Douglas M. Bates

lme4: Mixed-effects modeling with R

February 19, 2010

Springer

Contents

List of Figures

Chapter 1
A Simple, Linear, Mixed-effects Model

In this book we describe the theory behind a type of statistical model called *mixed-effects* models and the practice of fitting and analyzing such models using the `lme4` package for R. These models are used in many different disciplines. Because the descriptions of the models can vary markedly between disciplines, we begin by describing what mixed-effects models are and by exploring a very simple example of one type of mixed model, the *linear mixed model*.

This simple example allows us to illustrate the use of the `lmer` function in the `lme4` package for fitting such models and for analyzing the fitted model. We describe methods of assessing the precision of the parameter estimates and of visualizing the conditional distribution of the random effects, given the observed data.

1.1 Mixed-effects Models

Mixed-effects models, like many other types of statistical models, describe a relationship between a *response* variable and some of the *covariates* that have been measured or observed along with the response. In mixed-effects models at least one of the covariates is a *categorical* covariate representing experimental or observational "units" in the data set. In the example from the chemical industry that is given in this chapter, the observational unit is the batch of an intermediate product used in production of a dye. In medical and social sciences the observational units are often the human or animal subjects in the study. In agriculture the experimental units may be the plots of land or the specific plants being studied.

In all of these cases the categorical covariate or covariates are observed at a set of discrete *levels*. We may use numbers, such as subject identifiers, to designate the particular levels that we observed but these numbers are simply labels. The important characteristic of a categorical covariate is that, at each

observed value of the response, the covariate takes on the value of one of a set of distinct levels.

Parameters associated with the particular levels of a covariate are sometimes called the "effects" of the levels. If the set of possible levels of the covariate is fixed and reproducible we model the covariate using *fixed-effects* parameters. If the levels that we observed represent a random sample from the set of all possible levels we incorporate *random effects* in the model.

There are two things to notice about this distinction between fixed-effects parameters and random effects. First, the names are misleading because the distinction between fixed and random is more a property of the levels of the categorical covariate than a property of the effects associated with them. Secondly, we distinguish between "fixed-effects parameters", which are indeed parameters in the statistical model, and "random effects", which, strictly speaking, are not parameters. As we will see shortly, random effects are unobserved random variables.

To make the distinction more concrete, suppose that we wish to model the annual reading test scores for students in a school district and that the covariates recorded with the score include a student identifier and the student's gender. Both of these are categorical covariates. The levels of the gender covariate, male and female, are fixed. If we consider data from another school district or we incorporate scores from earlier tests, we will not change those levels. On the other hand, the students whose scores we observed would generally be regarded as a sample from the set of all possible students whom we could have observed. Adding more data, either from more school districts or from results on previous or subsequent tests, will increase the number of distinct levels of the student identifier.

Mixed-effects models or, more simply, *mixed models* are statistical models that incorporate both fixed-effects parameters and random effects. Because of the way that we will define random effects, a model with random effects always includes at least one fixed-effects parameter. Thus, any model with random effects is a mixed model.

We characterize the statistical model in terms of two random variables: a q-dimensional vector of random effects represented by the random variable $\mathscr{B}$ and an n-dimensional response vector represented by the random variable $\mathscr{Y}$. (We use upper-case "script" characters to denote random variables. The corresponding lower-case upright letter denotes a particular value of the random variable.) We observe the value, $\mathbf{y}$, of $\mathscr{Y}$. We do not observe the value of $\mathscr{B}$.

When formulating the model we describe the unconditional distribution of $\mathscr{B}$ and the conditional distribution, $(\mathscr{Y}|\mathscr{B} = \mathbf{b})$. The descriptions of the distributions involve the form of the distribution and the values of certain parameters. We use the observed values of the response and the covariates to estimate these parameters and to make inferences about them.

That's the big picture. Now let's make this more concrete by describing a particular, versatile class of mixed models called linear mixed models and by

studying a simple example of such a model. First we will describe the data in the example.

1.2 The Dyestuff and Dyestuff2 Data

Models with random effects have been in use for a long time. The first edition of the classic book, *Statistical Methods in Research and Production*, edited by O.L. Davies, was published in 1947 and contained examples of the use of random effects to characterize batch-to-batch variability in chemical processes. The data from one of these examples are available as the `Dyestuff` data in the `lme4` package. In this section we describe and plot these data and introduce a second example, the `Dyestuff2` data, described in Box and Tiao [1973].

1.2.1 The Dyestuff *Data*

The `Dyestuff` data are described in Davies and Goldsmith [1972, Table 6.3, p. 131], the fourth edition of the book mentioned above, as coming from

> an investigation to find out how much the variation from batch to batch in the quality of an intermediate product (H-acid) contributes to the variation in the yield of the dyestuff (Naphthalene Black 12B) made from it. In the experiment six samples of the intermediate, representing different batches of works manufacture, were obtained, and five preparations of the dyestuff were made in the laboratory from each sample. The equivalent yield of each preparation as grams of standard colour was determined by dye-trial.

To access these data within R we must first attach the `lme4` package to our session using

```
> library(lme4)
```

Note that the `">"` symbol in the line shown is the prompt in R and not part of what the user types. The `lme4` package must be attached before any of the data sets or functions in the package can be used. If typing this line results in an error report stating that there is no package by this name then you must first install the package.

In what follows, we will assume that the `lme4` package has been installed and that it has been attached to the R session before any of the code shown has been run.

The `str` function in R provides a concise description of the structure of the data

```
> str(Dyestuff)
```

```
'data.frame':        30 obs. of  2 variables:
 $ Batch: Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 2 2 2 2 2 ...
 $ Yield: num  1545 1440 1440 1520 1580 ...
```

from which we see that it consists of 30 observations of the `Yield`, the response variable, and of the covariate, `Batch`, which is a categorical variable stored as a `factor` object. If the labels for the factor levels are arbitrary, as they are here, we will use letters instead of numbers for the labels. That is, we label the batches as `"A"` through `"F"` rather than `"1"` through `"6"`. When the labels are letters it is clear that the variable is categorical. When the labels are numbers a categorical covariate can be mistaken for a numeric covariate, with unintended consequences.

It is a good practice to apply `str` to any data frame the first time you work with it and to check carefully that any categorical variables are indeed represented as factors.

The data in a data frame are viewed as a table with columns corresponding to variables and rows to observations. The functions `head` and `tail` print the first or last few rows (the default value of "few" happens to be 6 but we can specify another value if we so choose)

```
> head(Dyestuff)
```

```
  Batch Yield
1     A  1545
2     A  1440
3     A  1440
4     A  1520
5     A  1580
6     B  1540
```

or we could ask for a `summary` of the data

```
> summary(Dyestuff)
```

```
 Batch     Yield
 A:5   Min.   :1440
 B:5   1st Qu.:1469
 C:5   Median :1530
 D:5   Mean   :1528
 E:5   3rd Qu.:1575
 F:5   Max.   :1635
```

Although the `summary` does show us an important property of the data, namely that there are exactly 5 observations on each batch — a property that we will describe by saying that the data are *balanced* with respect to `Batch` — we usually learn much more about the structure of such data from plots like Fig. 1.1 than we can from numerical summaries.

In Fig. 1.1 we can see that there is considerable variability in yield, even for preparations from the same batch, but there is also noticeable batch-to-batch variability. For example, four of the five preparations from batch F provided lower yields than did any of the preparations from batches C and E.

This plot, and essentially all the other plots in this book, were created using Deepayan Sarkar's `lattice` package for R. In Sarkar [2008] he describes how one would create such a plot. Because this book was created using Sweave

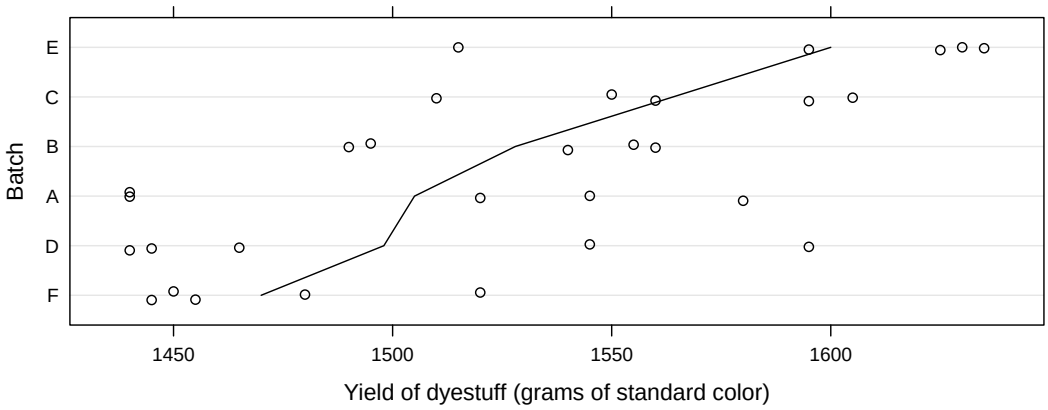


Fig. 1.1 Yield of dyestuff (Napthalene Black 12B) for 5 preparations from each of 6 batches of an intermediate product (H-acid). The line joins the mean yields from the batches, which have been ordered by increasing mean yield. The vertical positions are "jittered" slightly to avoid over-plotting. Notice that the lowest yield for batch A was observed for two distinct preparations from that batch.

[Leisch, 2002], the exact code used to create the plot, as well as the code for all the other figures and calculations in the book, is available on the web site for the book. In Sect. **??** we review some of the principles of lattice graphics, such as reordering the levels of the `Batch` factor by increasing mean response, that enhance the informativeness of the plot. At this point we will concentrate on the information conveyed by the plot and not on how the plot is created.

In Sect. 1.3.1 we will use mixed models to quantify the variability in yield between batches. For the time being let us just note that the particular batches used in this experiment are a selection or sample from the set of all batches that we wish to consider. Furthermore, the extent to which one particular batch tends to increase or decrease the mean yield of the process — in other words, the "effect" of that particular batch on the yield — is not as interesting to us as is the extent of the variability between batches. For the purposes of designing, monitoring and controlling a process we want to predict the yield from future batches, taking into account the batch-to-batch variability and the within-batch variability. Being able to estimate the extent to which a particular batch in the past increased or decreased the yield is not usually an important goal for us. We will model the effects of the batches as random effects rather than as fixed-effects parameters.

1.2.2 The `Dyestuff2` *Data*

The `Dyestuff2` data are simulated data presented in Box and Tiao [1973, Table 5.1.4, p. 247] where the authors state

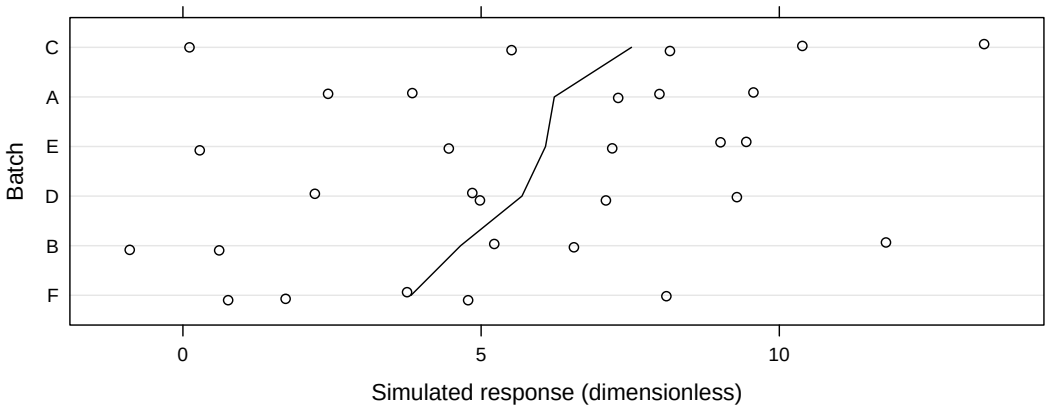


Fig. 1.2 Simulated data presented in Box and Tiao [1973] with a structure similar to that of the `Dyestuff` data. These data represent a case where the batch-to-batch variability is small relative to the within-batch variability.

These data had to be constructed for although examples of this sort undoubtedly occur in practice they seem to be rarely published.

The structure and summary

```
> str(Dyestuff2)

'data.frame':        30 obs. of  2 variables:
 $ Batch: Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 2 2 2 2 2 ...
 $ Yield: num  7.3 3.85 2.43 9.57 7.99 ...

> summary(Dyestuff2)

 Batch     Yield
 A:5   Min.   :-0.892
 B:5   1st Qu.: 2.765
 C:5   Median : 5.365
 D:5   Mean   : 5.666
 E:5   3rd Qu.: 8.151
 F:5   Max.   :13.434
```

are intentionally similar to those of the `Dyestuff` data. As can be seen in Fig. 1.2 the batch-to-batch variability in these data is small compared to the within-batch variability. In some approaches to mixed models it can be difficult to fit models to such data. Paradoxically, small "variance components" can be more difficult to estimate than large variance components.

The methods we will present are not compromised when estimating small variance components.

1.3 Fitting Linear Mixed Models

Before we formally define a linear mixed model, let's go ahead and fit models to these data sets using `lmer`. Like most model-fitting functions in R, `lmer` takes, as its first two arguments, a *formula* specifying the model and the *data* with which to evaluate the formula. This second argument, `data`, is optional but recommended. It is usually the name of a data frame, such as those we examined in the last section. Throughout this book all model specifications will be given in this formula/data format.

We will explain the structure of the formula after we have considered an example.

1.3.1 A Model For the `Dyestuff` *Data*

We fit a model to the `Dyestuff` data allowing for an overall level of the `Yield` and for an additive random effect for each level of `Batch`

```
> fm1 <- lmer(Yield ~ 1 + (1|Batch), Dyestuff)
> print(fm1)
```

```
Linear mixed model fit by REML
Formula: Yield ~ 1 + (1 | Batch)
   Data: Dyestuff
 REML
319.7
Random effects:
 Groups   Name        Variance Std.Dev.
 Batch    (Intercept) 1764.0   42.001
 Residual             2451.3   49.510
Number of obs: 30, groups: Batch, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  1527.50      19.38    78.8
```

In the first line we call the `lmer` function to fit a model with formula

```
Yield ~ 1 + (1 | Batch)
```

applied to the `Dyestuff` data and assign the result to the name `fm1`. (The name is arbitrary. I happen to use names that start with `fm`, indicating "fitted model".)

As is customary in R, there is no output shown after this assignment. We have simply saved the fitted model as an object named `fm1`. In the second line we display some information about the fitted model by applying `print` to `fm1`. In later examples we will condense these two steps into one but here it helps to emphasize that we save the result of fitting a model then apply various *extractor* functions to the fitted model to get a brief summary of the model fit or to obtain the values of some of the estimated quantities.

1.3.1.1 Details of the Printed Display

The printed display of a model fit with `lmer` has four major sections: a description of the model that was fit, some statistics characterizing the model fit, a summary of properties of the random effects and a summary of the fixed-effects parameter estimates. We consider each of these sections in turn.

The description section states that this is a linear mixed model in which the parameters have been estimated as those that minimize the REML criterion (explained in Sect. 5.5). The `formula` and `data` arguments are displayed for later reference. If other, optional arguments affecting the fit, such as a `subset` specification, were used, they too will be displayed here.

For models fit by the REML criterion the only statistic describing the model fit is the value of the REML criterion itself. An alternative set of parameter estimates, the maximum likelihood estimates, are obtained by specifying the optional argument `REML = FALSE`.

```
> (fm1ML <- lmer(Yield ~ 1 + (1|Batch), Dyestuff, REML = FALSE))

Linear mixed model fit by maximum likelihood
Formula: Yield ~ 1 + (1 | Batch)
   Data: Dyestuff
   AIC   BIC logLik deviance
 333.3 337.5 -163.7    327.3
Random effects:
 Groups   Name        Variance Std.Dev.
 Batch    (Intercept) 1388.3   37.26
 Residual             2451.3   49.51
Number of obs: 30, groups: Batch, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  1527.50      17.69   86.33
```

(Notice that this code fragment also illustrates a way to condense the assignment and the display of the fitted model into a single step. The redundant set of parentheses surrounding the assignment causes the result of the assignment to be displayed. We will use this device often in what follows.)

The display of a model fit by maximum likelihood provides several other model-fit statistics such as Akaike's Information Criterion (`AIC`) [Sakamoto et al., 1986], Schwarz's Bayesian Information Criterion (`BIC`) [Schwarz, 1978], the log-likelihood (`logLik`) at the parameter estimates, and the deviance (negative twice the log-likelihood) at the parameter estimates. These are all statistics related to the model fit and are used to compare different models fit to the same data.

At this point the important thing to note is that the default estimation criterion is the REML criterion. Generally the REML estimates of variance components are preferred to the ML estimates. However, when comparing models it is safest to refit all the models using the maximum likelihood criterion. We will discuss comparisons of model fits in Sect. 2.2.4.

The third section is the table of estimates of parameters associated with the random effects. There are two sources of variability in the model we have fit, a batch-to-batch variability in the level of the response and the residual or per-observation variability — also called the within-batch variability. The name "residual" is used in statistical modeling to denote the part of the variability that cannot be explained or modeled with the other terms. It is the variation in the observed data that is "left over" after we have determined the estimates of the parameters in the other parts of the model.

Some of the variability in the response is associated with the fixed-effects terms. In this model there is only one such term, labeled as the `(Intercept)`. The name "intercept", which is better suited to models based on straight lines written in a slope/intercept form, should be understood to represent an overall "typical" or mean level of the response in this case. (In case you are wondering about the parentheses around the name, they are included so that you can't accidentally create a variable with a name that conflicts with this name.) The line labeled `Batch` in the random effects table shows that the random effects added to the `(Intercept)` term, one for each level of the `Batch` factor, are modeled as random variables whose unconditional variance is estimated as 1764.05 g^2 in the REML fit and as 1388.33 g^2 in the ML fit. The corresponding standard deviations are 42.00 g for the REML fit and 37.26 g for the ML fit.

Note that the last column in the random effects summary table is the estimate of the variability expressed as a standard deviation rather than as a variance. These are provided because it is usually easier to visualize standard deviations, which are on the scale of the response, than it is to visualize the magnitude of a variance. The values in this column are a simple re-expression (the square root) of the estimated variances. Do not confuse them with the standard errors of the variance estimators, which are not given here. In Sect. 1.5 we explain why we do not provide standard errors of variance estimates.

The line labeled `Residual` in this table gives the estimate of the variance of the residuals (also in g^2) and its corresponding standard deviation. For the REML fit the estimated standard deviation of the residuals is 49.51 g and for the ML fit it is also 49.51 g (Generally these estimates do not need to be equal. They happen to be equal in this case because of the simple model form and the balanced data set.)

The last line in the random effects table states the number of observations to which the model was fit and the number of levels of any "grouping factors" for the random effects. In this case we have a single random effects term, `(1|Batch)`, in the model formula and the grouping factor for that term is `Batch`. There will be a total of six random effects, one for each level of `Batch`.

The final part of the printed display gives the estimates and standard errors of any fixed-effects parameters in the model. The only fixed-effects term in the model formula is the `1`, denoting a constant which, as explained above, is labeled as `(Intercept)`. For both the REML and the ML estimation criterion

the estimate of this parameter is 1527.5 g (equality is again a consequence of the simple model and balanced data set). The standard error of the intercept estimate is 19.38 g for the REML fit and 17.69 g for the ML fit.

1.3.2 A Model For the `Dyestuff2` *Data*

Fitting a similar model to the `Dyestuff2` data produces an estimate $\widehat{\sigma}_1 = 0$ in both the REML

```
> (fm2 <- lmer(Yield ~ 1 + (1|Batch), Dyestuff2))

Linear mixed model fit by REML
Formula: Yield ~ 1 + (1 | Batch)
   Data: Dyestuff2
 REML
161.8
Random effects:
 Groups   Name        Variance Std.Dev.
 Batch    (Intercept)  0.000   0.0000
 Residual             13.806   3.7157
Number of obs: 30, groups: Batch, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)   5.6656     0.6784   8.352
```

and the ML fits.

```
> (fm2ML <- update(fm2, REML = FALSE))

Linear mixed model fit by maximum likelihood
Formula: Yield ~ 1 + (1 | Batch)
   Data: Dyestuff2
   AIC   BIC logLik deviance
 168.9 173.1 -81.44    162.9
Random effects:
 Groups   Name        Variance Std.Dev.
 Batch    (Intercept)  0.000   0.0000
 Residual             13.346   3.6532
Number of obs: 30, groups: Batch, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)    5.666      0.667   8.494
```

(Note the use of the `update` function to re-fit a model changing some of the arguments. In a case like this, where the call to fit the original model is not very complicated, the use of `update` is not that much simpler than repeating the original call to `lmer` with extra arguments. For complicated model fits it can be.)

An estimate of 0 for σ_1 does not mean that there is no variation between the groups. Indeed Fig. 1.2 shows that there is some small amount of variability between the groups. The estimate, $\widehat{\sigma}_1 = 0$, simply indicates that the level of "between-group" variability is not sufficient to warrant incorporating random effects in the model.

The important point to take away from this example is that we must allow for the estimates of variance components to be zero. We describe such a model as being degenerate, in the sense that it corresponds to a linear model in which we have removed the random effects associated with `Batch`. Degenerate models can and do occur in practice. Even when the final fitted model is not degenerate, we must allow for such models when determining the parameter estimates through numerical optimization.

To reiterate, the model `fm2` corresponds to the linear model

```
> summary(fm2a <- lm(Yield ~ 1, Dyestuff2))

Call:
lm(formula = Yield ~ 1, data = Dyestuff2)
Residuals:
    Min      1Q  Median      3Q     Max
-6.5576 -2.9006 -0.3006  2.4854  7.7684

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.6656     0.6784   8.352 3.32e-09

Residual standard error: 3.716 on 29 degrees of freedom
```

because the random effects are inert, in the sense that they have a variance of zero, and can be removed.

Notice that the estimate of σ from the linear model (called the `Residual standard error` in the output corresponds to the estimate in the REML fit (`fm2`) but not that from the ML fit (`fm2ML`). The fact that the REML estimates of variance components in mixed models generalize the estimate of the variance used in linear models, in the sense that these estimates coincide in the degenerate case, is part of the motivation for the use of the REML criterion for fitting mixed-effects models.

1.3.3 Further Assessment of the Fitted Models

The parameter estimates in a statistical model represent our "best guess" at the unknown values of the model parameters and, as such, are important results in statistical modeling. However, they are not the whole story. Statistical models characterize the variability in the data and we must assess the effect of this variability on the parameter estimates and on the precision of predictions made from the model.

In Sect. 1.5 we introduce a method of assessing variability in parameter estimates using the "profiled deviance" and in Sect. 1.6 we show methods of characterizing the conditional distribution of the random effects given the data. Before we get to these sections, however, we should state in some detail the probability model for linear mixed-effects and establish some definitions and notation. In particular, before we can discuss profiling the deviance, we should define the deviance. We do that in the next section.

1.4 The Linear Mixed-effects Probability Model

In explaining some of parameter estimates related to the random effects we have used terms such as "unconditional distribution" from the theory of probability. Before proceeding further we should clarify the linear mixed-effects probability model and define several terms and concepts that will be used throughout the book.

1.4.1 Definitions and Results

In this section we provide some definitions and formulas without derivation and with minimal explanation, so that we can use these terms in what follows. In Chapter 5 we revisit these definitions providing derivations and more explanation.

As mentioned in Sect. 1.1, a mixed model incorporates two random variables: $\mathscr{B}$, the q-dimensional vector of random effects, and $\mathscr{Y}$, the n-dimensional response vector. In a linear mixed model the unconditional distribution of $\mathscr{B}$ and the conditional distribution, $(\mathscr{Y}|\mathscr{B}=\mathbf{b})$, are both multivariate Gaussian (or "normal") distributions,

$$\begin{aligned}(\mathscr{Y}|\mathscr{B}=\mathbf{b}) &\sim \mathscr{N}(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{b},\sigma^2\mathbf{I})\\ \mathscr{B} &\sim \mathscr{N}(\mathbf{0},\Sigma_{\boldsymbol{\theta}}).\end{aligned} \tag{1.1}$$

The *conditional mean* of $\mathscr{Y}$, given $\mathscr{B}=\mathbf{b}$, is the *linear predictor*, $\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{b}$, which depends on the p-dimensional *fixed-effects parameter*, $\boldsymbol{\beta}$, and on $\mathbf{b}$. The *model matrices*, $\mathbf{X}$ and $\mathbf{Z}$, of dimension $n\times p$ and $n\times q$, respectively, are determined from the formula for the model and the values of covariates. Although the matrix $\mathbf{Z}$ can be large (i.e. both n and q can be large), it is sparse (i.e. most of the elements in the matrix are zero).

The *relative covariance factor*, $\Lambda_{\boldsymbol{\theta}}$ is a $q\times q$ matrix, depending on the *variance-component parameter*, $\boldsymbol{\theta}$, and generating the symmetric $q\times q$ variance-covariance matrix, $\Sigma_{\boldsymbol{\theta}}$, according to

$$\Sigma_\theta = \sigma^2 \Lambda_\theta \Lambda_\theta^\mathsf{T}. \tag{1.2}$$

The *spherical random effects*, $\mathscr{U} \sim \mathscr{N}(\mathbf{0}, \sigma^2 \mathbf{I}_q)$, determine $\mathscr{B}$ according to

$$\mathscr{B} = \Lambda_\theta \mathscr{U}.$$

The *penalized residual sum of squares* (PRSS),

$$r^2(\theta, \beta, \mathbf{u}) = \left\{ \|\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta \mathbf{u}\|^2 + \|\mathbf{u}\|^2 \right\}, \tag{1.3}$$

is the sum of the residual sum of squares, measuring fidelity of the model to the data, and a penalty on the size of $\mathbf{u}$, measuring the complexity of the model. Minimizing r^2 with respect to $\mathbf{u}$,

$$r^2_{\beta,\theta} = \min_{\mathbf{u}} \left\{ \|\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta \mathbf{u}\|^2 + \|\mathbf{u}\|^2 \right\} \tag{1.4}$$

is a direct (i.e. non-iterative) computation for which we calculate the *sparse Cholesky factor*, $\mathbf{L}_\theta$, which is a lower triangular $q \times q$ matrix satisfying

$$\mathbf{L}_\theta \mathbf{L}_\theta^\mathsf{T} = \Lambda_\theta^\mathsf{T} \mathbf{Z}^\mathsf{T} \mathbf{Z} \Lambda_\theta + \mathbf{I}_q. \tag{1.5}$$

where $\mathbf{I}_q$ is the $q \times q$ *identity matrix*.

The *deviance* (negative twice the log-likelihood) of the parameters, given the data, $\mathbf{y}$, is

$$d(\theta, \beta, \sigma | \mathbf{y}) = n \log(2\pi\sigma^2) + \log(|\mathbf{L}_\theta|^2) + \frac{r^2_{\beta,\theta}}{\sigma^2}. \tag{1.6}$$

where $|\mathbf{L}_\theta|$ denotes the *determinant* of $\mathbf{L}_\theta$. Because $\mathbf{L}_\theta$ is triangular, its determinant is the product of its diagonal elements.

Because the conditional mean, μ, is a linear function of β and $\mathbf{u}$, minimization of the PRSS with respect to both β and $\mathbf{u}$ to produce

$$r^2_\theta = \min_{\beta,\mathbf{u}} \left\{ \|\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta \mathbf{u}\|^2 + \|\mathbf{u}\|^2 \right\} \tag{1.7}$$

is also a direct calculation. The values of $\mathbf{u}$ and β that provide this minimum are called, respectively, the *conditional mode*, $\tilde{\mathbf{u}}_\theta$, of the spherical random effects and the conditional estimate, $\widehat{\beta}_\theta$, of the fixed effects. At the conditional estimate of the fixed effects the deviance is

$$d(\theta, \widehat{\beta}_\theta, \sigma | \mathbf{y}) = n \log(2\pi\sigma^2) + \log(|\mathbf{L}_\theta|^2) + \frac{r^2_\theta}{\sigma^2}. \tag{1.8}$$

Minimizing this expression with respect to σ^2 produces the conditional estimate

$$\widehat{\sigma^2}_\theta = \frac{r^2_\theta}{n} \tag{1.9}$$

which provides the *profiled deviance*,

$$\tilde{d}(\theta|\mathbf{y}) = d(\theta, \widehat{\beta}_\theta, \widehat{\sigma}_\theta|\mathbf{y}) = \log(|\mathbf{L}_\theta|^2) + n\left[1 + \log\left(\frac{2\pi r_\theta^2}{n}\right)\right], \tag{1.10}$$

a function of θ alone.

The *maximum likelihood estimate* (MLE) of θ, written $\widehat{\theta}$, is the value that minimizes the profiled deviance (1.10). We determine this value by numerical optimization. In the process of evaluating $\tilde{d}(\widehat{\theta}|\mathbf{y})$ we determine $\widehat{\beta}$, $\tilde{\mathbf{u}}_{\widehat{\theta}}$ and $r^2_{\widehat{\theta}}$, from which we can evaluate $\widehat{\sigma} = \sqrt{r^2_{\widehat{\theta}}/n}$.

The elements of the conditional mode of $\mathscr{B}$, evaluated at the parameter estimates,

$$\tilde{b}_{\widehat{\theta}} = \Lambda_{\widehat{\theta}} \tilde{u}_{\widehat{\theta}} \tag{1.11}$$

are sometimes called the *best linear unbiased predictors* or BLUPs of the random effects. Although it has an appealing acronym, I don't find the term particularly instructive (what is a "linear unbiased predictor" and in what sense are these the "best"?) and prefer the term "conditional mode", which is explained in Sect. 1.6.

1.4.2 Matrices and Vectors in the Fitted Model Object

The optional argument, `verbose = TRUE`, in a call to `lmer` produces output showing the progress of the iterative optimization of $\tilde{d}(\theta|\mathbf{y})$.

```
> fm1ML <- lmer(Yield ~ 1|Batch, Dyestuff, REML = FALSE, verbose = TRUE)

  0:     327.76702:  1.00000
  1:     327.35312: 0.807151
  2:     327.33414: 0.725317
  3:     327.32711: 0.754925
  4:     327.32706: 0.752678
  5:     327.32706: 0.752578
  6:     327.32706: 0.752581
```

The algorithm converges in 6 iterations to a profiled deviance of 327.32706 at θ =0.752581.

The actual values of many of the matrices and vectors defined above are available in the *environment* of the fitted model object, accessed with the `env` function. For example, $\Lambda_{\widehat{\theta}}$ is

```
> env(fm1ML)$Lambda

6 x 6 diagonal matrix of class "ddiMatrix"
     [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,] 0.7525806         .         .         .         .         .
[2,]         . 0.7525806         .         .         .         .
[3,]         .         . 0.7525806         .         .         .
```

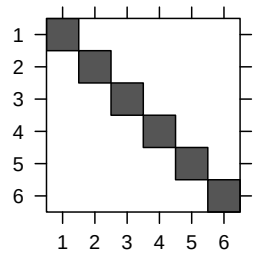


Fig. 1.3 Image of the relative covariance factor, $\Lambda_{\widehat{\theta}}$ for model `fm1ML`. The non-zero elements are shown as darkened squares. The zero elements are blank.

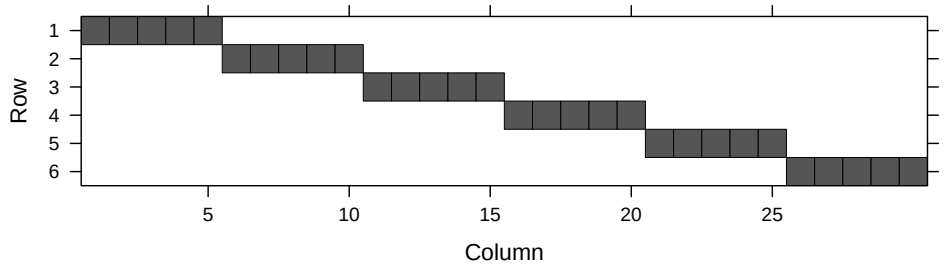


Fig. 1.4 Image of the transpose of the random-effects model matrix, $\mathbf{Z}$, for model `fm1`. The non-zero elements, which are all unity, are shown as darkened squares. The zero elements are blank.

```
[4,]         .         .         . 0.7525806         .         .
[5,]         .         .         .         . 0.7525806         .
[6,]         .         .         .         .         . 0.7525806
```

Often we will show the structure of sparse matrices as an image (Fig. 1.3). Especially for large sparse matrices, the image conveys the structure more compactly than does the printed representation.

In this simple model $\Lambda = \theta \mathbf{I}_6$ is a multiple of the identity matrix and the 30×6 model matrix $\mathbf{Z}$, whose transpose is shown in Fig. 1.4, consists of the indicator columns for `Batch`. Because the data are balanced with respect to `Batch`, the Cholesky factor, $\mathbf{L}$ is also a multiple of the identity (you can check this with `image(env(fm1ML)$L)`). The vectors $\mathbf{u}$ and $\mathbf{b}$ and the matrix $\mathbf{X}$ have the same names in `env(fm1ML)`. The vector β is called `fixef`.

1.5 Assessing the Variability of the Parameter Estimates

In this section we show how to create a *profile deviance* object from a fitted linear mixed model and how to use this object to evaluate confidence intervals on the parameters. We also discuss the construction and interpretation of

profile zeta plots for the parameters and *profile pairs* plots for parameter pairs.

1.5.1 Confidence Intervals on the Parameters

The mixed-effects model fit as `fm1` or `fm1ML` has three parameters for which we obtained estimates. These parameters are σ_1, the standard deviation of the random effects, σ, the standard deviation of the residual or "per-observation" noise term and β_0, the fixed-effects parameter that is labeled as `(Intercept)`.

The `profile` function systematically varies the parameters in a model, assessing the best possible fit that can be obtained with one parameter fixed at a specific value and comparing this fit to the *globally optimal fit*, which is the original model fit that allowed all the parameters to vary. The models are compared according to the change in the deviance, which is the *likelihood ratio test* (LRT) statistic. We apply a *signed square root* transformation to this statistic and plot the resulting function, called ζ, versus the parameter value. A ζ value can be compared to the quantiles of the *standard normal distribution*, $\mathcal{Z} \sim \mathcal{N}(0,1)$. For example, a 95% profile deviance confidence interval on the parameter consists of the values for which $-1.960 < \zeta < 1.960$.

Because the process of profiling a fitted model, which involves re-fitting the model many times, can be computationally intensive, one should exercise caution with complex models fit to very large data sets. Because the statistic of interest is a likelihood ratio, the model is re-fit according to the maximum likelihood criterion, even if the original fit is a REML fit. Thus, there is a slight advantage in starting with an ML fit.

```
> pr1 <- profile(fm1ML)
```

Plots of ζ versus the parameter being profiled (Fig. 1.5) are obtained with

```
> xyplot(pr1, aspect = 1.3)
```

We will refer to such plots as *profile zeta* plots. I usually adjust the aspect ratio of the panels in profile zeta plots to, say, `aspect = 1.3` and frequently set the layout so the panels form a single row (`layout = c(3,1)`, in this case).

The vertical lines in the panels delimit the 50%, 80%, 90%, 95% and 99% confidence intervals, when these intervals can be calculated. Numerical values of the endpoints are returned by the `confint` extractor.

```
> confint(pr1)
```

```
                  2.5 %      97.5 %
.sig01        12.197461   84.063361
.lsig          3.643624    4.214461
(Intercept) 1486.451506 1568.548494
```

By default the 95% confidence interval is returned. The optional argument, `level`, is used to obtain other confidence levels.

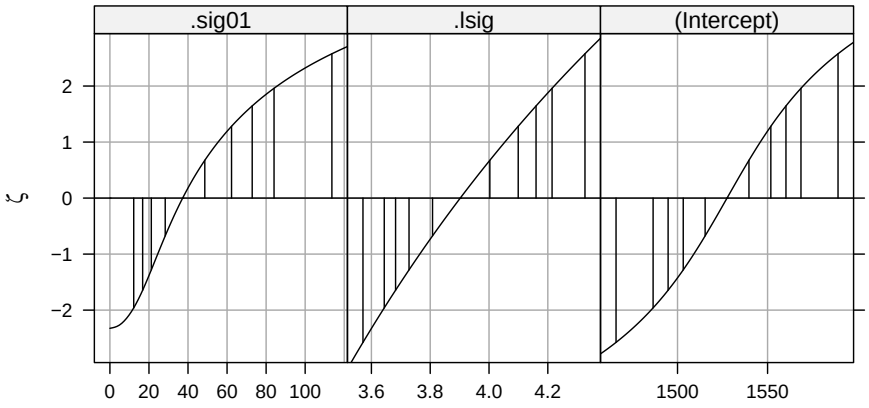


Fig. 1.5 Signed square root, ζ, of the likelihood ratio test statistic for each of the parameters in model `fm1ML`. The vertical lines are the endpoints of 50%, 80%, 90%, 95% and 99% confidence intervals derived from this test statistic.

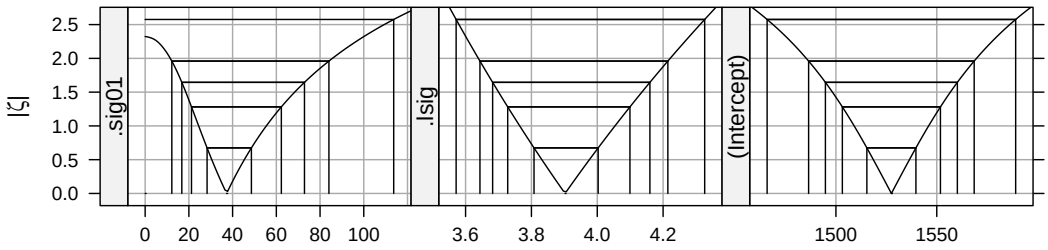


Fig. 1.6 Profiled deviance, on the scale $|\zeta|$, the square root of the change in the deviance, for each of the parameters in model `fm1ML`. The intervals shown are 50%, 80%, 90%, 95% and 99% confidence intervals based on the profile likelihood.

```
> confint(pr1, level = 0.99)

                   0.5 %      99.5 %
.sig01                NA  113.690280
.lsig          3.571290    4.326337
(Intercept) 1465.872875 1589.127125
```

Notice that the lower bound on the 99% confidence interval for σ_1 is not defined. Also notice that we profile $\log(\sigma)$ instead of σ, the residual standard deviation.

A plot of $|\zeta|$, the absolute value of ζ, versus the parameter (Fig. 1.6), obtained by adding the optional argument `absVal = TRUE` to the call to `xyplot`, can be more effective for visualizing the confidence intervals.

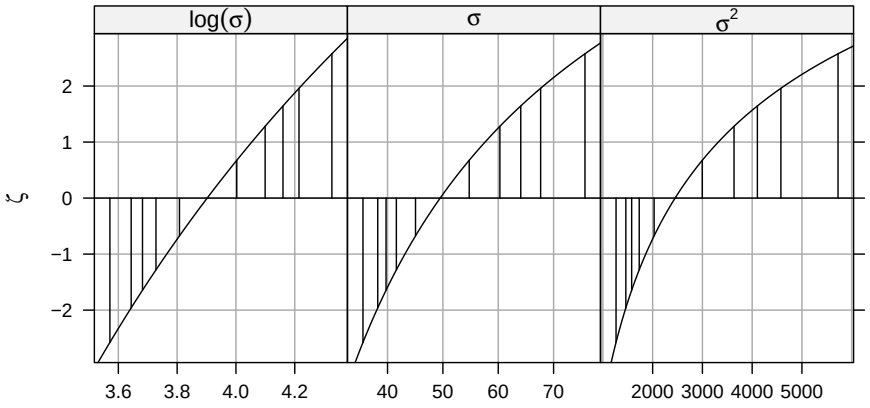


Fig. 1.7 Signed square root, ζ, of the likelihood ratio test statistic as a function of $\log(\sigma)$, of σ and of σ^2. The vertical lines are the endpoints of 50%, 80%, 90%, 95% and 99% confidence intervals.

1.5.2 Interpreting the Profile Zeta Plot

A profile zeta plot, such as Fig. 1.5, shows us the sensitivity of the model fit to changes in the value of particular parameters. Although this is not quite the same as describing the distribution of an estimator, it is a similar idea and we will use some of the terminology from distributions when describing these plots. Essentially we view the patterns in the plots as we would those in a normal probability plot of data values or residuals from a model.

Ideally the profile zeta plot will be close to a straight line over the region of interest, in which case we can perform reliable statistical inference based on the parameter's estimate, its standard error and quantiles of the standard normal distribution. We will describe such as situation as providing a good normal approximation for inference. The common practice of quoting a parameter estimate and its standard error assumes that this is always the case.

In Fig. 1.5 the profile zeta plot for $\log(\sigma)$ is reasonably straight so $\log(\sigma)$ has a good normal approximation. But this does not mean that there is a good normal approximation for σ^2 or even for σ. As shown in Fig. 1.7 the profile zeta plot for $\log(\sigma)$ is slightly skewed, that for σ is moderately skewed and the profile zeta plot for σ^2 is highly skewed. Deviance-based confidence intervals on σ^2 are quite asymmetric, of the form "estimate minus a little, plus a lot".

This should not come as a surprise to anyone who learned in an introductory statistics course that, given a random sample of data assumed to come from a Gaussian distribution, we use a χ^2 distribution, which can be quite skewed, to form a confidence interval on σ^2. Yet somehow there is a widespread belief that the distribution of variance estimators in much more complex situations should be well approximated by a normal distribution.

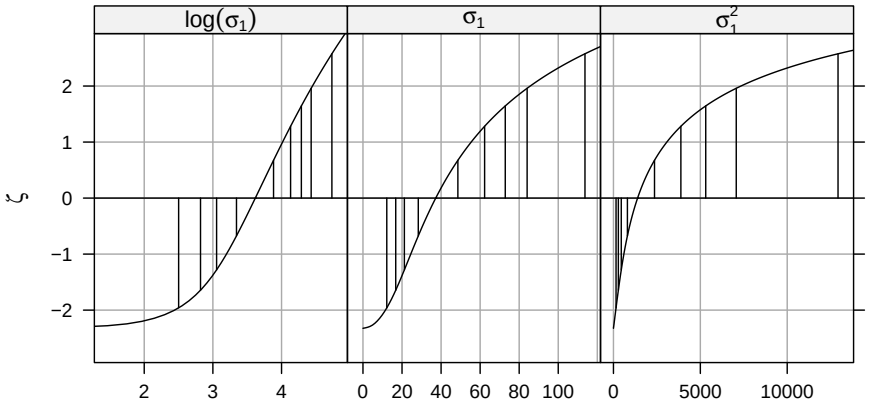


Fig. 1.8 Signed square root, ζ, of the likelihood ratio test statistic as a function of $\log(\sigma_1)$, of σ_1 and of σ_1^2. The vertical lines are the endpoints of 50%, 80%, 90%, 95% and 99% confidence intervals.

It is nonsensical to believe that. In most cases summarizing the precision of a variance component estimate by giving an approximate standard error is woefully inadequate.

The pattern in the profile plot for β_0 is sigmoidal (i.e. an elongated "S"-shape). The pattern is symmetric about the estimate but curved in such a way that the profile-based confidence intervals are wider than those based on a normal approximation. We characterize this pattern as symmetric but over-dispersed (relative to a normal distribution). Again, this pattern is not unexpected. Estimators of the coefficients in a linear model without random effects have a distribution which is a scaled Student's T distribution. That is, they follow a symmetric distribution that is over-dispersed relative to the normal.

The pattern in the profile zeta plot for σ_1 is more complex. Fig. 1.8 shows the profile zeta plot on the scale of $\log(\sigma_1)$, σ_1 and σ_1^2. Notice that the profile zeta plot for $\log(\sigma_1)$ is very close to linear to the right of the estimate but flattens out on the left. That is, σ_1 behaves like σ in that its profile zeta plot is more-or-less a straight line on the logarithmic scale, except when σ_1 is close to zero. The model loses sensitivity to values of σ_1 that are close to zero. If, as in this case, zero is within the "region of interest" then we should expect that the profile zeta plot will flatten out on the left hand side.

1.5.3 Profile Pairs Plots

A profiled deviance object, such as `pr1`, not only provides information on the sensitivity of the model fit to changes in parameters, it also tells us how the parameters influence each other. When we re-fit the model subject to a

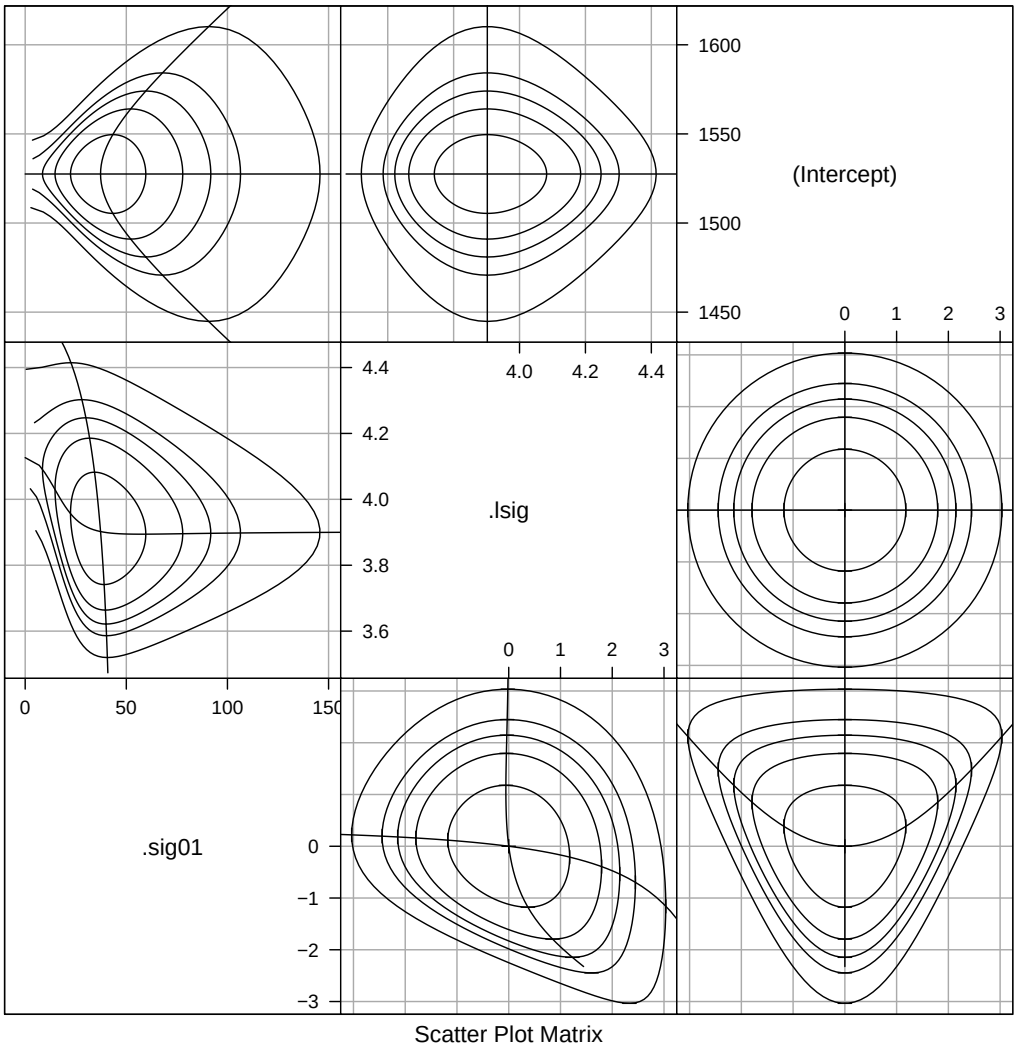


Fig. 1.9 Profile pairs plot for the parameters in model `fm1`. The contour lines correspond to two-dimensional 50%, 80%, 90%, 95% and 99% marginal confidence regions based on the likelihood ratio. Panels below the diagonal represent the (ζ_i, ζ_j) parameters; those above the diagonal represent the original parameters.

constraint such as, say, $\sigma_1 = 60$, we obtain the conditional estimates for the other parameters — σ and β_0 in this case. The conditional estimate of, say, σ as a function of σ_1 is called the *profile trace* of σ on σ_1. Plotting such traces provides valuable information on how the parameters in the model are influenced by each other.

The *profile pairs* plot, obtained as

```
> splom(pr1)
```

and shown in Fig. 1.9 shows the profile traces along with interpolated contours of the two-dimensional profiled deviance function. The contours are chosen to correspond to the two-dimensional marginal confidence regions at particular confidence levels.

Because this plot may be rather confusing at first we will explain what is shown in each panel. To make it easier to refer to panels we assign them (x,y) coordinates, as in a Cartesian coordinate system. The columns are numbered 1 to 3 from left to right and the rows are numbered 1 to 3 from bottom to top. Note that the rows are numbered from the bottom to the top, like the y-axis of a graph, not from top to bottom, like a matrix.

The diagonal panels show the ordering of the parameters: σ_1 first, then $\log(\sigma)$ then β_0. Panels above the diagonal are in the original scale of the parameters. That is, the top-left panel, which is the $(1,3)$ position, has σ_1 on the horizontal axis and β_0 on the vertical axis.

In addition to the contour lines in this panel, there are two other lines, which are the profile traces of σ_1 on β_0 and of β_0 on σ_1. The profile trace of β_0 on σ_1 is a straight horizontal line, indicating that the conditional estimate of β_0, given a value of σ_1, is constant. Again, this is a consequence of the simple model form and the balanced data set. The other line in this panel, which is the profile trace of σ_1 on β_0, is curved. That is, the conditional estimate of σ_1 given β_0 depends on β_0. As β_0 moves away from the estimate, $\widehat{\beta}_0$, in either direction, the conditional estimate of σ_1 increases.

We will refer to the two traces on a panel as the "horizontal trace" and "vertical trace". They are not always perfectly horizontal and vertical lines but the meaning should be clear from the panel because one trace will always be more horizontal and the other will be more vertical. The one that is more horizontal is the trace of the parameter on the y axis as a function of the parameter on the horizontal axis, and vice versa.

The contours shown on the panel are interpolated from the profile zeta function and the profile traces, in the manner described in Bates and Watts [1988, Chapter 6]. One characteristic of a profile trace, which we can verify visually in this panel, is that the tangent to a contour must be vertical where it intersects the horizontal trace and horizontal where it intersects the vertical trace.

The $(2,3)$ panel shows β_0 versus $\log(\sigma)$. In this case the traces actually are horizontal and vertical straight lines. That is, the conditional estimate of β_0 doesn't depend on $\log(\sigma)$ and the conditional estimate of $\log(\sigma)$ doesn't depend on β_0. Even in this case, however, the contour lines are not concentric ellipses, because the deviance is not perfectly quadratic in these parameters. That is, the zeta functions, $\zeta(\beta_0)$ and $\zeta(\log(\sigma))$, are not linear.

The $(1,2)$ panel, showing $\log(\sigma)$ versus σ_1 shows distortion along both axes and nonlinear patterns in both traces. When σ_1 is close to zero the conditional estimate of $\log(\sigma)$ is larger than when σ_1 is large. In other words small values of σ_1 inflate the estimate of $\log(\sigma)$ because the variability that would be explained by the random effects gets incorporated into the residual noise term.

Panels below the diagonal are on the ζ scale, which is why the axes on each of these panels span the same range, approximately -3 to $+3$, and the profile traces always cross at the origin. Thus the $(3,1)$ panel shows $\zeta(\sigma_1)$

on the vertical axis versus $\zeta(\beta_0)$ on the horizontal. These panels allow us to see distortions from an elliptical shape due to nonlinearity of the traces, separately from the one-dimensional distortions caused by a poor choice of scale for the parameter. The ζ scales provide, in some sense, the best possible set of single-parameter transformations for assessing the contours. On the ζ scales the extent of a contour on the horizontal axis is exactly the same as the extent on the vertical axis and both are centered about zero.

Another way to think of this is that, if we would have profiled σ_1^2 instead of σ_1, we would change all the panels in the first column but the panels on the first row would remain the same.

1.6 Assessing the Random Effects

In Sect. 1.4.1 we mentioned that what are sometimes called the BLUPs (or best linear unbiased estimators) of the random effects, $\mathscr{B}$, are the conditional modes evaluated at the parameter estimates, and that they can be calculated as $\tilde{b}_{\widehat{\theta}} = \Lambda_{\widehat{\theta}}\tilde{u}_{\widehat{\theta}}$.

These values are often considered as some sort of "estimates" of the random effects. It can be helpful to think of them this way but it can also be misleading. As we have stated, the random effects are not, strictly speaking, parameters—they are unobserved random variables. We don't estimate the random effects in the same sense that we estimate parameters. Instead, we consider the conditional distribution of $\mathscr{B}$ given the observed data, $(\mathscr{B}|\mathscr{Y} = \mathbf{y})$.

Because the unconditional distribution, $\mathscr{B} \sim \mathscr{N}(\mathbf{0}, \Sigma_\theta)$ is continuous, the conditional distribution, $(\mathscr{B}|\mathscr{Y} = \mathbf{y})$ will also be continuous. In general, the mode of a probability density is the point of maximum density, so the phrase "conditional mode" refers to the point at which this conditional density is maximized. Because this definition relates to the probability model, the values of the parameters are assumed to be known. In practice, of course, we don't know the values of the parameters (if we did there would be no purpose in forming the parameter estimates), so we use the estimated values of the parameters to evaluate the conditional modes.

Those who are familiar with the multivariate Gaussian distribution may recognize that, because both $\mathscr{B}$ and $(\mathscr{Y}|\mathscr{B} = \mathbf{b})$ are multivariate Gaussian, $(\mathscr{B}|\mathscr{Y} = \mathbf{y})$ will also be multivariate Gaussian and the conditional mode will also be the conditional mean of $\mathscr{B}$, given $\mathscr{Y} = \mathbf{y}$. This is the case for a linear mixed model but it does not carry over to other forms of mixed models. In the general case all we can say about $\tilde{\mathbf{u}}$ or $\tilde{\mathbf{b}}$ is that they maximize a conditional density, which is why we use the term "conditional mode" to describe these values. We will only use the term "conditional mean" and the symbol, μ, in reference to $\mathrm{E}(\mathscr{Y}|\mathscr{B} = \mathbf{b})$, which is the conditional mean of $\mathscr{Y}$ given $\mathscr{B}$, and an important part of the formulation of all types of mixed-effects models.

The `ranef` extractor returns the conditional modes.

```
> ranef(fm1ML)

$Batch
  (Intercept)
A  -16.628221
B    0.369516
C   26.974670
D  -21.801445
E   53.579824
F  -42.494343
```

Applying `str` to the result of `ranef`

```
> str(ranef(fm1ML))

List of 1
 $ Batch:'data.frame':        6 obs. of  1 variable:
  ..$ (Intercept): num [1:6] -16.628 0.37 26.975 -21.801 53.58 ...
 - attr(*, "class")= chr "ranef.mer"
```

shows that the value is a list of data frames. In this case the list is of length 1 because there is only one random-effects term, `(1|Batch)`, in the model and, hence, only one grouping factor, `Batch`, for the random effects. There is only one column in this data frame because the random-effects term, `(1|Batch)`, is a simple, scalar term.

To make this more explicit, random-effects terms in the model formula are those that contain the vertical bar (`"|"`) character. The `Batch` variable is the grouping factor for the random effects generated by this term. An expression for the grouping factor, usually just the name of a variable, occurs to the right of the vertical bar. If the expression on the left of the vertical bar is `1`, as it is here, we describe the term as a *simple, scalar, random-effects term*. The designation "scalar" means there will be exactly one random effect generated for each level of the grouping factor. A simple, scalar term generates a block of indicator columns — the indicators for the grouping factor — in $\mathbf{Z}$. Because there is only one random-effects term in this model and because that term is a simple, scalar term, the model matrix $\mathbf{Z}$ for this model is the indicator matrix for the levels of `Batch`.

In the next chapter we fit models with multiple simple, scalar terms and, in subsequent chapters, we extend random-effects terms beyond simple, scalar terms. When we have only simple, scalar terms in the model, each term has a unique grouping factor and the elements of the list returned by `ranef` can be considered as associated with terms or with grouping factors. In more complex models a particular grouping factor may occur in more than one term, in which case the elements of the list are associated with the grouping factors, not the terms.

Given the data, $\mathbf{y}$, and the parameter estimates, we can evaluate a measure of the dispersion of $(\mathcal{B}|\mathcal{Y} = \mathbf{y})$. In the case of a linear mixed model, this is the conditional standard deviation, from which we can obtain a prediction

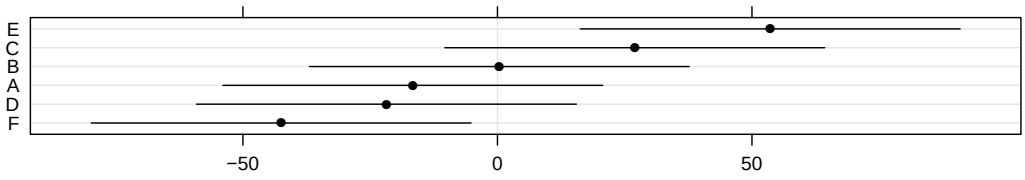
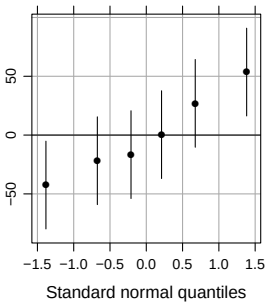


Fig. 1.10 95% prediction intervals on the random effects in `fm1ML`, shown as a dotplot.

Fig. 1.11 95% prediction intervals on the random effects in `fm1ML` versus quantiles of the standard normal distribution.



interval. The `ranef` extractor takes an optional argument, `postVar = TRUE`, which adds these dispersion measures as an attribute of the result. (The name stands for "posterior variance", which is a misnomer that had become established as an argument name before I realized that it wasn't the correct term.)

We can plot these prediction intervals using

```
> dotplot(ranef(fm1ML, postVar = TRUE))
```

(Fig. 1.10), which provides linear spacing of the levels on the y axis, or using

```
> qqmath(ranef(fm1ML, postVar=TRUE))
```

(Fig. 1.11), where the intervals are plotted versus quantiles of the standard normal.

The dotplot is preferred when there are only a few levels of the grouping factor, as in this case. When there are hundreds or thousands of random effects the `qqmath` form is preferred because it focuses attention on the "important few" at the extremes and de-emphasizes the "trivial many" that are close to zero.

1.7 Chapter Summary

A considerable amount of material has been presented in this chapter, especially considering the word "simple" in its title (it's the model that is simple, not the material). A summary may be in order.

A mixed-effects model incorporates fixed-effects parameters and random effects, which are unobserved random variables, $\mathscr{B}$. In a linear mixed model, both the unconditional distribution of $\mathscr{B}$ and the conditional distribution, $(\mathscr{Y}|\mathscr{B}=\mathbf{b})$, are multivariate Gaussian distributions. Furthermore, this conditional distribution is a spherical Gaussian with mean, μ, determined by the linear predictor, $\mathbf{Zb}+\mathbf{X}\beta$. That is,

$$(\mathscr{Y}|\mathscr{B}=\mathbf{b}) \sim \mathscr{N}(\mathbf{Zb}+\mathbf{X}\beta, \sigma^2\mathbf{I}_n).$$

The unconditional distribution of $\mathscr{B}$ has mean $\mathbf{0}$ and a parameterized $q \times q$ variance-covariance matrix, Σ_θ.

In the models we considered in this chapter, Σ_θ, is a simple multiple of the identity matrix, $\mathbf{I}_6$. This matrix is always a multiple of the identity in models with just one random-effects term that is a simple, scalar term. The reason for introducing all the machinery that we did is to allow for more general model specifications.

The maximum likelihood estimates of the parameters are obtained by minimizing the deviance. For linear mixed models we can minimize the profiled deviance, which is a function of θ only, thereby considerably simplifying the optimization problem.

To assess the precision of the parameter estimates, we profile the deviance function with respect to each parameter and apply a signed square root transformation to the likelihood ratio test statistic, producing a profile zeta function for each parameter. These functions provide likelihood-based confidence intervals for the parameters. Profile zeta plots allow us to visually assess the precision of individual parameters. Profile pairs plots allow us to visualize the pairwise dependence of parameter estimates and two-dimensional marginal confidence regions.

Prediction intervals from the conditional distribution of the random effects, given the observed data, allow us to assess the precision of the random effects.

Exercises

These exercises and several others in this book use data sets from the `MEMSS` package for R. You will need to ensure that this package is installed before you can access the data sets.

To load a particular data set, either attach the package

```
> library(MEMSS)
```

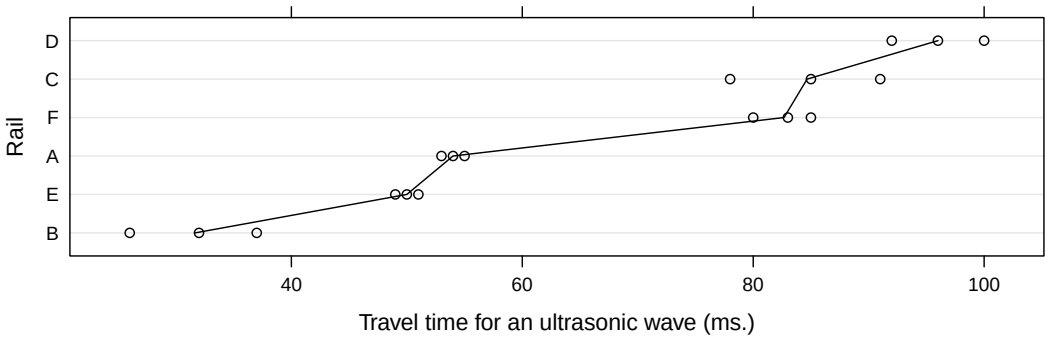


Fig. 1.12 Travel time for an ultrasonic wave test on 6 rails

or load just the one data set

```
> data(Rail, package = "MEMSS")
```

1.1. Check the documentation, the structure (`str`) and a summary of the `Rail` data (Fig. 1.12) from the `MEMSS` package. Note that if you used `data` to access this data set then you must use

```
> help(Rail, package = "MEMSS")
```

to display the documentation for it.

1.2. Fit a model with `travel` as the response and a simple, scalar random-effects term for the variable `Rail`. Use the REML criterion, which is the default. Create a dotplot of the conditional modes of the random effects.

1.3. Refit the model using maximum likelihood. Check the parameter estimates and, in the case of the fixed-effects parameter, its standard error. In what ways have the parameter estimates changed? Which parameter estimates have not changed?

1.4. Profile the fitted model and construct 95% profile-based confidence intervals on the parameters. Is the confidence interval on σ_1 close to being symmetric about the estimate? Is the corresponding interval on $\log(\sigma_1)$ close to being symmetric about its estimate?

1.5. Create the profile zeta plot for this model. For which parameters are there good normal approximations?

1.6. Create a profile pairs plot for this model. Does the shape of the deviance contours in this model mirror those in Fig. 1.9?

1.7. Plot the prediction intervals on the random effects from this model. Do any of these prediction intervals contain zero? Consider the relative magnitudes of $\widehat{\sigma_1}$ and $\widehat{\sigma}$ in this model compared to those in model `fm1` for the `Dyestuff` data. Should these ratios of σ_1/σ lead you to expect a different pattern of prediction intervals in this plot than those in Fig. 1.10?

Chapter 2
Models With Multiple Random-effects Terms

The mixed models considered in the previous chapter had only one random-effects term, which was a simple, scalar random-effects term, and a single fixed-effects coefficient. Although such models can be useful, it is with the facility to use multiple random-effects terms and to use random-effects terms beyond a simple, scalar term that we can begin to realize the flexibility and versatility of mixed models.

In this chapter we consider models with multiple simple, scalar random-effects terms, showing examples where the grouping factors for these terms are in completely crossed or nested or partially crossed configurations. For ease of description we will refer to the random effects as being crossed or nested although, strictly speaking, the distinction between nested and non-nested refers to the grouping factors, not the random effects.

2.1 A Model With Crossed Random Effects

One of the areas in which the methods in the `lme4` package for R are particularly effective is in fitting models to cross-classified data where several factors have random effects associated with them. For example, in many experiments in psychology the reaction of each of a group of subjects to each of a group of stimuli or items is measured. If the subjects are considered to be a sample from a population of subjects and the items are a sample from a population of items, then it would make sense to associate random effects with both these factors.

In the past it was difficult to fit mixed models with multiple, crossed grouping factors to large, possibly unbalanced, data sets. The methods in the `lme4` package are able to do this. To introduce the methods let us first consider a small, balanced data set with crossed grouping factors.

2.1.1 The `Penicillin` Data

The `Penicillin` data are derived from Table 6.6, p. 144 of Davies and Goldsmith [1972] where they are described as coming from an investigation to

> assess the variability between samples of penicillin by the *B. subtilis* method. In this test method a bulk-innoculated nutrient agar medium is poured into a Petri dish of approximately 90 mm. diameter, known as a plate. When the medium has set, six small hollow cylinders or pots (about 4 mm. in diameter) are cemented onto the surface at equally spaced intervals. A few drops of the penicillin solutions to be compared are placed in the respective cylinders, and the whole plate is placed in an incubator for a given time. Penicillin diffuses from the pots into the agar, and this produces a clear circular zone of inhibition of growth of the organisms, which can be readily measured. The diameter of the zone is related in a known way to the concentration of penicillin in the solution.

As with the `Dyestuff` data, we examine the structure

```
> str(Penicillin)

'data.frame':        144 obs. of  3 variables:
 $ diameter: num  27 23 26 23 23 21 27 23 26 23 ...
 $ plate   : Factor w/ 24 levels "a","b","c","d",..: 1 1 1 1 1 1 2 2 2 2..
 $ sample  : Factor w/ 6 levels "A","B","C","D",..: 1 2 3 4 5 6 1 2 3 4 ..
```

and a summary

```
> summary(Penicillin)

    diameter         plate     sample
 Min.   :18.00   a      :  6   A:24
 1st Qu.:22.00   b      :  6   B:24
 Median :23.00   c      :  6   C:24
 Mean   :22.97   d      :  6   D:24
 3rd Qu.:24.00   e      :  6   E:24
 Max.   :27.00   f      :  6   F:24
                 (Other):108
```

of the `Penicillin` data, then plot it (Fig. 2.1).

The variation in the diameter is associated with the plates and with the samples. Because each plate is used only for the six samples shown here we are not interested in the contributions of specific plates as much as we are interested in the variation due to plates and in assessing the potency of the samples after accounting for this variation. Thus, we will use random effects for the `plate` factor. We will also use random effects for the `sample` factor because, as in the dyestuff example, we are more interested in the sample-to-sample variability in the penicillin samples than in the potency of a particular sample.

In this experiment each sample is used on each plate. We say that the `sample` and `plate` factors are *crossed*, as opposed to *nested* factors, which we will describe in the next section. By itself, the designation "crossed" just

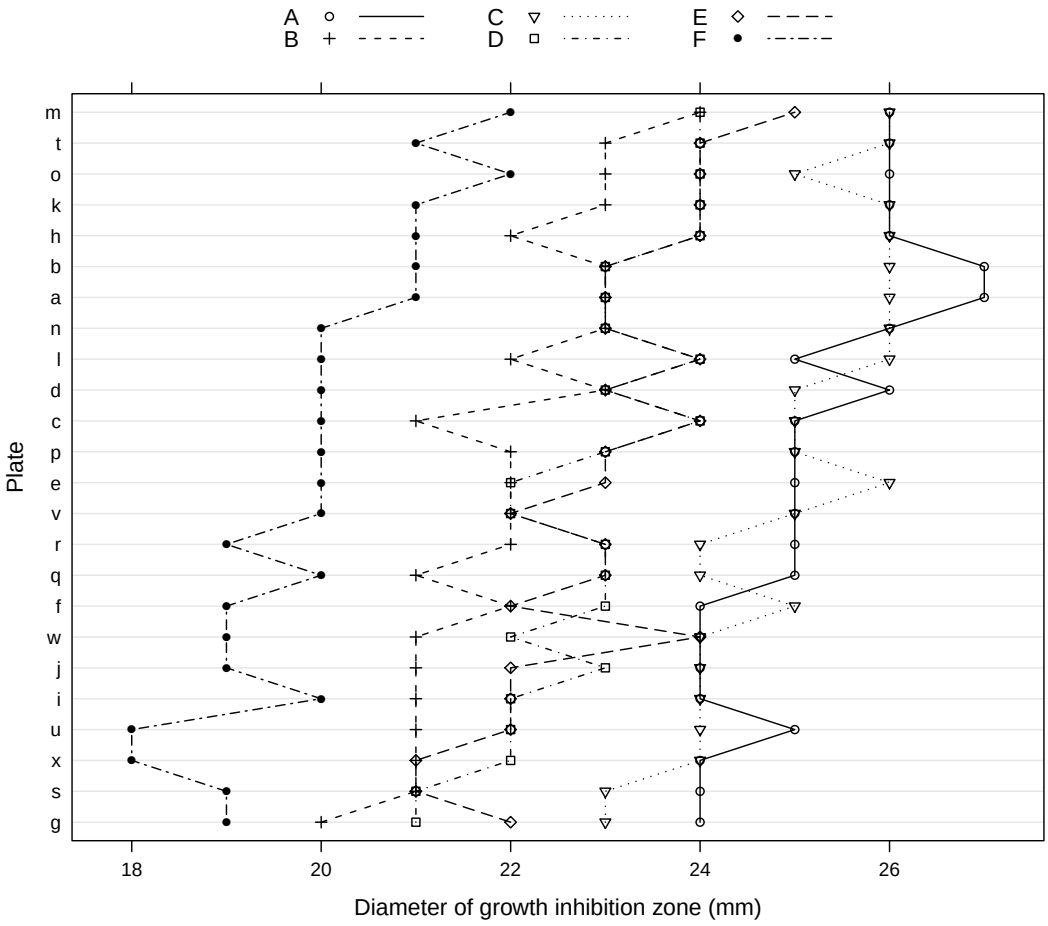


Fig. 2.1 Diameter of the growth inhibition zone (mm) in the *B. subtilis* method of assessing the concentration of penicillin. Each of 6 samples was applied to each of the 24 agar plates. The lines join observations on the same sample.

means that the factors are not nested. If we wish to be more specific, we could describe these factors as being *completely crossed*, which means that we have at least one observation for each combination of a level of `sample` and a level of `plate`. We can see this in Fig. 2.1 and, because there are moderate numbers of levels in these factors, we can check it in a cross-tabulation

```
> xtabs(~ sample + plate, Penicillin)
```

```
      plate
sample a b c d e f g h i j k l m n o p q r s t u v w x
     A 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
     B 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
     C 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
     D 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
     E 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
     F 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Like the `Dyestuff` data, the factors in the `Penicillin` data are balanced. That is, there are exactly the same number of observations on each plate and

for each sample and, furthermore, there is the same number of observations on each combination of levels. In this case there is exactly one observation for each combination of sample and plate. We would describe the configuration of these two factors as an unreplicated, completely balanced, crossed design.

In general, balance is a desirable but precarious property of a data set. We may be able to impose balance in a designed experiment but we typically cannot expect that data from an observation study will be balanced. Also, as anyone who analyzes real data soon finds out, expecting that balance in the design of an experiment will produce a balanced data set is contrary to "Murphy's Law". That's why statisticians allow for missing data. Even when we apply each of the six samples to each of the 24 plates, something could go wrong for one of the samples on one of the plates, leaving us without a measurement for that combination of levels and thus an unbalanced data set.

2.1.2 A Model For the `Penicillin` Data

A model incorporating random effects for both the `plate` and the `sample` is straightforward to specify — we include simple, scalar random effects terms for both these factors.

```
> (fm2 <- lmer(diameter ~ 1 + (1|plate) + (1|sample), Penicillin))

Linear mixed model fit by REML
Formula: diameter ~ 1 + (1 | plate) + (1 | sample)
   Data: Penicillin
 REML
330.9
Random effects:
 Groups   Name        Variance Std.Dev.
 plate    (Intercept) 0.71691  0.84671
 sample   (Intercept) 3.73097  1.93157
 Residual             0.30241  0.54992
Number of obs: 144, groups: plate, 24; sample, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  22.9722     0.8086   28.41
```

This model display indicates that the sample-to-sample variability has the greatest contribution, then plate-to-plate variability and finally the "residual" variability that cannot be attributed to either the sample or the plate. These conclusions are consistent with what we see in the `Penicillin` data plot (Fig. 2.1).

The prediction intervals on the random effects (Fig. 2.2) confirm that the conditional distribution of the random effects for `plate` has much less variability than does the conditional distribution of the random effects for `plate`, in the sense that the dots in the bottom panel have less variability than

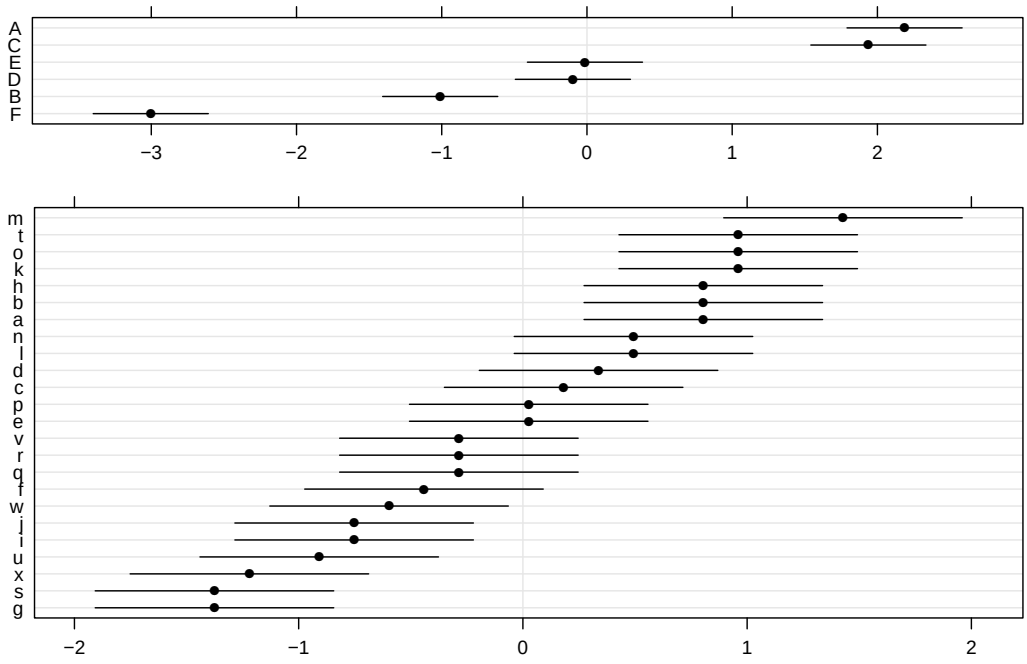


Fig. 2.2 95% prediction intervals on the random effects for model `fm2` fit to the `Penicillin` data.

those in the top panel. (Note the different horizontal axes for the two panels.) However, the conditional distribution of the random effect for a particular `sample`, say sample F, has less variability than the conditional distribution of the random effect for a particular plate, say plate m. That is, the lines in the bottom panel are wider than the lines in the top panel, even after taking the different axis scales into account. This is because the conditional distribution of the random effect for a particular sample depends on 24 responses while the conditional distribution of the random effect for a particular plate depends on only 6 responses.

In chapter 1 we saw that a model with a single, simple, scalar random-effects term generated a random-effects model matrix, $\mathbf{Z}$, that is the matrix of indicators of the levels of the grouping factor. When we have multiple, simple, scalar random-effects terms, as in model `fm2`, each term generates a matrix of indicator columns and these sets of indicators are concatenated to form the model matrix $\mathbf{Z}$. The transpose of this matrix, shown in Fig. 2.3, contains rows of indicators for each factor.

The relative covariance factor, Λ_θ, (Fig. 2.4, left panel) is no longer a multiple of the identity. It is now block diagonal, with two blocks, one of size 24 and one of size 6, each of which is a multiple of the identity. The diagonal elements of the two blocks are θ_1 and θ_2, respectively. The numeric values of these parameters can be obtained as

```
> env(fm2)$theta
```

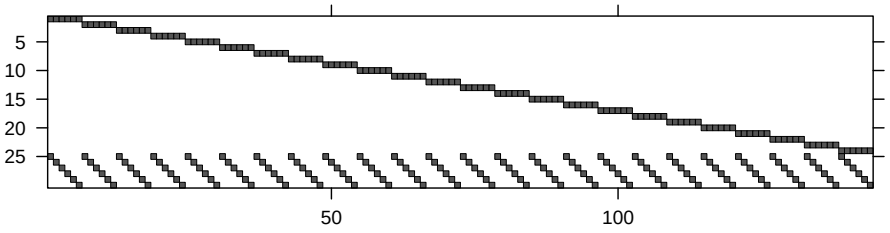


Fig. 2.3 Image of the transpose of the random-effects model matrix, $\mathbf{Z}$, for model `fm2`. The non-zero elements, which are all unity, are shown as darkened squares. The zero elements are blank.

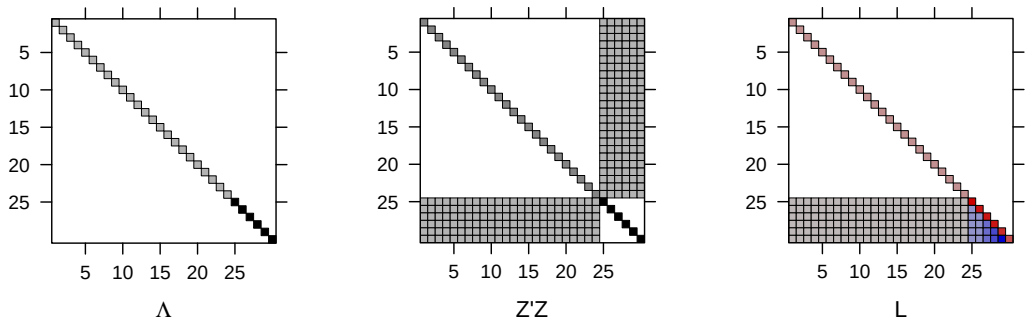


Fig. 2.4 Images of the relative covariance factor, Λ, the cross-product of the random-effects model matrix, $\mathbf{Z}^\top\mathbf{Z}$, and the sparse Cholesky factor, $\mathbf{L}$, for model `fm2`.

```
[1] 1.539683 3.512443
```

The first parameter is the relative standard deviation of the random effects for `plate`, which has the value $0.84671/0.54992 = 1.53968$ at convergence, and the second is the relative standard deviation of the random effects for `sample` ($1.93157/0.54992 = 3.512443$).

Because Λ_θ is diagonal, the pattern of non-zeros in $\Lambda_\theta^\top\mathbf{Z}^\top\mathbf{Z}\Lambda_\theta + \mathbf{I}$ will be the same as that in $\mathbf{Z}^\top\mathbf{Z}$, shown in the middle panel of Fig. 2.4. The sparse Cholesky factor, $\mathbf{L}$, shown in the right panel, is lower triangular and has non-zero elements in the lower right hand corner in positions where $\mathbf{Z}^\top\mathbf{Z}$ has systematic zeros. We say that "fill-in" has occurred when forming the sparse Cholesky decomposition. In this case there is a relatively minor amount of fill but in other cases there can be a substantial amount of fill and we shall take precautions so as to reduce this, because fill-in adds to the computational effort in determining the MLEs or the REML estimates.

A profile zeta plot (Fig. 2.5) for the parameters in model `fm2` leads to conclusions similar to those from Fig. 1.5 for model `fm1ML` in the previous chapter. The fixed-effect parameter, β_0, for the `(Intercept)` term has symmetric intervals and is over-dispersed relative to the normal distribution. The logarithm

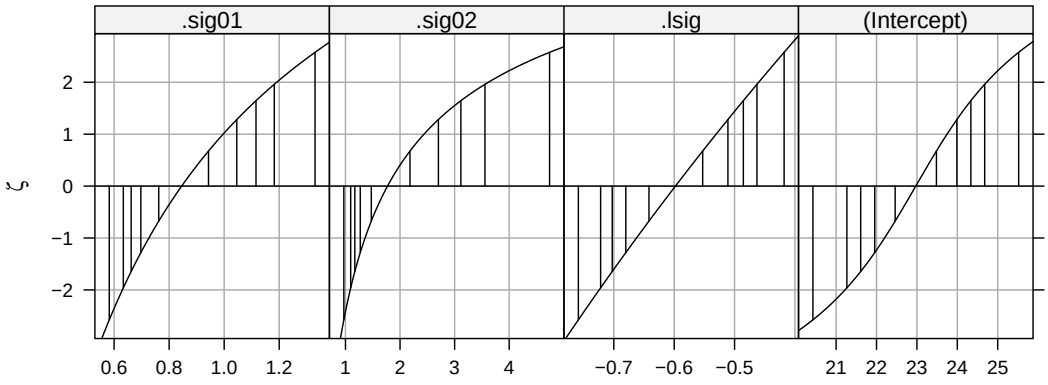


Fig. 2.5 Profile zeta plot of the parameters in model `fm2`.

of σ has a good normal approximation but the standard deviations of the random effects, σ_1 and σ_2, are skewed. The skewness for σ_2 is worse than that for σ_1, because the estimate of σ_2 is less precise than that of σ_1, in both absolute and relative senses. For an absolute comparison we compare the widths of the confidence intervals for these parameters.

```
> confint(pr2)
```

```
                2.5 %     97.5 %
.sig01       0.6335658  1.1821040
.sig02       1.0957822  3.5563194
.lsig       -0.7218645 -0.4629033
(Intercept) 21.2666274 24.6778176
```

In a relative comparison we examine the ratio of the endpoints of the interval divided by the estimate.

```
> confint(pr2)[1:2,]/c(0.8455722, 1.770648)
```

```
           2.5 %   97.5 %
.sig01 0.7492746 1.397993
.sig02 0.6188594 2.008485
```

The lack of precision in the estimate of σ_2 is a consequence of only having 6 distinct levels of the `sample` factor. The `plate` factor, on the other hand, has 24 distinct levels. In general it is more difficult to estimate a measure of spread, such as the standard deviation, than to estimate a measure of location, such as a mean, especially when the number of levels of the factor is small. Six levels are about the minimum number required for obtaining sensible estimates of standard deviations for simple, scalar random effects terms.

The profile pairs plot (Fig. 2.6) shows patterns similar to those in Fig. 1.9 for pairs of parameters in model `fm1` fit to the `Dyestuff` data. On the ζ scale

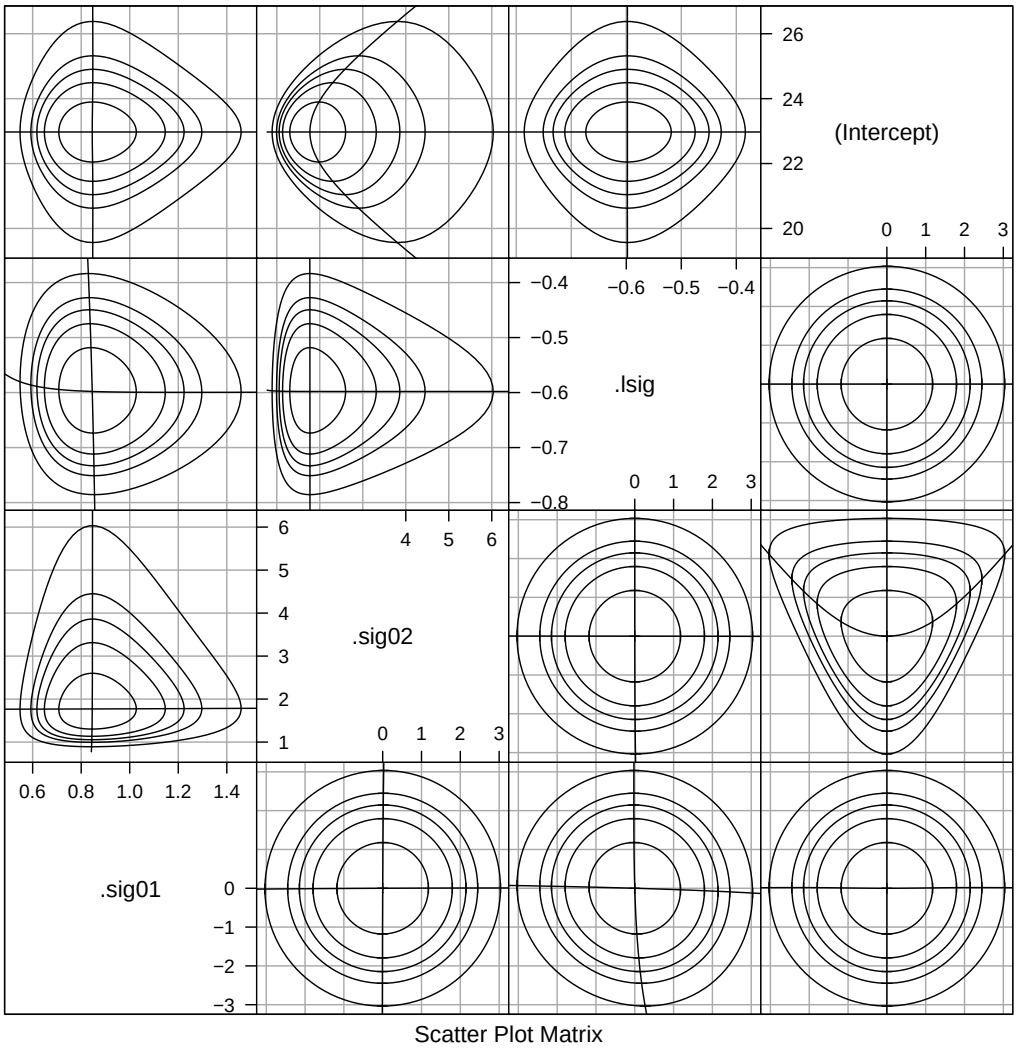


Fig. 2.6 Profile pairs plot for the parameters in model `fm2` fit to the `Penicillin` data.

(panels below the diagonal) the profile traces are nearly straight and orthogonal with the exception of the trace of $\zeta(\sigma_2)$ on $\zeta(\beta_0)$ (the horizontal trace for the panel in the $(4,2)$ position). The pattern of this trace is similar to the pattern of the trace of $\zeta(\sigma_1)$ on $\zeta(\beta_0)$ in Fig. 1.9. Moving β_0 from its estimate, $\widehat{\beta}_0$, in either direction will increase the residual sum of squares. The increase in the residual variability is reflected in an increase of one or more of the dispersion parameters. The balanced experimental design results in a fixed estimate of σ and the extra apparent variability must be incorporated into σ_1 or σ_2.

Contours in panels of parameter pairs on the original scales (i.e. panels above the diagonal) can show considerable distortion from the ideal elliptical shape. For example, contours in the σ_2 versus σ_1 panel (the $(1,2)$ position) and the $\log(\sigma)$ versus σ_2 panel (in the $(2,3)$ position) are dramatically non-

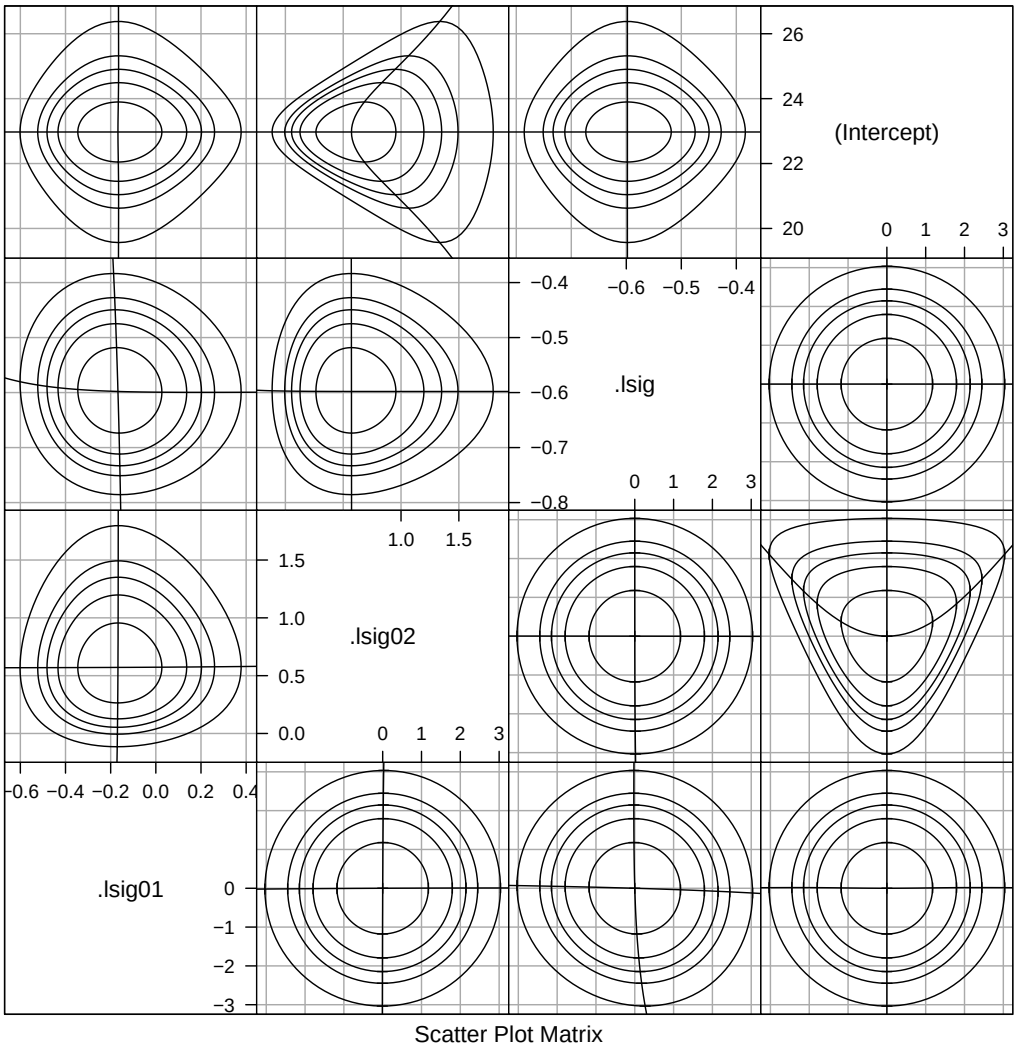


Fig. 2.7 Profile pairs plot for the parameters in model `fm2` fit to the `Penicillin` data. In this plot the parameters σ_1 and σ_2 are on the scale of the natural logarithm, as is the parameter σ in this and other profile pairs plots.

elliptical. However, the distortion of the contours is not due to these parameter estimates depending strongly on each other. It is almost entirely due to the choice of scale for σ_1 and σ_2. When we plot the contours on the scale of $\log(\sigma_1)$ and $\log(\sigma_2)$ instead (Fig. 2.7) they are much closer to the elliptical pattern.

Conversely, if we tried to plot contours on the scale of σ_1^2 and σ_2^2 (not shown), they would be hideously distorted.

2.2 A Model With Nested Random Effects

In this section we again consider a simple example, this time fitting a model with *nested* grouping factors for the random effects.

2.2.1 The Pastes *Data*

The third example from Davies and Goldsmith [1972, Table 6.5, p. 138] is described as coming from

> deliveries of a chemical paste product contained in casks where, in addition to sampling and testing errors, there are variations in quality between deliveries ... As a routine, three casks selected at random from each delivery were sampled and the samples were kept for reference. ... Ten of the delivery batches were sampled at random and two analytical tests carried out on each of the 30 samples.

The structure and summary of the `Pastes` data object are

```
> str(Pastes)
```

```
'data.frame':        60 obs. of  4 variables:
 $ strength: num  62.8 62.6 60.1 62.3 62.7 63.1 60 61.4 57.5 56.9 ...
 $ batch   : Factor w/ 10 levels "A","B","C","D",..: 1 1 1 1 1 1 2 2 2 2..
 $ cask    : Factor w/ 3 levels "a","b","c": 1 1 2 2 3 3 1 1 2 2 ...
 $ sample  : Factor w/ 30 levels "A:a","A:b","A:c",..: 1 1 2 2 3 3 4 4 5..
```

```
> summary(Pastes)
```

```
    strength         batch    cask       sample
 Min.   :54.20   A      : 6   a:20   A:a    : 2
 1st Qu.:57.50   B      : 6   b:20   A:b    : 2
 Median :59.30   C      : 6   c:20   A:c    : 2
 Mean   :60.05   D      : 6          B:a    : 2
 3rd Qu.:62.88   E      : 6          B:b    : 2
 Max.   :66.00   F      : 6          B:c    : 2
                 (Other):24          (Other):48
```

As stated in the description in Davies and Goldsmith [1972], there are 30 samples, three from each of the 10 delivery batches. We have labelled the levels of the `sample` factor with the label of the `batch` factor followed by 'a', 'b' or 'c' to distinguish the three samples taken from that batch. The cross-tabulation produced by the `xtabs` function, using the optional argument `sparse = TRUE`, provides a concise display of the relationship.

```
> xtabs(~ batch + sample, Pastes, drop = TRUE, sparse = TRUE)
```

```
10 x 30 sparse Matrix of class "dgCMatrix"
A 2 2 2 . . . . . . . . . . . . . . . . . . . . . . . . . . .
B . . . 2 2 2 . . . . . . . . . . . . . . . . . . . . . . . .
C . . . . . . 2 2 2 . . . . . . . . . . . . . . . . . . . . .
```

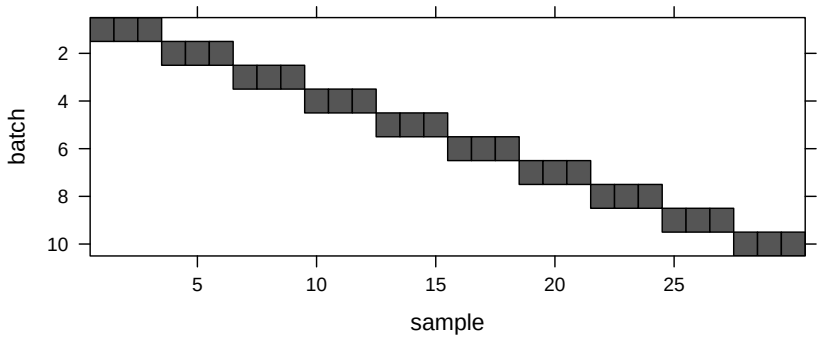


Fig. 2.8 Image of the cross-tabulation of the `batch` and `sample` factors in the `Pastes` data.

```
D . . . . . 2 2 2 . . . . . .
E . . . . . 2 2 2 . . . . . .
F . . . . . 2 2 2 . . . . . .
G . . . . . 2 2 2 . . . . . .
H . . . . . 2 2 2 . . . . . .
I . . . . . 2 2 2 . . . . . .
J . . . . . 2 2 2
```

Alternatively, we can use an image (Fig. 2.8) of this cross-tabulation to visualize the structure.

When plotting the `strength` versus `batch` and `sample` in the `Pastes` data we should remember that we have two strength measurements on each of the 30 samples. It is tempting to use the cask designation ('a', 'b' and 'c') to determine, say, the plotting symbol within a `batch`. It would be fine to do this within a batch but the plot would be misleading if we used the same symbol for cask 'a' in different batches. There is no relationship between cask 'a' in batch 'A' and cask 'a' in batch 'B'. The labels 'a', 'b' and 'c' are used only to distinguish the three samples within a batch; they do not have a meaning across batches.

In Fig. 2.9 we plot the two strength measurements on each of the samples within each of the batches and join up the average strength for each sample. The perceptive reader will have noticed that the levels of the factors on the vertical axis in this figure, and in Fig. 1.1 and 2.1, have been reordered according to increasing average response. In all these cases there is no inherent ordering of the levels of the covariate such as `batch` or `plate`. Rather than confuse our interpretation of the plot by determining the vertical displacement of points according to a random ordering, we impose an ordering according to increasing mean response. This allows us to more easily check for structure in the data, including undesirable characteristics like increasing variability of the response with increasing mean level of the response.

In Fig. 2.9 we order the samples within each batch separately then order the batches according to increasing mean strength.

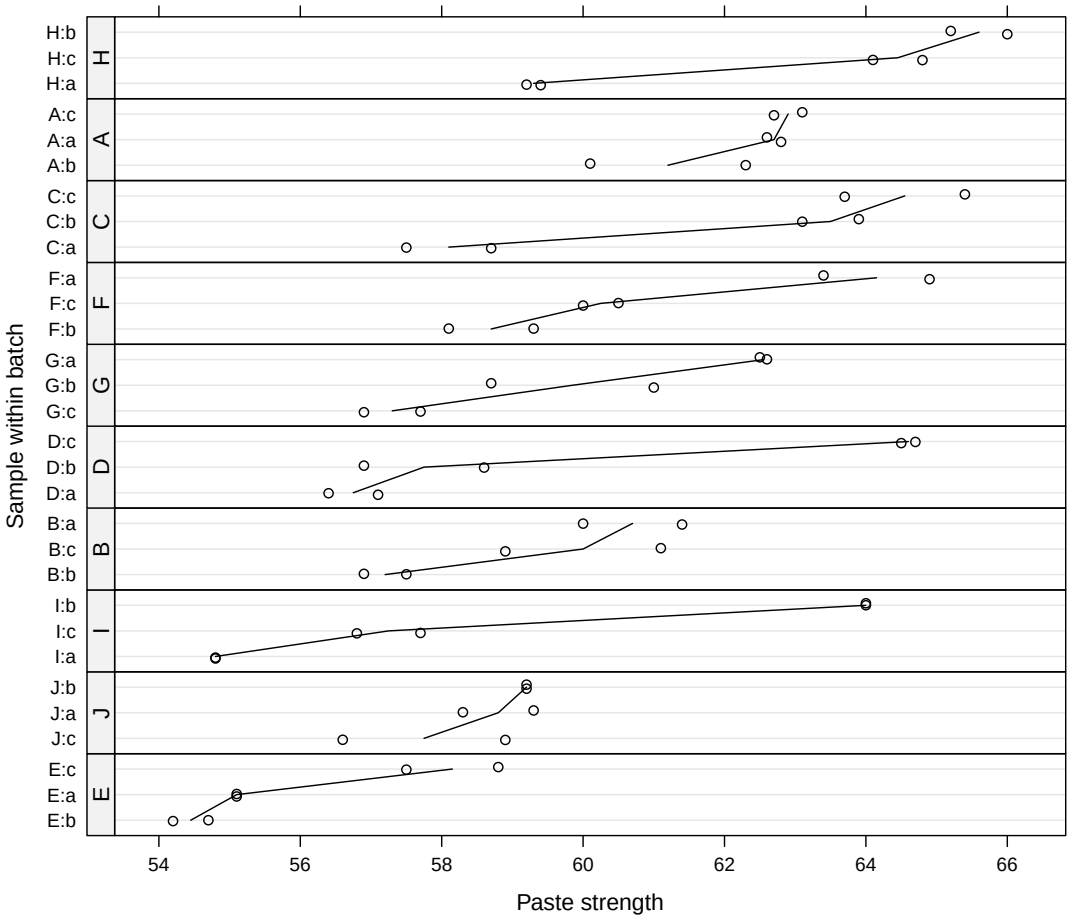


Fig. 2.9 Strength of paste preparations according to the `batch` and the `sample` within the batch. There were two strength measurements on each of the 30 samples; three samples each from 10 batches.

Figure 2.9 shows considerable variability in strength between samples relative to the variability within samples. There is some indication of variability between batches, in addition to the variability induced by the samples, but not a strong indication of a batch effect. For example, batches I and D, with low mean strength relative to the other batches, each contained one sample (I:b and D:c, respectively) that had high mean strength relative to the other samples. Also, batches H and C, with comparatively high mean batch strength, contain samples H:a and C:a with comparatively low mean sample strength. In Sect. 2.2.4 we will examine the need for incorporating batch-to-batch variability, in addition to sample-to-sample variability, in the statistical model.

2.2.1.1 Nested Factors

Because each level of `sample` occurs with one and only one level of `batch` we say that `sample` is *nested within* `batch`. Some presentations of mixed-effects models, especially those related to *multilevel modeling* [Rasbash et al., 2000] or *hierarchical linear models* [Raudenbush and Bryk, 2002], leave the impression that one can only define random effects with respect to factors that are nested. This is the origin of the terms "multilevel", referring to multiple, nested levels of variability, and "hierarchical", also invoking the concept of a hierarchy of levels. To be fair, both those references do describe the use of models with random effects associated with non-nested factors, but such models tend to be treated as a special case.

The blurring of mixed-effects models with the concept of multiple, hierarchical levels of variation results in an unwarranted emphasis on "levels" when defining a model and leads to considerable confusion. It is perfectly legitimate to define models having random effects associated with non-nested factors. The reasons for the emphasis on defining random effects with respect to nested factors only are that such cases do occur frequently in practice and that some of the computational methods for estimating the parameters in the models can only be easily applied to nested factors.

This is not the case for the methods used in the `lme4` package. Indeed there is nothing special done for models with random effects for nested factors. When random effects are associated with multiple factors exactly the same computational methods are used whether the factors form a nested sequence or are partially crossed or are completely crossed. A case of a nested sequence of "grouping factors" for the random effects (including the trivial case of only one such factor) is detected but this information does not change the course of the computation. It is available to be used as a diagnostic check. When the user knows that the grouping factors should be nested, she can check if they are indeed nested.

There is, however, one aspect of nested grouping factors that we should emphasize, which is the possibility of a factor that is *implicitly nested* within another factor. Suppose, for example, that the `sample` factor was defined as having three levels instead of 30 with the implicit assumption that `sample` is nested within `batch`. It may seem silly to try to distinguish 30 different batches with only three levels of a factor but, unfortunately, data are frequently organized and presented like this, especially in text books. The `cask` factor in the `Pastes` data is exactly such an implicitly nested factor. If we cross-tabulate `batch` and `cask`

```
> xtabs(~ cask + batch, Pastes)
```

```
    batch
cask A B C D E F G H I J
   a 2 2 2 2 2 2 2 2 2 2
   b 2 2 2 2 2 2 2 2 2 2
   c 2 2 2 2 2 2 2 2 2 2
```

we get the impression that the `cask` and `batch` factors are crossed, not nested. If we know that the cask should be considered as nested within the batch then we should create a new categorical variable giving the batch-cask combination, which is exactly what the `sample` factor is. A simple way to create such a factor is to use the interaction operator, ':', on the factors. It is advisable, but not necessary, to apply `factor` to the result thereby dropping unused levels of the interaction from the set of all possible levels of the factor. (An "unused level" is a combination that does not occur in the data.) A convenient code idiom is

```
> Pastes$sample <- with(Pastes, factor(batch:cask))
```

or

```
> Pastes <- within(Pastes, sample <- factor(batch:cask))
```

In a small data set like `Pastes` we can quickly detect a factor being implicitly nested within another factor and take appropriate action. In a large data set, perhaps hundreds of thousands of test scores for students in thousands of schools from hundreds of school districts, it is not always obvious if school identifiers are unique across the entire data set or just within a district. If you are not sure, the safest thing to do is to create the interaction factor, as shown above, so you can be confident that levels of the district:school interaction do indeed correspond to unique schools.

2.2.2 Fitting a Model With Nested Random Effects

Fitting a model with simple, scalar random effects for nested factors is done in exactly the same way as fitting a model with random effects for crossed grouping factors. We include random-effects terms for each factor, as in

```
> (fm3 <- lmer(strength ~ 1 + (1|sample) + (1|batch), Pastes, REML=0))

Linear mixed model fit by maximum likelihood
Formula: strength ~ 1 + (1 | sample) + (1 | batch)
   Data: Pastes
 AIC   BIC logLik deviance
 256 264.4   -124      248
Random effects:
 Groups   Name        Variance Std.Dev.
 sample   (Intercept) 8.4337   2.9041
 batch    (Intercept) 1.1992   1.0951
 Residual             0.6780   0.8234
Number of obs: 60, groups: sample, 30; batch, 10

Fixed effects:
            Estimate Std. Error t value
(Intercept)  60.0533     0.6421   93.52
```

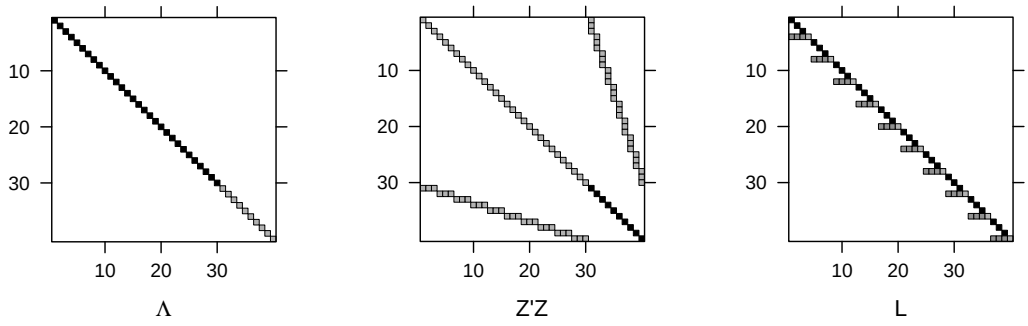


Fig. 2.10 Images of the relative covariance factor, Λ, the cross-product of the random-effects model matrix, $\mathbf{Z}^\mathsf{T}\mathbf{Z}$, and the sparse Cholesky factor, $\mathbf{L}$, for model `fm3`.

Not only is the model specification similar for nested and crossed factors, the internal calculations are performed according to the methods described in Sect. 1.4.1 for each model type. Comparing the patterns in the matrices Λ, $\mathbf{Z}^\mathsf{T}\mathbf{Z}$ and $\mathbf{L}$ for this model (Fig. 2.10) to those in Fig. 2.4 shows that models with nested factors produce simple repeated structures along the diagonal of the sparse Cholesky factor, $\mathbf{L}$, after reordering the random effects (we discuss this reordering later in Sect. 5.4.1). This type of structure has the desirable property that there is no "fill-in" during calculation of the Cholesky factor. In other words, the number of non-zeros in $\mathbf{L}$ is the same as the number of non-zeros in the lower triangle of the matrix being factored, $\Lambda^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda + \mathbf{I}$ (which, because Λ is diagonal, has the same structure as $\mathbf{Z}^\mathsf{T}\mathbf{Z}$).

Fill-in of the Cholesky factor is not an important issue when we have a few dozen random effects, as we do here. It is an important issue when we have millions of random effects in complex configurations, as has been the case in some of the models that have been fit using `lmer`.

2.2.3 Assessing Parameter Estimates in Model `fm3`

The parameter estimates are: $\widehat{\sigma_1}$ =2.904, the standard deviation of the random effects for `sample`; $\widehat{\sigma_2}$ =1.095, the standard deviation of the random effects for `batch`; $\widehat{\sigma}$ =0.823, the standard deviation of the residual noise term; and $\widehat{\beta_0}$ =60.053, the overall mean response, which is labeled `(Intercept)` in these models.

The estimated standard deviation for `sample` is nearly three times as large as that for `batch`, which confirms what we saw in Fig. 2.9. Indeed our con-

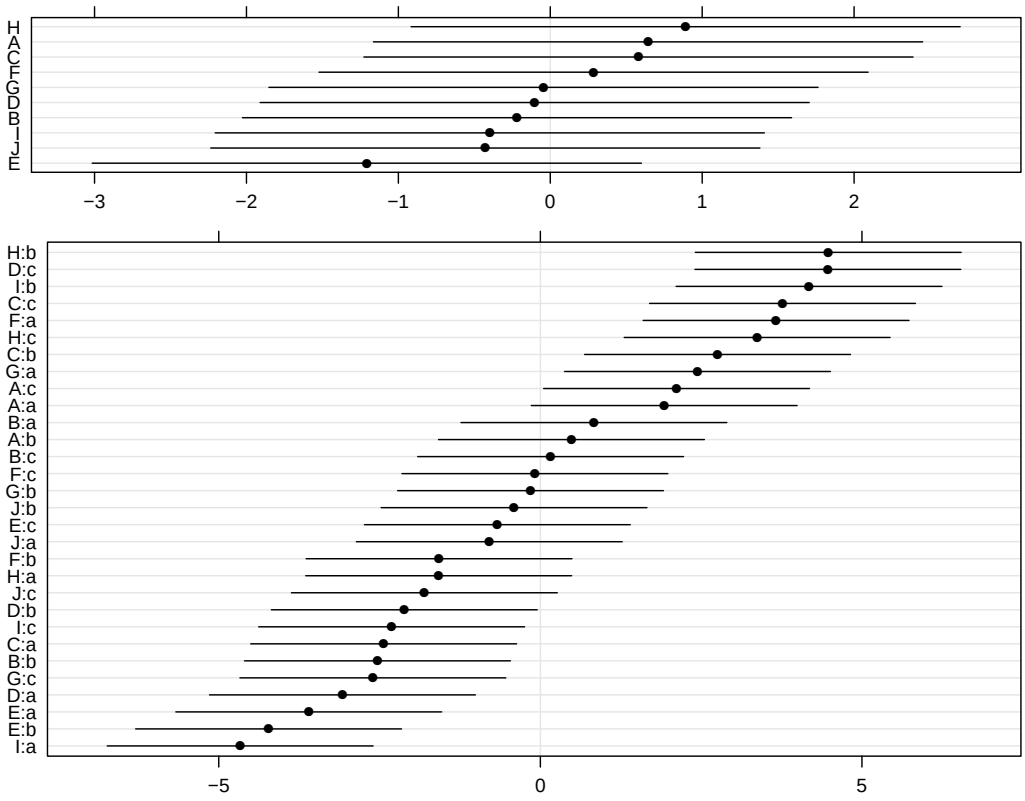


Fig. 2.11 95% prediction intervals on the random effects for model `fm3` fit to the `Pastes` data.

clusion from Fig. 2.9 was that there may not be a significant batch-to-batch variability in addition to the sample-to-sample variability.

Plots of the prediction intervals of the random effects (Fig. 2.11) confirm this impression in that all the prediction intervals for the random effects for `batch` contain zero. Furthermore, the profile zeta plot (Fig. 2.12) shows that the even the 50% profile-based confidence interval on σ_2 extends to zero.

Because there are several indications that σ_2 could reasonably be zero, resulting in a simpler model incorporating random effects for `batch` only, we perform a statistical test of this hypothesis.

2.2.4 Testing $H_0 : \sigma_2 = 0$ Versus $H_a : \sigma_2 > 0$

One of the many famous statements attributed to Albert Einstein is "Everything should be made as simple as possible, but not simpler." In statistical modeling this *principal of parsimony* is embodied in hypothesis tests comparing two models, one of which contains the other as a special case. Typically,

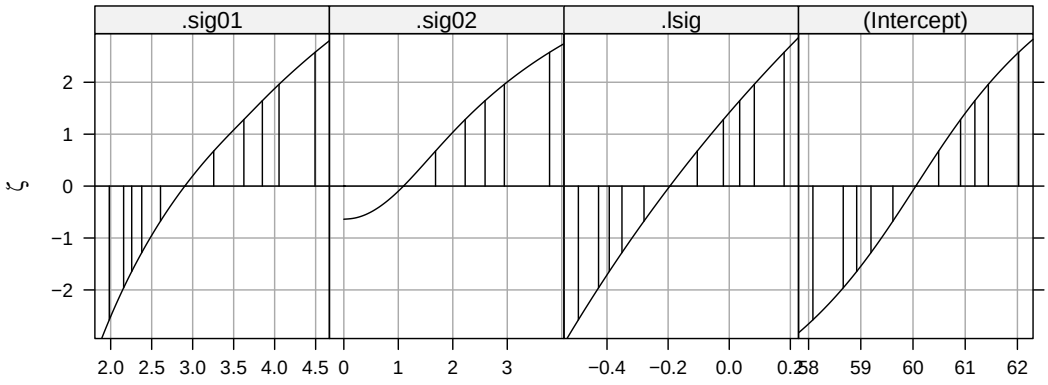


Fig. 2.12 Profile zeta plots for the parameters in model `fm3`.

one or more of the parameters in the more general model, which we call the *alternative hypothesis*, is constrained in some way, resulting in the restricted model, which we call the *null hypothesis*. Although we phrase the hypothesis test in terms of the parameter restriction, it is important to realize that we are comparing the quality of fits obtained with two nested models. That is, we are not assessing parameter values per se; we are comparing the model fit obtainable with some constraints on parameter values to that without the constraints.

Because the more general model, H_a, must provide a fit that is at least as good as the restricted model, H_0, our purpose is to determine whether the change in the quality of the fit is sufficient to justify the greater complexity of model H_a. This comparison is often reduced to a *p-value*, which is the probability of seeing a difference in the model fits as large as we did, or even larger, when, in fact, H_0 is adequate. Like all probabilities, a p-value must be between 0 and 1. When the p-value for a test is small (close to zero) we prefer the more complex model, saying that we "reject H_0 in favor of H_a". On the other hand, when the p-value is not small we "fail to reject H_0", arguing that there is a non-negligible probability that the observed difference in the model fits could reasonably be the result of random chance, not the inherent superiority of the model H_a. Under these circumstances we prefer the simpler model, H_0, according to the principal of parsimony.

These are the general principles of statistical hypothesis tests. To perform a test in practice we must specify the criterion for comparing the model fits, the method for calculating the p-value from an observed value of the criterion, and the standard by which we will determine if the p-value is "small" or not. The criterion is called the *test statistic*, the p-value is calculated from a *reference distribution* for the test statistic, and the standard for small p-values is called the *level* of the test.

In Sect. 1.5 we referred to likelihood ratio tests (LRTs) for which the test statistic is the difference in the deviance. That is, the LRT statistic is $d_0 - d_a$ where d_a is the deviance in the more general (H_a) model fit and d_0 is the deviance in the constrained (H_0) model. An approximate reference distribution for an LRT statistic is the χ^2_ν distribution where ν, the degrees of freedom, is determined by the number of constraints imposed on the parameters of H_a to produce H_0.

The restricted model fit

```
> (fm3a <- lmer(strength ~ 1 + (1|sample), Pastes, REML=0))

Linear mixed model fit by maximum likelihood
Formula: strength ~ 1 + (1 | sample)
   Data: Pastes
   AIC   BIC logLik deviance
 254.4 260.7 -124.2    248.4
Random effects:
 Groups   Name        Variance Std.Dev.
 sample   (Intercept) 9.6328   3.1037
 Residual             0.6780   0.8234
Number of obs: 60, groups: sample, 30

Fixed effects:
            Estimate Std. Error t value
(Intercept)  60.0533     0.5765   104.2
```

is compared to model `fm3` with the `anova` function

```
> anova(fm3a, fm3)

Data: Pastes
Models:
fm3a: strength ~ 1 + (1 | sample)
fm3: strength ~ 1 + (1 | sample) + (1 | batch)
     Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
fm3a  3 254.40 260.69 -124.20
fm3   4 255.99 264.37 -124.00 0.4072      1     0.5234
```

which provides a p-value of 0.5234. Because typical standards for "small" p-values are 5% or 1%, a p-value over 50% would not be considered significant at any reasonable level.

We do need to be cautious in quoting this p-value, however, because the parameter value being tested, $\sigma_2 = 0$, is on the boundary of set of possible values, $\sigma_2 \geq 0$, for this parameter. The argument for using a χ^2_1 distribution to calculate a p-value for the change in the deviance does not apply when the parameter value being tested is on the boundary. As shown in Pinheiro and Bates [2000, Sect. 2.5], the p-value from the χ^2_1 distribution will be "conservative" in the sense that it is larger than a simulation-based p-value would be. In the worst-case scenario the χ^2-based p-value will be twice as large as it should be but, even if that were true, an effective p-value of 26% would not cause us to reject H_0 in favor of H_a.

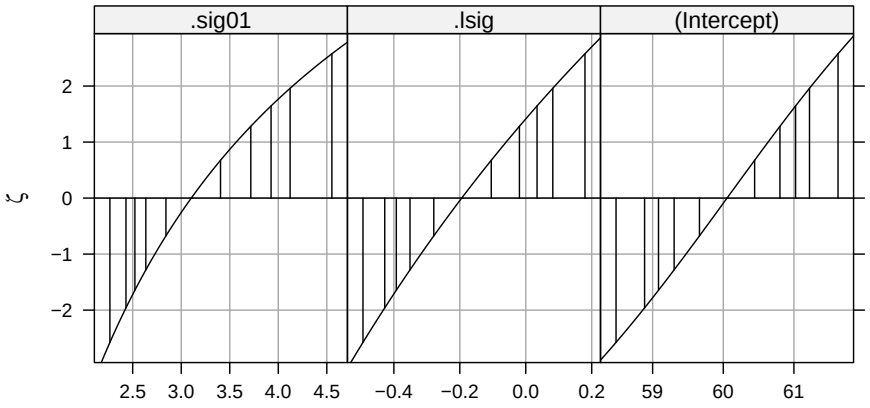


Fig. 2.13 Profile zeta plots for the parameters in model `fm3a`.

2.2.5 Assessing the Reduced Model, `fm3a`

The profile zeta plots for the remaining parameters in model `fm3a` (Fig. 2.13) are similar to the corresponding panels in Fig. 2.12, as confirmed by the numerical values of the confidence intervals.

```
> confint(pr3)
```

```
                 2.5 %      97.5 %
.sig01       2.1579337  4.05358895
.sig02              NA  2.94658928
.lsig       -0.4276761  0.08199287
(Intercept) 58.6636504 61.44301637
```

```
> confint(pr3a)
```

```
                 2.5 %      97.5 %
.sig01       2.4306377  4.12201052
.lsig       -0.4276772  0.08199277
(Intercept) 58.8861831 61.22048353
```

The confidence intervals on $\log(\sigma)$ and β_0 are similar for the two models. The confidence interval on σ_1 is slightly wider in model `fm3a` than in `fm3`, because the variability that is attributed to `batch` in `fm3` is incorporated into the variability due to `sample` in `fm3a`.

The patterns in the profile pairs plot (Fig. 2.14) for the reduced model `fm3a` are similar to those in Fig. 1.9, the profile pairs plot for model `fm1`.

2.3 A Model With Partially Crossed Random Effects

Especially in observational studies with multiple grouping factors, the configuration of the factors frequently ends up neither nested nor completely

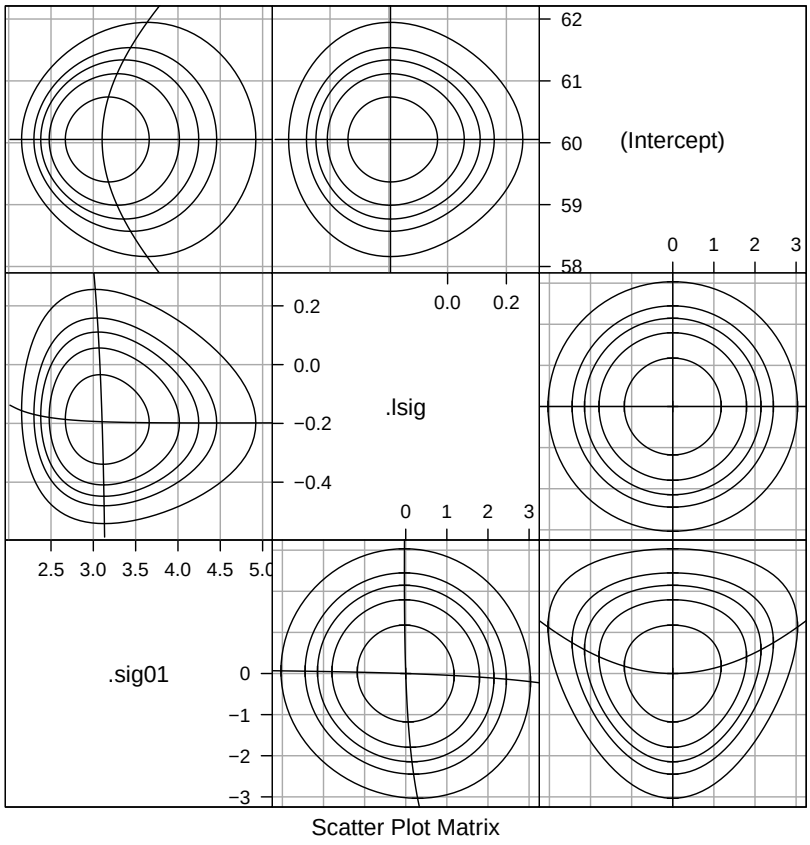


Fig. 2.14 Profile pairs plot for the parameters in model `fm3a` fit to the `Pastes` data.

crossed. We describe such situations as having *partially crossed* grouping factors for the random effects.

Studies in education, in which test scores for students over time are also associated with teachers and schools, usually result in partially crossed grouping factors. If students with scores in multiple years have different teachers for the different years, the student factor cannot be nested within the teacher factor. Conversely, student and teacher factors are not expected to be completely crossed. To have complete crossing of the student and teacher factors it would be necessary for each student to be observed with each teacher, which would be unusual. A longitudinal study of thousands of students with hundreds of different teachers inevitably ends up partially crossed.

In this section we consider an example with thousands of students and instructors where the response is the student's evaluation of the instructor's effectiveness. These data, like those from most large observational studies, are quite unbalanced.

2.3.1 *The* InstEval *Data*

The `InstEval` data are from a special evaluation of lecturers by students at the Swiss Federal Institute for Technology–Zürich (ETH–Zürich), to determine who should receive the "best-liked professor" award. These data have been slightly simplified and identifying labels have been removed, so as to preserve anonymity.

The variables

```
> str(InstEval)
```

```
'data.frame':        73421 obs. of  7 variables:
 $ s      : Factor w/ 2972 levels "1","2","3","4",..: 1 1 1 1 2 2 3 3 3 ..
 $ d      : Factor w/ 1128 levels "1","6","7","8",..: 525 560 832 1068 6..
 $ studage: Ord.factor w/ 4 levels "2"<"4"<"6"<"8": 1 1 1 1 1 1 1 1 1 1 ..
 $ lectage: Ord.factor w/ 6 levels "1"<"2"<"3"<"4"<..: 2 1 2 2 1 1 1 1 1..
 $ service: Factor w/ 2 levels "0","1": 1 2 1 2 1 1 2 1 1 1 ...
 $ dept   : Factor w/ 14 levels "15","5","10",..: 14 5 14 12 2 2 13 3 3 ..
 $ y      : int  5 2 5 3 2 4 4 5 5 4 ...
```

have somewhat cryptic names. Factor `s` designates the student and `d` the instructor. The `dept` factor is the department for the course and `service` indicates whether the course was a service course taught to students from other departments.

Although the response, `y`, is on a scale of 1 to 5,

```
> xtabs(~ y, InstEval)
```

```
y
    1     2     3     4     5
10186 12951 17609 16921 15754
```

it is sufficiently diffuse to warrant treating it as if it were a continuous response.

At this point we will fit models that have random effects for student, instructor, and department (or the `dept:service` combination) to these data. In the next chapter we will fit models incorporating fixed-effects for instructor and department to these data.

```
> (fm4 <- lmer(y ~ 1 + (1|s) + (1|d)+(1|dept:service), InstEval, REML=0))
```

```
Linear mixed model fit by maximum likelihood
Formula: y ~ 1 + (1 | s) + (1 | d) + (1 | dept:service)
   Data: InstEval
    AIC    BIC  logLik deviance
 237663 237709 -118827   237653
Random effects:
 Groups       Name        Variance Std.Dev.
 s            (Intercept) 0.105404 0.32466
 d            (Intercept) 0.262563 0.51241
 dept:service (Intercept) 0.012126 0.11012
 Residual                 1.384953 1.17684
```

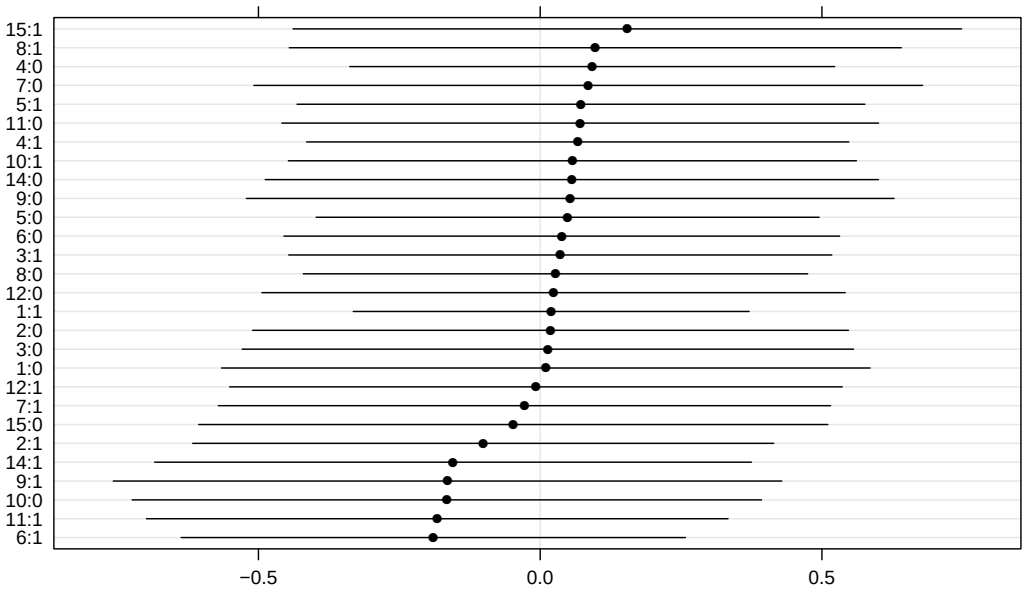


Fig. 2.15 95% prediction intervals on the random effects for the `dept:service` factor in model `fm4` fit to the `InstEval` data.

```
Number of obs: 73421, groups: s, 2972; d, 1128; dept:service, 28

Fixed effects:
            Estimate Std. Error t value
(Intercept)  3.25521    0.02824   115.3
```

(Fitting this complex model to a moderately large data set takes less than two minutes on a modest laptop computer purchased in 2006. Although this is more time than required for earlier model fits, it is a remarkably short time for fitting a model of this size and complexity. In some ways it is remarkable that such a model can be fit at all on such a computer.)

All three estimated standard deviations of the random effects are less than $\widehat{\sigma}$, with $\widehat{\sigma}_3$, the estimated standard deviation of the random effects for the `dept:service` interaction, less than one-tenth the estimated residual standard deviation.

It is not surprising that zero is within all of the prediction intervals on the random effects for this factor (Fig. 2.15). In fact, zero is close to the middle of all these prediction intervals. However, the p-value for the LRT of $H_0 : \sigma_3 = 0$ versus $H_a : \sigma_3 > 0$

```
> fm4a <- lmer(y ~ 1 + (1|s) + (1|d), InstEval, REML=0)
> anova(fm4a,fm4)

Data: InstEval
Models:
fm4a: y ~ 1 + (1 | s) + (1 | d)
fm4: y ~ 1 + (1 | s) + (1 | d) + (1 | dept:service)
```

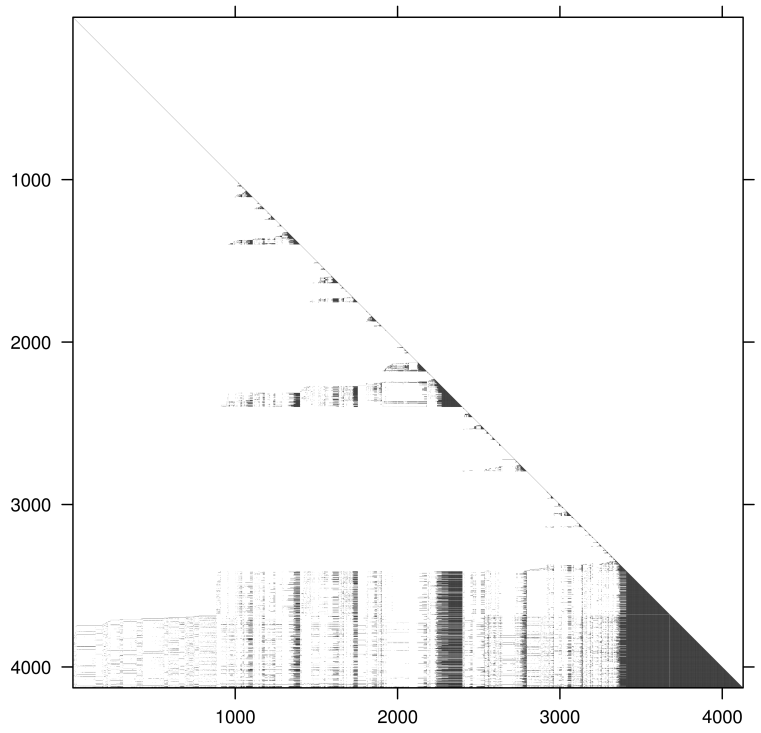


Fig. 2.16 Image of the sparse Cholesky factor, **L**, from model `fm4`

```
     Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
fm4a  4 237786 237823 -118889
fm4   5 237663 237709 -118827 124.43      1  < 2.2e-16
```

is highly significant. That is, we have very strong evidence that we should reject H_0 in favor of H_a.

The seeming inconsistency of these conclusions is due to the large sample size ($n = 73421$). When a model is fit to a very large sample even the most subtle of differences can be highly "statistically significant". The researcher or data analyst must then decide if these terms have practical significance, beyond the apparent statistical significance.

The large sample size also helps to assure that the parameters have good normal approximations. We could profile this model fit but doing so would take a very long time and, in this particular case, the analysts are more interested in a model that uses fixed-effects parameters for the instructors, which we will describe in the next chapter.

We could pursue other mixed-effects models here, such as using the `dept` factor and not the `dept:service` interaction to define random effects, but we will revisit these data in the next chapter and follow up on some of these variations there.

2.3.2 Structure of L for model fm4

Before leaving this model we examine the sparse Cholesky factor, $\mathbf{L}$, (Fig. 2.16), which is of size 4128×4128. Even as a sparse matrix this factor requires a considerable amount of memory,

```
> object.size(env(fm4)$L)

6904640 bytes

> unclass(round(object.size(env(fm4)$L)/2^20, 3))  # size in megabytes

[1] 6.585
```

but as a triangular dense matrix it would require nearly 10 times as much. There are $(4128 \times 4129)/2$ elements on and below the diagonal, each of which would require 8 bytes of storage. A packed lower triangular array would require

```
> (8 * (4128 * 4129)/2)/2^20   # size in megabytes

[1] 65.01965
```

megabytes. The more commonly used full rectangular storage requires

```
> (8 * 4128^2)/2^20   # size in megabytes

[1] 130.0078
```

megabytes of storage.

The number of nonzero elements in this matrix that must be updated for each evaluation of the deviance is

```
> nnzero(as(env(fm4)$L, "sparseMatrix"))

[1] 566960
```

Comparing this to 8522256, the number of elements that must be updated in a dense Cholesky factor, we can see why the sparse Cholesky factor provides a much more efficient evaluation of the profiled deviance function.

2.4 Chapter Summary

A simple, scalar random effects term in an `lmer` model formula is of the form `(1|fac)`, where `fac` is an expression whose value is the *grouping factor* of the set of random effects generated by this term. Typically, `fac` is simply the name of a factor, such as in the terms `(1|sample)` or `(1|plate)` in the examples in this chapter. However, the grouping factor can be the value of an expression, such as `(1|dept:service)` in the last example.

Because simple, scalar random-effects terms can differ only in the description of the grouping factor we refer to configurations such as crossed or nested as applying to the terms or to the random effects, although it is more accurate to refer to the configuration as applying to the grouping factors.

A model formula can include several such random effects terms. Because configurations such as nested or crossed or partially crossed grouping factors are a property of the data, the specification in the model formula does not depend on the configuration. We simply include multiple random effects terms in the formula specifying the model.

One apparent exception to this rule occurs with implicitly nested factors, in which the levels of one factor are only meaningful within a particular level of the other factor. In the `Pastes` data, levels of the `cask` factor are only meaningful within a particular level of the `batch` factor. A model formula of

```
strength ~ 1 + (1 | cask) + (1 | batch)
```

would result in a fitted model that did not appropriately reflect the sources of variability in the data. Following the simple rule that the factor should be defined so that distinct experimental or observational units correspond to distinct levels of the factor will avoid such ambiguity.

For convenience, a model with multiple, nested random-effects terms can be specified as

```
strength ~ 1 + (1 | batch/cask)
```

which internally is re-expressed as

```
strength ~ 1 + (1 | batch) + (1 | batch:cask)
```

We will avoid terms of the form `(1|batch/cask)`, preferring instead an explicit specification with simple, scalar terms based on unambiguous grouping factors.

The `InstEval` data, described in Sec. 2.3.1, illustrate some of the characteristics of the real data to which mixed-effects models are now fit. There is a large number of observations associated with several grouping factors; two of which, student and instructor, have a large number of levels and are partially crossed. Such data are common in sociological and educational studies but until now it has been very difficult to fit models that appropriately reflect such a structure. Much of the literature on mixed-effects models leaves the impression that multiple random effects terms can only be associated with nested grouping factors. The resulting emphasis on hierarchical or multilevel configurations is an artifact of the computational methods used to fit the models, not the models themselves.

The parameters of the models fit to small data sets have properties similar to those for the models in the previous chapter. That is, profile-based confidence intervals on the fixed-effects parameter, β_0, are symmetric about the estimate but overdispersed relative to those that would be calculated from a normal distribution and the logarithm of the residual standard deviation, $\log(\sigma)$, has a good normal approximation. Profile-based confidence intervals

for the standard deviations of random effects (σ_1, σ_2, etc.) are symmetric on a logarithmic scale except for those that could be zero.

Another observation from the last example is that, for data sets with a very large numbers of observations, a term in a model may be "statistically significant" even when its practical significance is questionable.

Exercises

These exercises use data sets from the `MEMSS` package for R. Recall that to access a particular data set, you must either attach the package

```
> library(MEMSS)
```

or load just the one data set

```
> data(ergoStool, package = "MEMSS")
```

We begin with exercises using the `ergoStool` data from the `MEMSS` package. The analysis and graphics in these exercises is performed in Chap. 3. The purpose of these exercises is to see if you can use the material from this chapter to anticipate the results quoted in the next chapter.

2.1. Check the documentation, the structure (`str`) and a summary of the `ergoStool` data from the `MEMSS` package. (If you are familiar with the Star Trek television series and movies, you may want to speculate about what, exactly, the "Borg scale" is.) Use

```
> xtabs(~ Type + Subject, ergoStool)
```

to determine if these factors are nested, partially crossed or completely crossed. Is this a replicated or an unreplicated design?

2.2. Create a plot, similar to Fig. 2.1, showing the effort by subject with lines connecting points corresponding to the same stool types. Order the levels of the `Subject` factor by increasing average `effort`.

2.3. The experimenters are interested in comparing these specific stool types. In the next chapter we will fit a model with fixed-effects for the `Type` factor and random effects for `Subject`, allowing us to perform comparisons of these specific types. At this point fit a model with random effects for both `Type` and `Subject`. What are the relative sizes of the estimates of the standard deviations, $\widehat{\sigma}_1$ (for `Subject`), $\widehat{\sigma}_2$ (for `Type`) and $\widehat{\sigma}$ (for the residual variability)?

2.4. Refit the model using maximum likelihood. Check the parameter estimates and, in the case of the fixed-effects parameter, β_0, its standard error. In what ways have the parameter estimates changed? Which parameter estimates have not changed?

2.5. Profile the fitted model and construct 95% profile-based confidence intervals on the parameters. (Note that you will get the same profile object whether you start with the REML fit or the ML fit. There is a slight advantage in starting with the ML fit.) Is the confidence interval on σ_1 close to being symmetric about its estimate? Is the confidence interval on σ_2 close to being symmetric about its estimate? Is the corresponding interval on $\log(\sigma_1)$ close to being symmetric about its estimate?

2.6. Create the profile zeta plot for this model. For which parameters are there good normal approximations?

2.7. Create a profile pairs plot for this model. Comment on the shapes of the profile traces in the transformed (ζ) scale and the shapes of the contours in the original scales of the parameters.

2.8. Create a plot of the 95% prediction intervals on the random effects for `Type` using

```
> dotplot(ranef(fm, which = "Type", postVar = TRUE), aspect = 0.2,
+         strip = FALSE)
```

(Substitute the name of your fitted model for `fm` in the call to `ranef`.) Is there a clear winner among the stool types? (Assume that lower numbers on the Borg scale correspond to less effort).

2.9. Create a plot of the 95% prediction intervals on the random effects for `Subject`.

2.10. Check the documentation, the structure (`str`) and a summary of the `Meat` data from the `MEMSS` package. Use a cross-tabulation to discover whether `Pair` and `Block` are nested, partially crossed or completely crossed.

2.11. Use a cross-tabulation

```
> xtabs(~ Pair + Storage, Meat)
```

to determine whether `Pair` and `Storage` are nested, partially crossed or completely crossed.

2.12. Fit a model of the `score` in the `Meat` data with random effects for `Pair`, `Storage` and `Block`.

2.13. Plot the prediction intervals for each of the three sets of random effects.

2.14. Profile the parameters in this model. Create a profile zeta plot. Does including the random effect for `Block` appear to be warranted. Does your conclusion from the profile zeta plot agree with your conclusion from examining the prediction intervals for the random effects for `Block`?

2.15. Refit the model without random effects for `Block`. Perform a likelihood ratio test of $H_0 : \sigma_3 = 0$ versus $H_a : \sigma_3 > 0$. Would you reject H_0 in favor of H_a or fail to reject H_0? Would you reach the same conclusion if you adjusted the p-value for the test by halving it, to take into account the fact that 0 is on the boundary of the parameter region?

2.16. Profile the reduced model (i.e. the one without random effects for `Block`) and create profile zeta and profile pairs plots. Can you explain the apparent interaction between $\log(\sigma)$ and σ_1? (This is a difficult question.)

Chapter 3
Models Incorporating Covariates

In the previous chapter we fit models having different configurations of simple, scalar random effects but always with the same fixed-effects specification of an intercept, or overall mean response, only.

It is common in practice to have several fixed-effects terms involving one or more covariates in the specification of a linear mixed model. Indeed, often the purpose of fitting a linear mixed model is to draw inferences about the effects of the covariates while appropriately accounting for different sources of variability in the responses.

In this chapter we demonstrate how fixed-effects terms are incorporated in a linear mixed model and how inferences about the effects of the covariates are drawn from a fitted linear mixed model.

3.1 Models for the `ergoStool` data

Problems 2.1 and 2.2 in Chap. 2 involve examining the structure of the `ergoStool` data from the `MEMSS` package

```
> str(ergoStool)
```

```
'data.frame':        36 obs. of  3 variables:
 $ effort : num  12 15 12 10 10 14 13 12 7 14 ...
 $ Type   : Factor w/ 4 levels "T1","T2","T3",..: 1 2 3 4 1 2 3 4 1 2 ...
 $ Subject: Factor w/ 9 levels "A","B","C","D",..: 1 1 1 1 2 2 2 2 3 3 ...
```

and plotting these data, as in Fig. 3.1. These data are from an ergometrics experiment where nine subjects evaluated the difficulty to arise from each of four types of stools. The measurements are on the scale of perceived exertion developed by the Swedish physician and researcher Gunnar Borg. Measurements on this scale are in the range 6-20 with lower values indicating less exertion.

From Fig. 3.1 we can see that all nine subjects rated type `T1` or type `T4` as requiring the least exertion and rated type `T4` as requiring the most exertion.

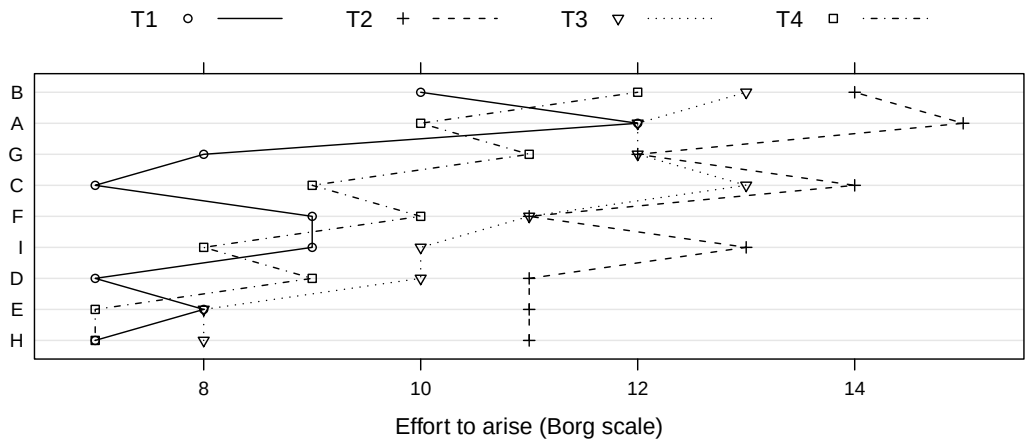


Fig. 3.1 Subjective evaluation of the effort required to arise (on the Borg scale) by 9 subjects, each of whom tried each of four types of stool.

Type `T3` was perceived as requiring comparatively little exertion by some subjects (H and E) and comparatively greater exertion by others (F, C and G).

Problem 2.3 involves fitting and evaluating a model in which the effects of both the `Subject` and the `Type` factors are incorporated as random effects. Such a model may not be appropriate for these data where we wish to make inferences about these particular four stool types. According to the distinction between fixed- and random-effects described in Sect. 1.1, if the levels of the `Type` factor are fixed and reproducible we generally incorporate the factor in the fixed-effects part of the model.

Before doing so, let's review the results of fitting a linear mixed model with random effects for both `Subject` and `Type`.

3.1.1 Random-effects for both `Subject` *and* `Type`

A model with random effects for both `Subject` and `Type` is fit in the same way that we fit models with random effects for multiple grouping factors in Chap. 2,

```
> (fm5 <- lmer(effort ~ 1 + (1|Subject) + (1|Type), ergoStool, REML=0))

Linear mixed model fit by maximum likelihood
Formula: effort ~ 1 + (1 | Subject) + (1 | Type)
   Data: ergoStool
   AIC   BIC logLik deviance
 144.0 150.4 -68.01    136.0
Random effects:
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 1.7040   1.3054
```

```
 Type     (Intercept) 2.2645   1.5048
 Residual             1.2128   1.1013
Number of obs: 36, groups: Subject, 9; Type, 4

Fixed effects:
            Estimate Std. Error t value
(Intercept)  10.2500     0.8883   11.54
```

from which we determine that the mean effort to arise, across stool types and across subjects, is 10.250 on this scale, with standard deviations of 1.305 for the random-effects for the `Subject` factor and 1.505 for the `Type` factor.

One question we would want to address is whether there are "significant" differences between stool types, taking into account the differences between subjects. We could approach this question by fitting a reduced model, without random effects for `Type`, and comparing this fit to model `fm5` using a likelihood-ratio test.

```
> fm5a <- lmer(effort ~ 1 + (1|Subject), ergoStool, REML=0)
> anova(fm5a, fm5)

Data: ergoStool
Models:
fm5a: effort ~ 1 + (1 | Subject)
fm5: effort ~ 1 + (1 | Subject) + (1 | Type)
     Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
fm5a  3 164.15 168.90 -79.075
fm5   4 144.02 150.36 -68.011 22.128      1  2.551e-06
```

The p-value in this test is very small, indicating that the more complex model, `fm5`, which allows for differences in the effort to arise for the different stool types, provides a significantly better fit to the observed data.

In Sect. 2.2.4 we indicated that, because the constraint on the reduced model, $\sigma_2 = 0$, is on the boundary of the parameter space, the p-value for this likelihood ratio test statistic calculated using a χ^2_1 reference distribution will be conservative. That is, the p-value one would obtain by, say, simulation from the null distribution, would be even smaller than the p-value, 0.0000026, reported by this test, which is already very small.

Thus the evidence against the null hypothesis ($H_0 : \sigma_2 = 0$) and in favor of the alternative, richer model ($H_a : \sigma_2 > 0$) is very strong.

Another way of addressing the question of whether it is reasonable for σ_2 to be zero is to profile `fm5` and examine profile zeta plots (Fig. 3.2) and the corresponding profile pairs plot (Fig. 3.3).

We can see from the profile zeta plot (Fig. 3.2) that although σ_2, the standard deviation of the random effect for `Type`, is safely non-zero, it is very poorly determined. That is, a 95% profile-based confidence interval on this parameter, obtained as

```
> confint(pr5)[".sig02",]

    2.5 %    97.5 %
0.7925434 3.7958504
```

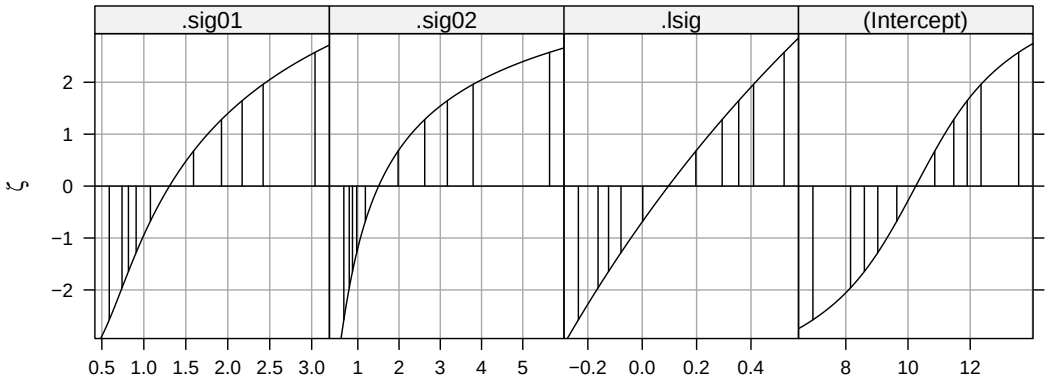


Fig. 3.2 Profile zeta plot for the parameters in model `fm5` fit to the `ergoStool` data

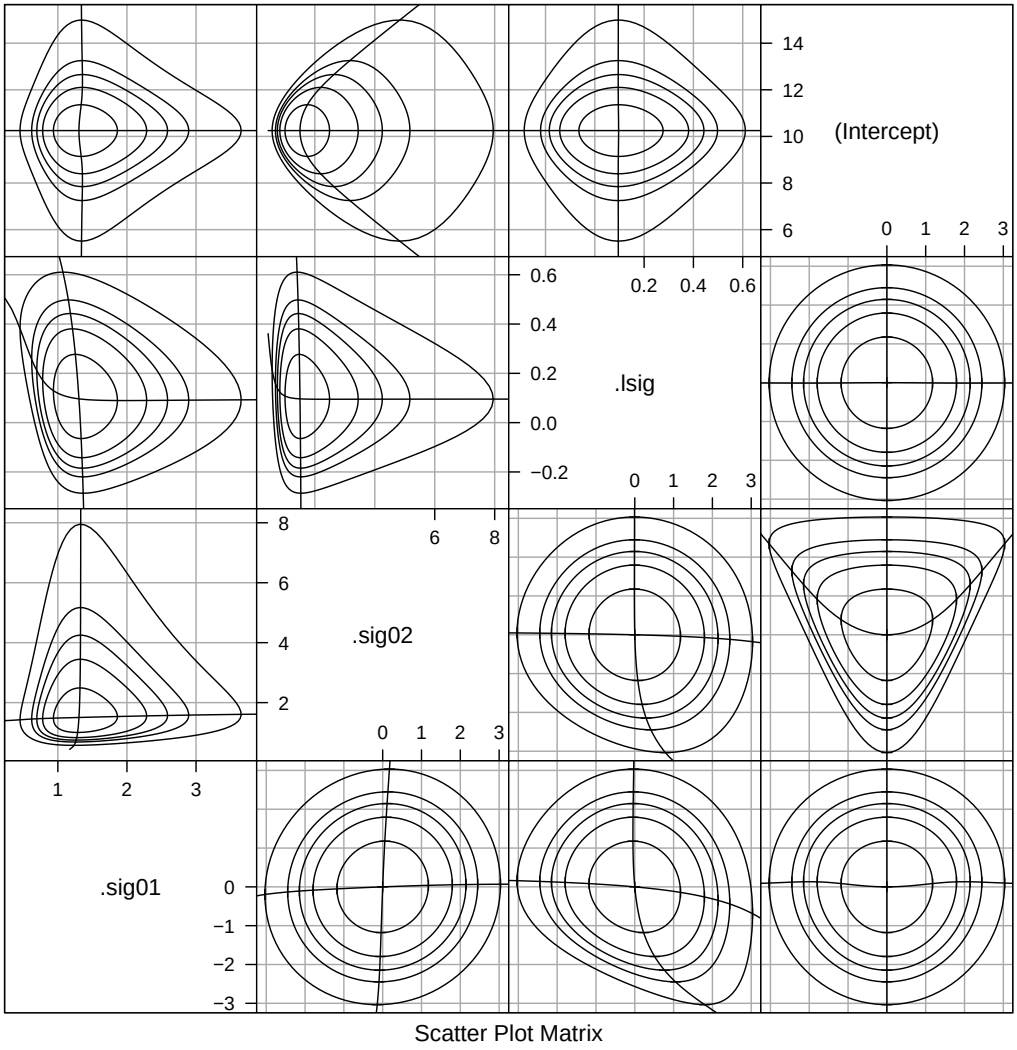


Fig. 3.3 Profile pairs plot for the parameters in model `fm5` fit to the `ergoStool` data

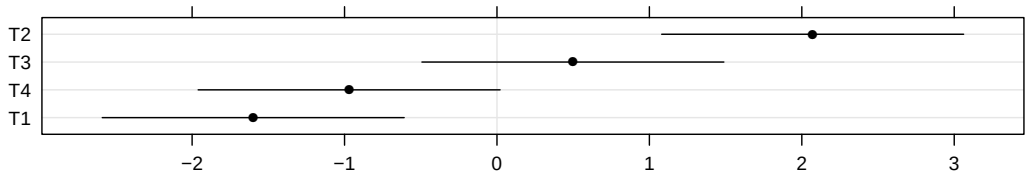


Fig. 3.4 95% prediction intervals on the random effects for `Type` from model `fm5` fit to the `ergoStool` data

is very wide. The upper end point of this 95% confidence interval, 3.796, is more than twice as large as the estimate, $\widehat{\sigma_2} = 1.505$

A plot of the prediction intervals on the random effects for `Type` (Fig. 3.4) confirms the impression from Fig. 3.1 regarding the stool types. Type `T2` requires the greatest effort and type `T1` requires the least effort. There is considerable overlap of the prediction intervals for types `T1` and `T4` and somewhat less overlap between types `T4` and `T3` and between types `T3` and `T2`.

Typically in an analysis like this we begin by asking if there are any significant differences between the stool types, which we answered for this model by testing the hypothesis $H_0 : \sigma_2 = 0$ versus $H_a : \sigma_2 > 0$. If we reject H_0 in favor of H_a — that is, if we conclude that the more complex model including random effects for `Type` provides a significantly better fit than the simpler model — then usually we want to follow up with the question, "Which stool types are significantly different from each other?". It is possible, though not easy, to formulate an answer to that question from a model fit such as `fm5` in which the stool types are modeled with random effects, but it is more straightforward to address that question when we model the stool types as fixed-effects parameters, which we do next.

3.1.2 Fixed-effects for `Type`*, Random Effects for* `Subject`

To incorporate the `Type` factor in the fixed-effects parameters, instead of as a grouping factor for random effects, we remove the random-effects term, `(1|Type)`, and add `Type` to the fixed-effects specification.

```
> (fm6 <- lmer(effort ~ 1 + Type + (1|Subject), ergoStool, REML = 0))

Linear mixed model fit by maximum likelihood
Formula: effort ~ 1 + Type + (1 | Subject)
   Data: ergoStool
   AIC   BIC logLik deviance
 134.1 143.6 -61.07    122.1
Random effects:
```

```
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 1.5782   1.2563
 Residual             1.0761   1.0374
Number of obs: 36, groups: Subject, 9

Fixed effects:
            Estimate Std. Error t value
(Intercept)   8.5556     0.5431  15.754
TypeT2        3.8889     0.4890   7.952
TypeT3        2.2222     0.4890   4.544
TypeT4        0.6667     0.4890   1.363

Correlation of Fixed Effects:
       (Intr) TypeT2 TypeT3
TypeT2 -0.450
TypeT3 -0.450  0.500
TypeT4 -0.450  0.500  0.500
```

It appears that the last three levels of the `Type` factor are now modeled as fixed-effects parameters in addition to the `(Intercept)` parameter, whose estimate has decreased markedly from that in model `fm5`. Furthermore, the estimates of the fixed-effects parameters labeled `TypeT2`, `TypeT3` and `TypeT4`, while positive, are very much smaller than would be indicated by the average responses for these types.

It turns out, of course, that the fixed-effects parameters generated by a factor covariate do not correspond to the overall mean and the effect for each level of the covariate. Although a model for an experiment such as this is sometimes written in a form like

$$y_{ij} = \mu + \alpha_i + b_j + \varepsilon_{ij}, \quad i = 1,\ldots,4,\ j = 1,\ldots 9 \tag{3.1}$$

where i indexes the stool type and j indexes the subject, the parameters $\{\mu, \alpha_1, \alpha_2, \alpha_3, \alpha_4\}$, representing the overall mean and the effects of each of the stool types, are redundant. If we have a set of estimates for these parameters we would not change the predictions from the model if, for example, we added one to μ and subtracted one from all the α's. In statistical terminology we say that this set of parameters is not *estimable* unless we impose some other conditions on them. The estimability condition $\sum_{i=1}^{4} \alpha_i = 0$ is often used in introductory texts.

The approach taken in R is not based on redundant parameters that are subject to estimability conditions. While this approach may seem reasonable initially, in complex models it quickly becomes unnecessarily complex to need to use constrained optimization for parameter estimation. Instead we incorporate the constraints into the parameters that we estimate. That is, we reduce the redundant set of parameters to an estimable set of *contrasts* between the levels of the factors.

3.1.2.1 The default contrasts generated for a factor

Although the particular set of contrasts used for a categorical factor can be controlled by the user, either as a global option for a session (see `?options`) or by the optional `contrasts` argument available in most model-fitting functions, most users do not modify the contrasts, preferring to leave them at the default setting, which is the "treatment" contrasts (`contr.treatment`) for an unordered factor and orthogonal polynomial contrasts (`contr.poly`) for an ordered factor. You can check the current global setting with

```
> getOption("contrasts")
```

```
        unordered           ordered
"contr.treatment"      "contr.poly"
```

Because these were the contrasts in effect when model `fm6` was fit, the particular contrasts used for the `Type` factor, which has four levels, correspond to

```
> contr.treatment(4)
```

```
  2 3 4
1 0 0 0
2 1 0 0
3 0 1 0
4 0 0 1
```

In this display the rows correspond to the levels of the `Type` factor and the columns correspond to the parameters labeled `TypeT2`, `TypeT3` and `TypeT4`.

The values of `Type` in the data frame, whose first few rows are

```
> head(ergoStool)
```

```
  effort Type Subject
1     12   T1       A
2     15   T2       A
3     12   T3       A
4     10   T4       A
5     10   T1       B
6     14   T2       B
```

combined with the contrasts produce the model matrix $\mathbf{X}$, whose first few rows are

```
> head(with(env(fm6), X))
```

```
6 x 4 Matrix of class "dgeMatrix"
     (Intercept) TypeT2 TypeT3 TypeT4
[1,]           1      0      0      0
[2,]           1      1      0      0
[3,]           1      0      1      0
[4,]           1      0      0      1
[5,]           1      0      0      0
[6,]           1      1      0      0
```

We see that the rows of $\mathbf{X}$ for observations on stool type `T1` have zeros in the last three columns; the rows for observations on stool type `T2` have a 1 in the second column and zeros in the last two columns, and so on. As before, the `(Intercept)` column is a column of 1's.

When we evaluate $\mathbf{X}\beta$ in the linear predictor expression, $\mathbf{X}\beta+\mathbf{Zb}$, we take the p elements of the fixed-effects parameter vector, β, whose estimate is

```
> fixef(fm6)
```

```
(Intercept)      TypeT2      TypeT3      TypeT4
  8.5555556   3.8888889   2.2222222   0.6666667
```

and the p elements of a row of the matrix $\mathbf{X}$, multiply corresponding components and sum the resulting products. For example, the fixed-effects predictor for the first observation (stool type `T1`) will be

$$8.5556\times 1+3.8889\times 0+2.2222\times 0+0.6667\times 0=8.5556$$

and the fixed-effects predictor for the second observation (stool type `T2`) will be

$$8.5556\times 1+3.8889\times 1+2.2222\times 0+0.6667\times 0=12.4444$$

We see that the parameter labeled `(Intercept)` is actually the fixed-effects prediction for the first level of `Type` (i.e. level `T1`) and the second parameter, labeled `TypeT2`, is the difference between the fixed-effects prediction for the second level (`T2`) and the first level (`T1`) of the `Type` factor.

Similarly, the fixed-effects predictions for the `T3` and `T4` levels of `Type` are $8.5556+2.2222=10.7778$ and $8.5556+0.6667=9.2222$, respectively, as can be verified from

```
> with(env(fm6), head(as.vector(X %*% fixef)))
```

```
[1]  8.555556 12.444444 10.777778  9.222222  8.555556 12.444444
```

The fact that the parameter labeled `TypeT2` is the difference between the fixed-effects prediction for levels `T2` and `T1` of the `Type` factor is why we refer to the parameters as being generated by *contrasts*. They are formed by contrasting the fixed-effects predictions for some combination of the levels of the factor. In this case the contrast is between levels `T2` and `T1`.

In general, the parameters generated by the "treatment" contrasts (the default for unordered factors) represent differences between the first level of the factor, which is incorporated into the `(Intercept)` parameter, and the subsequent levels. We say that the first level of the factor is the *reference* level and the others are characterized by their shift relative to this reference level.

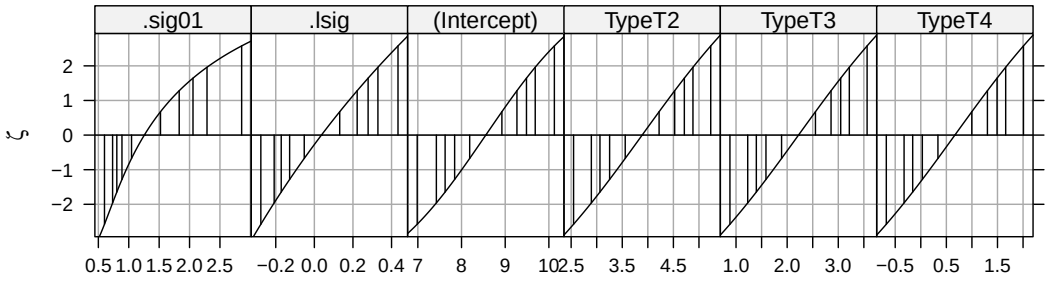


Fig. 3.5 Profile zeta plot for the parameters in model `fm5` fit to the `ergoStool` data

3.1.2.2 Profiling the contrasts

Because some of the contrasts that are of interest to us correspond to fixed-effects parameter values, we can profile the fitted model (Fig. 3.5) and check, say, the confidence intervals on these parameters.

```
> confint(pr6, c("TypeT2", "TypeT3", "TypeT4"))

            2.5 %   97.5 %
TypeT2  2.8953043 4.882473
TypeT3  1.2286377 3.215807
TypeT4 -0.3269179 1.660251
```

According to these intervals, and from what we see from Fig. 3.5, types `T2` and `T3` are significantly different from type T1 (the intervals do not contain zero) but type `T4` is not (the confidence interval on this contrast contains zero).

However, this process must be modified in two ways in two ways to provide a suitable answer. The most important modification is to take into account the fact that we are performing *multiple comparisons* simultaneously. We describe what this means and how to accomodate for it in the next subsection. The other problem is that this process only allows us to evaluate contrasts of the reference level, `T1`, with the other levels and the reference level is essentially arbitrary. For completeness we should evaluate all six possible contrasts of pairs of levels.

We can do this by refitting the model with a difference reference level for the `Type` factor and profiling the modified model fit. The `relevel` function allows us to change the reference level of a factor.

```
> pr6a <- profile(lmer(effort ~ 1 + Type + (1|Subject),
+                      within(ergoStool, Type <- relevel(Type, "T2")),
+                      REML = 0))
> pr6b <- profile(lmer(effort ~ 1 + Type + (1|Subject),
```

```
+                       within(ergoStool, Type <- relevel(Type, "T3")),
+                       REML = 0))
```

The other constrasts of interest are

```
> confint(pr6a, c("TypeT3", "TypeT4"))

           2.5 %     97.5 %
TypeT3 -2.660251 -0.6730821
TypeT4 -4.215807 -2.2286377

> confint(pr6b, "TypeT4")

          2.5 %    97.5 %
TypeT4 -2.54914 -0.561971
```

from which would conclude that type `T2` requires significantly greater effort than any of the other types at the 5% level (because none of the 95% confidence intervals on contrasts with `T2` contain zero) and that types `T3` and `T4` are significantly different at the 5% level.

However, we must take into account that we are performing multiple, simulataneous comparisons of levels.

3.1.2.3 Multiple comparisons

In the technical definition of a confidence interval we regard the end points as being random (because they are calculated from the random variable which is the observed data) and the value of the parameter or the contrast of interest as being fixed (because it is determined from the fixed, but unknown, values of the parameters). Thus we speak of the probability that an interval covers the true parameter value rather than the probability that the parameter falls in the interval. The distinction may seem, and probably is, somewhat pedantic. We introduce it here simply to clarify the term "coverage probability" used throughout this section.

We have evaluated six possible pairwise comparisons of the four levels of the `Type` factor. A 95% confidence interval on a particular contrast has, in theory, a 5% probability of failing to cover the true difference. That is, if the difference between two levels was in fact zero, there would still be a 5% probability that a 95% confidence interval on that contrast would not include zero. When we consider the coverage of the six intervals contrasting all possible pairs of stool types we usually have in mind that there should be a 95% probability of all six intervals covering the true, but unknown, differences in effort for the stool types. That is, we think of the coverage probability as applying to the simultaneous coverage of the family of intervals, not to the coverage of one specific interval.

But the intervals calculated in the previous section were based on 95% coverage for each specific interval. In the worst case scenario the family-wise coverage could be as low as $1 - 0.05 * 6 = 0.70$ or 70%. For factors with more

than four levels there are even more possible pairwise comparisons (for k levels there are $k(k-1)/2$ possible pairs) and this worst-case coverage probability is even further from the nominal level of 95%.

Several methods have been developed to compensate for multiple comparisons in the analysis of linear models with fixed effects only. One of the simplest, although somewhat coarse, compensations is the Bonferroni correction where the individual intervals are chosen to have a greater coverage probability in such a way that the "worst-case" probability is the desired level. With six comparisons to get a family-wise coverage probability of 95% the individual intervals are chosen to have coverage of

```
> (covrge <- 1 - 0.05/6)

[1] 0.9916667
```

or a little more than 99%. We can specify this coverage level for the individual intervals to ensure a family-wise coverage of at least 95%.

```
> rbind(confint(pr6, c("TypeT2","TypeT3","TypeT4"), covrge),
+       confint(pr6a, c("TypeT3","TypeT4"), covrge),
+       confint(pr6b, "TypeT4", covrge))

          0.417 %   99.583 %
TypeT2  2.5109497  5.2668280
TypeT3  0.8442830  3.6001613
TypeT4 -0.7112726  2.0446058
TypeT3 -3.0446059 -0.2887275
TypeT4 -4.6001614 -1.8442831
TypeT4 -2.9334948 -0.1776164
```

We again reach the conclusion that the only pair of stool types for which zero is within the confidence interval on the difference in effects is the (T1,T4) pair but, for these intervals, the family-wise coverage of all six intervals is at least 95%.

There are other, perhaps more effective, techniques for adjusting intervals to take into account multiple comparisons. The purpose of this section is to show that the profile-based confidence intervals can be extended to at least the Bonferroni correction.

The easiest way to apply other multiple comparison adjustment methods is to model both the Type and the Subject factors with fixed effects, which we do next.

3.1.3 Fixed-effects for both Type *and* Subject

Even though the subjects in this study are chosen as representatives of a population, many statisticians would regard Subject as a *blocking factor* in the experiment and fit a model with fixed-effects for both Type and Subject. A blocking factor is a known source of variability in the response. We are

not interested in the effects of the levels of the blocking factor—we only wish to accomodate for this source of variability when comparing levels of the experimental factor, which is the `Type` factor in this example.

We will discuss the advantages and disadvantages of the fixed- versus random-effects choice for the `Subject` factor at the end of this section. For the moment we proceed to fit the fixed-effects model, for which we could use the `lm` function or the `aov` function. These two functions produce exactly the same model fit but the `aov` function returns an object of class `"aov"` which extends the class `"lm"`, providing more options for examining the fitted model.

```
> summary(fm7 <- aov(effort ~ Subject + Type, ergoStool))

            Df Sum Sq Mean Sq F value    Pr(>F)
Subject      8 66.500  8.3125  6.8662 0.0001061
Type         3 81.194 27.0648 22.3556 3.935e-07
Residuals   24 29.056  1.2106
```

3.2 Covariates Affecting Mathematics Score Gain

West et al. [2007] provides comparisons of several different software systems for fitting linear mixed models by fitting sample models to different data sets using each of these software systems. The `lmer` function from the `lme4` package is not included in these comparisons because it was still being developed when that book was written.

In this section we will use `lmer` to fit models described in Chap. 4 of West et al. [2007] to data on the gain in mathematics scores for students in a selection of classrooms in several schools.

```
> str(classroom)

'data.frame':        1190 obs. of  11 variables:
 $ sex     : Factor w/ 2 levels "M","F": 2 1 2 1 1 2 1 1 2 1 ...
 $ minority: Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
 $ mathkind: int  448 460 511 449 425 450 452 443 422 480 ...
 $ mathgain: int  32 109 56 83 53 65 51 66 88 -7 ...
 $ ses     : num  0.46 -0.27 -0.03 -0.38 -0.03 0.76 -0.03 0.2 0.64 0.13 ...
 $ yearstea: num  1 1 1 2 2 2 2 2 2 2 ...
 $ mathknow: num  NA NA NA -0.11 -0.11 -0.11 -0.11 -0.11 -0.11 -0.11 ...
 $ housepov: num  0.082 0.082 0.082 0.082 0.082 0.082 0.082 0.082 0.082 0.082 ..
 $ mathprep: num  2 2 2 3.25 3.25 3.25 3.25 3.25 3.25 3.25 ...
 $ classid : Factor w/ 312 levels "1","2","3","4",..: 160 160 160 217 217 217 ..
 $ schoolid: Factor w/ 107 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...

> (fm7 <- lmer(mathgain ~ 1 + I(mathkind-450) + sex + minority + ses
+              + housepov + (1|classid) + (1|schoolid), classroom))

Linear mixed model fit by REML
Formula: mathgain ~ 1 + I(mathkind - 450) + sex + minority + ses + housepov +      (1 | class
   Data: classroom
 REML
```

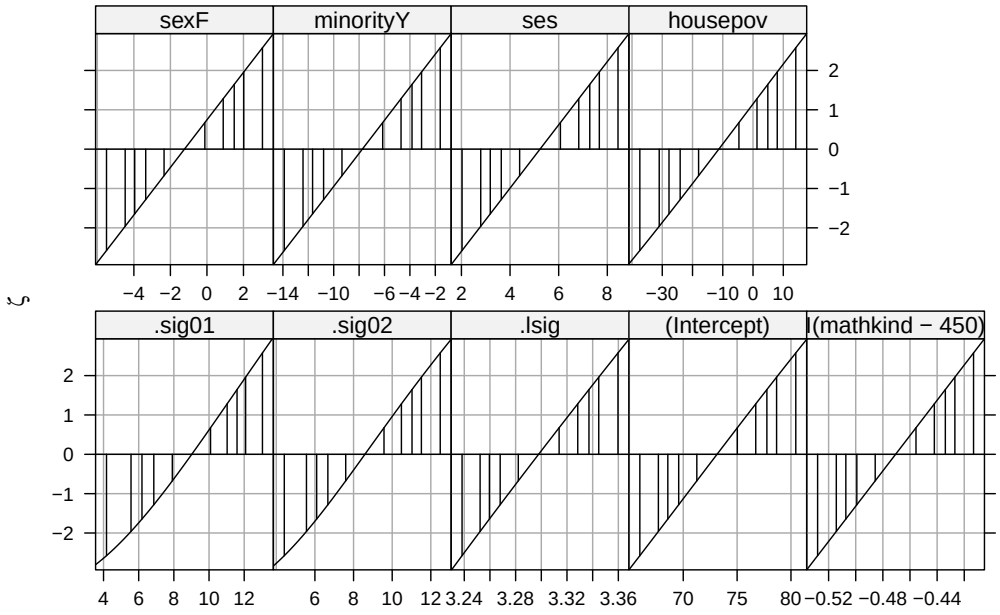


Fig. 3.6 Profile plot of the parameters in model `fm4`.

```
11378
Random effects:
 Groups   Name        Variance Std.Dev.
 classid  (Intercept)  81.555   9.0308
 schoolid (Intercept)  77.761   8.8182
 Residual             734.420  27.1002
Number of obs: 1190, groups: classid, 312; schoolid, 107

Fixed effects:
                  Estimate Std. Error t value
(Intercept)        73.17077    2.80273  26.107
I(mathkind - 450)  -0.47086    0.02228 -21.133
sexF               -1.23459    1.65743  -0.745
minorityY          -7.75587    2.38499  -3.252
ses                 5.23971    1.24497   4.209
housepov          -11.43920    9.93736  -1.151

Correlation of Fixed Effects:
            (Intr) I(-450 sexF   mnrtyY ses
I(mthk-450) -0.233
sexF        -0.279 -0.032
minorityY   -0.492  0.153 -0.015
ses         -0.105 -0.165  0.019  0.144
housepov    -0.555  0.035 -0.009 -0.184  0.078
```

A profile plot of the parameters in model `fm7` is shown in Fig. 3.6

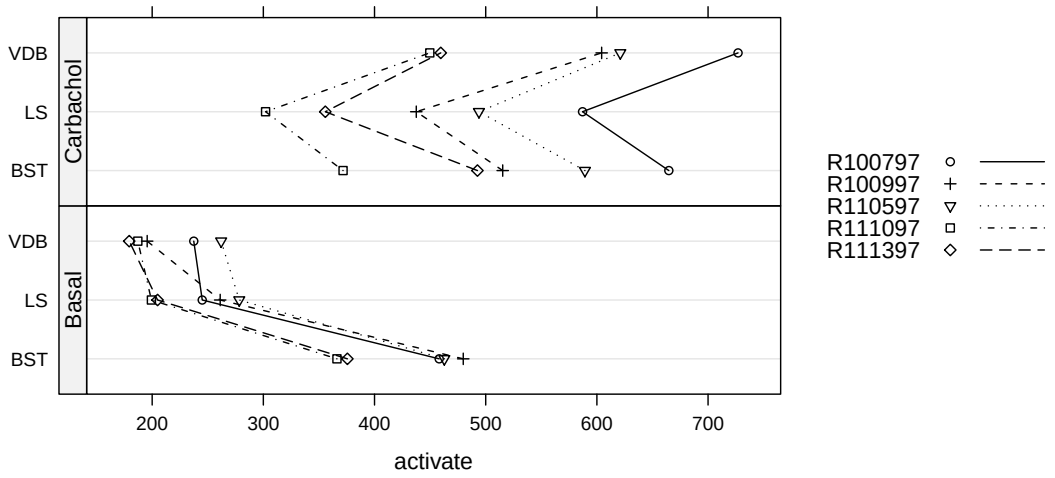


Fig. 3.7 Activation of brain regions in rats

3.3 Rat Brain example

```
> ftable(xtabs(activate ~ animal + treatment + region, ratbrain))

                 region    BST     LS    VDB
animal  treatment
R100797 Basal            458.16 245.04 237.42
        Carbachol        664.72 587.10 726.96
R100997 Basal            479.81 261.19 195.51
        Carbachol        515.29 437.56 604.29
R110597 Basal            462.79 278.33 262.05
        Carbachol        589.25 493.93 621.07
R111097 Basal            366.19 199.31 187.11
        Carbachol        371.71 302.02 449.70
R111397 Basal            375.58 204.85 179.38
        Carbachol        492.58 355.74 459.58
```

Description of the Rat Brain data should go here.

Chapter 4
Models for Longitudinal Data

Longitudinal data consist of repeated measurements on the same subject (or some other "experimental unit") taken over time. Generally we wish to characterize the time trends within subjects and between subjects. The data will always include the response, the time covariate and the indicator of the subject on which the measurement has been made. If other covariates are recorded, say whether the subject is in the treatment group or the control group, we may wish to relate the within- and between-subject trends to such covariates.

In this chapter we introduce graphical and statistical techniques for the analysis of longitudinal data by applying them to a simple example.

4.1 The `sleepstudy` Data

Belenky et al. [2003] report on a study of the effects of sleep deprivation on reaction time for a number of subjects chosen from a population of long-distance truck drivers. These subjects were divided into groups that were allowed only a limited amount of sleep each night. We consider here the group of 18 subjects who were restricted to three hours of sleep per night for the first ten days of the trial. Each subject's reaction time was measured several times on each day of the trial.

```
> str(sleepstudy)

'data.frame':        180 obs. of  3 variables:
 $ Reaction: num  250 259 251 321 357 ...
 $ Days    : num  0 1 2 3 4 5 6 7 8 9 ...
 $ Subject : Factor w/ 18 levels "308","309","310",..: 1 1 1 1 1 1 1 1 1..
```

In this data frame, the response variable `Reaction`, is the average of the reaction time measurements on a given subject for a given day. The two covariates are `Days`, the number of days of sleep deprivation, and `Subject`, the identifier of the subject on which the observation was made.

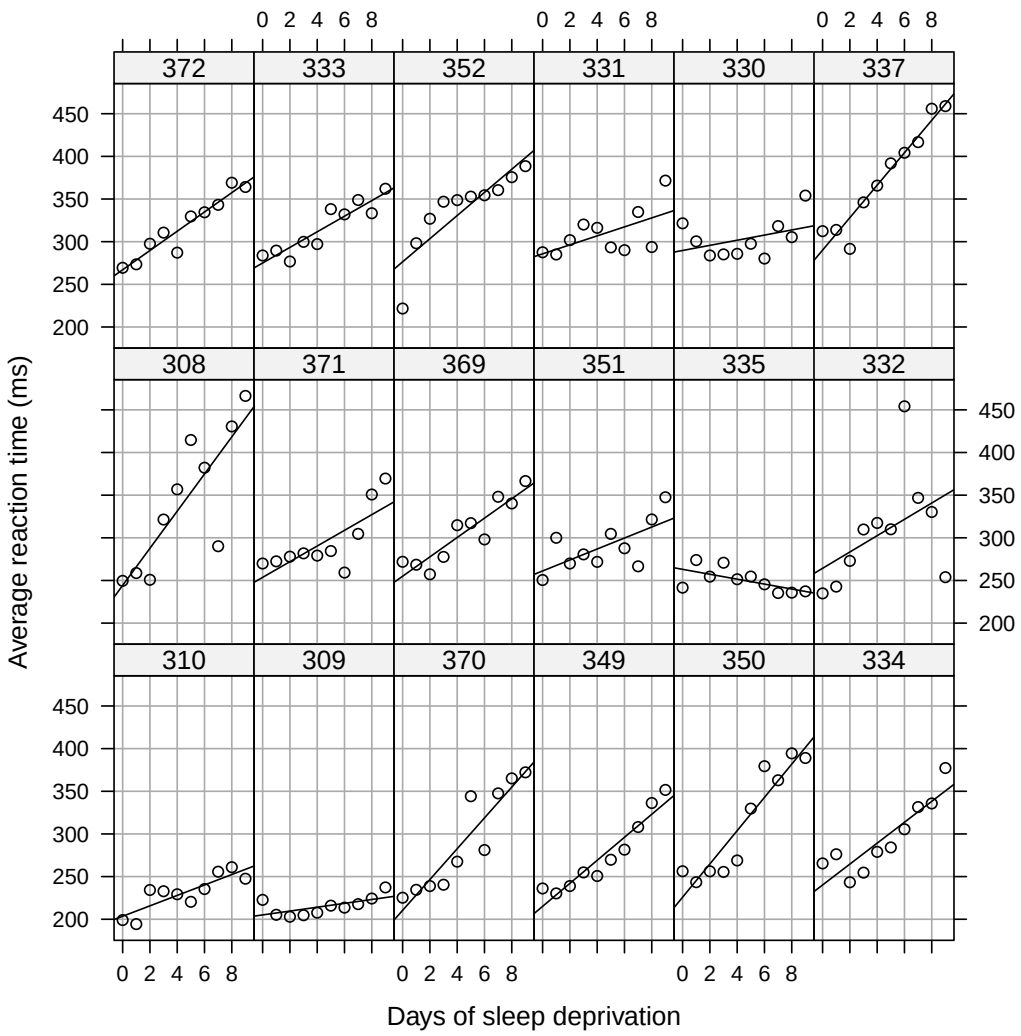


Fig. 4.1 A lattice plot of the average reaction time versus number of days of sleep deprivation by subject for the `sleepstudy` data. Each subject's data are shown in a separate panel, along with a simple linear regression line fit to the data in that panel. The panels are ordered, from left to right along rows starting at the bottom row, by increasing intercept of these per-subject linear regression lines. The subject number is given in the strip above the panel.

As recommended for any statistical analysis, we begin by plotting the data. The most important relationship to plot for longitudinal data on multiple subjects is the trend of the response over time by subject, as shown in Fig. 4.1. This plot, in which the data for different subjects are shown in separate panels with the axes held constant for all the panels, allows for examination of the time-trends within subjects and for comparison of these patterns between subjects. Through the use of small panels in a repeating pattern Fig. 4.1 conveys a great deal of information, the individual time trends for 18 subjects over 10 days — a total of 180 points — without being overly cluttered.

4.1.1 Characteristics of the sleepstudy Data Plot

The principles of "Trellis graphics", developed by Bill Cleveland and his coworkers at Bell Labs and implemented in the lattice package for R by Deepayan Sarkar, have been incorporated in this plot. As stated above, all the panels have the same vertical and horizontal scales, allowing us to evaluate the pattern over time for each subject and also to compare patterns between subjects. The line drawn in each panel is a simple least squares line fit to the data in that panel only. It is provided to enhance our ability to discern patterns in both the slope (the typical change in reaction time per day of sleep deprivation for that particular subject) and the intercept (the average response time for the subject when on their usual sleep pattern).

The aspect ratio of the panels (ratio of the height to the width) has been chosen, according to an algorithm described in Cleveland [1993], to facilitate comparison of slopes. The effect of choosing the aspect ratio in this way is to have the slopes of the lines on the page distributed around $\pm 45^\circ$, thereby making it easier to detect systematic changes in slopes.

The panels have been ordered (from left to right starting at the bottom row) by increasing intercept. Because the subject identifiers, shown in the strip above each panel, are unrelated to the response it would not be helpful to use the default ordering of the panels, which is by increasing subject number. If we did so our perception of patterns in the data would be confused by the, essentially random, ordering of the panels. Instead we use a characteristic of the data to determine the ordering of the panels, thereby enhancing our ability to compare across panels. For example, a question of interest to the experimenters is whether a subject's rate of change in reaction time is related to the subject's initial reaction time. If this were the case we would expect that the slopes would show an increasing trend (or, less likely, a decreasing trend) in the left to right, bottom to top ordering.

There is little evidence in Fig. 4.1 of such a systematic relationship between the subject's initial reaction time and their rate of change in reaction time per day of sleep deprivation. We do see that for each subject, except 335, reaction time increases, more-or-less linearly, with days of sleep deprivation. However, there is considerable variation both in the initial reaction time and in the daily rate of increase in reaction time. We can also see that these data are balanced, both with respect to the number of observations on each subject, and with respect to the times at which these observations were taken. This can be confirmed by cross-tabulating Subject and Days.

```
> xtabs(~ Subject + Days, sleepstudy)

       Days
Subject 0 1 2 3 4 5 6 7 8 9
    308 1 1 1 1 1 1 1 1 1 1
    309 1 1 1 1 1 1 1 1 1 1
    310 1 1 1 1 1 1 1 1 1 1
    330 1 1 1 1 1 1 1 1 1 1
```

```
331 1 1 1 1 1 1 1 1 1 1
332 1 1 1 1 1 1 1 1 1 1
333 1 1 1 1 1 1 1 1 1 1
334 1 1 1 1 1 1 1 1 1 1
335 1 1 1 1 1 1 1 1 1 1
337 1 1 1 1 1 1 1 1 1 1
349 1 1 1 1 1 1 1 1 1 1
350 1 1 1 1 1 1 1 1 1 1
351 1 1 1 1 1 1 1 1 1 1
352 1 1 1 1 1 1 1 1 1 1
369 1 1 1 1 1 1 1 1 1 1
370 1 1 1 1 1 1 1 1 1 1
371 1 1 1 1 1 1 1 1 1 1
372 1 1 1 1 1 1 1 1 1 1
```

In cases like this where there are several observations (10) per subject and a relatively simple within-subject pattern (more-or-less linear) we may want to examine coefficients from within-subject fixed-effects fits. However, because the subjects constitute a sample from the population of interest and we wish to drawn conclusions about typical patterns in the population and the subject-to-subject variability of these patterns, we will eventually want to fit mixed models and we begin by doing so. In section 4.4 we will compare estimates from a mixed-effects model with those from the within-subject fixed-effects fits.

4.2 Mixed-effects Models For the `sleepstudy` Data

Based on our preliminary graphical exploration of these data, we fit a mixed-effects model with two fixed-effects parameters, the intercept and slope of the linear time trend for the population, and two random effects for each subject. The random effects for a particular subject are the deviations in intercept and slope of that subject's time trend from the population values.

We will fit two linear mixed models to these data. One model, `fm8`, allows for correlation (in the unconditional distribution) of the random effects for the same subject. That is, we allow for the possibility that, for example, subjects with higher initial reaction times may, on average, be more strongly affected by sleep deprivation. The second model provides independent (again, in the unconditional distribution) random effects for intercept and slope for each subject.

4.2.1 A Model With Correlated Random Effects

The first model is fit as

```
> (fm8 <- lmer(Reaction ~ 1 + Days + (1 + Days|Subject), sleepstudy,
+              REML = 0))

Linear mixed model fit by maximum likelihood
Formula: Reaction ~ 1 + Days + (1 + Days | Subject)
   Data: sleepstudy
  AIC  BIC logLik deviance
 1764 1783   -876     1752
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 565.516  23.7806
          Days         32.682   5.7168  0.081
 Residual             654.941  25.5918
Number of obs: 180, groups: Subject, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept)  251.405      6.632   37.91
Days          10.467      1.502    6.97

Correlation of Fixed Effects:
     (Intr)
Days -0.138
```

From the display we see that this model incorporates both an intercept and a slope (with respect to `Days`) in the fixed effects and in the random effects. Extracting the conditional modes of the random effects

```
> head(ranef(fm8)[["Subject"]])

    (Intercept)       Days
308    2.815683  9.0755340
309  -40.048490 -8.6440671
310  -38.433156 -5.5133785
330   22.832297 -4.6587506
331   21.549991 -2.9445203
332    8.815587 -0.2352093
```

confirms that these are *vector-valued* random effects. There are a total of $q = 36$ random effects, two for each of the 18 subjects.

The random effects section of the model display,

```
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 565.516  23.7806
          Days         32.682   5.7168  0.081
 Residual             654.941  25.5918
```

indicates that there will be a random effect for the intercept and a random effect for the slope with respect to `Days` at each level of `Subject` and, furthermore, the unconditional distribution of these random effects, $\mathcal{B} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, allows for correlation of the random effects for the same subject.

We can confirm the potential for correlation of random effects within subject in the images of $\boldsymbol{\Lambda}$, $\boldsymbol{\Sigma}$ and $\mathbf{L}$ for this model (Fig. 4.2). The matrix $\boldsymbol{\Lambda}$ has

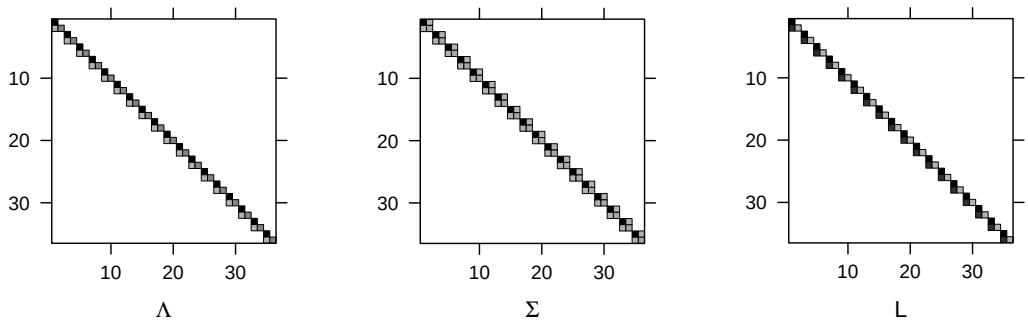


Fig. 4.2 Images of Λ, Σ and $\mathbf{L}$ for model `fm8`

18 triangular blocks of size 2 along the diagonal, generating 18 square, symmetric blocks of size 2 along the diagonal of Σ. The 18 symmetric blocks on the diagonal of Σ are identical. Overall we estimate two standard deviations and a correlation for a vector-valued random effect of size 2, as shown in the model summary.

Often the variances and the covariance of random effects are quoted, rather than the standard deviations and the correlation shown here. We have already seen that the variance of a random effect is a poor scale on which to quote the estimate because confidence intervals on the variance are so badly skewed. It is more sensible to assess the estimates of the standard deviations of random effects or, possibly, the logarithms of the standard deviations if we can be confident that 0 is outside the region of interest. We do display the estimates of the variances of the random effects but mostly so that the user can compare these estimates to those from other software or for cases where an estimate of a variance is expected (sometimes even required) to be given when reporting a mixed model fit.

We do not quote estimates of covariances of vector-valued random effects because the covariance is a difficult scale to interpret whereas a correlation has a fixed scale. A correlation must be between -1 and 1, allowing us to conclude that a correlation estimate close to those extremes indicates that Σ is close to singular and the model is not well formulated.

The estimates of the fixed effects parameters are $\widehat{\beta} = (251.41, 10.467)^{\mathsf{T}}$. These represent a typical initial reaction time (i.e. without sleep deprivation) in the population of about 250 milliseconds, or 1/4 sec., and a typical increase in reaction time of a little more than 10 milliseconds per day of sleep deprivation.

The estimated subject-to-subject variation in the intercept corresponds to a standard deviation of about 25 ms. A 95% prediction interval on this random variable would be approximately ± 50 ms. Combining this range with a population estimated intercept of 250 ms. indicates that we should not be

surprised by intercepts as low as 200 ms. or as high as 300 ms. This range is consistent with the reference lines shown in Figure 4.1.

Similarly, the estimated subject-to-subject variation in the slope corresponds to a standard deviation of about 6 ms./day so we would not be surprised by slopes as low as $10.5 - 2 \cdot 5.7 = -0.9$ ms./day or as high as $10.5 + 2 \cdot 6 = 21.9$ ms./day. Again, the conclusions from these rough, "back of the envelope" calculations are consistent with our observations from Fig. 4.1.

The estimated residual standard deviation is about 25 ms. leading us to expect a scatter around the fitted lines for each subject of up to ± 50 ms. From Figure 4.1 we can see that some subjects (309, 372 and 337) appear to have less variation than ± 50 ms. about their within-subject fit but others (308, 332 and 331) may have more.

Finally, we see the estimated within-subject correlation of the random effect for the intercept and the random effect for the slope is very low, 0.081, confirming our impression that there is little evidence of a systematic relationship between these quantities. In other words, observing a subject's initial reaction time does not give us much information for predicting whether their reaction time will be strongly affected by each day of sleep deprivation or not. It seems reasonable that we could get nearly as good a fit from a model that does not allow for correlation, which we describe next.

4.2.2 A Model With Uncorrelated Random Effects

In a model with uncorrelated random effects we have $\mathcal{B} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ where Σ is diagonal. We have seen models like this in previous chapters but those models had simple scalar random effects for all the grouping factors. Here we want to have a simple scalar random effect for `Subject` and a random effect for the slope with respect to `Days`, also indexed by `Subject`. We accomplish this by specifying two random-effects terms. The first, `(1|Subject)`, is a simple scalar term. The second has `Days` on the left hand side of the vertical bar.

It seems that the model formula we want should be

```
Reaction ~ 1 + Days + (1 | Subject) + (Days | Subject)
```

but it is not. Because the intercept is implicit in linear models, the second random effects term is equivalent to `(1+Days|Subject)` and will, by itself, produce correlated, vector-valued random effects.

We must suppress the implicit intercept in the second random-effects term which we do by writing it as `(0+Days|Subject)`, read as "no intercept and `Days` by `Subject`". An alternative expression for `Days` without an intercept by `Subject` is `(Days - 1 | Subject)`. Using the first form we have

```
> (fm9 <- lmer(Reaction ~ 1 + Days + (1|Subject) + (0+Days|Subject),
+              sleepstudy, REML = 0))
```

```
Linear mixed model fit by maximum likelihood
Formula: Reaction ~ 1 + Days + (1 | Subject) + (0 + Days | Subject)
   Data: sleepstudy
  AIC  BIC logLik deviance
 1762 1778   -876     1752
Random effects:
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 584.249  24.1713
 Subject  Days         33.633   5.7994
 Residual             653.116  25.5561
Number of obs: 180, groups: Subject, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept)  251.405      6.708   37.48
Days          10.467      1.519    6.89

Correlation of Fixed Effects:
     (Intr)
Days -0.194
```

As in model `fm8`, there are two random effects for each subject

```
> head(ranef(fm9)[["Subject"]])
```

```
    (Intercept)       Days
308    1.854653  9.2364353
309  -40.022293 -8.6174753
310  -38.723150 -5.4343821
330   23.903313 -4.8581932
331   22.396316 -3.1048397
332    9.051998 -0.2821594
```

but no correlation has been estimated

```
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 584.249  24.1713
 Subject  Days         33.633   5.7994
 Residual             653.116  25.5561
```

The `Subject` factor is repeated in the "Groups" column because there were two distinct terms generating these random effects and these two terms had the same grouping factor.

Images of the matrices Λ, Σ and $\mathbf{L}$ (Fig. 4.3) show that Σ is indeed diagonal. The order of the random effects in Σ and Λ for model `fm9` is different from the order in model `fm8`. In model `fm8` the two random effects for a particular subject were adjacent. In model `fm9` all the intercept random effects occur first then all the `Days` random effects. The sparse Cholesky decomposition, $\mathbf{L}$, has the same form in both models because the fill-reducing permutation (described in Sect. 5.4.1) calculated for model `fm9` provides a post-ordering to group random effects with similar structure in $\mathbf{Z}$.

Images of $\mathbf{Z}^\mathsf{T}$ for these two models (Fig. 4.4) shows that the columns of $\mathbf{Z}$ (rows of $\mathbf{Z}^\mathsf{T}$) from one model are the same those from the other model but in a different order.

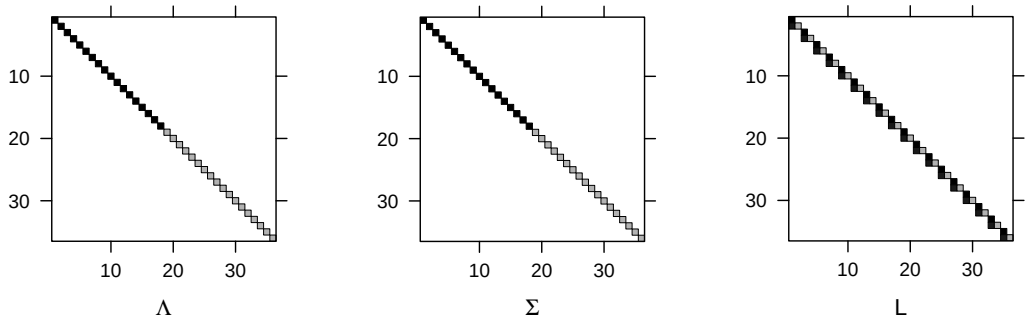


Fig. 4.3 Images of Λ, the relative covariance factor, Σ, the variance-covariance matrix of the random effects, and $\mathbf{L}$, the sparse Cholesky factor, in model `fm9`

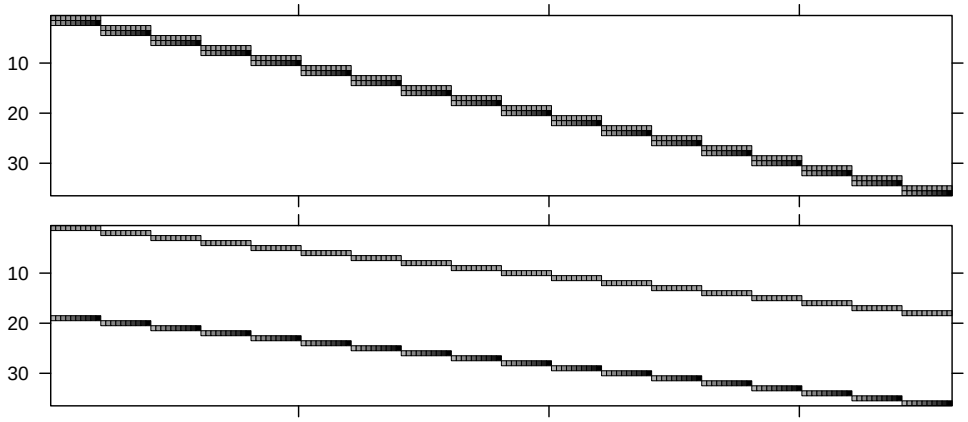


Fig. 4.4 Images of $\mathbf{Z}^\mathsf{T}$ for models `fm8` (upper panel) and `fm9` (lower panel)

4.2.3 Generating Z and Λ From Random-effects Terms

Let us consider these columns in more detail, starting with the columns of $\mathbf{Z}$ for model `fm9`. The first 18 columns (rows in the bottom panel of Fig. 4.4) are the indicator columns for the `Subject` factor, as we would expect from the simple, scalar random-effects term `(1|Subject)`. The pattern of zeros and non-zeros in the second group of 18 columns is determined by the indicators of the grouping factor, `Subject`, and the values of the non-zeros are determined by the `Days` factor. In other words, these columns are formed by the *interaction* of the numeric covariate, `Days`, and the categorical covariate, `Subject`.

The non-zero values in the model matrix $\mathbf{Z}$ for model `fm8` are the same as those for model `fm8` but the columns are in a different order. Pairs of columns associated with the same level of the grouping factor are adjacent. One way to think of the process of generating these columns is to extend the idea of an

interaction between a single covariate and the grouping factor to generating an "interaction" of a model matrix and the levels of the grouping factor. In other words, we begin with the two columns of the model matrix for the expression `1 + Days` and the 18 columns of indicators for the `Subject` factor. The result will have 36 columns that are considered as 18 adjacent pairs. The values within each of these pairs of columns are the values of the `1 + Days` columns, when the indicator is 1, otherwise they are zero.

We can now describe the general process of creating the model matrix, $\mathbf{Z}$, and the relative covariance factor, Λ from the random-effects terms in the model formula. Each random-effects term is of the form `(expr|fac)`. The expression `expr` is evaluated as a linear model formula, producing a model matrix with s columns. The expression `fac` is evaluated as a factor. Let k be the number of levels in this factor, after eliminating unused levels, if any. The ith term generates $s_i k_i$ columns in the model matrix, $\mathbf{Z}$, and a diagonal block of size $s_i k_i$ in the relative covariance factor, Λ. The $s_i k_i$ columns in $\mathbf{Z}$ have the pattern of the interaction of the s_i columns from the ith `expr` with the k indicator columns for the factor `fac`. The diagonal block in Λ is itself block diagonal, consisting of k_i blocks, each a lower triangular matrix of size s_i. In fact, these inner blocks are repetitions of the same lower triangular $s_i \times s_i$ matrix. The i term contributes $s_i(s_i+1)/2$ elements to the variance-component parameter, θ, and these are the elements in the lower triangle of this $s_i \times s_i$ template matrix.

Note that when counting the columns in a model matrix we must take into account the implicit intercept term. For example, we could write the formula for model `fm8` as

```
Reaction ~ Days + (Days | Subject)
```

realizing that the linear model expression, `Days`, actually generates two columns because of the implicit intercept.

Whether or not to include an explicit intercept term in a model formula is a matter of personal taste. Many people prefer to write the intercept explicitly in the formula so as to emphasize the relationship between terms in the formula and coefficients or random effects in the model. Others omit these implicit terms so as to economize on the amount of typing required. Either approach can be used. The important point to remember is that the intercept must be explicitly suppressed when you don't want it in a term.

Also, the intercept term must be explicit when it is the only term in the expression. That is, a simple, scalar random-effects term must be written as `(1|fac)` because a term like `(|fac)` is not syntactically correct. However, we can omit the intercept from the fixed-effects part of the model formula if we have any random-effects terms. That is, we could write the formula for model `fm1` in Chap. 1 as

```
Yield ~ (1 | Batch)
```

or even

```
Yield ~ 1 | Batch
```

although omitting the parentheses around a random-effects term is risky. Because of operator precedence, the vertical bar operator, `|`, takes essentially everything in the expression to the left of it as its first operand. It is advisable always to enclose such terms in parentheses so the scope of the operands to the `|` operator is clearly defined.

4.2.4 Comparing Models `fm9` *and* `fm8`

Returning to models `fm8` and `fm9` for the `sleepstudy` data, it is easy to see that these are nested models because `fm8` is reduced to `fm9` by constraining the within-group correlation of random effects to be zero (which is equivalent to constraining the element below the diagonal in the 2×2 lower triangular blocks of Λ in Fig. 4.2 to be zero).

We can use a likelihood ratio test to compare these fitted models.

```
> anova(fm9, fm8)
```

```
Data: sleepstudy
Models:
fm9: Reaction ~ 1 + Days + (1 | Subject) + (0 + Days | Subject)
fm8: Reaction ~ 1 + Days + (1 + Days | Subject)
    Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
fm9  5 1762.0 1778.0 -876.00
fm8  6 1763.9 1783.1 -875.97 0.0639      1     0.8004
```

The value of the χ^2 statistic, 0.0639, is very small, corresponding to a p-value of 0.80 and indicating that the extra parameter in model `fm8` relative to `fm9` does not produce a significantly better fit. By the principal of parsimony we prefer the reduced model, `fm9`.

This conclusion is consistent with the visual impression provided by Fig. 4.1. There does not appear to be a strong relationship between a subject's initial reaction time and the extent to which his or her reaction time is affected by sleep deprivation.

In this likelihood ratio test the value of the parameter being tested, a correlation of zero, is not on the boundary of the parameter space. We can be confident that the p-value from the LRT adequately reflects the underlying situation.

(**Note**: It is possible to extend profiling to the correlation parameters and we will do so but that has not been done yet.)

4.3 Assessing the Precision of the Parameter Estimates

Plots of the profile ζ for the parameters in model `fm9` (Fig. 4.5)show that

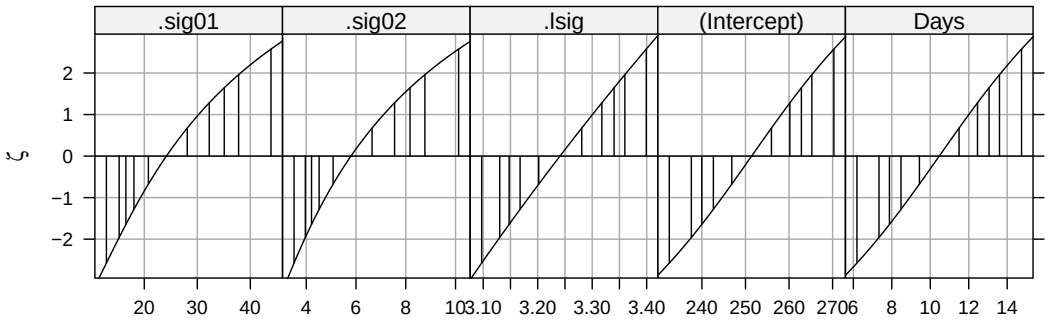


Fig. 4.5 Profile zeta plot for each of the parameters in model `fm9`. The vertical lines are the endpoints of 50%, 80%, 90%, 95% and 99% profile-based confidence intervals for each parameter.

confidence intervals on σ_1 and σ_2 will be slightly skewed; those for $\log(\sigma)$ will be symmetric and well-approximated by methods based on quantiles of the standard normal distribution and those for the fixed-effects parameters, β_1 and β_2 will be symmetric and slightly over-dispersed relative to the standard normal. For example, the 95% profile-based confidence intervals are

```
> confint(pr9)
```

```
                2.5 %     97.5 %
.sig01       15.258637  37.786532
.sig02        3.964074   8.769159
.lsig         3.130287   3.359945
(Intercept) 237.572148 265.238062
Days          7.334067  13.600505
```

The profile pairs plot (Fig. 4.6) shows, for the most part, the usual patterns. First, consider the panels below the diagonal, which are on the (ζ_i, ζ_j) scales. The ζ pairs for $\log(\sigma)$ and β_0, in the $(4,3)$ panel, and for $\log(\sigma)$ and β_1, in the $(5,3)$ panel, show the ideal pattern. The profile traces are straight and orthogonal, producing interpolated contours on the ζ scale that are concentric circles centered at the origin. When mapped back to the scales of $\log(\sigma)$ and β_0 or β_1, in panels $(3,4)$ and $(3,5)$, these circles become slightly distorted, but this is only due to the moderate nonlinearity in the profile ζ plots for these parameters.

Examining the profile traces on the ζ scale for $\log(\sigma)$ versus σ_1, the $(3,1)$ panel, or versus σ_2, the $(3,2)$ panel, and for σ_1 versus σ_2, the $(2,1)$ panel, we see that close to the estimate the traces are orthogonal but as one variance component becomes small there is usually an increase in the others. In some sense the total variability in the response will be partitioned across the contribution of the fixed effects and the variance components. In each of

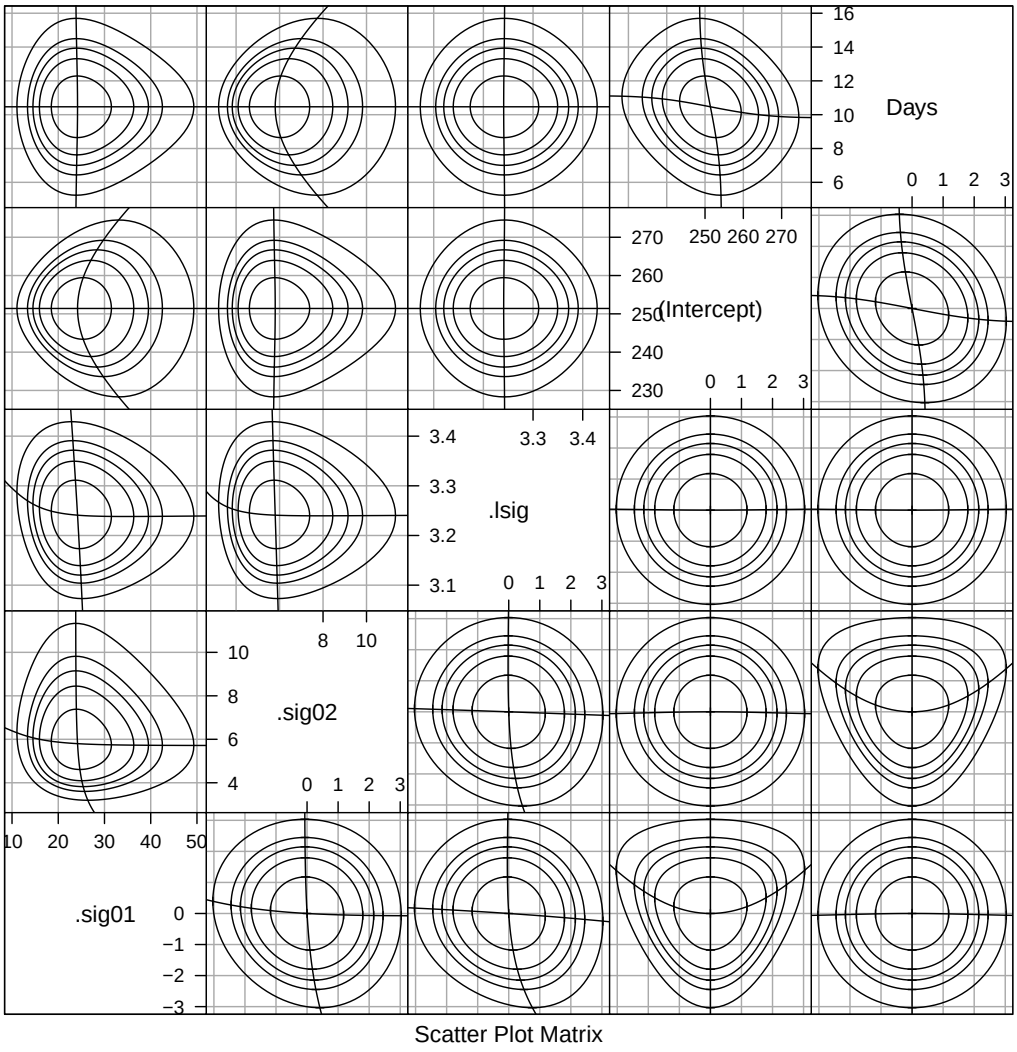


Fig. 4.6 Profile pairs plot for the parameters in model `fm9`. The contour lines correspond to marginal 50%, 80%, 90%, 95% and 99% confidence regions based on the likelihood ratio. Panels below the diagonal represent the (ζ_i, ζ_j) parameters; those above the diagonal represent the original parameters.

these panels the fixed-effects parameters are at their optimal values, conditional on the values of the variance components, and the variance components must compensate for each other. If one is made smaller then the others must become larger to compensate.

The patterns in the $(4,1)$ panel (σ_1 versus β_0, on the ζ scale) and the $(5,2)$ panel (σ_2 versus β_1, on the ζ scale) are what we have come to expect. As the fixed-effects parameter is moved from its estimate, the standard deviation of the corresponding random effect increases to compensate. The $(5,1)$ and $(4,2)$ panels show that changing the value of a fixed effect doesn't change the estimate of the standard deviation of the random effects corresponding to

the other fixed effect, which makes sense although the perfect orthogonality shown here will probably not be exhibited in models fit to unbalanced data.

In some ways the most interesting panels are those for the pair of fixed-effects parameters: (5,4) on the ζ scale and (4,5) on the original scale. The traces are not orthogonal. In fact the slopes of the traces at the origin of the (5,4) (ζ scale) panel are the correlation of the fixed-effects estimators (-0.194 for this model) and its inverse. However, as we move away from the origin on one of the traces in the (5,4) panel it curves back toward the horizontal axis (for the horizontal trace) or the vertical axis (for the vertical trace). In the ζ scale the individual contours are still concentric ellipses but their eccentricity changes from contour to contour. The innermost contours have greater eccentricity than the outermost contours. That is, the outermost contours are more like circles than are the innermost contours.

In a fixed-effects model the shapes of projections of deviance contours onto pairs of fixed-effects parameters are consistent. In a fixed-effects model the profile traces in the original scale will always be straight lines. For mixed models these traces can fail to be linear, as we see here, contradicting the widely-held belief that inferences for the fixed-effects parameters in linear mixed models, based on T or F distributions with suitably adjusted degrees of freedom, will be completely accurate. The actual patterns of deviance contours are more complex than that.

4.4 Examining the Random Effects and Predictions

The result of applying `ranef` to fitted linear mixed model is a list of data frames. The components of the list correspond to the grouping factors in the random-effects terms, not to the terms themselves. Model `fm9` is the first model we have fit with more than one term for the same grouping factor where we can see the combination of random effects from more than one term.

```
> str(rr1 <- ranef(fm9))
```

```
List of 1
 $ Subject:'data.frame':        18 obs. of  2 variables:
  ..$ (Intercept): num [1:18] 1.85 -40.02 -38.72 23.9 22.4 ...
  ..$ Days       : num [1:18] 9.24 -8.62 -5.43 -4.86 -3.1 ...
 - attr(*, "class")= chr "ranef.mer"
```

The `plot` method for `"ranef.mer"` objects produces one plot for each grouping factor. For scalar random effects the plot is a normal probability plot. For two-dimensional random effects, including the case of two scalar terms for the same grouping factor, as in this model, the plot is a scatterplot. For three or more random effects per level level of the grouping factor, the plot is a scatterplot matrix. The left hand panel in Fig. 4.7 was created with `plot(ranef(fm9))`.

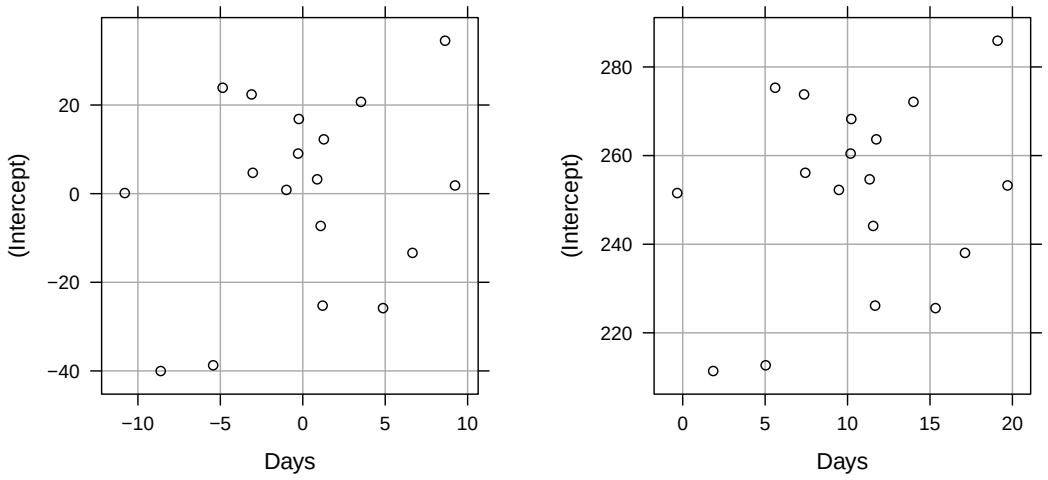


Fig. 4.7 Plot of the conditional modes of the random effects for model `fm9` (left panel) and the corresponding subject-specific coefficients (right panel)

The `coef` method for a fitted `lmer` model combines the fixed-effects estimates and the conditional modes of the random effects, whenever the column names of the random effects correspond to the names of coefficients. For model `fm9` the fixed-effects coefficients are `(Intercept)` and `Days` and the columns of the random effects match these names. Thus we can calculate some kind of per-subject "estimates" of the slope and intercept and plot them, as in the right hand panel of Fig. 4.7. By comparing the two panels in Fig. 4.7 we can see that the result of the `coef` method is simply the conditional modes of the random effects shifted by the coefficient estimates.

It is not entirely clear how we should interpret these values. They are a combination of parameter estimates with the modal values of random variables and, as such, are in a type of "no man's land" in the probability model. (In the Bayesian approach [Box and Tiao, 1973] to inference, however, both the parameters and the random effects are random variables and the interpretation of these values is straightforward.) Despite the difficulties of interpretation in the probability model, these values are of interest because they determine the fitted response for each subject.

Because responses for each individual are recorded on each of ten days we can determine the within-subject estimates of the slope and intercept (that is, the slope and intercept of each of the lines in Fig. 4.1). In Fig. 4.8 we compare the within-subject least squares estimates to the per-subject slope and intercept calculated from model `fm9`. We see that, in general, the per-subject slopes and intercepts from the mixed-effects model are closer to the population estimates than are the within-subject least squares estimates. This pattern is sometimes described as a *shrinkage* of coefficients toward the population values.

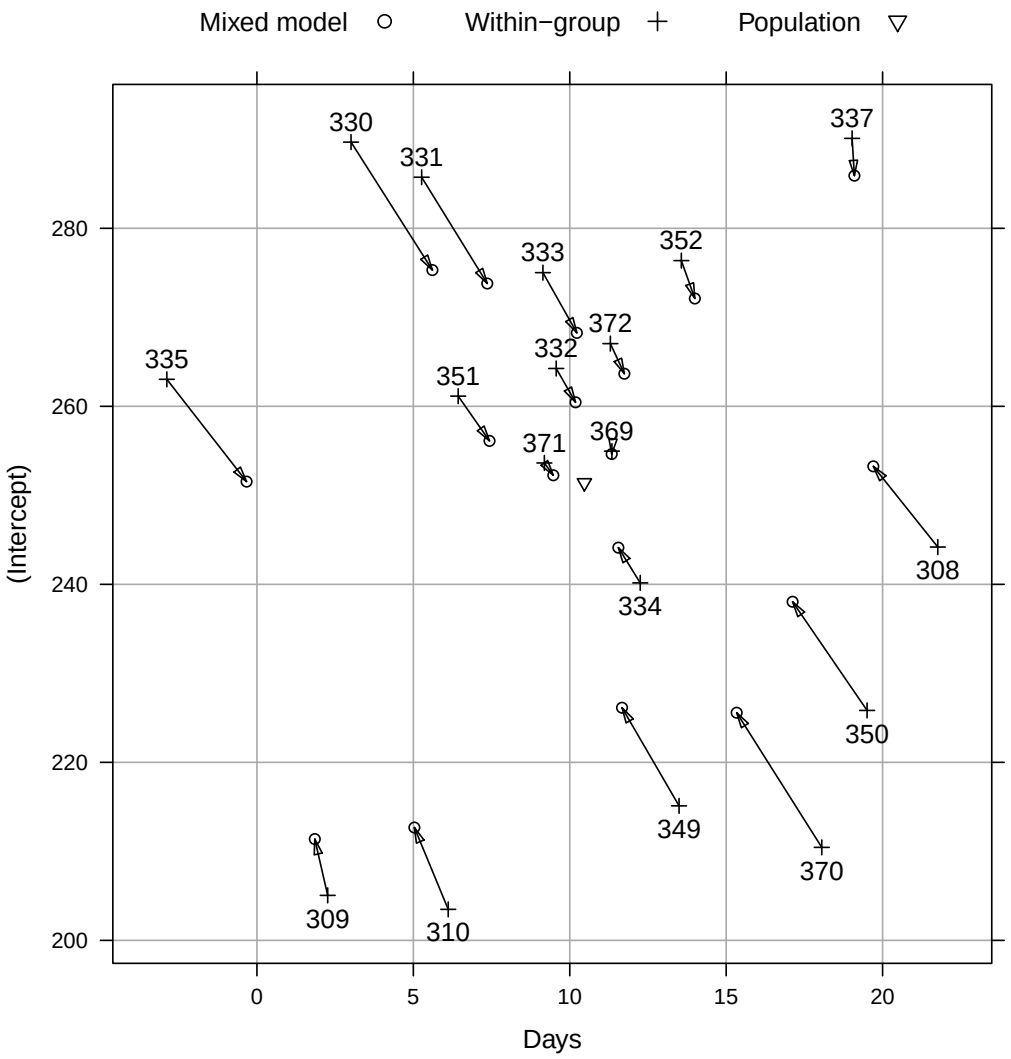


Fig. 4.8 Comparison of the within-subject estimates of the intercept and slope for each subject and the conditional modes of the per-subject intercept and slope. Each pair of points joined by an arrow are the within-subject and conditional mode estimates for the same subject. The arrow points from the within-subject estimate to the conditional mode for the mixed-effects model. The subject identifier number is at the head of each arrow.

The term "shrinkage" may have negative connotations. John Tukey preferred to refer to the process as the estimates for individual subjects "borrowing strength" from each other. This is a fundamental difference in the models underlying mixed-effects models versus strictly fixed-effects models. In a mixed-effects model we assume that the levels of a grouping factor are a selection from a population and, as a result, can be expected to share characteristics to some degree. Consequently, the predictions from a mixed-effects model are attenuated relative to those from strictly fixed-effects models.

The predictions from model `fm9` and from the within-subject least squares fits for each subject are shown in Fig. 4.9. In may seem that the shrinkage

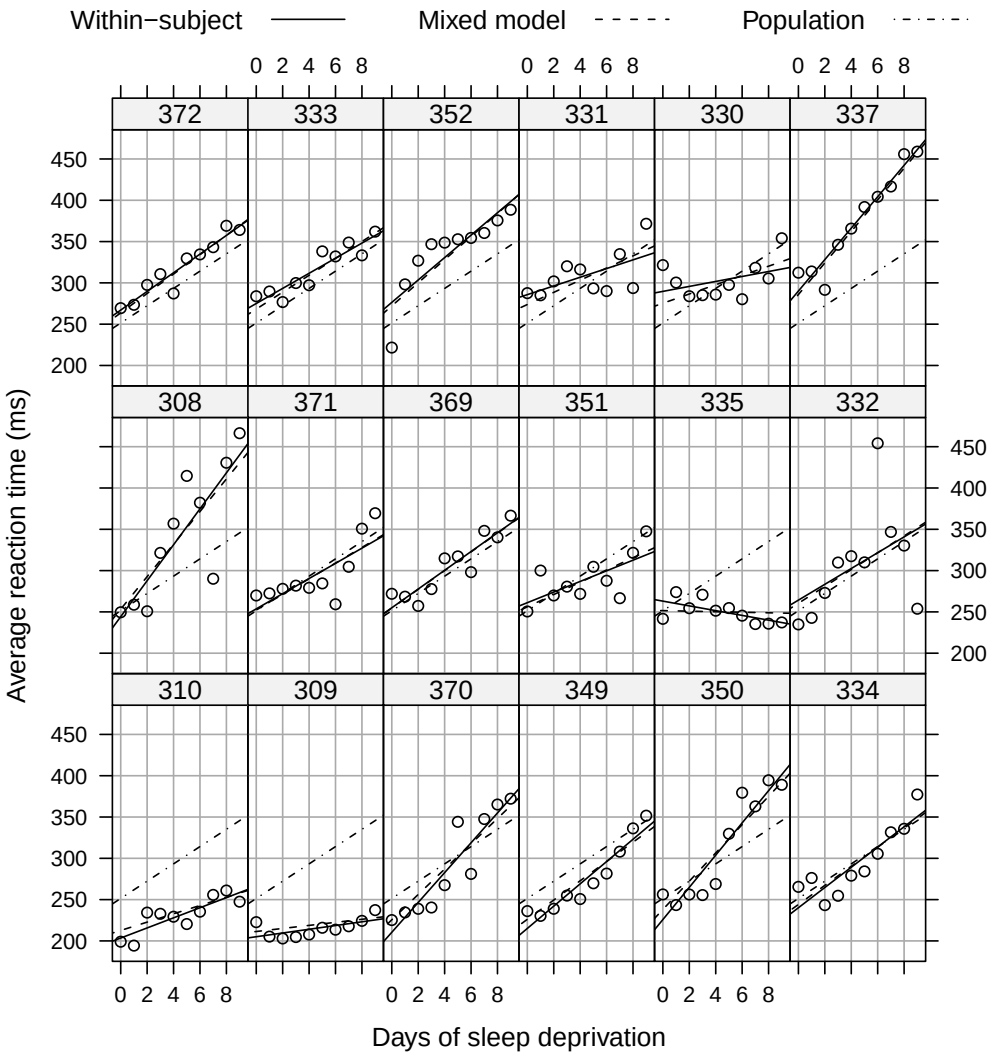


Fig. 4.9 Comparison of the predictions from the within-subject fits with those from the conditional modes of the subject-specific parameters in the mixed-effects model.

from the per-subject estimates toward the population estimates depends only on how far the per-subject estimates (solid lines) are from the population estimates (dot-dashed lines). However, careful examination of this figure shows that there is more at work here than a simple shrinkage toward the population estimates proportional to the distance of the per-subject estimates from the population estimates.

It is true that the mixed model estimates for a particular subject are "between" the within-subject estimates and the population estimates, in the sense that the arrows in Fig. 4.8 all point somewhat in the direction of the population estimate. However, the extent of the attenuation of the within-subject estimates toward the population estimates is not simply related to the distance between those two sets of estimates. Consider the two panels, labeled 330 and 337, at the top right of Fig. 4.9. The within-subject estimates for 337

are quite unlike the population estimates but the mixed-model estimates are very close to these within-subject estimates. That is, the solid line and the dashed line in that panel are nearly coincident and both are a considerable distance from the dot-dashed line. For subject 330, however, the dashed line is more-or-less an average of the solid line and the dot-dashed line, even though the solid and dot-dashed lines are not nearly as far apart as they are for subject 337.

The difference between these two cases is that the within-subject estimates for 337 are very well determined. Even though this subject had an unusually large intercept and slope, the overall pattern of the responses is very close to a straight line. In contrast, the overall pattern for 330 is not close to a straight line so the within-subject coefficients are not well determined. The multiple R^2 for the solid line in the 337 panel is 93.3% but in the 330 panel it is only 15.8%. The mixed model can pull the predictions in the 330 panel, where the data are quite noisy, closer to the population line without increasing the residual sum of squares substantially. When the within-subject coefficients are precisely estimated, as in the 337 panel, very little shrinkage takes place.

We see from Fig. 4.9 that the mixed-effects model smooths out the between-subject differences in the predictions by bringing them closer to a common set of predictions, but not at the expense of dramatically increasing the sum of squared residuals. That is, the predictions are determined so as to balance fidelity to the data, measured by the residual sum of squares, with simplicity of the model. The simplest model would use the same prediction in each panel (the dot-dashed line) and the most complex model, based on linear relationships in each panel, would correspond to the solid lines. The dashed lines are between these two extremes. We will return to this view of the predictions from mixed models balancing complexity versus fidelity in Sec. 5.3, where we make the mathematical nature of this balance explicit.

We should also examine the prediction intervals on the random effects (Fig. 4.10) where we see that many prediction intervals overlap zero but there are several that do not.. In this plot the subjects are ordered from bottom to top according to increasing conditional mode of the random effect for `(Intercept)`. The resulting pattern in the conditional modes of the random effect for `Days` reinforces our conclusion that the model `fm9`, which does not allow for correlation of the random effects for `(Intercept)` and `Days`, is suitable.

4.5 Chapter Summary

Problems

4.1. Check the structure of documentation, structure and a summary of the `Orthodont` data set from the `MEMSS` package.

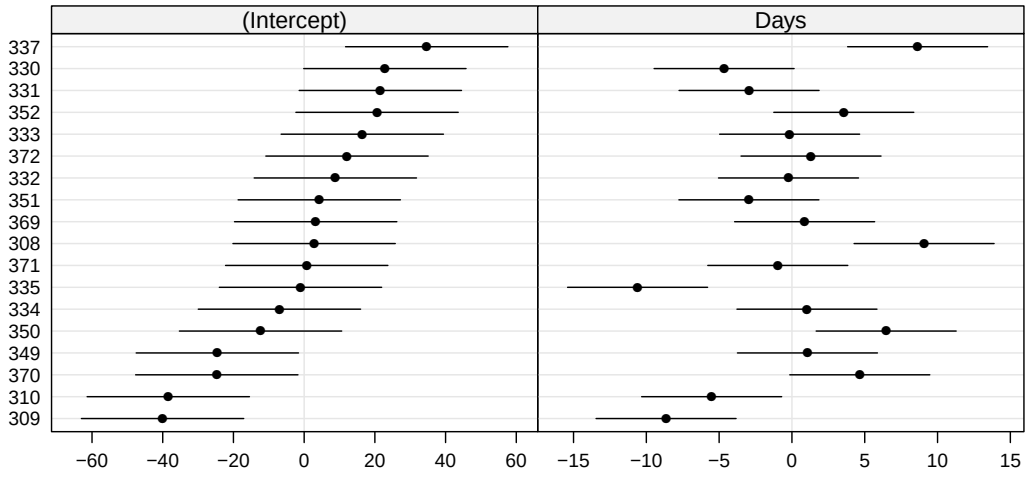
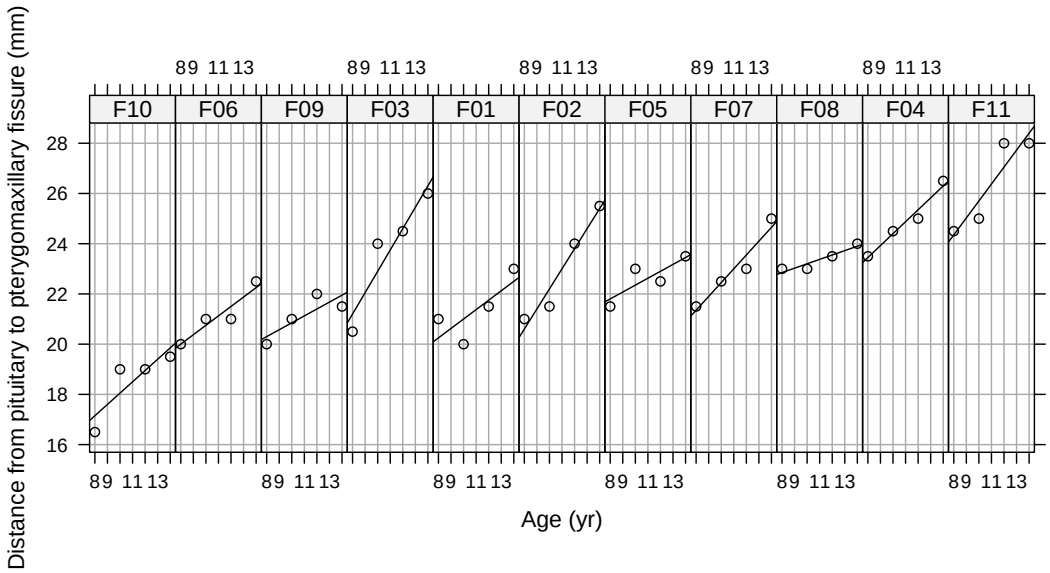


Fig. 4.10 Prediction intervals on the random effects for model `fm9`.

(a) Create an `xyplot` of the `distance` versus `age` by `Subject` for the female subjects only. You can use the optional argument `subset = Sex == "Female"` in the call to `xyplot` to achieve this. Use the optional argument `type = c("g","p","r")` to add reference lines to each panel.
(b) Enhance the plot by choosing an aspect ratio for which the typical slope of the reference line is around 45^o. You can set it manually (something like `aspect = 4`) or with an automatic specification (`aspect = "xy"`). Change the layout so the panels form one row (`layout = c(11,1)`).
(c) Order the panels according to increasing response at age 8. This is achieved with the optional argument `index.cond` which is a function of arguments `x` and `y`. In this case you could use `index.cond = function(x,y) y[x == 8]`. Add meaningful axis labels. Your final plot should be like



(d) Fit a linear mixed model to the data for the females only with random effects for the intercept and for the slope by subject, allowing for correlation of these random effects within subject. Relate the fixed effects and the random effects' variances and covariances to the variability shown in the figure.
(e) Produce a "caterpillar plot" of the random effects for intercept and slope. Does the plot indicate correlated random effects?
(f) Consider what the Intercept coefficient and random effects represents. What will happen if you center the ages by subtracting 8 (the baseline year) or 11 (the middle of the age range)?
(g) Repeat for the data from the male subjects.

4.2.
Fit a model to both the female and the male subjects in the `Orthodont` data set, allowing for differences by sex in the fixed-effects for intercept (probably with respect to the centered age range) and slope.

Chapter 5
Computational Methods for Mixed Models

In this chapter we describe some of the details of the computational methods for fitting linear mixed models, as implemented in the `lme4` package, and the theoretical development behind these methods. We also provide the basis for later generalizations to models for non-Gaussian responses and to models in which the relationship between the conditional mean, μ, and the linear predictor, $\gamma = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} = \mathbf{Z}\Lambda_\theta\mathbf{u} + \mathbf{X}\beta$, is a nonlinear relationship.

This material is directed at those readers who wish to follow the theory and methodology of linear mixed models and how both can be extended to other forms of mixed models. Readers who are less interested in the "how" and the "why" of fitting mixed models than in the results themselves should not feel obligated to master these details.

We begin by reviewing the definition of linear mixed-effects models and some of the basics of the computational methods, as given in Sect. 1.1.

5.1 Definitions and Basic Results

As described in Sect. 1.1, a linear mixed-effects model is based on two vector-valued random variables: the q-dimensional vector of random effects, $\mathscr{B}$, and the n-dimensional response vector, $\mathscr{Y}$. Equation (1.1) defines the unconditional distribution of $\mathscr{B}$ and the conditional distribution of $\mathscr{Y}$, given $\mathscr{B} = \mathbf{b}$, as multivariate Gaussian distributions of the form

$$\begin{aligned}(\mathscr{Y}|\mathscr{B} = \mathbf{b}) &\sim \mathscr{N}(\mathbf{X}\beta + \mathbf{Z}\mathbf{b}, \sigma^2\mathbf{I})\\ \mathscr{B} &\sim \mathscr{N}(\mathbf{0}, \Sigma_\theta).\end{aligned}$$

The $q \times q$, symmetric, variance-covariance matrix, $\mathrm{Var}(\mathscr{B}) = \Sigma_\theta$, depends on the *variance-component parameter vector*, θ, and is *positive semidefinite*, which means that

$$\mathbf{b}^\mathsf{T}\Sigma_\theta\mathbf{b} \geq 0, \quad \forall \mathbf{b} \neq \mathbf{0}. \tag{5.1}$$

(The symbol $\forall$ denotes "for all".) The fact that Σ_θ is positive semidefinite does not guarantee that Σ_θ^{-1} exists. We would need a stronger property, $\mathbf{b}^\mathsf{T}\Sigma_\theta\mathbf{b} > 0, \forall \mathbf{b} \neq \mathbf{0}$, called positive definiteness, to ensure that Σ_θ^{-1} exists.

Many computational formulas for linear mixed models are written in terms of Σ_θ^{-1}. Such formulas will become unstable as Σ_θ approaches singularity. And it can do so. It is a fact that singular (i.e. non-invertible) Σ_θ can and do occur in practice, as we have seen in some of the examples in earlier chapters. Moreover, during the course of the numerical optimization by which the parameter estimates are determined, it is frequently the case that the deviance or the REML criterion will need to be evaluated at values of θ that produce a singular Σ_θ. Because of this we will take care to use computational methods that can be applied even when Σ_θ is singular and are stable as Σ_θ approaches singularity.

As defined in (1.2) a relative covariance factor, Λ_θ, is any matrix that satisfies

$$\Sigma_\theta = \sigma^2\Lambda_\theta\Lambda_\theta^\mathsf{T}.$$

According to this definition, Σ depends on both σ and θ and we should write it as $\Sigma_{\sigma,\theta}$. However, we will blur that distinction and continue to write $\mathrm{Var}(\mathscr{B}) = \Sigma_\theta$. Another technicality is that the *common scale parameter*, σ, can, in theory, be zero. We will show that in practice the only way for its estimate, $\widehat{\sigma}$, to be zero is for the fitted values from the fixed-effects only, $\mathbf{X}\widehat{\beta}$, to be exactly equal to the observed data. This occurs only with data that have been (incorrectly) simulated without error. In practice we can safely assume that $\sigma > 0$. However, Λ_θ, like Σ_θ, can be singular.

Our computational methods are based on Λ_θ and do not require evaluation of Σ_θ. In fact, Σ_θ is explicitly evaluated only at the converged parameter estimates.

The spherical random effects, $\mathscr{U} \sim \mathscr{N}(\mathbf{0}, \sigma^2\mathbf{I}_q)$, determine $\mathscr{B}$ as

$$\mathscr{B} = \Lambda_\theta\mathscr{U}. \tag{5.2}$$

Although it may seem more intuitive to write $\mathscr{U}$ as a linear transformation of $\mathscr{B}$, we cannot do that when Λ_θ is singular, which is why (5.2) is in the form shown.

We can easily verify that (5.2) provides the desired distribution for $\mathscr{B}$. As a linear transformation of a multivariate Gaussian random variable, $\mathscr{B}$ will also be multivariate Gaussian. Its mean and variance-covariance matrix are straightforward to evaluate,

$$\mathrm{E}[\mathscr{B}] = \Lambda_\theta\mathrm{E}[\mathscr{U}] = \Lambda_\theta\mathbf{0} = \mathbf{0} \tag{5.3}$$

and

$$\begin{aligned}\mathrm{Var}(\mathcal{B}) &= \mathrm{E}\left[(\mathcal{B}-\mathrm{E}[\mathcal{B}])(\mathcal{B}-\mathrm{E}[\mathcal{B}])^\mathsf{T}\right] = \mathrm{E}\left[\mathcal{B}\mathcal{B}^\mathsf{T}\right] \\ &= \mathrm{E}\left[\Lambda_\theta \mathcal{U}\mathcal{U}^\mathsf{T}\Lambda_\theta^\mathsf{T}\right] = \Lambda_\theta\,\mathrm{E}[\mathcal{U}\mathcal{U}^\mathsf{T}]\Lambda_\theta^\mathsf{T} = \Lambda_\theta\,\mathrm{Var}(\mathcal{U})\Lambda_\theta^\mathsf{T} \\ &= \Lambda_\theta\,\sigma^2\mathbf{I}_q\Lambda_\theta^\mathsf{T} = \sigma^2\Lambda_\theta\Lambda_\theta^\mathsf{T} = \Sigma_\theta\end{aligned} \tag{5.4}$$

and have the desired form.

Just as we concentrate on how θ determines Λ_θ, not Σ_θ, we will concentrate on properties of $\mathcal{U}$ rather than $\mathcal{B}$. In particular, we now define the model according to the distributions

$$\begin{aligned}(\mathcal{Y}|\mathcal{U}=\mathbf{u}) &\sim \mathcal{N}(\mathbf{Z}\Lambda_\theta\mathbf{u}+\mathbf{X}\beta,\sigma^2\mathbf{I}_n) \\ \mathcal{U} &\sim \mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_q).\end{aligned} \tag{5.5}$$

To allow for extensions to other types of mixed models we distinguish between the *linear predictor*

$$\gamma = \mathbf{Z}\Lambda_\theta\mathbf{u}+\mathbf{X}\beta \tag{5.6}$$

and the *conditional mean* of $\mathcal{Y}$, given $\mathcal{U}=\mathbf{u}$, which is

$$\mu = \mathrm{E}[\mathcal{Y}|\mathcal{U}=\mathbf{u}]. \tag{5.7}$$

For a linear mixed model $\mu=\gamma$. In other forms of mixed models the conditional mean, μ, can be a nonlinear function of the linear predictor, γ. For some models the dimension of γ is a multiple of n, the dimension of μ and $\mathbf{y}$, but for a linear mixed model the dimension of γ must be n. Hence, the model matrix $\mathbf{Z}$ must be $n\times q$ and $\mathbf{X}$ must be $n\times p$.

5.2 The Conditional Distribution $(\mathcal{U}|\mathcal{Y}=\mathbf{y})$

In this chapter it will help to be able to distinguish between the observed response vector and an arbitrary value of $\mathcal{Y}$. For this chapter only we will write the observed data vector as $\mathbf{y}_{\mathrm{obs}}$, with the understanding that $\mathbf{y}$ without the subscript will refer to an arbitrary value of the random variable $\mathcal{Y}$.

The likelihood of the parameters, θ, β, and σ, given the observed data, $\mathbf{y}_{\mathrm{obs}}$, is the probability density of $\mathcal{Y}$, evaluated at $\mathbf{y}_{\mathrm{obs}}$. Although the numerical values of the probability density and the likelihood are identical, the interpretations of these functions are different. In the density we consider the parameters to be fixed and the value of $\mathbf{y}$ as varying. In the likelihood we consider $\mathbf{y}$ to be fixed at $\mathbf{y}_{\mathrm{obs}}$ and the parameters, θ, β and σ, as varying.

The natural approach for evaluating the likelihood is to determine the marginal distribution of $\mathcal{Y}$, which in this case amounts to determining the marginal density of $\mathcal{Y}$, and evaluate that density at $\mathbf{y}_{\mathrm{obs}}$. To follow this course

we would first determine the joint density of $\mathscr{U}$ and $\mathscr{Y}$, written $f_{\mathscr{U},\mathscr{Y}}(\mathbf{u},\mathbf{y})$, then integrate this density with respect $\mathbf{u}$ to create the marginal density, $f_{\mathscr{Y}}(\mathbf{y})$, and finally evaluate this marginal density at $\mathbf{y}_{\text{obs}}$.

To allow for later generalizations we will change the order of these steps slightly. We evaluate the joint density function, $f_{\mathscr{U},\mathscr{Y}}(\mathbf{u},\mathbf{y})$, at $\mathbf{y}_{\text{obs}}$, producing the *unnormalized conditional density*, $h(\mathbf{u})$. We say that h is "unnormalized" because the conditional density is a multiple of h

$$f_{\mathscr{U}|\mathscr{Y}}(\mathbf{u}|\mathbf{y}_{\text{obs}}) = \frac{h(\mathbf{u})}{\int_{\mathbb{R}^q} h(\mathbf{u})\,d\mathbf{u}}. \tag{5.8}$$

In some theoretical developments the normalizing constant, which is the integral in the denominator of an expression like (5.8), is not of interest. Here it is of interest because the normalizing constant is exactly the likelihood that we wish to evaluate,

$$L(\theta,\beta,\sigma|\mathbf{y}_{\text{obs}}) = \int_{\mathbb{R}^q} h(\mathbf{u})\,d\mathbf{u}. \tag{5.9}$$

For a linear mixed model, where all the distributions of interest are multivariate Gaussian and the conditional mean, μ, is a linear function of both $\mathbf{u}$ and β, the distinction between evaluating the joint density at $\mathbf{y}_{\text{obs}}$ to produce $h(\mathbf{u})$ then integrating with respect to $\mathbf{u}$, as opposed to first integrating the joint density then evaluating at $\mathbf{y}_{\text{obs}}$ is not terribly important. For other mixed models this distinction can be important. In particular, generalized linear mixed models, described in Sect. **??**, are often used to model a discrete response, such as a binary response or a count, leading to a joint distribution for $\mathscr{Y}$ and $\mathscr{U}$ that is discrete with respect to one variable, $\mathbf{y}$, and continuous with respect to the other, $\mathbf{u}$. In such cases there isn't a joint density for $\mathscr{Y}$ and $\mathscr{U}$. The necessary distribution theory for general $\mathbf{y}$ and $\mathbf{u}$ is well-defined but somewhat awkward to describe. It is much easier to realize that we are only interested in the observed response vector, $\mathbf{y}_{\text{obs}}$, not some arbitrary value of $\mathbf{y}$, so we can concentrate on the conditional distribution of $\mathscr{U}$ given $\mathscr{Y} = \mathbf{y}_{\text{obs}}$. For all the mixed models we will consider, the conditional distribution, $(\mathscr{U}|\mathscr{Y} = \mathbf{y}_{\text{obs}})$, is continuous and both the conditional density, $f_{\mathscr{U}|\mathscr{Y}}(\mathbf{u}|\mathbf{y}_{\text{obs}})$, and its unnormalized form, $h(\mathbf{u})$, are well-defined.

5.3 Integrating $h(\mathbf{u})$ in the Linear Mixed Model

The integral defining the likelihood in (5.9) has a closed form in the case of a linear mixed model but not for some of the more general forms of mixed models. To motivate methods for approximating the likelihood in more general situations, we describe in some detail how the integral can be evaluated using the sparse Cholesky factor, $\mathbf{L}_\theta$, and the conditional mode,

$$\tilde{\mathbf{u}} = \arg\max_{\mathbf{u}} f_{\mathscr{U}|\mathscr{Y}}(\mathbf{u}|\mathbf{y}_{\text{obs}}) = \arg\max_{\mathbf{u}} h(\mathbf{u}) = \arg\max_{\mathbf{u}} f_{\mathscr{Y}|\mathscr{U}}(\mathbf{y}_{\text{obs}}|\mathbf{u})\, f_{\mathscr{U}}(\mathbf{u}) \quad (5.10)$$

The notation $\arg\max_{\mathbf{u}}$ means that $\tilde{\mathbf{u}}$ is the value of $\mathbf{u}$ that maximizes the expression that follows.

In general, the *mode* of a continuous distribution is the value of the random variable that maximizes the density. The value $\tilde{\mathbf{u}}$ is called the conditional mode of $\mathbf{u}$, given $\mathscr{Y} = \mathbf{y}_{\text{obs}}$, because $\tilde{\mathbf{u}}$ maximizes the conditional density of $\mathscr{U}$ given $\mathscr{Y} = \mathbf{y}_{\text{obs}}$. The location of the maximum can be determined by maximizing the unnormalized conditional density because $h(\mathbf{u})$ is just a constant multiple of $f_{\mathscr{U}|\mathscr{Y}}(\mathbf{u}|\mathbf{y}_{\text{obs}})$. The last part of (5.10) is simply a re-expression of $h(\mathbf{u})$ as the product of $f_{\mathscr{Y}|\mathscr{U}}(\mathbf{y}_{\text{obs}}|\mathbf{u})$ and $f_{\mathscr{U}}(\mathbf{u})$. For a linear mixed model these densities are

$$f_{\mathscr{Y}|\mathscr{U}}(\mathbf{y}|\mathbf{u}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta \mathbf{u}\|^2}{2\sigma^2}\right) \quad (5.11)$$

$$f_{\mathscr{U}}(\mathbf{u}) = \frac{1}{(2\pi\sigma^2)^{q/2}} \exp\left(-\frac{\|\mathbf{u}\|^2}{2\sigma^2}\right) \quad (5.12)$$

with product

$$h(\mathbf{u}) = \frac{1}{(2\pi\sigma^2)^{(n+q)/2}} \exp\left(-\frac{\|\mathbf{y}_{\text{obs}} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta \mathbf{u}\|^2 + \|\mathbf{u}\|^2}{2\sigma^2}\right). \quad (5.13)$$

On the deviance scale we have

$$-2\log(h(\mathbf{u})) = (n+q)\log(2\pi\sigma^2) + \frac{\|\mathbf{y}_{\text{obs}} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta \mathbf{u}\|^2 + \|\mathbf{u}\|^2}{\sigma^2}. \quad (5.14)$$

Because (5.14) describes the negative log density, $\tilde{\mathbf{u}}$ will be the value of $\mathbf{u}$ that minimizes the expression on the right of (5.14).

The only part of the right hand side of (5.14) that depends on $\mathbf{u}$ is the numerator of the second term. Thus

$$\tilde{\mathbf{u}} = \arg\min_{\mathbf{u}} \|\mathbf{y}_{\text{obs}} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta \mathbf{u}\|^2 + \|\mathbf{u}\|^2. \quad (5.15)$$

The expression to be minimized, called the *objective function*, is described as a *penalized residual sum of squares* (PRSS) and the minimizer, $\tilde{\mathbf{u}}$, is called the *penalized least squares* (PLS) solution. They are given these names because the first term in the objective, $\|\mathbf{y}_{\text{obs}} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta \mathbf{u}\|^2$, is a sum of squared residuals, and the second term, $\|\mathbf{u}\|^2$, is a penalty on the length, $\|\mathbf{u}\|$, of $\mathbf{u}$. Larger values of $\mathbf{u}$ (in the sense of greater lengths as vectors) incur a higher penalty.

The PRSS criterion determining the conditional mode balances fidelity to the observed data (i.e. producing a small residual sum of squares) against simplicity of the model (small $\|\mathbf{u}\|$). We refer to this type of criterion as

a smoothing objective, in the sense that it seeks to smooth out the fitted response by reducing model complexity while still retaining reasonable fidelity to the observed data.

For the purpose of evaluating the likelihood we will regard the PRSS criterion as a function of the parameters, given the data, and write its minimum value as

$$r^2_{\theta,\beta} = \min_{\mathbf{u}} \|\mathbf{y}_{\text{obs}} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta \mathbf{u}\|^2 + \|\mathbf{u}\|^2. \tag{5.16}$$

Notice that β only enters the right hand side of (5.16) through the linear predictor expression. We will see that $\tilde{\mathbf{u}}$ can be determined by a direct (i.e. non-iterative) calculation and, in fact, we can minimize the PRSS criterion with respect to $\mathbf{u}$ and β simultaneously without iterating. We write this minimum value as

$$r^2_{\theta} = \min_{\mathbf{u},\beta} \|\mathbf{y}_{\text{obs}} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta \mathbf{u}\|^2 + \|\mathbf{u}\|^2. \tag{5.17}$$

The value of β at the minimum is called the conditional estimate of β given θ, written $\widehat{\beta}_\theta$.

5.4 Determining the PLS Solutions, $\tilde{\mathbf{u}}$ and $\widehat{\beta}_\theta$

One way of expressing a penalized least squares problem like (5.16) is by incorporating the penalty as "pseudo-data" in an ordinary least squares problem. We extend the "response vector", which is $\mathbf{y}_{\text{obs}} - \mathbf{X}\beta$ when we minimize with respect to $\mathbf{u}$ only, with q responses that are 0 and we extend the predictor expression, $\mathbf{Z}\Lambda_\theta\mathbf{u}$ with $\mathbf{I}_q\mathbf{u}$. Writing this as a least squares problem produces

$$\tilde{\mathbf{u}} = \arg\min_{\mathbf{u}} \left\| \begin{bmatrix} \mathbf{y}_{\text{obs}} - \mathbf{X}\beta \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{Z}\Lambda_\theta \\ \mathbf{I}_q \end{bmatrix} \mathbf{u} \right\|^2 \tag{5.18}$$

with a solution that satisfies

$$\left(\Lambda_\theta^\top \mathbf{Z}^\top \mathbf{Z}\Lambda_\theta + \mathbf{I}_q\right) \tilde{\mathbf{u}} = \Lambda_\theta^\top \mathbf{Z}^\top \left(\mathbf{y}_{\text{obs}} - \mathbf{X}\beta\right) \tag{5.19}$$

To evaluate $\tilde{\mathbf{u}}$ we form the *sparse Cholesky factor*, $\mathbf{L}_\theta$, which is a lower triangular $q \times q$ matrix that satisfies

$$\mathbf{L}_\theta \mathbf{L}_\theta^\top = \Lambda_\theta^\top \mathbf{Z}^\top \mathbf{Z}\Lambda_\theta + \mathbf{I}_q. \tag{5.20}$$

The actual evaluation of sparse Cholesky factor, $\mathbf{L}_\theta$, often incorporates a *fill-reducing permutation*, which we describe next.

5.4.1 The Fill-reducing Permutation, P

In earlier chapters we have seen that often the random effects vector is re-ordered before $\mathbf{L}_\theta$ is created. The re-ordering or permutation of the elements of $\mathbf{u}$ and, correspondingly, the columns of the model matrix, $\mathbf{Z}\Lambda_\theta$, does not affect the theory of linear mixed models but can have a profound effect on the time and storage required to evaluate $\mathbf{L}_\theta$ in large problems. We write the effect of the permutation as multiplication by a $q \times q$ *permutation matrix*, $\mathbf{P}$, although in practice we apply the permutation without ever constructing $\mathbf{P}$. That is, the matrix $\mathbf{P}$ is only a notational convenience only.

The matrix $\mathbf{P}$ consists of permuted columns of the identity matrix, $\mathbf{I}_q$, and it is easy to establish that the inverse permutation corresponds to multiplication by $\mathbf{P}^\mathsf{T}$. Because multiplication by $\mathbf{P}$ or by $\mathbf{P}^\mathsf{T}$ simply re-orders the components of a vector, the length of the vector is unchanged. Thus,

$$\|\mathbf{P}\mathbf{u}\|^2 = \|\mathbf{u}\|^2 = \|\mathbf{P}^\mathsf{T}\mathbf{u}\|^2 \tag{5.21}$$

and we can express the penalty in (5.17) in any of these three forms. The properties of $\mathbf{P}$ that it preserves lengths of vectors and that its transpose is its inverse are summarized by stating that $\mathbf{P}$ is an *orthogonal matrix*.

The permutation represented by $\mathbf{P}$ is determined from the structure of $\Lambda_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda_\theta + \mathbf{I}_q$ for some initial value of θ. The particular value of θ does not affect the result because the permutation depends only the positions of the non-zeros, not the numerical values at these positions.

Taking into account the permutation, the sparse Cholesky factor, $\mathbf{L}_\theta$, is defined to be the sparse, lower triangular, $q \times q$ matrix with positive diagonal elements satisfying

$$\mathbf{L}_\theta\mathbf{L}_\theta^\mathsf{T} = \mathbf{P}\left(\Lambda_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda_\theta + \mathbf{I}_q\right)\mathbf{P}^\mathsf{T}. \tag{5.22}$$

Note that we now require that the diagonal elements of Λ_θ be positive. Problems 5.1 and 5.2 indicate why we can require this. Because the diagonal elements of Λ_θ are positive, its determinant, $|\Lambda_\theta|$, which, for a triangular matrix such as Λ_θ, is simply the product of its diagonal elements, is also positive.

Many sparse matrix methods, including the sparse Cholesky decomposition, are performed in two stages: the *symbolic phase* in which the locations of the non-zeros in the result are determined and the *numeric phase* in which the numeric values at these positions are evaluated. The symbolic phase for the decomposition (5.22), which includes determining the permutation, $\mathbf{P}$, need only be done once. Evaluation of $\mathbf{L}_\theta$ for subsequent values of θ requires only the numeric phase, which typically is much faster than the symbolic phase.

The permutation, $\mathbf{P}$, serves two purposes. The first and most important purpose is to reduce the number of non-zeros in the factor, $\mathbf{L}_\theta$. The factor is potentially non-zero at every non-zero location in the lower triangle of the

matrix being decomposed. However, as we saw in Fig. 2.4 of Sect. 2.1.2, there may be positions in the factor that get filled-in even though they are known to be zero in the matrix being decomposed. The *fill-reducing permutation* is chosen according to certain heuristics to reduce the amount of fill-in. We use the approximate minimal degree (AMD) method described in Davis [1996]. After the fill-reducing permutation is determined, a "post-ordering" is applied. This has the effect of concentrating the non-zeros near the diagonal of the factor. See Davis [2006] for details.

The pseudo-data representation of the PLS problem, (5.18), becomes

$$\tilde{\mathbf{u}} = \arg\min_{\mathbf{u}} \left\| \begin{bmatrix} \mathbf{y}_{\text{obs}} - \mathbf{X}\beta \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{Z}\Lambda_\theta \mathbf{P}^\mathsf{T} \\ \mathbf{P}^\mathsf{T} \end{bmatrix} \mathbf{P}\mathbf{u} \right\|^2 \tag{5.23}$$

and the system of linear equations satisfied by $\tilde{\mathbf{u}}$ is

$$\mathbf{L}_\theta \mathbf{L}_\theta^\mathsf{T} \mathbf{P}\tilde{\mathbf{u}} = \mathbf{P}\left(\Lambda_\theta^\mathsf{T} \mathbf{Z}^\mathsf{T} \mathbf{Z}\Lambda_\theta + \mathbf{I}_q\right) \mathbf{P}^\mathsf{T} \mathbf{P}\tilde{\mathbf{u}} = \mathbf{P}\Lambda_\theta^\mathsf{T} \mathbf{Z}^\mathsf{T} \left(\mathbf{y}_{\text{obs}} - \mathbf{X}\beta\right). \tag{5.24}$$

Obtaining the Cholesky factor, $\mathbf{L}_\theta$, may not seem to be great progress toward determining $\tilde{\mathbf{u}}$ because we still must solve (5.24) for $\tilde{\mathbf{u}}$. However, it is the key to the computational methods in the `lme4` package. The ability to evaluate $\mathbf{L}_\theta$ rapidly for many different values of θ is what makes the computational methods in `lme4` feasible, even when applied to very large data sets with complex structure. Once we evaluate $\mathbf{L}_\theta$ it is straightforward to solve (5.24) for $\tilde{\mathbf{u}}$ because $\mathbf{L}_\theta$ is triangular.

In Sect. 5.6 we will describe the steps in determining this solution. First, though, we should show that the solution, $\tilde{\mathbf{u}}$, and the value of the objective at the solution, $r^2_{\theta,\beta}$, do allow us to evaluate the deviance.

5.4.2 The Value of the Deviance and Profiled Deviance

After evaluating $\mathbf{L}_\theta$ and using that to solve for $\tilde{\mathbf{u}}$, which also produces $r^2_{\beta,\theta}$, we can write the PRSS for a general $\mathbf{u}$ as

$$\|\mathbf{y}_{\text{obs}} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta \mathbf{u}\|^2 + \|\mathbf{u}\|^2 = r^2_{\theta,\beta} + \|\mathbf{L}_\theta^\mathsf{T}(\mathbf{u} - \tilde{\mathbf{u}})\|^2 \tag{5.25}$$

which finally allows us to evaluate the likelihood. We plug the right hand side of (5.25) into the definition of $h(\mathbf{u})$ and apply the change of variable

$$\mathbf{z} = \frac{\mathbf{L}_\theta^\mathsf{T}(\mathbf{u} - \tilde{\mathbf{u}})}{\sigma}. \tag{5.26}$$

The determinant of the Jacobian of this transformation,

$$\left|\frac{d\mathbf{z}}{d\mathbf{u}}\right| = \left|\frac{\mathbf{L}_\theta^\mathsf{T}}{\sigma}\right| = \frac{|\mathbf{L}_\theta|}{\sigma^q} \tag{5.27}$$

is required for the change of variable in the integral. We use the letter $\mathbf{z}$ for the transformed value because we will rearrange the integral to have the form of the integral of the density of the standard multivariate normal distribution. That is, we will use the result

$$\int_{\mathbb{R}^q} \frac{e^{-\|\mathbf{z}\|^2/2}}{(2\pi)^{q/2}}\, d\mathbf{z} = 1. \tag{5.28}$$

Putting all these pieces together gives

$$\begin{aligned}
L(\theta,\beta,\sigma) &= \int_{\mathbb{R}^q} h(\mathbf{u})\, d\mathbf{u} \\
&= \int_{\mathbb{R}^q} \frac{1}{(2\pi\sigma^2)^{(n+q)/2}} \exp\left(-\frac{r^2_{\theta,\beta} + \|\mathbf{L}_\theta^\mathsf{T}(\mathbf{u}-\tilde{\mathbf{u}})\|^2}{2\sigma^2}\right) d\mathbf{u} \\
&= \frac{\exp\left(-\frac{r^2_{\theta,\beta}}{2\sigma^2}\right)}{(2\pi\sigma^2)^{n/2}} \int_{\mathbb{R}^q} \frac{1}{(2\pi)^{q/2}} \exp\left(-\frac{\|\mathbf{L}_\theta^\mathsf{T}(\mathbf{u}-\tilde{\mathbf{u}})\|^2}{2\sigma^2}\right) \frac{|\mathbf{L}_\theta|}{|\mathbf{L}_\theta|}\frac{d\mathbf{u}}{\sigma^q} \\
&= \frac{\exp\left(-\frac{r^2_{\theta,\beta}}{2\sigma^2}\right)}{(2\pi\sigma^2)^{n/2}|\mathbf{L}_\theta|} \int_{\mathbb{R}^q} \frac{e^{-\|\mathbf{z}\|^2/2}}{(2\pi)^{q/2}}\, d\mathbf{z} \\
&= \frac{\exp\left(-\frac{r^2_{\theta,\beta}}{2\sigma^2}\right)}{(2\pi\sigma^2)^{n/2}|\mathbf{L}_\theta|}.
\end{aligned} \tag{5.29}$$

The deviance can now be expressed as

$$d(\theta,\beta,\sigma|\mathbf{y}_{\text{obs}}) = -2\log\left(L(\theta,\beta,\sigma|\mathbf{y}_{\text{obs}})\right) = n\log(2\pi\sigma^2) + 2\log|\mathbf{L}_\theta| + \frac{r^2_{\beta,\theta}}{\sigma^2},$$

as stated in (1.6). The maximum likelihood estimates of the parameters are those that minimize this deviance.

Equation (1.6) is a remarkably compact expression, considering that the class of models to which it applies is very large indeed. However, we can do better than this if we notice that β affects (1.6) only through $r^2_{\beta,\theta}$, and, for any value of θ, minimizing this expression with respect to β is just an extension of the penalized least squares problem. Let $\widehat{\beta}_\theta$ be the value of β that minimizes the PRSS simultaneously with respect to β and $\mathbf{u}$ and let r^2_θ be the PRSS at these minimizing values. If, in addition, we set $\widehat{\sigma^2}_\theta = r^2_\theta/n$, which is the value of σ^2 that minimizes the deviance for a given value of r^2_θ, then the *profiled deviance*, which is a function of θ only, becomes

$$\tilde{d}(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}}) = 2\log|\mathbf{L}_\theta| + n\left[1 + \log\left(\frac{2\pi r_\theta^2}{n}\right)\right]. \tag{5.30}$$

Numerical optimization (minimization) of $\tilde{d}(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}})$ with respect to $\boldsymbol{\theta}$ determines the MLE, $\widehat{\boldsymbol{\theta}}$. The MLEs for the other parameters, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\sigma}}$, are the corresponding conditional estimates evaluated at $\widehat{\boldsymbol{\theta}}$.

5.4.3 Determining r_θ^2 and $\widehat{\boldsymbol{\beta}}_\theta$

To determine $\tilde{\mathbf{u}}$ and $\widehat{\boldsymbol{\beta}}_\theta$ simultaneously we rearrange the terms in (5.23) as

$$\begin{bmatrix}\tilde{\mathbf{u}}\\ \widehat{\boldsymbol{\beta}}_\theta\end{bmatrix} = \arg\min_{\mathbf{u},\boldsymbol{\beta}} \left\| \begin{bmatrix}\mathbf{y}_{\text{obs}}\\ \mathbf{0}\end{bmatrix} - \begin{bmatrix}\mathbf{Z}\boldsymbol{\Lambda}_\theta\mathbf{P}^\mathsf{T} & \mathbf{X}\\ \mathbf{P}^\mathsf{T} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{P}\mathbf{u}\\ \boldsymbol{\beta}\end{bmatrix}\right\|^2. \tag{5.31}$$

The PLS values, $\tilde{\mathbf{u}}$ and $\widehat{\boldsymbol{\beta}}_\theta$, are the solutions to

$$\begin{bmatrix}\mathbf{P}\left(\boldsymbol{\Lambda}_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\boldsymbol{\Lambda}_\theta + \mathbf{I}_q\right)\mathbf{P}^\mathsf{T} & \mathbf{P}\boldsymbol{\Lambda}_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{X}\\ \mathbf{X}^\mathsf{T}\mathbf{Z}\boldsymbol{\Lambda}_\theta\mathbf{P}^\mathsf{T} & \mathbf{X}^\mathsf{T}\mathbf{X}\end{bmatrix}\begin{bmatrix}\mathbf{P}\tilde{\mathbf{u}}\\ \widehat{\boldsymbol{\beta}}_\theta\end{bmatrix} = \begin{bmatrix}\mathbf{P}\boldsymbol{\Lambda}_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{y}_{\text{obs}}\\ \mathbf{X}^\mathsf{T}\mathbf{y}_{\text{obs}}.\end{bmatrix} \tag{5.32}$$

To evaluate these solutions we decompose the system matrix as

$$\begin{bmatrix}\mathbf{P}\left(\boldsymbol{\Lambda}_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\boldsymbol{\Lambda}_\theta + \mathbf{I}_q\right)\mathbf{P}^\mathsf{T} & \mathbf{P}\boldsymbol{\Lambda}_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{X}\\ \mathbf{X}^\mathsf{T}\mathbf{Z}\boldsymbol{\Lambda}_\theta\mathbf{P}^\mathsf{T} & \mathbf{X}^\mathsf{T}\mathbf{X}\end{bmatrix} = \begin{bmatrix}\mathbf{L}_\theta & \mathbf{0}\\ \mathbf{R}_{ZX}^\mathsf{T} & \mathbf{R}_X^\mathsf{T}\end{bmatrix}\begin{bmatrix}\mathbf{L}_\theta^\mathsf{T} & \mathbf{R}_{ZX}\\ \mathbf{0} & \mathbf{R}_X\end{bmatrix} \tag{5.33}$$

where, as before, $\mathbf{L}_\theta$, the sparse Cholesky factor, is the sparse lower triangular $q \times q$ matrix satisfying (5.22). The other two matrices in (5.33): $\mathbf{R}_{ZX}$, which is a general $q \times p$ matrix, and $\mathbf{R}_X$, which is an upper triangular $p \times p$ matrix, satisfy

$$\mathbf{L}_\theta\mathbf{R}_{ZX} = \mathbf{P}\boldsymbol{\Lambda}_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{X} \tag{5.34}$$

and

$$\mathbf{R}_X^\mathsf{T}\mathbf{R}_X = \mathbf{X}^\mathsf{T}\mathbf{X} - \mathbf{R}_{ZX}^\mathsf{T}\mathbf{R}_{ZX} \tag{5.35}$$

Those familiar with standard ways of writing a Cholesky decomposition as either $\mathbf{L}\mathbf{L}^\mathsf{T}$ or $\mathbf{R}^\mathsf{T}\mathbf{R}$ ($\mathbf{L}$ is the factor as it appears on the left and $\mathbf{R}$ is as it appears on the right) will notice a notational inconsistency in (5.33). One Cholesky factor is defined as the lower triangular fractor on the left and the other is defined as the upper triangular factor on the right. It happens that in R the Cholesky factor of a dense positive-definite matrix is returned as the right factor, whereas the sparse Cholesky factor is returned as the left factor.

One other technical point that should be addressed is whether $\mathbf{X}^\mathsf{T}\mathbf{X} - \mathbf{R}_{ZX}^\mathsf{T}\mathbf{R}_{ZX}$ is positive definite. In theory, if $\mathbf{X}$ has full column rank, so that $\mathbf{X}^\mathsf{T}\mathbf{X}$ is positive definite, then the downdated matrix, $\mathbf{X}^\mathsf{T}\mathbf{X} - \mathbf{R}_{ZX}^\mathsf{T}\mathbf{R}_{ZX}$, must also be positive definite (see Prob. 5.4). In practice, the downdated matrix can

become computationally singular in ill-conditioned problems, in which case an error is reported.

The extended decomposition (5.33) not only provides for the evaluation of the profiled deviance function, $\tilde{d}(\theta)$, (5.30) but also allows us to define and evaluate the profiled REML criterion.

5.5 The REML Criterion

The so-called REML estimates of variance components are often preferred to the maximum likelihood estimates. ("REML" can be considered to be an acronym for "restricted" or "residual" maximum likelihood, although neither term is completely accurate because these estimates do not maximize a likelihood.) We can motivate the use of the REML criterion by considering a linear regression model,

$$\mathcal{Y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{I}_n), \tag{5.36}$$

in which we typically estimate σ^2 as

$$\widehat{\sigma^2_R} = \frac{\|\mathbf{y}_{\text{obs}} - \mathbf{X}\widehat{\beta}\|^2}{n-p} \tag{5.37}$$

even though the maximum likelihood estimate of σ^2 is

$$\widehat{\sigma^2_L} = \frac{\|\mathbf{y}_{\text{obs}} - \mathbf{X}\widehat{\beta}\|^2}{n}. \tag{5.38}$$

The argument for preferring $\widehat{\sigma^2_R}$ to $\widehat{\sigma^2_L}$ as an estimate of σ^2 is that the numerator in both estimates is the sum of squared residuals at $\widehat{\beta}$ and, although the residual vector, $\mathbf{y}_{\text{obs}} - \mathbf{X}\widehat{\beta}$, is an n-dimensional vector, the residual at $\widehat{\theta}$ satisfies p linearly independent constraints, $\mathbf{X}^\mathsf{T}(\mathbf{y}_{\text{obs}} - \mathbf{X}\widehat{\beta}) = \mathbf{0}$. That is, the residual at $\widehat{\theta}$ is the projection of the observed response vector, $\mathbf{y}_{\text{obs}}$, into an $(n-p)$-dimensional linear subspace of the n-dimensional response space. The estimate $\widehat{\sigma^2_R}$ takes into account the fact that σ^2 is estimated from residuals that have only $n-p$ *degrees of freedom.*

Another argument often put forward for REML estimation is that $\widehat{\sigma^2_R}$ is an *unbiased* estimate of σ^2, in the sense that the expected value of the estimator is equal to the value of the parameter. However, determining the expected value of an estimator involves integrating with respect to the density of the estimator and we have seen that densities of estimators of variances will be skewed, often highly skewed. It is not clear why we should be interested in the expected value of a highly skewed estimator. If we were to transform to a more symmetric scale, such as the estimator of the standard deviation or the estimator of the logarithm of the standard deviation, the REML estimator

would no longer be unbiased. Furthermore, this property of unbiasedness of variance estimators does not generalize from the linear regression model to linear mixed models. This is all to say that the distinction between REML and ML estimates of variances and variance components is probably less important that many people believe.

Nevertheless it is worthwhile seeing how the computational techniques described in this chapter apply to the REML criterion because the REML parameter estimates $\widehat{\theta}_R$ and $\widehat{\sigma^2_R}$ for a linear mixed model have the property that they would specialize to $\widehat{\sigma^2_R}$ from (5.37) for a linear regression model, as seen in Sect. 1.3.2.

Although not usually derived in this way, the REML criterion (on the deviance scale) can be expressed as

$$d_R(\theta,\sigma|\mathbf{y}_{\text{obs}}) = -2\log\int_{\mathbb{R}^p} L(\theta,\beta,\sigma|\mathbf{y}_{\text{obs}})\,d\beta. \tag{5.39}$$

The REML estimates $\widehat{\theta}_R$ and $\widehat{\sigma^2_R}$ minimize $d_R(\theta,\sigma|\mathbf{y}_{\text{obs}})$.

To evaluate this integral we form an expansion, similar to (5.25), of $r^2_{\theta,\beta}$ about $\widehat{\beta}_\theta$

$$r^2_{\theta,\beta} = r^2_\theta + \|\mathbf{R}_X(\beta-\widehat{\beta}_\theta)\|^2. \tag{5.40}$$

In the same way that (5.25) was used to simplify the integral in (5.29), we can derive

$$\int_{\mathbb{R}^p} \frac{\exp\left(-\frac{r^2_{\theta,\beta}}{2\sigma^2}\right)}{(2\pi\sigma^2)^{n/2}|\mathbf{L}_\theta|}\,d\beta = \frac{\exp\left(-\frac{r^2_\theta}{2\sigma^2}\right)}{(2\pi\sigma^2)^{(n-p)/2}|\mathbf{L}_\theta||\mathbf{R}_X|} \tag{5.41}$$

corresponding to a REML criterion on the deviance scale of

$$d_R(\theta,\sigma|\mathbf{y}_{\text{obs}}) = (n-p)\log(2\pi\sigma^2) + 2\log\left(|\mathbf{L}_\theta||\mathbf{R}_X|\right) + \frac{r^2_\theta}{\sigma^2}. \tag{5.42}$$

Plugging in the conditional REML estimate, $\widehat{\sigma^2}_R = r^2_\theta/(n-p)$, provides the profiled REML criterion

$$\tilde{d}_R(\theta|\mathbf{y}_{\text{obs}}) = 2\log\left(|\mathbf{L}_\theta||\mathbf{R}_X|\right) + (n-p)\left[1+\log\left(\frac{2\pi r^2_\theta}{n-p}\right)\right] \tag{5.43}$$

The REML estimate of θ is

$$\widehat{\theta}_R = \arg\min_\theta \tilde{d}_R(\theta|\mathbf{y}_{\text{obs}}), \tag{5.44}$$

and the REML estimate of σ^2 is the conditional REML estimate of σ^2 at $\widehat{\theta}_R$,

$$\widehat{\sigma^2_R} = r^2_{\widehat{\theta}_R}/(n-p). \tag{5.45}$$

It is not entirely clear how one would define a "REML estimate" of β because the REML criterion, $d_R(\theta, \sigma|\mathbf{y})$, defined in (5.42), does not depend on β. However, it is customary (and not unreasonable) to use $\widehat{\beta}_R = \widehat{\beta}_{\widehat{\theta}_R}$ as the REML estimate of β.

5.6 Step-by-step Evaluation of the Profiled Deviance

As we have seen, an object returned by `lmer` contains an environment, accessed with the `env` extractor. This environment contains several matrices and vectors that are used in the evaluation of the profiled deviance. In this section we use these matrices and vectors from one of our examples to explicitly trace the steps in evaluating the profiled deviance. This level of detail is provided for those whose style of learning is more of a "hands on" style and for those who may want to program modifications of this approach.

Consider our model `fm8`, fit as

```
> fm8 <- lmer(Reaction ~ 1 + Days + (1 + Days|Subject), sleepstudy,
+             REML = 0, verbose = TRUE)

  0:     1784.6423:  1.00000  0.00000  1.00000
  1:     1774.2946:  1.00042 -0.00836471 0.725280
  2:     1754.3212: 0.998969 -0.0239943 0.175808
  3:     1752.1500: 0.998284 0.00682229 0.243192
...
```

The environment of the model contains the converged parameter vector, θ (`theta`), the relative covariance factor, Λ_θ (`Lambda`), the sparse Cholesky factor, $\mathbf{L}_\theta$ (`L`), the matrices $\mathbf{R}_{ZX}$ (`RZX`) and $\mathbf{R}_X$ (`RX`), the conditional mode, $\tilde{\mathbf{u}}$ (`u`), and the conditional estimate, $\widehat{\beta}_\theta$ (`fixef`). The permutation represented by $\mathbf{P}$ is contained in the sparse Cholesky representation, `L`.

Although the model matrices, $\mathbf{X}$ (`X`) and $\mathbf{Z}^\mathsf{T}$ (`Zt`), and the response vector, $\mathbf{y}_{\text{obs}}$ (`y`), are available in the environment, many of the products that involve only these fixed values are precomputed and stored separately under the names `XtX` ($\mathbf{X}^\mathsf{T}\mathbf{X}$), `Xty`, `ZtX` and `Zty`.

To provide easy access to the objects in the environment of `fm8` we attach it to the search path.

```
> attach(env(fm8))
```

Please note that this is done here for illustration only. The practice of attaching a list or a data frame or, less commonly, an environment in an R session is overused, somewhat dangerous (because of the potential of forgetting to detach it later) and discouraged. The preferred practice is to use the `with` function to gain access by name to components of such composite objects. For this section of code, however, using `with` or `within` would quickly become very tedious and we use `attach` instead.

To update the matrix Λ_θ to a new value of θ we need to know which of the non-zeros in Λ are updated from which elements of θ. Recall that the dimension of θ is small (3, in this case) but Λ is potentially large (18×18 with 54 non-zeros). The environment contains an integer vector `Lind` that maps the elements of `theta` to the non-zeros in `Lambda`.

Suppose we wish to recreate the evaluation of the profiled deviance at the initial value of $\theta = (1, 0, 1)$. We begin by updating Λ_θ and forming the product $\mathbf{U}^\mathsf{T} = \Lambda_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}$

```
> str(Lambda)

Formal class 'dgCMatrix' [package "Matrix"] with 6 slots
  ..@ i       : int [1:54] 0 1 1 2 3 3 4 5 5 6 ...
  ..@ p       : int [1:37] 0 2 3 5 6 8 9 11 12 14 ...
  ..@ Dim     : int [1:2] 36 36
  ..@ Dimnames:List of 2
  .. ..$ : NULL
  .. ..$ : NULL
  ..@ x       : num [1:54] 0.9292 0.0182 0.2226 0.9292 0.0182 ...
  ..@ factors : list()

> str(Lind)

 int [1:54] 1 2 3 1 2 3 1 2 3 1 ...

> Lambda@x[] <- c(1,0,1)[Lind]
> str(Lambda@x)

 num [1:54] 1 0 1 1 0 1 1 0 1 1 ...

> Ut <- crossprod(Lambda, Zt)
```

The Cholesky factor object, `L`, can be updated from `Ut` without forming $\mathbf{U}^\mathsf{T}\mathbf{U} + \mathbf{I}$ explicitly. The optional argument `mult` to the `update` method specifies a multiple of the identity to be added to $\mathbf{U}^\mathsf{T}\mathbf{U}$

```
> L <- update(L, Ut, mult = 1)
```

Then we evaluate `RZX` and `RX` according to (5.34) and (5.35)

```
> RZX <- solve(L, solve(L, crossprod(Lambda, ZtX), sys = "P"), sys = "L")
> RX <- chol(XtX - crossprod(RZX))
```

Solving (5.32) for $\tilde{\mathbf{u}}$ and $\widehat{\beta}_\theta$ is done in stages. Writing $\mathbf{c}_u$ and $\mathbf{c}_\beta$ for the intermediate results that satisfy

$$\begin{bmatrix} \mathbf{L}_\theta & \mathbf{0} \\ \mathbf{R}_{ZX}^\mathsf{T} & \mathbf{R}_X^\mathsf{T} \end{bmatrix} \begin{bmatrix} \mathbf{c}_u \\ \mathbf{c}_\beta \end{bmatrix} = \begin{bmatrix} \mathbf{P}\Lambda_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{y}_{\text{obs}} \\ \mathbf{X}^\mathsf{T}\mathbf{y}_{\text{obs}}. \end{bmatrix} \tag{5.46}$$

we evaluate

```
> cu <- solve(L, solve(L, crossprod(Lambda, Zty), sys = "P"), sys = "L")
> cbeta <- solve(t(RX), Xty - crossprod(RZX, cu))
```

The next set of equations to solve is

$$\begin{bmatrix} \mathbf{L}_\theta^\mathsf{T} & \mathbf{R}_{ZX} \\ \mathbf{0} & \mathbf{R}_X \end{bmatrix} \begin{bmatrix} \mathbf{P}\tilde{\mathbf{u}} \\ \widehat{\beta}_\theta \end{bmatrix} = \begin{bmatrix} \mathbf{c}_U \\ \mathbf{c}_\beta \end{bmatrix} \tag{5.47}$$

```
> fixef <- as.vector(solve(RX, cbeta))
> u <- solve(L, solve(L, cu - RZX %*% fixef, sys = "Lt"), sys = "Pt")
```

We can now create the conditional mean, `mu`, the penalized residual sum of squares, `prss`, the logarithm of the square of the determinant of $\mathbf{L}$, `ldL2`, and the profiled deviance, which, fortuitously, equals the value shown earlier.

```
> mu <- gamma <- as.vector(crossprod(Ut, u) + X %*% fixef)
> prss <- sum(c(y - mu, as.vector(u))^2)
> ldL2 <- 2 * as.vector(determinant(L)$mod)
> (deviance <- ldL2 + nobs * (1 + log(2 * pi * prss/nobs)))
```

```
[1] 1784.642
```

The last step is detach the environment of `fm8` from the search list

```
> detach()
```

to avoid later name clashes.

In terms of the calculations performed, these steps describe exactly the evaluation of the profiled deviance in `lmer`. The actual function for evaluating the deviance, accessible as `fm8@setPars`, is a slightly modified version of what is shown above. However, the modifications are only to avoid creating copies of potentially large objects and to allow for cases where the model matrix, $\mathbf{X}$, is sparse. In practice, unless the optional argument `compDev = FALSE` is given, the profiled deviance is evaluated in compiled code, providing a speed boost, but the R code can be used if desired. This allows for checking the results from the compiled code and can also be used as a template for extending the computational methods to other types of models.

5.7 Generalizing to Other Forms of Mixed Models

In later chapters we cover the theory and practice of generalized linear mixed models (GLMMs), nonlinear mixed models (NLMMs) and generalized nonlinear mixed models (GNLMMs). Because quite a bit of the theoretical and computational methodology covered in this chapter extends to those models we will cover the common aspects here.

5.7.1 Descriptions of the Model Forms

We apply the name "generalized" to models in which the conditional distribution, $(\mathscr{Y}|\mathscr{U}=\mathbf{u})$, is not required to be Gaussian but does preserve some of the properties of the spherical Gaussian conditional distribution

$$(\mathscr{Y}|\mathscr{U}=\mathbf{u}) \sim \mathscr{N}(\mathbf{Z}\Lambda_\theta\mathbf{u}+\mathbf{X}\beta, \sigma^2\mathbf{I}_n)$$

from the linear mixed model. In particular, the components of $\mathscr{Y}$ are *conditionally independent*, given $\mathscr{U}=\mathbf{u}$. Furthermore, $\mathbf{u}$ affects the distribution only through the conditional mean, which we will continue to write as μ, and it affects the conditional mean only through the linear predictor, $\gamma=\mathbf{Z}\Lambda_\theta\mathbf{u}+\mathbf{X}\beta$.

Typically we do not have $\mu=\gamma$, however. The elements of the linear predictor, γ, can be positive or negative or zero. Theoretically they can take on any value between $-\infty$ and ∞. But many distributional forms used in GLMMs put constraints on the value of the mean. For example, the mean of a Bernoulli random variable, modeling a binary response, must be in the range $0<\mu<1$ and the mean of a Poisson random variable, modeling a count, must be positive. To achieve these constraints we write the conditional mean, μ as a transformation of the unbounded predictor, written η. For historical, and some theoretical, reasons the inverse of this transformation is called the *link function*, written

$$\eta=\mathbf{g}(\mu), \tag{5.48}$$

and the transformation we want is called the *inverse link*, written $\mathbf{g}^{-1}$.

Both $\mathbf{g}$ and $\mathbf{g}^{-1}$ are determined by scalar functions, g and g^{-1}, respectively, applied to the individual components of the vector argument. That is, η must be n-dimensional and the vector-valued function $\mu=\mathbf{g}^{-1}(\eta)$ is defined by the component functions $\mu_i=g^{-1}(\eta_i)$, $i=1,\ldots,n$. Among other things, this means that the Jacobian matrix of the inverse link, $\frac{d\mu}{d\eta}$, will be diagonal.

Because the link function, $\mathbf{g}$, and the inverse link, $\mathbf{g}^{-1}$, are nonlinear functions (there would be no purpose in using a linear link function) many people use the terms "generalized linear mixed model" and "nonlinear mixed model" interchangably. We reserve the term "nonlinear mixed model" for the type of models used, for example, in pharmacokinetics and pharmacodynamics, where the conditional distribution is a spherical multivariate Gaussian

$$(\mathscr{Y}|\mathscr{U}=\mathbf{u}) \sim \mathscr{N}(\mu, \sigma^2\mathbf{I}_n) \tag{5.49}$$

but μ depends nonlinearly on γ. For NLMMs the length of the linear predictor, γ, is a multiple, ns, of n, the length of μ.

Like the map from η to μ, the map from γ to μ has a "diagonal" property, which we now describe. If we use γ to fill the columns of an $n\times s$ matrix, Γ, then μ_i depends only on the ith row of Γ. In fact, μ_i is determined by a

nonlinear model function, f, applied to the i row of Γ. Writing $\mu = \mathbf{f}(\gamma)$ based on the component function f, we see that the Jacobian of $\mathbf{f}$, $\frac{d\mu}{d\gamma}$, will be the vertical concatenation of s diagonal $n \times n$ matrices.

Because we will allow for generalized nonlinear mixed models (GNLMMs), in which the mapping from γ to μ has the form

$$\gamma \rightarrow \eta \rightarrow \mu, \tag{5.50}$$

we will use (5.50) in our definitions.

5.7.2 Determining the Conditional Mode, $\tilde{\mathbf{u}}$

For all these types of mixed models, the conditional distribution, $(\mathscr{U}|\mathscr{Y} = \mathbf{y}_{\text{obs}})$, is a continuous distribution for which we can determine the unscaled conditional density, $h(\mathbf{u})$. As for linear mixed models, we define the conditional mode, $\tilde{\mathbf{u}}$ as the value that maximizes the unscaled condtional density.

Determining the conditional mode, $\tilde{\mathbf{u}}$, in a nonlinear mixed model is a penalized nonlinear least squares (PNLS) problem

$$\tilde{\mathbf{u}} = \arg\min_{\mathbf{u}} \|\mathbf{y}_{\text{obs}} - \mu\|^2 + \|\mathbf{u}\|^2 \tag{5.51}$$

which we solve by adapting the iterative techniques, such as the Gauss-Newton method [Bates and Watts, 1988, Sect. 2.2.1], used for nonlinear least squares. Starting at an initial value, $\mathbf{u}^{(0)}$, (the bracketed superscript denotes the iteration number) with conditional mean, $\mu^{(0)}$, we determine an increment $\delta^{(1)}$ by solving the penalized linear least squares problem,

$$\delta^{(1)} = \arg\min_{\delta} \left\| \begin{bmatrix} \mathbf{y}_{\text{obs}} - \mu^{(0)} \\ \mathbf{0} - \mathbf{u}^{(0)} \end{bmatrix} - \begin{bmatrix} \mathbf{U}^{(0)} \\ \mathbf{I}_q \end{bmatrix} \delta \right\|^2 \tag{5.52}$$

where

$$\mathbf{U}^{(0)} = \left.\frac{d\mu}{d\mathbf{u}}\right|_{\mathbf{u}^{(0)}}. \tag{5.53}$$

Naturally, we use the sparse Cholesky decomposition, $\mathbf{L}_\theta^{(0)}$, satisfying

$$\mathbf{L}_\theta^{(0)} \left(\mathbf{L}_\theta^{(0)}\right) = \mathbf{P}\left[\left(\mathbf{U}^{(0)}\right)^{\mathsf{T}} \mathbf{U}^{(0)} + \mathbf{I}_q\right] \mathbf{P}^{\mathsf{T}} \tag{5.54}$$

to determine this increment. The next iteration begins at

$$\mathbf{u}^{(1)} = \mathbf{u}^{(0)} + k\delta^{(1)} \tag{5.55}$$

where k is the step factor chosen, perhaps by step-halving [Bates and Watts, 1988, Sect. 2.2.1], to ensure that the penalized residual sum of squares decreases at each iteration. Convergence is declared when the orthogonality convergence criterion [Bates and Watts, 1988, Sect. 2.2.3] is below some prespecified tolerance.

The *Laplace approximation* to the deviance is

$$d(\theta,\beta,\sigma|\mathbf{y}_{\text{obs}}) \approx= n\log(2\pi\sigma^2)+2\log|\mathbf{L}_{\theta,\beta}|+\frac{r^2_{\theta,\beta}}{\sigma^2}, \tag{5.56}$$

where the Cholesky factor, $\mathbf{L}_{\theta,\beta}$, and the penalized residual sum of squares, $r^2_{\theta,\beta}$, are both evaluated at the conditional mode, $\tilde{\mathbf{u}}$. The Cholesky factor depends on θ, β and $\mathbf{u}$ for these models but typically the dependence on β and $\mathbf{u}$ is weak.

5.8 Chapter Summary

The definitions and the computational results for maximum likelihood estimation of the parameters in linear mixed models were summarized in Sect. 1.4.1. A key computation is evaluation of the sparse Cholesky factor, Λ_θ, satisfying eqn. 5.22,

$$\mathbf{L}_\theta\mathbf{L}_\theta^\mathsf{T} = \mathbf{P}\left(\Lambda_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda_\theta+\mathbf{I}_q\right)\mathbf{P}^\mathsf{T}.$$

where $\mathbf{P}$ represents the fill-reducing permutation determined during the symbolic phase of the sparse Cholesky decomposition.

An extended decomposition (eqn. 5.33) provides the $q\times p$ matrix $\mathbf{R}_{ZX}$ and the $p\times p$ upper triangular $\mathbf{R}_X$ that are used to determine the conditional mode $\tilde{\mathbf{u}}_\theta$, the conditional estimate $\widehat{\beta}_\theta$ and the minimum penalized residual sum of squares, r^2_θ from which the profiled deviance

$$\tilde{d}(\theta|\mathbf{y}_{\text{obs}}) = 2\log|\mathbf{L}_\theta|+n\left[1+\log\left(\frac{2\pi r^2_\theta}{n}\right)\right].$$

or the profile REML criterion

$$\tilde{d}_R(\theta|\mathbf{y}_{\text{obs}}) = 2\log\left(|\mathbf{L}_\theta||\mathbf{R}_X|\right)+(n-p)\left[1+\log\left(\frac{2\pi r^2_\theta}{n-p}\right)\right]$$

can be evaluated and optimized (minimized) with respect to θ.

Exercises

Unlike the exercises in other chapters, these exercises establish theoretical results, which do not always apply exactly to the computational results.

5.1. Show that the matrix $\mathbf{A}_\theta = \mathbf{P}\boldsymbol{\Lambda}_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\boldsymbol{\Lambda}_\theta\mathbf{P}^\mathsf{T} + \mathbf{I}_q$ is positive definite. That is, $\mathbf{b}^\mathsf{T}\mathbf{A}\mathbf{b} > 0, \forall \mathbf{b} \neq \mathbf{0}$.

5.2. (a) Show that $\boldsymbol{\Lambda}_\theta$ can be defined to have non-negative diagonal elements. (Hint: Show that the product $\boldsymbol{\Lambda}_\theta\mathbf{D}$ where $\mathbf{D}$ is a diagonal matrix with diagonal elements of ± 1 is also a Cholesky factor. Thus the signs of the diagonal elements can be chosen however we want.)

(b) Use the result of Prob. 5.1 to show that the diagonal elements of $\boldsymbol{\Lambda}_\theta$ must be non-zero. (Hint: Suppose that the first zero on the diagonal of $\boldsymbol{\Lambda}_\theta$ is in the ith position. Show that there is a solution $\mathbf{x}$ to $\boldsymbol{\Lambda}_\theta^\mathsf{T}\mathbf{x} = \mathbf{0}$ with $x_i = 1$ and $x_j = 0$, $j = i+1, \ldots, q$ and that this $\mathbf{x}$ contradicts the positive definite condition.)

5.3. Show that if $\mathbf{X}$ has full column rank, which means that there does not exist a $\boldsymbol{\beta} \neq \mathbf{0}$ for which $\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$, then $\mathbf{X}^\mathsf{T}\mathbf{X}$ is positive definite.

5.4. Show that if $\mathbf{X}$ has full column rank then

$$\begin{bmatrix} \mathbf{Z}\boldsymbol{\Lambda}_\theta\mathbf{P}^\mathsf{T} & \mathbf{X} \\ \mathbf{P}^\mathsf{T} & \mathbf{0} \end{bmatrix}$$

also must have full column rank. (Hint: First show that $\mathbf{u}$ must be zero in any vector $\begin{bmatrix} \mathbf{u} \\ \boldsymbol{\beta} \end{bmatrix}$ satisfying

$$\begin{bmatrix} \mathbf{Z}\boldsymbol{\Lambda}_\theta\mathbf{P}^\mathsf{T} & \mathbf{X} \\ \mathbf{P}^\mathsf{T} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\beta} \end{bmatrix} = \mathbf{0}.$$

Use this result and (5.33) to show that

$$\begin{bmatrix} \mathbf{L}_\theta & \mathbf{0} \\ \mathbf{R}_{ZX}^\mathsf{T} & \mathbf{R}_X^\mathsf{T} \end{bmatrix} \begin{bmatrix} \mathbf{L}_\theta^\mathsf{T} & \mathbf{R}_{ZX} \\ \mathbf{0} & \mathbf{R}_X \end{bmatrix}$$

is positive definite and, hence, $\mathbf{R}_X$ is non-singular.)

References

Douglas M. Bates and Donald G. Watts. *Nonlinear Regression Analysis and Its Applications*. Wiley, Hoboken, NJ, 1988. ISBN 0-471-81643-4.

Gregory Belenky, Nancy J. Wesensten, David R. Thorne, Maria L. Thomas, Helen C. Sing, Daniel P. Redmond, Michael B. Russo, and Thomas J. Balkin. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. *Journal of Sleep Research*, 12:1–12, 2003.

G.E.P. Box and G.C. Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA, 1973.

Bill Cleveland. *Visualizing Data*. Hobart Press, Summit, NJ, 1993.

Owen L. Davies and Peter L. Goldsmith, editors. *Statistical Methods in Research and Production*. Hafner, 4th edition, 1972.

Tim Davis. An approximate minimal degree ordering algorithm. *SIAM J. Matrix Analysis and Applications*, 17(4):886–905, 1996.

Tim Davis. *Direct Methods for Sparse Linear Systems*. SIAM, Philadelphia, PA, 2006.

Friedrich Leisch. Sweave: Dynamic generation of statistical reports using literate data analysis. In Wolfgang Härdle and Bernd Rönz, editors, *Compstat 2002 — Proceedings in Computational Statistics*, pages 575–580. Physica Verlag, Heidelberg, 2002. URL `http://www.stat.uni-muenchen.de/~leisch/Sweave`. ISBN 3-7908-1517-9.

José C. Pinheiro and Douglas M. Bates. *Mixed-effects Models in S and S-PLUS*. Springer, 2000.

J. Rasbash, W. Browne, H. Goldstein, M. Yang, and I. Plewis. *A User's Guide to MLwiN*. Multilevel Models Project, Institute of Education, University of London, London, 2000.

Stephen W. Raudenbush and Anthony S. Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage, 2nd edition, 2002. ISBN 0-7619-1904-X.

Y. Sakamoto, M. Ishiguro, and G. Kitagawa. *Akaike Information Criterion Statistics*. Reidel, Dordrecht, Holland, 1986.

Deepayan Sarkar. *Lattice: Multivariate Data Visualization with R*. Springer, 2008.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

Brady T. West, Kathleen B. Welch, and Andrzej T. Gałecki. *Linear Mixed Models: A Practical Guide Using Statistical Software*. Chapman and Hall/CRC, Boca Raton, FL, 2007. ISBN 1-58488-480-0.