Chapter 1

An Opinionated Survey of Methods for Mixed Linear Models

This chapter introduces the mixed-linear-model form using examples that would have been familiar in, say, the 1960s. Later chapters show some of the vast range of models that can be written in this form, few if any of which would have been recognizable in the 1960s. After the mixed-linear-model form is introduced in Section 1.1, Section 1.2 surveys the conventional (non-Bayesian) approach to analyzing these models, Section 1.3 surveys the Bayesian approach, and Section 1.4 summarizes differences between the two approaches and applies both to an example. Section 1.5 concludes with some comments on computing.

I sometimes refer to the times before and after the revolution in Bayesian computing created by and still consisting mostly of Markov chain Monte Carlo (MCMC) methods. The landmark papers in this great breakthrough were Geman & Geman (1984), Gelfand & Smith (1990), and Tierney (1994); for brevity I use 1990 as the time when Bayesian computing became broadly practicable.

1.1 Mixed Linear Models in the Standard Formulation

The notation defined here will be used throughout the book. Mixed linear models can be written in the form

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \varepsilon, \text{ where} \tag{1.1}$$

- the observation $\mathbf{y}$ is an n-vector;
- $\mathbf{X}$ is a known design matrix of size $n \times p$, for the so-called fixed effects;
- β, containing the fixed effects, is $p \times 1$ and unknown;
- $\mathbf{Z}$ is a known design matrix of size $n \times q$, for the so-called random effects;
- $\mathbf{u}$, containing the random effects, is $q \times 1$ and unknown;
- $\mathbf{u}$ is modeled as q-variate normal with mean zero and covariance $\mathbf{G}$, which is a function of unknowns in the vector ϕ_G, so $\mathbf{G} = \mathbf{G}(\phi_G)$;
- ε is an unobserved normally distributed error term with mean zero and covariance $\mathbf{R}$, which is a function of unknowns in the vector ϕ_R, so $\mathbf{R} = \mathbf{R}(\phi_R)$; and
- The unknowns in $\mathbf{G}$ and $\mathbf{R}$ are denoted $\phi = (\phi_G, \phi_R)$.

Most commonly, the errors in ε are independent and identically distributed so $\mathbf{R} = \sigma_e^2 \mathbf{I}_n$ for the unknown error variance $\phi_R = \sigma_e^2$ ($\mathbf{I}_n$ being the $n \times n$ identity matrix), but this is not always the case.

If $\mathbf{Zu}$ is eliminated from (1.1) or equivalently if $\mathbf{u}$'s covariance matrix $\mathbf{G}$ is set to zero, then (1.1) is the familiar linear model with a single error term, ε. Thus, the novelty in mixed linear models arises from $\mathbf{Zu}$ and the unknown ϕ_G. Writing a model, such as a penalized spline, in this form mostly involves specifying $\mathbf{Z}$ and $\mathbf{G}(\phi_G)$ and, as far as I know, *all* of the oddities and inconveniences examined in this book arise because ϕ_G is unknown.

The simplest useful example of a mixed linear model is the balanced one-way random-effects model. In this model, observations y_{ij} come in clusters indexed by $i = 1, \ldots, q$ with observations in each cluster indexed by $j = 1, \ldots, m$. One way to write this model is

$$y_{ij} = \mu + u_i + \varepsilon_{ij}, \tag{1.2}$$

where the u_i are independent and identically distributed (iid) $N(0, \sigma_s^2)$ and the ε_{ij} are iid $N(0, \sigma_e^2)$ and independent of the u_i. The y_{ij} have two components of variance, to use an old jargon term, with u_i capturing variation between clusters that affects all the observations in cluster i, and the ε_{ij} representing variation specific to observation (i, j). In the standard notation, $n = qm$, $\mathbf{y} = (y_{11}, \ldots, y_{qm})'$, $\mathbf{X} = \mathbf{1}_{qm}$, a qm-vector of 1's, $\beta = \mu$, $\mathbf{Z} = \mathbf{I}_q \otimes \mathbf{1}_m$, $\mathbf{u} = (u_1, \ldots, u_q)'$, $\varepsilon = (\varepsilon_{11}, \ldots, \varepsilon_{qm})'$, $\mathbf{G} = \sigma_s^2 \mathbf{I}_q$, and $\mathbf{R} = \sigma_e^2 \mathbf{I}_{qm}$, where $\otimes$ is the Kronecker product, defined for matrices $\mathbf{A} = (a_{ij})$ and $\mathbf{B}$ as $\mathbf{A} \otimes \mathbf{B} = (a_{ij}\mathbf{B})$.

The definitions above referred to "so-called fixed effects" and "so-called random effects" because the term *random effect* is no longer very well defined, to the point that at least one prominent statistician has advocated abolishing it (Gelman 2005a, Section 6). This issue is discussed at length in Chapter 13, which relies on material in the intervening chapters. For now, note that the term *random effect* originally referred to a very specific thing but now it includes a much larger and more miscellaneous collection of things:

- *Original meaning (old-style random effects)*: The levels of a random effect (in analysis-of-variance jargon) are draws from a population, and the draws are not of interest in themselves but only as samples from the larger population, which *is* of interest. In this original meaning, the random effects $\mathbf{Zu}$ provide a way to model sources of variation that affect several observations in a common way as in the one-way random effect model, where the cluster-specific random effect u_1 affects all the y_{1j}.
- *Current meaning (new-style random effects)*: In addition to the above, a random effect may have levels that are not draws from any population, or that are the entire population, or that may be a sample but a new draw from the random effect could not conceivably be drawn, and in all these cases the levels themselves are of interest. In this extended usage, the random effect $\mathbf{Zu}$ provides a model that is flexible because it is richly parameterized but that avoids overfitting because it constrains $\mathbf{u}$ by means of its covariance $\mathbf{G}$.

The original meaning given above is much like the definition in Scheffé (1959, p. 238) but differs in that Scheffé required a random sample from a population, not merely the more vague "draws" used above. (I use "draws" because outside of survey-sampling contexts, I have never seen anyone draw a genuinely random sample for analysis with a mixed linear model.) The current meaning of *random effects* has come into being implicitly as more statisticians notice that the mathematical *form* of a random effect can be used to model situations that do not fit the original meaning of random effect. Throughout this book, I will call these old-style and new-style random effects.

Later chapters show many examples of new-style random effects. The following three examples include old-style random effects; all arose in projects on which I collaborated. Besides exemplifying mixed-linear-model notation, they give a sense of the variety of models that fit in this framework even if we restrict ourselves to old-style random effects. Also, each example's analysis produced an inconvenient, mysterious, or plainly wrong result that no current theory of linear mixed models can explain or remedy. The second and third examples are dead ends in that as yet nobody can explain why these problems occur or offer advice more specific than "try something else." They suggest how readily such problems occur — and how many research opportunities await.

Example 1. Molecular structure of a virus. Dwight Anderson's lab at the University of Minnesota School of Dentistry used cryoelectron microscopy and other methods to develop a hypothesized molecular description of the outer shell (prohead) of the bacteriophage virus $\phi 29$ (Peterson et al. 2001). To test this hypothesized model, they estimated the count of each kind of protein in the prohead, to compare to counts they had hypothesized. They did so by breaking the prohead or phage into its constituent proteins, weighing those proteins, and converting the weight into a count of copies. I use as an example the major capsid protein, gp8. Counting copies of gp8 had four major steps:

- Selecting a *parent*; there were 2 prohead parents and 2 phage parents;
- Preparing a *batch* of the parent;
- On a given gel date, creating electrophoretic *gels*, separating the different proteins on the gels, and cutting out the piece of the gel relevant to gp8; and
- Burning several such gel pieces in an *oxidizer run* to get a gp8 weight from each gel, which was then converted into a count of gp8 copies.

For the gp8 counts, there were 4 parents, 9 batches, 11 gels, and 7 oxidizer runs for a total of 98 measurements. In the analyses published in Peterson et al. (2001), I treated each of these measurement steps as an old-style random effect, though now that I am older I recognize this is debatable at best for the four parents. The batches, gels, and oxidizer runs, however, are clearly old-style random effects: They can be viewed as draws from hypothetical infinite populations of batches, gels, and oxidizer runs; these 9, 11, and 7 draws, respectively, are of interest only for what they tell us about variation between batches, gels, and runs; and they have no interest in themselves. Table 1.1 shows the first 33 measurements of the gp8 count, each with its

Table 1.1: Molecular Structure of a Virus: First 33 Rows of Data for gp8 Weight

Parent	Batch	Gel Date	Oxidizer Run	gp8 Weight
1	1	1	1	244
1	1	2	1	267
1	1	2	1	259
1	1	2	1	286
1	3	1	1	218
1	3	2	1	249
1	3	2	1	266
1	3	2	1	259
1	7	4	3	293
1	7	4	3	277
1	7	4	3	286
1	7	4	3	297
1	7	5	4	315
1	7	5	4	302
1	7	5	4	312
1	7	5	4	319
1	7	5	4	316
1	7	5	4	321
1	7	5	4	293
1	7	5	4	283
1	7	7	4	311
1	7	7	4	282
1	7	7	4	283
1	7	7	4	276
1	7	7	4	331
1	7	7	4	252
1	7	7	4	326
1	7	7	4	334
2	2	1	1	272
2	2	2	1	223
2	4	1	1	208
2	4	2	1	226
2	4	2	1	223

parent, batch, gel date, and oxidizer run. (This design is far from ideal, but this is what was presented to me. In view of the ingenuity and labor that went into gathering these data, I waited until the paper was accepted and then gently advised my collaborators to talk to me before they did something like this again.)

To analyze these data, I used the oldest and simplest mixed linear model, the variance component model, in which the i^{th} measurement of the gp8 count, y_i, is

modeled as

$$\begin{aligned} y_i &= \mu + \text{parent}_{j(i)} + \text{batch}_{k(i)} + \text{gel}_{l(i)} + \text{run}_{m(i)} + \varepsilon_i \\ \text{parent}_{j(i)} &\overset{iid}{\sim} N(0,\sigma_p^2), j = 1,\ldots,4 \\ \text{batch}_{k(i)} &\overset{iid}{\sim} N(0,\sigma_b^2), k = 1,\ldots,9 \\ \text{gel}_{l(i)} &\overset{iid}{\sim} N(0,\sigma_g^2), l = 1,\ldots,11 \\ \text{run}_{m(i)} &\overset{iid}{\sim} N(0,\sigma_r^2), m = 1,\ldots,7 \\ \varepsilon_i &\overset{iid}{\sim} N(0,\sigma_e^2), i = 1,\ldots,98, \end{aligned} \tag{1.3}$$

where the deterministic functions $j(i), k(i), l(i)$, and $m(i)$ map i into the parent, batch, gel, and run indices, respectively.

In the standard formulation of the mixed linear model, equation (1.1), $\mathbf{X} = \mathbf{1}_{98}$, $\beta = \mu$,

$$\begin{array}{rcl} & \begin{array}{cccc} \text{parent (4 cols)} & \text{batch (9 cols)} & \text{gel (11 cols)} & \text{run (7 cols)} \end{array} & \\ \mathbf{Z} = & \begin{array}{cccc} \overbrace{1000} & \overbrace{10\ldots0} & \overbrace{10\ldots0} & \overbrace{10\ldots0} \\ 1000 & 10\ldots0 & 01\ldots0 & 10\ldots0 \\ \vdots & \vdots & \vdots & \vdots \\ 0001 & 00\ldots1 & 00\ldots1 & 00\ldots1 \end{array} & , \\ \mathbf{u} = & [\text{parent}_1,\ldots,\text{parent}_4,\text{batch}_1,\ldots,\text{batch}_9,\text{gel}_1,\ldots,\text{gel}_{11},\text{run}_1,\ldots,\text{run}_7]', & \end{array}$$

where $\mathbf{Z}$ is 98×31 and $\mathbf{u}$ is 31×1. $\mathbf{G}$, the covariance matrix of $\mathbf{u}$, is 31×31 block-diagonal with blocks $\sigma_p^2\mathbf{I}_4, \sigma_b^2\mathbf{I}_9, \sigma_g^2\mathbf{I}_{11}$, and $\sigma_r^2\mathbf{I}_7$, so that the unknowns in $\mathbf{G}$ are $\phi_G = (\sigma_p^2, \sigma_b^2, \sigma_g^2, \sigma_r^2)$. Finally, $\mathbf{R}$, the covariance matrix of the errors ε, is $\sigma_e^2\mathbf{I}_{98}$, so $\phi_R = \sigma_e^2$.

Similar models were used to analyze counts of the other proteins in the $\phi 29$ bacteriophage virus. In Section 1.4.2, we will see that the conventional analysis, maximizing the restricted likelihood, produces zero estimates of several of these variance components, which is obviously wrong and about which the theory of the conventional analysis is silent. Chapter 18 discusses zero variance estimates and associated practical questions, such as determining whether the restricted likelihood for ϕ or the data's contribution to the marginal posterior of ϕ is flat (contains little information) near zero.

Example 2. Testing a new system for imaging vocal folds. Ear-nose-and-throat doctors evaluate certain speech or larynx problems by taking video images of the inside of the larynx during speech and having trained raters assess the resulting images. As of 2008, the standard method used strobe lighting with a slightly longer period than the period of the vocal folds' vibration, a clever trick that produces an artificially slowed-down image of the folds' vibration. The new method under study used high-speed video (HSV) to allow very slow-motion replay of the video and thus a direct

Table 1.2: Imaging Vocal Folds: First 9 Subjects and 26 Rows of Data for "Percent Open Phase"

Subject ID	Imaging Method	Rater	% Open Phase
1	strobe	CR	56
1	strobe	KK	70
1	strobe	KK	70
1	HSV	KUC	70
1	HSV	KK	70
1	HSV	KK	60
2	strobe	KUC	50
2	HSV	CR	54
3	strobe	KUC	60
3	strobe	KUC	70
3	HSV	KK	56
4	strobe	CR	65
4	HSV	KK	56
5	strobe	KK	50
5	HSV	KUC	55
5	HSV	KUC	67
6	strobe	KUC	50
6	strobe	KUC	50
6	HSV	KUC	50
7	strobe	KK	50
7	HSV	KUC	57
8	strobe	CR	56
8	strobe	KUC	60
8	HSV	CR	50
9	strobe	CR	92
9	HSV	KUC	78

view of the folds' vibration. Katherine Kendall (2009) applied both imaging methods to each of 50 subjects, some healthy and some not. Each of the resulting 100 images was assessed by raters, labeled CR, KK, and KUC. Some images were rated by more than one rater and/or by a single rater more than once, in a haphazard design having 154 total ratings. (Again, this unfortunate design was presented to me as a *fait accompli*.) For each image, a rating consisted of 5 quantities on continuous scales. Dr. Kendall's interest was in differences between the new and old imaging methods and between raters, and in their interaction. Table 1.2 shows the design variables for the first 9 subjects (26 ratings), with the outcome "percent open phase."

This is a generalization (or corruption, if you prefer) of a repeated-measures design, where the subject effect is "subject ID" and the within-subject fixed effects are method, rater, and their interaction. I analyzed these data using a mixed linear model including those fixed effects and random effects for a subject's overall level, for the

interaction of subject and method (describing how the difference between methods varies between subjects), for the interaction of subject and rater (describing how the difference between raters varies between subjects), and a residual. The last could also be called the three-way interaction of subject, method, and rater (describing how the method-by-rater interaction varies between subjects) but conventionally it is called the residual and treated as an error term. As for Example 1, these three subject effects are old-style random effects that would have been recognized by Scheffé's contemporaries: They all arise from drawing subjects from the population of potential subjects, though not by a formal randomization, and the subjects themselves have no inherent interest.

To fit this analysis into the mixed-linear-model framework (1.1), the residual covariance matrix is $\mathbf{R} = \sigma_e^2 \mathbf{I}_{154}$ with unknown σ_e^2. The fixed-effect design matrix $\mathbf{X}$ and parameter β encode imaging method and rater and could use any of several parameterizations, each with its own specification of $\mathbf{X}$ and β. For the indicator parameterization (sometimes called "treatment contrasts") with strobe and KK as the reference (base) levels for method and rater respectively, the first 26 rows of $\mathbf{X}$ are given in Table 1.3. The values of β corresponding to these columns are, from left, the average for strobe rated by KK; HSV minus strobe for KK; CR minus KK for strobe; KUC minus KK for strobe; and two βs for interactions.

Regarding the random effects, with a design this messy it is easiest to think first of the elements of $\mathbf{u}$ and then construct the corresponding $\mathbf{Z}$. Here is the random-effect specification used by the JMP software in which I did the computations (which is the same specification as in SAS's MIXED procedure). The subject effect has 50 elements, $u_{s1}, \ldots, u_{s50}$; the subject-by-method effect has 100 elements $u_{t1}, u_{H1}, \ldots, u_{t50}, u_{H50}$, where u_{ti}, u_{Hi} are subject i's random effects for strobe and HSV respectively. For the interaction of subject and rater, we have one random effect (one u) for each unique combination of a rater and a subject ID appearing in the dataset, giving $u_{1,CR}, u_{1,KK}$, $u_{1,KUC}, u_{2,CR}, u_{2,KUK}, u_{3,KK}, u_{3,KUC}, u_{4,CR}, u_{4,KK}, \ldots, u_{49,CR}, u_{50,CR}$. Table 1.4 gives the columns of $\mathbf{Z}$ corresponding to the first three subjects in the dataset (the left three columns are labels, the right three clusters of columns are the columns of $\mathbf{Z}$).

This specification of the random effects differs from the specification of a balanced mixed model given by Scheffé (1959, e.g., Section 8.1), in which the random effects are defined to sum to zero across subscripts referring to fixed effects. The specification given here avoids quandaries arising from empty design cells, which the present problem has.

When I applied this analysis to the five outcome measures, the numerical routine performing the standard analysis (maximizing restricted likelihood) converged for four outcomes but not for the fifth. The designs were identical for the 5 outcomes $\mathbf{y}$ so the differences in convergence must arise from differences in the outcomes, not the design.[1] However, it is unknown *how* differences in the outcomes produce differences in convergence. Unfortunately, I have made zero progress on the latter

[1] Actually, they weren't quite identical. One of the outcomes for which the algorithm converged had a missing observation.

Table 1.3: Imaging Vocal Folds: First 26 Rows of Fixed-Effect Design Matrix **X**

1	0	1	0	0	0
1	0	0	0	0	0
1	0	0	0	0	0
1	1	0	1	0	1
1	1	0	0	0	0
1	1	0	0	0	0
1	0	0	1	0	0
1	1	1	0	1	0
1	0	0	1	0	0
1	0	0	1	0	0
1	1	0	0	0	0
1	0	1	0	0	0
1	1	0	0	0	0
1	0	0	0	0	0
1	1	0	1	0	1
1	1	0	1	0	1
1	0	0	1	0	0
1	0	0	1	0	0
1	1	0	1	0	1
1	0	0	0	0	0
1	1	0	1	0	1
1	0	1	0	0	0
1	0	0	1	0	0
1	1	1	0	1	0
1	0	1	0	0	0
1	1	0	1	0	1

Table 1.4: Imaging Vocal Folds: Rows and Non-zero Columns of the Random-Effect Design Matrix **Z** for the First Three Subjects

Subject ID	Method	Rater	Subject	Subject-by-Method	Subject-by-Rater
			$\overbrace{100}$	$\overbrace{100000}$	$\overbrace{1000000}$
1	strobe	CR	100	100000	1000000
1	strobe	KK	100	100000	0100000
1	strobe	KK	100	100000	0100000
1	HSV	KUC	100	010000	0010000
1	HSV	KK	100	010000	0100000
1	HSV	KK	100	010000	0100000
2	strobe	KUC	010	001000	0001000
2	HSV	CR	010	000100	0000100
3	strobe	KUC	001	000010	0000010
3	strobe	KUC	001	000010	0000010
3	HSV	KK	001	000001	0000001

question, so I have nothing more to say about this example. The dataset is on the book's web site; perhaps you can figure it out.

Example 3. Estimating glomerular filtration rate in kidney-transplant recipients. Glomerular filtration rate (GFR) describes the flow rate of filtered fluids through a person's kidneys. Direct measurements of GFR, such as iothalamate or iohexol GFR, are considered ideal but are very expensive and time-consuming. Outside of research settings, GFR is estimated by plugging standard clinical lab measurements into an equation that predicts GFR. Many such equations have been proposed, most commonly using serum creatinine. However, serum creatinine can be influenced by changes in muscle mass from steroid use, which may give misleading GFR estimates for the large number of kidney-transplant recipients whose immunosuppressive regimens include steroids.

Ward (2010) tested the accuracy and precision of 11 GFR prediction equations against iothalamate GFR (iGFR) using data from 153 subjects in a randomized trial enrolling kidney-transplant recipients. Of the 11 equations, 7 were functions of cystatin C only while 4 were functions of serum creatinine. I describe Ward's project in terms of a generic estimated GFR (eGFR). The study protocol specified that subjects would be randomized about a month after their transplant and have iGFR measured then and annually for the next 5 years. At these annual visits, standard labs were taken and eGFR was computed so each subject had up to 6 pairs of eGFR and the gold-standard iGFR, though many subjects had missed visits or had not finished follow-up at the time of these analyses. One question was: Does eGFR capture the trend over time in iGFR? Even if a given eGFR is biased high (say), if it is consistently biased high, it might accurately capture the trend over time in iGFR. To answer this question, each eGFR was compared separately to iGFR, using the following analysis. I describe the simplest analysis; variants included a covariate describing steroid use and analyses of subsets of subjects.

The dataset included two observations ("cases") per subject per annual visit, one for iGFR and one for eGFR. The mixed linear model fit a straight line in time (visit number) for each combination of a subject and a GFR method (iGFR or eGFR). For each subject these four quantities — a slope and intercept for each method — were treated as an iid draw from a 4-variate normal distribution. Thus the fixed effects design matrix $\mathbf{X}$ had one row per measurement and 4 columns corresponding to population-average values of the intercept for iGFR, the slope in visits for iGFR, the intercept for eGFR, and the slope in visits for eGFR. Table 1.5 shows the rows of $\mathbf{X}$ for two subjects, the first having all 6 visits and the second having only 3 visits. The corresponding fixed effects are $\beta' = (\beta_{0I}, \beta_{1I}, \beta_{0e}, \beta_{1e})$, with subscripts I and e indicating iGFR and eGFR, and subscripts 0 and 1 indicating slope and intercept. The substantive question is whether β_{1I} and β_{1e} differ.

The random-effects design matrix $\mathbf{Z}$ had 4 columns per subject, analogous to the columns for fixed effects. Each subject's columns in $\mathbf{Z}$ were the same as his/her columns in $\mathbf{X}$ for rows corresponding to his/her observations, and were zeros in rows corresponding to other subjects' observations. Table 1.5 shows the rows and columns of $\mathbf{Z}$ corresponding to the rows of $\mathbf{X}$ shown in the same table. Subject i's random

Table 1.5: Glomerular Filtration Rate: Fixed-Effect Design Matrix **X** and Random-Effect Design Matrix **Z** for the First Two Subjects

$$\mathbf{X} = \left[\begin{array}{cccc} 1&0&0&0\\ 1&1&0&0\\ 1&2&0&0\\ 1&3&0&0\\ 1&4&0&0\\ 1&5&0&0\\ 0&0&1&0\\ 0&0&1&1\\ 0&0&1&2\\ 0&0&1&3\\ 0&0&1&4\\ 0&0&1&5\\ \hline 1&0&0&0\\ 1&1&0&0\\ 1&2&0&0\\ 0&0&1&0\\ 0&0&1&1\\ 0&0&1&2\\ \vdots&\vdots&\vdots&\vdots \end{array}\right] \quad \mathbf{Z} = \left[\begin{array}{cccc|cccc} 1&0&0&0&0&0&0&0\ldots\\ 1&1&0&0&0&0&0&0\ldots\\ 1&2&0&0&0&0&0&0\ldots\\ 1&3&0&0&0&0&0&0\ldots\\ 1&4&0&0&0&0&0&0\ldots\\ 1&5&0&0&0&0&0&0\ldots\\ 0&0&1&0&0&0&0&0\ldots\\ 0&0&1&1&0&0&0&0\ldots\\ 0&0&1&2&0&0&0&0\ldots\\ 0&0&1&3&0&0&0&0\ldots\\ 0&0&1&4&0&0&0&0\ldots\\ 0&0&1&5&0&0&0&0\ldots\\ \hline 0&0&0&0&1&0&0&0\ldots\\ 0&0&0&0&1&1&0&0\ldots\\ 0&0&0&0&1&2&0&0\ldots\\ 0&0&0&0&0&0&1&0\ldots\\ 0&0&0&0&0&0&1&1\ldots\\ 0&0&0&0&0&0&1&2\ldots\\ \vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots \end{array}\right]$$

effects $\mathbf{u}_i' = (u_{0I}, u_{1I}, u_{0e}, u_{1e})$ are, respectively, that subject's deviations from the population averages of the intercept for iGFR, the slope in visits for iGFR, the intercept for eGFR, and the slope in visits for eGFR. These analyses assumed **u** was iid normal with mean zero and covariance matrix **G** that was a general 4×4 covariance matrix.

For this analysis, we judged that the errors ε might show two kinds of correlation: Between iGFR and eGFR at a given visit, and serially across visits for a given measurement method. Thus, the errors were modeled as independent between persons, and within persons were autocorrelated to order 1 (AR(1)) within each method and correlated between methods. In the syntax of SAS's MIXED procedure, the REPEATED command had TYPE = UN@AR(1). Thus **R** was block-diagonal with blocks corresponding to subjects and block size two times the number of visits a subject attended. Table 1.6 shows the block in **R** for the second subject, who attended the first 3 visits only.

I will not try to rationalize this analysis beyond the following. The iGFR outcomes were quite variable and a person-specific straight-line trend with serial correlation fit as well as anything could; the investigators did specifically ask about time trends, which clinicians usually interpret as rates, i.e., slopes; and as it turned out, all seven cystatin C equations gave as estimates nearly identical increasing trends for eGFR while iGFR showed a flat or slightly declining estimated trend (depending on

Table 1.6: Glomerular Filtration Rate: Block of Error Covariance Matrix **R** for the Second Subject, with 3 Visits

σ_1^2	$\sigma_1^2\rho$	$\sigma_1^2\rho^2$	σ_{12}	$\sigma_{12}\rho$	$\sigma_{12}\rho^2$
$\sigma_1^2\rho$	σ_1^2	$\sigma_1^2\rho$	$\sigma_{12}\rho$	σ_{12}	$\sigma_{12}\rho$
$\sigma_1^2\rho^2$	$\sigma_1^2\rho$	σ_1^2	$\sigma_{12}\rho^2$	$\sigma_{12}\rho$	σ_{12}
σ_{12}	$\sigma_{12}\rho$	$\sigma_{12}\rho^2$	σ_2^2	$\sigma_2^2\rho$	$\sigma_2^2\rho^2$
$\sigma_{12}\rho$	σ_{12}	$\sigma_{12}\rho$	$\sigma_2^2\rho$	σ_2^2	$\sigma_2^2\rho$
$\sigma_{12}\rho^2$	$\sigma_{12}\rho$	σ_{12}	$\sigma_2^2\rho^2$	$\sigma_2^2\rho$	σ_2^2

the group of subjects), while all four serum creatinine equations captured the iGFR time trend reasonably well.

For the present purpose, the analysis had a few interesting features. Although the 11 eGFRs were fairly highly correlated with each other for this group of measurements, they differed considerably in the ease with which restricted likelihood (as implemented in MIXED) fit the model to them, particularly for subsets of the data, e.g., excluding the 27 people who received continuous steroid treatment. For some eGFRs, the model fit happily with MIXED's defaults, while for other eGFRs the maximizing routine would not converge unless we specified particular starting values and allowed a large number of iterations, which makes me wonder whether there were multiple maxima. (We didn't find any.) Obvious tricks like centering covariates helped for some analyses but for some analyses that converged readily without centering, centering made it harder to get convergence.

Anybody with any experience knows that fitting this kind of model is a fussy business, but I am not aware of any detailed attempt to explain that fussiness. It is well-known that in the simplest version of this model — clustered data with a random effect for a cluster and compound symmetry correlation structure for the errors — the random effect variance and error correlation are not identified. The present model should be identified because we used AR(1) errors instead of compound-symmetric errors, but with relatively short series for each person, no one can say how close this model is to being non-identified. One sign of difficulty is that many of these fits gave estimates of **G** with correlations of 1. In fact, the TYPE=UN@AR1 specification in MIXED's RANDOM statement, which specifies the model for **G**, gave correlations greater than 1 in absolute value — before these analyses, I hadn't known MIXED would allow that — so we had to use the TYPE=UNR@AR1 specification to avoid illegal estimates for random-effect correlations.

Unfortunately, I cannot make Ward's dataset available. However, I get estimates on the boundary of legal values almost every time I fit a model in which two random effects are correlated, so you should have little trouble finding examples of your own. This is akin to the problem of zero variance estimates but it turns out to be harder to think about. I have made no progress on this example either and will not return to it.

The point of these examples — and of this book as a whole — is that we now have a tremendous ability to fit models of the form (1.1), but we have little understanding

of the resulting fits. Parts III and IV of this book are one attempt to begin developing a theory that would provide the necessary understanding, but it is only the barest beginning.

1.2 Conventional Analysis of the Mixed Linear Model

This section is not intended to be an exhaustive or even particularly full presentation of conventional analyses of the mixed linear model, which center on the restricted likelihood. More detailed treatments are available in many books, for example Searle et al. (1992); Snijders & Bosker (2012), which focuses on hierarchical (multi-level) models; Verbeke & Molenberghs (1997), which focuses on SAS's MIXED procedure; Ruppert et al. (2003), which focuses on penalized splines; and Diggle et al. (1994), which focuses on longitudinal analyses. Fahrmeier & Tutz (2010) consider a variety of analysis methods for a broad range of models that all have error distributions in the exponential family.

This section's purpose, besides defining notation and other things, is to emphasize three points:

- The theory and methods of mixed linear models are strongly connected to the theory and methods of linear models, though the differences are important;
- The restricted likelihood is the posterior distribution from a particular Bayesian analysis; and
- Conventional and Bayesian analyses are incomplete or problematic in many respects.

Regarding the first of these, we have an immense collection of tools and intuition from linear models and we should use them as much as possible in developing a theory of mixed linear models. Currently researchers appear to do this inconsistently; for example, ideas of outliers and influence have been carried over to mixed linear models but not collinearity, which features prominently in the present book's Part III.

Regarding incomplete or problematic features of the approaches, I emphasize these because so many people seem either unaware of or complacent about them. For the conventional analysis, Ruppert et al. (2003) give the best treatment of these problems that I know of, while most textbooks seem to ignore them, Snijders & Bosker (2012) being an exception. I often summarize and refer to Ruppert et al. (2003) rather than repeat them, and otherwise I emphasize problems that seem to receive little or no attention elsewhere.

1.2.1 Overview

If an ordinary linear model with a single homoscedastic error term (constant variance) is fit using ordinary least squares, the unknown error variance has no effect on estimates of the mean structure, i.e., the regression coefficients. Thus, the linear model literature emphasizes estimation, testing, etc., for the mean structure with the error variance being a secondary matter. This is not the case in mixed linear models. Estimates of the unknowns in the mean structure — and I include in this the random

effect $\mathbf{u}$ — depend on the unknowns ϕ in the random-effect covariance $\mathbf{G}$ and the error covariance $\mathbf{R}$.

The conventional analysis usually proceeds in three steps: Estimate ϕ; treating that estimate as if it is known to be true, estimate β and $\mathbf{u}$; and then compute tests, confidence intervals, etc., again treating the estimate of ϕ as if it is known to be true. This section begins with estimation of β and $\mathbf{u}$, then moves to estimating ϕ, and then on to tests, etc. The obvious problem here — treating the estimate of ϕ as if it is known to be true — is well known and is discussed briefly below.

1.2.2 Mean Structure Estimates

Mixed linear models have a long history and have accumulated a variety of jargon reflecting that history. In the traditional usage — which is still widely used (e.g., Ruppert et al. 2003) — one *estimates* the fixed effects β but *predicts* the random effects $\mathbf{u}$. This distinction is absent in Bayesian analyses of mixed linear models and appears to be losing ground in conventional non-Bayesian analyses. I maintain this distinction when it seems helpful to do so.

Contemporary conventional analyses can be interpreted (and I do so) as using a unified approach to estimating fixed effects and predicting random effects based on the likelihood function that results from treating ϕ as if it is known (except that, as we will see, this is not really a likelihood, at least by today's definition). The presentation here assumes normal distributions for errors and random effects but the approach is the same when other distributions are used.

To begin, then, we have the mixed linear model as in (1.1),

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \varepsilon, \qquad \mathbf{u} \sim N_q(0, \mathbf{G}(\phi_G)), \qquad \varepsilon \sim N_n(0, \mathbf{R}(\phi_R)). \tag{1.4}$$

In the conventional view, the random variables here are $\mathbf{u}$ and ε, but by implication $\mathbf{y}$ is also a random variable. Customarily, then, the analysis proceeds by writing the joint density of $\mathbf{y}$ and $\mathbf{u}$ as

$$f(\mathbf{y}, \mathbf{u}|\beta, \phi) = f(\mathbf{y}|\mathbf{u}, \beta, \phi_R) f(\mathbf{u}|\phi_G), \tag{1.5}$$

where f will be used generically to represent a probability density, in this case Gaussian densities. Taking the log of both sides gives

$$\begin{aligned} \log f(\mathbf{y}, \mathbf{u}|\beta, \phi) \quad = \quad & K - \frac{1}{2}\log|\mathbf{R}(\phi_R)| - \frac{1}{2}\log|\mathbf{G}(\phi_G)| \\ & -\frac{1}{2}\left\{(\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{u})'\mathbf{R}(\phi_R)^{-1}(\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{u}) + \mathbf{u}'\mathbf{G}(\phi_G)\mathbf{u}\right\}, \end{aligned} \tag{1.6}$$

where K is an unimportant constant.

Readers may note the awkwardness inherent in (1.6). It is tempting to call this a likelihood. However, the usual likelihood arises as the probability or probability density of an observed random variable treated as a function of unknown parameters, while in (1.6), the random variable $\mathbf{u}$ is unobserved. Lee et al. (2006, Chapters 1 to 4) try valiantly to resolve this difficulty in a way that allows them to say they do

not use prior distributions. However, I am persuaded by other work (Bayarri et al. 1988, Bayarri & DeGroot 1992) that the likelihood and prior distributions cannot be cleanly distinguished so that this difficulty is inherent and cannot be resolved tidily without adopting a Bayesian approach, in which the distinction is not important; see, e.g., Rappold et al. (2007) for a *bona fide* scientific application. Perhaps in the long run statisticians will not be so fussy about labeling things as likelihoods or priors. (This is arguably true already for conventional likelihood-based analyses of state-space models; see Chapter 6.) The conventional analysis, as practiced today at least, is already one step in this direction: In effect, it ignores this difficulty and treats (1.6) as if it were a likelihood function for the unknowns β and $\mathbf{u}$, with ϕ treated as if it were known. However, it is worth remembering that as of today, (1.6) is not considered a likelihood.

If we now maximize this not-really-a-likelihood ("quasi-likelihood" and "pseudo-likelihood" having already been claimed), then β and $\mathbf{u}$ are estimated by minimizing

$$\underbrace{\left[\mathbf{y}-\mathbf{C}\begin{pmatrix}\beta\\ \mathbf{u}\end{pmatrix}\right]'\mathbf{R}^{-1}\left[\mathbf{y}-\mathbf{C}\begin{pmatrix}\beta\\ \mathbf{u}\end{pmatrix}\right]}_{\text{"likelihood"}}+\underbrace{[\beta|\mathbf{u}]\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}^{-1}\end{pmatrix}\begin{bmatrix}\beta\\ \mathbf{u}\end{bmatrix}}_{\text{"penalty"}} \tag{1.7}$$

where the $\mathbf{0}$ are matrices of zeroes of the appropriate dimensions, I've suppressed the dependence on ϕ for simplicity, and $\mathbf{C}=[\mathbf{X}|\mathbf{Z}]$, where this notation denotes concatenation of the columns of $\mathbf{X}$ and $\mathbf{Z}$.

Equation (1.7) introduces two jargon terms commonly used in the conventional theory of mixed linear models and generalizations of it. Equation (1.7) is sometimes called a "penalized likelihood," where the first term fills the role of the likelihood — and it *would* be a likelihood, if $\mathbf{u}$ were simply a vector of unknown regression coefficients — and the second term, the "penalty," changes the likelihood by adding a penalty for values of $\mathbf{u}$ that are large relative to $\mathbf{G}$. The idea of a penalized likelihood is to allow a rich, flexible parameterization of the mean structure through a high-dimensional $\mathbf{u}$, but to avoid overfitting by pushing $\mathbf{u}$'s elements toward zero by means of the penalty. Here is more jargon: Pushing $\mathbf{u}$'s elements toward zero in this manner is called "shrinkage," a term that originated with Charles Stein's work in the late 1950s. Shrinkage is a key aspect of analyses of richly parameterized models and often recurs in later chapters.

It is easy to show that this maximization problem produces these point estimates for β and $\mathbf{u}$:

$$\begin{bmatrix}\tilde{\beta}\\ \tilde{\mathbf{u}}\end{bmatrix}_{\phi}=\left[\mathbf{C}'\mathbf{R}^{-1}\mathbf{C}+\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}^{-1}\end{pmatrix}\right]^{-1}\mathbf{C}'\mathbf{R}^{-1}\mathbf{y} \tag{1.8}$$

where the tildes above β and $\mathbf{u}$ mean these are estimates and the subscript ϕ indicates that the estimates depend on ϕ. (Derivation of (1.8) is an exercise at the end of the chapter.)

If the term arising from (1.7)'s penalty,

$$\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \end{bmatrix}, \tag{1.9}$$

were omitted from (1.8), $\tilde{\beta}$ and $\tilde{\mathbf{u}}$ would simply be the generalized least squares (GLS) estimates for the assumed value of ϕ_R. The penalty alters the GLS estimates of β and $\mathbf{u}$ by adding $\mathbf{G}^{-1}$ to the precision matrix for $\mathbf{u}$ from an ordinary regression, $\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}$. In this sense, the penalty term provides an extra piece of information about $\mathbf{u}$, namely $\mathbf{u} \sim N_q(\mathbf{0}, \mathbf{G})$.

The estimates (1.8) give fitted values:

$$\tilde{\mathbf{y}}_\phi = \mathbf{C} \begin{bmatrix} \tilde{\beta} \\ \tilde{\mathbf{u}} \end{bmatrix}_\phi \tag{1.10}$$

$$= \mathbf{C} \left[\mathbf{C}'\mathbf{R}^{-1}\mathbf{C} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \end{pmatrix} \right]^{-1} \mathbf{C}'\mathbf{R}^{-1}\mathbf{y}. \tag{1.11}$$

For a given value of ϕ, the fitted values $\tilde{\mathbf{y}}_\phi$ are simply the observations $\mathbf{y}$ pre-multiplied by a known square matrix.

In traditional usage, the estimates (1.8) of $\mathbf{u}$ are called the best linear unbiased predictors or predictions (BLUPs). In this usage, "bias" refers to hypothetical repeated sampling from the distributions of both $\mathbf{u}$ and ε. For new-style random effects, hypothetical repeated sampling over the distribution of $\mathbf{u}$ is often plainly meaningless, and BLUPs are in fact biased over repeated sampling from ε. In such cases, the term BLUP is somewhat misleading. This issue is treated at length in Chapters 3 and 13.

The original derivation of the estimates (1.8) used a two-step procedure, first estimating the fixed effects β and then using those estimates to "predict" the random effects $\mathbf{u}$. Specifically, the mixed linear model (1.1) was re-written by making $\mathbf{Zu}$ part of the error, so that $\mathbf{y} = \mathbf{X}\beta + \varepsilon^*$, where $\varepsilon^* = \mathbf{Zu} + \varepsilon$, so that $\mathbf{y} \sim N_n(\mathbf{X}\beta, \mathbf{V})$ for $\mathbf{V} = \text{cov}(\varepsilon^*) = \mathbf{ZGZ}' + \mathbf{R}$. For given $\mathbf{V}$, i.e., for given ϕ, the standard textbook estimator of β is the GLS estimator $\tilde{\beta}$. The so-called best linear prediction of $\mathbf{u}$ is then $\mathbf{GZ}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta)$ (Ruppert et al. 2003, Sections 4.4, 4.5), and when the unknown β is replaced by its estimate $\tilde{\beta}$, the resulting "prediction" of $\mathbf{u}$ is $\tilde{\mathbf{u}}$ as in (1.8).

More traditional jargon: The foregoing takes $\mathbf{G}(\phi_G)$ and $\mathbf{R}(\phi_R)$ as given. If an estimate of ϕ is plugged into (1.8), the resulting estimates of β and $\mathbf{u}$ are called "estimated BLUPs" or EBLUPs.

1.2.3 Estimating ϕ, the Unknowns in $\mathbf{G}$ *and* $\mathbf{R}$

1.2.3.1 Maximizing the, or rather a, likelihood

Perhaps the first thing that would come to most people's minds — at least, those who don't consider themselves Bayesians — would be to write down the likelihood and maximize it. In simpler problems, the likelihood is a function of all the unknowns in the model, but as noted in the previous section, if $\mathbf{u}$ is considered among the

unknowns, then it is unclear how the likelihood is defined for the mixed linear model or indeed whether it can be.

One pragmatic way to avoid this quandary is to get rid of the random effects $\mathbf{u}$ as in the traditional derivation of the point estimate $\tilde{\beta}$ shown in the previous section. With this approach, the model for the data $\mathbf{y}$ is

$$\mathbf{y} \sim N_n(\mathbf{X}\beta, \mathbf{V}(\phi)) \text{ where } \mathbf{V}(\phi) = \mathbf{Z}\mathbf{G}(\phi_G)\mathbf{Z}' + \mathbf{R}(\phi_R). \tag{1.12}$$

Having disposed of $\mathbf{u}$, (1.12) provides a well-defined likelihood that is a function of (β, ϕ) and which can be maximized to give estimates of β and ϕ.

Unfortunately, maximizing this likelihood produces a point estimate of ϕ with known flaws that can be quite severe. The simplest case is familiar from a first course in statistics. If $X_1, \ldots, X_n$ are independent and identically distributed (iid) as $N(\mu, \sigma_e^2)$ with unknown μ and σ_e^2, σ_e^2 has maximum likelihood estimate (MLE) $\sum(X_i - \bar{X})^2/n$, where $\bar{X}$ is the average of the X_i. This MLE is biased as an estimate of σ_e^2, with expected value $(n-1)\sigma_e^2/n$. This bias becomes trivial as n grows, but the same is not true for some simple elaborations of this problem, as Neyman and Scott showed in 1948 with the following example. If X_{i1} and X_{i2} are iid $N(\alpha_i, \sigma_e^2)$ for $i = 1, \ldots, n$, the MLE for σ_e^2 is $\sum_i (X_{i1} - X_{i2})^2/4n$ and has expected value $\sigma_e^2/2$ for all n. If α_i is treated as a random effect with variance σ_s^2, making this problem a mixed linear model, then the MLE of σ_e^2 becomes unbiased but now the MLE of σ_s^2 is biased. (Proofs are given as exercises.) In general, the problem is that the MLE of ϕ does not account for degrees of freedom — in this usage, linearly independent functions of the data $\mathbf{y}$ — that are used to estimate fixed effects.

1.2.3.2 Maximizing the Restricted (Residual) Likelihood

Dissatisfaction with the MLE's bias, among other things, prompted a search for alternatives that were unbiased, at least for the problems that were within reach at the time. This work, beginning in the 1950s, led to the present-day theory of the restricted likelihood, also called the residual likelihood. Various rationales have been given for the restricted likelihood, each providing some insight into how it differs from the likelihood in (1.12). Those seeking more detail about the rationales summarized here could start with Searle et al. (1992), Section 6.6, and references given there.

One rationale is that the restricted likelihood summarizes information about ϕ using the residuals from fitting the model $\mathbf{y} = \mathbf{X}\beta +$ [iid error] by ordinary least squares (hence "residual likelihood"). A closely related definition of the restricted likelihood is the likelihood arising from particular linear combinations of the data, $\mathbf{w} = \mathbf{K}'\mathbf{y}$, where $\mathbf{K}$ is a known, full-rank $n \times (n - \text{rank}(\mathbf{X}))$ matrix chosen so that $\mathbf{w}$'s distribution is independent of β. One such $\mathbf{K}$ has columns forming an orthonormal basis for the orthogonal complement of the column space of $\mathbf{X}$, i.e., the space of residuals from a least squares fit of $\mathbf{y}$ on $\mathbf{X}$. In general many such $\mathbf{K}$ exist but the likelihood of $\mathbf{K}'\mathbf{y}$ ("restricted likelihood") is invariant to the choice of $\mathbf{K}$ (Searle et al. 1992, Section 6.6). If the model for $\mathbf{y}$ is written as in equation (1.12), then for any $\mathbf{K}$, $\mathbf{w} = \mathbf{K}'\mathbf{y}$ is distributed as

$$\mathbf{w} = \mathbf{K}'\mathbf{y} \sim N\left(\mathbf{0}, \mathbf{K}'\mathbf{V}(\phi)\mathbf{K}\right). \tag{1.13}$$

where the random variable $\mathbf{w}$ and thus the covariance matrix $\mathbf{K}'\mathbf{V}(\phi)\mathbf{K}$ have dimension $n - \text{rank}(\mathbf{X})$.

Finally, the restricted likelihood can be derived as a particular "marginal likelihood." Begin with the likelihood arising from (1.12). This likelihood is a function of (β, ϕ); integrate β out of this likelihood to give a function of ϕ. The result, which we now explore in some detail, is the restricted likelihood. This integral has a Bayesian interpretation, to which I return below.

The likelihood arising from (1.12) is

$$L(\beta, \mathbf{V}) = K|\mathbf{V}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta)\right). \tag{1.14}$$

To integrate out β, expand the quadratic form in the exponent and complete the square (I am showing this instead of leaving it for an exercise because this technique is so useful):

$$\begin{aligned}(\mathbf{y} - \mathbf{X}\beta)'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta) &= \mathbf{y}'\mathbf{V}^{-1}\mathbf{y} + \beta'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\beta - 2\beta'\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\ &= \mathbf{y}'\mathbf{V}^{-1}\mathbf{y} - \tilde{\beta}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\tilde{\beta} + (\beta - \tilde{\beta})'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}(\beta - \tilde{\beta}),\end{aligned}$$

where $\tilde{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, the generalized least-squares estimate given $\mathbf{V}$. Integrating β out of (1.14) is then just the integral of a multivariate normal density, so

$$\int L(\beta, \mathbf{V})d\beta = K|\mathbf{V}|^{-\frac{1}{2}}|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left[\mathbf{y}'\mathbf{V}^{-1}\mathbf{y} - \tilde{\beta}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\tilde{\beta}\right]\right). \tag{1.15}$$

The natural logarithm of (1.15) is almost always used. Taking the log of (1.15) and substituting the expression for $\tilde{\beta}$ gives the log restricted likelihood

$$RL(\phi|\mathbf{y}) = K - 0.5\left(\log|\mathbf{V}| + \log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| + \mathbf{y}'\left[\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}\mathbf{V}^{-1}\right]\mathbf{y}\right), \tag{1.16}$$

where $\mathbf{V} = \mathbf{Z}\mathbf{G}(\phi_G)\mathbf{Z}' + \mathbf{R}(\phi_R)$ is a function of the unknown ϕ in the covariance matrices $\mathbf{G}$ and $\mathbf{R}$. (See, for example, Searle et al. 1992, Sections 8.3f and M.f.)

Within likelihood theory, this marginalizing integral is quite *ad hoc* and ultimately is justified only by the desirable properties of the resulting point estimates. This integral has a much more natural Bayesian interpretation: If the fixed effects β have an improper flat prior $\pi(\beta) \propto 1$, then (1.15), multiplied by a prior distribution for ϕ, is just the marginal posterior distribution of ϕ. Within Bayesian theory, there is nothing inherently unclean about this marginalizing integral although the improper flat prior on β can have striking consequences, some of which are pertinent to matters discussed in this book and will be noted when they arise.

1.2.4 Other Machinery of Conventional Statistical Inference

Now we have point estimates $\hat{\phi}$ of ϕ from maximizing the restricted likelihood; these are plugged into (1.8) to give estimates $(\hat{\beta}, \hat{\mathbf{u}})$ of $(\beta, \mathbf{u})$ which in turn are plugged into (1.10) to give fitted values $\hat{\mathbf{y}}$. I use hats $\hat{\bullet}$ instead of tildes $\tilde{\bullet}$ here to indicate that these

estimates and fitted values are computed using $\hat{\phi}$. These $(\hat{\beta},\hat{\mathbf{u}})$ and $\hat{\mathbf{y}}$ were derived from linear model theory by treating $\hat{\phi}$ as if it were the true value of ϕ and the rest of the conventional theory is derived from linear model theory the same way. These derivations are straightforward, so they have been left as exercises.

I believe it is not controversial to say that the conventional approach to analyzing mixed linear models has an unsatisfactory feature that is probably impossible to fix. Except for some special cases, there is no feasible alternative to treating the estimates $\hat{\phi}$ as if they were the true values of the unknowns in $\mathbf{G}$ and $\mathbf{R}$. This assumes away the usually large variation in $\hat{\phi}_G$ (or uncertainty about ϕ_G, if you prefer) and the often non-trivial variation in $\hat{\phi}_R$. Among the dissatisfied, some consider Bayesian analyses a promising alternative now that they can be computed readily (e.g., Ruppert et al. 2003, Section 4.7, p. 102). Thus, I give just a brief overview of the conventional machinery as implemented in widely used software, leaning heavily on Ruppert et al. (2003).

1.2.4.1 Standard Errors for Fixed and Random Effects

Standard errors for the fixed-effect estimates $\hat{\beta}$ can be derived from the model (1.12) in which the random effects $\mathbf{u}$ have been incorporated into the error term. Under that model, $\mathbf{y} \sim N_n(\mathbf{X}\beta, \mathbf{V}(\phi))$ where $\mathbf{V}(\phi) = \mathbf{Z}\mathbf{G}(\phi_G)\mathbf{Z}' + \mathbf{R}(\phi_R)$. If ϕ is set to $\hat{\phi}$ and treated as known, giving $\hat{\mathbf{V}}$, then $\hat{\beta}$ is the familiar GLS estimator $\hat{\beta} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$, which has the familiar covariance matrix $\mathrm{cov}(\hat{\beta}) = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$. This covariance — which is exact if $\mathbf{V}$ is known, and approximate otherwise (Section 1.2.4.2 below has more to say on this) — provides the standard errors for $\hat{\beta}$ given in, for example, SAS's MIXED procedure: $SE(\hat{\beta}_i) = \mathrm{cov}(\hat{\beta})_{ii}^{0.5}$, where the subscript ii indicates the i^{th} diagonal element of $\mathrm{cov}(\hat{\beta})$. (The same standard errors are obtained from (1.19) below.)

To see the deficiency of this usual standard error, note that

$$\mathrm{cov}(\hat{\beta}) = E(\mathrm{cov}\{\hat{\beta}|\phi\}) + \mathrm{cov}(E\{\hat{\beta}|\phi\}), \tag{1.17}$$

where the outer expectation and covariance are with respect to the distribution of $\hat{\phi}$. By the familiar fact that generalized least squares estimates are unbiased even if the covariance matrix is wrong, $E\{\hat{\beta}|\phi\} = \beta$, so $\mathrm{cov}(E\{\hat{\beta}|\phi\}) = 0$. The usual standard error for $\hat{\beta}$, given above, amounts to approximating the first term in (1.17) by $\mathrm{cov}\{\hat{\beta}|\hat{\phi}\}$. This deficiency is well known, e.g., it is noted in the documentation for the MIXED procedure.

Considering both the fixed and random effects simultaneously, recall from (1.8) that the estimates (EBLUPs) of the fixed and random effects are

$$\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}} \end{bmatrix} = \left[\mathbf{C}'\mathbf{R}(\hat{\phi}_R)^{-1}\mathbf{C} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}(\hat{\phi}_G)^{-1} \end{pmatrix}\right]^{-1} \mathbf{C}'\mathbf{R}(\hat{\phi}_R)^{-1}\mathbf{y}, \tag{1.18}$$

where, as before, $\mathbf{C} = [\mathbf{X}|\mathbf{Z}]$ and the estimates now have hats instead of tildes because they depend on $\hat{\phi}$. It will be useful to have two different covariances for $(\hat{\beta},\hat{\mathbf{u}})$. The

first involves the unconditional covariance of $\hat{\mathbf{u}} - \mathbf{u}$:

$$\operatorname{cov}\begin{pmatrix} \hat{\beta} \\ \hat{\mathbf{u}} - \mathbf{u} \end{pmatrix} = \left[\mathbf{C}'\mathbf{R}(\hat{\phi}_R)^{-1}\mathbf{C} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}(\hat{\phi}_G)^{-1} \end{pmatrix}\right]^{-1}, \tag{1.19}$$

This covariance is with respect to the distributions of both ε and $\mathbf{u}$ and is used to give standard errors for the random effects in SAS's MIXED procedure. (The derivation is left as an exercise.)

The second covariance conditions on $\mathbf{u}$. At this point in the book, it may seem odd to condition on $\mathbf{u}$ but it will seem more natural when we consider methods in which $\mathbf{u}$ is a new-style random effect, that is, an unknown parameter that happens to be constrained (by $\mathbf{G}$). This covariance has a key role in Chapter 3's discussion of confidence bands for penalized-spline fits. The covariance conditional on $\mathbf{u}$ is

$$\begin{aligned} \operatorname{cov}\left(\begin{array}{c|c} \hat{\beta} & \\ \hat{\mathbf{u}} & \mathbf{u} \end{array}\right) &= \left[\mathbf{C}'\mathbf{R}(\hat{\phi}_R)^{-1}\mathbf{C} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}(\hat{\phi}_G)^{-1} \end{pmatrix}\right]^{-1} \mathbf{C}'\mathbf{R}^{-1}\mathbf{C} \\ &\quad \times \left[\mathbf{C}'\mathbf{R}(\hat{\phi}_R)^{-1}\mathbf{C} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}(\hat{\phi}_G)^{-1} \end{pmatrix}\right]^{-1}. \end{aligned} \tag{1.20}$$

This expression is derived trivially by noting that conditional on $\mathbf{u}$, $\operatorname{cov}(\mathbf{y}) = \mathbf{R}$.

1.2.4.2 Testing and Intervals for Fixed and Random Effects

More complex tests involving the fixed and random effects use the covariances just given in Section 1.2.4.1 and again, the standard tests fix ϕ at its estimate $\hat{\phi}$ and treat it as if known. Thus, for example, SAS's test for whether $l'\beta$, a linear combination of β, is zero is the Wald test based on

$$z = \frac{l'\hat{\beta}}{(l'\operatorname{cov}(\hat{\beta})l)^{0.5}} \tag{1.21}$$

which is treated as if it is approximately distributed as $N(0,1)$. The 95% Wald confidence interval is constructed by replacing the numerator of z in (1.21) with $l'\hat{\beta} - l'\beta$.

The distribution in (1.21) is exact if ϕ is known, which, of course, it never is. On this point, Ruppert et al. (2003, p. 104) issue a sober warning:

> [T]he theoretical justification of [(1.21)] for general mixed models is somewhat elusive owing to the dependence in $\mathbf{y}$ imposed by the random effects. Theoretical back-up for [(1.21)] exists in certain special cases, such as those arising in analysis of variance and longitudinal data analysis. . . . For some mixed models, including many used in the subsequent chapters of this book [and the present book as well], justification of [(1.21)] remains an open problem.

In other words, use this at your own risk. Ruppert et al. (2003) note that as in simpler problems, the likelihood ratio test provides an alternative to the Wald test (1.21), but they also note that as for the Wald test, "its justification . . . is dependent on the type of

correlation structure induced by the **G** and **R** matrices" (Ruppert et al. 2003, p. 105), i.e., this doesn't solve the problem.

For linear functions of both β and **u**, similar Wald tests can be constructed using the joint covariance of β and **u** given in Section 1.2.4.1. Similarly, tests for entire effects, e.g., for the 4 degrees of freedom corresponding to a 5-category grouping variable, can be done as Wald tests (above) or as likelihood ratio tests. However, these tests have no better theoretical rationale than the test above that only involves β.

As an alternative to relying on large-sample approximations, Ruppert et al. (2003) propose a parametric bootstrap to produce reasonably accurate small-sample confidence intervals and P-values (e.g., p. 144 for fitted values).

1.2.4.3 Testing and Intervals for ϕ

In many applications of mixed linear models, the unknowns in ϕ are nuisances and there is little interest in estimates or intervals for them. In such cases, the only interesting question is whether the variance of a random effect, σ_s^2, is zero because if so, it can be omitted from the model. For example, in Example 2 about imaging vocal folds, it would be convenient if one of the variance components was zero, because removing that component would make the analysis simpler and better-behaved. Taken literally, of course, it seems obvious that none of these variance components could be exactly zero, but the same objection can be made about any sharp null hypothesis.

For testing whether $\sigma_s^2 = 0$, the conventional theory offers little beside likelihood ratio or restricted likelihood ratio tests, in which the maximized (restricted) likelihoods are compared for an unrestricted fit and a fit with σ_s^2 fixed at zero (e.g., Ruppert et al. 2003, Section 4.8). The usual large-sample approximate chi-square distribution for the (restricted) likelihood ratio test does not apply because the null hypothesis $\sigma_s^2 = 0$ is on the boundary of the parameter space. An asymptotic argument appropriate to the situation shows that in large samples, the (restricted) likelihood ratio statistic has a null distribution that is approximately an equal mixture of a point mass at zero and the chi-square distribution from the usual asymptotic theory. Unfortunately, this is well-known to be a poor approximation in mixed linear models for the same reasons discussed in the preceding section.[2] Thus Ruppert et al. (2003, p. 107), among others, recommend against using this asymptotic test and suggest instead a parametric bootstrap P-value.

Sometimes, however, the unknowns in ϕ are interesting in themselves. In Example 1, measuring the virus's structure, the elements of ϕ describe different components of measurement error and information about the variances of those components might guide design of future experiments or an effort to improve the measurement process. Again, the conventional theory offers little more than standard large-sample theory for maximum likelihood and its extension to restricted likelihood: The approximate covariance matrix for $\hat{\phi}$ is the inverse of -1 times the matrix of second derivatives of the log likelihood or restricted likelihood, evaluated at $\hat{\phi}$. See, for ex-

[2]However, references given in Snijders & Bosker (2012, Section 6.2.1) show that this approximate test has reasonably good properties for the special case of variances of intercepts and slopes in multi-level models, sometimes called random regressions.

ample, Searle et al. (1992), Section 6.3 for maximum likelihood and Section 6.6e for restricted likelihood. SAS's MIXED procedure uses this approach to provide standard errors for estimates of the unknowns in ϕ, and uses it to compute a z-statistic and P-value for testing whether each element of ϕ is zero.

The resulting covariance matrix and standard errors are, as far as I can tell, universally regarded as poor approximations; Ruppert et al. (2003) do not consider them worth mentioning. Even SAS's own documentation for MIXED says "tests and confidence intervals based on asymptotic normality can be obtained. However, these can be unreliable in small samples [and, of course, nobody knows what "small sample" means, though SAS is hardly alone in giving this sort of useless advice], especially for parameters such as variance components which have sampling distributions that tend to be skewed to the right" (v. 9.1.3 on-line manual). Other authors are less restrained: Verbeke & Molenberghs (1997) devote their Appendix B to criticizing MIXED's z-statistic and P-value for non-negative elements of ϕ.

One glaring problem with the large-sample approximation on the original scale is that it frequently gives standard errors for variances that are larger than the variance estimates, making Wald-style confidence intervals — estimate plus-or-minus 2 standard errors — generally inappropriate. A common alternative confidence interval uses a Satterthwaite approximation. As implemented in SAS's MIXED procedure (Milliken & Johnson 1992, p. 348), the Satterthwaite approximate confidence interval for a variance σ_s^2 is

$$\frac{\nu\hat{\sigma}_s^2}{\chi^2_{\nu,1-\alpha/2}} \leq \sigma_s^2 \leq \frac{\nu\hat{\sigma}_s^2}{\chi^2_{\nu,\alpha/2}}, \tag{1.22}$$

where the denominators are quantiles of a chi-squared distribution with degrees of freedom $\nu = 2z^2$ for $z = \hat{\sigma}_s^2/(\text{standard error } \hat{\sigma}_s^2)$. In MIXED, $\hat{\sigma}_s^2$ is obtained by maximizing the likelihood or restricted likelihood and its standard error is from the large-sample approximation. The Satterthwaite interval is derived by assuming z^2 is approximately distributed as chi-squared with r degrees of freedom, matching the variances of z^2 and the chi-squared distribution, and solving for r.

In some simple cases this distribution is exact and for some cases where it is approximate, Milliken & Johnson (1992) describe simulation experiments in which it performs well. However, they note (p. 349) "when there are very few levels [in the sense of ANOVA] associated with a random effect and consequently very few degrees of freedom [in the sense of independent pieces of information, not ν], the resulting confidence intervals are going to be very wide." Indeed, in my experience this approximation routinely gives intervals ranging effectively from zero to infinity, particularly when the variance estimate is barely positive. In fairness, such nonsensical results should probably be taken to mean "the approximation is not appropriate in this case."

The mixed linear model can be parameterized using log variances instead of variances and the large-sample approximation will undoubtedly work better with this parameterization. As far as I know, however, no statistical package does this.

1.3 Bayesian Analysis of the Mixed Linear Model

Deriving the conventional analysis involved various *ad hoc* acts. Such *ad hockery* is entirely consistent with the frequentist approach of deriving procedures by any convenient method and justifying them based on their operating characteristics, e.g., by the reduced bias of maximum restricted likelihood estimates compared to maximum likelihood estimates. As a pragmatist, I have no problem with this approach in general. For mixed linear models, however, we have already seen that the conventional analysis has several deficiencies and we will see more. These problems are bad enough that people who had no apparent interest in Bayesian methods before 1990 now see Bayesian methods as an elegant way to alleviate some problems of the conventional approach (e.g., Ruppert et al. 2003, Chapter 16). In this sense, those who are unconvinced by Bayesian ideology may view Bayesian methods as simply procedures derived using a particular general approach, which must be justified by their operating characteristics like any other procedure. This section presents Bayesian methods for mixed linear models from this viewpoint. Like the preceding section, this section is not intended to be exhaustive but is rather intended to define things and to emphasize aspects in which Bayesian analysis is incomplete, incompletely understood, or problematic.

1.3.1 A Very Brief Review of Bayesian Analysis in General

The Bayesian approach was for a long time considered revolutionary and thus disreputable to those opposed to it, and it still carries the lingering odors of both revolutionary zeal and disrepute. Thus, any treatment of it must begin with some kind of manifesto; this is mine. If you are unfamiliar with Bayesian analysis, you will almost certainly need to spend some time with a more general introduction, e.g., the early chapters of Carlin & Louis (2008) or Gelman et al. (2004), where you can read those authors' manifestos.

The fundamental idea of Bayesian analysis is to describe all uncertainty and variation using probability distributions. This contrasts with the conventional approach in the preceding section: There, for example, we did not know the values of the fixed effects β and although we produced probability statements about an *estimator* of β, $\hat{\beta}$, we made no probability statements about β itself. By contrast, a Bayesian analysis assigns probability distributions to entities like β that are conceived to be fixed in value but unknown. A Bayesian analysis also assigns probability distributions to observable quantities like $\mathbf{y}$ conditional on unknowns such as β, $\mathbf{u}$, and ϕ_R, just as in the conventional approach. Assigning a joint probability distribution to all observed and unobserved quantities allows a Bayesian analysis to proceed using *only* the probability calculus, with particular reliance on one theorem, Bayes's Theorem.

Using probability this way brings an undeniable theoretical tidiness and, with modern Bayesian computing, real practical advantages in certain problems. It has also made possible a great deal of confusion, some of which is especially relevant to contemporary uses of mixed linear models. To a person trained in conventional statistics, probability describes one type of thing — variation between replicate observations of a random mechanism (often hypothetical), like a coin flip — and "the

probability of the event A" describes the long-run frequency of the occurence of event A in a sequence of replicate observations from the random mechanism. In addition to this, Bayesian analysis uses probability to describe uncertainty about unknown and unobservable things which, like β, may have no real existence and which cannot in any sense be conceived as arising from a random mechanism, let alone one that can produce replicate observations. The random effects $\mathbf{u}$ have an ambiguous status between these two extremes: The traditional definition conceives of them as arising from a random mechanism like coin flips, but they are inherently unobservable. The newer varieties of random effect add to this difficulty because, as Chapter 13 argues in detail, their probability distributions cannot meaningfully be understood as describing variation between replicate realizations of a random mechanism.

One school of Bayesian thinking avoids this conceptual difficulty by arguing that the conventional notion of probability, as a long-run frequency, is so inconsistent with itself and with reality that it must be discarded. The leading figure in this school was Bruno de Finetti, whose magnum opus, *Theory of Probability* (1974–75, English translation by Antonio Machi and Adrian Smith), develops the view that probability can only be meaningfully interpreted as a quantification of an individual's personal uncertainty. From this beginning, statistical analysis becomes the process of updating personal beliefs expressed as probability distributions using the probability calculus, Bayes's Theorem in particular. For a long time, de Finetti's view was known to few people because his writing is — deliberately, it seems — nearly impenetrable even in a sympathetic translation and because the necessary calculations were intractable outside of a few simple problems. With the advent of modern Bayesian computing, far more people find Bayesian analyses worth pursuing because of their practical advantages but de Finetti has gained few if any adherents. To compress a complex subject perhaps excessively, it seems that few people find de Finetti's subjective approach relevant to their problems let alone convincing, though his devotees certainly remain devoted. (If you'd like to meet one, just say my previous sentence to a large mixed audience of academic statisticians.)

Indeed, as Kass (2011) argues, these days fewer and fewer people identify themselves as either Bayesian or frequentist but rather identify themselves as practicing statisticians who use tools from both approaches. This arises from the convergence of several trends: The increasing power and ease of data analysis using Bayesian methods; the growing belief that academic statisticians should become deeply involved in applications; the realization that the old Bayes-frequentist dispute was conducted in a theoretical world with no connection to the reality of statistical applications; and, as a consequence, growing acceptance of the view that "[statistical] procedures should be judged according to their performance under theoretical conditions thought to capture relevant real-world variation in a particular applied setting" (Kass 2011, p. 7) and not by the abstract properties with which advocates of the two schools have flogged each other. In particular, "it makes more sense to place in the center of our logical framework the match or mis-match of theoretical assumptions with the real world of data." Kass (2011) is, to my knowledge, the first explicit articulation of a view that he calls "statistical pragmatism" and while his view is, of course, not yet completely fleshed out, I endorse it heartily.

Having disposed of the obligatory manifesto, I now proceed in a manner I consider pragmatic.

Let $\mathbf{w}$ represent observed outcomes and let θ represent any unknown quantities. Very generally, Bayesian analysis uses the following items.

- *The Likelihood*. Denote the probability model for the data $\mathbf{w}$ given the unknowns θ as $\pi(\mathbf{w}|\theta)$, a function of $\mathbf{w}$ with θ treated as fixed (though unknown). The likelihood $L(\theta;\mathbf{w})$ is $\pi(\mathbf{w}|\theta)$ treated as a function of θ; $\mathbf{w}$ is treated as fixed once the data have been observed.
- *The Prior Distribution*. Information about θ external to $\mathbf{w}$ is represented by a probability distribution $\pi(\theta)$. This is called "the prior distribution" in the sense that it represents information or beliefs about θ *prior to* updating those beliefs in light of $\mathbf{w}$ using Bayes's Theorem (see the next item). However, the information summarized in $\pi(\theta)$ need not precede $\mathbf{w}$ in any temporal sense. For example, $\mathbf{w}$ could describe an epidemic that happened in the 1700s and $\pi(\theta)$ might capture present-day knowledge about the vectors that spread this particular disease.
- *Bayes's Theorem*. This standard and uncontroversial theorem states that if A and B are events in the sense of probability theory, $P(A|B) = P(B|A)P(A)/P(B)$, where "$P(A|B)$" means the conditional probability of event A given that the event B is supposed to have happened, while $P(A)$ is the unconditional probability of event A. Another version of this theorem applies to situations in which the events A and B refer to quantities that take values on a continuous scale, where probability densities for those continuous variates are denoted by π: $\pi(\theta|\mathbf{w}) = \pi(\mathbf{w}|\theta)\pi(\theta)/\pi(\mathbf{w})$, where $\pi(\mathbf{w}) = \int \pi(\mathbf{w}|\theta)\pi(\theta)d\theta$.
- *The Posterior Distribution*. Suppose the observable data $\mathbf{w}$ has probability density $\pi(\mathbf{w}|\theta)$ depending on the unknown θ. Suppose further that θ has prior distribution $\pi(\theta)$. (This common wording glosses over the miraculous origin of $\pi(\theta)$, to which we will return.) Then we can use Bayes's theorem to update the information in the prior distribution, giving the posterior distribution $\pi(\theta|\mathbf{w}) \propto L(\theta;\mathbf{w})\pi(\theta)$, where the likelihood $L(\theta;\mathbf{w})$ fills the role of $\pi(\mathbf{w}|\theta)$. Again, "posterior" need not have a temporal meaning; it merely means *after* accounting for the data $\mathbf{w}$.
- *The Fundamental Principle*. In a Bayesian analysis, *all* inferential or predictive statements about θ depend on the data $\mathbf{w}$ *only* through the posterior distribution.

A few comments are in order.

For people trained in conventional statistics, The Fundamental Principle seems like a needless constraint. In conventional statistics, you can use any method you like to devise procedures that give inferential or predictive statements about θ; all that matters is their performance. From a Bayesian viewpoint, however, procedures that do not conform to The Fundamental Principle should be avoided because they are subject to technical defects, called *incoherence* in aggregate, while Bayesian procedures are not (if they use proper priors, and in special cases if they use improper priors). In practice, even people who call themselves Bayesian use all manner of non-conforming methods for model building, i.e., for selecting $\pi(\mathbf{w}|\theta)$, and leading advocates of Bayesian methods have sanctioned this (e.g., Smith 1986) even though in theory one could use only Bayesian methods even for model-building. One rea-

son they sanction non-Bayesian methods for model building is that using Bayesian methods for this purpose is like playing the piano with your feet, in most if not all cases. Since the advent of modern Bayesian computing, one rarely hears objections to using non-Bayesian methods to specify $\pi(\mathbf{w}|\theta)$.

For decades, while the Foundations of Statistics were a live topic of debate, anti-Bayesians criticized the prior distribution as introducing a needless element of subjectivity into the analysis. Those toeing de Finetti's line countered that the prior *adds* no subjectivity because objectivity in statistical analysis is a delusion. More conciliatory pro-Bayesians argued that it is usually possible to pick a prior that adds little or no information to the analysis, as in Jimmy Savage's principle of stable estimation. By the time Bayesian analyses became practical and widespread in the early 1990s, this argument had mostly died along with the disputants. (But not completely; see Lee et al. 2006, Lee & Nelder 2009.) In practice, it is rare to see prior distributions that describe subjective personal belief. Sometimes the prior represents *bona fide* external information, e.g., information about a medical device based on clinical trials of similar devices. Most often, the prior simply conditions the likelihood $L(\theta;\mathbf{w}) = \pi(\theta|\mathbf{w})$ by restricting the range of θ or by shortening the tails of $L(\theta;\mathbf{w})$. In either role, a prior can improve a procedure's performance markedly. However, most often people would like to use a minimally informative prior and unfortunately no consensus exists on what such priors might be. If you want to do a Bayesian analysis, you must supply a prior distribution and selection of priors is a live issue in analysis of mixed linear models.

1.3.2 Bayesian Analysis of Mixed Linear Models

Applying the preceding section's terminology to the mixed linear model as in (1.1):

- The unknown θ includes
  - unknowns in the mean structure, β and $\mathbf{u}$, and
  - unknowns in the variance structure, $\phi = (\phi_G, \phi_R)$.
- The data $\mathbf{w}$ are, by convention, the outcome $\mathbf{y}$.
- The likelihood is specified by (1.1) and the lines immediately following it; $\pi(\mathbf{y}|\beta, \mathbf{u}, \phi_R)$ is Gaussian (normal) with mean $\mathbf{X}\beta + \mathbf{Z}\mathbf{u}$ and covariance $\mathbf{R}(\phi_R)$.
- The prior for $\mathbf{u}$ given ϕ_G is $N(0, \mathbf{G}(\phi_G))$; β is generally given a multivariate normal prior, most commonly an improper flat prior $\pi(\beta) \propto 1$. Usually β and $\mathbf{u}$ are independent conditional on ϕ though this is not necessary. Finally, (ϕ_G, ϕ_R) are given a prior $\pi(\phi_G, \phi_R)$ which will be discussed at length below.

Although in common usage $\mathbf{X}$ and $\mathbf{Z}$ are considered part of the *dataset* for a mixed-linear-model analysis, for historical reasons they are not part of the *data* $\mathbf{y}$. Rather, they are simply treated as known quantities. The foregoing specifications permit a Bayesian analysis.

1.3.2.1 Tests and Intervals for Unknowns

By The Fundamental Principle, inferential statements about any unknowns depend on the data only through their posterior distributions. Thus, all inferential statements

about β depend on the data only through the marginal posterior distribution of β, $\pi(\beta|\mathbf{y})$; similarly for individual β_j, for $\mathbf{u}$ or individual u_j, and for ϕ or individual elements of it. Posterior intervals or regions analogous to confidence intervals are obtained from these marginal distributions. Any interval or region containing the desired probability can be used. Conventionally this is 95%, although in the dismal Bayesian tradition of offering naïve consumers something for nothing, narrower 90% intervals are becoming more common. These intervals naturally account for uncertainty about the unknowns ϕ, in distinct contrast to the conventional approach, which treats $\hat{\phi}$ as if it were the true ϕ.

Point estimates for elements of β and $\mathbf{u}$ are generally posterior means or medians, which are usually quite similar if not identical. Selection of point estimates is much more complicated for elements of ϕ, because their posteriors are not always unimodal and rarely if ever symmetric. The most commonly used point estimates are the mean or median of marginal distributions of individual elements of ϕ. The simulation study in He & Hodges (2008) strongly suggests that the posterior mean should not be used as a point estimate because it is extremely sensitive to the tails of the posterior distribution, which for elements of ϕ are often long and poorly determined by the data. The posterior median is less sensitive than the mean to the distribution's tails but can still be sensitive if the upper tail is long. The simulation study in He & Hodges (2008) reached the surprising conclusion that the posterior mode is a better point estimate for standard deviations, variances, and precisions than the posterior mean or median because the posterior mode is quite insensitive to tail behavior. This grates against the standard intuition that modes are unstable but that intuition derives from histograms of data, not from posterior distributions. A more serious objection to the posterior mode as a point estimate is that the posterior for ϕ can be multimodal (Chapter 14 gives an example and Chapter 19 explores this further). Given our generally poor understanding of ϕ's posterior distribution, perhaps this odd finding is best interpreted as showing the complexity of that posterior distribution.

In the Bayesian approach, tests for entire effects, e.g., testing a 5-category grouping variable, can be specified but are somewhat more problematic. The conventional analysis allows Wald or likelihood ratio tests for entire effects; the Bayesian analog to Wald tests is to compute a 95% posterior region for the vector of coefficients in β and/or $\mathbf{u}$ corresponding to the effect and reject the null hypothesis of no effect if the region excludes the origin.

More commonly these days, an effect is tested implicitly by doing the Bayesian analysis with and without the effect and comparing the two model fits using the Deviance Information Criterion (DIC, Spiegelhalter et al. 2002). DIC is a Bayesian analog to Akaike's Information Criterion (AIC) and the Schwarz Criterion (also known as the Bayesian Information Criterion or BIC) in that each has the form of a goodness-of-fit measure penalized by a measure of model complexity. Although the DIC became popular very quickly after its introduction as an all-purpose method for comparing models, it is rather *ad hoc* as its originators readily admit, bless their hearts (see especially the discussion and rejoinder following Spiegelhalter et al. 2002). The DIC's measure of model complexity in particular has been criticized for its *ad hoc* definition and some odd behavior such as occasional negative values

(Celeux et al. 2006, Plummer 2008, Cui et al. 2010). Thus, although DIC is widely used, in my judgment it is of questionable value. Leaving aside such qualms, DIC, like AIC and its many variants, has no associated formal testing machinery so there is no calibration for differences between models in DIC analogous to P-values or chi-squared statistics.

Finally, Bayes factors provide a more traditional sort of Bayesian test. The Bayes factor comparing two hypotheses, say the null hypothesis that an effect is zero versus the alternative hypothesis that it is not, is the ratio of the marginal probability of the data under the null divided by the marginal probability of the data under the alternative (see, e.g., Bernardo & Smith 1994, pp. 389–395; Kass & Raftery 1995 give a review). If the Bayes factor is multiplied by the ratio of the prior probability of the null hypothesis divided by the prior probability of the alternative hypothesis, the result is the ratio of the posterior probability of the null and alternative; the Bayes factor allows a user to avoid putting prior probabilities on the competing models. Although these Bayesian entities have a long history, they have not come into widespread use because they depend on the prior distributions placed on the unknowns in the null and alternative models and the extent of that dependence is rather generally unknown (though Kass & Raftery 1995 take a more sanguine view). One approximate Bayes factor that turns up in a number of applications is computed from the model-selection criterion called the Schwarz Criterion or Bayesian Information Criterion (BIC); see Kass & Raftery (1995, p. 778). Raftery and his collaborators use this approximation in several applications, e.g., the R package BMA, which does Bayesian model averaging for linear and generalized linear models and Cox regressions. As far as I know, no software package uses this approximation for testing effects in mixed linear models; an exercise involves deriving such Bayes factors.

1.3.2.2 Some Comments on the Bayesian Analysis

In constrast to the conventional analysis, a Bayesian analysis makes no particular distinction between β and $\mathbf{u}$: Each is just a vector of unknown constants. Customarily, β and $\mathbf{u}$ are independent *a priori* so the prior distribution for $(\beta, \mathbf{u})$ factors as $\pi(\beta, \mathbf{u}) = \pi(\beta)\pi(\mathbf{u})$. This book follows that custom, but it is not necessary.

The list above described $\mathbf{u}$'s distribution as part of the prior, a label used by many people who call themselves Bayesian. However, as noted, Bayarri et al. (1988) and Bayarri & DeGroot (1992) have argued convincingly, using mixed linear models as an example, that it is not possible to distinguish cleanly the likelihood and the prior. I prefer to call $\mathbf{u}$'s distribution part of the model for the data and to reserve the term "prior" for $\pi(\beta, \phi)$. This choice is admittedly arbitrary but makes it easier to communicate with the large majority of statisticians who were primarily trained in conventional statistics, to whom Bayesian jargon is like a foreign language. In particular, this usage makes it possible to communicate with the declining but still non-trivial group of statisticians who consider Bayesian methods to be somewhere between unnecessary and loathsome. In my first job after graduate school, I had an older colleague who became visibly upset when the conversation turned to Bayesian methods, as if he was in the presence of an abomination. By using the term *model*

for assumptions like "$\mathbf{u} \sim N(0, \mathbf{G})$," I was able to maintain a conversation with this well-read colleague and learn many useful things.

Fortunately, this labeling quandary creates no problem for a Bayesian analysis: No matter what you call $\mathbf{u}$'s probability distribution, the Bayesian calculations are the same. Even if you eliminate $\mathbf{u}$ as in (1.12), no problems arise within the Bayesian framework because $\mathbf{u}$ is eliminated simply by marginalizing with respect to it, i.e., integrating $\mathbf{u}$ out of the model or the posterior, either of which gives the same marginal posterior distribution of (β, ϕ). This is an example of the coherence of Bayesian analysis.

Before discussing specific prior distributions for ϕ, I offer some tentative generalizations about their importance. In analyses of mixed linear models with $\mathbf{R} = \sigma_e^2 \mathbf{I}$, the data generally provide fairly strong information about σ_e^2. The theory in Part IV of this book gives good reasons to expect this and except for tiny datasets, I have never seen a case in which the posterior for σ_e^2 depended much on its prior — that is, when the prior was varied moderately within the classes described below, because obviously you can make the posterior for σ_e^2 do almost anything if you make the prior sufficiently extreme.

However, for normal-error mixed linear models at least, the data only provide weak information about ϕ_G. Indeed, for these models ϕ_G is Bayesianly unidentified by the definition of Gelfand and Sahu (1999) because $\pi(\phi_G|\mathbf{u}, \mathbf{y}) = \pi(\phi_G|\mathbf{u})$, i.e., "observing the data $\mathbf{y}$ does not increase our prior knowledge about $[\phi_G]$ given $[\mathbf{u}]$" (Gelfand & Sahu 1999, p. 248). This differs from the conventional notion of identification, which Gelfand and Sahu call "likelihood identification," and apart from exceptions arising from, e.g., extremely small datasets, the *marginal* posterior of ϕ_G differs from the *marginal* prior of ϕ_G, so the data do provide information about ϕ_G. Generally speaking, however, the posterior of ϕ_G — and thus also the restricted likelihood — is not very well determined by the data. The theory in Part IV of this book gives some explanation for this empirical observation for an interesting subset of mixed linear models. Thus the prior for ϕ_G seems to be important in considerable generality and the consequences of different choices are poorly understood. Although some relevant literature exists, it is not large and "what is a good prior for general-purpose use?" is very much an open question. This is given as an exercise, of doctoral-dissertation scale, at the end of the chapter.

(A reader commented that the posterior predictive distribution of $\mathbf{y}$ or a new $\mathbf{y}_0$ is probably much less sensitive to ϕ_G's prior than the marginal posterior of β or $\mathbf{u}$, so the importance of ϕ_G's prior may depend on whether the goal is parameter estimation or prediction. This is certainly plausible but I would say unproven; it may depend on specifics of the model or data.)

Part of the problem of specifying priors for ϕ is the tremendous variety of covariance structures included in the class of mixed linear models. Such literature as exists has been almost exclusively about priors for individual variances or for covariance matrices.

1.3.2.3 Prior Distributions for Variances

For mixed linear models, the gamma prior for the precision — the reciprocal of the variance — is conditionally conjugate: It gives a gamma posterior for that precision conditional on all other unknowns. This prior became popular before 1990 when computation was rarely possible with non-conjugate priors but it continues to be popular even though modern Bayesian computing has made conjugacy unnecessary.

Such gamma priors can insert specific information into the analysis by specifying the distribution's mean and variance or by specifying a gamma distribution with particular percentiles, e.g., the 2.5^{th} and 97.5^{th} percentiles. However, most commonly gamma priors for the precision are intended to insert minimal information into the analysis. The most frequently used prior of this sort is a gamma distribution with parameters $(\varepsilon, \varepsilon)$, which has mean 1 and variance $1/\varepsilon$. (This distribution transforms to an inverse gamma prior on the variance, although the inverse gamma distribution has no mean or higher moments when $\varepsilon \leq 1$.) Small ε gives this gamma prior a large variance, which is generally taken to mean that the prior inserts little information into the analysis. One of this chapter's exercises shows another sense in which this prior inserts little information into the analysis, by interpreting it as being worth ε observations.

Common values for ε in practice are 0.01 and 0.001, which give gamma distributions with mean 1 and variance 100 and 1000, respectively. However, for a positive random variable with a fixed mean, increasing the variance does not give the distribution a roughly constant probability density, as it does for random variables taking values on the real line. Instead, the variance of a positive random variable can be increased with fixed mean only by concentrating probability near zero and in an extremely long right tail. This results in weird distributions: The Gamma(0.001,0.001) distribution has mean 1, variance 1000, and 95^{th} percentile 3×10^{-20}, which seems anything but diffuse. When first presented with this 95^{th} percentile, most people refuse to believe it. See for yourself: The R command is "qgamma(0.95, shape=0.001, rate=0.001)."

The Jeffreys prior for a variance σ^2 is sometimes used. It is a special case of the gamma prior on the precision with $\varepsilon = 0$, which is equivalent to an improper flat distribution on the real line for $\log \sigma^2$. As we will see, for some models the likelihood's contribution to the posterior distribution is a strange function of variances in ϕ, so this impropriety can cause problems.

Gelman has proposed at least two alternatives to gamma priors; the following two were proposed in Gelman (2006), which gives a nice survey of priors for variance parameters. The first idea is a flat prior on the standard deviation, the rationale being the natural desire to let the posterior reflect the likelihood's information about the standard deviation, which has the same units as the random variable it describes. This prior's weakness is that a user must specify upper and lower limits for its sample space. The lower limit of this interval is usually zero or close to it (sometimes the prior is bounded away from zero to avoid computing problems) and the upper limit is usually somewhat arbitrary. Estimates and intervals for variances based on this prior can be sensitive to the upper limit when the likelihood has a flat upper tail, as

it often does. This prior performed reasonably well in the simulation studies in He et al. (2007) and He & Hodges (2008) but otherwise little systematic information is available about its performance.

Gelman (2006) also proposed a half-t prior for the standard deviation of a random effect; this can also be described as the distribution of the absolute value of a central t variate. The half-Cauchy distribution is a special case. The full conditional posterior implied by this prior is a folded noncentral t distribution, so if the half-t is understood as a particular folded noncentral t, then it is conditionally conjugate for a standard deviation. This prior has largely been studied as a way to improve the mixing behavior in MCMC routines (see references in Gelman 2006, p. 519) but little is known about its statistical performance.

He et al. (2007) proposed a rather different prior for the common case in which $\mathbf{G}$ is diagonal with two or more unknown variances on its diagonal and $\mathbf{R} = \sigma_e^2\mathbf{I}$. The prior is defined in terms of the precisions, i.e., the reciprocals of the variances. Suppose the vector of unknown precisions is $(\tau_e, \tau_1, \ldots, \tau_m)$, where $\tau_e = 1/\sigma_e^2$ is the error precision and $\tau_1, \ldots, \tau_m$ are the precisions of the m random effects. Reparameterize the m random effect precisions as follows. First, define the total relative precision $\lambda = \sum_{k=1}^m \tau_k/\tau_e$, where "relative" means "precision of the random effects relative to the error precision." Now define the allocation of total relative precision among the m random effects as $\gamma = (\gamma_1, \ldots, \gamma_m)$, where

$$\gamma_k = \frac{\tau_k/\tau_e}{\lambda} = \frac{\tau_k}{\sum_{j=1}^m \tau_j}, \tag{1.23}$$

so $\tau_k = \tau_e\gamma_k\lambda$. Note that $\sum_{k=1}^m \gamma_k = 1$, and $\gamma = (\gamma_1, \ldots, \gamma_m)$ takes values in the m-dimensional simplex. This parameterization has two nice features: γ is invariant to changes in the data's scale and it takes values in a compact set. Thus a flat prior on γ is proper and treats the random-effect variances as exchangeable. Total relative precision, a positive real number, is better determined by the data than are the individual random-effect precisions, so its prior has relatively little impact. The simulation studies by He et al. (2007) and He & Hodges (2008) suggest this prior has good properties but it too has received little study.

A third approach was proposed by Hodges & Sargent (2001) and extended in Cui et al. (2010). This approach uses a measure of the complexity of the model's fit, degrees of freedom, which Section 2.2 discusses in detail so for now we'll use a simplified description and defer a detailed discussion. For the case in which $\mathbf{R} = \sigma_e^2\mathbf{I}$ and $\mathbf{G}$ is a diagonal matrix with L unknown variances on its diagonal, so $\phi_G = (\sigma_1^2, \ldots, \sigma_L^2)$, the degrees of freedom in the mixed linear model's fit is the sum of the degrees of freedom in the L components of the fit corresponding to the L variances in ϕ_G. The vector of component-specific degrees of freedom, $(\nu_1, \ldots, \nu_L)$, is a function of the L ratios $(\sigma_1^2/\sigma_e^2, \ldots, \sigma_L^2/\sigma_e^2)$; thus a prior distribution on $(\nu_1, \ldots, \nu_L)$ induces a prior distribution on $(\sigma_1^2/\sigma_e^2, \ldots, \sigma_L^2/\sigma_e^2)$. (All of this is well-defined under conditions given in Section 2.2.2.3.) Specifying a prior distribution on the error variance σ_e^2 completes the specification of a prior on ϕ. Published examples include Hodges et al. (2007), Cui et al. (2010) and, for a spatial model, Paciorek & Schervish (2006).

Putting a prior distribution on the degrees-of-freedom scale has two main advantages. First, for models with new-style random effects such as penalized splines, the variances in ϕ_G are merely a device to implement smoothing and it is nearly impossible to develop intuition or external information about them. By contrast, the degrees of freedom in a fit is subject to much more direct interpretation and thus intuition. In such uses, a prior on the degrees of freedom allows a user to directly condition the smoothness of the fit. Chapter 2 gives an example. Second, the prior on degrees of freedom is invariant to a large range of scale changes and re-parameterizations of the mixed linear model (Section 2.2.2.3), which is generally not true for priors specified directly on variances in ϕ. For the balanced one-way ANOVA model, a flat prior on the degrees of freedom in the fit is equivalent to the uniform-shrinkage prior (Daniels 1999).

The performance of this prior was examined in a simulation study in Hodges et al. (2007). Once again, however, little is known about this prior's properties outside of very simple cases.

1.3.2.4 Prior Distributions for Covariance Matrices

For unstructured $\mathbf{G}$, in practice it is rare to see a prior distribution other than the Wishart distribution for $\mathbf{G}^{-1}$; see e.g., Bernardo & Smith (1994, pp. 138–140). The Wishart is a generalization of the gamma distribution and is conditionally conjugate for $\mathbf{G}^{-1}$. This distribution has two parameters, degrees of freedom α and a symmetric non-singular matrix $\mathbf{B}$, where $\alpha > (q-1)/2$, q being the dimension of $\mathbf{G}$, and where the Wishart's expectation is $\alpha\mathbf{B}^{-1}$. Most commonly, these are specified as $\mathbf{B} = \mathbf{I}_q$ and α the smallest integer that specifies a proper Wishart. The Wishart's popularity appears to be a relic of the time before modern Bayesian computing because it resists intuition. Under the Wishart, the marginal distributions of individual precisions are chi-squared, and the marginal distributions of individual correlations are beta-distributed on the interval $[-1,1]$ (Barnard et al. 2000, Section 2), so the correlations can be given marginal uniform priors by setting the Wishart's parameters a certain way. However, I know of nothing that provides intuition about the correlations between unknowns implied by the Wishart. Little is known about the performance of posterior distributions obtained using this prior, though it performs poorly in some cases in the simulation studies of Daniels & Kass (1999, 2001).

There is a modest literature of alternatives to the Wishart. Much of this work was motivated by simple problems, e.g., X_i iid $N(\mu,\Sigma)$ with unknown μ and Σ, but it is applicable to mixed linear models. Daniels & Kass (1999, 2001) review the literature up to 2001. They discuss the Jeffreys prior and Berger-Bernardo priors for a covariance matrix, but these priors perform poorly in their simulations. The poor performace of the Jeffreys prior has some theoretical basis. Eaton & Freedman (2004) showed that it gives incoherent predictive distributions in the iid $N(\mu,\Sigma)$ case and surveyed literature with similar results. Eaton & Sudderth (2004) showed the same thing as a by-product of a more general argument about priors for covariance matrices. Incoherence also holds for the estimation problem by a similar argument (M.L. Eaton, personal communication). I do not know whether the Berger-Bernardo

prior has a similar flaw. In any case, the Jeffreys prior and Berger-Bernardo priors will not be discussed further here.

Three alternatives to the Wishart were proposed by Daniels & Kass (1999), Barnard, McCulloch, & Meng (2000), and Chen & Dunson (2003). Each depends on a decomposition of the covariance matrix; the first two are non-conjugate while the third is partially conjugate.

Daniels & Kass (1999, 2001) intended their prior to shrink the covariance toward a diagonal matrix and to shrink the eigenvalues of the covariance toward each other. It uses the spectral decomposition $\mathbf{G} = \Gamma\mathbf{D}\Gamma'$, where the orthogonal matrix Γ has $\mathbf{G}$'s eigenvectors as its columns, and the diagonal matrix $\mathbf{D}$ has $\mathbf{G}$'s eigenvalues as its diagonal entries. Daniels & Kass (1999) propose a flat prior on the eigenvalues, which has the effect of shrinking them toward each other (p. 1256). The $q \times q$ orthogonal matrix Γ is written as the product of $q(q-1)$ simple matrices, each of which is a function of a so-called Givens angle θ_{ij} which "may be considered a rotation in the plane spanned by [the] i and j components of the basis defining the matrix [Γ]." Each $\theta_{ij} \in (-\pi/2, \pi/2)$; the logit of each θ_{ij} receives a normal prior centered at zero.

Barnard et al. (2000) write $\mathbf{G} = \text{diag}(\mathbf{S})\ \mathbf{B}\ \text{diag}(\mathbf{S})$, where $\mathbf{S}$ is a q-vector of standard deviations, the square roots of the diagonals of $\mathbf{G}$, and $\mathbf{B}$ is the correlation matrix corresponding to $\mathbf{G}$. The motivation is that practitioners are trained to think in terms of standard deviations, which are on the same scale as the data, and correlations. Although these authors discuss a variety of possibilities within this framework, they favor making $\mathbf{S}$ and $\mathbf{B}$ independent *a priori*. For $\log(\mathbf{S})$, they favor a multivariate normal with a diagonal covariance. For $\mathbf{B}$, they focus on two priors: The marginal distribution of the correlation matrix $\mathbf{B}$ implied by an inverse-Wishart on $\mathbf{G}$, which for a specific setting of the Wishart parameters gives a uniform marginal distribution for each correlation in $\mathbf{B}$; and a flat proper prior on the space of legal correlation matrices $\mathbf{B}$, which is uniform on the correlations jointly but not uniform marginally for individual correlations.

The prior of Chen & Dunson (2003) is somewhat like that of Barnard et al. (2000) but uses a Cholesky decomposition. For the covariance matrix $\mathbf{G}$ of the random effects in the mixed linear model (1.1), let $\mathbf{G} = \mathbf{L}\mathbf{L}'$ be the unique Cholesky decomposition, where $\mathbf{L}$ is lower triangular with nonnegative diagonal elements. Let $\mathbf{L} = \Lambda\Gamma$, where Λ is diagonal with non-negative elements and Γ is lower triangular with 1's in the diagonal entries. Then the random effects of (1.1) can be rewritten as $\mathbf{Zu} = \mathbf{Z}\Lambda\Gamma\mathbf{b}$, where $\mathbf{b}$ is a vector of iid standard normal latent variables. This re-expression of the random effects is handy because Λ contains the standard deviations of the random effects $\mathbf{u}$ and the lower-triangular Γ, which carries information about correlations between random effects, can be interpreted as regression coefficients. Because of the latter, Chen & Dunson (2003) give it a multivariate normal prior conditional on Λ, which is conditionally conjugate. It would then be possible to follow Barnard et al. (2000) by putting a multivariate normal prior on $\log(\Lambda)$, but Chen & Dunson's purpose is to do variable selection for random effects, so instead their prior treats the elements of Λ as independent and gives each element a mixture distribution with positive prior probability on zero. Pourahmadi proposed a non-Bayesian method based on a Cholesky decomposition with 1's on the diagonal, though it treats the standard

deviations differently than Chen & Dunson (2003). Pourahmadi (2007) gives a brief review and contrasts his approach with the approach of Chen & Dunson (2003).

Simulation studies in Daniels & Kass (1999, 2001) show that the Givens-angle prior and Barnard et al.'s decomposition perform well in a range of cases and perform much better than the Wishart prior in some cases. Pourahmadi (2007, p. 1006) claims the Cholesky-based approach "has been used effectively" in several areas but none of the citations gives information about its performance apart from illustrative examples.

1.3.3 Computing for Bayesian Analyses

This section is not intended to be a survey of an area, Bayesian computing, that has simply exploded in the last two decades. Rather, it emphasizes one method, Markov chain Monte Carlo, which brought Bayes to the masses and dominates Bayesian applications (which was considered an oxymoron when I was in graduate school).

Before Gelfand and Smith (1990), explicit expressions for $\pi(\beta,\mathbf{u}|\mathbf{y})$, $\pi(\beta|\mathbf{y})$, $\pi(\mathbf{u}|\mathbf{y})$, or $\pi(\mathbf{X}\beta+\mathbf{Z}\mathbf{u}|\mathbf{y})$ were only tractable in simple or unrealistic special cases. It is easy to derive $\pi(\phi_G,\phi_R|\mathbf{y})$, the marginal posterior of the unknowns in the covariances $\mathbf{G}$ and $\mathbf{R}$; this is given as an exercise. Similarly, if $\mathbf{R}=\sigma_e^2\mathbf{I}$ and $1/\sigma_e^2$ has a gamma prior, it is easy to derive the marginal posterior of the unknowns in $\frac{1}{\sigma_e^2}\mathbf{G}$; this is also given as an exercise. Unfortunately, these marginal posteriors have nonstandard forms so it is hard to compute posterior intervals with pre-1990 methods except when ϕ_G has a single dimension. Thus while these results are sometimes useful in developing theory and methods, otherwise they have little practical utility.

In practice, Bayesian analyses since 1990 have almost always been performed by drawing samples from the posterior distribution of the unknowns and summarizing those draws as (estimates of) posterior percentiles, means, etc. By far the most common method of sampling from the posterior is Markov chain Monte Carlo (MCMC). If $\theta=(\beta,\mathbf{u},\phi)$ is the vector of unknowns in a mixed linear model, an MCMC method produces a sequence of draws from the joint posterior distribution of θ, $\theta_{(b)}, b=1,\ldots,B$. The sequence is a Markov chain because the distribution of $\theta_{(b)}$ is a function only of the preceding draw $\theta_{(b-1)}$ (and $\mathbf{y}$, $\mathbf{X}$, and $\mathbf{Z}$, of course) but not of any earlier draws in the chain $\theta_{(j)}, j\le b-2$. With a few exceptions, it is usually much easier to make draws from certain Markov chains than to make iid draws from the posterior; the price is that the sequence of MCMC draws is serially correlated. One extremely useful feature of MCMC is that once you have drawn a sequence $\theta_{(1)},\theta_{(2)},\ldots,\theta_{(B)}$, then for *any* function $g(\theta)$, the sequence $g(\theta_{(1)}),g(\theta_{(2)}),\ldots,g(\theta_{(B)})$ is a sequence of MCMC draws from the posterior of $g(\theta)$.

The MCMC literature is enormous and I will not try to summarize it even for mixed linear models. I give an example below of an MCMC algorithm for mixed linear models which is as well understood as any MCMC algorithm for an interesting class of models but which nobody claims is optimal. Interested readers looking for more background on MCMC can consult the relatively friendly introductions in Car-

lin & Louis (2008) or Gilks et al. (1996) although new texts are published routinely and you might find it worthwhile to shop around.

Here is an MCMC algorithm for mixed linear models; in MCMC jargon, this is a Gibbs sampler that blocks on $(\beta, \mathbf{u})$. Given a starting value for θ, $\theta_{(0)}$, one draw of the full vector θ is made by drawing, in order, subsets of θ conditional on the most recent draws for the rest of θ. In the Gibbs sampler shown here, a draw is first made from the so-called full conditional of $(\beta, \mathbf{u})$, i.e., the posterior distribution of $(\beta, \mathbf{u})$ conditional on the previous draw of ϕ (the full conditionals shown here are straightforward to derive and given as an exercise):

$$\begin{aligned}
(\beta, \mathbf{u})|\mathbf{y}, \phi_{(b-1)} \quad &\sim \quad \text{normal with mean} \\
&\left[\mathbf{C}'\mathbf{R}^{-1}_{(b-1)}\mathbf{C} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1}_{(b-1)} \end{pmatrix}\right]^{-1} \mathbf{C}'\mathbf{R}^{-1}_{(b-1)}\mathbf{y} \\
\text{and covariance} \quad &\left[\mathbf{C}'\mathbf{R}^{-1}_{(b-1)}\mathbf{C} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1}_{(b-1)} \end{pmatrix}\right]^{-1}, \qquad (1.24)
\end{aligned}$$

where, as before, $\mathbf{C} = [\mathbf{X}|\mathbf{Z}]$, the $\mathbf{0}$ are appropriate-sized matrices of zeroes, and the subscripts on $\mathbf{R}$ and $\mathbf{G}$ indicate that they are computed using the $(b-1)^{\text{th}}$ draw of ϕ. The conditional posterior mean is formally identical to Section 1.2.2's equation (1.8), the BLUPs, which is not a coincidence. Having drawn $(\beta_{(b)}, \mathbf{u}_{(b)})$ from this conditional distribution, $\phi_{(b)}$ is now drawn from its full conditional distribution given $(\beta_{(b)}, \mathbf{u}_{(b)})$:

$$\begin{aligned}
&\pi(\phi_G, \phi_R|\mathbf{y}, \beta_{(b)}, \mathbf{u}_{(b)}) \\
\propto \quad &|\mathbf{R}(\phi_R)|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\beta_{(b)} - \mathbf{Z}\mathbf{u}_{(b)})'\mathbf{R}(\phi_R)^{-1}(\mathbf{y} - \mathbf{X}\beta_{(b)} - \mathbf{Z}\mathbf{u}_{(b)})\right) \\
&\times|\mathbf{G}(\phi_G)|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{u}'_{(b)}\mathbf{G}(\phi_G)^{-1}\mathbf{u}_{(b)}\right)\pi(\phi_G, \phi_R), \qquad (1.25)
\end{aligned}$$

where $\pi(\phi_G, \phi_R)$ is the prior of (ϕ_G, ϕ_R). Note that if ϕ_G and ϕ_R are independent *a priori*, so $\pi(\phi_G, \phi_R) = \pi(\phi_G)\pi(\phi_R)$, then ϕ_G and ϕ_R are conditionally independent *a posteriori*, which is convenient computationally. For many problems (1.25) is a well-known distributional form. For example, if $\mathbf{G} = \sigma_s^2\mathbf{I}_q$ and $\mathbf{R} = \sigma_e^2\mathbf{I}_n$ so $\phi_G = \sigma_s^2$ and $\phi_R = \sigma_e^2$, then the full conditional distribution for (σ_s^2, σ_e^2) is

$$\begin{aligned}
&\pi(\sigma_s^2, \sigma_e^2|\mathbf{y}, \beta_{(b)}, \mathbf{u}_{(b)}) \\
\propto \quad &(\sigma_e^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma_e^2}(\mathbf{y} - \mathbf{X}\beta_{(b)} - \mathbf{Z}\mathbf{u}_{(b)})'(\mathbf{y} - \mathbf{X}\beta_{(b)} - \mathbf{Z}\mathbf{u}_{(b)})\right) \\
&\times(\sigma_s^2)^{-\frac{q}{2}} \exp\left(-\frac{1}{2\sigma_s^2}\mathbf{u}'_{(b)}\mathbf{u}_{(b)}\right)\pi(\sigma_s^2, \sigma_e^2), \qquad (1.26)
\end{aligned}$$

where $\pi(\sigma_s^2, \sigma_e^2)$ is the prior of (σ_s^2, σ_e^2). If the precisions $1/\sigma_s^2$ and $1/\sigma_e^2$ have the conventional gamma prior, then their full conditionals are also gamma distributions.

Ruppert et al. (2003, Section 16.6) give another MCMC algorithm for the simple case just above, which generalizes straightforwardly to all mixed linear models. In this so-called Rao-Blackwellized algorithm, MCMC draws are first made from the *marginal* posterior of (ϕ_G, ϕ_R), i.e., the posterior of (ϕ_G, ϕ_R) after having integrated $(\beta, \mathbf{u})$ out of the joint posterior. Then the marginal posterior density of $(\beta, \mathbf{u})$ is estimated by averaging the full-conditional density (1.24), for each value of $(\beta, \mathbf{u})$, over the MCMC draws of (ϕ_G, ϕ_R). This gives an efficient estimate of the marginal posterior density of $(\beta, \mathbf{u})$. Taking advantage of normal theory applied to (1.24), Rao-Blackwellization also allows similarly efficient estimates of posterior moments and marginal densities of linear functions of $(\beta, \mathbf{u})$.

That's the good news: In the glorious new world of Bayesian computing, there are algorithms for pretty much any problem and many possible algorithms for mixed linear models. Now for the not-so-good news: For innocent-looking mixed linear models, it is easy to write down legal MCMC samplers that have extremely poor properties. For example, Chapter 17 examines in detail a simple model yielding very strange marginal posteriors for ϕ, which was a nasty but productive surprise (Reich et al. 2007). "Poor properties" refers to several related things: The sequence of MCMC draws has high serial autocorrelation; the series of draws moves slowly around the space in which θ takes values; the sampler can "get stuck" in a small region of values of θ for many draws; and so on. Presence of any of these problems means the sampler provides little information about the posterior distribution for a given amount of computing. These problems afflict MCMC generally and many diagnostics have been proposed for detecting them. As early as 1996, Cowles & Carlin (1996) gave detailed reviews of 13 such diagnostics but concluded that "all ... can fail to detect the sorts of convergence failure that they were designed to identify."

Recently, encouraging progress has been made in an approach pioneered by Jim Hobert, Galin Jones, and their students. For models with $\mathbf{G} = \sigma_s^2\mathbf{I}$ and $\mathbf{R} = \sigma_e^2\mathbf{I}$ and inverse gamma priors on the two variances, Johnson & Jones (2010) gave sufficent conditions on the gamma priors under which the blocked Gibbs sampler described above is geometrically ergodic. Roman & Hobert (2012) extended those results to include models with multiple random effects and a wider class of priors, including a flat (improper) prior on the fixed effects (β) and inverse gammas on the random-effect variances including hyperparameter values that give improper priors. Johnson & Geyer (2012) extended this approach in a different direction by showing how to change variables in the posterior to obtain a posterior density that satisfies the conditions for geometric ergodicity, so that running an MCMC on the transformed variable and back-transforming gives a geometrically ergodic sampler for the original variable. This approach has very wide applicability. These geometric ergodicity results permit derivation of asymptotic MCMC standard errors for expectations of functions of the unknown parameters, which gives a way to determine the number of MCMC iterations B needed to obtain an estimate of the posterior mean of a scalar function $g(\theta)$ with a given (small) MCMC error (Flegel et al. 2008). One of these methods, the consistent batch means method (Jones et al. 2006), is extremely easy to use and is available in the R package mcmcse, written by James Flegal and John Hughes.

The results just described establish asymptotic properties of the MCMC sequence and there is no guarantee that in "small samples" (i.e., any practical B), a given chain has converged. However, this work is progressing quickly and I think it is not unrealistic to look forward to Bayesian software for a respectable class of mixed linear models that is simple for users and as foolproof as statistical software can be.

A final comment on MCMC: Proponents of Bayesian methods sometimes claim that analyses using MCMC or another sampling method are exact in that they do not rely on large-sample approximations, as conventional analyses do for mixed linear models. This is true: If you run a well-functioning MCMC long enough, you can obtain an arbitrarily precise estimate of any posterior quantity. However, nobody knows how long is "long enough" and there are examples in which "long enough" is on the order of 10^{20} iterations. In other words, this claim should be taken as hype, not fact, at least for the near future.

As noted at the beginning of this section, Bayesian computing continues to develop quickly and now includes other sampling methods and powerful approximations such as integrated nested Laplace approximations (Rue et al. 2009). Interested readers looking for a place to start might consider the latest edition of Carlin & Louis's text.

1.4 Conventional and Bayesian Approaches Compared

This section summarizes the advantages and disadvantages of the two approaches, then illustrates them using analyses I did for Example 1, the molecular structure of a virus (Peterson et al. 2001).

1.4.1 Advantages and Disadvantages of the Two Approaches

Some key advantages of the conventional approach, based on maximizing the restricted likelihood, are as follows:

- Maximizing is usually simpler than integrating; this is the case for mixed models.
- As a result, flexible mixed-linear-model analyses using the conventional approach are available in every major statistical package.
- Maximizing the restricted likelihood is often much faster than Bayesian analyses using MCMC, if it's done right.

The following are some key disadvantages:

- It is somewhere between difficult and impossible to account for variability in the estimate $\hat{\phi}$ (or account for uncertainty about ϕ, if you prefer) in intervals and tests for β and $\mathbf{u}$. To my knowledge, no software package does this.
- Maximizing the restricted likelihood routinely produces zero estimates for variances in $\mathbf{G}$. When this happens, conventional theory offers no way to construct an interval describing likely values of the variance that was estimated to be zero.

The statement "Maximizing the restricted likelihood is [very fast] if it's done right" requires some comment. In maximizing the restricted likelihood, some soft-

ware packages use brute force to invert the matrices involved, which can be prohibitively large. The MIXED procedure in SAS does this and it is not hard to specify relatively simple problems that MIXED cannot practically compute. By contrast, for mixed linear models with $\mathbf{R} = \sigma_e^2 \mathbf{I}_n$, the algorithm in the lme4 package in R (Bates & DebRoy 2004) uses a clever matrix decomposition to compute extremely quickly, even for some problems that are effectively impossible to compute in MIXED. Unfortunately, the restriction to $\mathbf{R} = \sigma_e^2 \mathbf{I}_n$ excludes many models that can be specified in MIXED.

Regarding zero variance estimates, I use mixed linear models to analyze perhaps 25 datasets a year, fitting maybe 200 mixed linear models, and I get zero variance estimates routinely. Most of these analyses are some version of repeated-measures or variance-components analysis. In some cases, the random effects are present merely to account for clustering in computing standard errors for β, so the difference between a zero variance and a small variance has little effect and the zero variance estimate seems innocuous. In many cases, however, the variance component itself is of interest and in such cases a zero estimate is simply wrong — the variance component is not zero, though it may be small — and the absence of an interval of plausible values is worse than useless.

To me, the single most striking thing about text and reference books for mixed linear models is their silence on this subject. I have read Ruppert et al. (2003) very carefully and found exactly one mention of zero variance estimates, on p. 177, which "we cannot explain."[3] Perhaps these authors rarely obtain zero estimates in the problems they work, although my students get zero variance estimates often enough when they fit the same models (penalized splines). Ruppert et al. (2003) do discuss testing whether a variance component is zero (e.g., pp. 106–107), with laudable warnings about relying on large-sample approximations. I have found no mention of zero variance estimates in Lee et al. (2006), Fahrmeir & Tutz (2010), Searle et al. (1992, which does discuss negative variance estimates produced by obsolete methods), Verbeke & Molenberghs (1997), or Diggle et al. (1994), though I might have missed some brief mentions. Snijders & Bosker (2012) is alone in forthrightly stating that zero variance estimates are common and that standard software gives a misleading standard error of zero for such zero variance estimates (Sections 4.7, 5.4). Otherwise, no text or monograph, as far as I know, gives prominent treatment to this routine annoyance and research in the conventional analysis has produced no methods to handle it. In fairness to these authors and others, most books in this field focus on specifying and fitting models and interpreting the results of tidy examples, which may be the right emphasis given that this class of models has not yet penetrated the general statistics user market as much as, say, ordinary linear models or Cox regression. Nonetheless, it is an odd hole in a field for which model-fitting methods have been intensively developed for at least 40 years. Zero variance estimates are present in the example in Section 1.4.2 below, while Chapter 18 of the present book explores this topic in detail.

Some key advantages of the Bayesian approach are as follows:

[3] Actually, I missed this lone mention but my student Lisa Henn didn't.

- A Bayesian analysis naturally accounts for uncertainty about ϕ in tests and intervals for β and $\mathbf{u}$.
- The Bayesian analyses specified above do not give zero point estimates for variances in ϕ_G (but see the related comment below) and naturally provide intervals of plausible values in cases in which the conventional analysis would give a zero point estimate and no interval.

Some key disadvantages are as follows:

- The results depend on prior distributions, particularly for ϕ_G, to an extent that is not close to being well understood. Sensitivity analyses can address this but I've never seen a sensitivity analysis for priors that really explored a range of alternative priors (including, alas, my own published analyses, such as the example given below).
- With existing software, a Bayesian analysis is harder to compute than a conventional analysis, although presumably this will improve as time passes.
- MCMC routines for mixed linear models are not well understood in any generality and in a strict view would be considered experimental for all cases not meeting the sufficient conditions described above.

It is not difficult to specify Bayesian analyses that allow a variance's posterior distribution to have positive mass on zero; this is one Bayesian way to test whether that variance is zero. (The method of Chen & Dunson 2003, discussed above in Section 1.3.2.4, is an example.) Such specifications are not exotic today but they are not widely used and little is known about how they perform.

Saying that a Bayesian analysis "naturally gives intervals of plausible values" glosses over the question of how much you should trust those intervals. In problems where maximizing the restricted likelihood gives a zero estimate for a variance, the restricted likelihood can decline very slowly from zero as that variance is increased. I would like to say this happens frequently, but no one knows whether that is true. Chapters 14 and 18 explore an example of a zero variance estimate in which the restricted likelihood is quite flat near zero. In this example, I only discovered that the restricted likelihood was flat near zero because I did Bayesian analyses with many different priors and happened to notice that for each prior, the prior and posterior had very similar percentiles up to about the median. Thus, while a Bayesian analysis avoids the obviously wrong zero estimates for variances and provides a plausible interval, in cases where the restricted likelihood — the data contribution to the marginal distribution of ϕ — is flat near that peak at zero, the posterior is largely determined by the prior. If you begin with a Bayesian analysis and are not alerted to the possibility of this problem, there is no way to know that the data contribution to the marginal distribution of ϕ (the restricted likelihood) is quite flat unless you run many analyses with different priors. In other words, the ability to do a Bayesian analysis here makes it easy for a user to fool himself.

1.4.2 Conventional and Bayesian Analyses for the Viral-Structure Example

To be fair, I chose this example to make the conventional approach look bad: It emphasizes neglected consequences of the zero variance estimates routinely produced by maximizing the restricted likelihood. I also chose it because the analysis appeared in a good non-statistical journal. Be aware that in 2000 the Bayesian analyses presented here took me about 10 times as long to execute as the conventional analyses and that I have no compelling rationale for the prior distributions.

Peterson et al. (2001) collected data to estimate numbers of each of six proteins in the prohead of the bacteriophage virus $\phi 29$. The model I used to analyze these data was described in Section 1.1 and given in equation (1.3). Among the six proteins, gp8 had the most measurements with 98, while the other proteins had fewer than half as many measurements, so that parts of model (1.3) were omitted in analyzing them.

The conventional analyses were done in SAS's VARCOMP procedure with no fixed effects. The 95% interval in Table 1.7 is the estimate for the intercept plus and minus two times the standard error given by VARCOMP. (I should have used VARCOMP's confidence interval, which is based on the t distribution and thus wider.)

For the Bayesian analyses, the prior distributions were an improper flat prior for the intercept (β); for $1/\sigma_e^2$, gamma with mean 0.025 and standard deviation 0.05 (i.e., shape 0.25, scale 10); for $1/\sigma_j^2$, for $j = p,b,g,r$, gamma with mean 0.5 and standard deviation 1 (i.e., shape 0.25, scale 0.5). Peterson et al. (2001) gave this rationale for the prior on ϕ: "The prior distributions for the variances provide information equivalent to a quarter of a measurement [i.e., shape 0.25] and, apart from pure measurement error [σ_e^2], treat the components of variation equally." Subject to treating $1/\sigma_j^2, j = p,b,g,r$ equally, I chose this prior because it was consistent with what I knew about σ_e^2 and induced a nearly flat prior distribution on the degrees of freedom in the model fit. (Chapter 2 discusses degrees of freedom in the fitted values of mixed linear models and priors based on them.) For computations, I used the Gibbs sampler described in Section 1.3.3 implemented in an S+ program that I wrote myself. I ran the sampler for 5100 cycles and discarded the first 100 for burn-in. The point estimate and 95% interval are the average and the 2.5^{th} and 97.5^{th} percentiles of the 5000 retained Gibbs draws, respectively. WARNING: *Do not* mimic this MCMC sampler! I did this a long time ago when I was so ignorant about MCMC that I didn't know how ignorant I was. There was no need to write my own code; the number of iterations should be determined using objective methods like consistent batch means (Jones et al. 2006); discarding iterations for burn-in is unnecessary (Flegal et al. 2008, Flegal & Jones 2011); and the posterior mean is a worse point estimator than the posterior median for a variance, standard deviation, or precision (Section 1.3.2.1).

Table 1.7 gives the estimates and nominal 95% intervals for numbers ("Copies") of each molecule, as published in Peterson et al. (2001, Table II). The Bayesian interval is wider for all six molecules. The difference between the Bayesian and conventional ("REML") analyses is negligible for gp8 but quite large for the other five molecules. It is not a coincidence that gp8 has the largest sample size by quite a bit, but this difference between the two approaches is also affected by the efficiency of the design, in this case by the crossing and nesting of random effects.

Table 1.7: Molecular Structure of a Virus: Estimates and Intervals for Copy Numbers

Molecule	N Measurements	Method	Copies	95% Interval
gp7	31	REML	25.8	(19.0,32.6)
		Bayesian	26.9	(13.6,40.9)
gp8	98	REML	231.9	(200.5,263.3)
		Bayesian	232.4	(199.8,266.5)
gp8.5	46	REML	53.5	(37.7,69.3)
		Bayesian	52.6	(29.1,74.4)
gp9	40	REML	9.1	(7.9,10.3)
		Bayesian	9.1	(3.3,15.2)
gp11	40	REML	11.2	(10.3,12.1)
		Bayesian	11.2	(5.6,16.6)
gp12	40	REML	58.6	(51.9,65.3)
		Bayesian	58.2	(44.7,71.2)

The difference between the Bayesian and conventional intervals in Table 1.7 can be understood by considering the estimates of the variance components, which Table 1.8 gives for three molecules. For gp8, where the conventional and Bayesian intervals for "Copies" are quite similar, the REML estimate for a variance component's standard deviation is larger for some components while the Bayesian estimate is larger for others. The same is true for gp8.5, where the Bayesian interval for copies (Table 1.7) is 43% wider than the conventional interval and REML gives a zero estimate for the variance of the oxidizer-run random effect. For gp9, where the Bayesian interval for copies is nearly five times as wide as the conventional interval, the posterior medians are much larger than the REML estimates for the four variance components that are present. For all three molecules, the Bayesian 95% interval is extremely wide for all variance components except error (σ_e), reflecting a great deal of uncertainty. This uncertainty is ignored in the conventional analysis; this tends to make the conventional intervals narrower than the Bayesian intervals.

Table 1.8 shows some other things that, in my experience, are common in analyses of mixed linear models. For gp8 and gp8.5, the REML estimate and the two Bayesian estimates of σ_e are quite similar, while for gp9 the estimates differ less than they do for other variance components. The Bayesian interval for σ_e is also fairly narrow (the conventional interval would be, too, if it were shown). These observations are, in my experience, true of the error variance quite generally and Chapters 15 and 16 gives some theory suggesting — I won't go so far as to say "explaining" — why the data provide better information about the error variance than about random-

Table 1.8: Molecular Structure of a Virus: Estimates and Intervals for Variance Components (Expressed as Standard Deviations)

	σ_e	σ_p	σ_b	σ_g	σ_r
gp8					
REML estimate	20.3	23.8	20.6	0	15.1
Posterior mean	20.6	15.5	26.4	5.3	11.9
Posterior median	20.6	9.7	25.5	3.3	10.4
Bayes 95% interval	(17.7,24.1)	(0.6,63)	(4.0,53)	(0.5,20)	(0.7,34)
gp8.5					
REML estimate	4.3	9.1	7.6	3.6	0
Posterior mean	4.5	8.3	9.7	3.6	2.3
Posterior median	4.4	3.2	8.1	3.1	1.6
Bayes 95% interval	(3.6,5.6)	(0.5,49)	(2.6,27)	(0.8,9.1)	(0.5,9.3)
gp9					
REML estimate	0.96	—	0.85	0	0
Posterior mean	1.2	—	2.7	1.3	1.4
Posterior median	1.2	—	1.4	0.95	1.1
Bayes 95% interval	(0.99,1.6)	—	(0.5,12)	(0.4,4.1)	(0.4,4.9)

effect variances, even in mixed linear models that have no replication within design cells.

On the other hand, the intervals for variance components other than error are all quite wide. They are especially wide here because the sample sizes are small, but in my experience this is true in considerable generality. The theory in Chapters 15 and 16 also suggests why this happens.

Note also the similarity of the point estimates of copies produced by the conventional and Bayesian methods (Table 1.7). In this simple case, differences arise because the two approaches tend to give somewhat different estimates of variance components and thus to weight individual observations differently. Fixed effect estimates from the two approaches are not this similar in general. I have noticed, though, that even in cases where two variance-structure models fit quite differently according to standard criteria like a restricted likelihood ratio test or the Bayesian Information (Schwarz) criterion, that rarely has much effect on tests of fixed effects. There must be instances in which this fails, that is, apart from cases in which one of the models is grossly inappropriate, but I don't know what conditions are required.

Finally, the posterior means and medians differ a great deal for some of the random effects' standard deviations. This happens when the marginal posterior has an extremely long upper tail. The simulation study by He & Hodges (2008) suggests that in some generality the posterior mean is unstable as an estimate of a variance, standard error, or precision so it should not be used as a point estimate.

1.5 A Few Words about Computing

I am nowhere near expert on computing for mixed linear models or anything else. I have included this section, weak as it is, for the benefit of readers who know even less than I do and who are looking for a place to start. I say a bit about packages I have actually used and briefly mention other packages that I have not used but which have a following. Let the reader beware. Snijders & Bosker (2012, Chapter 18) give a survey of software for hierarchical (multi-level) models that includes almost all software mentioned here and quite a few packages not mentioned here.

The MIXED program in SAS performs the conventional analysis for a large range of mixed linear models using the methods and approximations discussed in previous sections. The REPEATED command specifies $\mathbf{R}$, the covariance of the errors ε, and RANDOM commands specify $\mathbf{G}$, the covariance of $\mathbf{u}$. It has taken me many years to become comfortable with the syntax of these two commands but MIXED can specify a very large collection of models. SAS's documentation for MIXED does not include many of the models discussed in this book; Ruppert et al. (2003, Appendix B.3) gives a SAS macro calling MIXED that fits many of the penalized splines in their book and their book's web site gives more SAS code. No doubt a web search would turn up a lot more SAS code for models in the present book. MIXED does not do tests comparing models with different variance structures but provides the maximized log restricted likelihood from which such tests can be computed. Computing for the conventional approach is by brute force and can be very slow. MIXED does Bayesian analyses using MCMC but only for the small subset of its models in which $\mathbf{R} = \sigma_e^2\mathbf{I}$ and $\mathbf{G}$ is diagonal (the documentation calls these "variance components models"). The easiest way to access documentation about MIXED is at SAS's web site, www.sas.com.

The R system has a few contributed packages that do computations for mixed linear models. The two that I have used are nlme and lme4, which also do computations for some generalized mixed linear models. Doug Bates was a co-author of both packages but only lme4, the later of the two, uses the matrix decompositions that speed up the necessary matrix inversions. Also, the two packages have very different syntaxes for specifying random effects. Although neither package has built-in capability for the same range of models as SAS's MIXED procedure, each allows user-specified variance functions. The lme4 package includes some Bayesian capability but I have never used it. Pinheiro & Bates (2000) gives a very friendly introduction to an earlier version of nlme implemented in the S-plus system. Documentation for the R version of nlme and for lme4 are available at the CRAN web site, under contributed packages (http://cran.r-project.org/web/packages/). The web site for Ruppert et al. (2003) has some code suitable for R or S-plus. I am advised that as of 2010 the SemiPar R package associated with their book is no longer actively maintained so you should probably not use it, although it was available on the CRAN web site the last time I checked. Finally, I am advised that the R package mgcv is popular and powerful but I have never done more than peruse its documentation.

Although MCMC methods seem quite simple for many problems, if it's practicable you are better off using existing MCMC engines instead of writing your own code because they're far less likely to have undiscovered bugs. For larger problems,

unfortunately, there may be no practicable alternative to writing your own code but given how little we understand posterior distributions even for mixed linear models, this is a hazardous undertaking.

The best-known Bayesian software is produced by the BUGS project and available in many forms. The Windows version is WinBUGS, OpenBUGS works on many platforms, and the BRugs package allows WinBUGS to be called from within the R system. Current information is available on the web site of The BUGS Project, http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml. Unlike SAS or R, BUGS does not have packages for particular classes of commonly used models analogous to MIXED or lme4, but rather provides a flexible programming environment in which models can be specified. BUGS's internals determine an MCMC routine and then run it. The latter is no small feat and the BUGS team has been extremely open about their blunders, for all of which they deserve immense credit. However, this general design means that you have to write a new program for every new model you want to fit instead of being able, as in SAS or lme4, to specify entirely new models very quickly using an interface specialized for mixed linear models. No doubt if I were a better programmer I would find this aspect of BUGS less onerous but for me it is a real disincentive to doing Bayesian analyses. I only make the effort of doing a Bayesian analysis in collaborative work when I think it likely to give a substantially different result from the conventional analysis which could affect the answer to the subject-matter question, as in the viral-structure example in Section 1.4.2. BUGS uses Gibbs or Metropolis-Hastings sampling, generally sampling one unknown at a time although it has a limited facility for sampling multivariate normal vectors. In this sense, BUGS's great strength, its generality, is also its great weakness because it is unlikely to give an efficient MCMC algorithm for any given problem. However, it is usually much easier than writing your own MCMC sampler in R or C++ and far less likely to suffer from coding errors. The R system also has a contributed MCMC package called JAGS, written by Martyn Plummer, which has nearly the same syntax as BUGS but has very different internals so it can be used as a check on BUGS. JAGS has the advantage that it will run on any platform that R runs on, so a computer ignoramus like me can run it on the Macintosh operating system without having to figure out how to use the Mac's Windows emulator, which I would need to use the BRugs package to access WinBUGS. JAGS documentation is available from the CRAN R web site.

Those are the software systems I have used. Other popular packages with extensive facilities for doing the conventional analysis include the general-purpose packages SPSS and STATA. Two popular specialized packages are HLM and MLwiN. These were originally designed to do analyses for hierarchical or multilevel models, which is a special case of mixed linear models for normal errors, though both systems handle some families of non-normal errors. Both systems do some form of both conventional as well as Bayesian analyses though neither does Bayesian analyses for the range of models that BUGS handles. Snijders & Bosker (2012, Chapter 18) discuss these packages in detail. Also, Rue and his collaborators (Rue et al. 2009) have developed an R package for doing approximate Bayesian calculations using integrated nested Laplace approximations (INLA); Section 7.2 gives a bit more information.

If you want to do conventional analyses and are already comfortable in SPSS and STATA you are probably better off learning their mixed linear model functionality than switching to a new system. If you are keen to do Bayesian analyses, you probably need to learn BUGS or JAGS if you want to do Bayesian analyses for classes of models besides mixed linear models (unless you are a good or keen programmer, I suppose).

Exercises

Regular Exercises

1. (Section 1.2.2) Derive the conventional point estimate (1.8) of $(\beta, \mathbf{u})$ given $\mathbf{G}$ and $\mathbf{R}$.
2. (Section 1.2.3) Neyman-Scott paradox: If Y_{i1} and Y_{i2} are iid $N(\alpha_i, \sigma_e^2)$ for each of $i = 1, \ldots, n$, show that the MLE for σ_e^2 is $\sum_i (Y_{i1} - Y_{i2})^2 / 4n$ and show that it has expected value $\sigma_e^2/2$ for all n. Show that the maximum restricted likelihood estimate for σ_e^2 is unbiased. Now turn this into a mixed linear model by adding $\alpha_i \sim$ iid $N(0, \sigma_s^2)$. Show that the MLE for σ_e^2 in this mixed linear model is unbiased — using the expectation over the distributions of $\mathbf{u}$ and ε — but the MLE for σ_s^2 has a bias of order $1/n$. Finally, show that maximizing the restricted likelihood for this mixed linear model gives unbiased estimates for both variances. In deriving these results, allow the variance estimates to take negative values. (Hint: Re-parameterize (σ_e^2, σ_s^2) as (σ_e^2, γ), where $\gamma = \sigma_e^2 + 2\sigma_s^2$.)
3. (Section 1.2.4.1) Derive $\operatorname{cov}\begin{pmatrix} \hat{\beta} \\ \hat{\mathbf{u}} - \mathbf{u} \end{pmatrix}$ in (1.19).
4. (Section 1.3.2.3) For the simple problem where $X_i, i = 1, \ldots, n$ are iid $N(0, 1/\tau)$, τ being the precision, derive the posterior distribution of τ for the prior $\tau \sim Gamma(a, b)$ with density $\pi(\tau) \propto \tau^{a-1} e^{-b\tau}$, so τ has prior mean a/b. Note the role n plays in τ's posterior and use that to interpret the parameter a of the gamma prior as the "value" of the prior measured in terms of observations. Note also how b enters into the posterior. I think this, more than anything else, accounts for the popularity of the gamma(ε,ε) prior for precisions: It appears to contribute almost no information to the posterior. But it is a *very* weird distribution.
5. (Section 1.3.3) For the Bayesian analysis, derive the marginal posterior of ϕ for a general prior on ϕ, i.e., the result should have the form (function of ϕ) $\times$ (prior on ϕ). Hint: Integrate out $(\beta, \mathbf{u})$ with a single integral.
6. (Following the preceding exercise.) Assume $\mathbf{R} = \sigma_e^2 \mathbf{I}$. Re-parameterize ϕ from (σ_e^2, ϕ_G) to (σ_e^2, ϕ_G^*), where ϕ_G^* is the vector of unknowns in $\mathbf{G}/\sigma_e^2$. Assume $1/\sigma_e^2$ has a gamma prior distribution, independent of the prior for ϕ_G^*. Derive the marginal posterior of ϕ_G^*. Hint: If this seems hard, you're doing it the wrong way.
7. (Section 1.3.3) For the blocked Gibbs sampler given in this section, derive the full conditional posteriors of $(\beta, \mathbf{u})$, ϕ_G, ϕ_R. This should be easy, but if you're getting stuck, Ruppert et al. (2003) Section 16.3 shows the full conditionals for a special case.

Open Questions

1. (Section 1.2.4.2) For an interesting class of models, e.g., in one of Chapters 3 to 6, derive better large-sample approximations for the various test statistics and pivotal quantities used in the conventional theory.
2. (Sections 1.4.1 and 1.4.2) For cases in which maximizing the restricted likelihood gives a zero estimate for a variance, derive a simple diagnostic test that detects when the restricted likelihood is quite flat near zero, and a simple one-sided interval for that variance with reasonably accurate coverage.
3. (Sections 1.3.2.3 and 1.3.2.4) For an interesting class of models, find a prior for ϕ that gives point estimates with good properties and posterior intervals with at worst near-nominal frequentist coverage. For an example of what such a project might look like, see He et al. (2007) and He & Hodges (2008). I consider these papers *components* of an examination of a prior's performance but by no means complete.
4. (Section 1.3.2.1) For an interesting class of models, examine the effect of priors for (β, ϕ) on Bayes factors for testing effects with multiple degrees of freedom, and if possible, specify tests using these priors with more or less reliable frequentist properties.
5. (Section 1.3.2.1) Derive an approximate Bayes factor for an effect in a mixed linear model based on the Schwarz Criterion/BIC approximation in Kass & Raftery (1995, p. 778). See if you can extend this to testing elements of ϕ. Be careful: The asymptotic theory that rationalizes this approximation may not apply for some tests of interest.
6. (Section 1.3.2.3 and 1.4.1) For an interesting class of models, examine the effect of priors for (β, ϕ) on Bayes factors for testing whether a variance component in ϕ_G is zero, and if possible, specify tests with more or less reliable frequentist properties.