Douglas M. Bates

lme4: Mixed-effects modeling with R

January 11, 2010

Springer

To Phyl and Dave, with thanks for their kind hospitality.

Contents

List of Figures

List of Tables

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 analyzing the fitted model. Building from the example we describe the general form of linear mixed models that can be fit using `lmer`.

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 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 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 (3).

1.2.1 The Dyestuff *Data*

The `Dyestuff` data are described in (5), 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(lme4a)
```

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 Figure~1.1 than we can from numerical summaries.

In Figure~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 (13) he describes how one would create such a plot. Because this book was created using Sweave (9), 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.

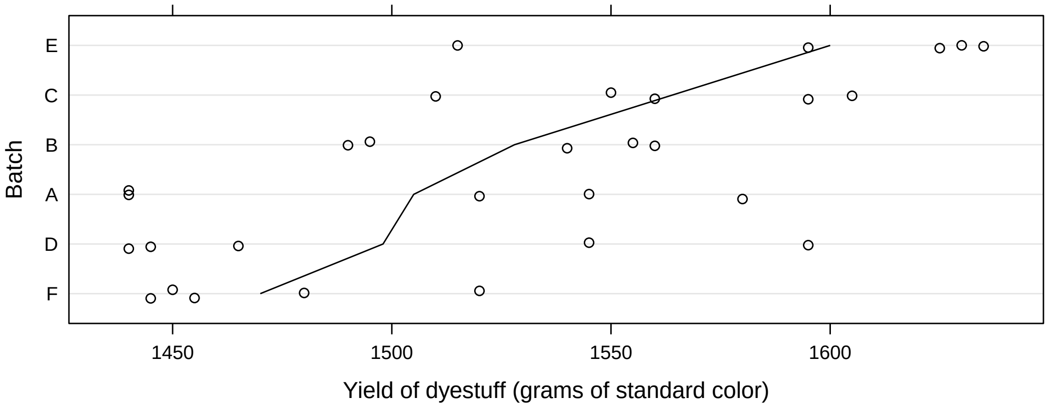


Fig. 1.1 Yield of dyestuff (Napthalene Black 12B) for 5 preparations from each of 6 batches of an intermediate product (H-acid). The line joins the mean yields from the batches, which have been ordered by increasing mean yield. The vertical positions are "jittered" slightly to avoid over-plotting. Notice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

In section˜**??** 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 section 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 (3), Table 5.1.4, p. 247 where the authors state

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

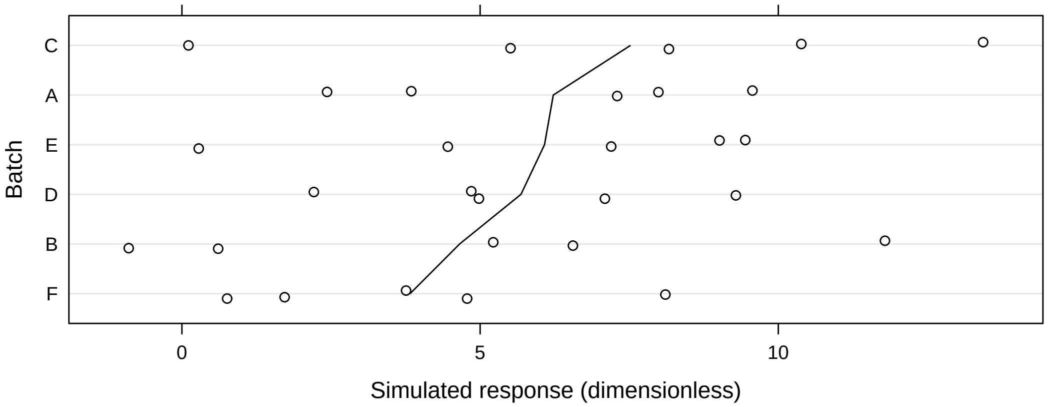


Fig. 1.2 Simulated data presented in (3) with a structure similar to that of the `Dyestuff` data. These data represent a case where the batch-to-batch variability is small relative to the within-batch variability.

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 Figure˜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 section **??**). 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. This code fragment also illustrates, in the second line, the continuation prompt, `"+"`. When an incomplete R expression is submitted, the prompt is changed from `">"` to `"+"`, indicating that additional input is required before the expression can be evaluated.)

The display of a model fit by maximum likelihood provides several other model-fit statistics such as Akaike's Information Criterion (`AIC`)˜(12), Schwarz's Bayesian Information Criterion (`BIC`)˜(14), 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 later in section~**??**.

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 gm.^{2} in the REML fit and as 1388.33 gm.^{2} in the ML fit. The corresponding standard deviations are 42.00 gm. for the REML fit and 37.26 gm. 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 section~**??** 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 gm.^{2}) and its corresponding standard deviation. For the REML fit the estimated standard deviation of the residuals is 49.51 gm. and for the ML fit it is also 49.51 gm. (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 gm. (equality is again a consequence of the simple model and balanced data set). The standard error of the intercept estimate is 19.38 gm. for the REML fit and 17.69 gm. for the ML fit.

1.3.2 Assessing the Variability of the Parameter Estimates

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 labelled as the intercept. A standard error for the estimate $\widehat{\beta_0}$ is provided but no standard errors are given for the estimates $\widehat{\sigma_1}$ and $\widehat{\sigma}$.

The reason that we do not simply quote standard errors for the estimated standard deviations is because they are misleading. Even more misleading is the common practice of quoting a standard error for an estimated variance. Quoting a parameter estimate and a standard error for the estimator gives the impression that we can summarize the precision of the estimate with a series of confidence intervals that are symmetric about the estimate. In other words we are assuming that decreasing a parameter from its estimate by an amount δ has the same effect on the quality of the model fit as does increasing the parameter by δ. It is not realistic to expect that this is true of measures of variability, such as standard deviations and variances.

Instead of making the unrealistic assumption that the precision of these parameter estimates can be summarized by standard errors we evaluate the quality of the fit for nearby values of the parameter and form confidence intervals based on the change in the deviance. This is called *profiling* the deviance with respect to a parameter. We plot the profiles, obtained by applying the function `profile` to the fitted model, on the scale of the square root of the change in the deviance, as shown in Figure˜1.3.

Because we will use such plots of the profiled deviance extensively for assessing the uncertainty in parameter estimates, we will describe this particular plot qualitatively, but in some detail, here. Formal definitions and equations will be given in the next section.

Recall that one of the measures of the quality of a model fit is the *deviance*, which is negative twice the log-likelihood evaluated at the parameter

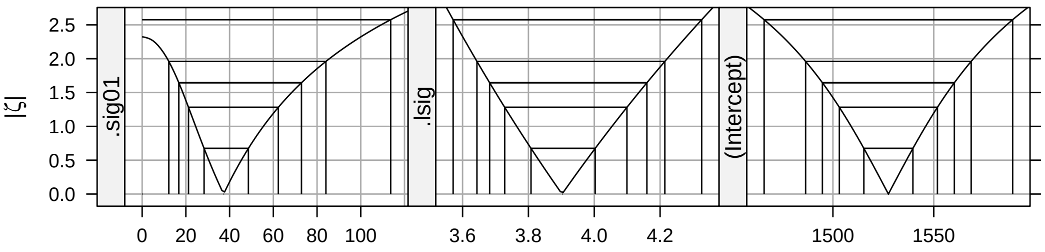


Fig. 1.3 Profiled deviance, on the scale $|\zeta|$, the square root of the change in the deviance, for each of the parameters in model `fm1ML`. The intervals shown are 50%, 80%, 90%, 95% and 99% confidence intervals based on the profile likelihood.

estimates. The maximum likelihood estimates are the parameter values that minimize the deviance, $d(\sigma_1,\sigma,\beta_0)$, over all feasible values of the parameters.

To perform a hypothesis test of the form $H_0 : \sigma_1 = 80$ versus $H_a : \sigma_1 \neq 80$ we could consider the minimum deviance achieved when $\sigma_1 = 80$ minus the global minimum deviance. This difference,

$$\min_{\sigma,\beta_0} d(80,\sigma,\beta_0) - d(\widehat{\sigma}_1,\widehat{\sigma},\widehat{\beta}_0),$$

is the likelihood ratio test statistic. (The deviance is a multiple of the logarithm of the likelihood so ratios of likelihoods correspond to differences in the deviance.) This test statistic explicitly compares how well we can fit the model for this specific value of one of the parameters to the best possible fit of the model. Note that this quantity is based on two model fits: the unconstrained fit and a second fit subject to the constraint $\sigma_1 = 80$. We are not basing our test on how well we think we should be able to fit the constrained model — based on the global parameter estimates and approximate standard errors — we are actually re-fitting the model subject to the constraint.

The theory of likelihood ratio tests states that the test statistic would be compared to the quantiles of the χ^2_k distribution where k is the number of constraints, which is 1 in this case. But a χ^2_1 distribution is simply the square of the standard normal distribution, $\mathscr{N}(0,1)$, so we plot its square root, which we denote as

$$|\zeta(\sigma_1)| = \sqrt{\min_{\sigma,\beta_0} d(\sigma_1,\sigma,\beta_0) - d(\widehat{\sigma}_1,\widehat{\sigma},\widehat{\beta}_0)}, \tag{1.1}$$

with similar definitions for the other parameters. (The perceptive reader will have notice that the second profile plot is on the scale of $\log(\sigma)$, not σ. We will discuss this choice of scale later.)

The intervals shown on the plots are 50%, 80%, 90%, 95% and 99% confidence intervals based on this likelihood ratio test. That is, the bounds on

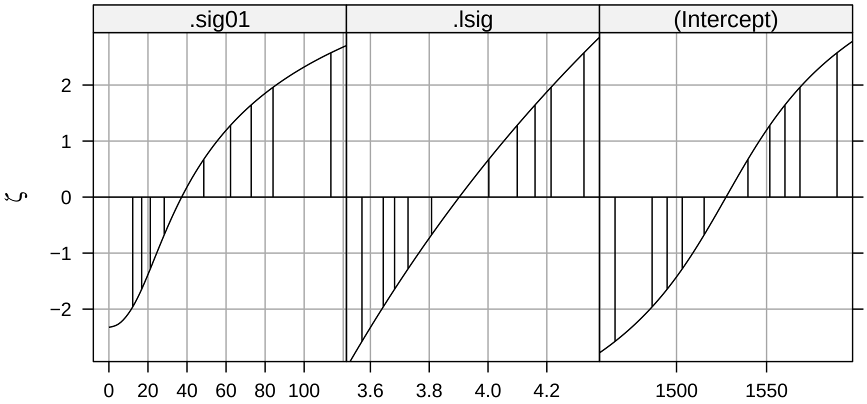


Fig. 1.4 Signed square root, ζ, of the likelihood ratio test statistic for each of the parameters in model `fm1ML`. The vertical lines are the endpoints of 50%, 80%, 90%, 95% and 99% confidence intervals derived from this test statistic.

the 95% confidence intervals are the points where $|\zeta| = 1.960$. We use the R function `profile` to create the profiled deviance object and `xyplot` to display the plot.

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

The `confint` extractor is used to obtain the endpoints of a particular confidence interval.

```
> confint(pr1)
```

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

Often the *signed square root* provides a more meaningful scale on which to plot the likelihood ratio test statistic, as in Figure˜1.4. The signed square root assigns a negative sign to ζ for parameter values less than the estimate and a positive sign for parameter values greater than the estimate providing a curve whose interpretation is similar to that of a normal probability plot of data or residuals.

1.3.2.1 Interpretation of the Shape of the ζ Plot

We will use ϕ to represent a parameter or a transformation of a parameter in the model. For the model we are considering ϕ can be σ_1 or σ_1^2 or $\log(\sigma)$ or σ or σ^2 or β_0 or one of many other possibilities. In a plot of $\zeta(\phi)$ versus ϕ the estimate, $\widehat{\phi}$, is the point where $\zeta = 0$ and the slope of the curve at $\widehat{\phi}$ is the inverse of the standard error of ϕ. There is more than one way of calculating a standard error of a parameter estimate and in chapter~?? we will clarify that this standard error is based on a quantity called "the observed information matrix". That level of detail is not necessary here. All we need to know is that the inverse slope, written s_ϕ, is that standard error of the parameter estimate, which is a measure of its precision.

If the plot of ζ versus ϕ is close to a straight line over the region of interest then our $(1-\alpha)$ confidence interval on ϕ will be close to

$$\widehat{\phi} \pm z_{\alpha/2} s_\phi \tag{1.2}$$

where $z_{\alpha/2}$ is the $\alpha/2$ upper quantile of the standard normal (also called Gaussian) distribution. For example, $z_{0.025} = 1.960$ so a 95% confidence interval on the parameter ϕ would be calculated as

$$\widehat{\phi} \pm 1.960 s_\phi \tag{1.3}$$

or, effectively, the estimate plus-or-minus two standard errors of the estimate.

We should realize that confidence intervals of the form (1.2) are based on an approximation. Similarly, performing a hypothesis test of the form $H_0 : \phi = \phi_0$ vs. $H_a : \phi \neq \phi_0$ by calculating the observed "z-statistic"

$$z_{\text{obs}} = \frac{\widehat{\phi} - \phi_0}{s_\phi} \tag{1.4}$$

and evaluating a p-value as $2 \cdot \mathrm{P}[\mathscr{Z} > |z_{\text{obs}}|]$ is also an approximation.

Of course, we don't need to perform hypothesis tests of the form $H_0 : \phi = \phi_0$ *vs.* $H_a : \phi \neq \phi_0$ in this way because can fit the model without any constraints on the parameters then fit the model subject to the constraint $\phi = \phi_0$ and compare the two model fits using the likelihood ratio. Furthermore the confidence intervals we obtain are based on "inverting" such hypothesis tests. Because we are fitting the model under both conditions and comparing the quality of the fits we obtain more realistic information regarding whether the parameter value ϕ_0 is reasonable. We do not need to resort to fitting the model once only and keeping our fingers crossed while using approximations to perform statistical inference.

This situation is unfortunately characteristic of modern statistics. The well-known statistical methods shown in many texts are often based on unnecessary approximations. We have the ability to fit models many times subject to different constraints on the parameters so we can actually examine the effective variability in the parameter estimates rather than resorting to approximations.

Returning to consideration of confidence intervals of the form (1.2) and hypothesis tests based on (1.4), these will be suitable methods when the profile plot (the plot of ζ versus ϕ) is reasonably straight. In those cases, of course, the methods based on ζ will produce confidence intervals or p-values for hypothesis tests of the expected form. The important cases are when the profile plot is not reasonably straight over the region of interest. In those cases the confidence intervals are not symmetric or not determined from quantiles of the standard normal nor are p-values based on the standard normal distribution.

As shown above, we can create confidence intervals on parameters or transformed parameter values using the profile and the `confint` extractor function so we don't really need to examine the profile plot. However, the profile plots do provide us with a valuable visual evaluation of the precision of the parameter estimates and we will use them frequently.

We can read a profile plot similar to the way that we read a quantile-quantile plot, such as a normal probability plot, of observed data or of residuals from a fitted model. That is, we evaluate a profile plot for skewness and for over-dispersion. We will speak of these properties as applying to the distribution of the parameter estimator. Technically that is not quite correct (the profile plot indicates the sensitivity of the model fit to the parameter values) but I find it a convenient way of discussing patterns in the profile plots.

With this in mind we can examine the patterns in Figure˜1.4. The profile plot for $\log(\sigma)$ is quite straight indicating that the local approximation based on a standard error is quite good. Notice, however, that this pattern is with respect to $\log(\sigma)$, not σ or, even worse, σ^2 see that the sigmoidal (i.e. like an elongated "S" curve) pattern for the profile plot of β_0 in Figure˜1.4 the parameter β_0 is very close to symmetric about the estimate $\widehat{\beta_0}$ but over-dispersed relative to a normal distribution.

1.3.2.2 Precision of the Variance Components

When we summarize our knowledge of a parameter in a fitted model by giving its estimate and a standard error of this estimate we are assuming that the quality of fit for different values of this parameter, as measured by the signed square root, η, say, will be symmetric about the estimate. Furthermore, if we are to use a z test" or confidence intervals based on quantiles of the standard normal, then we are assuming that the plot of ζ will be reasonably close to a straight line. In that case the standard error of the parameter corresponds to the inverse of the slope of the line.

We can see that the only panel in Figure˜1.4 that is at all close to a straight line is the middle panel, which shows ζ versus $\log(\sigma)$, not ζ versus σ or, even worse, ζ versus σ^2, as shown in Figure˜**??**.

1.3.3 Profile Pairs Plots

We can use the information from the profiled deviance to produce approximate

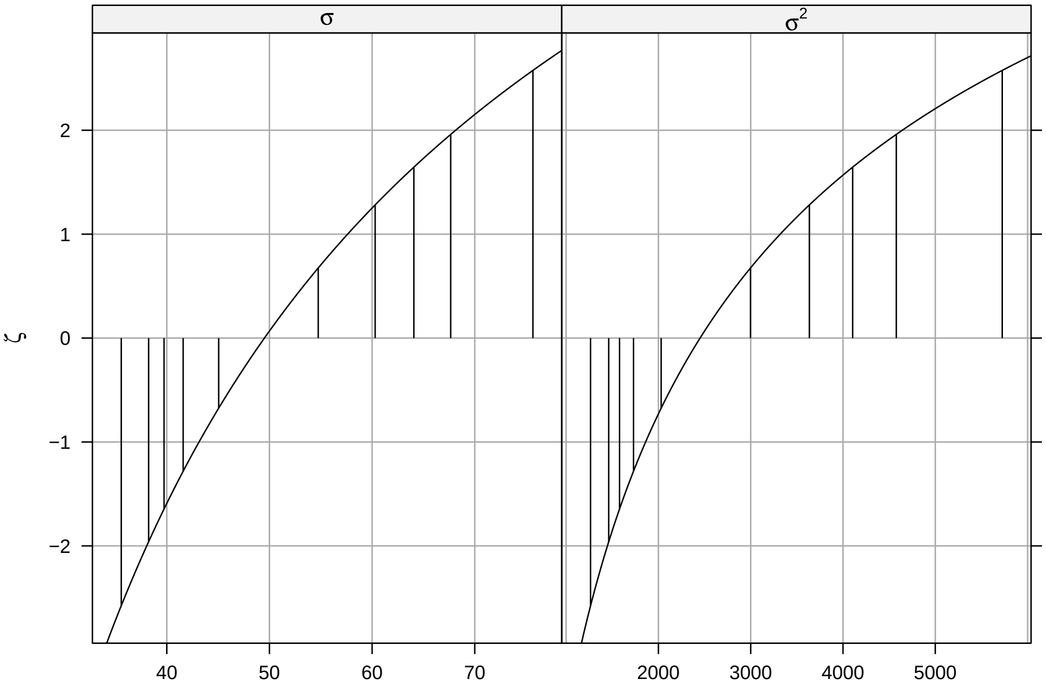


Fig. 1.5 Signed square root, ζ, of the likelihood ratio test statistic as a function of σ and as a function of σ^2. The vertical lines are the endpoints of 50%, 80%, 90%, 95% and 99% confidence intervals.

contours of the profiled deviance with respect to pairs of parameters. Such a plot is shown in Figure˜1.6. Because we will use such plots throughout this book to examine the pairwise dependence of parameters, we will describe this plot in some detail.

The labels along the diagonal indicate the parameter shown on a particular axis. Each parameter is represented on two scales, the original scale and the transformation determined by ζ. Panels above the diagonal are on the original scale; those below the diagonal are on the (ζ_i, ζ_j) scale. There are two lines drawn on each panel. These lines intersect at the estimated parameter values.

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. In the first part of this section we define several terms and

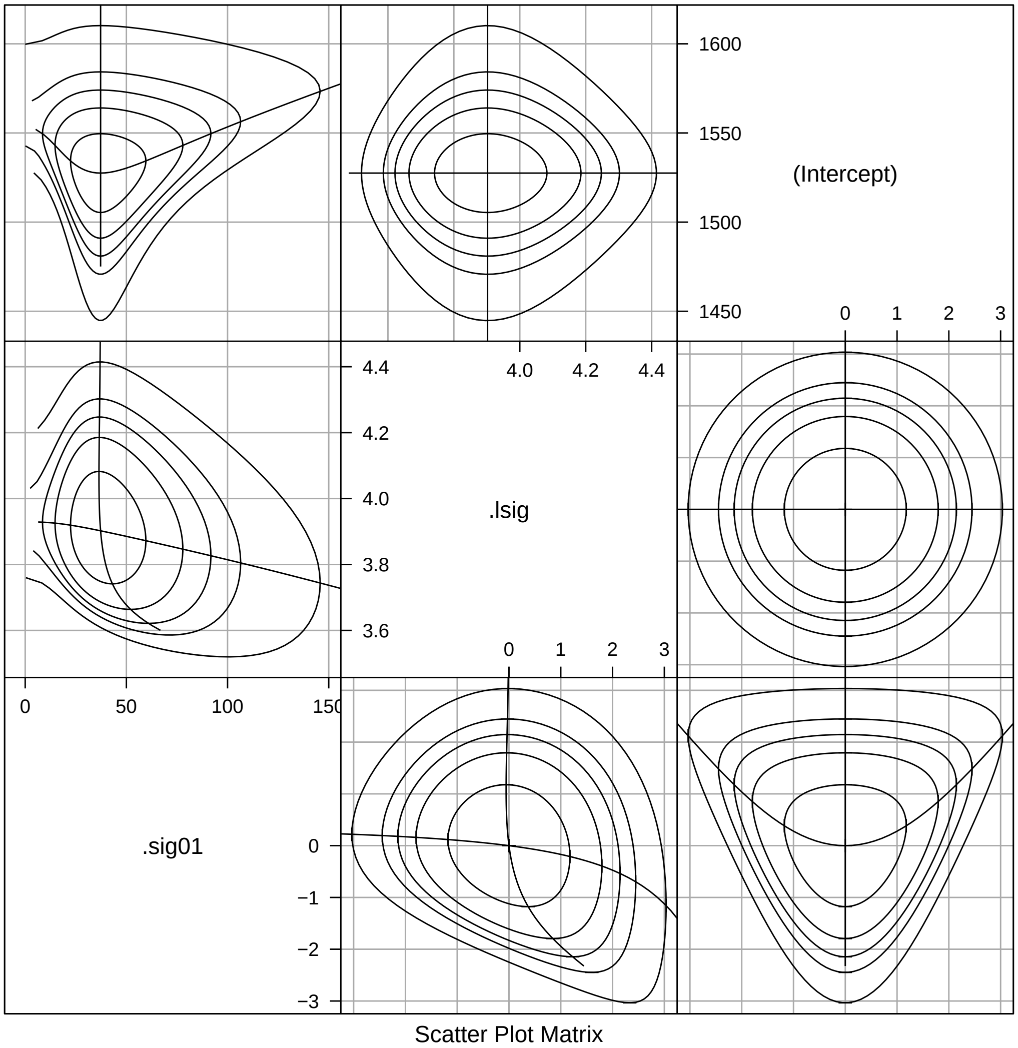


Fig. 1.6 Profile pairs plot for the parameters in model `fm1`. The contour lines correspond to marginal 50%, 80%, 90%, 95% and 99% marginal confidence regions based on the likelihood ratio. Panels below the diagonal represent the (ζ_i, ζ_j) parameters; those above the diagonal represent the original parameters.

concepts that will be used throughout the book. I would recommend that subsection to all readers.

1.4.1 Definitions of the Random Variables

As mentioned in section˜1.1, the mixed-effects probability model is based on two vector-valued random variables: the q-dimensional vector of random effects, $\mathscr{B}$, and the n-dimensional response vector, $\mathscr{Y}$. In our model for the

`Dyestuff` data the dimension of the response vector is $n = 30$ and the dimension of the random effects vector is $q = 6$.

In all the forms of mixed models that we will consider, the random effects vector has a multivariate normal (also called Gaussian) distribution with mean vector $\mathbf{0}$ and with a parameterized, $q \times q$, symmetric, variance-covariance matrix that we will write as $\Sigma_{\boldsymbol{\theta}}$. The notation indicates that this symmetric $q \times q$ matrix Σ depends on a parameter vector that is written as $\boldsymbol{\theta}$. Even though some of the elements of $\boldsymbol{\theta}$ may determine covariances and not variances, we shall refer to them collectively as the *variance-component parameters.*

We write this distribution as

$$\mathscr{B} \sim \mathscr{N}(\mathbf{0}, \Sigma_{\boldsymbol{\theta}}). \tag{1.5}$$

Because this distribution does not depend on the value, $\mathbf{y}$, of the random variable, $\mathscr{Y}$, we say it is the *unconditional* distribution of $\mathscr{B}$.

The probability model for the response, $\mathscr{Y}$, is most easily described by stating the conditional distribution $(\mathscr{Y}|\mathscr{B} = \mathbf{b})$, which is the distribution of the response vector, $\mathscr{Y}$, assuming that the value of the random effects vector, $\mathscr{B}$, is known to be $\mathbf{b}$. (In practice we never know the value $\mathbf{b}$ but, for the purpose of formulating the model, we will assume that we do.) For all the forms of mixed models that we will consider, $\mathbf{b}$ changes the conditional distribution of $\mathscr{Y}$ by changing the conditional mean, $\mu_{\mathscr{Y}|\mathscr{B}=\mathbf{b}}$. Furthermore, the conditional mean response depends on $\mathbf{b}$ and on the fixed-effects parameter vector β only through the *linear predictor*

$$\mathbf{Z}\mathbf{b} + \mathbf{X}\beta \tag{1.6}$$

where $\mathbf{Z}$ and $\mathbf{X}$ are known *model matrices* of the appropriate size.

This description applies to all forms of mixed models considered in this book. In the case of a linear mixed model we can be more specific about the conditional mean and the conditional distribution. For linear mixed models the conditional mean, $\mu_{\mathscr{Y}|\mathscr{B}=\mathbf{b}}$, is exactly the linear predictor, and the conditional distribution is a "spherical" Gaussian distribution. That is

$$\mu_{\mathscr{Y}|\mathscr{B}=\mathbf{b}} = \mathbf{Z}\mathbf{b} + \mathbf{X}\beta \tag{1.7}$$

and

$$(\mathscr{Y}|\mathscr{B} = \mathbf{b}) \sim \mathscr{N}\left(\mathbf{Z}\mathbf{b} + \mathbf{X}\beta, \sigma^2\mathbf{I}_n\right). \tag{1.8}$$

The expression $\mathbf{I}_n$ denotes the identity matrix of size n. This is the $n \times n$ matrix whose diagonal elements are all unity and whose off-diagonal elements are all zero. The parameter σ is called the *common scale parameter*. Its square is the variance of the residual "noise" terms that cannot be explained by other parts of the model. The name "spherical" is applied to Gaussian distributions of the form (1.8) because the contours of constant probability density are spheres centered at $\mathbf{Zb}+\mathbf{X}\beta$ in the n-dimensional response space

Because the conditional mean must be an n-dimensional vector, the model matrix $\mathbf{Z}$ must be $n \times q$ and the model matrix $\mathbf{X}$ must be $n \times p$, where p is the dimension of the fixed-effects parameter vector, β. For the model fit to the `Dyestuff` data, $p = 1$ and the matrix $\mathbf{X}$ is a 30×1 matrix, all of whose elements are unity. The fixed-effects term `1` in a model formula generates a column of ones in the fixed-effects model matrix, $\mathbf{X}$. For the model being considered here, this column of ones is the only column in $\mathbf{X}$.

The form of the random-effects model matrix, $\mathbf{Z}$, and the form of the variance-covariance matrix, $\Sigma_{\boldsymbol{\theta}}$, and the method by which $\Sigma_{\boldsymbol{\theta}}$ is determined from the value of $\boldsymbol{\theta}$ are all based on the random-effects terms in the model formula. As stated earlier, there is one random effects term, `(1 | Batch)`, in the formula for this model. Random-effects terms are those that contain the vertical bar, `"|"`, character. The `Batch` variable is the grouping factor for the random effects described 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`.

The transpose of this matrix, of dimension 6×30, is stored as a sparse matrix in the environment of the fitted model as `Zt`.

```
> env(fm1)$Zt

6 x 30 sparse Matrix of class "dgCMatrix"

A 1 1 1 1 1 . . . . . . . . . . . . . . . . . . . . . . . . .
B . . . . . 1 1 1 1 1 . . . . . . . . . . . . . . . . . . . .
C . . . . . . . . . . 1 1 1 1 1 . . . . . . . . . . . . . . .
D . . . . . . . . . . . . . . . 1 1 1 1 1 . . . . . . . . . .
E . . . . . . . . . . . . . . . . . . . . 1 1 1 1 1 . . . . .
F . . . . . . . . . . . . . . . . . . . . . . . . . 1 1 1 1 1
```

A sparse matrix is one in which most of the elements are known to be zero. These elements are represented by a `"."` in this display. The nature of an indicator matrix is such that only one element in each row of the indicator matrix, corresponding to a column of the transpose of the indicator matrix that is shown above, is non-zero. Sparse matrix methods~(8) for numerical linear algebra provide special techniques for storing and manipulating such matrices. These techniques are the basis of the numerical methods implemented in `lmer`.

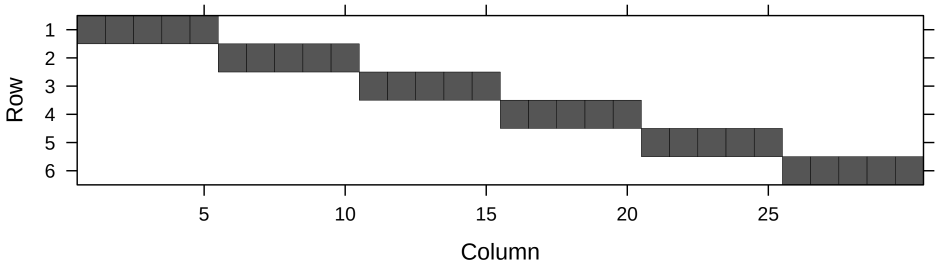


Fig. 1.7 Image of the transpose of the random-effects model matrix, $\mathbf{Z}$, for model `fm1`. The non-zero elements, which are all unity, are shown as darkened squares. The zero elements are blank.

Often we will show the structure of sparse matrices as an image like Figure˜1.7. Especially for large matrices, the image of the sparse matrix conveys its structure more compactly than does the printed representation.

When the model contains only one random-effects term and that term is a simple, scalar term, then the variance-covariance matrix Σ_θ is a non-negative multiple of $\mathbf{I}_q$, the $q \times q$ identity matrix. Although we will defer until later the discussion of the exact form of Σ_θ for other models, we will now introduce a transformation of Σ_θ that we will use throughout.

1.4.1.1 The relative covariance factor

A variance-covariance matrix like Σ_θ is required to be symmetric and, in addition, to be *positive semi-definite*, which means, in effect, that Σ_θ has the matrix equivalent of a "square root". Recall that a scalar variance, such as σ^2 in (1.8), must be non-negative and that the standard deviation, $\sigma \geq 0$, is the square root of the variance. A variance-covariance matrix has a similar property. There must be a matrix, say Λ_θ, such that when it is multiplied by its transpose it produces the original matrix, Σ_θ. This is not quite like "squaring" the matrix because the factor, Λ_θ, is multiplied by its transpose, so as to create a product that is symmetric, as Σ_θ must be.

It turns out that we can simplify many subsequent formulas if we define Λ_θ to be a multiple of the square-root factor, which we call the *relative covariance factor*, defined to satisfy

$$\Sigma_\theta = \sigma^2 \Lambda_\theta \Lambda_\theta^\mathsf{T}. \tag{1.9}$$

The symbol $^\mathsf{T}$ denotes the transpose of a matrix. The scalar σ^2 is the square of the common scale parameter introduced in (1.8).

The term "relative" indicates that Λ_θ is the factor of the variance, Σ_θ, of $\mathscr{B}$ relative to the variance, σ^2, in the conditional distribution $(\mathscr{Y}|\mathscr{B} = \mathbf{b})$. It happens that the way that we generate Λ_θ it ends up being lower triangular and corresponds to the left Cholesky factor of Σ_θ/σ^2. A left Cholesky factor is often written $\mathbf{L}$; because this matrix contains parameter values we denote it with the corresponding Greek letter, Λ.

For the `Dyestuff` model, which has only one simple, scalar random-effects term, Σ_θ and Λ_θ are both multiples of $\mathbf{I}_6$, the identity matrix of size six. Furthermore, θ is one-dimensional so we will write it as θ. We set

$$\Lambda_\theta = \theta \mathbf{I}_6 \tag{1.10}$$

subject to the constraint that $\theta \geq 0$.

The latter part of this section covers some of the finicky but important details of the matrix representation of the model and the computational methods employed. Readers who would prefer to avoid exposure to such details should feel free to do so.

1.4.2 A penalized least squares problem

The general form of a linear mixed model incorporates two parameter vectors, β and θ, and one scalar parameter, σ^2. To evaluate the maximum likelihood estimates, $\widehat{\beta}$, $\widehat{\theta}$ and $\widehat{\sigma^2}$, we could attempt to optimize the likelihood of the parameters, given the model and the observed data, with respect to all of these parameters simultaneously. However, for this particular model we can simplify the optimization problem because, for any value of θ, we can determine the conditional estimates, $\widehat{\beta}_\theta$ and $\widehat{\sigma^2}_\theta$, of the other parameters in a direct computation.

Furthermore, this computation, determining the solution of the penalized linear least squares problem

$$\min_{\mathbf{u},\beta} \left\| \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{Z}\Lambda_\theta & \mathbf{X} \\ \mathbf{I}_q & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \beta \end{bmatrix} \right\|^2 = \min_{\mathbf{u},\beta} \left(\|\mathbf{y} - \mathbf{Z}\Lambda_\theta \mathbf{u} - \mathbf{X}\beta\|^2 + \|\mathbf{u}\|^2 \right) \tag{1.11}$$

as the vectors, $\widehat{\beta}_\theta$ and $\tilde{\mathbf{u}}_\theta$, that satisfy

$$\begin{bmatrix} \Lambda_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda_\theta + \mathbf{I}_q & \Lambda_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{X} \\ \mathbf{X}^\mathsf{T}\mathbf{Z}\Lambda_\theta & \mathbf{X}^\mathsf{T}\mathbf{X} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{u}} \\ \widehat{\beta}_\theta \end{bmatrix} = \begin{bmatrix} \Lambda_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{y} \\ \mathbf{X}^\mathsf{T}\mathbf{y} \end{bmatrix}, \tag{1.12}$$

provides, as we have seen, the conditional mean, $\mu_{\mathscr{B}|\mathbf{Y}=\mathbf{y}}$, and conditional variance of the random effects, $\mathscr{B}$, given the observed data, $\mathscr{Y} = \mathbf{y}$, plus a concise expression for the log-likelihood and the REML criterion.

An effective way of determining the solution to (1.12) is to create the Cholesky decomposition of the matrix on the left. In considering this step, however, we must bear in mind that the matrix $\mathbf{Z}$ is quite sparse and, in some cases, can be very large. I have myself fit linear mixed models in which both the number of rows and the number of columns in $\mathbf{Z}$ exceeded one million.

Dealing with matrices of this size requires considerable care in deciding how to structure the computation. The key to fitting mixed models with a complex structure to very large data sets using the `lme4` package is the ability to determine the sparse Cholesky factor of $\boldsymbol{\Lambda}_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\boldsymbol{\Lambda}_\theta + \mathbf{I}_q$ for many different values of $\boldsymbol{\theta}$.

The Cholesky decomposition of large, sparse matrices like this has the interesting characteristic that reordering the rows and columns of the matrix can change, sometimes dramatically, the number of nonzero elements in the factor. Decreasing the number of nonzeros in the factor will increase the speed of the decomposition, which is important when this needs to be done for many values of $\boldsymbol{\theta}$. We will write a *fill-reducing permutation*, determined in this case by the Approximate Minimal Degree algorithm˜(6), as a $q \times q$ permutation matrix, $\mathbf{P}$. (In practice it is only the permutation, which is a reordering of the numbers 1 to q, that is created and stored but it is convenient to represent this permutation as a matrix in the formulas.) The permutation depends only on the positions of the nonzeros in the matrix $\mathbf{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta}_0)$ where $\boldsymbol{\theta}_0$ is an initial estimate or "starting value" for the parameter $\boldsymbol{\theta}$.

To allow for later generalizations of the linear mixed model we write the decomposition as

$$\begin{bmatrix}\mathbf{L}_\theta & \mathbf{0}\\ \mathbf{R}_{ZX}^\mathsf{T} & \mathbf{R}_X^\mathsf{T}\end{bmatrix}\begin{bmatrix}\mathbf{L}_\theta^\mathsf{T} & \mathbf{R}_{ZX}\\ \mathbf{0} & \mathbf{R}_X\end{bmatrix} = \begin{bmatrix}\mathbf{U}_\theta^\mathsf{T}\mathbf{U}_\theta + \mathbf{I}_q & \mathbf{U}_\theta^\mathsf{T}\mathbf{V}\\ \mathbf{V}^\mathsf{T}\mathbf{U}_\theta & \mathbf{V}^\mathsf{T}\mathbf{V}\end{bmatrix} \tag{1.13}$$

where

$$\mathbf{U}_\theta = \mathbf{Z}\boldsymbol{\Lambda}_\theta \text{ and } \mathbf{V} = \mathbf{X}. \tag{1.14}$$

The solution to (1.12), which would now be written as,

$$\begin{bmatrix}\mathbf{P}^\mathsf{T}\mathbf{L}_\theta & \mathbf{0}\\ \mathbf{R}_{ZX}^\mathsf{T} & \mathbf{R}_X^\mathsf{T}\end{bmatrix}\begin{bmatrix}\mathbf{L}_\theta^\mathsf{T}\mathbf{P} & \mathbf{R}_{ZX}\\ \mathbf{0} & \mathbf{R}_X\end{bmatrix} = \begin{bmatrix}\tilde{\mathbf{u}}_\theta\\ \widehat{\boldsymbol{\beta}}_\theta\end{bmatrix} = \begin{bmatrix}\mathbf{U}_\theta^\mathsf{T}\mathbf{y}\\ \mathbf{V}^\mathsf{T}\mathbf{y}\end{bmatrix}, \tag{1.15}$$

produces both the conditional estimate, $\widehat{\boldsymbol{\beta}}_\theta$, of the fixed-effects parameters and the conditional mean $\boldsymbol{\mu}_{\mathscr{B}|\mathscr{Y}=\mathbf{y}} = \boldsymbol{\Lambda}_\theta\tilde{\mathbf{u}}$, of the random effects. If r_θ^2 is the minimum penalized residual sum of squares,

$$r_\theta^2 = \|\mathbf{y} - \mathbf{U}_\theta\tilde{\mathbf{u}}_\theta - \mathbf{V}\widehat{\boldsymbol{\beta}}_\theta\|^2 + \|\tilde{\mathbf{u}}_\theta\|^2 \tag{1.16}$$

then the conditional maximum likelihood estimate of σ^2 is $\widehat{\sigma^2}_\theta = \frac{r^2}{n}$ and the conditional REML estimate of σ^2 is $\widehat{\sigma^2}_R(\theta) = \frac{r^2}{n-p}$.

Although we have not written this explicitly, the matrices $\mathbf{L}_\theta$, $\mathbf{R}_{ZX}$ and $\mathbf{R}_X$ all depend on the value of $\boldsymbol{\theta}$. Both the $q \times q$ matrix $\mathbf{L}_\theta$ and the $p \times p$ matrix $\mathbf{R}_X(\boldsymbol{\theta})$ are triangular: $\mathbf{L}_\theta$ is lower triangular and $\mathbf{R}_X(\boldsymbol{\theta})$ is upper triangular. Because these matrices are square and triangular their determinants, written $|\mathbf{L}|$ and $|\mathbf{R}_X|$, are easily evaluated as the product of the diagonal elements in the matrix. The determinant of a matrix is the measure of the size of the matrix that is used in evaluating probability densities of random variables after transformation.

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. 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. In the next chapter we consider models with random-effects terms that are more complex than a simple, scalar term.

2.1 Data with crossed grouping factors

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 set of subjects to each of a group of stimuli or items is measured. If the subjects are considered to be a sample from a particular population and the items are a sample from a population 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 (5) 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 (Figure~2.1).

The variation in the diameter is associated with the plates and with the samples. Because each plate is used for only the six samples shown here we will use random effects for the plate. 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. Hence we will use random effects for the sample too.

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 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 Figure~2.1 and, because there are a moderate number of levels in these factors, we can check it in a cross-tabulation

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

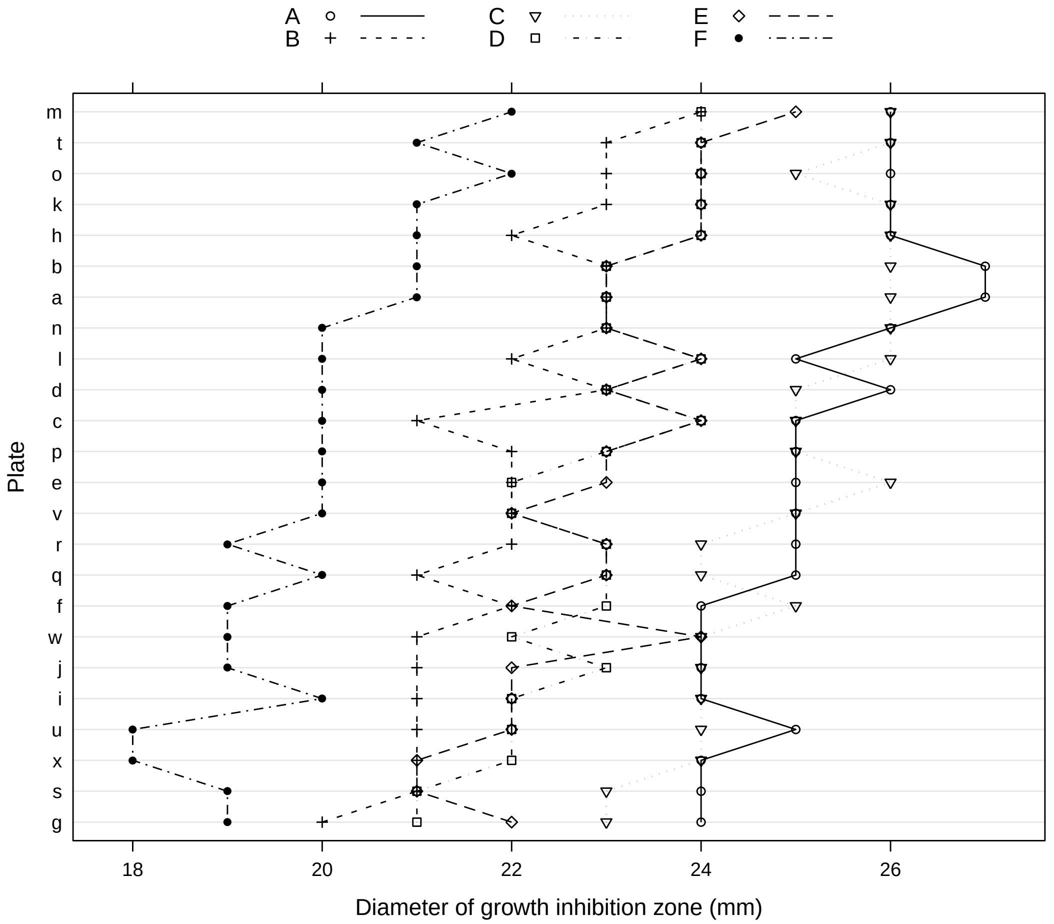


Fig. 2.1 Diameter of the growth inhibition zone (mm) in the *B. subtilis* method of assessing the concentration of penicillin. Each of 6 samples was applied to each of the 24 agar plates. The lines join observations on the same sample.

```
      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 (Figure~2.1).

The prediction intervals on the random effects (Figure~2.2) show that the conditional distribution of the random effects for `sample` has much less variability than does the conditional distribution of the random effects for `sample`.

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 all the sets of indicator columns are concatenated

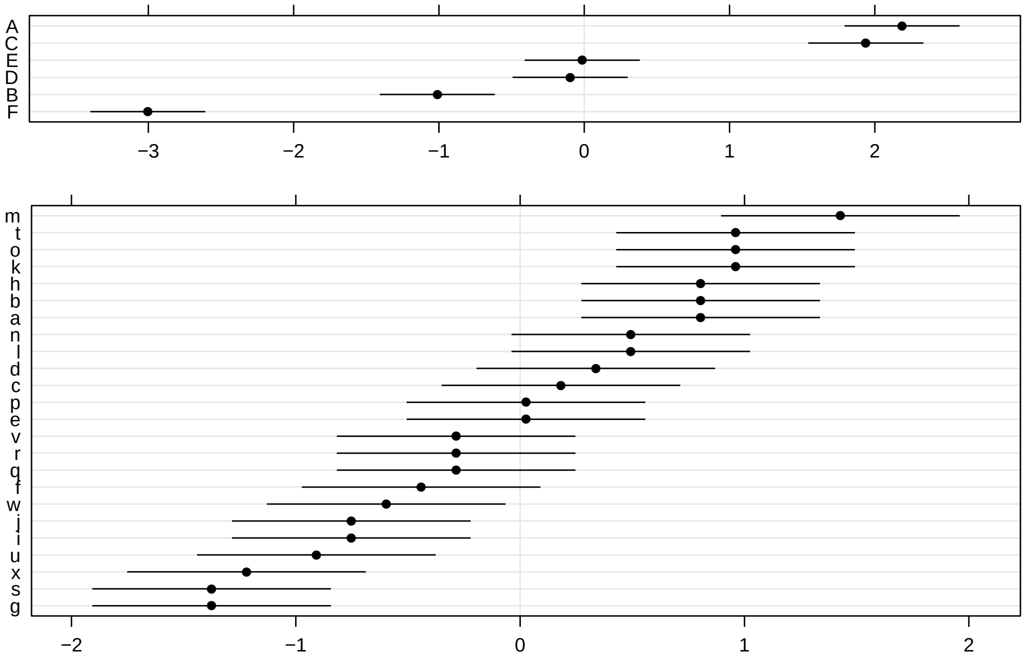


Fig. 2.2 95% prediction intervals on the random effects for model `fm2` fit to the `Penicillin` data.

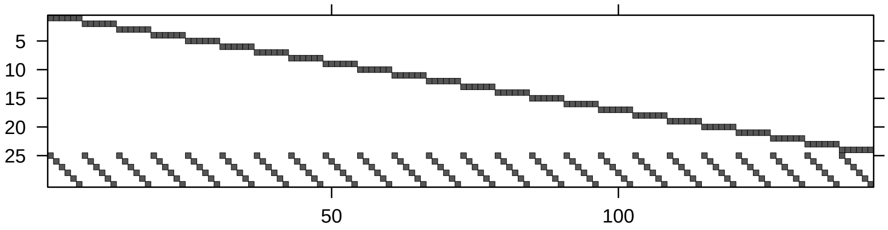


Fig. 2.3 Image of the transpose of the random-effects model matrix, **Z**, for model `fm1`. The non-zero elements, which are all unity, are shown as darkened squares. The zero elements are blank.

to for the model matrix **Z**. The transpose of this matrix contains rows of indicators for each factor, as shown in Figure˜2.3.

The parameter $\boldsymbol{\theta}$ for this model is two-dimensional.

```
> fm2@getPars()
```

```
[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) at convergence, and the second is the relative standard deviation of the random effects for `sample` (1.93157/0.54992).

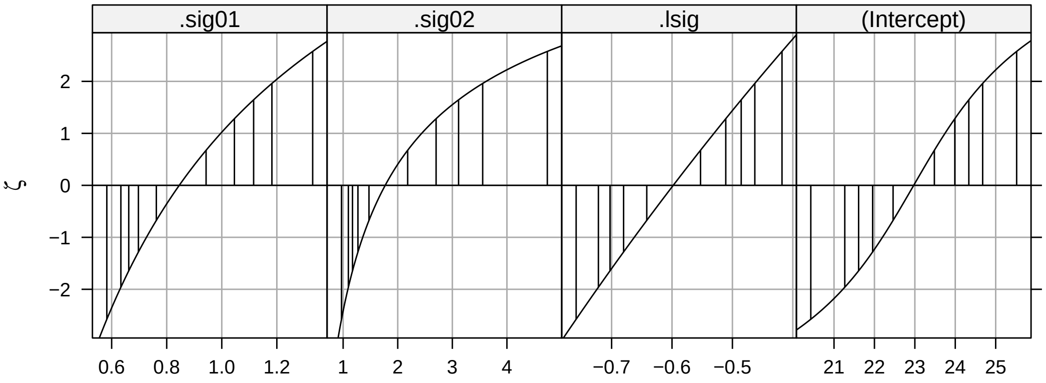


Fig. 2.4 Profile plot of the parameters in model `fm2`.

A profile plot of the parameters in model `fm2` is shown in Figure˜2.4 and the profile pairs plot in Figure˜**??**. The contours in the profile pairs plot correspond to pairwise marginal confidence regions with confidence levels 50%, 80%, 90%, 95% and 99%.

2.1.3 The `Pastes` *data*

The third example from (5), 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 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
```

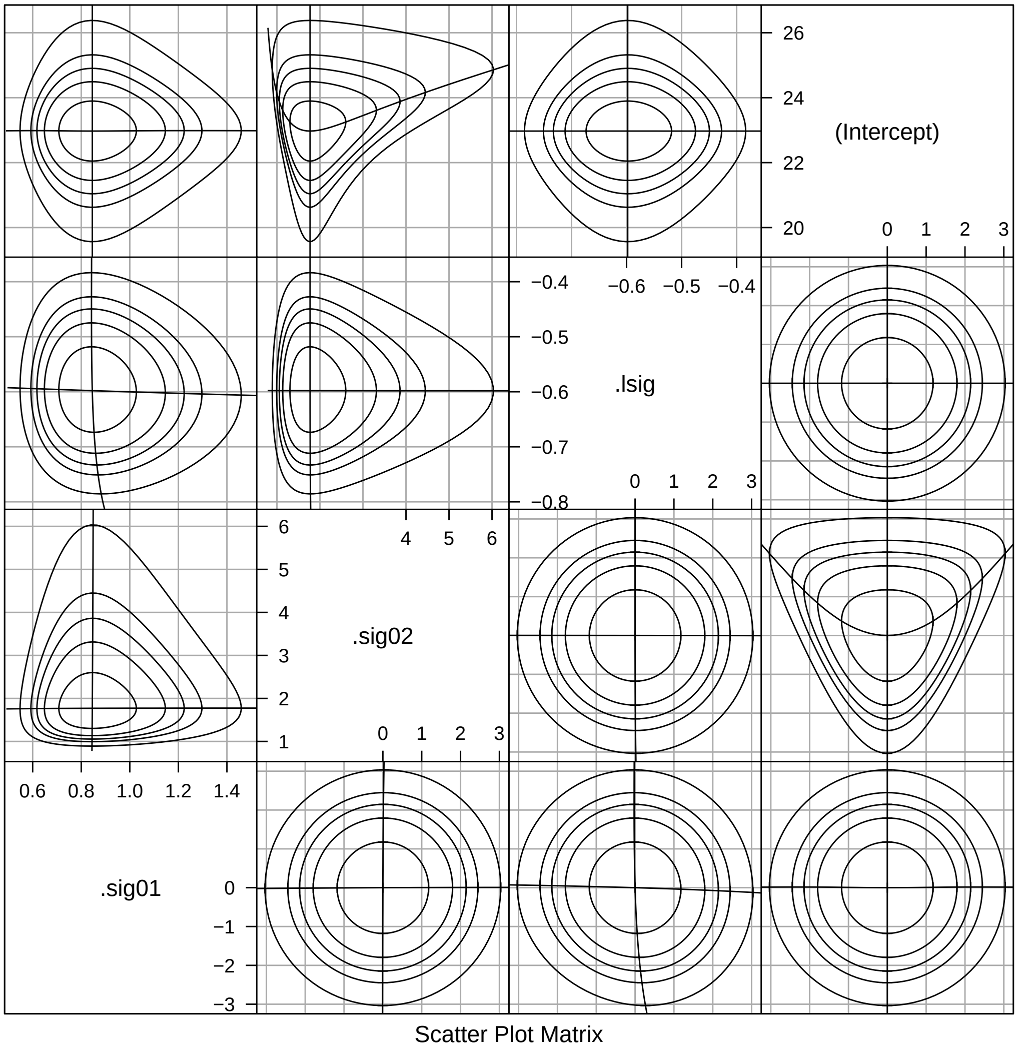


Fig. 2.5 Profile pairs plot of the parameters in model `fm2`.

```
 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 (5), 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 . . . . . . . . . . . . . . . . . . . . . . . .
```

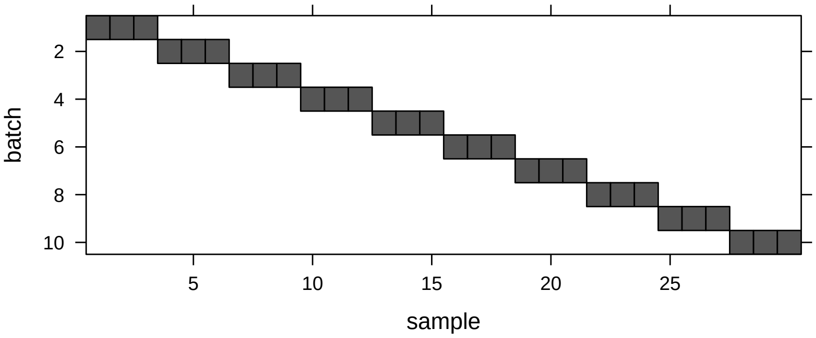


Fig. 2.6 Image of the cross-tabulation of the `batch` and `sample` factors in the `Pastes` data.

```
C . . . . . . 2 2 2 . . . . . . . . . . . . . . . . . . . . .
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 (Figure˜2.6) of this cross-tabulation to visualize the structure. Images like this are often an effective way of visualizing the structure of large, sparse matrices.

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 Figure˜2.7 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 Figures˜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.

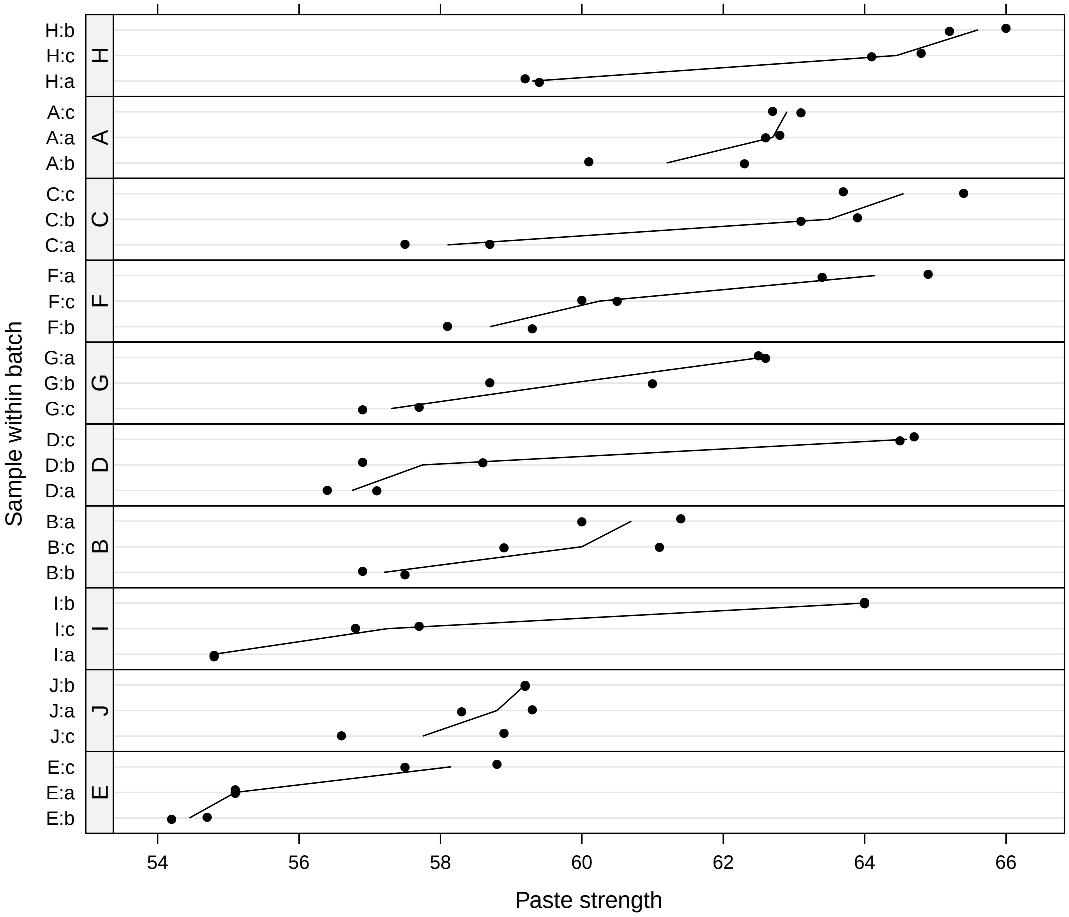


Fig. 2.7 Strength of paste preparations according to the `batch` and the `sample` within the batch. There were two strength measurements on each of the 30 samples; three samples each from 10 batches.

In Figure˜2.7 we order the samples within each batch separately then order the batches according to increasing mean strength.

Figure˜2.7 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, each of batches I and D, with low mean strength relative to the other batches, 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 section **??** we will examine the need for incorporating batch-to-batch variability in a statistical model in addition to sample-to-sample variability.

2.1.3.1 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*~(10) or *hierarchical linear models*~(11), 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 drop unused levels of the interaction factor in the process of creating it. (An "unused level" is a combination that did not occur in the data.) A convenient code idiom is

```
> Pastes$sample <- with(Pastes, (batch:cask)[drop = TRUE])
```

or

```
> Pastes <- within(Pastes, sample <- (batch:cask)[drop = TRUE])
```

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.

2.1.4 Fitting a model with random-effects for nested factors

```
> (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
```

A profile plot of the parameters in model `fm3` is shown in Figure~2.8 and the profile pairs plot in Figure~**??**. The contours in the profile pairs plot

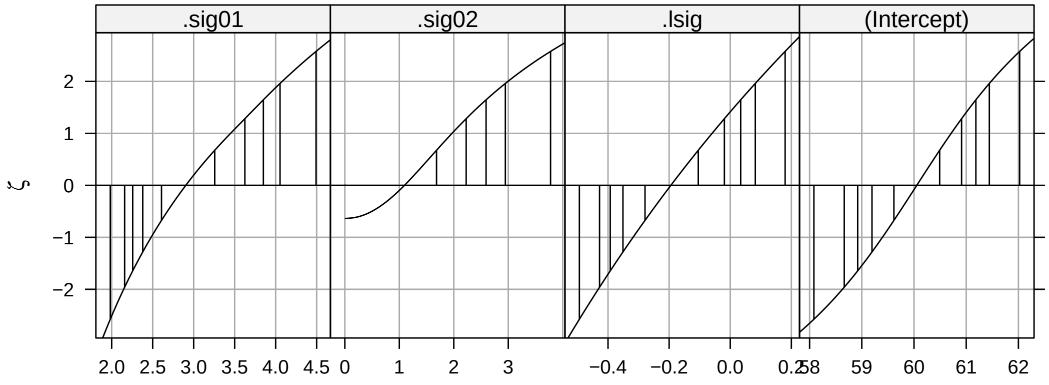


Fig. 2.8 Profile plot of the parameters in model `fm3`.

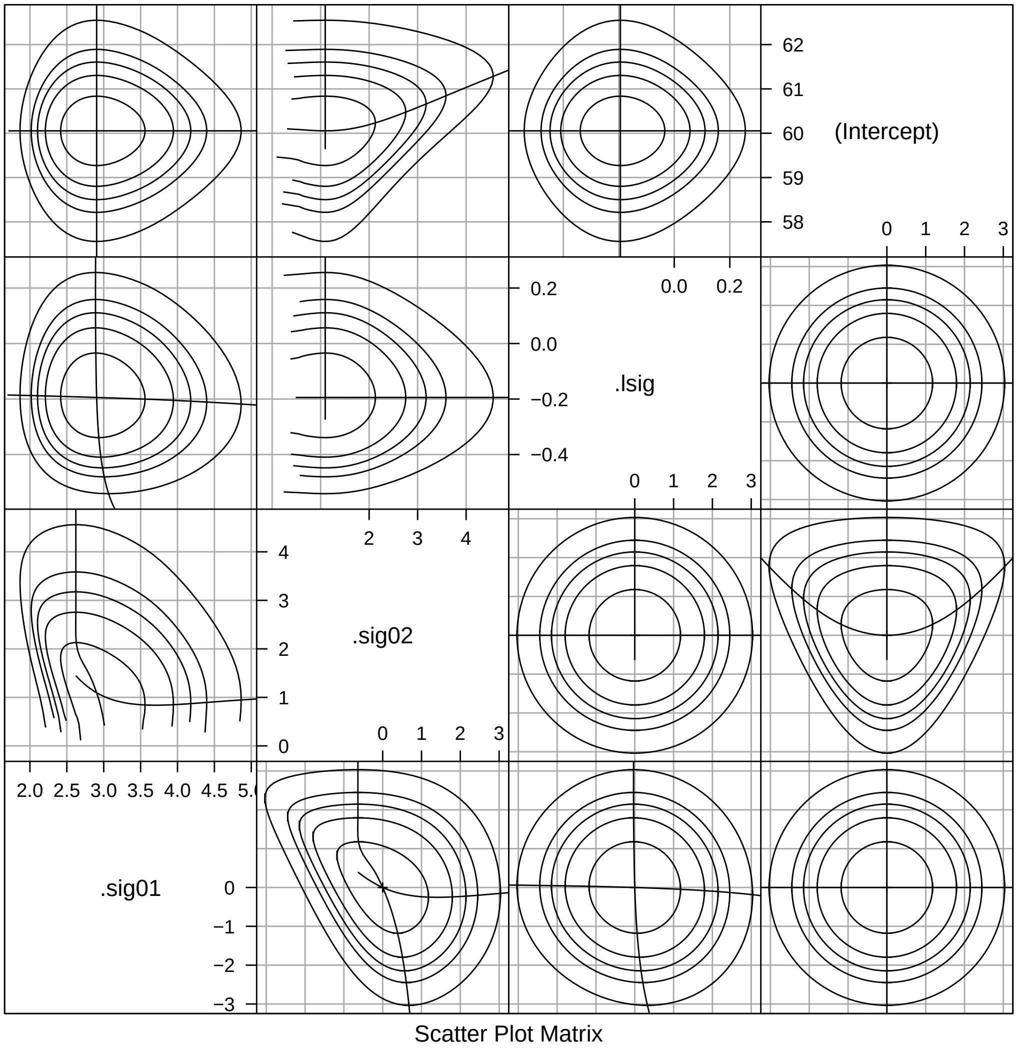


Fig. 2.9 Profile pairs plot of the parameters in model `fm3`.

correspond to pairwise marginal confidence regions with confidence levels 50%, 80%, 90%, 95% and 99%.

2.2 Models incorporating covariates

```
> (fm4 <- lmer(mathgain ~ I(mathkind - 450) + sex + minority + ses +
+     housepov + (1 | classid) + (1 | schoolid), classroom))

Linear mixed model fit by REML
Formula: mathgain ~ I(mathkind - 450) + sex + minority + ses + housepov +      (1 | classid)
   Data: classroom
 REML
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 `fm4` is shown in Figure~2.10 and the profile pairs plot in Figure~**??**. The contours in the profile pairs plot correspond to pairwise marginal confidence regions with confidence levels 50%, 80%, 90%, 95% and 99%.

2.3 Rat Brain example

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

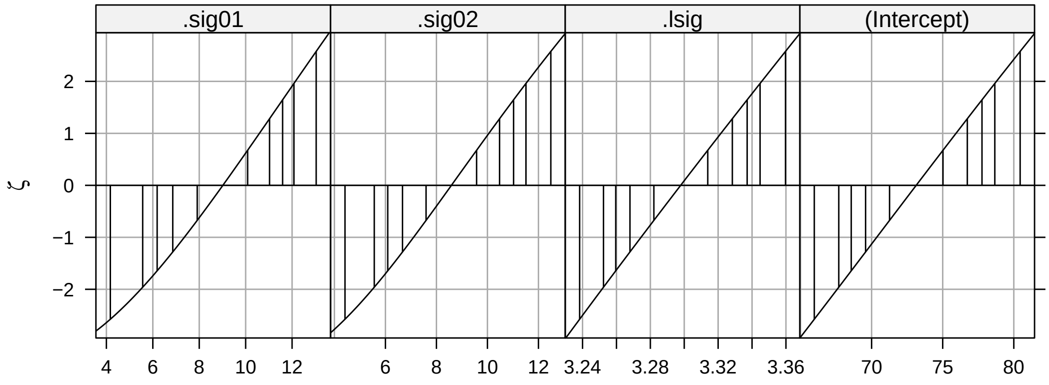


Fig. 2.10 Profile plot of the parameters in model `fm4`.

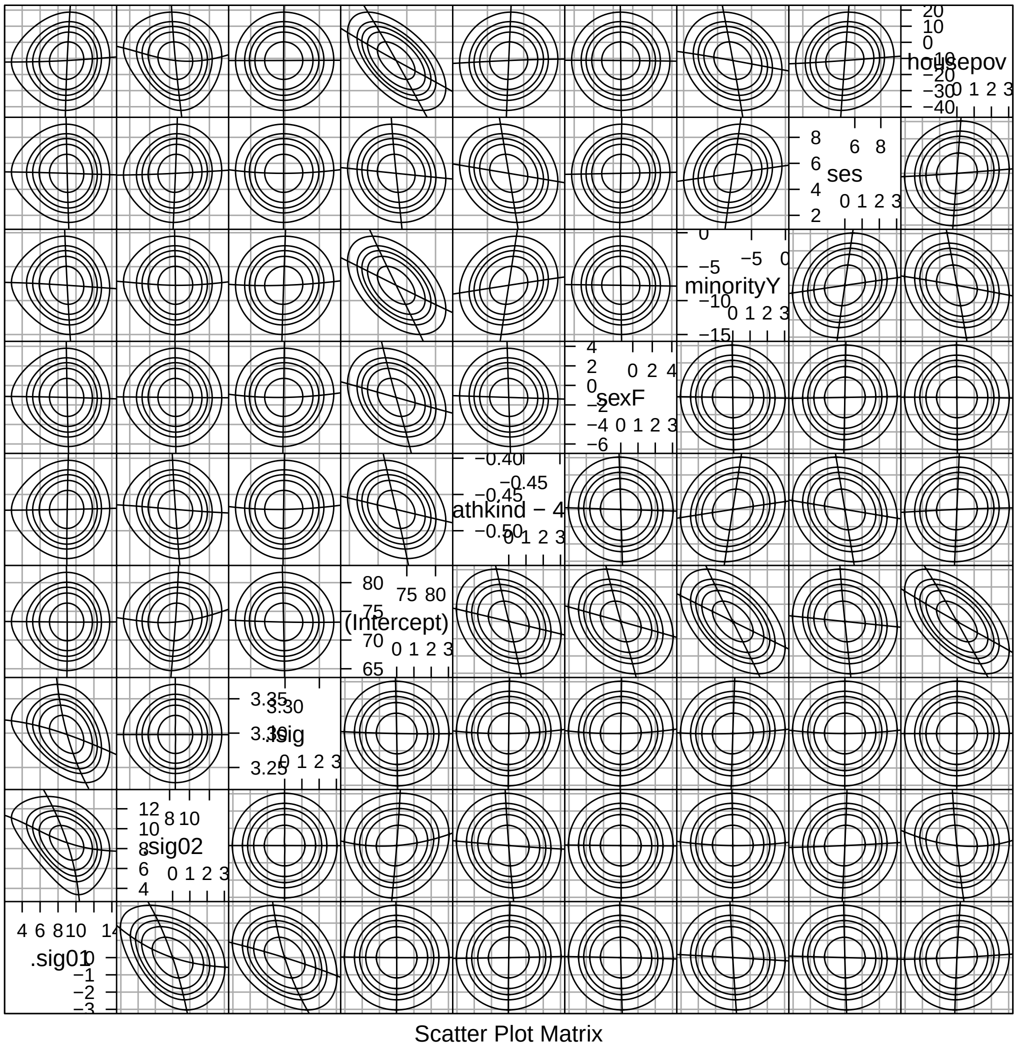


Fig. 2.11 Profile pairs plot of the parameters in model `fm4`.

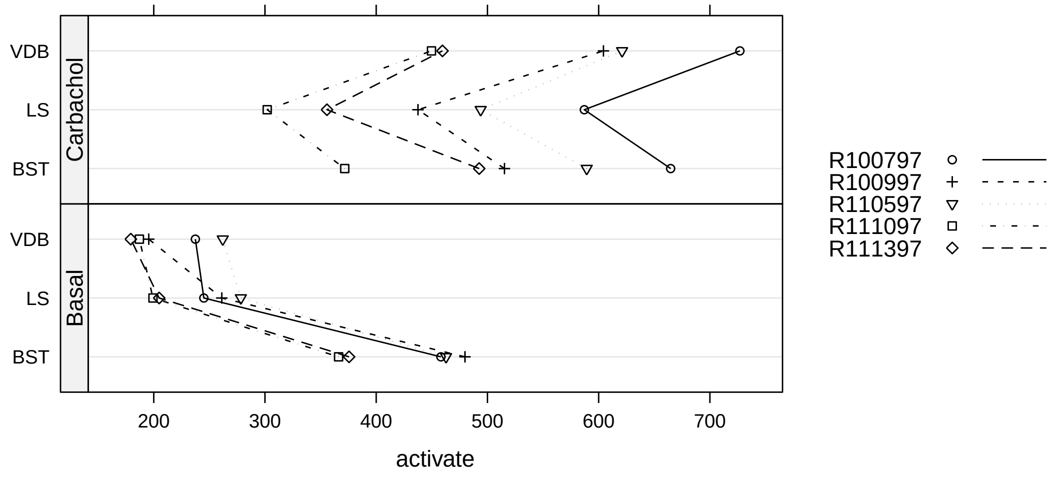


Fig. 2.12 Activation of brain regions in rats

```
                  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.

2.3.1 Structure of the formula used in `lmer`

At this point all that we need to know about the formula is that it consists of two expressions separated by a tilde ($\sim$), which is read as “is modeled by”. The expression on the left of the tilde is the response — typically just the name of a variable in the data set but more complicated expressions are possible. The expression on the right consists of one or more *terms*, separated by plus (+) operators. A random-effects term consists of two expressions separated by the vertical bar (|) operator, read as “given” or “by”. The expression on the right of the vertical bar is evaluated as a factor, called the *grouping factor* for the random effects. The random effects are associated with the levels of this factor. In this section all the models we will fit have with *simple, scalar random effects*, meaning that there is exactly one random effect associated with each level of the grouping factor and it is an additive random effect. In these cases the expression on the left of the vertical bar is “`1`”, which denotes a constant.

Chapter 3
Simple 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 relate the within- and between-subject trends to such covariates.

In this chapter we introduce graphical and statistical techniques for multilevel analysis by applying them to a simple longitudinal example. Because our purpose is to motivate the later presentation of the statistical and computational techniques we simply provide the results here and delay until later chapters the detailed explanations of how the plots are created and how the statistical models are fit to the data.

3.1 The sleepstudy data

(2) 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.

The data, available as the data frame `sleepstudy` in the `"Matrix"` package, consist of the response variable `Reaction`, which is the average of the reaction time measurements on a given subject for a given day, and two covariates: `Days`, the number of days of sleep deprivation, and `Subject`, the identifier of the subject on which the observation was made.

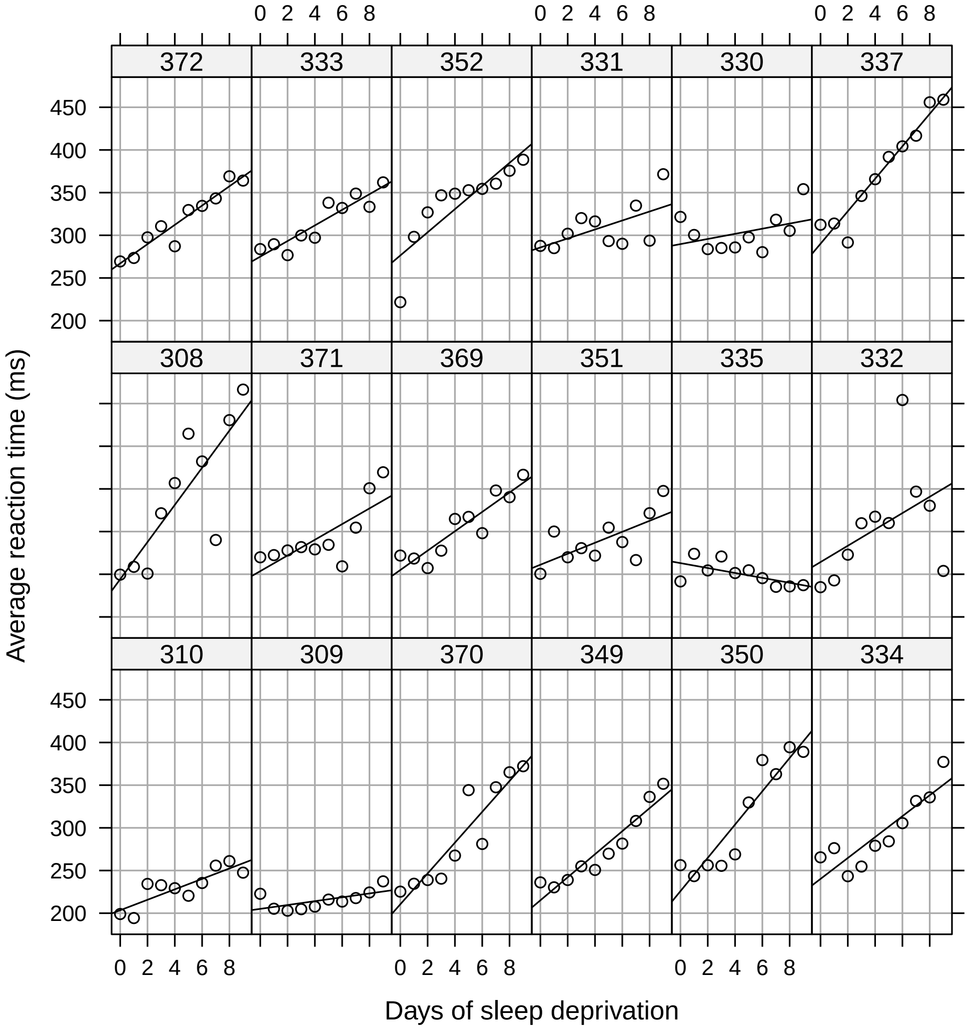


Fig. 3.1 A lattice plot of the average reaction time versus number of days of sleep deprivation by subject for the `sleepstudy` data. Each subject's data are shown in a separate panel, along with a simple linear regression line fit to the data in that panel. The panels are ordered, from left to right 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. For data such as these we use the multipanel style of plots pioneered by Bill Cleveland and his coworkers at Bell Labs as "Trellis graphics" and implemented for R in the `"lattice"` package by Deepayan Sarkar.

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 Figure~3.1. 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 Figure˜3.1 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.

In section **??** we discuss the details of producing this style of plot for longitudinal data and give other examples of such plots. Here we concentrate on some of the aspects of this particular plot; specifically, the use of a reference line, the choice of the aspect ratio and the ordering of the panels.

A simple linear regression line has been added to each panel of Figure˜3.1 to help assess the linearity of the relationship and to allow comparison of slopes and intercepts. The aspect ratio of the panels (ratio of the height to the width) has been chosen, according to an algorithm described in (4), 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 ±45°, thereby making it easier to detect systematic changes in slopes.

We can see that for all the subjects except subject 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. A question of interest to the experimenters is whether a subject's rate of increase is related to the subject's initial reaction time. To aid in assessing possible relationships between the per-subject intercept and slope, the panels in Figure˜3.1 have been ordered by increasing intercept of the within-panel regression lines, from left to right within rows starting on the bottom row. With such an ordering of the panels, a systematic relationship between the intercept and the slope would show up as a systematic trend in slopes across the panels. There is little evidence in Figure˜3.1 of such a systematic relationship between the subject's initial reaction time and their rate of increase in reaction time with days of sleep deprivation.

Another way of assessing the extent of such a relationship is to plot confidence intervals on the fitted slopes and intercepts (Figure˜3.2). In this plot we have again ordered the subjects according to the estimated intercept. This produces an obvious pattern in the intervals for the intercept. If there was a systematic relationship between the intercepts and the slopes we should also see a pattern in the slopes. Although Figure˜3.2 does show considerable variation between subjects in both the intercept and the slope, there is little evidence of a systematic relationship between the intercept and the slope.

One notable characteristic of Figure˜3.2 is the consistency of the widths of the intervals, a result of the data being balanced and of the use of a common ("pooled") estimate of the variance of the random noise term in the regression model.

We say that longitudinal data are *balanced* if each subject is observed the same number of times if the times of these observations are consistent across subjects. In a designed experiment, such as generated these data, we can reasonably expect the data to be balanced. An observational study, on the other hand, will almost never generate balanced data. Although the methods that we describe in this book apply to balanced or unbalanced data, highly

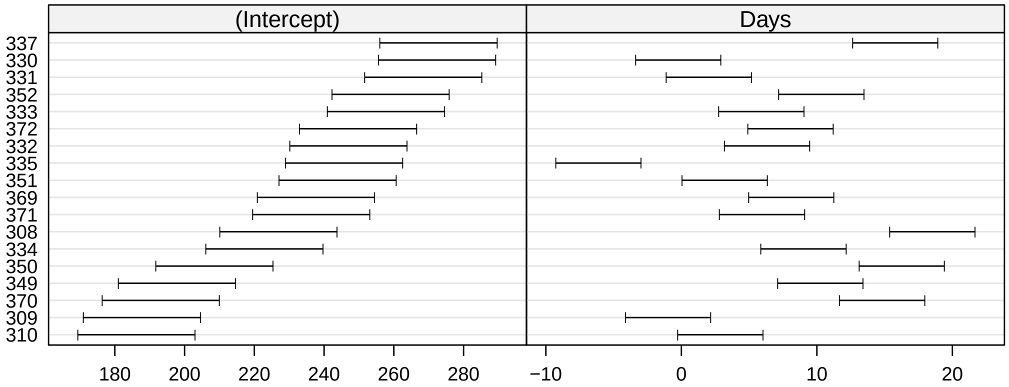


Fig. 3.2 95% confidence intervals on the intercept and slope of linear regression models fitted to the data for each subject in the sleep study. A pooled estimate of the residual standard deviation was used, resulting in a constant width of the intervals across subjects for this balanced data set. The subjects have been ordered according to increasing estimated intercept.

unbalanced data can make interpretation of the results from a fitted model more difficult, as we shall see in later chapters.

3.2 Mixed-effects models for the sleep data

Based on our preliminary graphical and analytical 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.

One way of writing such a model is

$$y_{ij} = \beta_1 + \beta_2 x_{ij} + b_{i1} + b_{i2} x_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, 18, j = 1, \ldots, 10 \tag{3.1}$$

where y_{ij} is the jth reaction time observed on subject i, taken on day x_{ij}, $\beta = (\beta_1, \beta_2)^{\mathsf{T}}$ is the vector of fixed effects, $\mathbf{b}_i = (b_{i1}, b_{i2})^{\mathsf{T}}$ is the vector of random effects for the ith subject and ε_{ij} is the unexplained or "residual" random noise associated with this observation.

We assume that the ε_{ij} are independently distributed as $\varepsilon_{ij} \sim \mathscr{N}(0, \sigma^2)$. In the first form of the mixed-effects model we assume that the $\mathbf{b}_i$ are independently distributed as $\mathbf{b}_i \sim \mathscr{N}(\mathbf{0}, \Sigma_1)$ where Σ_1 is a general 2×2 variance-covariance matrix. That is, in the first model the random effects for the intercept and the slope of the same subject are allowed to be correlated. Because our graphical evaluation of the data in §3.1 indicates that it may be reasonable to model these data with independent random effects for intercept and slope within subject, we will later fit a second model that has such independent random effects then compare the two fitted models.

When fitting mixed-effects models using `lmer` we specify the model to be fit using a formula. For the first model the formula can be written

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

indicating that the variable `Reaction` is modeled by an intercept and a slope with respect to the variable `Days` plus random effects, grouped by `Subject`, for the intercept and the slope.

That is, the left hand side of the formula, `Reaction` in this case, specifies the response. Usually the left hand side is the name of a variable, as it is here, but it can be a more general expression like `log(Response)`. Expressions on the right hand side of the formula, `1 + Days + (1 + Days|Subject)` in this case, together specify the fixed effects and the random effects formulation. A conditional expression, which is any expression that contains the vertical bar, `|`, designate random effects for the terms on the left of the `|` grouped according to the factor on the right of the `|`. Expressions on the right hand side of the formula that do not include the vertical bar specify fixed effects.

The formula as given above explicitly indicates the presence of the intercept terms. However, it is common to write a formula like this as

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

taking advantage of the fact that the formula language implicitly includes an intercept term in the fixed effects and in each random effects specification.

To specify a model with independent random effects by subject for the intercept and the slope we will use two random-effects terms and write the formula as

```
Reaction ~ Days + (1 | Subject) + (0 + Days | Subject)
```

In this formula the random effects for the intercept and the random effects for the slope are modeled as independent random variables because they are declared in different expressions. Note that to specify a random intercept given subject we must explicitly include the intercept term `1` in `(1|Subject)` because there are no other terms. Similarly in the second expression we suppress the implicit intercept term by using `(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)`.

We delay further discussion on the mathematical form of the model and the interpretation of the model formulae until §**??**. At this point let us fit these models and examine the parameter estimates.

We can fit the first model, store the result as `sm1` (sleepstudy model 1), and ask for a brief display of the results with

```
> (sm1 <- lmer(Reaction ~ Days + (Days | Subject), sleepstudy))
```

```
Linear mixed model fit by REML
Formula: Reaction ~ Days + (Days | Subject)
   Data: sleepstudy
REML
1744

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 612.091  24.7405
          Days         35.072   5.9221  0.066
 Residual             654.941  25.5918
Number of obs: 180, groups: Subject, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept)  251.405      6.825   36.84
Days          10.467      1.546    6.77

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

(The extra set of parentheses surrounding the assignment causes the fitted model to display itself. Normally the result of an assignment is not displayed.)

This brief display includes information on the criterion used to fit the model (restricted maximum likelihood or REML, see §**??** for a formal definiton), the model formula and the data to which it was fit, some information on the quality of the fit and information on the parameter estimates. For the moment we will concentrate on the parameter estimates.

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 variance-covariance matrix for the random effects is displayed by giving the variances of these random variables, the corresponding standard deviations and any estimated correlations. Note that the columns labeled `Variance` and `Std.Dev.` in this section are redundant in that each entry in the `Std.Dev.` column is simply the square root of the corresponding variance estimate. These estimates are expressed in both the variance scale and the standard deviation scale because both are useful in interpretation. (Some readers may be tempted to interpret the elements of the `Std.Dev.` column as standard errors of the variance estimates. Don't do that. These are not standard errors.)

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˜3.1 and the intervals shown in Figure˜3.2.

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 $10.5-2\cdot 6=-1.5$ ms./day or as high as $10.5+2\cdot 6=22.5$ ms./day. Again, the conclusions from these rough, "back of the envelope" calculations are consistent with our observations from Figures˜3.1 and 3.2.

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˜3.1 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, 0.066, 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.

By fitting model `sm2` with independent random effects for intercept and slope and comparing this fitted model to `sm1` we can assess this claim using a statistical hypothesis test.

```
> (sm2 <- lmer(Reaction ~ Days + (1 | Subject) + (0 + Days | Subject),
+     sleepstudy))
```

```
Linear mixed model fit by REML
Formula: Reaction ~ Days + (1 | Subject) + (0 + Days | Subject)
   Data: sleepstudy
REML
1744

Random effects:
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 627.569  25.0513
 Subject  Days         35.858   5.9882
 Residual             653.584  25.5653
Number of obs: 180, groups: Subject, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept)  251.405      6.885   36.51
Days          10.467      1.560    6.71

Correlation of Fixed Effects:
     (Intr)
Days -0.184
```

```
> anova(sm2, sm1)
```

```
Data: sleepstudy
Models:
```

```
sm2: Reaction ~ Days + (1 | Subject) + (0 + Days | Subject)
sm1: Reaction ~ Days + (Days | Subject)
    Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
sm2  5 1762.0 1778.0 -876.00
sm1  6 1763.9 1783.1 -875.97 0.0639      1     0.8004
```

We can see that the fitted model sm2 is quite similar to sm1 except for the obvious difference that there is no within-subject correlation of the random effects in sm2. The estimates of all the other parameters, which are common to the two models, are practically unchanged.

The call to anova compares the two fitted models using a likelihood ratio test, which evaluates the change in the quality of the fits, as measured by the deviance (defined in §**??**), relative to the change in the number of parameters. The results of this test indicate that the model sm1 does not fit significantly better than the model sm2 and hence we prefer the model sm2 which has fewer parameters.

We conclude that there is significant variation between subjects in both the initial reaction time and in the rate of change in reaction time with respect to days of sleep deprivation but that these changes are not correlated. That is knowing a person's initial reaction time does not help us to predict their response to sleep deprivation.

3.3 Assessing the precision of the parameter estimates

Plots of the profile ζ (Figure˜3.3) show that confidence intervals σ_1 and

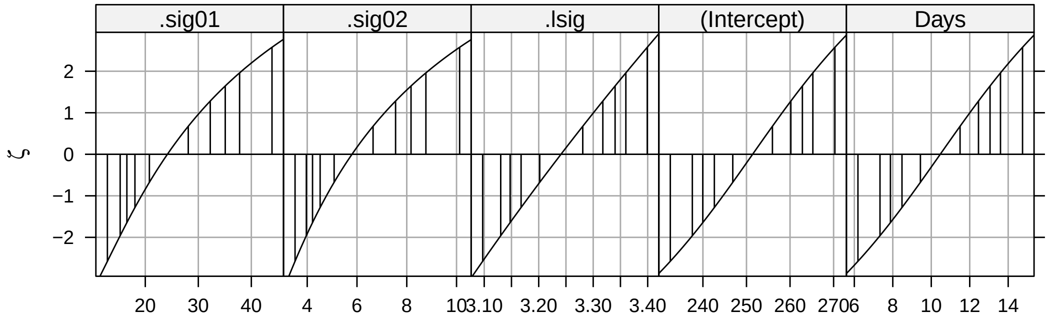


Fig. 3.3 Signed square root, ζ, of the likelihood ratio test statistic for each of the parameters in model fm2. The vertical lines are the endpoints of 50%, 80%, 90%, 95% and 99% confidence intervals derived from this test statistic.

σ_2 will be slightly skewed; those for $\log(\sigma)$ will be symmetric and wellapproximated 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% confidence intervals from the profile ζ are

```
> confint(pr2)

                 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 (Figure˜3.4)

In the previous section we used a likelihood ratio test to assess whether the covariance of the random effects within subject is significantly different from zero. This is an example of a statistical hypothesis test, which is one form of statistical inference. Another, related, form of statisticial inference is assessing the precision of the parameter estimates, say by forming confidence intervals or more general confidence regions.

For some statistical models it is possible to derive the theoretical distributions of the parameter estimates and use these theoretical distributions to create confidence intervals or regions. At present, the theoretical tools to analyze the general form of a linear mixed model in this "exact" approach are not available and confidence intervals and regions are typically created using approximations, especially what are called "asymptotic" approximations that are expected to perform well for large data sets.

Recently another approach to statistical inference, using Markov chain Monte Carlo (MCMC) samples from the (Bayesian) posterior distribution of the parameters, has gained popularity. Although MCMC is a computationally intensive approach to inference the current availability of inexpensive, powerful computers has made it feasible for models such as linear mixed models.

An advantage of using MCMC samples to assess the precision of the parameter estimates is that this approach uses the actual distribution of the parameters and not an approximation. Furthermore we can use graphical techniques to visualize these distributions and provide insight into the behavior of the parameters in the model.

MCMC sampling methods are based on a Bayesian formulation of the linear mixed model in which the parameters are considered to be random variables that have a *prior* distribution (prior in the sense of "before the data are known") and a *posterior*, or "after the data are known" distribution. Instead of confidence intervals on the parameters we will formulate *highest posterior density* (HPD) intervals (3). An 95% HPD interval on a parameter is the shortest interval that contains 95% of the probability content of the posterior distribution. It is the Bayesian equivalent of a 95% confidence interval on the parameter.

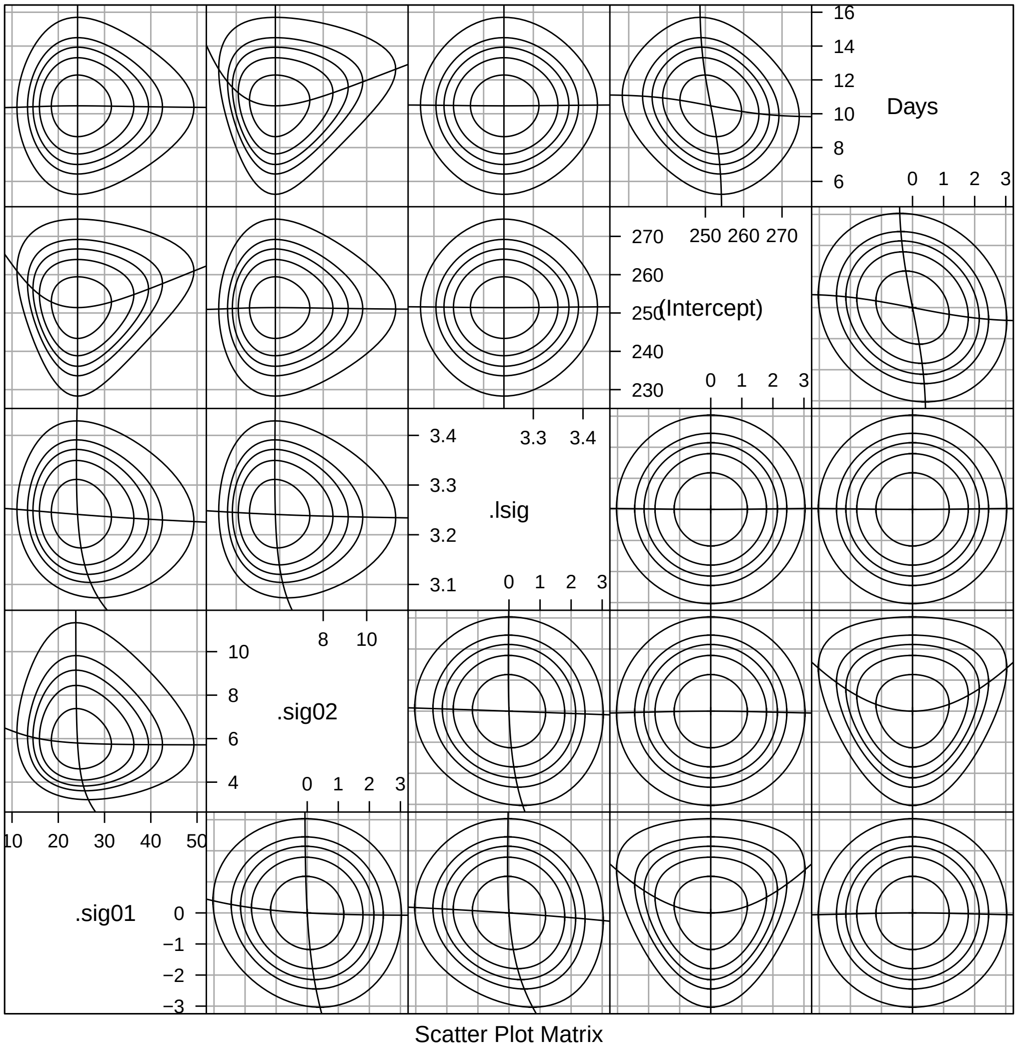


Fig. 3.4 Profile pairs plot for the parameters in model `fm2`. The contour lines correspond to marginal 50%, 80%, 90%, 95% and 99% confidence regions based on the likelihood ratio. Panels below the diagonal represent the (ζ_i, ζ_j) parameters; those above the diagonal represent the original parameters.

Details of the particular Bayesian formulation of the linear mixed model that we use, including the choice of prior distributions for the parameters, are given in §**??**.

3.3.1 Posterior distributions from model `sm2`

The function `mcmcsamp` applied to a fitted lmer model produces an MCMC sample from the posterior distribution of the parameter estimates, from which we can evaluate HPD intervals. Let us create and store a sample of size 10,000 from the posterior distribution of the parameters in model `sm2`.

The `HPDinterval` function creates HPD intervals on each of the parameters in the sample. By default it returns intervals whose empirical probability content is 95% (this can be changed, if desired).

Notice that in the intervals all the variance parameters are reported on the logarithm scale. The reason for taking this transformation is because the logarithm of a variance tends to be symmetrically distributed as shown by the density plots in Figure˜**??**

Not only are the posterior distributions on this scale symmetric, they are very close to normal distributions as shown by their normal probability plots (Figure˜**??**)

3.4 Examining the random effects

- Although the random effects $\mathbf{b}$ behave like parameters in the linear predictor, technically they are not parameters in the model.
- Instead of referring to "estimates" of the random effects it is customary to refer to "predictors" - in particular, the best linear unbiased predictors or BLUPs.
- These values are also the modes of the conditional distribution (i.e. given the data $\mathbf{y}$ and the estimates of $\boldsymbol{\beta}$, σ^2 and Σ) of $\mathbf{b}$.
- For linear mixed model the conditional distribution $[\mathbf{b}|\mathbf{y},\sigma^2,\Sigma]$ is normal (Gaussian) hence the modes are also the conditional means.
- The `ranef` extractor function returns these conditional modes evaluated at the parameter estimates.

```
> (rr1 <- ranef(sm1))

$Subject
    (Intercept)        Days
308   2.2585637   9.1989719
309 -40.3985926  -8.6197003
310 -38.9602618  -5.4488771
330  23.6905107  -4.8143332
331  22.2602137  -3.0698963
332   9.0395301  -0.2721713
333  16.8404388  -0.2236257
334  -7.2325828   1.0745766
335  -0.3336928 -10.7521593
337  34.8903640   8.6282815
349 -25.2101221   1.1734162
350 -13.0699646   6.6142061
351   4.5778381  -3.0152576
352  20.8636008   3.5360118
369   3.2754542   0.8722164
370 -25.6128825   4.8224666
```

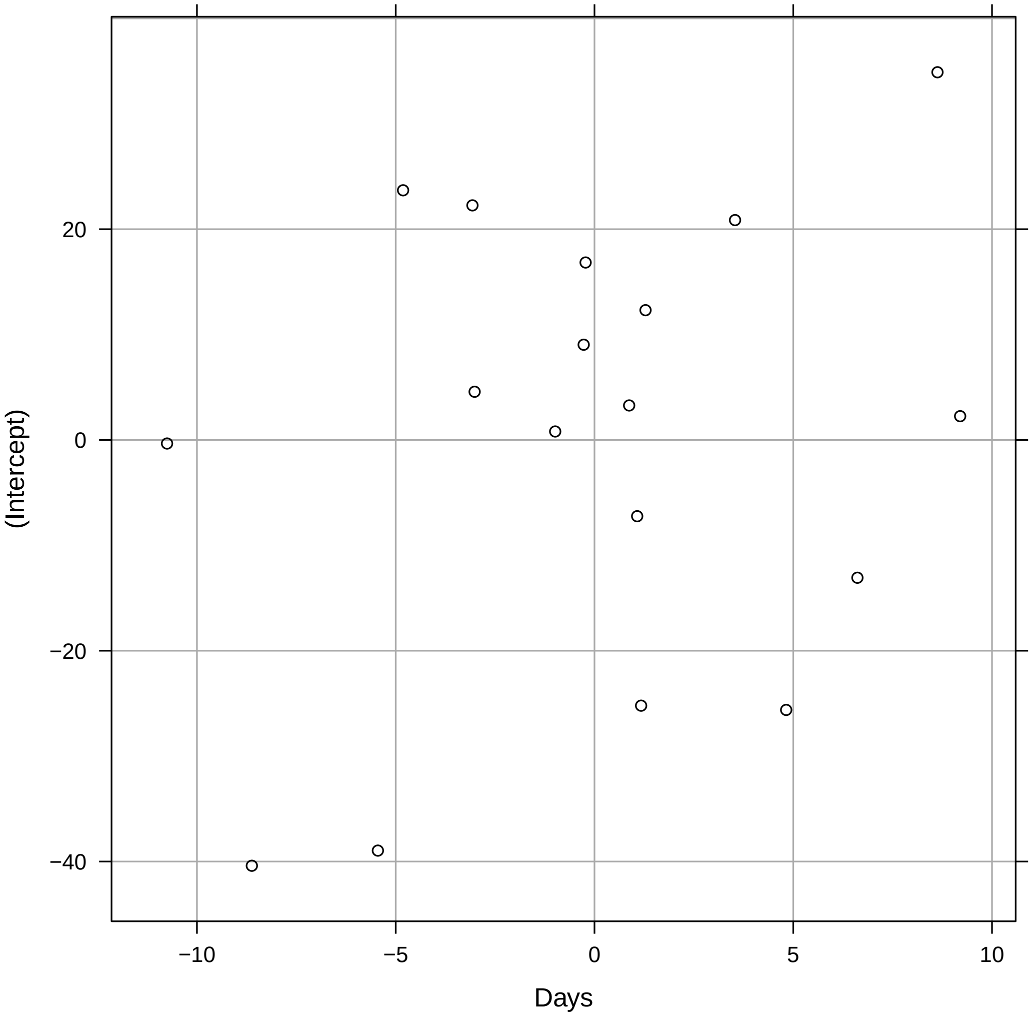


Fig. 3.5 Scatterplot of the conditional modes, or BLUPs, of the random effects for model `Sm1`. Each point represents the mode of the distribution of the random effects for the intercept and slope associated with one of the subjects.

```
371   0.8070404  -0.9881552
372  12.3145445   1.2840288
```

- For this model we can combine the BLUPs of the random effects and the estimates of the fixed effects to get BLUPs for the within-subject coefficients.
- These BLUPs will be “shrunken” towards the fixed-effects estimates relative to the estimated coefficients from only that subjects data. John Tukey called this “borrowing strength” between subjects.
- Plotting the shrinkage of the within-subject coefficients shows that some of the coefficients are considerably shrunken toward the fixed-effects estimates.

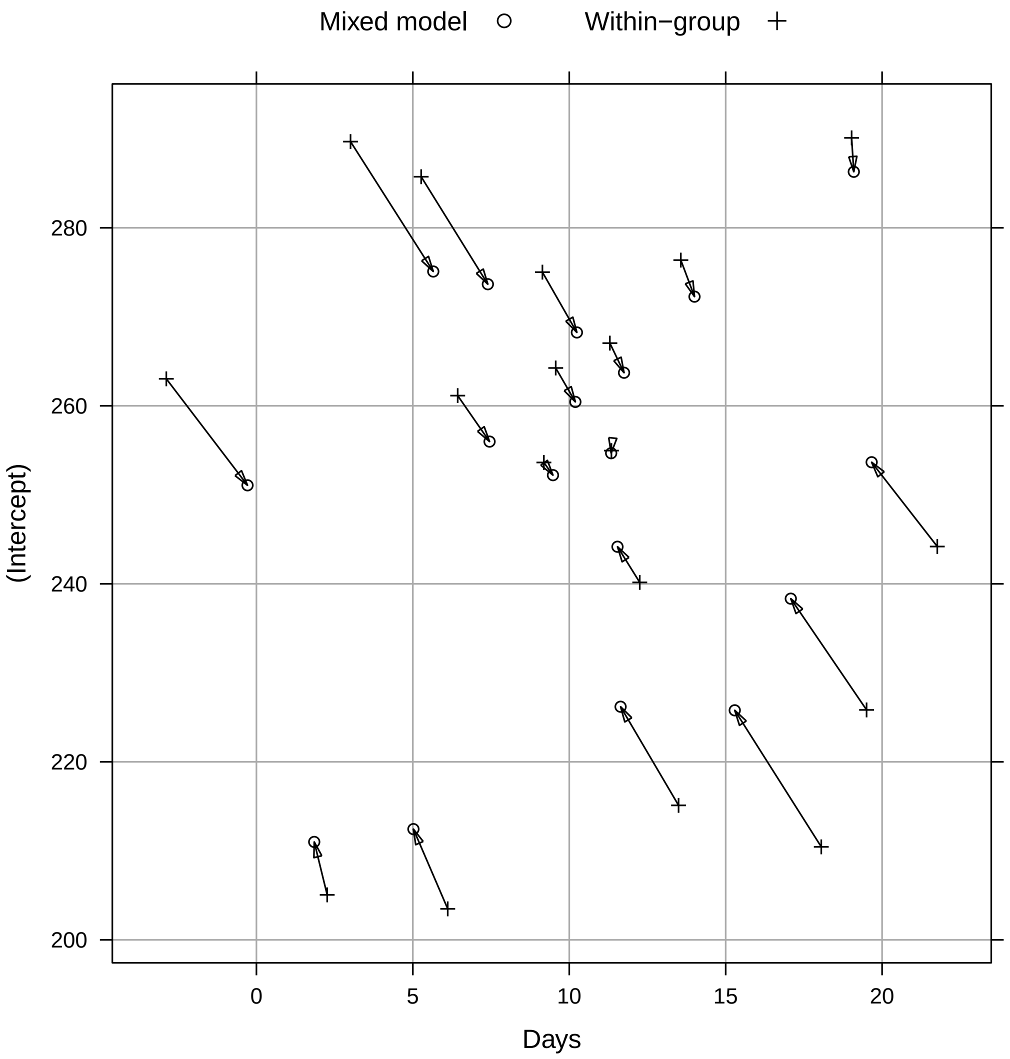


Fig. 3.6 Comparison of the within-subject estimates of the intercept and slope for each subject and the conditional modes of the per-subject intercept and slope. Each pair of points joined by an arrow are the within-subject and conditional mode estimates for the same subject. The arrow points from the within-subject estimate to the conditional mode for the mixed-effects model.

- However, comparing the within-group and mixed model fitted lines shows that large changes in coefficients occur in the noisy data. Precisely estimated within-group coefficients are not changed substantially.

3.4.1 Prediction intervals on the random effects

- For the linear mixed model we can calculate both the means and the variances of the random-effects conditional on the estimated values of the

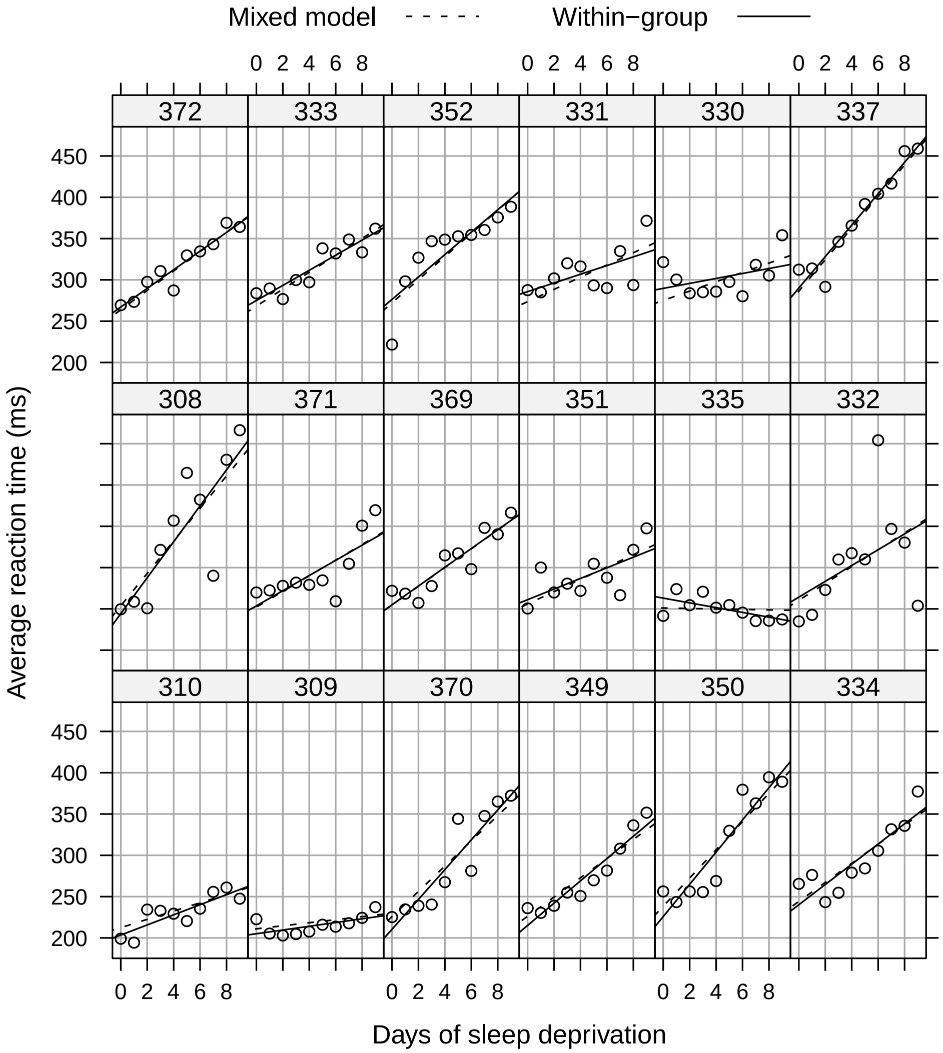


Fig. 3.7 Comparison of the predictions from the within-subject fits with those from the conditional modes of the subject-specific parameters in the mixed-effects model.

Fig. 3.8 Prediction intervals on the random effects per subject.

model parameters, which allows us to calculate prediction intervals on the values of individual random effects.

- We plot the prediction intervals as a normal probabity plot so we can see the overall shape of the distribution of the means and which of the random effects are "significantly different" from zero.
- Note that failure of the conditional means of the random effects to look like a normal (Gaussian) distribution is not terribly alarming. It is the "prior" distribution of the random effects that is assumed to be normal. The conditional means or BLUPs are strongly influenced by the data and may appear non-normal.

3.5 Model specification for `lmer`

A linear mixed-effects model to be fit by `lmer` is specified by the `formula` argument. For model `sm1` the formula is

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

which can be read as "`Reaction` is modeled by `Days` and `Days` given `Subject`". That is, the response, which is the variable named `Reaction`, is to be modeled by one conditional term, `(Days|Subject)`, and one unconditional term, `Days`.

A conditonal term (any term including the vertical bar, `|`) contributes to the random effects specification. An unconditional term contributes to the fixed effects specification.

The unconditional terms, together with the data to be fit, generate an $n \times p$ fixed-effects model matrix $\mathbf{X}$ according to the rules that we describe in §**??**. The dimensions n and p are the number of observations and the dimension of the fixed-effects parameter vector $\boldsymbol{\beta}$, respectively. In this case of model `sm1` the only unconditional term, `Days`, is the name of a numeric variable, which is incorporated as a column, labelled `Days` in $\mathbf{X}$. By convention, the intercept term, which generates a column of 1's labelled `(Intercept)`, is included implicitly in the model specification. (This column can be suppressed if desired, as shown below.)

The first few rows of the 180×2 model matrix $\mathbf{X}$ for model `sm1` are

```
> head(model.matrix(sm1))

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

Each conditional term in the model formula generates a set of random effects and variance-covariance matrix for these random effects. In a conditional term the expression on the right hand side of the `|` is evaluated as a factor called the *grouping factor* for the term. Because a factor associates one of a finite set of levels with each observation, we can consider a factor as dividing the observations into groups corresponding to the levels of the factor. Observations within such groups share a set of random effects. The expression on the left of the `|` in a conditional term is evaluated as a linear model formula and determines the number and form of the random effects associated with each level of the grouping factor. Thus `(Days|Subject)` designates an (implicit) intercept coefficient and a `Days` coefficient for each level of the `Subject` grouping factor producing, as we have seen, 36 random effects — two for each of the 18 subjects.

Random effects generated by different conditional terms are independent, as are random effects corresponding to different levels of the grouping factor in the same conditional term.

Each group of random effects models some of the variation in the response. There is one further level of variation in the model - the “per-observation” or “residual” noise. It is the unexplained variation or what is “left over” after we have modeled all the other sources of variation in the model. This level of variation is modeled as an n-dimensional vector $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ where $\mathbf{I}$ is an identity matrix. That is, the elements of $\boldsymbol{\varepsilon}$ is independent and identical normal random variates with mean zero and variance σ^2.

The general model allows for multiple conditional terms in a model specification generating multiple groups of random effects. Let k be the number of conditional terms, $n_i, i = 1, \ldots, k$ be the number of levels of the grouping factor for the ith such term and q_i be the number of random effects associated with each level of the grouping factor.

If we represent all the random effects as a vector $\mathbf{b}$, of length $q = \sum_{i=1}^{k} q_i n_i$, we can write the model as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \varepsilon, \quad \mathbf{b} \sim \mathcal{N}(\mathbf{0}, \Sigma), \varepsilon \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right), \mathbf{b} \perp \varepsilon \tag{3.2}$$

where $\mathbf{Z}$ is an $n \times q$ model matrix generated from the conditional terms, Σ is a $q \times q$ variance-covariance matrix, also generated from the conditional terms, and the symbol $\perp$ denotes independent random variables.

Although the total number of random effects, q, can be large and hence the dimensions of $\mathbf{Z}$ and Σ in the general form (3.2) can be very large, these matrices are sparse and patterned and thus are determined by a relatively small number of values.

For model `sm1`, $k = 1$, $q_1 = 2$ and $n_i = 18$ so, as we have seen, $q = 36$. However, the assumptions of independence of random effects associated with different subjects means that the 36×36 matrix Σ consists of the 2×2 matrix Σ_1 repeated 18 times in the pattern

$$\Sigma = \begin{bmatrix} \Sigma_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \Sigma_1 & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \Sigma_1 \end{bmatrix} \tag{3.3}$$

Also, although the model matrix $\mathbf{Z}$ has dimension 180×36, all but two of the elements in any given row are known to be zero. Each row corresponds to one and only one subject and the model for that row incorporates only the two random effects associated with that subject. The other 34 random effects associated with other subjects are multiplied by zero.

The matrix $\mathbf{Z}$ is evaluated and stored as a sparse matrix. The first few rows of $\mathbf{Z}$ for model `sm1` are

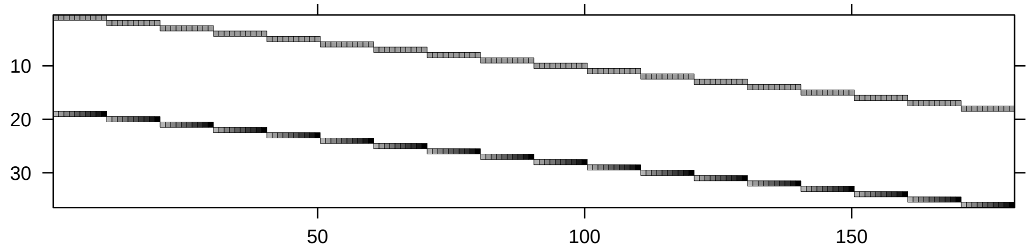


Fig. 3.9 Image of $\mathbf{Z}^{\mathsf{T}}$, the transpose of $\mathbf{Z}$, the random effects model matrix in model `sm2`

```
6 x 36 sparse Matrix of class "dgCMatrix"

[1,] 1 0 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
[2,] 1 1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
[3,] 1 2 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
[4,] 1 3 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
[5,] 1 4 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
[6,] 1 5 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
```

These rows correspond to the first 6 observations on the first subject. In the representation as a sparse matrix an element that is known to be zero prints as a '.'. (The (1,2) element of this matrix does have the value zero because the value of the `Days` variable is zero for this observation. It could have another value if, for example, we renumbered the `Days` and thus is not a systematic zero in the matrix.)

Easier to understand, perhaps, is the image of $\mathbf{Z}^{\mathsf{T}}$ in Figure˜3.9.

In the model $\mathbf{b}$ is a random variable

$$\mathbf{b} \sim \mathscr{N}(\mathbf{0}, \Sigma) \tag{3.4}$$

and it is the elements of Σ that we estimate. Because we assume that random effects associated with different grouping factors are indep which we can write as If we write the model matrix for the entire random effects vector as $\mathbf{Z}$ and the variance-cov matrix, which we call $\mathbf{Z}$, for the random effects vector.

contribute to another model matrix, which we call $\mathbf{Z}$, through a slightly more complicated mechanism. In a conditional term the expression on the left of the `|` is interpreted as a linear model formula and used to create a model matrix while the expression on the right of the `|` is evaluated as a factor, called the *grouping factor* for the term. We allow multiple conditional terms in a model specification so we will refer to the ith conditional term even though there is only one such conditional term in the model specification for model `sm1`.

If the model matrix from the expression on the left of the ith conditional has q_i columns and the grouping factor has n_i levels then the expression contributes $q_i n_i$ columns to the matrix $\mathbf{Z}$.

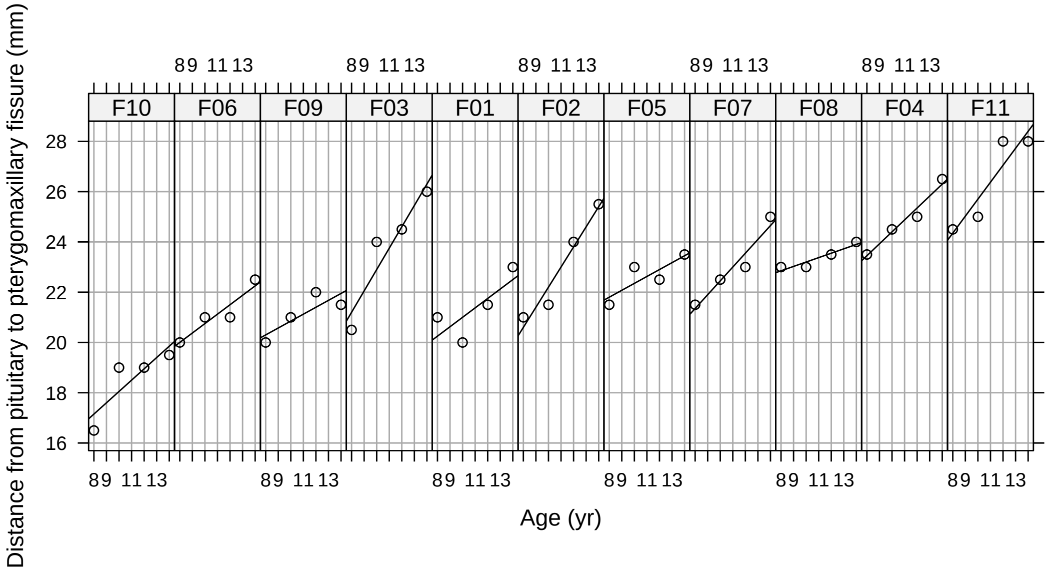
3.6 Conclusions from the example

- Carefully plotting the data is enormously helpful in formulating the model.
- It is relatively easy to fit and evaluate models to data like these, from a balanced designed experiment.
- For a linear mixed model the estimates of the fixed effects typically have a symmetric distribution close to a Gaussian distribution.
- The distribution of the variance components or the covariances are not symmetric, which is why we transform these parameters to a symmetric scale.
- We use the MCMC sample to create confidence (actually HPD) intervals on the fixed-effects parameters. We could also use the parameter estimates and standard errors.
- The "estimates" (actually BLUPs) of the random effects can be considered as penalized estimates of these parameters in that they are shrunk towards the origin.
- Most of the prediction intervals for the random effects overlap zero.

Problems

3.1. Check the structure of documentation, structure and a summary of the `Orthodont` data set.

1. 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.
2. 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)`).
3. 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



4. 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.
5. Produce a "caterpillar plot" of the random effects for intercept and slope. Does the plot indicate correlated random effects?
6. 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)?
7. Repeat for the data from the male subjects.

3.2.

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

Chapter 4
Computational methods

4.1 Introduction

The `lme4` package provides `R` functions to fit and analyze linear mixed models, generalized linear mixed models and nonlinear mixed models. These models are called *mixed-effects models* or, more simply, *mixed models* because they incorporate both *fixed-effects* parameters, which apply to an entire population or to certain well-defined and repeatable subsets of a population, and *random effects*, which apply to the particular experimental units or observational units in the study. Such models are also called *multilevel* models because the random effects represent levels of variation in addition to the per-observation noise term that is incorporated in common statistical models such as linear regression models, generalized linear models and nonlinear regression models.

We begin by describing common properties of these mixed models and the general computational approach used in the `lme4` package. The estimates of the parameters in a mixed model are determined as the values that optimize an objective function — either the likelihood of the parameters given the observed data, for maximum likelihood (ML) estimates, or a related objective function called the REML criterion. Because this objective function must be evaluated at many different values of the model parameters during the optimization process, we focus on the evaluation of the objective function and a critical computation in this evalution — determining the solution to a penalized, weighted least squares (PWLS) problem.

The dimension of the solution of the PWLS problem can be very large, perhaps in the millions. Furthermore, such problems must be solved repeatedly during the optimization process to determine parameter estimates. The whole approach would be infeasible were it not for the fact that the matrices determining the PWLS problem are sparse and we can use sparse matrix storage formats and sparse matrix computations (8). In particular, the whole computational approach hinges on the extraordinarily efficient methods for determining the Cholesky decomposition of sparse, symmetric, positive-definite matrices embodied in the CHOLMOD library of C functions (7).

In the next section we describe the general form of the mixed models that can be represented in the `lme4` package and the computational approach embodied in the package. In the following section we describe a particular form of mixed model, called a linear mixed model, and the computational details for those models. In the fourth section we describe computational methods for generalized linear mixed models, nonlinear mixed models and generalized nonlinear mixed models.

4.2 Formulation of mixed models

A mixed-effects model incorporates two vector-valued random variables: the n-dimensional response vector, $\mathscr{Y}$, and the q-dimensional random effects vector, $\mathscr{B}$. We observe the value, $\mathbf{y}$, of $\mathscr{Y}$. We do not observe the value of $\mathscr{B}$.

The random variable $\mathscr{Y}$ may be continuous or discrete. That is, the observed data, $\mathbf{y}$, may be on a continuous scale or they may be on a discrete scale, such as binary responses or responses representing a count. In our formulation, the random variable $\mathscr{B}$ is always continous.

We specify a mixed model by describing the unconditional distribution of $\mathscr{B}$ and the conditional distribution $(\mathscr{Y}|\mathscr{B} = \mathbf{b})$.

4.2.1 The unconditional distribution of $\mathscr{B}$

In our formulation, the unconditional distribution of $\mathscr{B}$ is always a q-dimensional multivariate Gaussian (or “normal”) distribution with mean $\mathbf{0}$ and with a parameterized covariance matrix,

$$\mathscr{B} \sim \mathscr{N}\left(\mathbf{0}, \sigma^2 \Lambda(\theta)\Lambda^{\mathsf{T}}(\theta)\right). \tag{4.1}$$

The scalar, σ, in (4.1), is called the *common scale parameter*. As we will see later, not all types of mixed models incorporate this parameter. We will include σ^2 in the general form of the unconditional distribution of $\mathscr{B}$ with the understanding that, in some models, $\sigma \equiv 1$.

The $q \times q$ matrix $\Lambda(\theta)$, which is a left factor of the covariance matrix (when $\sigma = 1$) or the relative covariance matrix (when $\sigma \neq 1$), depends on an m-dimensional parameter θ. Typically $m \ll q$; in the examples we show below it is always the case that $m < 5$, even when q is in the thousands. The fact that m is very small is important because, as we shall see, determining the parameter estimates in a mixed model can be expressed as an optimization problem with respect to θ only.

The parameter $\boldsymbol{\theta}$ may be, and typically is, subject to constraints. For ease of computation, we require that the constraints be expressed as "box" constraints of the form $\theta_{iL} \leq \theta_i \leq \theta_{iU}, i = 1, \ldots, m$ for constants θ_{iL} and $\theta_{iU}, i = 1, \ldots, m$. We shall write the set of such constraints as $\boldsymbol{\theta}_L \leq \boldsymbol{\theta} \leq \boldsymbol{\theta}_R$. The matrix $\Lambda(\boldsymbol{\theta})$ is required to be non-singular (i.e. invertible) when $\boldsymbol{\theta}$ is not on the boundary.

4.2.2 The conditional distribution, $(\mathscr{Y}|\mathscr{B} = \mathbf{b})$

The conditional distribution, $(\mathscr{Y}|\mathscr{B} = \mathbf{b})$, must satisfy:

1. The conditional mean, $\mu_{\mathscr{Y}|\mathscr{B}}(\mathbf{b}) = \mathrm{E}[\mathscr{Y}|\mathscr{B} = \mathbf{b}]$, depends on $\mathbf{b}$ only through the value of the *linear predictor*, $\mathbf{Zb} + \mathbf{X}\beta$, where β is the p-dimensional *fixed-effects* parameter vector and the *model matrices*, $\mathbf{Z}$ and $\mathbf{X}$, are fixed matrices of the appropriate dimension. That is, the two model matrices must have the same number of rows and must have q and p columns, respectively. The number of rows in $\mathbf{Z}$ and $\mathbf{X}$ is a multiple of n, the dimension of $\mathbf{y}$.
2. The scalar distributions, $(\mathscr{Y}_i|\mathscr{B} = \mathbf{b}), i = 1, \ldots, n$, all have the same form and are completely determined by the conditional mean, $\mu_{\mathscr{Y}|\mathscr{B}}(\mathbf{b})$ and, at most, one additional parameter, σ, which is the common scale parameter.
3. The scalar distributions, $(\mathscr{Y}_i|\mathscr{B} = \mathbf{b}), i = 1, \ldots, n$, are independent. That is, the components of $\mathscr{Y}$ are *conditionally independent* given $\mathscr{B}$.

An important special case of the conditional distribution is the multivariate Gaussian distribution of the form

$$(\mathscr{Y}|\mathscr{B} = \mathbf{b}) \sim \mathscr{N}(\mathbf{Zb} + \mathbf{X}\beta, \sigma^2\mathbf{I}_n) \tag{4.2}$$

where $\mathbf{I}_n$ denotes the identity matrix of size n. In this case the conditional mean, $\mu_{\mathscr{Y}|\mathscr{B}}(\mathbf{b})$, is exactly the linear predictor, $\mathbf{Zb} + \mathbf{X}\beta$, a situation we will later describe as being an "identity link" between the conditional mean and the linear predictor. Models with conditional distribution (4.2) are called *linear mixed models.*

4.2.3 A change of variable to "spherical" random effects

Because the conditional distribution $(\mathscr{Y}|\mathscr{B} = \mathbf{b})$ depends on $\mathbf{b}$ only through the linear predictor, it is easy to express the model in terms of a linear transformation of $\mathscr{B}$. We define the linear transformation from a q-dimensional "spherical" Gaussian random variable, $\mathscr{U}$, to $\mathscr{B}$ as

$$\mathscr{B} = \Lambda(\theta)\mathscr{U}, \quad \mathscr{U} \sim \mathscr{N}(\mathbf{0}, \sigma^2\mathbf{I}_q). \tag{4.3}$$

(The term "spherical" refers to the fact that contours of constant probability density for $\mathscr{U}$ are spheres centered at the mean — in this case, $\mathbf{0}$.)

When θ is not on the boundary this is an invertible transformation. When θ is on the boundary the transformation can fail to be invertible. However, we will only need to be able to express $\mathscr{B}$ in terms of $\mathscr{U}$ and that transformation is well-defined, even when θ is on the boundary.

The linear predictor, as a function of $\mathbf{u}$, is

$$\gamma(\mathbf{u}) = \mathbf{Z}\Lambda(\theta)\mathbf{u} + \mathbf{X}\beta. \tag{4.4}$$

When we wish to emphasize the role of the model parameters, θ and β, in the formulation of γ, we will write the linear predictor as $\gamma(\mathbf{u}, \theta, \beta)$.

4.2.4 The conditional density $(\mathscr{U}|\mathscr{Y} = \mathbf{y})$

Because we observe $\mathbf{y}$ and do not observe $\mathbf{b}$ or $\mathbf{u}$, the conditional distribution of interest, for the purposes of statistical inference, is $(\mathscr{U}|\mathscr{Y} = \mathbf{y})$ (or, equivalently, $(\mathscr{B}|\mathscr{Y} = \mathbf{y})$). This conditional distribution is always a continuous distribution with conditional probability density $f_{\mathscr{U}|\mathscr{Y}}(\mathbf{u}|\mathbf{y})$.

We can evaluate $f_{\mathscr{U}|\mathscr{Y}}(\mathbf{u}|\mathbf{y})$, up to a constant, as the product of the unconditional density, $f_{\mathscr{U}}(\mathbf{u})$, and the conditional density (or the probability mass function, whichever is appropriate), $f_{\mathscr{Y}|\mathscr{U}}(\mathbf{y}|\mathbf{u})$. We write this unnormalized conditional density as

$$h(\mathbf{u}|\mathbf{y}, \theta, \beta, \sigma) = f_{\mathscr{Y}|\mathscr{U}}(\mathbf{y}|\mathbf{u}, \theta, \beta, \sigma) f_{\mathscr{U}}(\mathbf{u}|\sigma). \tag{4.5}$$

We say that h is the "unnormalized" conditional density because all we know is that the conditional density is proportional to $h(\mathbf{u}|\mathbf{y}, \theta, \beta, \sigma)$. To obtain the conditional density we must normalize h by dividing by the value of the integral

$$L(\theta, \beta, \sigma|\mathbf{y}) = \int_{\mathbb{R}^q} h(\mathbf{u}|\mathbf{y}, \theta, \beta, \sigma)\, d\mathbf{u}. \tag{4.6}$$

We write the value of the integral (4.6) as $L(\theta, \beta, \sigma|\mathbf{y})$ because it is exactly the *likelihood* of the parameters θ, β and σ, given the observed data $\mathbf{y}$. The *maximum likelihood (ML) estimates* of these parameters are the values that maximize L.

4.2.5 Determining the ML estimates

The general problem of maximizing $L(\boldsymbol{\theta},\boldsymbol{\beta},\sigma|\mathbf{y})$ with respect to $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and σ can be formidable because each evaluation of this function involves a potentially high-dimensional integral and because the dimension of $\boldsymbol{\beta}$ can be large. However, this general optimization problem can be split into manageable subproblems. Given a value of $\boldsymbol{\theta}$ we can determine the *conditional mode*, $\tilde{\mathbf{u}}(\boldsymbol{\theta})$, of $\mathbf{u}$ and the *conditional estimate*, $\tilde{\beta}(\boldsymbol{\theta})$ simultaneously using *penalized, iteratively re-weighted least squares* (PIRLS). The conditional mode and the conditional estimate are defined as

$$\begin{bmatrix}\tilde{\mathbf{u}}(\boldsymbol{\theta})\\ \tilde{\beta}(\boldsymbol{\theta})\end{bmatrix} = \arg\max_{\mathbf{u},\beta} h(\mathbf{u}|\mathbf{y},\boldsymbol{\theta},\boldsymbol{\beta},\sigma). \tag{4.7}$$

(It may look as if we have missed the dependence on σ on the left-hand side but it turns out that the scale parameter does not affect the location of the optimal values of quantities in the linear predictor.)

As is common in such optimization problems, we re-express the conditional density on the *deviance scale*, which is negative twice the logarithm of the density, where the optimization becomes

$$\begin{bmatrix}\tilde{\mathbf{u}}(\boldsymbol{\theta})\\ \tilde{\beta}(\boldsymbol{\theta})\end{bmatrix} = \arg\min_{\mathbf{u},\beta} -2\log\left(h(\mathbf{u}|\mathbf{y},\boldsymbol{\theta},\boldsymbol{\beta},\sigma)\right). \tag{4.8}$$

It is this optimization problem that can be solved quite efficiently using PIRLS. In fact, for linear mixed models, which are described in the next section, $\tilde{\mathbf{u}}(\boldsymbol{\theta})$ and $\tilde{\beta}(\boldsymbol{\theta})$ can be directly evaluated.

The second-order Taylor series expansion of $-2\log h$ at $\tilde{\mathbf{u}}(\boldsymbol{\theta})$ and $\tilde{\beta}(\boldsymbol{\theta})$ provides the Laplace approximation to the profiled deviance. Optimizing this function with respect to $\boldsymbol{\theta}$ provides the ML estimates of $\boldsymbol{\theta}$, from which the ML estimates of $\boldsymbol{\beta}$ and σ (if used) are derived.

4.3 Methods for linear mixed models

As indicated in the introduction, a critical step in our methods for determining the maximum likelihood estimates of the parameters in a mixed model is solving a penalized, weighted least squares (PWLS) problem. We will motivate the general form of the PWLS problem by first considering computational methods for linear mixed models that result in a penalized least squares (PLS) problem.

Recall from §4.2.2 that, in a linear mixed model, both the conditional distribution, $(\mathscr{Y}|\mathscr{U}=\mathbf{u})$, and the unconditional distribution, $\mathscr{U}$, are spherical Gaussian distributions and that the conditional mean, $\mu_{\mathscr{Y}|\mathscr{U}}(\mathbf{u})$, is the linear predictor, $\gamma(\mathbf{u})$. Because all the distributions determining the model are continuous distributions, we consider their densities. On the deviance scale these are

$$\begin{aligned}
-2\log(f_{\mathscr{U}}(\mathbf{u})) &= q\log(2\pi\sigma^2) + \frac{\|\mathbf{u}\|^2}{\sigma^2} \\
-2\log(f_{\mathscr{Y}|\mathscr{U}}(\mathbf{y}|\mathbf{u})) &= n\log(2\pi\sigma^2) + \frac{\|\mathbf{y}-\mathbf{Z}\Lambda(\theta)\mathbf{u}-\mathbf{X}\beta\|^2}{\sigma^2} \\
-2\log(h(\mathbf{u}|\mathbf{y},\theta,\beta,\sigma)) &= (n+q)\log(2\pi\sigma^2) + \frac{\|\mathbf{y}-\gamma(\mathbf{u},\theta,\beta)\|^2+\|\mathbf{u}\|^2}{\sigma^2} \\
&= (n+q)\log(2\pi\sigma^2) + \frac{d(\mathbf{u}|\mathbf{y},\theta,\beta)}{\sigma^2}
\end{aligned} \tag{4.9}$$

In (4.9) the *discrepancy* function,

$$d(\mathbf{u}|\mathbf{y},\theta,\beta) = \|\mathbf{y}-\gamma(\mathbf{u},\theta,\beta)\|^2+\|\mathbf{u}\|^2 \tag{4.10}$$

has the form of a penalized residual sum of squares in that the first term, $\|\mathbf{y}-\gamma(\mathbf{u},\theta,\beta)\|^2$ is the residual sum of squares for $\mathbf{y}$, $\mathbf{u}$, θ and β and the second term, $\|\mathbf{u}\|^2$, is a penalty on the size of $\mathbf{u}$. Notice that the discrepancy does not depend on the common scale parameter, σ.

4.3.1 The canonical form of the discrepancy

Using a so-called "pseudo data" representation, we can write the discrepancy as a residual sum of squares for a regression model that is linear in both $\mathbf{u}$ and β

$$d(\mathbf{u}|\mathbf{y},\theta,\beta) = \left\| \begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix} - \begin{bmatrix}\mathbf{Z}\Lambda(\theta) & \mathbf{X}\\ \mathbf{I}_q & \mathbf{0}\end{bmatrix} \begin{bmatrix}\mathbf{u}\\ \beta\end{bmatrix} \right\|^2. \tag{4.11}$$

The term "pseudo data" reflects the fact that we have added q "pseudo observations" to the observed response, $\mathbf{y}$, and to the linear predictor, $\gamma(\mathbf{u},\theta,\beta) = \mathbf{Z}\Lambda(\theta)\mathbf{u}+\mathbf{X}\beta$, in such a way that their contribution to the overall residual sum of squares is exactly the penalty term in the discrepancy.

In the form (4.11) we can see that the discrepancy is a quadratic form in both $\mathbf{u}$ and β. Furthermore, because we require that $\mathbf{X}$ has full column rank, the discrepancy is a positive-definite quadratic form in $\mathbf{u}$ and β that is minimized at $\tilde{\mathbf{u}}(\theta)$ and $\tilde{\beta}(\theta)$ satisfying

$$\begin{bmatrix}\Lambda^{\mathsf{T}}(\theta)\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\Lambda(\theta)+\mathbf{I}_q & \Lambda^{\mathsf{T}}(\theta)\mathbf{Z}^{\mathsf{T}}\mathbf{X}\\ \mathbf{X}^{\mathsf{T}}\mathbf{Z}\Lambda(\theta) & \mathbf{X}^{\mathsf{T}}\mathbf{X}\end{bmatrix} \begin{bmatrix}\tilde{\mathbf{u}}(\theta)\\ \tilde{\beta}(\theta)\end{bmatrix} = \begin{bmatrix}\Lambda^{\mathsf{T}}(\theta)\mathbf{Z}^{\mathsf{T}}\mathbf{y}\\ \mathbf{X}^{\mathsf{T}}\mathbf{y}\end{bmatrix} \tag{4.12}$$

An effective way of determining the solution to a sparse, symmetric, positive definite system of equations such as (4.12) is the sparse Cholesky decomposition (8). If $\mathbf{A}$ is a sparse, symmetric positive definite matrix then the sparse Cholesky factor with fill-reducing permutation $\mathbf{P}$ is the lower-triangular matrix $\mathbf{L}$ such that

$$\mathbf{L}\mathbf{L}^{\mathsf{T}} = \mathbf{P}\mathbf{A}\mathbf{P}^{\mathsf{T}}. \tag{4.13}$$

(Technically, the factor $\mathbf{L}$ is only determined up to changes in the sign of the diagonal elements. By convention we require the diagonal elements to be positive.)

The fill-reducing permutation represented by the permutation matrix $\mathbf{P}$, which is determined from the pattern of nonzeros in $\mathbf{A}$ but does not depend on particular values of those nonzeros, can have a profound impact on the number of nonzeros in $\mathbf{L}$ and hence on the speed with which $\mathbf{L}$ can be calculated from $\mathbf{A}$.

In most applications of linear mixed models the matrix $\mathbf{Z}\Lambda(\theta)$ is sparse while $\mathbf{X}$ is dense or close to it so the permutation matrix $\mathbf{P}$ can be restricted to the form

$$\mathbf{P} = \begin{bmatrix} \mathbf{P_Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{P_X} \end{bmatrix} \tag{4.14}$$

without loss of efficiency. In fact, in most cases we can set $\mathbf{P_X} = \mathbf{I}_p$ without loss of efficiency.

Let us assume that the permutation matrix is required to be of the form (4.14) so that we can write the Cholesky factorization for the positive definite system (4.12) as

$$\begin{bmatrix} \mathbf{L_Z} & \mathbf{0} \\ \mathbf{L_{XZ}} & \mathbf{L_X} \end{bmatrix} \begin{bmatrix} \mathbf{L_Z} & \mathbf{0} \\ \mathbf{L_{XZ}} & \mathbf{L_X} \end{bmatrix}^{\mathsf{T}} = \\ \begin{bmatrix} \mathbf{P_Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{P_X} \end{bmatrix} \begin{bmatrix} \Lambda^{\mathsf{T}}(\theta)\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\Lambda(\theta) + \mathbf{I}_q & \Lambda^{\mathsf{T}}(\theta)\mathbf{Z}^{\mathsf{T}}\mathbf{X} \\ \mathbf{X}^{\mathsf{T}}\mathbf{Z}\Lambda(\theta) & \mathbf{X}^{\mathsf{T}}\mathbf{X} \end{bmatrix} \begin{bmatrix} \mathbf{P_Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{P_X} \end{bmatrix}^{\mathsf{T}}. \tag{4.15}$$

The discrepancy can now be written in the canonical form

$$d(\mathbf{u}|\mathbf{y},\theta,\beta) = \tilde{d}(\mathbf{y},\theta) + \left\| \begin{bmatrix} \mathbf{L_Z^{\mathsf{T}}} & \mathbf{L_{XZ}^{\mathsf{T}}} \\ \mathbf{0} & \mathbf{L_X^{\mathsf{T}}} \end{bmatrix} \begin{bmatrix} \mathbf{P_Z}(\mathbf{u}-\tilde{\mathbf{u}}) \\ \mathbf{P_X}(\beta-\tilde{\beta}) \end{bmatrix} \right\|^2 \tag{4.16}$$

where

$$\tilde{d}(\mathbf{y},\theta) = d(\tilde{\mathbf{u}}(\theta)|\mathbf{y},\theta,\tilde{\beta}(\theta)) \tag{4.17}$$

is the minimum discrepancy, given θ.

4.3.2 The profiled likelihood for linear mixed models

Substituting (4.16) into (4.9) provides the unnormalized conditional density $h(\mathbf{u}|\mathbf{y},\theta,\beta,\sigma)$ on the deviance scale as

$$
\begin{aligned}
&-2\log(h(\mathbf{u}|\mathbf{y},\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\sigma}))\\
&\qquad = (n+q)\log(2\pi\sigma^2) + \frac{\tilde{d}(\mathbf{y},\theta) + \left\| \begin{bmatrix} \mathbf{L}_\mathbf{Z}^\mathsf{T} & \mathbf{L}_\mathbf{XZ}^\mathsf{T} \\ \mathbf{0} & \mathbf{L}_\mathbf{X}^\mathsf{T} \end{bmatrix} \begin{bmatrix} \mathbf{P}_\mathbf{Z}(\mathbf{u}-\tilde{\mathbf{u}}) \\ \mathbf{P}_\mathbf{X}(\beta-\tilde{\beta}) \end{bmatrix} \right\|^2}{\sigma^2}. \qquad (4.18)
\end{aligned}
$$

As shown in Appendix **??**, the integral of a quadratic form on the deviance scale, such as (4.18), is easily evaluated, providing the log-likelihood, $\ell(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\sigma}|\mathbf{y})$, as

$$
\begin{aligned}
&-2\ell(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\sigma}|\mathbf{y})\\
&\qquad = -2\log\left(L(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\sigma}|\mathbf{y})\right)\\
&\qquad = n\log(2\pi\sigma^2) + \log(|\mathbf{L}_\mathbf{Z}|^2) + \frac{\tilde{d}(\mathbf{y},\theta) + \left\| \mathbf{L}_\mathbf{X}^\mathsf{T}\mathbf{P}_\mathbf{X}(\beta-\tilde{\beta}) \right\|^2}{\sigma^2}, \qquad (4.19)
\end{aligned}
$$

from which we can see that the conditional estimate of $\boldsymbol{\beta}$, given $\boldsymbol{\theta}$, is $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and the conditional estimate of $\boldsymbol{\sigma}$, given $\boldsymbol{\theta}$, is

$$
\tilde{\sigma}^2(\theta) = \frac{\tilde{d}(\theta|\mathbf{y})}{n}. \qquad (4.20)
$$

Substituting these conditional estimates into (4.19) produces the *profiled likelihood*, $\tilde{L}(\boldsymbol{\theta}|\mathbf{y})$, as

$$
-2\tilde{\ell}(\theta|\mathbf{y})) = \log(|\mathbf{L}_\mathbf{Z}(\theta)|^2) + n\left(1+\log\left(\frac{2\pi\tilde{d}(\mathbf{y},\theta)}{n}\right)\right). \qquad (4.21)
$$

The maximum likelihood estimate of $\boldsymbol{\theta}$ can then be expressed as

$$
\widehat{\theta}_L = \arg\min_\theta\left(-2\tilde{\ell}(\theta|\mathbf{y})\right). \qquad (4.22)
$$

from which the ML estimates of $\boldsymbol{\sigma}^2$ and $\boldsymbol{\beta}$ are evaluated as

$$
\begin{aligned}
\widehat{\sigma^2_L} &= \frac{\tilde{d}(\widehat{\theta}_L,\mathbf{y})}{n} \qquad (4.23)\\
\widehat{\beta}_L &= \tilde{\beta}(\widehat{\theta}_L). \qquad (4.24)
\end{aligned}
$$

The important thing to note about optimizing the profiled likelihood, (4.21), is that it is a m-dimensional optimization problem and typically m is very small.

4.3.3 The REML criterion

In practice 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,

$$\mathscr{Y} \sim \mathscr{N}(\mathbf{X}\beta, \sigma^2\mathbf{I}_n), \tag{4.25}$$

in which we typically estimate σ^2 by

$$\widehat{\sigma^2_R} = \frac{\|\mathbf{y}-\mathbf{X}\widehat{\beta}\|^2}{n-p} \tag{4.26}$$

even though the maximum likelihood estimate of σ^2 is

$$\widehat{\sigma^2_L} = \frac{\|\mathbf{y}-\mathbf{X}\widehat{\beta}\|^2}{n}. \tag{4.27}$$

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 $\mathbf{y}-\mathbf{X}\beta$ is an n-dimensional vector, the residual at $\widehat{\theta}$ satisfies p linearly independent constraints, $\mathbf{X}^{\mathsf{T}}(\mathbf{y}-\mathbf{X}\widehat{\beta}) = \mathbf{0}$. That is, the residual at $\widehat{\theta}$ is the projection of the observed response vector, $\mathbf{y}$, 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*.

The REML criterion for determining parameter estimates $\widehat{\theta}_R$ and $\widehat{\sigma^2_R}$ in a linear mixed model has the property that these estimates would specialize to $\widehat{\sigma^2_R}$ from (4.26) for a linear regression model. Although not usually derived in this way, the REML criterion can be expressed as

$$c_R(\theta,\sigma|\mathbf{y}) = -2\log\int_{\mathbb{R}^p} L(\mathbf{u}|\mathbf{y},\theta,\beta,\sigma)\,d\beta \tag{4.28}$$

on the deviance scale. The REML estimates $\widehat{\theta}_R$ and $\widehat{\sigma^2_R}$ minimize $c_R(\theta,\sigma|\mathbf{y})$.

The profiled REML criterion, a function of θ only, is

$$\tilde{c}_R(\theta|\mathbf{y}) = \log(|\mathbf{L_Z}(\theta)|^2|\mathbf{L_X}(\theta)|^2) + (n-p)\left(1+\log\left(\frac{2\pi\tilde{d}(\theta|\mathbf{y})}{n-p}\right)\right) \tag{4.29}$$

and the REML estimate of θ is

$$\widehat{\theta}_R = \arg\min_{\theta} \tilde{c}_R(\theta, \mathbf{y}). \tag{4.30}$$

The REML estimate of σ^2 is $\widehat{\sigma_R^2} = \tilde{d}(\widehat{\theta}_R|\mathbf{y})/(n-p)$.

It is not entirely clear how one would define a "REML estimate" of β because the REML criterion, $c_R(\theta,\sigma|\mathbf{y})$, defined in (4.28), does not depend on β. However, it is customary (and not unreasonable) to use $\widehat{\beta}_R = \tilde{\beta}(\widehat{\theta}_R)$ as the REML estimate of β.

Note that the profiled REML criterion can be evaluated from a sparse Cholesky decomposition like that in (4.15) but without the requirement that the permutation can be applied to the columns of $\mathbf{Z}\Lambda(\theta)$ separately from the columnns of $\mathbf{X}$. That is, we can use a general fill-reducing permutation rather than the specific form (4.14) with separate permutations represented by $\mathbf{P_Z}$ and $\mathbf{P_X}$. This can be useful in cases where both $\mathbf{Z}$ and $\mathbf{X}$ are large and sparse.

4.3.4 Summary for linear mixed models

A linear mixed model is characterized by the conditional distribution

$$(\mathscr{Y}|\mathscr{U}=\mathbf{u}) \sim \mathscr{N}(\gamma(\mathbf{u},\theta,\beta),\sigma^2\mathbf{I}_n) \text{ where } \gamma(\mathbf{u},\theta,\beta) = \mathbf{Z}\Lambda(\theta)\mathbf{u}+\mathbf{X}\beta \tag{4.31}$$

and the unconditional distribution $\mathscr{U} \sim \mathscr{N}(\mathbf{0},\sigma^2\mathbf{I}_q)$. The discrepancy function,

$$d(\mathbf{u}|\mathbf{y},\theta,\beta) = \|\mathbf{y}-\gamma(\mathbf{u},\theta,\beta)\|^2 + \|\mathbf{u}\|^2,$$

is minimized at the conditional mode, $\tilde{\mathbf{u}}(\theta)$, and the conditional estimate, $\tilde{\beta}(\theta)$, which are the solutions to the sparse, positive-definite linear system

$$\begin{bmatrix} \Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda(\theta)+\mathbf{I}_q & \Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}\mathbf{X} \\ \mathbf{X}^\mathsf{T}\mathbf{Z}\Lambda(\theta) & \mathbf{X}^\mathsf{T}\mathbf{X} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{u}}(\theta) \\ \tilde{\beta}(\theta) \end{bmatrix} = \begin{bmatrix} \Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}\mathbf{y} \\ \mathbf{X}^\mathsf{T}\mathbf{y} \end{bmatrix}.$$

In the process of solving this system we create the sparse left Cholesky factor, $L_\mathbf{Z}(\theta)$, which is a lower triangular sparse matrix satisfying

$$\mathbf{L_Z}(\theta)\mathbf{L_Z}(\theta)^\mathsf{T} = \mathbf{P_Z}\left(\Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda(\theta)+\mathbf{I}_q\right)\mathbf{P_Z}^\mathsf{T}$$

where $\mathbf{P_Z}$ is a permutation matrix representing a fill-reducing permutation formed from the pattern of nonzeros in $\mathbf{Z}\Lambda(\theta)$ for any θ not on the boundary of the parameter region. (The values of the nonzeros depend on θ but the pattern doesn't.)

The profiled log-likelihood, $\tilde{\ell}(\theta|\mathbf{y})$, is

$$-2\tilde{\ell}(\theta|\mathbf{y}) = \log(|\mathbf{L_Z}(\theta)|^2) + n\left(1+\log\left(\frac{2\pi\tilde{d}(\mathbf{y},\theta)}{n}\right)\right)$$

where $\tilde{d}(\mathbf{y},\theta) = d(\tilde{\mathbf{u}}(\theta)|\mathbf{y},\tilde{\beta}(\theta),\theta)$.

4.4 Generalizing the discrepancy function

Because one of the factors influencing the choice of implementation for linear mixed models is the extent to which the methods can also be applied to other mixed models, we describe several other classes of mixed models before discussing the implementation details for linear mixed models. At the core of our methods for determining the maximum likelihood estimates (MLEs) of the parameters in the mixed model are methods for minimizing the discrepancy function with respect to the coefficients $\mathbf{u}$ and β in the linear predictor $\gamma(\mathbf{u},\theta,\beta)$.

In this section we describe the general form of the discrepancy function that we will use and a penalized iteratively reweighted least squares (PIRLS) algorithm for determining the conditional modes $\tilde{\mathbf{u}}(\theta)$ and $\tilde{\beta}(\theta)$. We then describe several types of mixed models and the form of the discrepancy function for each.

4.4.1 A weighted residual sum of squares

As shown in §4.3.1, the discrepancy function for a linear mixed model has the form of a penalized residual sum of squares from a linear model (4.11). In this section we generalize that definition to

$$d(\mathbf{u}|\mathbf{y},\theta,\beta) = \left\| \mathbf{W}^{1/2}(\mu)\left[\mathbf{y} - \mu_{\mathscr{Y}|\mathscr{U}}(\mathbf{u},\theta,\beta)\right] \right\|^2 + \|\mathbf{0}-\mathbf{u}\|^2. \tag{4.32}$$

where $\mathbf{W}$ is an $n \times n$ diagonal matrix, called the *weights matrix*, with positive diagonal elements and $\mathbf{W}^{1/2}$ is the diagonal matrix with the square roots of the weights on the diagonal. The ith weight is inversely proportional to the conditional variances of $(\mathscr{Y}|\mathscr{U}=\mathbf{u})$ and may depend on the conditional mean, $\mu_{\mathscr{Y}|\mathscr{U}}$.

We allow the conditional mean to be a nonlinear function of the linear predictor, but with certain restrictions. We require that the mapping from $\mathbf{u}$ to $\mu_{\mathscr{Y}|\mathscr{U}=\mathbf{u}}$ be expressed as

$$\mathbf{u} \rightarrow \gamma \rightarrow \eta \rightarrow \mu \tag{4.33}$$

where $\gamma = \mathbf{Z}\Lambda(\theta)\mathbf{u} + \mathbf{X}\theta$ is an ns-dimensional vector ($s > 0$) while η and μ are n-dimensional vectors.

The map $\eta \rightarrow \mu$ has the property that μ_i depends only on η_i, $i = 1,\dots,n$. The map $\gamma \rightarrow \eta$ has a similar property in that, if we write γ as an $n \times s$ matrix

■ such that

$$\gamma = \blacksquare \tag{4.34}$$

(i.e. concatenating the columns of ■ produces γ) then η_i depends only on the ith row of ■, $i = 1, \dots, n$. Thus the Jacobian matrix $\frac{d\mu}{d\eta^{\mathsf{T}}}$ is an $n \times n$ diagonal matrix and the Jacobian matrix $\frac{d\eta}{d\gamma^{\mathsf{T}}}$ is the horizontal concatenation of s diagonal $n \times n$ matrices.

For historical reasons, the function that maps η_i to μ_i is called the *inverse link* function and is written $\mu = g^{-1}(\eta)$. The *link function*, naturally, is $\eta = g(\mu)$. When applied component-wise to vectors μ or η we write these as $\eta = \mathbf{g}(\mu)$ and $\mu = \mathbf{g}^{-1}(\eta)$.

Recall that the conditional distribution, $(\mathscr{Y}_i|\mathscr{U} = \mathbf{u})$, is required to be independent of $(\mathscr{Y}_j|\mathscr{U} = \mathbf{u})$ for $i, j = 1, \dots, n, i \neq j$ and that all the component conditional distributions must be of the same form and differ only according to the value of the conditional mean.

Depending on the family of the conditional distributions, the allowable values of the μ_i may be in a restricted range. For example, if the conditional distributions are Bernoulli then $0 \le \mu_i \le 1, i = 1, \dots, n$. If the conditional distributions are Poisson then $0 \le \mu_i, i = 1, \dots, n$. A characteristic of the link function, g, is that it must map the restricted range to an unrestricted range. That is, a link function for the Bernoulli distribution must map $[0,1]$ to $[-\infty, \infty]$ and must be invertible within the range.

The mapping from γ to η is defined by a function $m : \mathbb{R}^s \to \mathbb{R}$, called the *nonlinear model* function, such that $\eta_i = m(\gamma_i), i = 1, \dots, n$ where γ_i is the ith row of ■. The vector-valued function is $\eta = \mathbf{m}(\gamma)$.

Determining the conditional modes, $\tilde{\mathbf{u}}(\mathbf{y}|\theta)$, and $\tilde{\beta}(\mathbf{y}|\theta)$, that jointly minimize the discrepancy,

$$\begin{bmatrix} \tilde{\mathbf{u}}(\mathbf{y}|\theta) \\ \tilde{\beta}(\mathbf{y}|\theta) \end{bmatrix} = \arg\min_{\mathbf{u},\beta} \left[(\mathbf{y} - \mu)^{\mathsf{T}} \mathbf{W} (\mathbf{y} - \mu) + \|\mathbf{u}\|^2 \right] \tag{4.35}$$

becomes a weighted, nonlinear least squares problem except that the weights, $\mathbf{W}$, can depend on μ and, hence, on $\mathbf{u}$ and β.

In describing an algorithm for linear mixed models we called $\tilde{\beta}(\theta)$ the *conditional estimate*. That name reflects that fact that this is the maximum likelihood estimate of β for that particular value of θ. Once we have determined the MLE, $\widehat{(\theta)}_L$ of θ, we have a "plug-in" estimator, $\widehat{\beta}_L = \tilde{\beta}(\theta)$ for β.

This property does not carry over exactly to other forms of mixed models. The values $\tilde{\mathbf{u}}(\theta)$ and $\tilde{\beta}(\theta)$ are conditional modes in the sense that they are the coefficients in γ that jointly maximize the unscaled conditional density $h(\mathbf{u}|\mathbf{y}, \theta, \beta, \sigma)$. Here we are using the adjective "conditional" more in the sense of conditioning on $\mathscr{Y} = \mathbf{y}$ than in the sense of conditioning on θ, although these values are determined for a fixed value of θ.

4.4.2 The PIRLS algorithm for $\tilde{\mathbf{u}}$ and $\tilde{\beta}$

The penalized, iteratively reweighted, least squares (PIRLS) algorithm to determine $\tilde{\mathbf{u}}(\boldsymbol{\theta})$ and $\tilde{\beta}(\boldsymbol{\theta})$ is a form of the Fisher scoring algorithm. We fix the weights matrix, $\mathbf{W}$, and use penalized, weighted, nonlinear least squares to minimize the penalized, weighted residual sum of squares conditional on these weights. Then we update the weights to those determined by the current value of μ and iterate.

To describe this algorithm in more detail we will use parenthesized superscripts to denote the iteration number. Thus $\mathbf{u}^{(0)}$ and $\beta^{(0)}$ are the initial values of these parameters, while $\mathbf{u}^{(i)}$ and $\beta^{(i)}$ are the values at the ith iteration. Similarly $\gamma^{(i)} = \mathbf{Z}\Lambda(\theta)\mathbf{u}^{(i)} + \mathbf{X}\beta^{(i)}$, $\eta^{(i)} = \mathbf{m}(\gamma^{(i)})$ and $\mu^{(i)} = \mathbf{g}^{-1}(\eta^{(i)})$.

We use a penalized version of the Gauss-Newton algorithm (1, ch.˜2) for which we define the weighted Jacobian matrices

$$\mathbf{U}^{(i)} = \mathbf{W}^{1/2} \left.\frac{d\mu}{d\mathbf{u}^\mathsf{T}}\right|_{\mathbf{u}=\mathbf{u}^{(i)},\beta=\beta^{(i)}} = \mathbf{W}^{1/2} \left.\frac{d\mu}{d\eta^\mathsf{T}}\right|_{\eta^{(i)}} \left.\frac{d\eta}{d\gamma^\mathsf{T}}\right|_{\gamma^{(i)}} \mathbf{Z}\Lambda(\theta) \tag{4.36}$$

$$\mathbf{V}^{(i)} = \mathbf{W}^{1/2} \left.\frac{d\mu}{d\beta^\mathsf{T}}\right|_{\mathbf{u}=\mathbf{u}^{(i)},\beta=\beta^{(i)}} = \mathbf{W}^{1/2} \left.\frac{d\mu}{d\eta^\mathsf{T}}\right|_{\eta^{(i)}} \left.\frac{d\eta}{d\gamma^\mathsf{T}}\right|_{\gamma^{(i)}} \mathbf{X} \tag{4.37}$$

of dimension $n \times q$ and $n \times p$, respectively. The increments at the ith iteration, $\delta_{\mathbf{u}}^{(i)}$ and $\delta_{\beta}^{(i)}$, are the solutions to

$$\begin{bmatrix} \mathbf{U}^{(i)}{}^\mathsf{T}\mathbf{U}^{(i)} + \mathbf{I}_q & \mathbf{U}^{(i)}{}^\mathsf{T}\mathbf{V}^{(i)} \\ \mathbf{V}^{(i)}{}^\mathsf{T}\mathbf{U}^{(i)} & \mathbf{V}^{(i)}{}^\mathsf{T}\mathbf{V}^{(i)} \end{bmatrix} \begin{bmatrix} \delta_{\mathbf{u}}^{(i)} \\ \delta_{\beta}^{(i)} \end{bmatrix} = \begin{bmatrix} \mathbf{U}^{(i)}{}^\mathsf{T}\mathbf{W}^{1/2}(\mathbf{y} - \mu^{(i)}) - \mathbf{u}^{(i)} \\ \mathbf{U}^{(i)}{}^\mathsf{T}\mathbf{W}^{1/2}(\mathbf{y} - \mu^{(i)}) \end{bmatrix} \tag{4.38}$$

providing the updated parameter values

$$\begin{bmatrix} \mathbf{u}^{(i+1)} \\ \beta^{(i+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{u}^{(i)} \\ \beta^{(i)} \end{bmatrix} + \lambda \begin{bmatrix} \delta_{\mathbf{u}}^{(i)} \\ \delta_{\beta}^{(i)} \end{bmatrix} \tag{4.39}$$

where $\lambda > 0$ is a step factor chosen to ensure that

$$(\mathbf{y} - \mu^{(i+1)})^\mathsf{T}\mathbf{W}(\mathbf{y} - \mu^{(i+1)}) + \|\mathbf{u}^{(i+1)}\|^2 < (\mathbf{y} - \mu^{(i)})^\mathsf{T}\mathbf{W}(\mathbf{y} - \mu^{(i)}) + \|\mathbf{u}^{(i)}\|^2. \tag{4.40}$$

In the process of solving for the increments we form the sparse, lower triangular, Cholesky factor, $\mathbf{L}^{(i)}$, satisfying

$$\mathbf{L}^{(i)}\mathbf{L}^{(i)}{}^\mathsf{T} = \mathbf{P}_\mathbf{Z}\left(\mathbf{U}^{(i)}{}^\mathsf{T}\mathbf{U}^{(i)} + \mathbf{I}_n\right)\mathbf{P}_\mathbf{Z}^\mathsf{T}. \tag{4.41}$$

After each successful iteration, determining new values of the coefficients, $\mathbf{u}^{(i+1)}$ and $\beta^{(i+1)}$, that reduce the penalized, weighted residual sum of squqres, we update the weights matrix to $\mathbf{W}(\mu^{(i+1)})$ and the weighted Jacobians, $\mathbf{U}^{(i+1)}$ and $\mathbf{V}^{(i+1)}$, then iterate. Convergence is determined according to the orthogonality convergence criterion~(1, ch.~2), suitably adjusted for the weights matrix and the penalty.

4.4.3 Weighted linear mixed models

One of the simplest generalizations of linear mixed models is a weighted linear mixed model where $s = 1$, the link function, g, and the nonlinear model function, m, are both the identity, the weights matrix, $\mathbf{W}$, is constant and the conditional distribution family is Gaussian. That is, the conditional distribution can be written

$$(\mathscr{Y}|\mathscr{U} = \mathbf{u}) \sim \mathscr{N}(\gamma(\mathbf{u}, \theta, \beta), \sigma^2\mathbf{W}^{-1}) \tag{4.42}$$

with discrepancy function

$$d(\mathbf{u}|\mathbf{y}, \theta, \beta) = \left\| \mathbf{W}^{1/2}(\mathbf{y} - \mathbf{Z}\Lambda(\theta)\mathbf{u} - \mathbf{X}\theta) \right\|^2 + \|\mathbf{u}\|^2. \tag{4.43}$$

The conditional mode, $\tilde{\mathbf{u}}(\theta)$, and the conditional estimate, $\tilde{\beta}(\theta)$, are the solutions to

$$\begin{bmatrix} \Lambda^{\mathsf{T}}(\theta)\mathbf{Z}^{\mathsf{T}}\mathbf{W}\mathbf{Z}\Lambda(\theta) + \mathbf{I}_q & \Lambda^{\mathsf{T}}(\theta)\mathbf{Z}^{\mathsf{T}}\mathbf{W}\mathbf{X} \\ \mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{Z}\Lambda(\theta) & \mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{u}}(\theta) \\ \tilde{\beta}(\theta) \end{bmatrix} = \begin{bmatrix} \Lambda^{\mathsf{T}}(\theta)\mathbf{Z}^{\mathsf{T}}\mathbf{W}\mathbf{y} \\ \mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{y} \end{bmatrix}, \tag{4.44}$$

which can be solved directly, and the Cholesky factor, $\mathbf{L}_{\mathbf{Z}}(\theta)$, satisfies

$$\mathbf{L}_{\mathbf{Z}}(\theta)\mathbf{L}_{\mathbf{Z}}(\theta)^{\mathsf{T}} = \mathbf{P}_{\mathbf{Z}} \left(\Lambda^{\mathsf{T}}(\theta)\mathbf{Z}^{\mathsf{T}}\mathbf{W}\mathbf{Z}\Lambda(\theta) + \mathbf{I}_q \right) \mathbf{P}_{\mathbf{Z}}^{\mathsf{T}}. \tag{4.45}$$

The profiled log-likelihood, $\tilde{\ell}(\theta|\mathbf{y})$, is

$$-2\tilde{\ell}(\theta|\mathbf{y}) = \log\left(\frac{|\mathbf{L}_{\mathbf{Z}}(\theta)|^2}{|\mathbf{W}|} \right) + n\left(1 + \log\left(\frac{2\pi\tilde{d}(\mathbf{y}, \theta)}{n} \right) \right). \tag{4.46}$$

If the matrix $\mathbf{W}$ is fixed then we can ignore the term $|\mathbf{W}|$ in (4.46) when determining the MLE, $\widehat{\theta}_L$. However, in some models, we use a parameterized weight matrix, $\mathbf{W}(\phi)$, and wish to determine the MLEs, $\widehat{\phi}_L$ and $\widehat{\theta}_L$ simultaneously. In these cases we must include the term involving $|\mathbf{W}(\phi)|$ when evaluating the profiled log-likelihood.

Note that we must define the parameterization of $\mathbf{W}(\phi)$ such that σ^2 and ϕ are not a redundant parameterization of $\sigma^2\mathbf{W}(\phi)$. For example, we could require that the first diagonal element of $\mathbf{W}$ be unity.

4.4.4 Nonlinear mixed models

In an unweighted, nonlinear mixed model the conditional distribution is Gaussian, the link, g, is the identity and the weights matrix, $\mathbf{W} = \mathbf{I}_n$. That is,

$$(\mathscr{Y}|\mathscr{U} = \mathbf{u}) \sim \mathscr{N}(\mathbf{m}(\gamma), \sigma^2\mathbf{I}_n) \tag{4.47}$$

with discrepancy function

$$d(\mathbf{u}|\mathbf{y},\theta,\beta) = \|\mathbf{y}-\mu\|^2 + \|\mathbf{u}\|^2. \tag{4.48}$$

For a given value of θ we determine the conditional modes, $\tilde{\mathbf{u}}(\theta)$ and $\tilde{\beta}(\theta)$, as the solution to the penalized nonlinear least squares problem

$$\begin{bmatrix}\tilde{\mathbf{u}}(\theta)\\ \tilde{\beta}(\theta)\end{bmatrix} = \arg\min_{\mathbf{u},\theta} d(\mathbf{u}|\mathbf{y},\theta,\beta) \tag{4.49}$$

and we write the minimum discrepancy, given $\mathbf{y}$ and θ, as

$$\tilde{d}(\mathbf{y},\theta) = d(\tilde{\mathbf{u}}(\theta)|\mathbf{y},\theta,\tilde{\beta}(\theta)). \tag{4.50}$$

Let $\tilde{\mathbf{L}}_Z(\theta)$ and $\tilde{\mathbf{L}}_X(\theta)$ be the Cholesky factors at θ, $\tilde{\beta}(\theta)$ and $\tilde{\mathbf{u}}(\theta)$. Then the *Laplace approximation* to the log-likelihood is

$$-2\ell_P(\theta,\beta,\sigma|\mathbf{y}) \approx n\log(2\pi\sigma^2) + \log(|\tilde{\mathbf{L}}_\mathbf{Z}|^2) + \frac{\tilde{d}(\mathbf{y},\theta) + \left\|\tilde{\mathbf{L}}_\mathbf{X}^\mathsf{T}(\beta-\tilde{\beta})\right\|^2}{\sigma^2}, \tag{4.51}$$

producing the approximate profiled log-likelihood, $\tilde{\ell}_P(\theta|\mathbf{y})$,

$$-2\tilde{\ell}_P(\theta|\mathbf{y}) \approx \log(|\tilde{\mathbf{L}}_\mathbf{Z}|^2) + n\left(1+\log(2\pi\tilde{d}(\mathbf{y},\theta)/n)\right). \tag{4.52}$$

4.4.4.1 Nonlinear mixed model summary

In a nonlinear mixed model we determine the parameter estimate, $\widehat{\theta}_P$, from the Laplace approximation to the log-likelihood as

$$\widehat{\theta}_P = \arg\max_\theta \tilde{\ell}_P(\theta|\mathbf{y}) = \arg\min_\theta \log(|\tilde{\mathbf{L}}_\mathbf{Z}|^2) + n\left(1+\log(2\pi\tilde{d}(\mathbf{y},\theta)/n)\right). \tag{4.53}$$

Each evaluation of $\tilde{\ell}_P(\theta|\mathbf{y})$ requires a solving the penalized nonlinear least squares problem (4.49) simultaneously with respect to both sets of coefficients, $\mathbf{u}$ and β, in the linear predictor, γ.

For a weighted nonlinear mixed model with fixed weights, $\mathbf{W}$, we replace the unweighted discrepancy function $d(\mathbf{u}|\mathbf{y},\theta,\beta)$ with the weighted discrepancy function,

4.5 Details of the implementation

4.5.1 Implementation details for linear mixed models

The crucial step in implementing algorithms for determining ML or REML estimates of the parameters in a linear mixed model is evaluating the factorization (4.15) for any $\boldsymbol{\theta}$ satisfying $\boldsymbol{\theta}_L \leq \boldsymbol{\theta} \leq \boldsymbol{\theta}_U$. We will assume that $\mathbf{Z}$ is sparse as is $\mathbf{Z}\Lambda(\boldsymbol{\theta})$.

When $\mathbf{X}$ is not sparse we will use the factorization (4.15) setting $\mathbf{P_X} = \mathbf{I}_p$ and storing $\mathbf{L_{XZ}}$ and $\mathbf{L_X}$ as dense matrices. The permutation matrix $\mathbf{P_Z}$ is determined from the pattern of non-zeros in $\mathbf{Z}\Lambda(\boldsymbol{\theta})$ which is does not depend on $\boldsymbol{\theta}$, as long as $\boldsymbol{\theta}$ is not on the boundary. In fact, in most cases the pattern of non-zeros in $\mathbf{Z}\Lambda(\boldsymbol{\theta})$ is the same as the pattern of non-zeros in $\mathbf{Z}$. For many models, in particular models with scalar random effects (described later), the matrix $\Lambda(\boldsymbol{\theta})$ is diagonal.

Given a value of $\boldsymbol{\theta}$ we determine the Cholesky factor $\mathbf{L_Z}$ satisfying

$$\mathbf{L_Z}\mathbf{L_Z}^{\mathsf{T}} = \mathbf{P_Z}(\Lambda^{\mathsf{T}}(\theta)\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\Lambda(\theta) + \mathbf{I}_q)\mathbf{P_Z}^{\mathsf{T}}. \tag{4.54}$$

The CHOLMOD package allows for $\mathbf{L_Z}$ to be calculated directly from $\Lambda^{\mathsf{T}}(\theta)\mathbf{Z}^{\mathsf{T}}$ or from $\Lambda^{\mathsf{T}}(\theta)\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\Lambda(\theta)$. The choice in implementation is whether to store $\mathbf{Z}^{\mathsf{T}}$ and update it to $\Lambda^{\mathsf{T}}(\theta)\mathbf{Z}$ or to store $\mathbf{Z}^{\mathsf{T}}\mathbf{Z}$ and use it to form $\Lambda^{\mathsf{T}}(\theta)\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\Lambda(\theta)$ at each evaluation.

In the `lme4` package we store $\mathbf{Z}^{\mathsf{T}}$ and use it to form $\Lambda^{\mathsf{T}}(\theta)\mathbf{Z}^{\mathsf{T}}$ from which $\mathbf{L_Z}$ is evaluated. There are two reasons for this choice. First, the calculations for the more general forms of mixed models cannot be reduced to calculations involving $\mathbf{Z}^{\mathsf{T}}\mathbf{Z}$ and by expressing these calculations in terms of $\Lambda(\boldsymbol{\theta})\mathbf{Z}^{\mathsf{T}}$ for linear mixed models we can reuse the code for the more general models. Second, the calculation of $\Lambda(\boldsymbol{\theta})^{\mathsf{T}}\left(\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\right)\Lambda(\boldsymbol{\theta})$ from $\mathbf{Z}^{\mathsf{T}}\mathbf{Z}$ is complicated compared to the calculation of $\Lambda(\boldsymbol{\theta})^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}$ from $\mathbf{Z}^{\mathsf{T}}$.

This choice is disadvantageous when $n \gg q$ because $\mathbf{Z}^{\mathsf{T}}$ is much larger than $\mathbf{Z}^{\mathsf{T}}\mathbf{Z}$, even when they are stored as sparse matrices. Evaluation of $\mathbf{L_Z}$ directly from $\mathbf{Z}^{\mathsf{T}}$ requires more storage and more calculation that evaluating $\mathbf{L_Z}$ from $\mathbf{Z}^{\mathsf{T}}\mathbf{Z}$.

Next we evaluate $\mathbf{L_{XZ}}^{\mathsf{T}}$ as the solution to

$$\mathbf{L_Z}\mathbf{L_{XZ}}^{\mathsf{T}} = \mathbf{P_Z}\Lambda^{\mathsf{T}}(\theta)\mathbf{Z}^{\mathsf{T}}\mathbf{X}. \tag{4.55}$$

Again we have the choice of calculating and storing $\mathbf{Z}^{\mathsf{T}}\mathbf{X}$ or storing $\mathbf{X}$ and using it to reevaluate $\mathbf{Z}^{\mathsf{T}}\mathbf{X}$. In the `lme4` package we store $\mathbf{X}$, because the calculations for the more general models cannot be expressed in terms of $\mathbf{Z}^{\mathsf{T}}\mathbf{X}$.

Finally $\mathbf{L_X}$ is evaluated as the (dense) solution to

$$\mathbf{L_X}\mathbf{L_X}^{\mathsf{T}} = \mathbf{X}^{\mathsf{T}}\mathbf{X} - \mathbf{L_{XZ}}\mathbf{L_{XZ}}. \tag{4.56}$$

from which $\tilde{\beta}$ can be determined as the solution to dense system

$$\mathbf{L_X L_X}\tilde{\beta} = \mathbf{X}^{\mathsf{T}}\mathbf{y} \tag{4.57}$$

and $\tilde{\mathbf{u}}$ as the solution to the sparse system

$$\mathbf{L_Z L_Z}\tilde{u} = \Lambda^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}\mathbf{y} \tag{4.58}$$

For many models, in particular models with scalar random effects, which are described later, the matrix $\Lambda(\boldsymbol{\theta})$ is diagonal. For such a model, if both $\mathbf{Z}$ and $\mathbf{X}$ are sparse and we plan to use the REML criterion then we create and store

$$\mathbf{A} = \begin{bmatrix} \mathbf{Z}^{\mathsf{T}}\mathbf{Z} & \mathbf{Z}^{\mathsf{T}}\mathbf{X} \\ \mathbf{X}^{\mathsf{T}}\mathbf{Z} & \mathbf{X}^{\mathsf{T}}\mathbf{X} \end{bmatrix} \quad \text{and} \quad \mathbf{c} = \begin{bmatrix} \mathbf{Z}^{\mathsf{T}}\mathbf{y} \\ \mathbf{X}^{\mathsf{T}}\mathbf{y} \end{bmatrix} \tag{4.59}$$

and determine a fill-reducing permutation, $\mathbf{P}$, for $\mathbf{A}$. Given a value of $\boldsymbol{\theta}$ we create the factorization

$$\mathbf{L}(\theta)\mathbf{L}(\theta)^{\mathsf{T}} = \mathbf{P}\left(\begin{bmatrix} \Lambda(\theta) & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_p \end{bmatrix} \mathbf{A} \begin{bmatrix} \Lambda(\theta) & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_p \end{bmatrix} + \begin{bmatrix} \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right)\mathbf{P}^{\mathsf{T}} \tag{4.60}$$

solve for $\tilde{\mathbf{u}}(\boldsymbol{\theta})$ and $\tilde{\beta}(\boldsymbol{\theta})$ in

$$\mathbf{L}\mathbf{L}^{\mathsf{T}}\mathbf{P}\begin{bmatrix} \tilde{\mathbf{u}}(\theta) \\ \tilde{\beta}(\theta) \end{bmatrix} = \mathbf{P}\begin{bmatrix} \Lambda(\theta) & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_p \end{bmatrix}\mathbf{c} \tag{4.61}$$

then evaluate $\tilde{d}(\mathbf{y}|\boldsymbol{\theta})$ and the profiled REML criterion as

$$\tilde{d}_R(\boldsymbol{\theta}|\mathbf{y}) = \log(|\mathbf{L}(\theta)|^2) + (n-p)\left(1 + \log\left(\frac{2\pi\tilde{d}(\mathbf{y}|\boldsymbol{\theta})}{n-p}\right)\right). \tag{4.62}$$

References

[1] Bates, D.M., Watts, D.G.: Nonlinear Regression Analysis and Its Applications. Wiley, Hoboken, NJ (1988)

[2] Belenky, G., Wesensten, N.J., Thorne, D.R., Thomas, M.L., Sing, H.C., Redmond, D.P., Russo, M.B., Balkin, T.J.: 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)

[3] Box, G., Tiao, G.: Bayesian Inference in Statistical Analysis. Addison-Wesley, Reading, MA (1973)

[4] Cleveland, B.: Visualizing Data. Hobart Press, Summit, NJ (1993)

[5] Davies, O.L., Goldsmith, P.L. (eds.): Statistical Methods in Research and Production, 4th edn. Hafner (1972)

[6] Davis, T.: An approximate minimal degree ordering algorithm. SIAM J. Matrix Analysis and Applications **17**(4), 886–905 (1996)

[7] Davis, T.: CHOLMOD: sparse supernodal Cholesky factorization and update/downdate. http://www.cise.ufl.edu/research/sparse/cholmod (2005)

[8] Davis, T.: Direct Methods for Sparse Linear Systems. SIAM, Philadelphia, PA (2006)

[9] Leisch, F.: Sweave: Dynamic generation of statistical reports using literate data analysis. In: W.~Härdle, B.~Rönz (eds.) Compstat 2002 — Proceedings in Computational Statistics, pp. 575–580. Physica Verlag, Heidelberg (2002). URL `http://www.stat.uni-muenchen.de/~leisch/Sweave`. ISBN 3-7908-1517-9

[10] Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I.: A User's Guide to MLwiN. Multilevel Models Project, Institute of Education, University of London, London (2000)

[11] Raudenbush, S.W., Bryk, A.S.: Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd edn. Sage (2002)

[12] Sakamoto, Y., Ishiguro, M., Kitagawa, G.: Akaike Information Criterion Statistics. Reidel, Dordrecht, Holland (1986)

[13] Sarkar, D.: Lattice: Multivariate Data Visualization with R. Springer (2008)

[14] Schwarz, G.: Estimating the dimension of a model. Annals of Statistics **6**, 461–464 (1978)

Index