- One would like to be assured that a value of n can be reached at which the current estimate is ***sufficiently close*** to the true value and will never "wander away" again after further data collection.
- Strong consistency ensures this – for n large enough, the probability that $\widehat{\eta}_n$ will stay arbitrarily close to η_0 is 1.
- Weak consistency does not – it states state that the probability that $\widehat{\eta}_n$ will wander away again is "small."

This argument seems to suggest that we should ***always prefer*** strong consistency. However, statisticians are usually willing to settle for weak consistency; most are content that an estimator is "good" if we can make the probability of $\widehat{\eta}_n$ being "close to" the true value "large" (rather than equal to 1).

The unqualified term ***consistency"*** in most statistical literature almost always refers to ***weak*** consistency. In this course, we are satisfied with weak consistency.

TECHNICAL NOTES:

- We have presented consistency under the conditions that there is a ***statistical model*** involving a parameter η, and this model is ***correctly specified***. We can thus think of this model as indexed by values of η, and there is a true value of η, η_0, that is responsible for the data we have seen. Interest in statistical problems is of course in estimating this true value under these conditions. Usually, the term ***consistency*** is meant to imply this situation; i.e., that the true value of some parameter generating the data is correctly identified in the probabilistic sense.
- It is not always the case that the model is correctly specified (although we may not be aware of this). In this situation, we may still forge ahead as if it were correct, and conceive of it as being ***indexed*** by a parameter η, and we can deduce estimators for η.

However, there may not be a "true value" for η, as the model does not coincide with the true ***data generating mechanism***. The estimator $\widehat{\eta}_n$ can still be defined for each n (as a function of the sample data), and it may still converge in probability (or almost surely) to some quantity, η_*, say. Such a $\widehat{\eta}_n$ is sometimes referred to as being "consistent" for the value η_*, which can be confusing.

- Alternatively, even in the context of a correctly specified model, it is possible to define estimators $\widehat{\eta}_n$ that ***do not*** "estimate the right stuff;" i.e., that are ***not consistent*** in that $\widehat{\eta}_n \xrightarrow{p} \eta_*$, where $\eta_* \neq \eta_0$. In this case, $\widehat{\eta}_n$ is said to be ***inconsistent***.

- In general, the goal is (a) to identify a statistical model that is ***correct*** (i.e., contains the true distribution generating the data) and (b) identify a ***consistent estimator*** for the true value of a parameter η indexing this model. If (a) is not carried out, (b) may not be possible.

- In most studies of properties of estimators, that the model is correctly specified is taken as a ***starting point***. We take this perspective initially; however, we will also investigate what happens when certain components of models are ***not correctly specified***.

ORDER IN PROBABILITY: This notation can appear confusing initially, but once mastered, is useful for streamlining presentation of large sample results. Again, let $\{Y_n\}$ denote a generic sequence of random variables/vectors indexed by n.

DEFINITION C.3 "Big" O_p. Y_n is ***at most of order in probability*** n^k if, for all $\epsilon > 0$, there exist constants $n_\epsilon, M_\epsilon > 0$ such that

$$P(n^{-k}\|Y_n\| < M_\epsilon) > 1 - \epsilon$$

for all $n > n_\epsilon$. Here, $\|\cdot\|$ is some ***norm*** measuring magnitude in the case of vector Y_n; if Y_n is scalar, then this is just absolute value.

The notation is $Y_n = O_p(n^k)$

- The definition says that the magnitude of $n^{-k}Y_n$ ***stays bounded with high probability*** if n is large enough. The cases of most interest to us are when k is ***nonpositive***.

- For example, if $k = -1/2$, then $n^{1/2}Y_n$ stays bounded as n gets large with high probability. In particular, with high probability, Y_n is bounded by $M_\epsilon n^{-1/2}$. This means that Y_n itself is getting "small" as n gets large. A practical interpretation is that Y_n ***behaves like*** $n^{-1/2}$ with high probability for n large enough; i.e., becomes negligible in the same "way" $n^{-1/2}$ does.

- In fact, as $n^{-1/2} \to 0$ as $n \to \infty$, this says that Y_n itself "approaches" (***converges in probability*** to) zero at the same ***rate*** as $n^{-1/2}$.
- If $k = 0$, then $Y_n = O_p(1)$. From Definition **C.3**, this says that Y_n remains bounded by the constant M_ϵ for n large with high probability. In this case, Y_n is said to be ***bounded in probability***.
- Practically speaking, this says that, as n gets large, Y_n does not become negligible, nor does it "blow up." Instead, it is "nicely behaved."

DEFINITION C.4 "Little" o_p. Y_n is said to be ***of smaller order in probability*** than n^k if

$$n^{-k} Y_n \xrightarrow{p} 0 \quad \text{as } n \to \infty.$$

The notation is $Y_n = o_p(n^k)$.

- The case of most interest to us is $k = 0$. It is ***shorthand*** for saying that Y_n converges in probability to zero; that is, for n sufficiently large, Y_n stays arbitrarily "close" to zero with high probability. If we can show an expression is $o_p(1)$, it can be "ignored" as "negligible."
- More generally, the case $k \leq 0$ is the most interesting. For example, if $k = -1/2$, $Y_n = o_p(n^{-1/2})$, then $n^{1/2} Y_n \xrightarrow{p} 0$. This shorthand notation says that we can multiply Y_n by a factor that acts like $n^{1/2}$, and the entire product will still be ***negligible***. Thus, this notation is a useful way of expressing ***how quickly*** $Y_n \xrightarrow{p} 0$; i.e. if $Y_n = o_p(n^{-1/2})$, then it is "faster" than $n^{-1/2}$.

FACTS: The following facts can be deduced from the definitions.

- $Y_n = O_p(n^{-\delta})$ for $\delta > 0$ implies that $Y_n = o_p(1)$. This is intuitively obvious – if Y_n ***acts like*** $n^{-\delta}$ when n is large with high probability, then it must "go to zero."

  Stating that $Y_n = O_p(n^{-\delta})$ is ***more informative*** than just saying $Y_n = o_p(1)$; the former not only tells us that Y_n becomes negligible with high probability, but at what ***rate***.
- If $Y_n = O_p(n^k)$ and $X_n = O_p(n^j)$, then $X_n Y_n = O_p(n^{k+j})$. The same holds true if O_p is replaced by o_p, and for combinations of O_p and o_p. Thus, if we know the order in probability of each of two quantities, we can deduce how their ***product*** behaves.
- A useful ***special case*** is when $Y_n = O_p(1)$, referred to as ***bounded in probability*** , and $X_n = o_p(1)$ ($X_n \xrightarrow{p} 0$). Then $X_n Y_n = o_p(1)$. Intuitively, this makes sense; Y_n is "well behaved," neither getting small nor blowing up and is multiplied by something that is getting small. Thus, the product would be expected to get small. Of course, this means that $X_n Y_n \xrightarrow{p} 0$.

It is important to keep in mind that $Y_n = O_p(n^{-1/2})$, say, only means that the magnitude of $n^{1/2}Y_n$ is ***bounded*** by ***some constant***. That constant could be ***huge***, so that n must be very, very large for Y_n to become negligible for practical purposes. This explains in part why, sometimes, large sample approximations ***do not seem relevant*** in practice.

We now turn to concepts useful in deducing approximations to ***sampling distributions of estimators***. Continue to regard Y_n and Y as generic random variables/vectors.

DEFINITION C.5 Convergence in distribution. Suppose Y_n has ***cumulative distribution function*** (cdf) and that Y has cdf F. Y_n is said to ***converge in distribution*** (or ***law***) to Y if and only if, for each continuity point of F,

$$\lim_{n\to\infty} F_n(y) = F(y).$$

The standard notation is $Y_n \xrightarrow{D} Y$ or $Y_n \xrightarrow{\mathcal{L}} Y$; we use the latter.

PRACTICAL INTERPRETATION: If $Y_n \xrightarrow{\mathcal{L}} Y$, this implies roughly that, for large n, ***except*** at a few points, the ***distribution*** of Y_n (and thus probabilities associated with Y_n) is ***the same*** as that of Y. Thus, if we are interested in probability and distributional statements about Y_n, we can ***approximate*** these with statements about Y.

In the context of estimation, if we are interested in approximating the ***sampling distribution*** of an estimator, we are interested in the ***convergence in distribution*** of the estimator (or some function thereof).

FACTS: The following can be deduced from Definition **C.5** and previous definitions.

- If $Y_n \xrightarrow{p} Y$, then $Y_n \xrightarrow{\mathcal{L}} Y$. This says that if, for large n, the ***probability*** that Y_n differs from Y is small, then we would expect the ***probabilities*** with which they take on values to be "close," and thus expect them to have distributions that are "close."
- However, $Y_n \xrightarrow{\mathcal{L}} Y$ ***DOES NOT*** imply $Y_n \xrightarrow{p} Y$ in general. For example, suppose that Y_n and Y have the ***same*** distribution for each n, but Y_n and Y are ***independent*** for each n. Then a realization of Y_n is ***totally unrelated*** to a realization of Y!
- $Y_n \xrightarrow{\mathcal{L}} y$, where y is a ***constant***, ***does*** imply $Y_n \xrightarrow{p} y$. Intuitively, because the distribution of Y_n ***collapses*** to a single point, a realization of Y_n must ***also*** approach that point.

Of course, if $Y_n \xrightarrow{\mathcal{L}} y$, a constant, then the distribution is ***degenerate***, which is not particularly interesting if one seeks to deduce a ***sampling distribution*** to be used for constructing confidence intervals and hypothesis tests.

SAMPLING DISTRIBUTION OF AN ESTIMATOR: Return now to our situation of interest, where $\widehat{\eta}_n$ is an ***estimator*** for a parameter η in a (correct) ***statistical model*** with true value η_0. Showing that $\widehat{\eta}_n \xrightarrow{p} \eta_0$ thus implies that $\widehat{\eta}_n \xrightarrow{\mathcal{L}} \eta_0$. However, this knowledge that the distribution of $\widehat{\eta}_n$ collapses to the single point η_0 is ***not very useful*** for the usual inferential goals described above. In particular, this result does not even give information on ***precision of estimation***.

To gain insight and to provide a basis for the standard inferential objectives, we must pursue a ***more refined assessment*** of large sample behavior. Instead of considering the properties of $\widehat{\eta}_n$ itself, we instead consider a suitable function of $\widehat{\eta}_n$ whose properties are "more interesting" and relevant. For most estimators solving ***estimating equations***, a standard approach to deriving an approximate sampling distribution that is more useful applies.

DEFINITION C.6 Asymptotic normality. We present this definition in the scalar case; the vector case is similar. Classically speaking, a random variable Y_n is said to be ***asymptotically normal*** if we can find sequences $\{a_n\}$ and $\{c_n\}$ such that

$$c_n(Y_n - a_n) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

By this notation, we mean that the right hand side of this expression is a ***standard normal random variable***.

DEFINITION* C.6** implies that, although the limit distribution of Y_n itself may be uninteresting, if we ***center and scale Y_n appropriately, this "standardized" version of Y_n has an ***interesting limit distribution***.

- In particular, a_n is called the ***asymptotic mean*** and c_n is called the ***asymptotic variance***, and ***DEFINITION* C.6** can be interpreted to mean that, approximately for large n

$$Y_n \dot{\sim} \mathcal{N}(a_n, c_n^{-2}).$$

  The usefulness of this result for approximating a sampling distribution is thus evident.

How is this applied in estimation situations of interest to us? As we will see, because many estimators of interest to us are ***not available in a closed form***, things are not as simple as immediately identifying $\widehat{\eta}_n$ with Y_n and then determining appropriate centering and scaling constants. Instead, what is done is to find an approximation to an ***appropriate centered and scaled*** version of $\widehat{\eta}_n$ by applying a ***Taylor series*** to the ***estimating equation*** that defines $\widehat{\eta}_n$ implicitly. This approximation then forms the basis for deducing behavior like that in Definition **C.6**.

Some important tools for deducing this behavior are the following. After we state these important results, we sketch how they are used in this way.

There are numerous versions of ***central limit theorems*** that characterize the ***convergence in distribution*** of appropriately standardized ***sums of independent random variables/vectors***. These can be extended to random vectors in a number of ways to allow generalization of univariate results to multivariate ones. We do not discuss the technicalities behind this. Instead, we state a particular ***multivariate central limit theorem*** that is useful for our purposes.

MULTIVARIATE CENTRAL LIMIT THEOREM: Let $\boldsymbol{Z}_i$ be independent random vectors with $E(\boldsymbol{Z}_i) = \boldsymbol{\mu}_i$ and $\text{var}(\boldsymbol{Z}_i) = \boldsymbol{\Sigma}_i$, $i = 1, \ldots, n$, such that

$$\lim_{n\to\infty} n^{-1}(\boldsymbol{\Sigma}_1 + \cdots + \boldsymbol{\Sigma}_n) = \lim_{n\to\infty} n^{-1} \sum_{j=1}^{n} \boldsymbol{\Sigma}_i = \boldsymbol{\Sigma},$$

say, letting F_i be the cdf of $\boldsymbol{Z}_i$,

$$n^{-1} \sum_{i=1}^{n} \int_{\|Z_i - \mu_i\| \geq \epsilon n^{1/2}} \|\boldsymbol{z} - \boldsymbol{\mu}_i\|^2 \, dF_i(\boldsymbol{z}) \longrightarrow \mathbf{0} \quad \text{as } n \to \infty. \tag{C.1}$$

Then

$$n^{-1/2} \sum_{i=1}^{n} (\boldsymbol{Z}_i - \boldsymbol{\mu}_i) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}).$$

The ***Lindeberg condition*** (C.1) effectively restricts the ***tail behavior*** of $\boldsymbol{Z}_i$ and does not appear particularly intuitive. It turns out that (C.1) may be shown to hold if the ***third moments*** of $\boldsymbol{Z}_i$ exist and are finite (the so-called ***Lyapunov condition***). We generally assume that higher moments of the response ***exist*** and are ***finite***, so that the moments of relevant quantities to which this theorem will be applied can be assumed to exist and be finite. Thus, condition (C.1) is assumed without comment when we apply the multivariate central limit theorem.

Additional key results of which we will make heavy use are the following.

SLUTSKY'S THEOREM: Suppose that $Y_n \xrightarrow{\mathcal{L}} Y$ and $V_n \xrightarrow{p} c$, where c is a constant. Then

$$Y_n + V_n \xrightarrow{\mathcal{L}} Y + c, \quad Y_n V_n \xrightarrow{\mathcal{L}} cY, \quad Y_n / V_n \xrightarrow{\mathcal{L}} Y/c,$$

where in the last expression $c \neq 0$ is required. These ***extend*** readily to random vectors: If $\boldsymbol{Y}_n \xrightarrow{\mathcal{L}} \boldsymbol{Y}$ and $\boldsymbol{\Sigma}_n \xrightarrow{p} \boldsymbol{C}$, where $\boldsymbol{\Sigma}_n$ and $\boldsymbol{C}$ are matrices,

$$\boldsymbol{Y}_n + \boldsymbol{\Sigma}_n \xrightarrow{\mathcal{L}} \boldsymbol{Y} + \boldsymbol{C}, \quad \boldsymbol{\Sigma}_n \boldsymbol{Y}_n \xrightarrow{\mathcal{L}} \boldsymbol{C}\boldsymbol{Y}, \quad \boldsymbol{\Sigma}_n^{-1} \boldsymbol{Y}_n \xrightarrow{\mathcal{L}} \boldsymbol{C}^{-1}\boldsymbol{Y}.$$

Thus, when $E(\boldsymbol{Y}) = \boldsymbol{\mu}$ and $\text{var}(\boldsymbol{Y}) = \boldsymbol{\Sigma}$, say, then we have that $\boldsymbol{\Sigma}_n \boldsymbol{Y}_n \xrightarrow{\mathcal{L}}$ to random vector with mean $\boldsymbol{C}\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^T$.

Slutsky's theorem may be invoked repeatedly so that if $\boldsymbol{U}_n \xrightarrow{p} \boldsymbol{D}$ as well, then

$$\boldsymbol{\Sigma}_n \boldsymbol{Y}_n + \boldsymbol{U}_n \xrightarrow{\mathcal{L}} \boldsymbol{C}\boldsymbol{Y} + \boldsymbol{D}.$$

WEAK LAW OF LARGE NUMBERS: We state this in the scalar case, but it ***extends straightforwardly*** to vectors. Suppose Z_i are independent (or uncorrelated) random variables and a_i are constants. Then, if $n^{-2} \sum_{i=1}^n \text{var}(Z_i) a_i^2 \to 0$,

$$n^{-1} \sum_{i=1}^n a_i Z_i - n^{-1} \sum_{i=1}^n a_i E(Z_i) \xrightarrow{p} 0.$$

- The condition $n^{-2} \sum_{i=1}^n \text{var}(Z_i) a_i^2 \to 0$ is satisfied if $n^{-1} \sum_{i=1}^n \text{var}(Z_i) a_i^2 \to c$ for some constant c, which is often reasonable (and similar to the requirement for the central limit theorem).
- If we furthermore know that $n^{-1} \sum_{i=1}^n a_i E(Z_i) \to d$, say, then we can conclude that $n^{-1} \sum_{i=1}^n a_i Z_i \xrightarrow{p} d$, as

$$n^{-1} \sum_{i=1}^n a_i Z_i - d = \{n^{-1} \sum_{i=1}^n a_i Z_i - n^{-1} \sum_{i=1}^n a_i E(Z_i)\} + \{n^{-1} \sum_{i=1}^n a_i E(Z_i) - d\} \xrightarrow{p} 0.$$

We are now in a position to describe how all of this is used in more detail. We drop the n subscript on our generic estimator and treat it and the parameter of interest as ***vectors***, writing $\boldsymbol{\eta}$ and $\widehat{\boldsymbol{\eta}}$.

For ***estimating equations*** for a parameter $\boldsymbol{\eta}$ of interest with solution $\widehat{\boldsymbol{\eta}}$, we can deduce using ***Taylor series*** and some additional conditions that

$$n^{1/2}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) = \boldsymbol{A}_n^{-1} \boldsymbol{C}_n + o_p(1),$$

where $\boldsymbol{C}_n = n^{-1/2} \sum_{i=1}^n$ (function of data), $\boldsymbol{A}_n = n^{-1} \sum_{i=1}^n$ (function of data), and $o_p(1)$ represents terms that converge in probability to zero.

We then

- Apply the ***central limit theorem*** to $\boldsymbol{C}_n$ to show that it ***converges in distribution*** to a normal random vector.
- Apply the ***weak law of large numbers*** to $\boldsymbol{A}_n$ to show that it ***converges in probability*** to a constant matrix.
- Apply ***Slutsky's theorem*** to $\boldsymbol{A}_n^{-1}\boldsymbol{C}_n$ to conclude that $n^{1/2}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)$ ***converges in distribution*** to a normal random vector with ***mean zero*** and some ***covariance matrix*** $\boldsymbol{\Sigma}$; i.e.,

$$n^{1/2}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{C.2}$$

  say. This is often interpreted as $\widehat{\boldsymbol{\eta}} \dot{\sim} \mathcal{N}(\boldsymbol{\eta}_0, n^{-1}\boldsymbol{\Sigma})$; i.e., $\widehat{\boldsymbol{\eta}}$ is asymptotically normal with mean $\boldsymbol{\eta}_0$ and covariance matrix $n^{-1}\boldsymbol{\Sigma}$.

From these steps, we can then deduce an approximate (normal) ***sampling distribution*** for $\widehat{\boldsymbol{\eta}}$.

Suppose that we have ***two competing estimators*** for $\boldsymbol{\eta}$ ($k \times 1$), $\widehat{\boldsymbol{\eta}}^{(1)}$ and $\widehat{\boldsymbol{\eta}}^{(2)}$, say, so that we have

$$n^{1/2}(\widehat{\boldsymbol{\eta}}^{(1)} - \boldsymbol{\eta}_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_1) \quad \text{and} \quad n^{1/2}(\widehat{\boldsymbol{\eta}}^{(2)} - \boldsymbol{\eta}_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_2)$$

for some matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$.

- Both $\widehat{\boldsymbol{\eta}}^{(1)}$ and $\widehat{\boldsymbol{\eta}}^{(2)}$ are ***consistent***. It is a general fact that, if a random vector ***converges in distribution***, then it is ***bounded in probability***. Thus, we have

$$n^{1/2}(\widehat{\boldsymbol{\eta}}^{(\ell)} - \boldsymbol{\eta}_0) = O_p(1)$$

  for $\ell = 1, 2$. This may be expressed equivalently as

$$(\widehat{\boldsymbol{\eta}}^{(\ell)} - \boldsymbol{\eta}_0) = O_p(n^{-1/2}).$$

- Thus, both estimators "estimate the right stuff" and approach it at the same rate. On this basis, then, they are entirely ***comparable***.
- As this ***does not distinguish*** the two estimators from one another, consider their ***precision***. In finite sample, exact theory, the estimator that is ***more precise*** is to be preferred. Here, we approximate the covariance matrices of the estimators by $n^{-1}\boldsymbol{\Sigma}_1$ and $n^{-1}\boldsymbol{\Sigma}_2$, respectively. This suggests comparing $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$.

- In the case $k = 1$, so that Σ_1 and Σ_2 are scalar variances, this suggests preferring the estimator with the smaller variance. That is, prefer $\hat{\eta}_2$ to $\hat{\eta}_1$ if $\Sigma_2 < \Sigma_1$. If $\Sigma_2 = \Sigma_1$, then the two estimators are of equal precision.

DEFINITION C.7 Asymptotic relative efficiency. For scalars, the ***asymptotic relative efficiency*** of $\hat{\eta}_1$ to $\hat{\eta}_2$ is defined as

$$ARE = \Sigma_2/\Sigma_1.$$

With this definition, if $ARE = 1$, the estimators are equally precise. If $ARE < 1$, then $\hat{\eta}_1$ is ***inefficient*** relative to $\hat{\eta}_2$, and if $ARE > 1$, then $\hat{\eta}_1$ offers a gain in efficiency relative to $\hat{\eta}_2$.

Often, one constructs the ratio with the potentially "better" estimator's variance in the ***numerator*** , so that $ARE < 1$ is "good" for showing another estimator is ***inefficient relative to*** it. However, many texts and authors do this in the reverse, so that larger-than-one values are preferred.

The ***extension*** of the definition to $k > 1$ is that $\widehat{\boldsymbol{\eta}}^{(2)}$ is preferable to $\widehat{\boldsymbol{\eta}}^{(1)}$ if the covariance matrix $\boldsymbol{\Sigma}_2$ is "smaller" than $\boldsymbol{\Sigma}_1$ in some sense. To formalize this, if $(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)$ is ***nonnegative definite*** , then, for all $(k \times 1)$ $\boldsymbol{\lambda}$,

$$\boldsymbol{\lambda}^T\boldsymbol{\Sigma}_2\boldsymbol{\lambda} \leq \boldsymbol{\lambda}^T\boldsymbol{\Sigma}_1\boldsymbol{\lambda}.$$

By choosing $\boldsymbol{\lambda}$ in turn to be the vector with a 1 in one position and zeroes elsewhere, we see that this implies that the variances on the diagonal of $\boldsymbol{\Sigma}_2$ must be smaller than those on the diagonal of $\boldsymbol{\Sigma}_1$, so that the (approximate) variance of each component of $\widehat{\boldsymbol{\eta}}^{(2)}$ is ***smaller*** than that of $\widehat{\boldsymbol{\eta}}^{(1)}$.

If $(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)$ is nonnegative definite, it follows that

$$|\boldsymbol{\Sigma}_2| \leq |\boldsymbol{\Sigma}_1|.$$

Thus, the asymptotic relative efficiency of $\widehat{\boldsymbol{\eta}}^{(1)}$ to $\widehat{\boldsymbol{\eta}}^{(2)}$ is generally defined for $k > 1$ as

$$ARE = \{|\boldsymbol{\Sigma}_2|/|\boldsymbol{\Sigma}_1|\}^{1/k}.$$

This comparison is sometimes simplified in that it turns out that $\boldsymbol{\Sigma}_1 = \alpha_1\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_2 = \alpha_2\boldsymbol{\Sigma}$ for some scalars α_ℓ, $\ell = 1, 2$, and common matrix $\boldsymbol{\Sigma}$. In this case, ARE reduces to α_2/α_1, as $|\alpha\boldsymbol{\Sigma}| = \alpha^k|\boldsymbol{\Sigma}|$.

It is often argued that that one estimator is efficient relative to another by examining the difference $\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2$. However, simply noting that this difference is nonnegative definite does not give insight into ***how much*** more efficient. The calculation of ARE quantifies "how much better." In complex statistical models, ARE usually depends on the design, parameter values, and functions involved, so that a "global" statement of relative efficiency can not be made.

Appendix D: Brief Review of Monte Carlo Simulation

The following is a brief review of the basics of a Monte Carlo simulation study.

When analytical arguments are intractable, a popular way to learn about the finite-sample properties of estimators is by Monte Carlo simulation. The objective of a simulation is to approximate the sampling distribution of an estimator by generating (via random deviate generation routines) some large number S independent data sets from a known situation and computing the estimator for each data set. The sample mean of the estimates over all S data sets is an estimate of the mean of the sampling distribution of the estimator; similarly, the standard deviation of the estimates over the S data sets is an estimate of the standard deviation of the sampling distribution (how good these quantities are at capturing the true features of the sampling distribution obviously depends on the size of S).

To carry out a simulation to evaluate to properties of several competing estimators for some parameter β ($p \times 1$) in a statistical model and for how well large sample approximations to the true sampling distribution work, the following are basic steps in a simulation.

- Generate S data sets from a scenario of interest. This scenario represents the true statistical model generating the data, with true value of β equal to some β_0.
- For each data set, estimate β using each of the competing methods under consideration.
- Also obtain standard errors for the components of each estimator for β using the accompanying large sample theory approximation to the sampling distribution for that estimator.

Let $\beta_{k,0}$ be the kth component of the true value β_0, $k = 1, \ldots, p$. Let $\widehat{\beta}$ be one of the estimators, and let $\widehat{\beta}_k$ be its k component, $k = 1, \ldots, p$. Let $\widehat{\beta}_s$ be the estimate obtained from the sth data set, $s = 1, \ldots, S$, and let $\widehat{\beta}_{k,s}$ be its kth element, $k = 1, \ldots, p$.

Then, for each estimator, do the following.

(i) If the estimators are consistent, we would hope that they would be approximately unbiased in finite samples. Thus, we would hope that the mean of the sampling distribution is close to the true value of β, β_0. with only minimal bias. To assess this based on the S observations from the sampling distribution, calculate the *Monte Carlo bias* for each component of an estimator $\widehat{\beta}$,

defined for the kth component as

$$S^{-1}\sum_{s=1}^{S}\widehat{\beta}_{k,s}-\beta_{0,k}.$$

It is standard to report this "raw" Monte Carlo bias. It is also standard to report this bias relative to to the true value (so report for each k

$$\frac{S^{-1}\sum_{s=1}^{S}\widehat{\beta}_{k,s}-\beta_{0,k}}{\beta_{0,k}},$$

which can of course be problematic when the true value is very close to 0) and relative to the Monte Carlo standard deviation (see (iii) below), so that the size of the bias relative to the variation in the estimator can be assessed.

(ii) To compare the precision of two competing estimators based on the S estimates of each, we could compare their sample variances, thus mimicking the idea of asymptotic relative efficiency. However, because the estimators may exhibit some finite sample *bias* for finite m in our case, it is standard instead routine to take this into account and compute the *Monte Carlo mean square error* (MSE) for each estimator. The estimated MSE based on the S estimates $\widehat{\beta}_s$ for the kth component is defined as

$$S^{-1}\sum_{s=1}^{S}(\widehat{\beta}_{k,s}-\beta_{0,k})^2 = S^{-1}\sum_{s=1}^{S}(\widehat{\beta}_{k,s}-\bar{\beta}_k)^2+(\bar{\beta}_k-\beta_{0,k})^2,$$

where $\bar{\beta}_k$ is the sample average of the $\widehat{\beta}_{k,s}$. Note that MSE may thus be interpreted as sample variance over the S estimates plus observed bias, squared.

The ratio of estimated MSE values may be used as a measure of relative precision, similar to asymptotic relative efficiency. One ordinarily calculates MSE for each component k and then forms the ratio for each k. It is customary to put the MSE for the estimator that is thought to be *more efficient* in the numerator, so that a MSE ratio less than 1 reflects the relative inefficiency of the estimator whose MSE is in the denominator. It can of course be done either way as long as the user defines the MSE ratio so that it can be interpreted appropriately.

(iii) To assess how well the estimated standard errors approximate the true sampling variation, one can compare the sample standard deviation of each component of the S estimates $\widehat{\beta}$, that is, the *Monte Carlo standard deviation*, to the average of the estimated standard errors for that component found using the large sample theory. If the theory is relevant, we would expect the sample standard deviation, an approximation to the true sampling variation, and the average of estimated standard errors, to be "close." Sometimes, the ratio of the two is formed for each component k to get a sense of this.

(iv) To assess further how well the approximate large sample sampling distribution approximates the true sampling distribution, for each estimator, one might calculate for each of the S data sets 95% Wald confidence intervals (using the usual critical value from the standard normal distribution of 1.96) for the true values of each component of β, and record the proportion of times that the intervals contain the true values. These proportions are *Monte Carlo coverage probabilities* – if the Wald intervals are reliable, we would expect them to be close to the nominal coverage probability of 0.95. If the Monte Carlo values are not close to 0.95, then using the large-sample approximation may be unreliable.

Appendix E: SAS PROC MIXED Syntax

We summarize the basic syntax of SAS `proc mixed`. The usual linear mixed effects model is

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i,$$

$$E(\boldsymbol{b}_i|\boldsymbol{x}_i) = \boldsymbol{0}, \quad \text{var}(\boldsymbol{b}_i|\boldsymbol{x}_i) = \boldsymbol{D}, \quad E(\boldsymbol{e}_i|\boldsymbol{x}_i, \boldsymbol{b}_i) = E(\boldsymbol{e}_i|\boldsymbol{x}_i) = \boldsymbol{0}, \quad \text{var}(\boldsymbol{e}_i|\boldsymbol{x}_i, \boldsymbol{b}_i) = \text{var}(\boldsymbol{e}_i|\boldsymbol{x}_i) = \boldsymbol{R}_i$$

(`proc mixed` does not directly accommodate specifications of $\boldsymbol{R}_i$ that depend on $\boldsymbol{b}_i$). The model is usually written in software documentation in a streamlined form by "stacking" the contributions from each individual. Define

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_m \end{pmatrix}, \quad \boldsymbol{e} = \begin{pmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \vdots \\ \boldsymbol{e}_m \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_m \end{pmatrix}, \quad \boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{R}_m \end{pmatrix},$$

$$\boldsymbol{b} = \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \vdots \\ \boldsymbol{b}_m \end{pmatrix}, \quad \boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{Z}_m \end{pmatrix}, \quad \widetilde{\boldsymbol{D}} = \begin{pmatrix} \boldsymbol{D} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{D} \end{pmatrix}.$$

Here, $\widetilde{\boldsymbol{D}}$ has been displayed in the case where $\text{var}(\boldsymbol{b}_i|\boldsymbol{x}_i) = \boldsymbol{D}$ for all individuals (so is independent of $\boldsymbol{x}_i$), but can be modified if this is relaxed, as in the dental study with the girls and boys having different matrices $\boldsymbol{D}_G$ and $\boldsymbol{D}_B$ (so depending on $\boldsymbol{a}_i$). The model can be written consisely with these definitions as

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{e}, \quad E(\boldsymbol{Y}|\tilde{\boldsymbol{x}}) = \boldsymbol{X}\beta, \quad \text{var}(\boldsymbol{Y}|\tilde{\boldsymbol{x}}) = \boldsymbol{V} = \boldsymbol{Z}\widetilde{\boldsymbol{D}}\boldsymbol{Z}^T + \boldsymbol{R}. \tag{B.1}$$

The SAS documentation refers to $\widetilde{\boldsymbol{D}}$ as $\boldsymbol{G}$.

The syntax for `proc mixed` is geared to the ***subject-specific*** linear mixed effects model; however, the procedure can also be used to fit ***population-averaged*** linear models of the form

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \epsilon_i$$

or in "stacked" form

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon, \quad E(\boldsymbol{Y}|\tilde{\boldsymbol{x}}) = \boldsymbol{X}\beta, \quad \text{var}(\epsilon|\tilde{\boldsymbol{x}}) = \text{var}(\boldsymbol{Y}|\tilde{\boldsymbol{x}}) = \boldsymbol{V},$$

where $\boldsymbol{V}$ does not have a specific structure; the structure of $\boldsymbol{V}$ is specified fully by the analyst and is not induced by the model.

For either type of model, the form of the ***population mean*** is either induced (SS model) or is specified explicitly (PA model). The `model` statement is of course the mechanism by which the analyst specifies the form the population mean, which is $\boldsymbol{X}_i\beta$ for i or $\boldsymbol{X}\beta$ for all individuals, stacked, in the usual SAS way.

In the context of a PA model, we have used the `repeated` statement to specify the overall covariance matrix $\boldsymbol{V}$. For the SS linear mixed effects model, the `repeated` statement is used to specify the form of the ***within-individual*** covariance model $\boldsymbol{R}_i$ or, equivalently, $\boldsymbol{R}$ above. For this model the `random` statement is used to specify the assumption on var($\boldsymbol{b}_i|\boldsymbol{x}_i$) ($\widetilde{\boldsymbol{D}}$).

Here is a summary of the basic form of a call to `proc mixed`.

```
proc mixed data=dataset method= (ML,REML);
class  classification variables;
model response =  columns of X / solution;
random columns of Z / type=  subject=  group=   ;
repeated  classification variable for time / type=  subject=  group=   ;
run;
```

`proc mixed` statement

- `method=REML` is the default; no `method=` required in this case

`model` statement

- ***columns of X*** are variables (`class` or continuous) corresponding to variables associated with fixed effects β
- Intercept is assumed unless `noint` option after slash
- `solution` is an option and must be invoked to get the estimates of β

`random` statement

- Describes the matrix $\widetilde{\boldsymbol{D}} = \text{var}(\boldsymbol{b}|\tilde{\boldsymbol{x}})$ (i.e. the matrices $\text{var}(\boldsymbol{b}_i|\boldsymbol{x}_i)$ making up the blocks of $\widetilde{\boldsymbol{D}}$
- ***columns of Z*** are variables (`class` or continuous), i.e. variables associated with random effects ***b***
- `subject=` tells `mixed` what `class` variable denotes the grouping determining the ***individuals***
- `type=` allows choice of matrix (e.g. `un`, unstructured)
- `group=` allows ***D*** to be different according to this `class` variable (e.g. dental study, boys, girls)

`repeated` statement

- Describes the matrix $\boldsymbol{R} = \text{var}(\boldsymbol{e}|\tilde{\boldsymbol{x}})$ (i.e. the matrices $\boldsymbol{R}_i = \text{var}(\boldsymbol{e}_i|\boldsymbol{x}_i)$)
- If $\text{var}(\boldsymbol{e}_i\boldsymbol{x}_i) = \sigma^2\boldsymbol{I}_{n_i}$ is the same for all i `repeated` statement is ***NOT*** needed
- `subject=` tells `mixed` what `class` variable denotes the grouping determining the individuals
- `type=` allows choice other than diagonal (e.g. `ar(1), cs, etc.`)
- The optional classification variable before the slash is for situations with unbalanced data and nondiagonal `type` so that observations can be correctly attributed to the times at which they were taken
- `group=` allows $\boldsymbol{R}_i$ to be different depending on group membership (e.g. dental study, $\text{var}(\boldsymbol{e}_i|\boldsymbol{x}_i) = \sigma^2_G\boldsymbol{I}_{n_i}$ girls, $\text{var}(\boldsymbol{e}_i|\boldsymbol{x}_i) = \sigma^2_B\boldsymbol{I}_{n_i}$ boys)

The foregoing syntax makes clear that, to implement a linear PA model using `proc mixed` with the `repeated` statement, we simply make a correspondence between this model and the model (B.1) with ***no*** random effects ***b***, From purely ***operational*** point of view (but ***not*** an ***interpretation*** point of view), the models have the same structure – a mean plus a deviation with components of length n_i, each of which has a covariance matrix. Thus, purely to specify these covariance matrices for the PA model, the `repeated` statement can be used.

See the SAS documentation for `proc mixed` for much more detail on the use of these statements and available options.

Appendix F: Writing a Data Analysis Report

BACKGROUND: In the 21st century, "team science," in which individuals possessing different disciplinary expertise integrate their skills to formulate and address subject matter questions, is the primary mechanism by which the advance of knowledge takes place. Statisticians are key members of such multidisciplinary teams, and almost every PhD statistician will engage in such collaborations with domain science experts. Collaborations with subject matter experts lead to inspiration for novel methodological research by statisticians whose primary responsibility is to engage in statistical research (mainly but not exclusively in academia), as, often, the need for new statistical methods to handle nonstandard challenges for which existing approaches are not approrpraiate becomes apparent in the course of collaborative projects. Collaborative work is the primary activity for statisticians in industry, research institutions, and government. Frankly, statistics as a discipline would not exist except for the challenges in collection, analysis, and interpretation of data that arise in other fields, so by nature relies on such collaborations to advance.

To be an effective collaborator, a statistician must possess outstanding oral and written communication skills. It is critical that the statistician is able not only to determine appropriate statistical methods to use and to carry out analyses using them, but to interpret the results in the context of the subject matter and to communicate them to his or her nonstatistician collaborators in a way that they can understand. Typically, such collaborators may have some familiarity with basic statistical models and methods at the level of an introductory statistics course or of particular methods that are predominant in their areas of expertise, but not of the more advanced and specialized models and methods that are often required to address the questions. Thus, the statistician must be adept at explaining the rationale for his or her choice of models and methods and why more familiar methods may not be appropriate.

When the questions of interest require longitudinal data models and methods, collaborators may not be familiar with the methods you as a statistician might choose. Collaborators may be familiar with, for example, classical linear regression and possibly logistic regression, and they may have heard of more complex methods for longitudinal and clustered data, but their knowledge of the latter is usually superficial. In some subject matter areas, classical analysis of variance methods may still be favored, which the collaborators believe they understand, but they usually do not appreciate the limitations of these methods and why you as the statistician may recommend and use more modern approaches. In fact, collaborators may sometimes lobby vigorously to use methods other than those that you feel

are appropriate because the journals in their area may not accept a paper for publication reporting on the results of a study if it does not use the "accepted" methods.

One key mechanism by which we as statisticians communicate the results of an analysis we have designed to answer specific scientific questions posed by our collaborators is through a formal data analysis report. Such a report documents systematically what was done and why and explains the results and their interpretation in terms of the subject matter. Here, we give some tips on how to structure and write a good data analysis report.

AUDIENCE: Ordinarily, the intended audience for a data analysis report is your collaborators. It should be written primarily for them and not for you or other statisticians with your level of statistical knowledge. This audience wants to understand what you did, why you did it, and what the conclusions are and, while interested in the features of the methods you used and why you used them, does not want to see technical details, lots of symbols and specialized terminology, and computer code and output.

However, there are other interested audiences. You yourself a key audience, because you will want to have a careful and detailed record of what you did and why you did it if you need to revisit a study in the future and get up to speed quickly if new questions arise or if you are working on another project in which similar issues and questions are to be addressed. You may also have reason to share the report with other statisticians facing similar challenges. Thus, the report must also contain a sufficient level of detail for you and other statistician colleagues.

BASIC STRUCTURE: A good report usually follows this basic structure:

1. Introduction
2. Main Body
3. Summary/Conclusions
4. Appendix
5. References (if needed; see below).

This organization makes it easy for readers with different interests to find information presented at a level appropriate for their backgrounds. We now describe what would ordinarily go into each of these sections.

INTRODUCTION: Even if everyone who will be interested in the report is deeply familiar with the problem, data, and questions, it is always important to provide the following, so that the report stands entirely on its own:

- Subject matter background – what is broad scientific context and what are the challenges and unresolved issues? Here, the report should give a short description of the subject matter problem and why it is important in the domain science area.
- A brief summary of the study carried out to address the challenges.
- A statement of each of the specific scientific questions to that will be addressed.
- A "high-level" summary of the conclusions of the data analysis in the context of the subject matter area.
- A brief roadmap for the rest of the report indicating what can be found in each subsequent section.

MAIN BODY: The main portion of the report can be organized in whatever way makes the most sense to you given the nature of the study and questions. Here is one standard way.

- Detailed summary of the data. For continuous longitudinal data, at the very least, spaghetti plots of the data, perhaps separately by natural groupings (e.g., treatment group), should be presented. If the data are very large, random samples can be plotted. For discrete data, other summaries (e.g., for binary data, sample proportions at each time point) can be presented in tables. Missing data and any other features should be noted. Any information on what led to these should be discussed and implications for interpretation of summaries and plots should be noted (e.g., plots of mean outcome at different time points may not be based on the same numbers of individuals).
- If addressing each question involves a different statistical model, you may wish to have a separate section for each question in which you present the question, the model in which it will be addressed, and the analysis, and the results. If all questions are addressed within the same statistical model framework, you may wish to present the model first in its own dedicated section and then have a separate section for each question, stating the question, the analysis, and the results. You should use your best judgment as to what makes it easiest for a reader.

- For each statistical model, you should present a description of the basic features of the model, an explanation for why it is appropriate, and a summary of the assumptions it embodies and the extent to which those assumptions are satisfied for the data at hand. In particular, for longitudinal analysis, you need to explain why specialized statistical models and methods are required and why other, more familiar methods are not appropriate. You also need to explain the basic features of the model and how the question(s) can be stated in terms of the model. Modeling choices and assumptions and your rationale for them should be clearly stated; you might include plots or other summaries that support the model and assumptions. These should be related back to the subject matter; so, for example, if a model assumes constant within-individual variance of the outcome, why that assumption makes sense in the subject matter context and/or is supported by the data should be stated.

  In general, investigators are not interested in seeing lots of equations, formulæ, matrices, and mathematical symbols. Some who are more sophisticated statistically may be able to tolerate some symbols and equations, but many who are not well-versed in statistics will prefer a description of the model that is mainly in words, with perhaps a "hand wavy" equation or two that helps give a sense of the framework but is not precise. Be sure to define clearly any symbols you do use. Any statistical terminology and concepts that may be unfamiliar to the investigators should be defined and explained. So, for example, don't just say "within-individual correlation" without explaining what that term/concept represents from a subject matter point of view and why it is important.

  Describe the method used to fit the model and mention the software that was used (investigators will want to state this in any papers they write reporting on the study). The investigators will likely not be familiar with the methods, so provide an description in nontechnical terms of the basic premise of the method; e.g., "this is similar to least squares for fitting a conventional linear regression model, but takes into account the correlation among longitudinal outcomes on the same subject."

- For a given question, state how the question can be represented in terms of the model. Present relevant numerical results and interpret them in the context of the subject matter. So, for example, do not give a p-value and say "so we reject H_0." Explain results in terms of the science, and present all important information. Never present an estimate without an associated standard error, and comment on the quality and precision of the results. If there are graphs or other data summaries that shed light on a result, present/discuss them.

- For each analysis, point out any limitations or caveats. For example, if results are predicated on certain assumptions embodied in the model for which there is not substantial support, comment on how robust a conclusion might be to violation of those assumptions.

- ***Do not*** include code or raw output! It is fine to summarize results in a table, but the table should not just be the raw output from software. All columns and entries should be explained in a table caption, with additional explanation if needed in the text. Code and output belong in an appendix to the report; see below. Never instruct or expect readers to go look at these; everything that investigators need to read should be in the main report.

SUMMARY/CONCLUSIONS: As with any report on any topic, there should be a section providing an overarching summary of the objectives and results. This final section should present the scientific questions again, the conclusions of the analysis, and the interpretation of them in terms of the subject matter. Discussion of the implications of the results for the science; any additional observations or findings that, while not directly related to the questions, seem interesting; and possible future studies suggested by the analyses and results can be given here.

APPENDIX: An appendix to the report contains technical details and supporting information. Typical things that would be presented include:

- A detailed description of all statistical models, with all symbols precisely defined, and precise statements of the assumptions that were made. This can be written for a statistician with general knowledge of statistics but perhaps not deep familiarity with the models and methods used.

- A precise, technical explanation of the methods that were used, including how they were implemented (e.g., software used, with any special options or assumptions noted).

- Code implementing the methods; it is prudent to document all code with extensive descriptive comments, which will make it easier for interested readers to understand it (and for you to understand what you did if you revisit the project or want to use it as an example of what you wish to do in a future project). Output should also be included; if this is voluminous, you may want to present only the key parts of the output, but at the very least all important output should be included.

- Additional data summaries, tables, figures, that might be of interest but are not directly relevant to the results.

It is fine to refer readers to the appendix in the main part of the report for more information and details, but looking at the appendix should not be required. All necessary information for investigators to understand what was done and the results should be in the main report. As above, never refer investigators to output or code.

MISCELLANEOUS: Some additional items:

- If any literature (papers, books, software documentation) is cited, there should be a References section with full information on each, in a consistent format.
- It goes without saying that a report should be typed.
- There should be no misspellings or grammatical errors; always spell check any report before finalizing it!
- The entire report should be organized in a logical fashion, with section headings that make it easy for a reader interested in a particular result or question to locate that portion.
- Keep the writing straightforward and to the point, but not to the point that it is so brief that important information is obscured or missing. Check for run-on sentences and avoid language that is too flowery and wordy. Do not use terminology or words that are unfamiliar to a likely reader unless it is necessary, and, if you do, define them. If you use acroynms, present the entire term the first time you use it and define the acronym; e.g., “generalized estimating equation (GEE).”
- The narrative should flow naturally and “tell a story,” so should be easy to follow. To not go off on tangents regarding details; place details that are not central to the flow of the narrative in the appendix and refer to them.

Overall, good, clear writing is essential. Good report writing, both the writing itself and the knack for organizing the information in a sensible, logical way, is a skill that some people are born with but most people must learn. Use each report you write as an opportunity to develop and hone your skill. Good report writing skills will serve you well in not only collaborative work but in writing statistical research papers.

References

Full citations for all books, monographs, and journal articles referenced in the notes are given here. Also included are references to texts from which material in the notes was adapted. The books and monographs cited are all useful resources for learning about futher developments in the analysis of repeated measurement data.

Adams, B.M., Banks, H.T., Davidian, M., and Rosenberg, E.S. (2007). Model fitting and prediction with HIV treatment interruption data. *Bulletin for Mathematical Biology*, 69, 563–584.

Beal, S.L. and Sheiner, L.B. (1985). Methodology of population pharmacokinetics. In *Drug Fate and Metabolism – Methods and Techniques* (eds. E.R. Garrett and J.L. Hirtz). New York: Marcel Dekker.

Bennett, J.E., Racine-Poon, A., and Wakefield, J.C. (1996). MCMC for nonlinear hierarchical models. In *Markov Chain Monte Carlo in Practice*, (eds. W.R. Gilks, S. Richardson, and D.J. Spiegelhalter). London: Chapman & Hall.

Boos, D.D. (1992). On generalized score tests. *The American Statistician*, 46, 327–333.

Booth, J.G. and Hobert, J.P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B*, 61, 265–285.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9–25.

Breslow, N.E. and Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, 82, 81–91.

Carroll, R.J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. New York: Chapman and Hall/CRC Press.

Carroll, R.J., Ruppert, D., Stefanski, L.A., and Crainiceanu, C.M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd edition. New York: Chapman and Hall/CRC Press.

Crowder, M.J. (1995). On use of a working correlation matrix in using generalized linear models for repeated measures. *Biometrika*, 82, 407–410.

Crowder, M.J. and Hand, D.J. (1990). *Analysis of Repeated Measures*. London: Chapman and Hall/CRC Press.

Davidian, M. and Carroll, R.J. (1987). Variance function estimation. *Journal of the American Statistical Association*, 82, 1079–1091.

Davidian, M. and Gallant, A.R. (1992). Smooth nonparametric maximum likelihood for pop- ulation pharmacokinetics, with application to quinidine. *Journal of Pharmacokinetics and Biopharmaceutics*, 20, 529–556.

Davidian, M. and Gallant, A.R. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika*, 80, 475–488.

Davidian, M. and Giltinan, D.M. (1993). Some simle methods for estimating intra-individual variability in nonlinear mixed effects models. *Biometrics*, 49, 59–73.

Davidian, M. and Giltinan, D.M. (1995). *Nonlinear Models for Repeated Measurement Data*. London: Chapman and Hall/CRC Press.

Davidian, M. and Giltinan, D.M. (2003). Nonlinear models for repeated measurement data: An overview and update. *Journal of Agricultural, Biological, and Environmental Statistics*, 8, 387–419.

Diggle, P.J., Heagerty, P., Liang, K.-Y., and Zeger, S.L. (2002). *Analysis of Longitudinal Data*, 2nd edition. New York: Oxford University Press.

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2009). *Longitudinal Data Analysis*. Boca Raton: Chapman and Hall/CRC Press.

Fitzmaurice, G.M., Laird, N.M., and Ware, J.H. (2011). *Applied Longitudinal Analysis*, 2nd edition. New York: Wiley.

Fitzmaurice, G.M., Molenberghs, G., and Lipsitz, S.R. (1995). Regression models for longitudinal binary responses with informative dropouts. *Journal of the Royal Statistical Society, Series B*, 57, 691–704.

Fuller, W.A. (1987). *Measurement Error Models*. New York: John Wiley and Sons.

Gelman, A., Bois, F., and Jiang, L.M. (1996). Physiological pharmacokinetic analysis using population modelling and informative prior distributions. *Journal of the American Statistical Association*, 91, 1400–1412.

Gibaldi, M. and Perrier, D. (1982). *Pharamcokinetics* (2nd edn.). New York: Marcel-Dekker.

Giltinan, D.M. (2014). Pharmacokinetics and pharmacodynamics. *Wiley StatsRef: Statistics Reference Online*. Wiley, DOI: 10.1002/9781118445112.stat05078.

Gumpertz, M. and Pantula, S.G. (1989). A simple approach to inference in random coefficient models. *The American Statistician*, 43, 203-210.

Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J., and Ostrowski, E. (1994). *A Handbook of Small Data Sets*. London: Chapman and Hall/CRC Press.

Harville, D.A. (1997). *Matrix Algebra from a Statistician's Perspective*. New York: Springer.

Henderson, C.R. (1984). *Applications of Linear Models in Animal Breeding*. Guelph, Ontario, Canada: University of Guelph Press.

Johnson, R.A. and Wichern, D.W. (2002). *Applied Multivariate Statistical Analysis, Fifth Edition*. Englewood Cliffs, New Jersey: Prentice Hall.

Laird, N.M. and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics*, 38, 963–974.

Liang, K.-Y., Zeger, S.L., and Qaqish, G. (1992). Multivariate regression analysis for categorical data. *Journal of the Royal Statistical Society, Series B*, 54, 3–40.

Lin, X. and Breslow, N.E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91, 1007–1016.

Lindsey, J.K. (1993). *Models for Repeated Measurements*. New York: Oxford University Press.

Lindstrom, M.J. and Bates, D.M. (1990). Nonlinear mixed effects models for repeated measurement data. *Biometrics*, 46, 673–687.

Lipsitz, S.R., Laird, N.M., and Harrington, D.P. (1991). Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika*, 78, 153–160.

Lipsitz, S.R., Laird, N.M., and Harrington, D.P. (1992). A three-stage estimator for studies with repeated and possibly missing binary outcomes. *Applied Statistics*, 41, 203–213.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D., and Schabenberger, O. (2006). *SAS System for Mixed Models, Second Edition*, Cary NC: SAS Institute, Inc.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd edition. New York: Wiley.

Longford, N.T. (1993). *Random Coefficient Models*. New York: Oxford University Press.

Louis, T. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association*, 79, 393–398.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd edition. London: Chapman and Hall/CRC Press.

McCulloch, C.E. (1997). Maximum likelihood algorthms for generalized linear mixed models. *Journal of the American Statistical Association*, 92, 162–170.

Molenberghs, G. and Kenward, M. G. (2007). *Missing Data in Clinical Studies*. Chichester, UK: Wiley.

Müller, P. and Rosner, G.L. (1997). A Bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association*, 92, 1279-1292.

Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58, 545–554.

Pepe, M.S. and Anderson, G. L. (1994). A cautionary note on inference in marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics – Simulation and Computation*, 24, 939–951.

Pinheiro, J.C. and Bates, D.M. (1995). Approximations to the log-likelihood function in the nonlinear mixed effects model. *Journal of Computational and Graphical Statistics*, 4, 12–35.

Pinheiro, J.C. and Bates, D.M. (2000). *Mixed Effects Models in S and S-PLUS*. New York: Springer.

Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44, 1033–1048.

Prentice, R.L. and Zhao, L.P. (1991). Estimating equations for parameters in means and covariances of discrete and continuous responses. *Biometrics*, 47, 825–838.

Potthoff, R.F. and Roy, S.N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51, 313–326.

Rabe-Hesketh, S. and Skrondal, A. (2009). Generalized linear mixed-effects models. In *Longitudinal Data Analysis* (eds. G. Fitzmaurice, M. Davidian, G. Verbeke, G. Molenberghs). Boca Raton: CRC Press/Chapman and Hall.

Robins, J.M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics – Theory and Methods*, 23, 379–2412.

Robins, J.M., Greenland, S., and Hu, F.-C. (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association*, 94, 687–712.

Robins, J.M., Hernán, M.A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11, 550-560.

Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106–121.

Robinson, G.K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6, 15–51.

Rosner, G.L. and Müller, P. (1994). Pharmacodynamic/pharmacokinetic analysis of hematologic profiles. *Journal of Pharmacokinetics and Biopharmaceutics*, 22, 499–524.

Rotnitzky, A. and Jewell, N.P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, 77, 485–497.

Rowell, J.G. and Walters, D.E. (1976). Analyzing data with repeated observations on each experimental unit. *Journal of Agricultural Science*, 87, 423–432.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581–591.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 40, 917–927.

Searle, S.R., Casella, G., and McCulloch, C.E. (2006). *Variance Components*. Hoboken, New Jersey: Wiley.

Self, S.G. and Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio test statistics under nonstandard conditions. *Journal of the American Statistical Association*, 82, 605–610.

Shen, W. and Louis, T.A. (1991). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society, Series B*, 60, 455–471.

Stefanski, L.A. and Boos, D.D. (2002). The calculus of M-estimation. *The American Statistician*, 56, 29–38.

Stram, D.O. and Lee, J.W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50, 1171-1177. (Correction (1995), *Biometrics*, 94, 1196).

Thall, P.F. and Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, 46, 657–671.

Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Vansteelandt, S. and Joffe, M. (2014). Structural nested models and G-estimation: The partially realized promise. *Statistical Science*, 29, 707–731.

Verbeke, G. and Molenberghs, G. (1997). *Linear Mixed Models in Practice: A SAS-Oriented Approach¿* Lecture Notes in Statistics 126. New York: Springer.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics*, 59, 254–262.

Vonesh, E.F. (1996). A note on the use of Laplace's approximation for nonlinear mixed effects models. *Biometrika*, 83, 447–452.

Vonesh, E.F. and Chinchilli, V.M. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. New York: Marcel Dekker.

Wakefield, J. (1996). The Bayesian analysis of population pharmacokinetic models. *Journal of the American Statistical Association*, 91,62–75.

Wakefield, J., Smith, A.F.M., Racine-Poon, A., and Gelfand, A.E. (1994). Bayesian analysis of linear and nonlinear population models by using the Gibbs sampler. *Applied Statistics*, 43, 201–221.

Walker, S.G. (1996). An EM algorithm for non-linear random effects models. *Biometrics*, 52, 934–944.

Weiss, R.E. (2005). *Modeling Longitudinal Data*. New York: Springer.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika*, 80, 791–795.

Wolfinger, R. and Lin, X. (1997). Two Taylor-series approximation methods for nonlinear mixed models. *Computational Statistics and Data Analysis*,25, 465–490.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48, 233–243.

Zeger, S.L. and Karim, M.R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association*, 86, 79–86.

Zeger, S.L., Liang, K.-Y., and Albert, P.S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44, 1049–1066.

Zhang, D. and Davidian, M. (2001). Linear mixed models with exible distributions of random effects for longitudinal data. *Biometrics*, 57, 795–802.

Zhao, L.P. and Prentice, R.L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, 77, 642–648.