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.

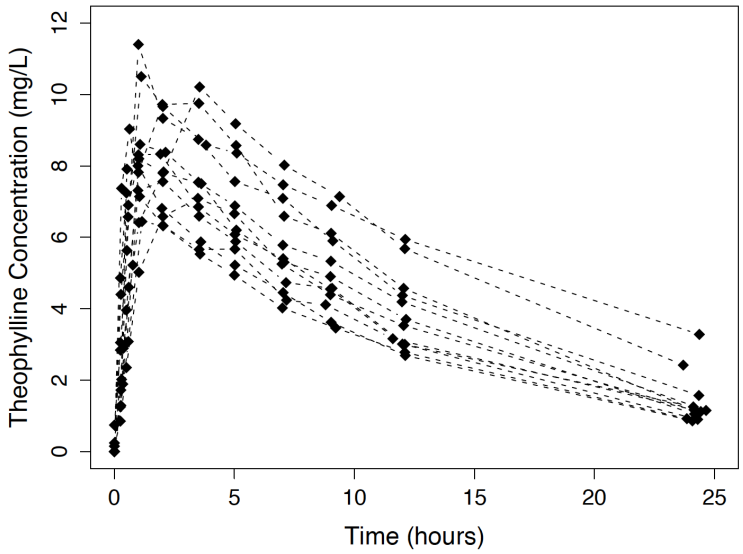

Figure 9.1: *Theophylline concentration-time profiles for $m = 12$ subjects receiving an oral dose of theophylline at time 0.*

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.

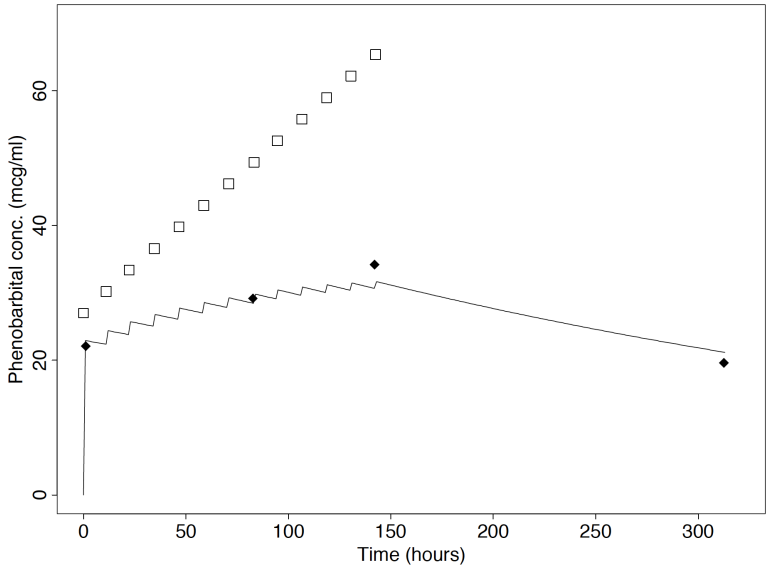

Figure 9.2: *Phenobarbital data for infant 9. The diamonds are phenobarbital concentrations, the open squares represent cumulative dose (μg/kg), and the solid line is a fit of the model* (9.30).

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.

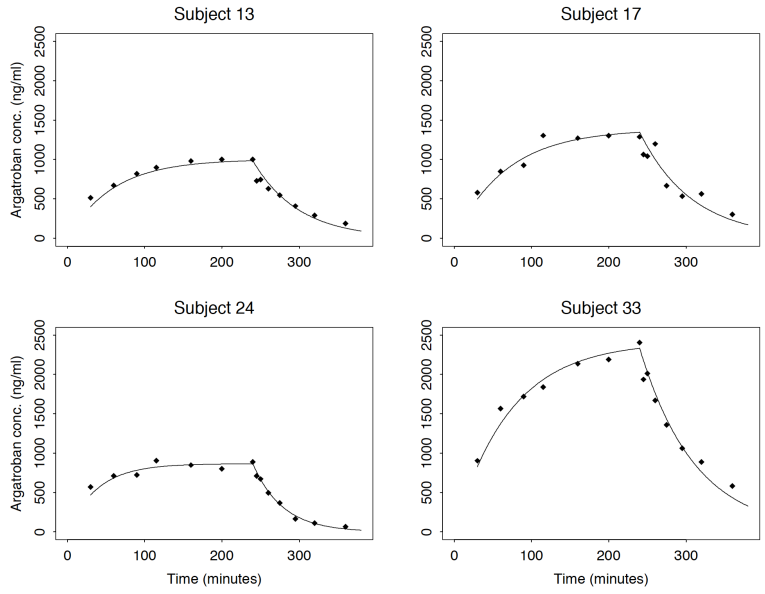

Figure 9.3: *Concentration-time data for four subjects from the argatroban pharmacokinetic study.*

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.

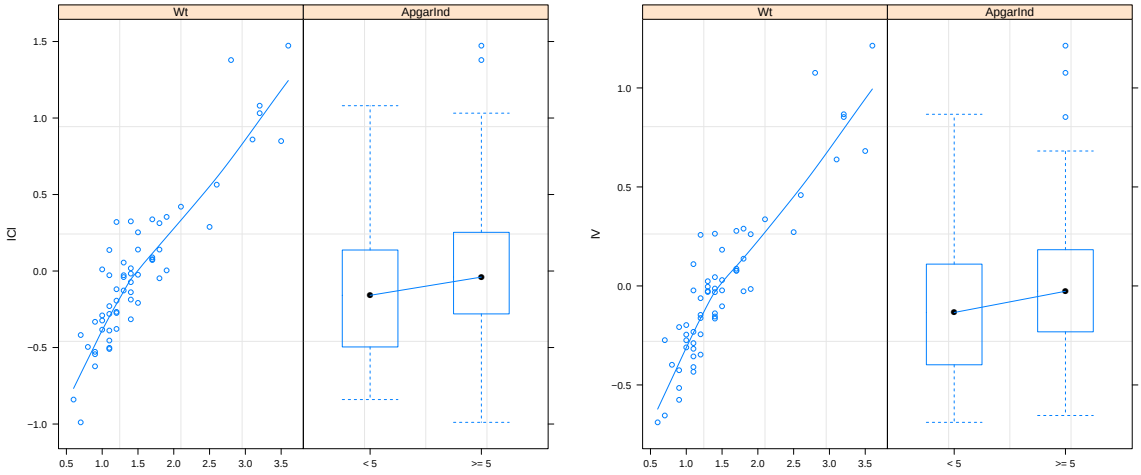

Figure 9.4: *Empirical Bayes estimates of the random effects b_{1i} and b_{2i} from a fit of model (i) plotted against the covariates birthweight and Apgar score category..*

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.