Chapter 2

Linear Mixed Models

In general, linear mixed models extend the linear model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2),$$

to

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \quad \mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\psi}_\theta), \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Lambda}_\theta),$$

where random vector, $\mathbf{b}$, contains *random effects*, with zero expected value and covariance matrix $\boldsymbol{\psi}_\theta$, with unknown parameters in $\boldsymbol{\theta}$; $\mathbf{Z}$ is a model matrix for the random effects. $\boldsymbol{\Lambda}_\theta$ is a positive definite matrix which can be used to model residual autocorrelation: its elements are usually determined by some simple model, with few (or no) unknown parameters (here considered to be part of $\boldsymbol{\theta}$). Often $\boldsymbol{\Lambda}_\theta = \mathbf{I}\sigma^2$. $\mathbf{b}$ and $\boldsymbol{\epsilon}$ are independent.

The idea is to allow a linear model structure for the random component of the response data, $\mathbf{y}$, which is as rich as the linear model structure used to model the systematic component. Except in the special case of the analysis of balanced designed experiments, inference for such models relies on some general theory for maximum likelihood estimation. It is also worth noting the resemblance of the distributional assumption for $\mathbf{b}$ to the specification of a prior in Bayesian analysis: some of the computations performed with this model will also correspond to Bayesian computations, although the inferential framework is not really Bayesian.

2.1 Mixed models for balanced data

This section briefly covers why random effects models are useful, and the topic of linear mixed models for balanced data. The latter is not essential to the rest of this book, but does provide an understanding of the degrees of freedom used in classical mixed model ANOVA computations, for example. Readers who skip straight to section 2.3 will not miss anything essential to the rest of the book.

2.1.1 A motivating example

Plant leaves have tiny holes, called 'stomata', through which they take up air, but also lose water. Most non-tropical plants photosynthesize in such a way that, on sunny days, they are limited by how much carbon dioxide they can obtain through

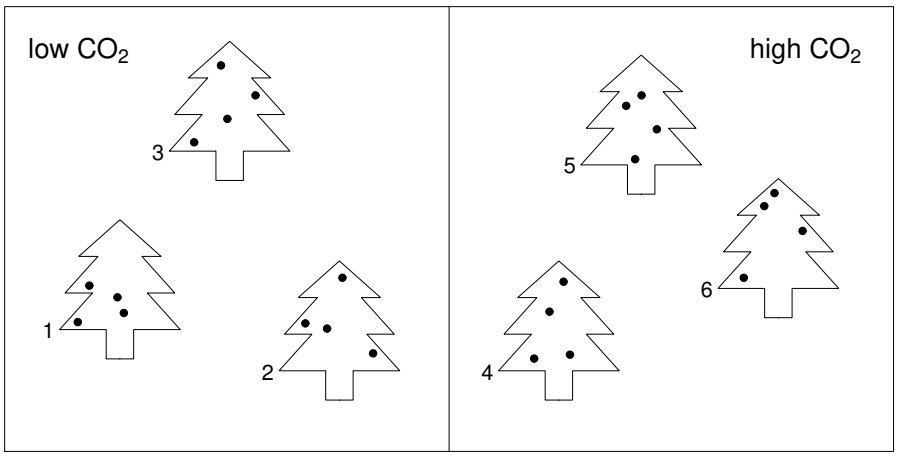

Figure 2.1 *Schematic diagram of the* CO_2 *experiment.*

these stomata. The 'problem' for a plant is that if its stomata are too small, it will not be able to get enough carbon dioxide, and if they are too large it will lose too much water on sunny days. Given the importance of this to such plants, it seems likely that stomatal size will depend on the concentration of carbon dioxide in the atmosphere. This may have climate change implications if increasing the amount of CO_2 in the atmosphere causes plants to release less water: water vapour is the most important greenhouse gas.

Consider an experiment* in which tree seedlings are grown under 2 levels of carbon dioxide concentration, with 3 trees assigned to each treatment, and suppose that after 6 months' growth stomatal area is measured at each of 4 random locations on each plant (the sample sizes here are artificially small). Figure 2.1 shows the experimental layout, schematically.

The wrong approach: A fixed effects linear model

A model of these data should include a (2 level) factor for CO_2 treatment, but also a (6 level) factor for individual tree, since we have multiple measurements on each tree and must expect some variability in stomatal area from tree to tree. So a suitable linear model is

$$y_i = \alpha_j + \beta_k + \epsilon_i \text{ if observation } i \text{ is for } CO_2 \text{ level } j, \text{ tree } k,$$

where y_i is the i^{th} stomatal area measurement, α_j is the population mean stomatal area at CO_2 level j, β_k is the deviation of tree k from that mean and the ϵ_i are independent $N(0, \sigma^2)$ random variables. Now if this is a fixed effects model, we have two problems:

1. The α_j's and β_k's are completely confounded. Trees are 'nested' within treatment,

*One important part of the design of such an experiment would be to ensure that the trees are grown under natural, *variable* light levels. At constant *average* light levels the plants are not CO_2 limited.

with 3 trees in one treatment and 3 in the other: any number you like could be added to α_1 and simultaneously subtracted from β_1, β_2 and β_3, without changing the model predictions at all, and the same goes for α_2 and the remaining β_k's.

2. We really want to learn about trees in general, but this is not possible with a model in which there is a fixed effect for each particular tree: unless the tree effects happen to turn out to be negligible, we cannot use the model to predict what happens to a tree other than one of the six in the experiment.

The following R session illustrates problem 1. First compare models with and without the tree factor (β_k):

```
> m1 <- lm(area ~ CO2 + tree, stomata)
> m0 <- lm(area ~ CO2, stomata)
> anova(m0,m1)
Analysis of Variance Table

Model 1: area ~ CO2
Model 2: area ~ CO2 + tree
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     22 2.1348
2     18 0.8604  4    1.2744 6.6654 0.001788 **
```

Clearly, there is strong evidence for tree to tree differences, which means that with this model we can not tell whether CO_2 had an effect or not. To re-emphasize this point, here is what happens if we attempt to test for a CO_2 effect:

```
> m2 <- lm(area ~ tree, stomata)
> anova(m2,m1)
Analysis of Variance Table

Model 1: area ~ tree
Model 2: area ~ CO2 + tree
  Res.Df    RSS Df  Sum of Sq F Pr(>F)
1     18 0.8604
2     18 0.8604  0 -2.220e-16
```

The confounding of the CO_2 and tree factors means that the models being compared here are really the same model: as a result, they give the same residual sum of squares and have the same residual degrees of freedom — 'comparing' them tells us nothing about the effect of CO_2.

In many ways this problem comes about because our model is simply too flexible. Individual tree effects are allowed to take any value whatsoever, which amounts to saying that each individual tree is completely different to every other individual tree: having results for 6 trees will tell us nothing whatsoever about a 7^{th}. This is not a sensible starting point for a model aimed at analysing data like these. We really expect trees of a particular species to behave in broadly similar ways so that a representative (preferably random) sample of trees, from the wider population of such trees, *will* allow us to make inferences about that wider population of trees. Treating the individual trees, not as completely unique individuals but as a random sample from the target population of trees, will allow us to estimate the CO_2 effect *and* to generalize beyond the 6 trees in the experiment.

The right approach: A mixed effects model

The key to establishing whether CO_2 has an effect is to recognise that the CO_2 factor and tree factors are different in kind. The CO_2 effects are fixed characteristics of the whole population of trees that we are trying to learn about. In contrast, the tree effect will vary randomly from tree to tree in the population. We are not primarily interested in the values of the tree effect for the particular trees used in the experiment: if we had used a different 6 trees these effects would have taken different values anyway. But we can not simply ignore the tree effect without inducing dependence between the response observations (area), and hence violating the independence assumption of the linear model. In this circumstance it makes sense to model the distribution of tree effects across the population of trees, and to suppose that the particular tree effects that occur in the experiment are just independent observations from this distribution. That is, the CO_2 effect will be modelled as a fixed effect, but the tree effect will be modelled as a random effect. Here is a model set up in this way:

$$y_i = \alpha_j + b_k + \epsilon_i \text{ if observation } i \text{ is for } CO_2 \text{ level } j \text{ and tree } k, \tag{2.1}$$

where $b_k \sim N(0, \sigma_b^2)$, $\epsilon_i \sim N(0, \sigma^2)$ and all the b_k and ϵ_i are mutually independent random variables. Now testing for tree effects can proceed exactly as it did in the fixed effects case, by comparing the least squares fits of models with and without the tree effects. But this mixed effects model also lets us test CO_2 effects, whether or not there is evidence for a tree effect.

All that is required is to average the data at each level of the random effect, i.e., at each tree. For balanced data, such as we have here, the key feature of a mixed effects model is that this 'averaging out' of a random effect automatically implies a simplified mixed effects model for the aggregated data: the random effect is absorbed into the independent residual error term. It is easy to see that the model for the average stomatal area per tree must be

$$\bar{y}_k = \alpha_j + e_k \text{ if tree } k \text{ is for } CO_2 \text{ level } j, \tag{2.2}$$

where the e_k are independent $N(0, \sigma_b^2 + \sigma^2/4)$ random variables.

Now it is a straightforward matter to test for a CO_2 effect in R. First aggregate the data for each tree:

```
> st <- aggregate(data.matrix(stomata),
+       by=list(tree=stomata$tree),mean)
> st$CO2 <- as.factor(st$CO2);st
  tree     area CO2 tree
1    1 1.623374   1    1
2    2 1.598643   1    2
3    3 1.162961   1    3
4    4 2.789238   2    4
5    5 2.903544   2    5
6    6 2.329761   2    6
```

and then fit the model implied by the aggregation.

```
> m3 <- lm(area ~ CO2, st)
```

```
> anova(m3)
Analysis of Variance Table

Response: area
          Df  Sum Sq Mean Sq F value   Pr(>F)
CO2        1 2.20531 2.20531  27.687 0.006247 **
Residuals  4 0.31861 0.07965
```

There is strong evidence for a CO_2 effect here, and we would now proceed to examine the estimate of this fixed effect (e.g., using `summary(m3)`). Usually with a mixed model the variances of the random effects are of more interest than the effects themselves, so in this example σ_b^2 should be estimated.

Let RSS_i stand for the residual sum of squares for model i. From the usual theory of linear models we have that:

$$\hat{\sigma}^2 = \mathrm{RSS}_1/18$$

(RSS_1 is the residual sum of squares from fitting (2.1)) and

$$\widehat{\sigma_b^2 + \sigma^2}/4 = \mathrm{RSS}_3/4$$

(RSS_3 is the residual sum of squares from fitting (2.2)). Both estimators are unbiased. Hence, an unbiased estimator for σ_b^2 is

$$\hat{\sigma}_b^2 = \mathrm{RSS}_3/4 - \mathrm{RSS}_1/72.$$

This can easily be evaluated in R.

```
> summary(m3)$sigma^2 - summary(m1)$sigma^2/4
[1] 0.06770177
```

2.1.2 General principles

To see how the ideas from the previous section generalize, consider data from a designed experiment and an associated linear mixed model for the data, in which the response variable depends only on factor variables and their interactions (which may be random or fixed). Assume that the data are *balanced* with respect to the model, meaning that for each factor or interaction in the model, the same number of data have been collected at each of its levels. In this case:

- Aggregated data, obtained by averaging the response at each level of any factor or interaction, will be described by a mixed model, derived from the original mixed model by the averaging process.
- Models for different aggregations will enable inferences to be made about different fixed and random factors, using standard methods for ordinary linear models. Note that not all aggregations will be useful, and the random effects themselves can not be 'estimated' in this way.
- The variances of the random effects can be estimated from combinations of the the usual residual variance estimates from models for different aggregations.

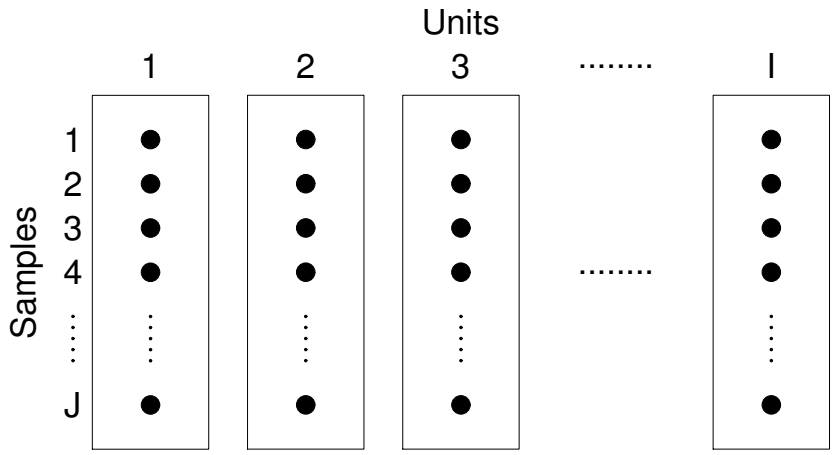

Figure 2.2 *Schematic illustration of the balanced one-way experimental layout discussed in section 2.1.3. Rectangles are experimental units and •'s indicate measurements.*

These principles are useful for two reasons. Firstly, the classical mixed model analyses for designed experiments can be derived using them. Secondly, they provide a straightforward explanation for the degrees of freedom of the reference distributions used in mixed model hypothesis testing: the degrees of freedom are always those that apply to the aggregated model appropriate for testing hypotheses about the effect concerned. For example, in the CO_2 analysis the hypothesis tests about the CO_2 effect were conducted with reference to an $F_{1,4}$ distribution, with these degrees of freedom being those appropriate to the aggregated model, used for the test.

To illustrate and reinforce these ideas two further simple examples of the analysis of 'standard designs' will be covered, before returning to the general mixed models that are of more direct relevance to GAMs.

2.1.3 A single random factor

Consider an experimental design in which you have J measurements from each of I units, illustrated schematically in figure 2.2. Suppose that we are interested in establishing whether there are differences between the units, but are not really interested in the individual unit effects: rather in quantifying how much variability can be ascribed to differences between units. This would suggest using a random effect term for units.

A concrete example comes from animal breeding. For a breeding program to be successful we need to know that variability in the targeted trait has a substantial enough genetic component that we can expect to alter it by selective breeding. Consider a pig breeding experiment in which I pregnant sows are fed a standard diet, and the fat content of J of each of their piglets is measured. The interesting questions here are: is there evidence for litter to litter variability in fat content (which would be consistent with genetic variation in this trait) and if so how large is this component, in relation to the piglet to piglet variability within a litter? Notice here that we are

not interested in how piglet fat content varies from particular sow to particular sow in the experiment, but rather in the variability between sows in general. This suggests using a random effect for sow in a model for such data.

So a suitable model is

$$y_{ij} = \alpha + b_i + \epsilon_{ij}, \tag{2.3}$$

where α is the fixed parameter for the population mean, $i = 1 \ldots I$, $j = 1 \ldots J$, $b_i \sim N(0, \sigma_b^2)$, $\epsilon_{ij} \sim N(0, \sigma^2)$ and all the b_i and ϵ_{ij} terms are mutually independent.

The first question of interest is whether $\sigma_b^2 > 0$, i.e., whether there is evidence that the factor variable contributes to the variance of the response. Formally we would like to test $\mathrm{H}_0 : \sigma_b^2 = 0$ against $\mathrm{H}_1 : \sigma_b^2 > 0$. To do this, simply note that the null hypothesis is exactly equivalent to $\mathrm{H}_0 : b_i = 0 \; \forall \; i$, since both formulations of H_0 imply that the data follow,

$$y_{ij} = \alpha + \epsilon_{ij}. \tag{2.4}$$

Hence we can test the null hypothesis by using standard linear modelling methods to compare (2.4) to (2.3) using an F-ratio test (ANOVA).

Fitting (2.3) to the data will also yield the usual estimate of σ^2,

$$\hat{\sigma}^2 = \mathrm{RSS}_1/(n - I),$$

where RSS_1 is the residual sum of squares from fitting the model to data, $n = IJ$ is the number of data, and $n - I$ is the residual degrees of freedom from this model fit.

So far the analysis with the mixed model has been identical to what would have been done with a fixed effects model, but now consider estimating σ_b^2. The 'obvious' method of just using the sample variance of the $\hat{b}_i$'s, 'estimated' by least squares, is not to be recommended, as such estimators are biased. Instead we make use of the model that results from averaging at each level of the factor:

$$\bar{y}_{i\cdot} = \alpha + b_i + \frac{1}{J}\sum_{j=1}^{J} \epsilon_{ij}.$$

Now define a new set of I random variables,

$$e_i = b_i + \frac{1}{J}\sum_{j=1}^{J} \epsilon_{ij}.$$

The e_i's are clearly mutually independent, since their constituent random variables are independent and no two e_i's share a constituent random variable. They are also zero mean normal random variables, since each is a sum of zero mean normal random variables. It is also clear that

$$\mathrm{var}(e_i) = \sigma_b^2 + \sigma^2/J.$$

Hence, the model for the aggregated data becomes

$$\bar{y}_{i\cdot} = \alpha + e_i, \tag{2.5}$$

where the e_i are i.i.d. $N(0, \sigma_b^2 + \sigma^2/J)$ random variables. If RSS_2 is the residual sum of squares when this model is fitted by least squares, then the usual unbiased residual variance estimate gives

$$\mathrm{RSS}_2/(I-1) = \hat{\sigma}_b^2 + \hat{\sigma}^2/J.$$

Re-arrangement and substitution of the previous $\hat{\sigma}^2$ implies that

$$\hat{\sigma}_b^2 = \mathrm{RSS}_2/(I-1) - \hat{\sigma}^2/J$$

is an unbiased estimator of σ_b^2.

Now consider a practical industrial example. An engineering test for longitudinal stress in rails involves measuring the time it takes certain ultrasonic waves to travel along the rail. To be a useful test, engineers need to know the average travel time for rails and the variability to expect between rails, as well as the variability in the measurement process. The `Rail` data frame available with R package `nlme` provides 3 measurements of travel time for each of 6 randomly chosen rails. This provides an obvious application for model (2.3). First examine the data.

```
> library(nlme)  # load nlme 'library', which contains data
> data(Rail)     # load data
> Rail
   Rail travel
1     1     55
2     1     53
3     1     54
4     2     26
5     2     37
.     .     .
.     .     .
17    6     85
18    6     83
```

Now fit model (2.3) as a fixed effects model, and use this model to test $\mathrm{H}_0 : \sigma_b^2 = 0$, i.e., to test for evidence of differences between rails.

```
> m1 <- lm(travel ~ Rail,Rail)
> anova(m1)
Analysis of Variance Table

Response: travel
          Df Sum Sq Mean Sq F value    Pr(>F)
Rail       5 9310.5  1862.1  115.18 1.033e-09 ***
Residuals 12  194.0    16.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So there is strong evidence to reject the null hypothesis and accept rail to rail differences as real. As we saw theoretically, so far the analysis does not differ from that for a fixed effects model, but to estimate σ_b^2 involves averaging at each level of the random effect and fitting model (2.5) to the resulting averages. R function `aggregate` will achieve the required averaging.

```
> rt <-  # average over Rail effect
+ aggregate(data.matrix(Rail),by=list(Rail$Rail),mean)
> rt
  Group.1 Rail   travel
1       2    1 31.66667
2       5    2 50.00000
3       1    3 54.00000
4       6    4 82.66667
5       3    5 84.66667
6       4    6 96.00000
```

It is now possible to fit (2.5) and calculate $\hat{\sigma}_b$ and $\hat{\sigma}$, as described above:

```
> m0 <- lm(travel ~ 1,rt)   # fit model to aggregated data
> sig <- summary(m1)$sigma # sig^2 is resid. var. component
> sigb <- (summary(m0)$sigma^2 - sig^2/3)^0.5
>               # sigb^2 is the variance component for rail
> sigb
[1] 24.80547
> sig
[1] 4.020779
```

So, there is a fairly large amount of rail to rail variability, whereas the measurement error is relatively small. In this case the model intercept, α, is confounded with the random effects, b_j, so α must be estimated from the fit of model (2.5).

```
> summary(m0)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    66.50      10.17   6.538  0.00125 **
```

Model checking proceeds by looking at residual plots, from the fits to both the original and the aggregated data, since, approximately, these should look like samples of i.i.d. normal random variables. However, there would have to be a really grotesque violation of the normality assumption for the b_j before you could hope to pick it up from examination of 6 residuals.

2.1.4 A model with two factors

Now consider an experiment in which each observation is grouped according to two factors. A schematic diagram of such a design is shown in figure 2.3. Suppose that one factor is to be modelled as a fixed effect and one as a random effect. A typical example is a randomized block design for an agricultural field trial, testing different fertilizer formulations. The response variable would be yield of the crop concerned, assessed by harvesting at the end of the experiment. Because crop yields depend on many uncontrolled soil related factors, it is usual to arrange the experiment in blocks, within which it is hoped that the soil will be fairly homogeneous. Treatments are randomly arranged within the blocks. For example, a field site might be split into 4 adjacent blocks with 15 plots in each block, each plot being randomly assigned one of five replicates of each of 3 fertilizer treatments. The idea is that differences within blocks should be smaller than differences between blocks — i.e., variability

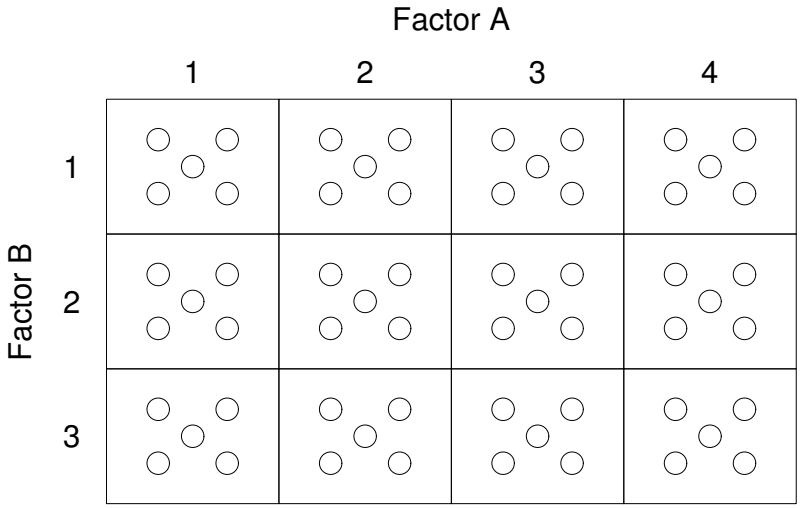

Figure 2.3 *A schematic diagram of a two factor design of the sort discussed in section 2.1.4, with 3 levels of one factor, 4 levels of another and 5 observations for each combination of factor levels. Note that this diagram is not intended to represent the actual physical layout of any experiment.*

in conditions within a block will be smaller than variability across the whole field. A suitable model for the data would include a block effect, to account for this block to block variability, and it makes sense to treat it as a random effect since we are not in the least interested in the particular values of the block effects, but view them as representing variability in environment with location. The treatments, on the other hand, would be modelled as fixed effects, since the values of the treatment effects are fixed properties of the crop population in general. (If we repeated the experiment in a different location, the fertilizer effects should be very similar, whereas the particular values of the block effects would be unrelated to the block effects in the first experiment, apart from having a similar distribution.)

So, a model for the k^{th} observation at level i of fixed effect A and level j of random effect B is

$$y_{ijk} = \mu + \alpha_i + b_j + (\alpha b)_{ij} + \epsilon_{ijk}, \tag{2.6}$$

where $b_j \sim N(0, \sigma_b^2)$, $(\alpha b)_{ij} \sim N(0, \sigma_{\alpha b}^2)$ and $\epsilon_{ijk} \sim N(0, \sigma^2)$, and all these random variables are mutually independent. μ is the overall population mean, the α_i are the I fixed effects for factor A, the b_j are the J random effects for factor B, and the $(\alpha b)_{ij}$ are the IJ interaction terms for the interaction between the factors (an interaction term involving a random effect must also be a random term).

Testing $\mathrm{H}_0 : \sigma_{\alpha b}^2 = 0$ is logically equivalent to testing $\mathrm{H}_0 : (\alpha b)_{ij} = 0 \; \forall \; ij$, in a fixed effects framework. Hence this hypothesis can be tested by the usual ANOVA/F-ratio test comparison of models with and without the interaction terms. If RSS_1 now denotes the residual sum of squares from fitting (2.6) by least squares then

$$\hat{\sigma}^2 = \mathrm{RSS}_1/(n - IJ).$$

In a purely fixed effects context it only makes sense to test for main effects if the

interaction terms are not significant, and can hence be treated as zero. In the mixed effects case, because the interaction is a random effect, it is possible to make inferences about the main effects whether or not the interaction terms are significant. This can be done by averaging the K data at each level of the interaction. The averaging, together with model (2.6), implies the following model for the averages:

$$\bar{y}_{ij\cdot} = \mu + \alpha_i + b_j + (\alpha b)_{ij} + \frac{1}{K}\sum_{k=1}^{K} \epsilon_{ijk}.$$

Defining

$$e_{ij} = (\alpha b)_{ij} + \frac{1}{K}\sum_{k=1}^{K} \epsilon_{ijk},$$

it is clear that, since the e_{ij} are each sums of zero mean normal random variables, they are also zero mean normal random variables. Also, since the $(\alpha b)_{ij}$ and ϵ_{ijk} are mutually independent random variables, and no $(\alpha b)_{ij}$ or ϵ_{ijk} is a component of more than one e_{ij}, the e_{ij} are mutually independent. Furthermore

$$\mathrm{var}(e_{ij}) = \sigma^2_{\alpha b} + \sigma^2/K.$$

Hence the simplified model is

$$\bar{y}_{ij\cdot} = \mu + \alpha_i + b_j + e_{ij}, \tag{2.7}$$

where the e_{ij} are i.i.d. $N(0, \sigma^2_{\alpha b} + \sigma^2/K)$ random variables. The null hypothesis, $\mathrm{H}_0 : \alpha_i = 0 \; \forall \; i$, is tested by comparing the least squares fits of (2.7) and $\bar{y}_{ij\cdot} = \mu + b_j + e_{ij}$, in the usual way, by F-ratio testing. Similarly $\mathrm{H}_0 : \sigma^2_b = 0$ is logically equivalent to $\mathrm{H}_0 : b_j = 0 \; \forall \; j$, and is hence tested by F-ratio test comparison of (2.7) and $\bar{y}_{ij\cdot} = \mu + \alpha_i + e_{ij}$. The residual sum of squares for model (2.7), RSS_2, say, is useful for unbiased estimation of the interaction variance:

$$\hat{\sigma}^2_{\alpha b} + \hat{\sigma}^2/K = \mathrm{RSS}_2/(IJ - I - J + 1)$$

and hence,

$$\hat{\sigma}^2_{\alpha b} = \mathrm{RSS}_2/(IJ - I - J + 1) - \hat{\sigma}^2/K.$$

Averaging the data once more, over the levels of factor B, induces the model

$$\bar{y}_{\cdot j\cdot} = \mu + \frac{1}{I}\sum_{i=1}^{I} \alpha_i + b_j + \frac{1}{I}\sum_{i=1}^{I} e_{ij}.$$

Defining $\mu' = \mu + \frac{1}{I}\sum_i \alpha_i$ and $e_j = b_j + \frac{1}{I}\sum_i e_{ij}$ this model becomes

$$\bar{y}_{\cdot j\cdot} = \mu' + e_j, \text{ where } e_j \sim N(0, \sigma^2_b + \sigma^2_{\alpha b}/I + \sigma^2/(IK)). \tag{2.8}$$

Hence, if RSS_3 is the residual sum of squares of model (2.8), an unbiased estimator of σ^2_b is given by

$$\hat{\sigma}^2_b = \mathrm{RSS}_3/(J-1) - \hat{\sigma}^2_{\alpha b}/I - \hat{\sigma}^2/(IK).$$

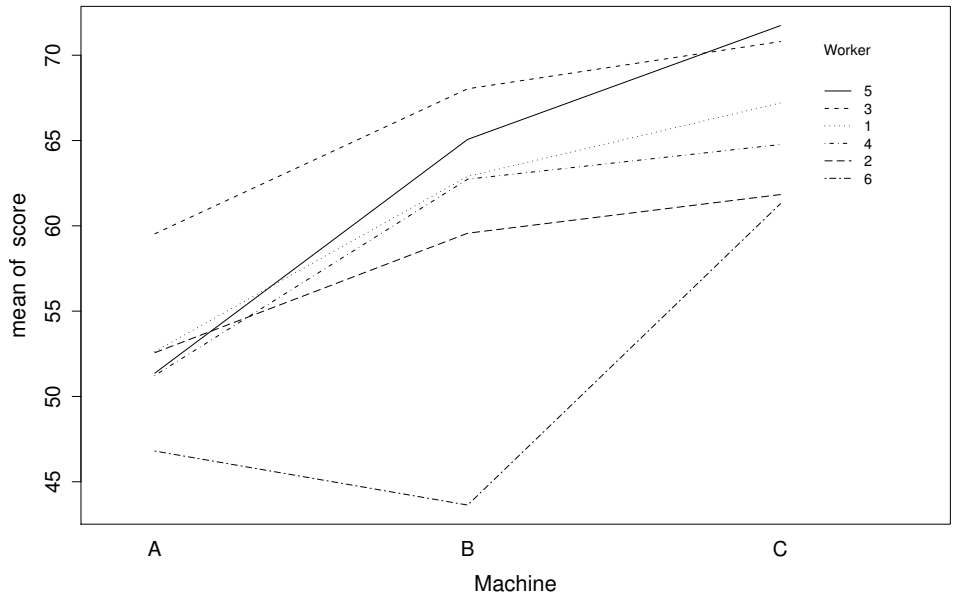

Figure 2.4 *Plot of the* `Machines` *data discussed in section 2.1.4.*

Now consider a practical example. The `Machines` data frame, from the `nlme` package, contains data from an industrial experiment comparing 3 different machine types. The aim of the experiment was to determine which machine type resulted in highest worker productivity. 6 workers were randomly selected to take part in the trial, with each worker operating each machine 3 times (presumably after an appropriate period of training designed to eliminate any 'learning effect'). The following produces the plot shown in figure 2.4

```
> library(nlme)
> data(Machines)
> names(Machines)
[1] "Worker"  "Machine" "score"
> attach(Machines)  # make data available without 'Machines$'
> interaction.plot(Machine,Worker,score)
```

From the experimental aims, it is clear that fixed machine effects and random worker effects are appropriate. We are interested in the effects of these particular machine types, but are only interested in the worker effects in as much as they reflect variability between workers in the population of workers using this type of machine. Put another way, if the experiment were repeated somewhere else (with different workers) we would expect the estimates of the machine effects to be quite close to the results obtained from the current experiment, while the individual worker effects would be quite different (although with similar variability, we hope). So model (2.6) is appropriate, with the α_i representing the fixed machine effects, b_j representing the random worker effects, and $(\alpha b)_{ij}$ representing the worker machine interaction (i.e., the fact that different workers may work better on different machines).

Fitting the full model, we can immediately test $\mathrm{H}_0 : \sigma^2_{\alpha b} = 0$.

```
> m1 <- lm(score ~ Worker*Machine,Machines)
> m0 <- lm(score ~ Worker + Machine,Machines)
> anova(m0,m1)
Analysis of Variance Table
Model 1: score ~ Worker + Machine
Model 2: score ~ Worker + Machine + Worker:Machine
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1     46 459.82
2     36  33.29 10    426.53 46.13 < 2.2e-16 ***
```

We must accept $H_1 : \sigma^2_{\alpha b} \neq 0$. There is very strong evidence for an interaction between machine and worker. σ^2 can now be estimated:

```
> summary(m1)$sigma^2
[1] 0.9246296
```

To examine the main effects we can aggregate at each level of the interaction,

```
Mach <- aggregate(data.matrix(Machines),by=
        list(Machines$Worker,Machines$Machine),mean)
Mach$Worker <- as.factor(Mach$Worker)
Mach$Machine <- as.factor(Mach$Machine)
```

and fit model (2.7) to the resulting data.

```
> m0 <- lm(score ~ Worker + Machine,Mach)
> anova(m0)
Analysis of Variance Table
Response: score
          Df Sum Sq Mean Sq F value    Pr(>F)
Worker     5 413.96   82.79  5.8232 0.0089495 **
Machine    2 585.09  292.54 20.5761 0.0002855 ***
Residuals 10 142.18   14.22
```

The very low p-values again indicate that $H_0 : \sigma^2_b = 0$ and $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ should be rejected in favour of the obvious alternatives. There is strong evidence for differences between machine types and for variability between workers. Going on to examine the fixed effect estimates, using standard fixed effects methods, indicates that machine C leads to substantially increased productivity.

Estimation of $\sigma^2_{\alpha b}$, the interaction variance, is straightforward.

```
> summary(m0)$sigma^2 - summary(m1)$sigma^2/3
[1] 13.90946
```

Aggregating once more and fitting (2.8), we can estimate the worker variance component, σ^2_b.

```
> M <- aggregate(data.matrix(Mach),by=list(Mach$Worker),mean)
> m00 <- lm(score ~ 1, M)
> summary(m00)$sigma^2 - (summary(m0)$sigma^2)/3
[1] 22.96118
```

Residual plots should be checked for `m1`, `m0` and `m00`. If this is done, then it is tempting to try and see how robust the results are to the omission of worker 6 on machine B (see figure 2.4), but this requires methods that can cope with unbalanced data, which will be considered shortly.

2.1.5 Discussion

Although practical examples were presented above, this theory for mixed models of balanced experimental data is primarily of theoretical interest, for understanding the results used in classical mixed model ANOVA tables, and for motivating the use of particular reference distributions when conducting hypothesis tests for mixed models. In practice the somewhat cumbersome analysis based on aggregating data would usually be eschewed in favour of using specialist mixed modelling software, such as that accessible via R function `lme` from the `nlme` library.

Before leaving the topic of balanced data altogether, it is worth noting the reason that ordinary linear model theory can be used for inference with balanced models. The first reason relates to the estimator of the residual variance, $\hat{\sigma}^2$. In ordinary linear modelling, $\hat{\sigma}^2$ does not depend in any way on the values of the model parameters $\boldsymbol{\beta}$ and is independent of $\hat{\boldsymbol{\beta}}$ (see section 1.3.3). This fact is not altered if some elements of $\boldsymbol{\beta}$ are themselves random variables. Hence the usual estimator of $\hat{\sigma}^2$, based on the least squares estimate of a linear model, remains valid, and unbiased, for a linear mixed model.

The second reason relates to the estimators of the fixed effect parameters. In a fixed effects setting, consider two subsets $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ of the parameter vector, $\boldsymbol{\beta}$, with corresponding model matrix columns $\mathbf{X}_1$ and $\mathbf{X}_2$. If $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal, meaning that $\mathbf{X}_1^{\mathsf{T}}\mathbf{X}_2 = \mathbf{0}$, then the least squares estimators $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ will be independent: so inferences about $\boldsymbol{\beta}_1$ do not depend in any way on the value of $\boldsymbol{\beta}_2$. This situation is unaltered if we now move to a mixed effects model, and suppose that $\boldsymbol{\beta}_2$ is actually a random vector. Hence, in a mixed model context, we can still use fixed effects least squares methods for inferences about any fixed effects whose estimators are independent of all the random effects in the model. So, when successively aggregating data (and models), we can use least squares methods to make inferences about a fixed effect as soon as the least squares estimator of that fixed effect becomes independent of all random effects in the aggregated model. Generally such independence only occurs for balanced data from designed experiments.

Finally, note that least squares methods are not useful for 'estimating' the random effects. This is, in part, because identifiability constraints are generally required in order to estimate effects, but imposing such constraints on random effects fundamentally modifies the model, by changing the random effect distributions.

2.2 Maximum likelihood estimation

We have come as far as we can relying only on least squares. A more general approach to mixed models, as well as the generalized linear models of the next chapter, will require use of general large sample results from the theory of maximum likelihood estimation. Appendix A derives these results. This section simply summarises what we need in order to proceed.

A statistical model for a data vector, $\mathbf{y}$, defines a probability density (or mass) function for the random vector of which $\mathbf{y}$ is an observation, $f_\theta(\mathbf{y})$. $\boldsymbol{\theta}$ denotes the unknown parameters of the model. The aim is to make inferences about $\boldsymbol{\theta}$ based on $\mathbf{y}$. Of course, f_θ may depend on other observed variables (covariates) and known

parameters, but there is no need to make this explicit in the notation. The key idea behind maximum likelihood estimation is:

> Values of $\boldsymbol{\theta}$ that make $f_\theta(\mathbf{y})$ larger for the observed $\mathbf{y}$ are more *likely* to be correct than values that make $f_\theta(\mathbf{y})$ smaller.

So we judge the goodness of fit of a value for $\boldsymbol{\theta}$ using the log likelihood function[†]

$$l(\boldsymbol{\theta}) = \log f_\theta(\mathbf{y}),$$

that is the log of the p.d.f. (or p.m.f.) of $\mathbf{y}$ evaluated at the observed $\mathbf{y}$, and considered as a function of $\boldsymbol{\theta}$. The maximum likelihood estimate of $\boldsymbol{\theta}$ is then simply

$$\hat{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmax}}\, l(\boldsymbol{\theta}).$$

Subject to some regularity conditions and in the large sample limit

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \boldsymbol{\mathcal{I}}^{-1}), \tag{2.9}$$

where $\boldsymbol{\theta}$ denotes the true parameter value, and $\boldsymbol{\mathcal{I}}$ is the negative expected Hessian matrix of the log likelihood, so that $\mathcal{I}_{ij} = -\mathbb{E}(\partial^2 l/\partial\theta_i\partial\theta_j)$. $\boldsymbol{\mathcal{I}}$ is known as the (Fisher) information matrix, and actually the same result holds if we replace it by the 'observed information matrix', $\hat{\boldsymbol{\mathcal{I}}}$, where $\hat{\mathcal{I}}_{ij} = -\left.\partial^2 l/\partial\theta_i\partial\theta_j\right|_{\hat{\theta}}$. This result can be used to compute approximate confidence intervals for elements of $\boldsymbol{\theta}$, or to obtain p-values for tests about elements of $\boldsymbol{\theta}$.

There are also large sample results useful for model selection. Consider two models, where model 0 is a reduced version of model 1 (i.e., the models are 'nested'), and suppose we want to test the null hypothesis that model 0 is correct. Let p_j be the number of identifiable parameters for model j. If both models are estimated by MLE then in the large sample limit, assuming a regular likelihood,

$$2\{l(\hat{\boldsymbol{\theta}}_1) - l(\hat{\boldsymbol{\theta}}_0)\} \sim \chi^2_{p_1-p_0} \text{ if model 0 is correct,}$$

i.e., under repeated sampling of $\mathbf{y}$ the generalized log likelihood ratio statistic, $2\{l(\hat{\boldsymbol{\theta}}_1) - l(\hat{\boldsymbol{\theta}}_0)\}$, should follow a $\chi^2_{p_1-p_0}$ distribution, if model 0 is correct. Otherwise $l(\hat{\boldsymbol{\theta}}_0)$ should be sufficiently much lower than $l(\hat{\boldsymbol{\theta}}_1)$ that the statistic will be too large for consistency with $\chi^2_{p_1-p_0}$. As with any hypothesis test, consistency of the data with the null model is judged using a p-value, here $\Pr[2\{l(\hat{\boldsymbol{\theta}}_1) - l(\hat{\boldsymbol{\theta}}_0)\} \leq \chi^2_{p_1-p_0}]$ (the l.h.s. of the inequality is the *observed* test statistic).

A popular alternative model selection approach is to select between models on the basis of an estimate of their closeness to $f_0(\mathbf{y})$, the true density (or mass) function of $\mathbf{y}$. The idea is to try to estimate the expected Kullback-Leibler divergence

$$\mathbb{E}_{\hat{\theta}} \int \{\log f_0(\mathbf{y}) - \log f_{\hat{\theta}}(\mathbf{y})\} f_0(\mathbf{y}) d\mathbf{y}.$$

[†]The *log* likelihood is used for reasons of computational convenience (the likelihood itself tends to underflow to zero), and because the key distributional results involve the log likelihood.

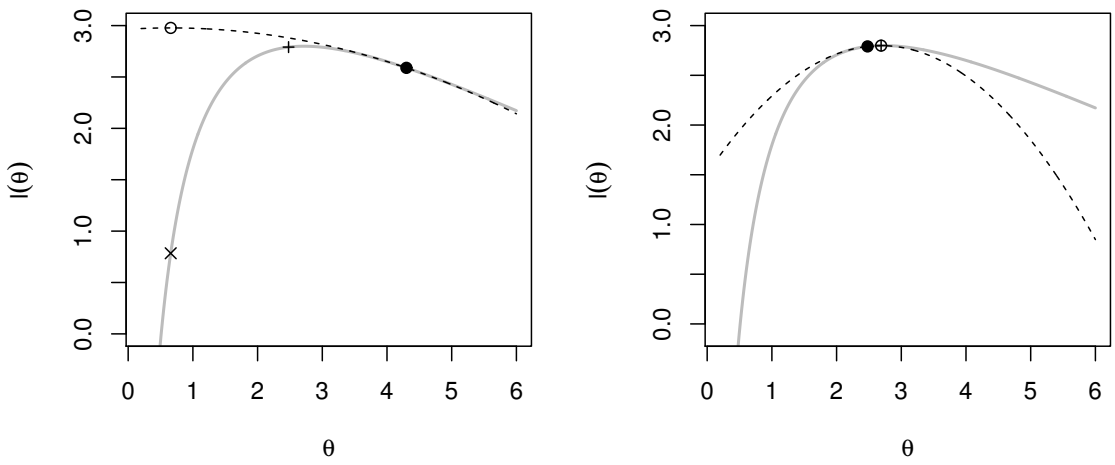

Figure 2.5 *Illustration of two steps of Newton's method as described in section 2.2.1. In both panels the grey curve is the log likelihood to be maximised, and the dashed curve is the quadratic Taylor approximation about the point indicated by •. The maximum of the quadratic approximation is indicated by ∘. Left: the first Newton step is initially too large, resulting in decreased likelihood (×) so that step halving is needed to obtain the improved likelihood (+). Right: for the next step the quadratic approximation is excellent, with a maximum corresponding to that of the log-likelihood.*

A particular estimate of this quantity is minimized by whichever model minimizes

$$\text{AIC} = -2l(\hat{\boldsymbol{\theta}}) + 2p,$$

where p is the number of identifiable model parameters (usually the dimension of $\boldsymbol{\theta}$). From a set of not necessarily nested models we select the one with the lowest AIC.

2.2.1 Numerical likelihood maximization

It is unusual to be able maximize the log likelihood directly, and numerical optimization methods are usually required. Some variant of Newton's method is often the method of choice, for reasons of speed and reliability, and because it is based on the same Hessian of the log likelihood that appears in the large sample result (2.9). Note that optimization software often minimizes by default, but maximizing $l(\boldsymbol{\theta})$ is exactly equivalent to minimizing $-l(\boldsymbol{\theta})$.

The basic principle of Newton's method is to approximate $l(\boldsymbol{\theta})$ by a second order Taylor expansion about the current parameter, guess, $\boldsymbol{\theta}_0$. The maximizer of the approximating quadratic is taken as the next trial parameter vector $\boldsymbol{\theta}'$. If $l(\boldsymbol{\theta}') < l(\boldsymbol{\theta}_0)$ then we repeatedly set $\boldsymbol{\theta}' \leftarrow (\boldsymbol{\theta}' + \boldsymbol{\theta}_0)/2$ until the new log likelihood is not worse than the old one. The whole processes is repeated until $\partial l/\partial \boldsymbol{\theta} \simeq \mathbf{0}$. See figure 2.5.

Formally we are using the approximation

$$l(\boldsymbol{\theta}_0 + \boldsymbol{\Delta}) \simeq l(\boldsymbol{\theta}_0) + \nabla l^{\mathsf{T}} \boldsymbol{\Delta} + \boldsymbol{\Delta}^{\mathsf{T}} \nabla^2 l \boldsymbol{\Delta}/2,$$

where $\nabla l = \partial l/\partial \boldsymbol{\theta}|_{\theta_0}$ and $\nabla^2 l = \partial^2 l/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}|_{\theta_0}$. Assuming $-\nabla^2 l$ is positive

definite, the maximum of the right hand side is found by differentiating and setting to zero, to obtain

$$\boldsymbol{\Delta} = -(\nabla^2 l)^{-1}\nabla l,$$

so $\boldsymbol{\theta}' = \boldsymbol{\theta}_0 + \boldsymbol{\Delta}$. An important detail is that $\boldsymbol{\Delta}$ will be an ascent direction if *any* positive definite matrix, $\mathbf{H}$, is used in place of $-(\nabla^2 l)^{-1}$. This is because, for small enough (positive) step size α, a first order Taylor expansion then implies that $l(\boldsymbol{\theta}_0 + \alpha\boldsymbol{\Delta}) \rightarrow l(\boldsymbol{\theta}_0) + \alpha\nabla l^{\mathsf{T}}\mathbf{H}\nabla l$, and the second term on the right hand side is positive if $\mathbf{H}$ is positive definite.

This means that if the Hessian of the negative log likelihood is indefinite, then it should be perturbed to make it positive definite, in order to guarantee that the Newton step will actually increase the log likelihood. This is sometimes achieved by adding a small multiple of the identity matrix to the Hessian, or by eigen-decomposing the Hessian and re-setting any eigenvalues ≤ 0 to positive values. Other consequences of not requiring the exact Hessian matrix to get an ascent direction are that it is possible to substitute the information matrix for the Hessian when this is convenient ('Fisher scoring'), and also that finite difference approximations to the Hessian often work well. Quasi-Newton methods are a variant on Newton's method in which an approximation to the Hessian is updated at each step, using only the gradients of the log likelihood at the start and end of the step. See (Wood, 2015, §5.1) for more.

2.3 Linear mixed models in general

Recall that the general linear mixed model can conveniently be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \quad \mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\psi}_\theta), \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Lambda}_\theta) \tag{2.10}$$

where $\boldsymbol{\psi}_\theta$ is a positive definite covariance matrix for the random effects $\mathbf{b}$, and $\mathbf{Z}$ is a matrix of fixed coefficients describing how the response variable, $\mathbf{y}$, depends on the random effects (it is a model matrix for the random effects). $\boldsymbol{\psi}_\theta$ depends on some parameters, $\boldsymbol{\theta}$, which will be the prime target of statistical inference about the random effects (the exact nature of the dependence is model specific). Finally, $\boldsymbol{\Lambda}_\theta$ is a positive definite matrix which usually has a simple structure depending on few or no unknown parameters: often it is simply $\mathbf{I}\sigma^2$, or sometimes the covariance matrix of a simple auto-regressive residual model, yielding a banded $\boldsymbol{\Lambda}_\theta^{-1}$ and hence efficient computation. Notice that the model states that $\mathbf{y}$ is a linear combination of normal random variables, implying that it has a multivariate normal distribution: $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\boldsymbol{\psi}_\theta\mathbf{Z}^{\mathsf{T}} + \boldsymbol{\Lambda}_\theta)$.

As a simple example of this general formulation, recall the rails example from section 2.1.3. The model for the j^{th} response on the i^{th} rail is

$$y_{ij} = \alpha + b_i + \epsilon_{ij}, \quad b_i \sim N(0, \sigma_b^2), \quad \epsilon_{ij} \sim N(0, \sigma^2), \tag{2.11}$$

with all the b_i and ϵ_{ij} mutually independent. There were 6 rails with 3 measurements

on each. In the general linear mixed model form the model is therefore

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{41} \\ y_{42} \\ y_{43} \\ y_{51} \\ y_{52} \\ y_{53} \\ y_{61} \\ y_{62} \\ y_{63} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} \alpha \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{31} \\ \epsilon_{32} \\ \epsilon_{33} \\ \epsilon_{41} \\ \epsilon_{42} \\ \epsilon_{43} \\ \epsilon_{51} \\ \epsilon_{52} \\ \epsilon_{53} \\ \epsilon_{61} \\ \epsilon_{62} \\ \epsilon_{63} \end{bmatrix}$$

where $\mathbf{b} \sim N(\mathbf{0}, \mathbf{I}_6\sigma_b^2)$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}_{18}\sigma^2)$. In this case the parameter vector, $\boldsymbol{\theta}$, contains σ_b^2 and σ^2.

2.4 Maximum likelihood estimation for the linear mixed model

The likelihood for the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ of the linear mixed model (2.10) could in principle be based on the p.d.f. implied by the fact that $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\boldsymbol{\psi}_\theta\mathbf{Z}^\mathsf{T} + \boldsymbol{\Lambda}_\theta)$, but this involves the inverse of the $n \times n$ matrix $\mathbf{Z}\boldsymbol{\psi}_\theta\mathbf{Z}^\mathsf{T} + \boldsymbol{\Lambda}_\theta$ which is computationally unattractive. A more convenient expression results by obtaining the marginal distribution, $f(\mathbf{y}|\boldsymbol{\beta})$ (which is the likelihood for $\boldsymbol{\beta}$), by integrating out $\mathbf{b}$ from the joint density of $\mathbf{y}$ and $\mathbf{b}$, $f(\mathbf{y}, \mathbf{b}|\boldsymbol{\beta})$.

From standard probability theory, $f(\mathbf{y}, \mathbf{b}|\boldsymbol{\beta}) = f(\mathbf{y}|\mathbf{b}, \boldsymbol{\beta})f(\mathbf{b})$, and from (2.10),

$$f(\mathbf{y}|\mathbf{b}, \boldsymbol{\beta}) = (2\pi)^{-n/2}|\boldsymbol{\Lambda}_\theta|^{-1/2} \exp\{-\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}\|^2_{\Lambda_\theta^{-1}}/2\},$$

where $\|\mathbf{x}\|^2_{\Lambda^{-1}} = \mathbf{x}^\mathsf{T}\boldsymbol{\Lambda}^{-1}\mathbf{x}$, while, if p is the dimension of $\mathbf{b}$,

$$f(\mathbf{b}) = (2\pi)^{-p/2}|\boldsymbol{\psi}_\theta|^{-1/2} \exp\{-\mathbf{b}^\mathsf{T}\boldsymbol{\psi}_\theta^{-1}\mathbf{b}/2\}.$$

Now let $\hat{\mathbf{b}}$ be the maximizer of $\log f(\mathbf{y}, \mathbf{b}|\boldsymbol{\beta})$ (and hence $f(\mathbf{y}, \mathbf{b}|\boldsymbol{\beta})$) for a given $\boldsymbol{\beta}$, and consider evaluating the likelihood $f(\mathbf{y}|\boldsymbol{\beta})$ by integration:

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\beta}) &= \int f(\mathbf{y}, \mathbf{b}|\boldsymbol{\beta})d\mathbf{b} = \int \exp\{\log f(\mathbf{y}, \mathbf{b}|\boldsymbol{\beta})\}d\mathbf{b} \\ &= \int \exp\left\{\log f(\mathbf{y}, \hat{\mathbf{b}}|\boldsymbol{\beta}) + \frac{1}{2}(\mathbf{b} - \hat{\mathbf{b}})^\mathsf{T}\frac{\partial^2 \log f(\mathbf{y}, \mathbf{b}|\boldsymbol{\beta})}{\partial\mathbf{b}\partial\mathbf{b}^\mathsf{T}}(\mathbf{b} - \hat{\mathbf{b}})\right\} d\mathbf{b} \end{aligned}$$

where the second line is obtained by Taylor expansion about $\hat{\mathbf{b}}$, and there is no remainder term because the higher order derivatives of $\log f(\mathbf{y},\mathbf{b}|\boldsymbol{\beta})$ w.r.t. $\mathbf{b}$ are identically zero (since $f(\mathbf{b})$ and $f(\mathbf{y}|\mathbf{b},\boldsymbol{\beta})$ are Gaussian). Hence

$$f(\mathbf{y}|\boldsymbol{\beta}) = f(\mathbf{y},\hat{\mathbf{b}}|\boldsymbol{\beta}) \int \exp\{-(\mathbf{b}-\hat{\mathbf{b}})^{\mathsf{T}}(\mathbf{Z}^{\mathsf{T}}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z}+\boldsymbol{\psi}_\theta^{-1})(\mathbf{b}-\hat{\mathbf{b}})/2\}d\mathbf{b}. \quad (2.12)$$

Like any p.d.f. a multivariate normal p.d.f. must integrate to 1, i.e.,

$$\int \frac{|\boldsymbol{\Sigma}^{-1}|^{1/2}}{(2\pi)^{p/2}} \exp\left\{-\frac{1}{2}(\mathbf{b}-\hat{\mathbf{b}})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{b}-\hat{\mathbf{b}})\right\} d\mathbf{b} = 1$$

$$\Rightarrow \int \exp\left\{-\frac{1}{2}(\mathbf{b}-\hat{\mathbf{b}})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{b}-\hat{\mathbf{b}})\right\} d\mathbf{b} = \frac{(2\pi)^{p/2}}{|\boldsymbol{\Sigma}^{-1}|^{1/2}}.$$

Applying this result to the integral in (2.12) we obtain the likelihood for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$,

$$f(\mathbf{y}|\boldsymbol{\beta}) = f(\mathbf{y},\hat{\mathbf{b}}|\boldsymbol{\beta})\frac{(2\pi)^{p/2}}{|\mathbf{Z}^{\mathsf{T}}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z}+\boldsymbol{\psi}_\theta^{-1}|^{1/2}}.$$

Explicitly, twice the log likelihood ($l = \log f(\mathbf{y}|\boldsymbol{\beta})$) is therefore

$$2l(\boldsymbol{\beta},\boldsymbol{\theta}) = -\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\hat{\mathbf{b}}\|^2_{\Lambda_\theta^{-1}} - \hat{\mathbf{b}}^{\mathsf{T}}\boldsymbol{\psi}_\theta^{-1}\hat{\mathbf{b}}$$
$$- \log|\boldsymbol{\Lambda}_\theta| - \log|\boldsymbol{\psi}_\theta| - \log|\mathbf{Z}^{\mathsf{T}}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z}+\boldsymbol{\psi}_\theta^{-1}| - n\log(2\pi), \quad (2.13)$$

where $\hat{\mathbf{b}}$ is dependent on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

Since only the first two terms on the r.h.s. of (2.13) depend on $\boldsymbol{\beta}$ (don't forget that $\hat{\mathbf{b}}$ depends on $\boldsymbol{\beta}$) it follows that for any $\boldsymbol{\theta}$, the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ can be found by simply minimizing

$$\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{b}\|^2_{\Lambda_\theta^{-1}} + \mathbf{b}^{\mathsf{T}}\boldsymbol{\psi}_\theta^{-1}\mathbf{b}, \quad (2.14)$$

jointly w.r.t. $\boldsymbol{\beta}$ and $\mathbf{b}$ (also yielding $\hat{\mathbf{b}}$). So, given $\boldsymbol{\theta}$ values, $\hat{\boldsymbol{\beta}}$ can be found by a quadratic optimization which has an explicit solution. This fact can be exploited by basing inference about $\boldsymbol{\theta}$ on the *profile likelihood* $l_p(\boldsymbol{\theta}) = l(\boldsymbol{\theta},\hat{\boldsymbol{\beta}}_\theta)$, where $\hat{\boldsymbol{\beta}}_\theta$ is the minimizer of (2.14) given $\boldsymbol{\theta}$. Exploiting the direct computation of $\hat{\boldsymbol{\beta}}$ in this way reduces the computational burden of iteratively seeking the MLE of $\boldsymbol{\theta}$ by numerical methods.

2.4.1 The distribution of $\mathbf{b}|\mathbf{y},\hat{\boldsymbol{\beta}}$ *given* $\boldsymbol{\theta}$

Now consider the distribution of $\mathbf{b}|\mathbf{y},\hat{\boldsymbol{\beta}}$. We know that $f(\mathbf{b}|\mathbf{y},\hat{\boldsymbol{\beta}}) \propto f(\mathbf{y},\mathbf{b}|\hat{\boldsymbol{\beta}})$, so, defining $\tilde{\mathbf{y}} = \mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}$,

$$\begin{aligned}\log f(\mathbf{b}|\mathbf{y},\hat{\boldsymbol{\beta}}) &= -(\tilde{\mathbf{y}}-\mathbf{Z}\mathbf{b})^{\mathsf{T}}\boldsymbol{\Lambda}_\theta^{-1}(\tilde{\mathbf{y}}-\mathbf{Z}\mathbf{b})/2 - \mathbf{b}^{\mathsf{T}}\boldsymbol{\psi}_\theta^{-1}\mathbf{b}/2 + k_1\\ &= -(\tilde{\mathbf{y}}^{\mathsf{T}}\boldsymbol{\Lambda}_\theta^{-1}\tilde{\mathbf{y}} - 2\mathbf{b}^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}\boldsymbol{\Lambda}_\theta^{-1}\tilde{\mathbf{y}} + \mathbf{b}^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z}\mathbf{b} + \mathbf{b}^{\mathsf{T}}\boldsymbol{\psi}_\theta^{-1}\mathbf{b})/2 + k_1\\ &= -\{\mathbf{b}-(\mathbf{Z}^{\mathsf{T}}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z}+\boldsymbol{\psi}_\theta^{-1})^{-1}\mathbf{Z}^{\mathsf{T}}\boldsymbol{\Lambda}_\theta^{-1}\tilde{\mathbf{y}}\}^{\mathsf{T}}(\mathbf{Z}^{\mathsf{T}}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z}+\boldsymbol{\psi}_\theta^{-1})\\ &\qquad \{\mathbf{b}-(\mathbf{Z}^{\mathsf{T}}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z}+\boldsymbol{\psi}_\theta^{-1})^{-1}\mathbf{Z}^{\mathsf{T}}\boldsymbol{\Lambda}_\theta^{-1}\tilde{\mathbf{y}}\}/2 + k_2,\end{aligned}$$

where k_1 and k_2 are constants not involving $\mathbf{b}$. The final expression is recognisable as the kernel of a multivariate normal p.d.f. so

$$\mathbf{b}|\mathbf{y}, \hat{\boldsymbol{\beta}} \sim N(\hat{\mathbf{b}}, (\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z} + \boldsymbol{\psi}_\theta^{-1})^{-1}), \tag{2.15}$$

where $\hat{\mathbf{b}} = (\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z}+\boldsymbol{\psi}_\theta^{-1})^{-1}\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}_\theta^{-1}\tilde{\mathbf{y}}$, which is readily seen to be the minimiser of (2.14). $\hat{\mathbf{b}}$ is sometimes referred to as the *maximum a posteriori*, or MAP, estimate of the random effects or alternatively as the *predicted* random effects vector.

2.4.2 *The distribution of $\hat{\boldsymbol{\beta}}$ given $\boldsymbol{\theta}$*

Finally, consider the distribution of $\hat{\boldsymbol{\beta}}$ for a given $\boldsymbol{\theta}$. An obvious way to approach this is to use the log likelihood based on $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\boldsymbol{\psi}\mathbf{Z}^\mathsf{T} + \boldsymbol{\Lambda})$ (where $\boldsymbol{\theta}$ subscripts have been dropped to avoid clutter). Then for a given $\boldsymbol{\theta}$ the MLE is the weighted least squares estimate $\hat{\boldsymbol{\beta}} = \{\mathbf{X}^\mathsf{T}(\mathbf{Z}\boldsymbol{\psi}\mathbf{Z}^\mathsf{T} + \boldsymbol{\Lambda})^{-1}\mathbf{X}\}^{-1}\mathbf{X}^\mathsf{T}(\mathbf{Z}\boldsymbol{\psi}\mathbf{Z}^\mathsf{T} + \boldsymbol{\Lambda})^{-1}\mathbf{y}$. Since $\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, and because $\hat{\boldsymbol{\beta}}$ is a linear transformation of $\mathbf{y}$, which has covariance matrix $\mathbf{Z}\boldsymbol{\psi}\mathbf{Z}^\mathsf{T} + \boldsymbol{\Lambda}$, the covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned}\{\mathbf{X}^\mathsf{T}(\mathbf{Z}\boldsymbol{\psi}\mathbf{Z}^\mathsf{T} + \boldsymbol{\Lambda})^{-1}\mathbf{X}\}^{-1}\mathbf{X}^\mathsf{T}(\mathbf{Z}\boldsymbol{\psi}\mathbf{Z}^\mathsf{T} + \boldsymbol{\Lambda})^{-1}(\mathbf{Z}\boldsymbol{\psi}\mathbf{Z}^\mathsf{T} + \boldsymbol{\Lambda})(\mathbf{Z}\boldsymbol{\psi}\mathbf{Z}^\mathsf{T} + \boldsymbol{\Lambda})^{-1}\mathbf{X}&\\ \{\mathbf{X}^\mathsf{T}(\mathbf{Z}\boldsymbol{\psi}\mathbf{Z}^\mathsf{T} + \boldsymbol{\Lambda})^{-1}\mathbf{X}\}^{-1} = \{\mathbf{X}^\mathsf{T}(\mathbf{Z}\boldsymbol{\psi}\mathbf{Z}^\mathsf{T} + \boldsymbol{\Lambda})^{-1}\mathbf{X}\}^{-1}.&\end{aligned}$$

So

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \{\mathbf{X}^\mathsf{T}(\mathbf{Z}\boldsymbol{\psi}\mathbf{Z}^\mathsf{T} + \boldsymbol{\Lambda})^{-1}\mathbf{X}\}^{-1}). \tag{2.16}$$

Again, the inverse of $\mathbf{Z}\boldsymbol{\psi}\mathbf{Z}^\mathsf{T} + \boldsymbol{\Lambda}$ is potentially computationally costly, with $O(n^3)$ floating point computational cost for n data. However, it turns out that an identical covariance matrix can be obtained by treating $\boldsymbol{\beta}$ as a vector of random effects with improper uniform (prior) distributions, reducing the cost to $O(np^2)$, where p is the number of model coefficients. In this case $\boldsymbol{\beta}$ is effectively part of $\mathbf{b}$ so that (2.15) applies and can be re-written as

$$\begin{bmatrix} \mathbf{b}|\mathbf{y} \\ \boldsymbol{\beta}|\mathbf{y} \end{bmatrix} \sim N\left(\begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\boldsymbol{\beta}} \end{bmatrix}, \begin{bmatrix} \mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}^{-1}\mathbf{Z} + \boldsymbol{\psi}^{-1} & \mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}^{-1}\mathbf{X} \\ \mathbf{X}^\mathsf{T}\boldsymbol{\Lambda}^{-1}\mathbf{Z} & \mathbf{X}^\mathsf{T}\boldsymbol{\Lambda}^{-1}\mathbf{X} \end{bmatrix}^{-1}\right). \tag{2.17}$$

It turns out that the block of the above covariance matrix relating to $\boldsymbol{\beta}$ is identical to the frequentist covariance matrix for $\hat{\boldsymbol{\beta}}$ in (2.16). Using a standard result on the inverse of a symmetric partitioned matrix[‡] the covariance matrix block of (2.17) corresponding to $\boldsymbol{\beta}$ is $[\mathbf{X}^\mathsf{T}\{\boldsymbol{\Lambda}^{-1}-\boldsymbol{\Lambda}^{-1}\mathbf{Z}(\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}^{-1}\mathbf{Z}+\boldsymbol{\psi}^{-1})^{-1}\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}^{-1}\}\mathbf{X}]^{-1}$. This turns out to be identical to $\{\mathbf{X}^\mathsf{T}(\mathbf{Z}\boldsymbol{\psi}\mathbf{Z}^\mathsf{T} + \boldsymbol{\Lambda})^{-1}\mathbf{X}\}^{-1}$, because $\boldsymbol{\Lambda}^{-1} - \boldsymbol{\Lambda}^{-1}\mathbf{Z}(\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}^{-1}\mathbf{Z}+\boldsymbol{\psi}^{-1})^{-1}\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}^{-1}$ is equal to the inverse of $\mathbf{Z}\boldsymbol{\psi}\mathbf{Z}^\mathsf{T} + \boldsymbol{\Lambda}$, as the following shows,

[‡] Defining $\mathbf{D} = \mathbf{B} - \mathbf{C}^\mathsf{T}\mathbf{A}^{-1}\mathbf{C}$, it is easy, but tedious, to check that

$$\begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^\mathsf{T} & \mathbf{B} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{C}\mathbf{D}^{-1}\mathbf{C}^\mathsf{T}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{C}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}^\mathsf{T}\mathbf{A}^{-1} & \mathbf{D}^{-1} \end{bmatrix}.$$

$$
\begin{aligned}
\{\boldsymbol{\Lambda}^{-1} - \boldsymbol{\Lambda}^{-1}\mathbf{Z}(\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}^{-1}\mathbf{Z} + \boldsymbol{\psi}^{-1})^{-1}\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}^{-1}\}(\mathbf{Z}\boldsymbol{\psi}\mathbf{Z}^\mathsf{T} + \boldsymbol{\Lambda}) &= \\
\boldsymbol{\Lambda}^{-1}\mathbf{Z}\boldsymbol{\psi}\mathbf{Z}^\mathsf{T} + \mathbf{I} - \boldsymbol{\Lambda}^{-1}\mathbf{Z}(\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}^{-1}\mathbf{Z} + \boldsymbol{\psi}^{-1})^{-1}\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}^{-1}\mathbf{Z}\boldsymbol{\psi}\mathbf{Z}^\mathsf{T} & \\
-\boldsymbol{\Lambda}^{-1}\mathbf{Z}(\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}^{-1}\mathbf{Z} + \boldsymbol{\psi}^{-1})^{-1}\mathbf{Z}^\mathsf{T} &= \\
\mathbf{I} + \boldsymbol{\Lambda}^{-1}\mathbf{Z}\{\boldsymbol{\psi} - (\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}^{-1}\mathbf{Z} + \boldsymbol{\psi}^{-1})^{-1}(\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}^{-1}\mathbf{Z}\boldsymbol{\psi} + \mathbf{I})\}\mathbf{Z}^\mathsf{T} &= \\
\mathbf{I} + \boldsymbol{\Lambda}^{-1}\mathbf{Z}\{\boldsymbol{\psi} - \boldsymbol{\psi}(\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}^{-1}\mathbf{Z}\boldsymbol{\psi} + \mathbf{I})^{-1}(\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}^{-1}\mathbf{Z}\boldsymbol{\psi} + \mathbf{I})\}\mathbf{Z}^\mathsf{T} &= \mathbf{I}.
\end{aligned}
$$

2.4.3 *The distribution of $\hat{\boldsymbol{\theta}}$*

Inference about $\boldsymbol{\theta}$ is reliant on the large sample result that

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \hat{\boldsymbol{\mathcal{I}}}_p^{-1}) \tag{2.18}$$

where $\hat{\boldsymbol{\mathcal{I}}}_p$ is the negative Hessian of the log profile likelihood, with i, j^{th} element $-\partial^2 l_p / \partial\theta_i \partial\theta_j|_{\hat{\theta}}$. l_r from section (2.4.5) can also be used in place of l_p.

2.4.4 *Maximizing the profile likelihood*

In practice $l_p(\boldsymbol{\theta})$ is maximized numerically, with each trial value for $\boldsymbol{\theta}$ requiring (2.14) to be minimized for $\hat{\mathbf{b}}, \hat{\boldsymbol{\beta}}$. The $\hat{\boldsymbol{\beta}}$ computed at convergence is the MLE, of course, while the $\hat{\mathbf{b}}$ are the predicted random effects. To make this concrete, let us repeat the simple rail example from section 2.1.3 and 2.3 using a maximum likelihood approach. Here is a function to evaluate the negative log profile likelihood, l_r. In this case $\boldsymbol{\Lambda}_\theta = \mathbf{I}\sigma^2$ and $\boldsymbol{\psi}_\theta = \mathbf{I}\sigma_b^2$, so $\boldsymbol{\theta} = (\log\sigma_b, \log\sigma)^\mathsf{T}$. The log parameterization ensures that the variance components remain positive.

```
llm <- function(theta,X,Z,y) {
  ## untransform parameters...
  sigma.b <- exp(theta[1])
  sigma <- exp(theta[2])
  ## extract dimensions...
  n <- length(y); pr <- ncol(Z); pf <- ncol(X)
  ## obtain \hat \beta, \hat b...
  X1 <- cbind(X,Z)
  ipsi <- c(rep(0,pf),rep(1/sigma.b^2,pr))
  b1 <- solve(crossprod(X1)/sigma^2+diag(ipsi),
              t(X1)%*%y/sigma^2)
  ## compute log|Z'Z/sigma^2 + I/sigma.b^2|...
  ldet <- sum(log(diag(chol(crossprod(Z)/sigma^2 +
              diag(ipsi[-(1:pf)])))))
  ## compute log profile likelihood...
  l <- (-sum((y-X1%*%b1)^2)/sigma^2 - sum(b1^2*ipsi) -
  n*log(sigma^2) - pr*log(sigma.b^2) - 2*ldet - n*log(2*pi))/2
  attr(l,"b") <- as.numeric(b1) ## return \hat beta and \hat b
  -l
}
```

Notice how (2.14) is minimized to find `b1` which contains $\hat{\boldsymbol{\beta}}$ followed by $\hat{\mathbf{b}}$. The determinant of a positive definite matrix is the square of the product of the terms on the leading diagonal of its Cholesky factor (see appendix B.7), and this is used to compute $\log|\mathbf{Z}^{\mathsf{T}}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z}+\boldsymbol{\psi}_\theta^{-1}|$. We then have all the ingredients to evaluate $l(\boldsymbol{\theta},\hat{\boldsymbol{\beta}})$ using (2.13). Before returning, the estimates/predictions $\hat{\boldsymbol{\beta}},\hat{\mathbf{b}}$ are attached to the log profile likelihood value as an attribute. Finally the negative of the log likelihood is returned, since maximization of l_p is equivalent to minimizing $-l_p$ and most optimization routines minimize by default.

The following fits the rail model by maximizing l_p using R function `optim`.[§]

```
> library(nlme) ## for Rail data
> options(contrasts=c("contr.treatment","contr.treatment"))
> Z <- model.matrix(~ Rail$Rail - 1) ## r.e. model matrix
> X <- matrix(1,18,1)                ## fixed model matrix
> ## fit the model...
> rail.mod <- optim(c(0,0),llm,hessian=TRUE,
                            X=X,Z=Z,y=Rail$travel)
> exp(rail.mod$par) ## variance components
[1] 22.629166  4.024072
> solve(rail.mod$hessian) ## approx cov matrix for theta
              [,1]          [,2]
[1,]  0.0851408546 -0.0004397245
[2,] -0.0004397245  0.0417347933
> attr(llm(rail.mod$par,X,Z,Rail$travel),"b")
[1]  66.50000 -34.46999 -16.32789 -12.36961  15.99803
[7]  17.97717  29.19229
```

The estimated variance components and intercept term should be compared to those obtained in section 2.1.3. `optim` can return the Hessian of its objective function, which is $\hat{\mathcal{I}}_p$, and can therefore be inverted to estimate the covariance matrix of $\hat{\boldsymbol{\theta}}$. The final line of output is $\hat{\beta}$ followed by $\hat{\mathbf{b}}$.

Of course, for practical analysis we would usually use specialist software to estimate a linear mixed model. A major reason for this is computational efficiency. In many applications $\mathbf{Z}$ has a very large number of columns, but also has a sparse structure with many zero elements. Similarly, $\boldsymbol{\psi}$ or its inverse are often sparse (block diagonal, for example). Exploiting this sparse structure (i.e., avoiding multiplications and additions between number pairs containing a zero) is essential for efficient computation. The major packages for linear mixed modelling in R are `nlme` and `lme4`. The former is designed to exploit the sparsity that arises when models have a nested structure, while the latter uses sparse direct matrix methods (e.g., Davis, 2006) to exploit any sparsity pattern. Before looking at these packages there are some more theoretical issues to deal with.

[§]The setting of contrast options ensures that we get the desired form for $\mathbf{Z}$ since `Rail$Rail` is originally declared as an ordered factor.

2.4.5 *REML*

A problem with maximum likelihood estimation of variance components is that it tends to underestimate them. The most obvious example of this is the MLE of σ^2 for the linear model, which is $\hat{\sigma}^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2/n$. This is clearly biased as shown by comparison with the unbiased estimator (1.8) derived in section 1.3.3 (p. 13). This tendency is not limited to the residual variance and gets worse as the number of fixed effects increase. Patterson and Thompson (1971) proposed REML (restricted maximum likelihood) as a bias reducing alternative to maximum likelihood. The original approach is motivated by considering the estimation of particular contrasts, but here it is more helpful to follow Laird and Ware (1982). They observed that the restricted likelihood can be obtained by integrating $\mathbf{b}$ *and* $\boldsymbol{\beta}$ out of $f(\mathbf{y}, \mathbf{b}|\boldsymbol{\beta})$.

The integral can be performed using exactly the same approach taken to arrive at (2.13) in section 2.4, resulting in

$$2l_r(\boldsymbol{\theta}) = -\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{b}}\|^2_{\Lambda_\theta^{-1}} - \hat{\mathbf{b}}^\mathsf{T}\boldsymbol{\psi}_\theta^{-1}\hat{\mathbf{b}} - \log|\boldsymbol{\Lambda}_\theta| - \log|\boldsymbol{\psi}_\theta| - \log\left|\begin{array}{cc} \mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z} + \boldsymbol{\psi}_\theta^{-1} & \mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{X} \\ \mathbf{X}^\mathsf{T}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z} & \mathbf{X}^\mathsf{T}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{X} \end{array}\right| - (n - M)\log(2\pi), \quad (2.19)$$

where M is the dimension of $\boldsymbol{\beta}$ and, as before, $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{b}}$ are dependent on $\boldsymbol{\theta}$ and must be recomputed afresh for each value of $\boldsymbol{\theta}$ for which l_r is evaluated. $l_r(\boldsymbol{\theta})$ can be used exactly as $l_p(\boldsymbol{\theta})$ is used with one exception. l_r can not be used to compare models with differing fixed effect structures (e.g., in a generalized likelihood ratio test statistic or AIC comparison): l_r is simply not comparable between models with different fixed effect structures. $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{b}}$ are also used exactly as before (although some people object to referring to $\hat{\boldsymbol{\beta}}$ as 'REML estimates' of $\boldsymbol{\beta}$, since l_r is not a function of $\boldsymbol{\beta}$).

2.4.6 *Effective degrees of freedom*

A notion that will prove helpful in the context of smoothing is the *effective degrees of freedom* of a model. To see why the idea might be useful, consider the example of a p dimensional random effect $\mathbf{b} \sim (\mathbf{0}, \mathbf{I}\sigma_b^2)$. How many degrees of freedom are associated with $\mathbf{b}$? Clearly if $\sigma_b = 0$ then $\mathbf{b}$ makes no contribution to the model and the answer must be 0. On the other hand if $\sigma_b \to \infty$, then $\mathbf{b}$ will behave like a fixed effect parameter, and the answer is presumably p. This suggests that the effective degrees of freedom for $\mathbf{b}$ should increase with σ_b, from 0 up to p.

One approach to arriving at a quantitative definition of the effective degrees of freedom is to consider REML estimation of σ^2 when $\boldsymbol{\Lambda} = \mathbf{I}\sigma^2$. To save ink write $\mathbf{b}$ and $\boldsymbol{\beta}$ in a single vector $\mathcal{B}^\mathsf{T} = (\mathbf{b}^\mathsf{T}, \boldsymbol{\beta}^\mathsf{T})$, and define corresponding model matrix and precision matrix

$$\mathcal{X} = (\mathbf{Z}, \mathbf{X}) \text{ and } \mathbf{S} = \begin{bmatrix} \boldsymbol{\psi}_\theta^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Then the parts of (2.19) dependent on σ^2 can be written

$$-\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{b}}\|^2/\sigma^2 - \hat{\mathbf{b}}^\mathsf{T}\boldsymbol{\psi}_\theta^{-1}\hat{\mathbf{b}} - n\log\sigma^2 - \log|\mathcal{X}^\mathsf{T}\mathcal{X}/\sigma^2 + \mathbf{S}|.$$

Differentiating with respect to σ^2 and equating to zero yields[¶]

$$\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{b}}\|^2/\sigma^4 - n/\sigma^2 + \text{tr}\{(\mathcal{X}^\mathsf{T}\mathcal{X}/\sigma^2 + \mathbf{S})^{-1}\mathcal{X}^\mathsf{T}\mathcal{X}/\sigma^4\} = 0,$$

which implies that

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{b}}\|^2}{n - \tau}$$

where $\tau = \text{tr}\{(\mathcal{X}^\mathsf{T}\mathcal{X}/\hat{\sigma}^2 + \mathbf{S})^{-1}\mathcal{X}^\mathsf{T}\mathcal{X}/\hat{\sigma}^2\}$. By comparison with the usual unbiased estimator of σ^2, (1.8) p. 13, this suggests considering τ as the *effective* degrees of freedom of the mixed model. It is relatively straightforward to show that $M \leq \tau \leq M + p$ where M and p are the number of fixed and random effects, respectively.

2.4.7 The EM algorithm

The profile log likelihood, l_p, or restricted log likelihood, l_r, can be numerically optimized using Newton's method. However Newton's method can sometimes be slow if the starting values for the parameters are poor, and it can therefore be useful to start fitting using another approach that converges more rapidly when far from the optimum values. The *EM algorithm* (Dempster et al., 1977) is such a method (see Davison, 2003; Wood, 2015, for introductions).

Starting from parameter guesses, $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}$, the following steps are iterated:

1. Find the distribution of $\mathbf{b}|\mathbf{y}$ according to the current parameter estimates.
2. Treating the distribution from 1 as fixed (rather than depending on $\boldsymbol{\theta}, \boldsymbol{\beta}$), find an expression for $Q(\boldsymbol{\theta}, \boldsymbol{\beta}) = \mathbb{E}_{\mathbf{b}|\mathbf{y}}\{\log f(\mathbf{y}, \mathbf{b}|\boldsymbol{\beta})\}$ as a function of $\boldsymbol{\theta}, \boldsymbol{\beta}$, using the distribution from 1. The $\mathbf{y}$ are treated as fixed, here. (This is the E-step.)
3. Maximize the expression for $Q(\boldsymbol{\theta}, \boldsymbol{\beta})$ w.r.t. the parameters to obtain updated estimates $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}$. (This is the M-step.)

Note that the expectation in step 2 is taken with respect to the *fixed* distribution from step 1, which depends on the current parameter *estimates*. When evaluating $Q(\boldsymbol{\theta}, \boldsymbol{\beta})$, we view $\log f(\mathbf{y}, \mathbf{b}|\boldsymbol{\beta})$ as a function of $\boldsymbol{\theta}, \boldsymbol{\beta}$, but do not treat the distribution of $\mathbf{b}|\mathbf{y}$ as depending on these parameters.

There are two key points about this algorithm.

1. Each step of the algorithm can be shown to increase the log-likelihood $l(\boldsymbol{\theta}, \boldsymbol{\beta})$ (until a turning point of the likelihood is reached, hopefully the MLE).
2. $Q(\boldsymbol{\theta}, \boldsymbol{\beta})$ is often much easier to evaluate and maximize than $l(\boldsymbol{\theta}, \boldsymbol{\beta})$ itself.

To appreciate exactly what the E-step involves, it helps to derive Q. First note that if $k(\boldsymbol{\theta}) = -\log|\boldsymbol{\Lambda}_\theta|/2 - \log|\boldsymbol{\psi}_\theta|/2 - (n-p)\log(2\pi)/2$, then

$$\begin{aligned}
\log f(\mathbf{y}, \mathbf{b}|\boldsymbol{\beta}) &= -\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}\|^2_{\boldsymbol{\Lambda}_\theta^{-1}}/2 - \mathbf{b}^\mathsf{T}\boldsymbol{\psi}_\theta^{-1}\mathbf{b}/2 + k(\boldsymbol{\theta}) \\
&= -\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2_{\boldsymbol{\Lambda}_\theta^{-1}}/2 - \mathbf{b}^\mathsf{T}\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z}\mathbf{b}/2 + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\mathsf{T}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z}\mathbf{b} \\
&\quad -\mathbf{b}^\mathsf{T}\boldsymbol{\psi}_\theta^{-1}\mathbf{b}/2 + k(\boldsymbol{\theta}).
\end{aligned}$$

[¶]The derivatives of $-\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}\|^2/\sigma^2 - \mathbf{b}^\mathsf{T}\boldsymbol{\psi}_\theta^{-1}\mathbf{b}$ w.r.t. $\mathbf{b}$ are all 0 at $\hat{\mathbf{b}}$, which is why $\hat{\mathbf{b}}$'s dependence on σ^2 adds nothing to the derivative.

To find the expectation of $\log f(\mathbf{y}, \mathbf{b}|\boldsymbol{\beta})$ w.r.t. $f(\mathbf{b}|\mathbf{y})$ requires use of the standard results $\mathbb{E}(\mathbf{b}^\mathsf{T}\mathbf{A}\mathbf{b}) = \mathrm{tr}\{\mathbf{A}\mathbb{E}(\mathbf{b}\mathbf{b}^\mathsf{T})\} = \mathrm{tr}\{\mathbf{A}\mathbf{V}_b\} + \mathbb{E}(\mathbf{b})^\mathsf{T}\mathbf{A}\mathbb{E}(\mathbf{b})$, where $\mathbf{V}_b$ is the covariance matrix of $\mathbf{b}$. From (2.15) in section 2.4.1, $\mathbb{E}_{b|y}(\mathbf{b}) = \hat{\mathbf{b}}$ and $\mathbf{V}_b = (\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z} + \boldsymbol{\psi}_\theta^{-1})^{-1}$, so taking the required expectations we have

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\beta}) &= -\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2_{\Lambda_\theta^{-1}}/2 - \mathrm{tr}\{\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z}(\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z} + \boldsymbol{\psi}_\theta^{-1})^{-1}\}/2 \\ &\quad -\hat{\mathbf{b}}^\mathsf{T}\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z}\hat{\mathbf{b}}/2 + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\mathsf{T}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z}\hat{\mathbf{b}} - \hat{\mathbf{b}}^\mathsf{T}\boldsymbol{\psi}_\theta^{-1}\hat{\mathbf{b}}/2 \\ &\quad -\mathrm{tr}\{\boldsymbol{\psi}_\theta^{-1}(\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z} + \boldsymbol{\psi}_\theta^{-1})^{-1}\}/2 + k(\boldsymbol{\theta}) \\ &= -\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2_{\Lambda_\theta^{-1}}/2 - \hat{\mathbf{b}}^\mathsf{T}\mathbf{Z}^\mathsf{T}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z}\hat{\mathbf{b}}/2 + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\mathsf{T}\boldsymbol{\Lambda}_\theta^{-1}\mathbf{Z}\hat{\mathbf{b}} \\ &\quad -\hat{\mathbf{b}}^\mathsf{T}\boldsymbol{\psi}_\theta^{-1}\hat{\mathbf{b}}/2 - p/2 + k(\boldsymbol{\theta}). \end{aligned}$$

Remember that when optimizing $Q(\boldsymbol{\theta}, \boldsymbol{\beta})$ w.r.t. $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, $\hat{\mathbf{b}}$ remains fixed at its value from step 1 of the iteration. $\hat{\mathbf{b}}$ only changes at the next step 1. This contrasts to the situation when optimizing l_p or l_r, when $\hat{\mathbf{b}}$ and $\hat{\boldsymbol{\beta}}$ must be re-computed for each new trial value for $\boldsymbol{\theta}$.

The algorithm is a very reliable way of maximizing the likelihood and can also be used with the REML criterion, but is rather slow to converge near the MLE. Hence it is often best to start optimization off using EM steps, before switching to Newton's method (Pinheiro and Bates, 2000).

2.4.8 Model selection

The results of section 2.4.2 provide the basis for obtaining approximate confidence intervals for $\boldsymbol{\beta}$, albeit by fixing $\boldsymbol{\theta}$ at its estimated value. It is also possible to use the results to construct simple Wald tests for fixed effects.

The random effect parameters, $\boldsymbol{\theta}$, are more awkward. The fundamental difficulty is that many tests of interest restrict some parameters to the edge of the feasible parameter space, invalidating the usual large sample generalized likelihood ratio tests and other simple testing procedures.

AIC is also problematic for model comparison (Greven and Kneib, 2010). The difficulty is that we can only base AIC on the the log restricted likelihood, l_r, if the fixed effect structure of all models is identical, but if we use the log likelihood, l_p, instead then variance parameters are biased downwards, biasing AIC model selection towards models with a simpler random effects structure. See section 6.11 (p. 301).

Despite the difficulties matters are far from hopeless. Confidence intervals for elements of $\boldsymbol{\theta}$ can be computed using (3.4.3), and if those intervals suggest that variance components are bounded comfortably away from zero then it is clear that they are needed in the model. Similarly if a variance component is estimated as being effectively zero, then we can safely drop that term. In cases of doubt we may also use a GLRT as a rough guide for model comparison: very large or small p-values still suggest a clear-cut result and it is only p-values of similar size to our decision threshold for inclusion/exclusion that are problematic. Note also that some interesting tests do not involve restricting parameters to the edge of the feasible space, and in that case there is no problem.

There are also several reliable tests available for testing variance components for equality to zero. Crainiceanu and Ruppert (2004) produced an exact test for a linear mixed model with one variance component and Greven et al. (2008) proposed an approximate method to extend this to models with multiple variance components. Wood (2013b) proposed an alternative test exploiting the link between mixed models and penalized regression which is also applicable beyond the Gaussian setting (see section 6.12.2, p. 309). However, at time of writing, none of these are directly available in the major linear mixed modelling packages in R.

2.5 Linear mixed models in R

There are several packages for linear mixed modelling in R, of which `nlme` and the `lme4` are particularly noteworthy. Package `mgcv` can also fit linear mixed models with relatively simple random effects structures. `nlme` also provides nonlinear mixed models (see Pinheiro and Bates, 2000), while `lme4` and `mgcv` also provide generalized linear mixed models.

2.5.1 *Package* `nlme`

The main model fitting function of interest is called `lme`. A call to the `lme` function is similar to a call to `lm`, except that an extra argument specifying the random effects structure must also be supplied to the model. By default, `lme` works with a slightly more restricted structure for linear mixed models than the very general form (2.10), given in section 2.3. Specifically `lme` assumes that your data are grouped according to the levels of some factor(s), and that the same random effects structure is required for each group, with random effects independent between groups. Assuming just one level of grouping, the model for the data in the i^{th} group is then

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad \mathbf{b}_i \sim N(\mathbf{0}, \boldsymbol{\psi}_\theta), \quad \boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Lambda}_\theta\sigma^2). \tag{2.20}$$

Careful attention should be paid to which terms have an i subscript, and hence depend on group, and which are common to all groups. Note, in particular, that $\boldsymbol{\beta}$, $\boldsymbol{\psi}_\theta$ and $\boldsymbol{\Lambda}_\theta$, the unknowns for the fixed effects and random effects, respectively, are assumed to be the same for all groups, as is the residual variance, σ^2. This form of mixed effects model, which was introduced by Laird and Ware (1982), is a sensible default because it is very common in practical applications, and because model fitting for this structure is more efficient than for the general form (2.10). However, it is important to realize that (2.20) is only a special case of (2.10). Indeed if we treat all the data as belonging to a single group then (2.20) is exactly (2.10), with no special structure imposed.

Because of `lme`'s default behaviour, you need to provide two parts to the random effects specification: a part that specifies $\mathbf{Z}_i$ and a part specifying the grouping factor(s). By default, $\boldsymbol{\psi}_\theta$ is assumed to be a general positive definite matrix to be estimated, but it is also possible to specify that it should have a more restricted form (for example $\mathbf{I}\sigma_b^2$). The simplest way to specify the random effects structure is with a one sided formula. For example `~x | g` would set up $\mathbf{Z}_i$ according to the `~x` part

of the formula while the levels of the factor variable, `g`, would be used to split the data into groups (i.e., the levels of `g` are effectively the group index, i). The random effects formula is one sided, because there is no choice about the response variable — it must be whatever was specified in the fixed effects formula. So an example call to `lme` looks something like this:

```
lme(y ~ x + z, dat, ~ x|g)
```

where the response is `y`, the fixed effects depend on `x` and `z`, the random effects depend only on `x`, the data are grouped according to factor `g`, and all data are in data frame `dat`. An alternative way of specifying the same model is:

```
lme(y ~ x + z, dat, list(g = ~x))
```

and in fact this latter form is the one we will eventually use with GAMMs.

As an example, model (2.11), from sections 2.1.3 and 2.3, can easily be fitted.

```
> library(nlme)
> lme(travel ~ 1, Rail, list(Rail = ~ 1))
Linear mixed-effects model fit by REML
  Data: Rail
  Log-restricted-likelihood: -61.0885
  Fixed: travel ~ 1
(Intercept)
       66.5

Random effects:
 Formula: ~ 1 | Rail
        (Intercept) Residual
StdDev:    24.80547 4.020779

Number of Observations: 18
Number of Groups: 6
```

Because REML has been used for estimation the results are identical to those obtained in section 2.1.3. If we had used MLE, by specifying `method="ML"` in the call to `lme`, then the results would have corresponded to those obtained in section 2.4.4.

2.5.2 *Tree growth: An example using* `lme`

The `nlme` package includes a data frame called `Loblolly`, containing growth data on Loblolly pine trees. `height`, in feet (data are from the US), and `age`, in years, are recorded for 14 individual trees. A factor variable `Seed`, with 14 levels, indicates the identity of individual trees. Interest lies in characterising the population level mean growth trajectory of Loblolly pines, but it is clear that we would expect a good deal of tree to tree variation, and probably also some degree of autocorrelation in the random component of height.

From examination of data plots, the following initial model might be appropriate

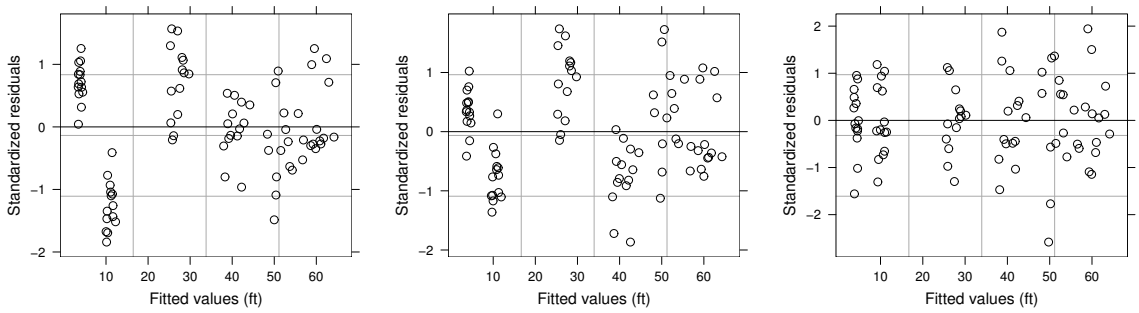

Figure 2.6 *Default residual plots for models* `m0`*,* `m1` *and* `m2` *(left to right). There is a clear trend in the mean of the residuals for the first two models, which model* `m2` *eliminates.*

for the i^{th} measurement on the j^{th} tree:

$$\texttt{height}_{ji} = \beta_0 + \beta_1 \texttt{age}_{ji} + \beta_2 \texttt{age}_{ji}^2 + \beta_3 \texttt{age}_{ji}^3 + b_0 + b_{j1}\texttt{age}_{ji} + b_{j2}\texttt{age}_{ji}^2 + b_{j3}\texttt{age}_{ji}^3 + \epsilon_{j,i}$$

where the $\epsilon_{j,i}$ are zero mean normal random variables, with correlation given by $\rho(\epsilon_{j,i}, \epsilon_{j,i-k}) = \phi^k$, and ϕ is an unknown parameter: this $\epsilon_{j,i}$ model is an autoregressive model of order 1 (if the ages had been unevenly spaced then a continuous generalization of this is available, which would then be more appropriate). The ϵ terms are independent between different trees. As usual $\boldsymbol{\beta}$ denotes the fixed effects and $\mathbf{b}_j \sim N(0, \boldsymbol{\psi})$ denotes the random effects.

This model can be estimated using `lme`, but to avoid convergence difficulties in the following analysis, two preparatory steps are useful. Firstly, it is worth centring the `age` variable as follows:

```
Loblolly$age <- Loblolly$age - mean(Loblolly$age)
```

without such centring, polynomial terms can become highly correlated which can cause numerical difficulties. An alternative to centering would be to use the `poly` function to set up an orthogonal polynomial basis.

Secondly, for this analysis the default fitting method fails without some adjustment. `lme` fits start by using the EM algorithm to get reasonably close to the optimal parameter estimates, and then switch to Newton's method, which converges more quickly. The number of EM steps to take, and the maximum number of Newton steps to allow, are both controllable via the `control` argument of `lme`. The `lmeControl` function offers a convenient way of producing a `control` list, with some elements modified from their default. For example

```
lmc <- lmeControl(niterEM=500,msMaxIter=100)
```

produces a `control` list in which the number of EM iterations is set to 500, and the maximum number of Newton iterations is set to 100. For future reference, note that `niterEM` should rarely be increased from its default 0 when calling `gamm`.

The model can now be estimated.

```
m0 <- lme(height ~ age + I(age^2) + I(age^3),Loblolly,
          random = list(Seed = ~ age + I(age^2) + I(age^3)),
          correlation = corAR1(form = ~ age|Seed),control=lmc)
```

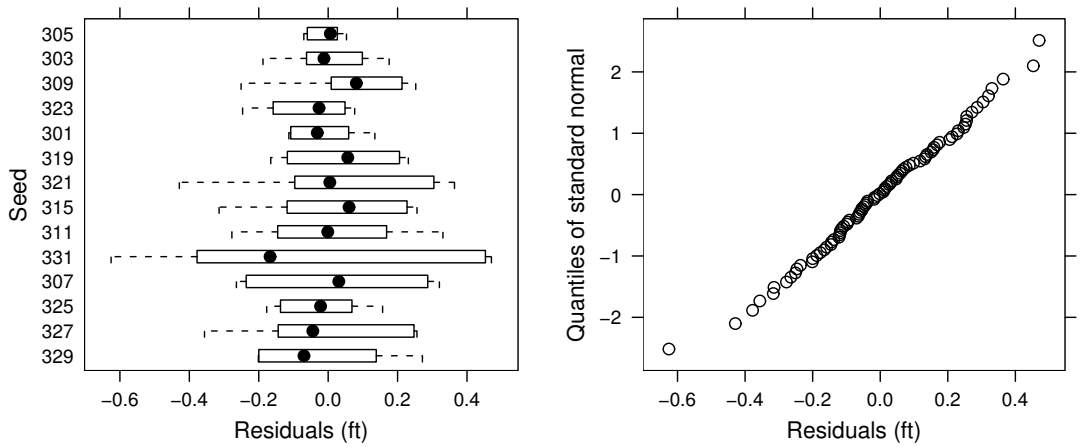

Figure 2.7 *Further residual plots for model* `m2`. *The left panel shows boxplots of the residuals for each tree, while the right plot is a normal QQ-plot for the residuals.*

The `random` argument specifies that there should be a different cubic term for each tree, while the `correlation` argument specifies an autoregressive model for the residuals for each tree. `form=~age|Seed` indicates that `age` is the variable determining the ordering of residuals, and that the correlation applies within measurements made on one tree, but not between measurements on different trees.

The command `plot(m0)` produces the default residual plot shown in the left panel of figure 2.6. The plot shows a clear trend in the mean of the residuals: the model seems to underestimate the first group of measurements, made at age 5, and then overestimate the next group, made at age 10, before somewhat underestimating the next group, which correspond to year 15. This suggests a need for a more flexible model, so fourth and fifth order polynomials were also tried.

```
m1 <- lme(height ~ age + I(age^2) + I(age^3) + I(age^4),
          Loblolly, list(Seed = ~ age + I(age^2) + I(age^3)),
          cor = corAR1(form = ~age|Seed), control=lmc)
plot(m1)
m2 <- lme(height ~ age +I(age^2) +I(age^3) +I(age^4) +I(age^5),
          Loblolly,list(Seed = ~ age + I(age^2) + I(age^3)),
          cor = corAR1(form = ~ age|Seed), control=lmc)
plot(m2)
```

The resulting residuals plots are shown in the middle and right panels of figure 2.6. `m1` does lead to a slight improvement, but only `m2` is really satisfactory. Further model checking plots can now be produced for `m2`.

```
plot(m2,Seed~resid(.))
qqnorm(m2,~resid(.))
qqnorm(m2,~ranef(.))
```

The resulting plots are shown in figures 2.7 and 2.8, and suggest that the model is reasonable.

An obvious question is whether the elaborate model structure, with random cubic and autocorrelated within-tree errors, is really required. First try dropping the autocorrelation component.

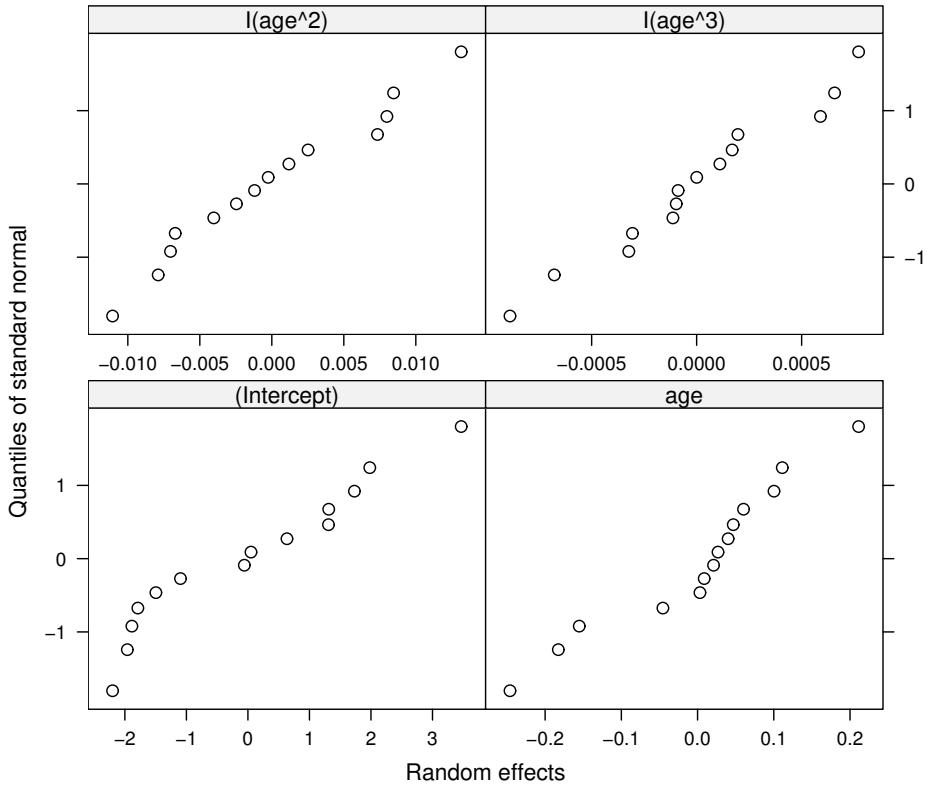

Figure 2.8 *Normal QQ-plots for the predicted random effects from model* `m2`. *The plots should look like correlated random scatters around straight lines, if the normality assumptions for the random effects are reasonable: only* $\hat{b}_1$ *shows any suggestion of any problem, but it is not enough to cause serious concern.*

```
> m3 <- lme(height ~ age+I(age^2)+I(age^3)+I(age^4)+I(age^5),
+    Loblolly,list(Seed = ~ age+I(age^2)+I(age^3)),control=lmc)
> anova(m3,m2)
   Model df    AIC    BIC  logLik   Test L.Ratio p-value
m3     1 17 250.46 290.53 -108.23
m2     2 18 239.36 281.78 -101.68 1 vs 2 13.1041   3e-04
```

The `anova` command is actually conducting a generalized likelihood ratio test here, which rejects `m3` in favour of `m2`. Note that in this case the GLRT assumptions are met: `m3` is effectively setting the autocorrelation parameter ϕ to zero, which is in the middle of its possible range, not on a boundary. `anova` also reports the AIC for the models, which also suggest that `m2` is preferable. There seems to be strong evidence for auto-correlation in the within-tree residuals.

Perhaps the random effects model could be simplified, by dropping the dependence of tree-specific growth on the cube of age.

```
> m4 <- lme(height~age+I(age^2)+I(age^3)+I(age^4)+I(age^5),
+           Loblolly,list(Seed=~age+I(age^2)),
+           correlation=corAR1(form=~age|Seed),control=lmc)
```

```
> anova(m4,m2)
   Model df    AIC    BIC  logLik   Test L.Ratio p-value
m4     1 14 253.76 286.75 -112.88
m2     2 18 239.36 281.78 -101.68 1 vs 2 22.4004   2e-04
```

Recall that the GLRT test is somewhat problematic here, since `m4` is `m2` with some variance parameters set to the edge of the feasible parameter space; however, a likelihood ratio statistic so large that it would have given rise to a p-value of .0002, for a standard GLRT, is strong grounds for rejecting `m4` in favour of `m2` in the current case. Comparison of AIC scores (which could also have been obtained using `AIC(m4,m2)`) suggests quite emphatically that `m2` is the better model.

Another obvious model to try is one with a less general random effects structure. The models so far have allowed the random effects for any tree to be correlated in a very general way: it has simply been assumed that $\mathbf{b}_j \sim N(0, \boldsymbol{\psi}_\theta)$, where the only restriction on the matrix $\boldsymbol{\psi}_\theta$, is that it should be positive definite. Perhaps a less flexible model would suffice: for example, $\boldsymbol{\psi}_\theta$ might be a diagonal matrix (with positive diagonal elements). Such a structure (and indeed many other structures) can be specified in the call to `lme`.

```
> m5 <- lme(height~age+I(age^2)+I(age^3)+I(age^4)+I(age^5),
+          Loblolly,list(Seed=pdDiag(~age+I(age^2)+I(age^3))),
+          correlation=corAR1(form=~age|Seed),control=lmc)
```

Here the `pdDiag` function indicates that the covariance matrix for the random effects at each level of `Seed` should have a (positive definite) diagonal structure. `m5` can be compared to `m2`.

```
> anova(m2,m5)
   Model df      AIC      BIC    logLik   Test L.Ratio p-value
m2     1 18 239.3576 281.7783 -101.6788
m5     2 12 293.7081 321.9886 -134.8540 1 vs 2 66.3505  <.0001
```

Again, both the GLRT test and AIC comparison favour the more general model `m2`. In this case the GLRT assumptions are met: `m5` amounts to setting the random effects covariances in `m2` to zero, but since covariances can be positive or negative this is not on the boundary of the parameter space and the GLRT assumptions hold. The `nlme` package includes very many useful utilities for examining and plotting grouped data, one of which is the following, for plotting data and model predictions together on a unit by unit basis. See figure 2.9.

```
plot(augPred(m2))
```

2.5.3 *Several levels of nesting*

When using mixed models it is quite common to have several levels of nesting present in a model. For example, in the machine type and worker productivity model (2.6), of section 2.1.4, there are random effects for worker and each worker-machine combination. `lme` can accommodate such structures as follows

```
> lme(score ~ Machine,Machines,list(Worker = ~1, Machine = ~1))
Linear mixed-effects model fit by REML
  Data: Machines
  Log-restricted-likelihood: -107.8438
```

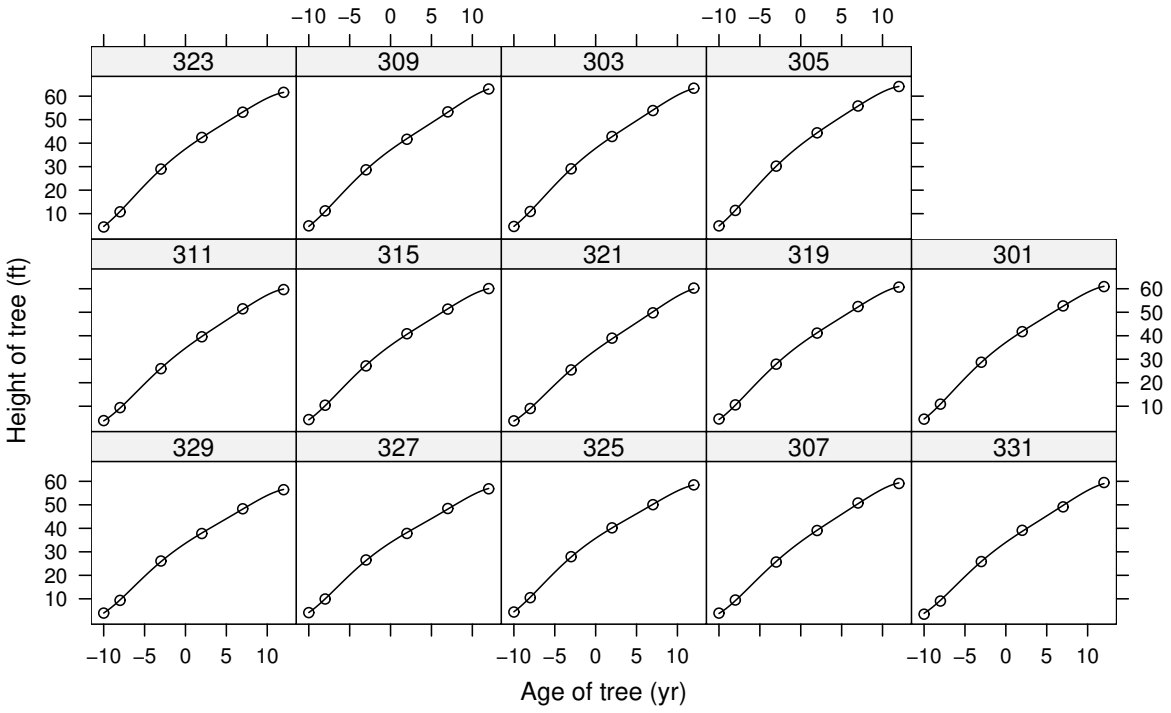

Figure 2.9 *Model predictions from* `m4` *at the individual tree level, overlaid on individual Loblolly pine growth data. The panel titles are the value of the* `Seed` *tree identifier.*

```
  Fixed: score ~ Machine
(Intercept)    MachineB    MachineC
  52.355556    7.966667   13.916667

Random effects:
 Formula: ~1 | Worker
        (Intercept)
StdDev:    4.781049

 Formula: ~1 | Machine %in% Worker
        (Intercept)  Residual
StdDev:    3.729536 0.9615768

Number of Observations: 54
Number of Groups:
             Worker Machine %in% Worker
                  6                  18
```

Notice how any grouping factor in the random effects list is assumed to be nested within the grouping factors to its left.

This section can only hope to scratch the surface of what is possible with `lme`: for a much fuller account, see Pinheiro and Bates (2000).

2.5.4 *Package* lme4

The nlme package is designed to efficiently exploit the nested structure of linear models such as

$$y_{ijk} = \alpha + b_i + c_{ij} + \epsilon_{ijk}$$

where b_i and c_{ij} are Gaussian random effects. In consequence nlme is less computationally efficient with non-nested models such as

$$y_{ijk} = \alpha + b_i + c_j + \epsilon_{ijk},$$

even when the random effects model matrices for both models have the same proportion of zero entries (the same 'sparsity').

The lme4 package is designed to be equally efficient in both nested and non nested cases, by using direct sparse matrix methods (e.g., Davis, 2006) for the model estimation. Sparse matrix methods are designed to be memory efficient by only storing the non-zero elements of matrices, and floating point efficient, by only computing the elements of matrix results that are not structurally zero. Use of sparse matrix methods is complicated by the problem of *infill*: that the result of a matrix operation on one or more sparse matrices need not be sparse. However, the linear mixed model likelihood can be efficiently computed using sparse methods, by using sparse Cholesky decomposition as the basis for the computations (see, e.g., the Matrix package in R). For example, the solve and determinant calculations in the simple code in section 2.4.4 would both be replaced with code based on sparse Cholesky decomposition.

The lmer function from package lme4 is designed for linear mixed modelling, and its use is similar to other R modelling functions, such as lm. Model specification uses a single model formula combining fixed and random effects. Fixed effects are specified exactly as for lm, whereas random effects are of the form (x|g) where g is a grouping factor, and x is interpreted as the right hand side of a model formula specifying the random effect model matrix nested in each level of g. In contrast to lme, the ordering of the random effects is unimportant. At time of writing there are no facilities for the sort of correlation structure that lme supports.

Consider repeating the Machines model from the previous section.

```
> library(lme4)
> a1 <- lmer(score ~ Machine + (1|Worker) + (1|Worker:Machine),
+            data=Machines)
> a1
Linear mixed model fit by REML ['lmerMod']
Formula: score ~ Machine + (1 | Worker) + (1 | Worker:Machine)
   Data: Machines
REML criterion at convergence: 215.6876
Random effects:
 Groups         Name        Std.Dev.
 Worker:Machine (Intercept) 3.7295
 Worker         (Intercept) 4.7811
 Residual                   0.9616
Number of obs: 54, groups:  Worker:Machine, 18; Worker, 6
```

```
Fixed Effects:
(Intercept)     MachineB     MachineC
     52.356        7.967       13.917
```

The estimates are the same as those from `lme`, of course. Now try an alternative model in which the random effects for machine are correlated between machines used by the same worker, but independent between workers, and compare the models using a generalized likelihood ratio test, and AIC.

```
> a2 <- lmer(score ~ Machine + (1|Worker) + (Machine-1|Worker),
+            data=Machines)
> AIC(a1,a2)
   df      AIC
a1  6 227.6876
a2 11 230.3112
> anova(a1,a2)
refitting model(s) with ML (instead of REML)
Data: Machines
Models:
a1: score ~ Machine + (1 | Worker) + (1 | Worker:Machine)
a2: score ~ Machine + (1 | Worker) + (Machine - 1 | Worker)
   Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
a1  6 237.27 249.2 -112.64   225.27
a2 11 238.42 260.3 -108.21   216.42 8.8516      5     0.1151
```

AIC suggests that this new model is overcomplicated. This is confirmed by the generalized likelihood ratio test. The latter is valid because the null model, `a1`, is restricting the variances of the `Machine` effects to be equal (not zero), and the correlations between machines to be zero (which is not at the edge of the feasible parameter space). See the `lme4` help files and package vignettes for more.

2.5.5 Package `mgcv`

Function `gam` from package `mgcv` can also fit mixed models provided they have simple independent Gaussian random effects. `mgcv` does not exploit sparsity of the random effects model matrix at all, but its relatively efficient optimizers mean that it can still be computationally competitive for modest numbers of random effects. Random effects are specified as special cases of 'smooths' using terms like `s(z,x,v,bs="re")` in the model formula arguments of functions `gam` or `bam`. The model matrix specified by such a term is whatever results from `model.matrix(~z:x:v-1)` in R, and the corresponding random effects covariance matrix is simply $\mathbf{I}\sigma_b^2$.

Again refit the simple `Machines` model.

```
> library(mgcv)
> b1 <- gam(score ~ Machine + s(Worker,bs="re") +
+       s(Machine,Worker,bs="re"),data=Machines,method="REML")

> gam.vcomp(b1)
```

```
Standard deviations and 0.95 confidence intervals:

                     std.dev     lower     upper
s(Worker)          4.7810626 2.2498659 10.159965
s(Machine,Worker)  3.7295240 2.3828104  5.837371
scale              0.9615766 0.7632535  1.211432
```

Within the confines of simple i.i.d. random effects an alternative model might allow different worker variances for each machine.

```
> b2 <- gam(score ~ Machine + s(Worker,bs="re") +
+     s(Worker,bs="re",by=Machine),data=Machines,method="REML")
> gam.vcomp(b2)

Standard deviations and 0.95 confidence intervals:

                      std.dev     lower     upper
s(Worker)           3.7859468 1.7987315  7.968612
s(Worker):MachineA  1.9403242 0.2531895 14.869726
s(Worker):MachineB  5.8740228 2.9883339 11.546281
s(Worker):MachineC  2.8454688 0.8299327  9.755842
scale               0.9615766 0.7632536  1.211432

Rank: 5/5
> AIC(b1,b2)
         df      AIC
b1 18.85995 165.1905
b2 18.98557 165.6204
```

The AIC comparison here still suggests that the simpler model is marginally preferable, but note that it is a 'conditional' AIC as opposed to the 'marginal' AIC produced by `lme4`. The distinction is covered in section 6.11 (p. 301).

2.6 Exercises

1. A pig breeding company was interested in investigating litter to litter variability in piglet weight (after a fixed growth period). Six sows were selected randomly from the company's breeding stock, impregnated and 5 (randomly selected) piglets from each resulting litter were then weighed at the end of a growth period. The data were entered into an R data frame, `pig`, with weights recorded in column `w` and a column, `sow`, containing a factor variable indicating which litter the piglet came from. The following R session is part of the analysis of these data using a simple mixed model for piglet weight.

   ```
   > pig$w
    [1]   9.6 10.1 11.2 11.1 10.5  9.5  9.6  9.4  9.5  9.5 11.5
    [12] 10.9 10.8 10.7 11.7 10.7 11.2 11.2 10.9 10.5 12.3 12.1
    [23] 11.2 12.3 11.7 11.2 10.3  9.9 11.1 10.5
   > pig$sow
    [1]  1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 4 4 4 4 4 5 5 5 5 5 6 6
    [28] 6 6 6
   Levels: 1 2 3 4 5 6
   ```

```
> m1<-lm(w~sow,data=pig)
> anova(m1)
Analysis of Variance Table

Response: w
          Df  Sum Sq Mean Sq F value    Pr(>F)
sow        5 15.8777  3.1755  14.897 1.086e-06 ***
Residuals 24  5.1160  0.2132

> piggy<-aggregate(data.matrix(pig),
+                  by=list(sow=pig$sow),mean)
> m0<-lm(w~1,data=piggy)
> summary(m1)$sigma^2
[1] 0.2131667
> summary(m0)$sigma^2
[1] 0.6351067
```

(a) The full mixed model being used in the R session has a random effect for litter/sow and a fixed mean. Write down a full mathematical specification of the model.

(b) Specify the hypothesis being tested by the `anova` function, both in terms of the parameters of the mixed model, and in words.

(c) What conclusion would you draw from the printed ANOVA table? Again state your conclusions both in terms of the model parameters and in terms of what this tells you about pigs.

(d) Using the given output, obtain an (unbiased) estimate of the between litter variance in weight, in the wider population of pigs.

2. Consider a model with two random effects of the form:

$$y_{ij} = \alpha + b_i + c_j + \epsilon_{ij}$$

where $i = 1, \ldots, I$, $j = 1, \ldots, J$, $b_i \sim N(0, \sigma_b^2)$, $c_j \sim N(0, \sigma_c^2)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$ and all these r.v.'s are mutually independent. If the model is fitted by least squares then

$$\hat{\sigma}^2 = RSS/(IJ - I - J + 1)$$

is an unbiased estimator of σ^2, where RSS is the residual sum of squares from the model fit.

(a) Show that, if the above model is correct, the averages $\bar{y}_{i\cdot} = \sum_j y_{ij}/J$ are governed by the model:

$$\bar{y}_{i\cdot} = a + e_i$$

where the e_i are i.i.d. $N(0, \sigma_b^2 + \sigma^2/J)$ and a is a random intercept term. Hence suggest how to estimate σ_b^2.

(b) Show that the averages $\bar{y}_{\cdot j} = \sum_i y_{ij}/I$ are governed by the model:

$$\bar{y}_{\cdot j} = a' + e'_j$$

where the e'_j are i.i.d. $N(0, \sigma_c^2 + \sigma^2/I)$ and a' is a random intercept parameter. Suggest an estimator for σ_c^2.

3. Data were collected on blood cholesterol levels and blood pressure for a group of patients regularly attending an outpatient clinic for a non-heart-disease related condition. Measurements were taken each time the patient attended the clinic. A possible model for the resulting data is

$$y_{ij} = \mu + a_i + \beta x_{ij} + \epsilon_{ij}, \quad a_i \sim N(0, \sigma_a^2) \text{ and } \epsilon_{ij} \sim N(0, \sigma^2),$$

where y_{ij} is the j^{th} blood pressure measurement for the i^{th} patient and x_{ij} is the corresponding cholesterol measurement. β is a fixed parameter relating blood pressure to cholesterol concentration and a_i is a random coefficient for the i^{th} patient. Assume (somewhat improbably) that the same number of measurements are available for each patient.

   (a) Explain how you would test $\mathrm{H}_0 : \sigma_a^2 = 0$ vs. $\mathrm{H}_1 : \sigma_a^2 > 0$ and test $\mathrm{H}_0 : \beta = 0$ vs. $\mathrm{H}_1 : \beta \neq 0$, using standard software for ordinary linear modelling.

   (b) Explain how β and σ_a^2 could be estimated. You should write down the models involved, but should assume that these would be fitted using standard linear modelling software.

4. Write out the following three models in the general form,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \quad \mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\psi}_\theta) \text{ and } \boldsymbol{\epsilon} \sim N(0, \mathbf{I}\sigma^2),$$

where $\mathbf{Z}$ is a matrix containing known coefficients which determine how the response, $\mathbf{y}$, depends on the random effects $\mathbf{b}$ (i.e. it is a 'model matrix' for the random effects). $\boldsymbol{\psi}_\theta$ is the covariance matrix of the random effects $\mathbf{b}$. You should ensure that $\mathbf{X}$ is specified so that the fixed effects are identifiable (you don't need to do this for $\mathbf{Z}$) and don't forget to specify $\boldsymbol{\psi}_\theta$.

   (a) The model from question 3, assuming four patients and two measurements per patient.

   (b) The mixed effects model from section 2.1.1, assuming only two measurements per tree.

   (c) Model (2.6) from section 2.1.4, assuming that $I = 2$, $J = 3$ and $K = 3$.

5. (a) Show that if $\mathbf{X}$ and $\mathbf{Z}$ are independent random vectors, both of the same dimension, and with covariance matrices $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_z$, then the covariance matrix of $\mathbf{X} + \mathbf{Z}$ is $\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_z$.

   (b) Consider a study examining patients' blood insulin levels 30 minutes after eating, y, in relation to sugar content, x, of the meal eaten. Suppose that each of 3 patients had their insulin levels measured for each of 3 sugar levels, and that an appropriate linear mixed model for the j^{th} measurement on the i^{th} patient is

$$y_{ij} = \alpha + \beta x_{ij} + b_i + \epsilon_{ij}, \quad b_i \sim N(0, \sigma^2), \text{ and } \epsilon_{ij} \sim N(0, \sigma^2),$$

where all the random effects and residuals are mutually independent.

i. Write this model out in matrix vector form.
ii. Find the covariance matrix for the response vector $\mathbf{y}$.

6. The R data frame `Oxide` from the `nlme` library contains data from a quality control exercise in the semiconductor industry. The object of the exercise was to investigate sources of variability in the thickness of oxide layers in silicon wafers. The dataframe contains the following columns:

   `Thickness` is the thickness of the oxide layer (in nanometres, as far as I can tell).

   `Source` is a two level factor indicating which of two possible suppliers the sample came from.

   `Site` is a 3 level factor, indicating which of three sites on the silicon wafer the thickness was measured.

   `Lot` is a factor variable with levels indicating which particular batch of Silicon wafers this measurement comes from.

   `Wafer` is a factor variable with levels labelling the individual wafers examined.

   The investigators are interested in finding out if there are systematic differences between the two sources, and expect that thickness may vary systematically across the three sites; they are only interested in the lots and wafers in as much as they are representative of a wider population of lots and wafers.

   (a) Identify which factors you would treat as random and which as fixed, in a linear mixed model analysis of these data.
   (b) Write down a model that might form a suitable basis for beginning to analyse the `Oxide` data.
   (c) Perform a complete analysis of the data, including model checking. Your aim should be to identify the sources of thickness variability in the data and any fixed effects causing thickness variability.

7. Starting from model (2.6) in section 2.1.4, re-analyse the `Machines` data using `lme`. Try to find the most appropriate model, taking care to examine appropriate model checking plots. Make sure that you test whether the interaction in (2.6) is appropriate. Similarly test whether a more complex random effects structure would be appropriate: specifically one in which the machine-worker interaction is correlated within worker. If any data appear particularly problematic in the checking plots, repeat the analysis, and see if the conclusions change.

8. This question follows on from question 7. *Follow-up multiple comparisons* are a desirable part of some analyses. This question is about how to do this in practice. In the analysis of the Machines data the ANOVA table for the fixed effects indicates that there are significant differences between machine types, so an obvious follow-up analysis would attempt to assess exactly where these differences lie. Obtaining Bonferroni corrected intervals for each of the 3 machine to machine differences would be one way to proceed, and this is easy to do.
   First note that provided you have set up the default contrasts using

   ```
   options(contrasts=c("contr.treatment","contr.treatment"))
   ```

(*before calling* `lme`*, of course*) then `lme` will set your model up in such a way that the coefficients associated with the `Machine` effect correspond to the *difference* between the second and first machines, and between the third and first machines. Hence the `intervals` function can produce two of the required comparisons automatically. However, by default the `intervals` function uses the 95% confidence level, which needs to be modified if you wish to Bonferroni correct for the fact that 3 comparisons are being made. If your model object is `m1` then

```
intervals(m1,level=1-0.05/3,which="fixed")
```

will produce 2 of the required intervals. Note the Bonferroni correction '3'. The option `which="fixed"` indicates that only fixed effect intervals are required. The third comparison, between machines B and C, can easily be obtained by changing the way that the factor variable `Machine` is treated, so that machine type B or C count as the 'first machine' when setting up the model. The `relevel` function can be used to do this.

```
levels(Machines$Machine) # check the level names
## reset levels so that 'first level' is "B" ...
Machines$Machine<-relevel(Machines$Machine,"B")
```

Now re-fit the model and re-run the `intervals` function for the new fit. This will yield the interval for the remaining comparison (plus one of the intervals you already have, of course). What are the Bonferroni corrected 95% intervals for the 3 possible comparisons? How would you interpret them?

9. The data frame `Gun` (library `nlme`) is from a trial examining methods for firing naval guns. Two firing methods were compared, with each of a number of teams of 3 gunners; the gunners in each team were matched to have similar physique (Slight, Average or Heavy). The response variable `rounds` is rounds fired per minute, and there are 3 explanatory factor variables, `Physique` (levels `Slight`, `Medium` and `Heavy`); `Method` (levels `M1` and `M2`) and `Team` with 9 levels. The main interest is in determining which method and/or physique results in the highest firing rate and in quantifying team-to-team variability in firing rate.

   (a) Identify which factors should be treated as random and which as fixed, in the analysis of these data.

   (b) Write out a suitable mixed model as a starting point for the analysis of these data.

   (c) Analyse the data using `lme` in order to answer the main questions of interest. Include any necessary follow-up multiple comparisons (as in the previous question) and report your conclusions.