There have been substantial changes to the lme4 package since the 3rd edition appeared, which are reflected in this draft

10

Multi-level Models, and Repeated Measures

This chapter further extends the discussion of models that are a marked departure from the independent errors models of Chapters 5 to 8. In the models that will be discussed in this chapter, there is a hierarchy of variation that corresponds to groupings within the data. The groups are nested. For example, students might be sampled from different classes, that in turn are sampled from different schools. Or, crop yields might be measured on multiple parcels of land at each of a number of different sites.

After fitting such models, predictions can be made at any of the different levels. For example crop yield could be predicted at new sites, or new parcels. Prediction for a new parcel at one of the existing sites is likely to be more accurate than a prediction for a totally new site. Multi-level models, i.e. models which have multiple *error* (or *noise*) terms, are able to account for such differences in predictive accuracy.

Repeated measures models are multi-level models where measurements consist of multiple profiles in time or space; each profile can be viewed as a time series. Such data may arise in a clinical trial, and animal or plant growth curves are common examples; each "individual" is measured at several different times. Typically, the data exhibit some form of time dependence that the model should accommodate.

By contrast with the data that typically appear in a time series model, repeated measures data consist of a multiple profiles through time. Relative to the length of time series that is required for a realistic analysis, each individual repeated measures profile can and often will have values for a small number of time points only. Repeated measures data have, typically, multiple time series that are of short duration.

Ideas that will be central to the discussion of these different models are:

- fixed and random effects,
- variance components, and their connection, in special cases, with expected values of mean squares,
- the specification of mixed models with a simple error structure,
- sequential correlation in repeated measures profiles.

Multi-level model and repeated measures analyses will make extensive use of the function `lmer()` from the package *lme4*, which must be installed. The initial focus will be on examples that can be handled using the more limited abilities of the function `aov()` (base R, *stats*), comparing and contrasting output from `aov()` with output from `lmer()`. The function `lmer()` is a partial replacement for `lme()`, from the older *nlme* package. For later reference, note that objects returned by the function `lmer()` have the class `merMod`.

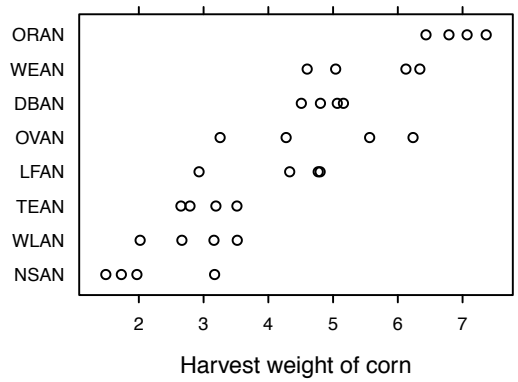

Figure 10.1: Corn yields for four parcels of land in each of eight sites on the Caribbean island of Antigua. Data are in Table 10.1. They are a summarized version (parcel measurements are block means) of a subset of data given in Andrews and Herzberg 1985, pp.339-353. Sites have been reordered according to the magnitude of the site means.

The data set `Orthodont` that is used for the analyses of Subsection 10.7.2, and several data sets that appear in the exercises, are in the *MEMSS* package.

Corn yield measurements example

An especially simple multi-level model is the random effects model for the one way layout. Thus, consider the data frame `ant111b` in the *DAAG* package, based on an agricultural experiment on the Caribbean island of Antigua. Corn yield measurements were taken on four parcels of land within each of eight sites. Figure 10.1 is a visual summary.

Code for Figure 10.1 is:

```
library(lattice); library(DAAG)
Site <- with(ant111b, reorder(site, harvwt, FUN=mean))
stripplot(Site ~ harvwt, data=ant111b, scales=list(tck=0.5),
          xlab="Harvest weight of corn")
```

Figure 10.1 suggests that, as might be expected, parcels on the same site will be relatively similar, while parcels on different sites will be relatively less similar. A farmer whose farm was close to one of the experimental sites might take data from that site as indicative of what he/she might expect. In other cases it may be more appropriate for a farmer to regard his/her farm as a new site, distinct from the experimental sites, so that the issue is one of generalizing to a new site. Prediction for a new parcel at one of the existing sites is more accurate than prediction for a totally new site.

There are two levels of random variation. They are site, and parcel within site. Variation between sites may be due, for example, to differences in elevation or proximity to bodies of water. Within a site, one might expect different parcels to be somewhat similar in terms of elevation and climatic conditions; however, differences in soil fertility and drainage may still have a noticeable effect on yield. (Use of information on such effects, not available as part of the present data, might allow more accurate modeling.)

The model will need: (a) a random term that accounts for variation within sites, and (b) a second superimposed random term that allows variability between parcels that are on different sites to be greater than variation between parcels within sites. The different random terms are known as *random effects*.

The model can be expressed as:

$$\text{yield} \;=\; \text{overall mean} + \underset{\text{(random)}}{\text{site effect}} + \underset{\text{(random)}}{\text{parcel effect (within site)}} \qquad (10.1)$$

Because of the balance (there are four parcels per site), analysis of variance using `aov()` is entirely satisfactory for these data. Section 10.1 that now follows will demonstrate the analysis that uses `aov()`.

It will then be instructive, in Subsection 10.2 below, to set set results from use of `aov()` alongside results from the function `lmer()` (from *lme4*). The comparison is between a traditional analysis of variance approach, which is fine for data from a balanced experimental design, and a general multi-level modeling approach that can in principle handle both balanced and unbalanced designs.

10.1 Corn Yield Data — Analysis Using `aov()`

In the above model, the overall mean is assumed to be a fixed constant, while the site and parcel effects are both assumed to be random. In order to account for the two levels of variation, the model formula must include an `Error(site)` term, thus:

```
library(DAAG)
ant111b.aov <- aov(harvwt ~ 1 + Error(site), data=ant111b)
```

Explicit mention of the "within site" level of variation is unnecessary. (Use of the error term `Error(site/parcel)`, which explicitly identifies parcels within sites, is however allowed.) Output is:

```
> summary(ant111b.aov)

Error: site
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  7  70.34   10.05

Error: Within
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 24 13.861   0.578
```

The analysis of variance (*anova*) table breaks the total sum of squares about the mean into two parts – variation within sites, and variation between site means. Since there are eight sites, the variation between sites is estimated from seven degrees of freedom, after estimating the overall mean. Within each site, estimation of the site mean leaves three degrees of freedom for estimating the variance for that site. Three degrees of freedom at each of eight sites yields 24 degrees of freedom for estimating within site variation.

Table 10.1: *The leftmost column has harvest weights (*`harvwt`*), for the parcels in each site, for the Antiguan corn data. Each of these harvest weights can be expressed as the sum of the overall mean (= 4.29), site effect (fourth column), and residual from the site effect (final column). The information in the fourth and final columns can be used to generate the sums of squares and mean squares for the analysis of variance table.*

Site	Parcel measurements	Site		Residuals from
		means	effects	site mean
DBAN	5.16, 4.8, 5.07, 4.51	4.88	+0.59	0.28, −0.08, 0.18, −0.38
LFAN	2.93, 4.77, 4.33, 4.8	4.21	−0.08	−1.28, 0.56, 0.12, 0.59
NSAN	1.73, 3.17, 1.49, 1.97	2.09	−2.2	−0.36, 1.08, −0.6, −0.12
ORAN	6.79, 7.37, 6.44, 7.07	6.91	+2.62	−0.13, 0.45, −0.48, 0.15
OVAN	3.25, 4.28, 5.56, 6.24	4.83	+0.54	−1.58, −0.56, 0.73, 1.4
TEAN	2.65, 3.19, 2.79, 3.51	3.03	−1.26	−0.39, 0.15, −0.25, 0.48
WEAN	5.04, 4.6, 6.34, 6.12	5.52	+1.23	−0.49, −0.93, 0.81, 0.6
WLAN	2.02, 2.66, 3.16, 3.52	2.84	−1.45	−0.82, −0.18, 0.32, 0.68

Interpreting the mean squares

The division of the sum of squares into two parts mirrors the two different types of prediction that can be based on these data.

First, suppose that measurements are taken on four new parcels at one of the existing sites. How much might the mean of the four measurements be expected to vary, between one such set of measurements and another. For this, the only source of uncertainty is parcel to parcel variation within the existing site. Recall that standard errors of averages can be estimated by dividing the (within) residual mean square by the sample size (in this case, four), and taking the square root. Thus the relevant standard error is $\sqrt{0.578/4} = 0.38$. (Note that this is another form of the pooled variance estimate discussed in Chapter 4.)

Second, for prediction of an average of four parcels at some different site, distinct from the original eight, the relevant standard error can be calculated in the same way, but using the between site mean square; it is $\sqrt{10.05/4} = 1.6$.

Details of the calculations

This subsection may be omitted by readers who already understand the mean square calculations. Table 10.1 contains the data and gives an indication of the mean square calculations used to produce the anova table.

First, the overall mean is calculated. It is 4.29 for this example. Then site means are calculated using the parcel measurements. Site effects are calculated by subtracting the overall mean from the site means. The parcel effects are the residuals after subtracting the site means from the individual parcel measurements.

The between site sum of squares is obtained by squaring the site effects, summing, and multiplying by four. This last step reflects the number of parcels per site. Dividing by the degrees of freedom (8 - 1 = 7) gives the mean square.

The within site sum of squares is obtained by squaring the residuals (parcel effects), summing, and dividing by the degrees of freedom ($8 \times (4\text{-}1) = 24$).

Practical use of the analysis of variance results

Treating site as random when we do the analysis does not at all commit us to treating it as random for purposes of predicting results from a new site. Rather, it allows us this option, if this seems appropriate. Consider how a person who has newly come to the island, and intends to purchase a farming property, might assess the prospects of a farming property that is available for purchase. Two extremes in the range of possibilities are:

1. The property is similar to one of the sites for which data are available, so similar in fact that yields would be akin to those from adding new parcels that together comprise the same area on that site.
2. It is impossible to say with any assurance where the new property should be placed within the range of results from experimental sites. The best that can be done is to treat it as a random sample from the population of all possible sites on the island.

Given adequate local knowledge (and ignoring changes that have taken place since these data were collected!) it might be possible to classify most properties on the island as likely to give yields that are relatively close to those from one or more of the experimental sites. Given such knowledge, it is then possible to give a would-be purchaser advice that is more finely tuned. The standard error (for the mean of four parcels) is likely to be much less than 1.6, and may for some properties be closer to 0.38. In order to interpret analysis results with confidence, and give the would-be purchaser high quality advice, a fact-finding mission to the island of Antigua may be expedient!

Random effects vs. fixed effects

The random effects model bears some resemblance to the one way model considered in Section 4.5. The important difference is that in Section 4.5 the interest was in differences between the *fixed* levels of the nutrient treatment that were used in the experiment. Generalization to other possible nutrient treatments was not of interest, and would not have made sense. The only predictions that were possible were for nutrient treatments considered in the study.

The random effects model allows for predictions at two levels: (1) for agricultural yield at a new location within an existing site, or (2) for locations in sites that were different from any of the sites that were included in the original study.

Nested factors – a variety of applications

Random effects models apply in any situation where there is more than one level of random variability. In many situations, one source of variability is *nested* within the other – thus parcels are nested within sites.

Other examples include: variation between houses in the same suburb, as against variation between suburbs, variation between different clinical assessments of the same patients,

as against variation between patients; variation within different branches of the same business, as against variation between different branches; variations in the bacterial count between subsamples of a sample taken from a lake, as opposed to variation between different samples; variation between the drug prescribing practices of clinicians in a particular specialty in the same hospital, as against variation between different clinicians in different hospitals; and so on. In all these cases, the accuracy with which predictions are possible will depend on which of the two levels of variability is involved. These examples can all be extended in fairly obvious ways to include more than two levels of variability.

Sources of variation can also be *crossed*. For example, different years may be crossed with different sites. Years are not nested in sites, nor are sites nested in years. In agricultural yield trials these two sources of variation may be comparable; see for example Talbot (1984).

10.1.1 A More Formal Approach

Consider now a formal mathematical description of the model. The model is:

$$y_{ij} = \mu + \underset{\text{(site, random)}}{\alpha_i} + \underset{\text{(parcel, random)}}{\beta_{ij}} \quad (i = 1,\ldots,8;\ j = 1,\ldots,4) \tag{10.2}$$

with $\text{var}[\alpha_i] = \sigma^2_{\text{L}}$, $\text{var}[\beta_{ij}] = \sigma^2_{\text{W}}$. The quantities σ^2_L (L=location, another term for site) and σ^2_W (W=within) are referred to as *variance components*.

Variance components allow inferences that are not immediately available from the information in the analysis of variance table. Importantly, the variance components provide information that can help design another experiment.

Relations between variance components and mean squares

The expected values of the mean squares are, in suitably balanced designs such as this, linear combinations of the variance components. The discussion that now follows demonstrates how to obtain the variance components from the analysis of variance calculations. In an unbalanced design, this is not usually possible.

Consider, again, prediction of the average of four parcels within the *i*th existing site. This average can be written as

$$\bar{y}_i = \mu + \alpha_i + \bar{\beta}_i$$

where $\bar{\beta}_i$ denotes the average of the four parcel effects within the *i*th site. Since μ and α_i are constant for the *i*th site (in technical terms, we *condition* on the site being the *i*th), $\text{var}[\bar{y}_i]$ is the square root of $\text{var}[\bar{\beta}_i]$, which equals $\sigma_W/\sqrt{4}$.

In the `aov()` output, the expected mean square for `Error: Within`, i.e., at the within site (between packages) level, is σ^2_W. Thus $\widehat{\sigma^2_W} = 0.578$ and $\text{SE}[\bar{y}_i]$ is estimated as $\widehat{\sigma_W}/\sqrt{4} = \sqrt{0.578/4} = 0.38$.

Next, consider prediction of the average yield at four parcels within a new site. The expected mean square at the site level is $4\sigma^2_L + \sigma^2_W$, i.e., the between site mean square, which in the `aov()` output is 10.05, estimates $4\sigma^2_L + \sigma^2_W$. The standard error for the prediction of

the average yield at four parcels within a new site is

$$\sqrt{\sigma_L^2 + \sigma_W^2/4} = \sqrt{(4\sigma_L^2 + \sigma_W^2)/4}$$

The estimate for this is $\sqrt{10.05/4} = 1.59$.

Finally, note how, in this balanced case, σ_L^2 can be estimated from the analysis of variance output. Equating the expected between site mean square to the observed mean square:

$$4\widehat{\sigma_L^2} + \widehat{\sigma_W^2} = 10.05,$$

i.e.,

$$4\widehat{\sigma_L^2} + 0.578 = 10.05,$$

so that $\widehat{\sigma_L^2}$ = (10.05-0.578)/4 = 2.37.

Interpretation of variance components

In summary, here is how the variance components can be interpreted, for the Antiguan data. Plugging in numerical values ($\widehat{\sigma_W^2}$ = 0.578 and $\widehat{\sigma_L^2}$ = 2.37), take-home messages from this analysis are:

- For prediction for a new parcel at one of the existing sites, the standard error is $\widehat{\sigma_W} = \sqrt{0.578} = 0.76$
- For prediction for a new parcel at a new site, the standard error is $\sqrt{\sigma_L^2 + \sigma_W^2} = \sqrt{2.37 + 0.578} = 1.72$
- For prediction of the mean of n parcels at a new site, the standard error is $\sqrt{\sigma_L^2 + \sigma_W^2/n} = \sqrt{2.37 + 0.578/n}$
  [Notice that while σ_W^2 is divided by n, σ_L^2 is not. This is because the site effect is the same for all n parcels.]
- For prediction of the total of n parcels at a new site, the standard error is $\sqrt{\sigma_L^2 n + \sigma_W^2} = \sqrt{2.37n + 0.578}$

Additionally

- The variance of the difference between two such parcels at the same site is $2\sigma_W^2$
  [Both parcels have the same site effect α_i, so that var(α_i) does not contribute to the variance of the difference.]
- The variance of the difference between two parcels that are in different sites is

$$2(\sigma_L^2 + \sigma_W^2)$$

Thus, where there are multiple levels of variation, the predictive accuracy can be dramatically different, depending on what is to be predicted. Similar issues arise in repeated measures contexts, and in time series.

Intra-class correlation

According to the model, two observations at different sites are uncorrelated. Two observations at the same site are correlated, by an amount that has the name *intra-class correlation* correlation. Here, it equals $\sigma_L^2/(\sigma_L^2 + \sigma_W^2)$. This is the proportion of residual variance explained by differences between sites.

Plugging in the variance component estimates, the intra-class correlation for the corn yield data is $2.37/(2.37 + 0.578) = .804$. Roughly 80% of the yield variation is due to differences between sites.

10.2 Analysis using `lmer()`, from the *lme4* package

In output from the function `lmer()`, the assumption of two nested random effects, i.e., a hierarchy of three levels of variation, is explicit. Variation between sites (this appeared first in the anova table in Subsection 10.1) is the "lower" of the two levels. Here, the *nlme* convention will be followed, and this will be called level 1. Variation between parcels in the same site (this appeared second in the anova table, under "Residuals") is at the "higher" of the two levels, conveniently called level 2.

The modeling command takes the form:

```
library(lme4)
ant111b.lmer <- lmer(harvwt ~ 1 + (1 | site), data=ant111b)
```

The only fixed effect is the overall mean. The `(1 | site)` term fits random variation between sites. Variation between the individual units that are nested within sites, i.e., between parcels, are by default treated as random. Here is the default output:

```
> ## Note that there is no degrees of freedom information.
> print(ant111b.lmer, ranef.comp="Variance", digits=3)
Linear mixed model fit by REML ['lmerMod']
Formula: harvwt ~ 1 + (1 | site)
   Data: ant111b
REML criterion at convergence: 94.4163
Random effects:
 Groups   Name        Variance
 site     (Intercept) 2.368
 Residual             0.578
Number of obs: 32, groups:  site, 8
Fixed Effects:
(Intercept)
       4.29
```

Observe that, according to `lmer()`, $\widehat{\sigma_W^2} = 0.578$, and $\widehat{\sigma_L^2} = 2.368$. Observe also that $\widehat{\sigma_W^2} = 0.578$ is the mean square from the analysis of variance table. The mean square at level 1 does not appear in the output from the `lmer()` analysis.

The processing of output from `lmer()`

The function `coef()` will be used, with output from `summary()`, to obtain estimates of fixed effect coefficients and their standard errors. Thus, for the model `ant111b.lmer`, we obtain:

```
> print(coef(summary(ant111b.lmer)), digits=3)
            Estimate Std. Error t value
(Intercept)     4.29       0.56    7.66
```

Users who require approximate *p*-values can use the function `mixed()` from the *afex* package. A call to `mixed()` replaces the call to `lmer()`. This uses abilities from the *pbkrtest* package to process output from `lmer()`. If called with `method="KR"`, the Kenward-Roger approximation is used to calculate degrees of freedom for statistics in the `t value` column in the output from `lmer()`. With degrees of freedom thus given, the *t*-values are treated as *t* statistics and approximate *p*-values determined.

Objects returned by the function `lmer()` have the class `merMod`. Objects returned by the `summary()` method for `merMod` objects have class `summary.merMod`. Objects returned by `VarCorr()`, used in the sequel for extracting variance component estimates, have class `VarCorr.merMod`.

See (`help(merMod)` for details of methods for `merMod` and `summary.merMod` objects. Note in particular the `print()` methods, with arguments that control the details of what ia printed.[1]

Fitted values and residuals in `lmer()`

In hierarchical multi-level models, fitted values can be calculated at each level of variation that the model allows. Corresponding to each level of fitted values, there is a set of residuals that is obtained by subtracting the fitted values from the observed values.

The default, and at the time of writing the only option, is to calculate fitted values and residuals that adjust for all random effects except the residual. Here, these are estimates of the site expected values. They differ slightly from the site means, as will be seen below. Such fitted values are known as BLUPs (Best Linear Unbiased Predictors). Among linear unbiased predictors of the site means, the BLUPs are designed to have the smallest expected error mean square.

Relative to the site means $\bar{y}_{i.}$, the BLUPs are pulled in toward the overall mean $\bar{y}_{..}$. The most extreme site means will on average, because of random variation, be more extreme than the corresponding "true" means for those sites. For the simple model considered here, the fitted value $\widehat{\alpha_i}$ for the ith site is given by

$$\widehat{y_{i\cdot}} = \bar{y}_{..} + \frac{n\widehat{\sigma_L^2}}{n\widehat{\sigma_L^2} + \widehat{\sigma_W^2}}(\bar{y}_{i\cdot} - \bar{y}_{..}).$$

Shrinkage is substantial, i.e., a shrinkage factor much less than 1.0, when $n^{-1}\widehat{\sigma_W^2}$ is large

[1] Thus, for use of `print()` with `merMod` and `summary.merMod` objects, the argument `ranef.comp` can be set to any combination of `comp="Variance"` and `comp="Std.Dev."`. For use of `print()` with `VarCorr.merMod` objects, the same alternatives are available for the `comp` argument.

relative to $\widehat{\sigma_L^2}$. (For the notation, refer back to equation 10.2.)

As a check, compare the BLUPs given by the above formula with the values from `fitted(ant111b.lmer)`:

```
> s2W <- 0.578; s2L <- 2.37; n <- 4
> sitemeans <- with(ant111b, sapply(split(harvwt, site), mean))
> grandmean <- mean(sitemeans)
> shrinkage <- (n*s2L)/(n*s2L+s2W)
> grandmean + shrinkage*(sitemeans - grandmean)
 DBAN  LFAN  NSAN  ORAN  OVAN  TEAN  WEAN  WLAN
4.851 4.212 2.217 6.764 4.801 3.108 5.455 2.925
> ##
> ## More directly, use fitted() with the lmer object
> unique(fitted(ant111b.lmer))
[1] 4.851 4.212 2.217 6.764 4.801 3.108 5.455 2.925
> ##
> ## Compare with site means
> sitemeans
 DBAN  LFAN  NSAN  ORAN  OVAN  TEAN  WEAN  WLAN
4.885 4.207 2.090 6.915 4.832 3.036 5.526 2.841
```

Observe that the fitted values differ slightly from the site means. For site means below the overall mean (4.29), the fitted values are larger (closer to the overall mean), and for site means above the overall mean, the fitted values are smaller.

Notice that `fitted()` has given the fitted values at level 1, i.e., it adjusts for the single random effect. The fitted value at level 0 is the overall mean, given by `fixef(ant111b.lmer)`. Residuals can be also defined on several levels. At level 0, they are the differences between the observed responses and the overall mean. At level 1, they are the differences between the observed responses and the fitted values at level 1 (which are the BLUPs for the sites).

**Uncertainty in the parameter estimates — profile likelihood and alternatives*

The limits of acceptance of a likelihood ratio test for the null hypothesis of no change in a parameter value can be used as approximate 95% confidence limits for that parameter. Where the likelihood is a function of more than one parameter, the profile likelihood may be used. For any parameter ψ, the profile likelihood is the function of ψ that is obtained by maximizing the likelihood, for each value of ψ, over values of other parameters.[2]

The function `confint()` can be used to pull together the profile information, calculated using the profile method for `merMod` objects, to create approximate confidence intervals:

```
> prof.lmer <- profile(ant111b.lmer)
> CI95 <- confint(prof.lmer, level=0.95)
> rbind("sigmaL\^{}2"=CI95{[}1,{]}\^{}2, "sigma\^{}2"=CI95{[}2,{]}\^{}2)
            2.5 % 97.5 %
sigmaL^2    0.796   6.94
sigma^2     0.344   1.08
```

[2]Note that convergence problems will sometimes occur in the calculation of the profile likelihood, generating warning messages.

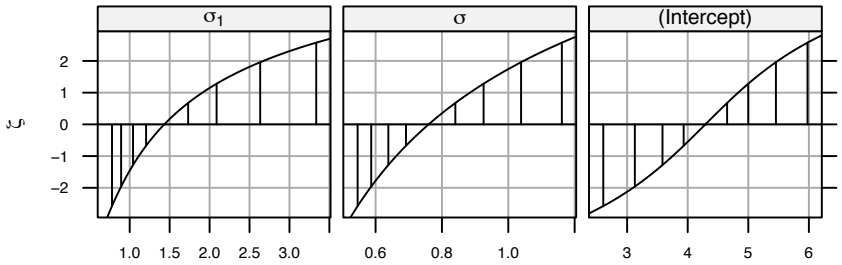

Figure 10.2: Profile likelihoods for the two random and one fixed parameter in the model `ant111b.lmer`. The horizontal scales are σ_1, labeled σ_L in the text,σ, labeled σ_W in the text, and `(Intercept)`. On the vertical scale, the confidence interval limits are labeled according to the equivalent normal deviates. The 95% confidence interval limits are thus at -1.96 and 1.96. The vertical bars are placed at 50%, 80%, 95% and 99% limits.

A 95% confidence interval for the intercept is:

```
> CI95[3,]
             2.5 % 97.5 %
(Intercept) 3.128   5.46
```

The function `confint()`, as used here, returned confidence intervals for σ_L (row label `.sig01`, random), for σ (row label `.sigma`, random), and for `(Intercept)` (fixed). The `(Intercept)` is the intercept in the fitted model, which estimates the overall mean.

The profile likelihoods, scaled so that the lower 2.5% limit transforms to -1.96 and the upper lower 97.5% limit, are shown in Figure 10.2. Code is:

```
library(lattice)
print(xyplot(prof.lmer, conf=c(50, 80, 95, 99)/100,
      aspect=0.8, between=list(x=0.35)))
```

For variances, the horizontal scales show `Std.Dev.` $= \sqrt{\text{Variance}}$. For details of this and other displays that can be used for the output from the `profile()` method for `merMod` objects, see `help(xyplot.thpr)`.

See `help(confint.merMod)` for details of the `confint` method for `merMod` objects. Alternatives to `method="profile"`) are `method="Wald"` or `method="boot"`. The `Wald` method is fast, but based on approximations that can be highly inaccurate. The `boot` method uses repeated fits to suitably constructed bootstrap samples, and can be time consuming. The trustworthiness of results from this method may be questioned if more than an occasional fit fails. See `help(bootMer)` and `help(simulate.merMod)` for further details of `method="boot"`, and for references.

Handling more than two levels of random variation

There can be variation at each of several nested levels. In the example in the next section, atttitude to science scores, on a scale that measured extent of `like`, were obtained for 1385 year 7 students, divided between 66 classes which in turn were divided between 41 schools.

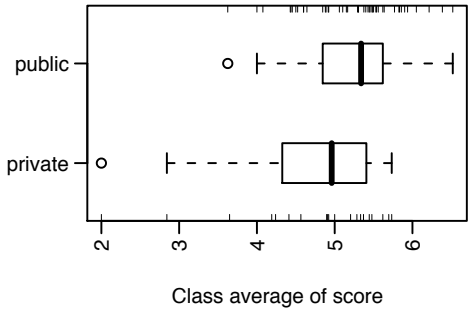

Figure 10.3: Average scores for class, compared between public and private schools.

The analysis in Section 10.3 will treat both school and class as random effects. Using the terminology of the *nlme* package, there are then three levels of random variation — level 3 is pupil, level 2 is class, and level 1 is school. (Note however that the `lmer()` function is not limited to the hierarchical models to which this terminology applies, and does not make formal use of the "levels" terminology.)

The model will also take account of two "fixed effects". One of these accounts for a possible difference between sexes, and the other for a possible differences between public and private schools. Much of the interest is in the implications of the random effects for the accuracy of the fixed effect estimates.

The random effects are in each case assumed to be independent normal variables — one set for schools, one for classes, and one for pupils, operating independently. Careful analysts will be on the watch any indication that failure in some part of this framework of assumptions may compromise the analysis.

10.3 Survey Data, with Clustering

The data that will now be explored are from the data frame `science` (*DAAG*). They are measurements of attitudes to science, from a survey where there were results from 20 classes in 12 private schools and 46 classes in 29 public (i.e. state) schools, all in and around Canberra, Australia. Results are from a total of 1385 year 7 students. The variable `like` is a summary score based on two of the questions. It is on a scale from 1 (dislike) to 12 (like). The number in each class from whom scores were available ranged from 3 to 50, with a median of 21.5. Figure 10.3 compares results for public schools with those for private schools.[3]

[3]
```
## Means of like (data frame science: DAAG), by class
classmeans <- with(science,
                   aggregate(like, by=list(PrivPub, Class), mean) )
  # NB: Class identifies classes independently of schools
  #     class identifies classes within schools
names(classmeans) <- c("PrivPub", "Class", "avlike")
with(classmeans, {
     ## Boxplots: class means by Private or Public school
     boxplot(split(avlike, PrivPub), horizontal=TRUE, las=2,
                   xlab = "Class average of score", boxwex = 0.4)
     rug(avlike[PrivPub == "private"], side = 1)
     rug(avlike[PrivPub == "public"], side = 3)
})
```

10.3.1 Alternative models

Within any one school, we might have

$$y = \text{class effect} + \text{pupil effect}$$

where y represents the attitude measure.

Within any one school, we might use a one-way analysis of variance to estimate and compare class effects. However, this study has the aim of generalizing beyond the classes in the study to all of some wider population of classes, not just in the one school, but in a wider population of schools from which the schools in the study were drawn. In order to be able to generalize in this way, we treat school (`school`), and class (`class`) within school, as random effects. We are interested in possible differences between the sexes (`sex`), and between private and public schools (`PrivPub`). The two sexes are not a sample from some larger population of possible sexes (!), nor are the two types of school (for this study at least) a sample from some large population of types of school. Thus they are fixed effects. The interest is in the specific fixed differences between males and females, and between private and public schools.

The preferred approach is a multi-level model analysis. While it is sometimes possible to treat such data using an approximation to the analysis of variance as for a balanced experimental design, it may be hard to know how good the approximation is. We specify sex (`sex`) and school type (`PrivPub`) as fixed effects, while school (`school`) and class (`class`) are specified as random effects. Class is *nested* within school; within each school there are several classes. The model is

$$y = \underset{\text{(fixed)}}{\text{sex effect}} + \underset{\text{(fixed)}}{\text{type (private or public)}} + \underset{\text{(random)}}{\text{school effect}} + \underset{\text{(random)}}{\text{class effect}} + \underset{\text{(random).}}{\text{pupil effect}}$$

Questions we might ask are:

- Are there differences between private and public schools?
- Are there differences between females and males?
- Clearly there are differences among pupils. Are there differences between classes within schools, and between schools, greater than pupil to pupil variation within classes would lead us to expect?

```
science.lmer <- lmer(like ~ sex + PrivPub + (1 | school) +
                     (1 | school:class), data = science,
                     na.action=na.exclude)
```

The components of variance estimates are:

```
> print(VarCorr(science.lmer), comp="Variance", digits=3)
 Groups       Name        Variance
 school:class (Intercept) 0.321
 school       (Intercept) 0.000
 Residual                 3.052
```

The table of estimates and standard errors for the coefficients of the fixed component is similar to that from an `lm()` (single level) analysis.

```
> print(coef(summary(science.lmer)), digits=2)
              Estimate Std. Error t value
(Intercept)       4.72      0.162    29.1
sexm              0.18      0.098     1.9
PrivPubpublic     0.41      0.186     2.2
```

Groups within the 1383 observations that are included are:

```
> summary(science.lmer)$ngrps
school:class       school
          66           41
```

Degrees of freedom are as follows:

- Between types of school: 41 (number of schools) - 2 = 39
- Between sexes: 1383 - 1 (overall mean) - 1 (differences between males and females) - 65 (differences between the 66 `school:class` combinations) = 1316

The comparison between types of schools compares 12 private schools with 29 public schools, comprising 41 algebraically independent items of information. However because the numbers of classes and class sizes differ between schools, the three components of variance contribute differently to these different accuracies, and the 39 degrees of freedom are for a statistic that has only an approximate *t*-distribution. On the other hand, schools are made up of mixed males and female classes. The between pupils level of variation, where the only component of variance is that for the `Residual` in the output above, is thus the relevant level for the comparison between males and females. The *t*-test for this comparison is, under model assumptions, an exact test with 1316 degrees of freedom.

There are three variance components:

```
Between schools  (school)        0.000
Between classes  (school:class)  0.321
Between students (Residual)      3.052
```

It is important to note that these variances form a nested hierarchy. Variation between students contributes to variation between classes. Variation between students and between classes both contribute to variation between schools. The modest-sized between class component of variance tells us that differences between classes are greater than would be expected from differences between students alone. This will be further discussed shortly.

As the estimate for the between schools component of variance is zero, it can be omitted from the model, leading to the following simpler analysis. Notice that the variance component estimates are, to three decimal places, the same as before:

```
science1.lmer <- lmer(like ~ sex + PrivPub + (1 | school:class),
                      data = science, na.action=na.exclude)
```

Estimates of random and fixed effects are:

```
> print(VarCorr(science1.lmer), comp="Variance", digits=3)
 Groups       Name        Variance
 school:class (Intercept) 0.321
```

```
 Residual                 3.052
> print(coef(summary(science1.lmer)), digits=2)
             Estimate Std. Error t value
(Intercept)      4.72      0.162    29.1
sexm             0.18      0.098     1.9
PrivPubpublic    0.41      0.186     2.2
```

Approximate p-values, if required, can be obtained thus:

```
> library(afex)
> mixed(like ~ sex + PrivPub + (1 | school:class),
        data = na.omit(science), method="KR")
Contrasts set to contr.sum for the following variables: sex, PrivPub, school, class
. . . .
   Effect    F ndf     ddf F.scaling p.value
1     sex 3.44   1 1379.49      1.00     .06
2 PrivPub 4.91   1   60.44      1.00     .03
```

In the output, `ddf` is an acronym for "denominator degrees of freedom".

More detailed examination of the output

Now use the function `confint()` to get approximate 95% confidence intervals for the variance components:

```
> ## Use profile likelihood
> pp <- profile(science1.lmer, which="theta_")
>   # which="theta_": all random parameters
>   # which="beta_": fixed effect parameters
> var95 <- confint(pp, level=0.95)^2
>   # Square to get variances in place of SDs
> rownames(var95) <- c("sigma_Class^2", "sigma^2")
> signif(var95, 3)
              2.5 % 97.5 %
sigma_Class^2 0.178  0.511
sigma^2       2.830  3.300
```

To what extent do differences between classes affect the attitude to science? A measure of the effect is the *intra-class correlation*, which is the proportion of variance that is explained by differences between classes. Here, it equals $0.321/(0.321 + 3.052) = 0.095$. The main influence comes from outside the class that the pupil attends, e.g. from home, television, friends, inborn tendencies, etc.

Do not be tempted to think that, because 0.321 is small relative to the within class component variance of 3.05, it is of little consequence. The variance for the mean of a class that is chosen at random is $0.321 + 3.05/n$. Thus, with a class size of 20, the between class component makes a bigger contribution than the within class component. If all classes were the same size, then the standard error of the difference between class means for public schools and those for private schools would, as there were 20 private schools and 46 public

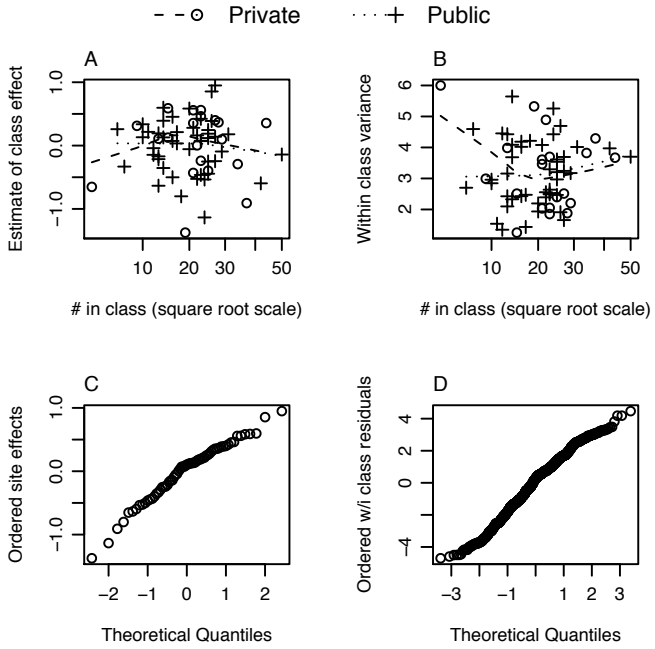

Figure 10.4: Panel A plots class effects against number in the class. Panel B plots within class variance against number in the class. Panels C and D show normal probability plots, for the class effect and for the level 1 residuals (adjusting for the class effect) respectively.

schools, be

$$\sqrt{(0.321 + 3.05/n)\left(\frac{1}{20} + \frac{1}{46}\right)}.$$

From the output table of coefficients and standard errors, we note that the standard error of difference between public and private schools is 0.1857. For this to equal the expression just given, we require $n = 19.1$. Thus the sampling design is roughly equivalent to a balanced design with 19.1 pupils per class.

Figure 10.4 displays information that may be helpful in the assessment of the model. A simplified version of the code is:

```
science1.lmer <- lmer(like ~ sex + PrivPub + (1 | school:class),
                      data = science, na.action=na.omit)
ranf <- ranef(obj = science1.lmer, drop=TRUE)[["school:class"]]
flist <- science1.lmer@flist[["school:class"]]
privpub <- science[match(names(ranf), flist), "PrivPub"]
num <- unclass(table(flist)); numlabs <- pretty(num)
## Panel A: Plot effect estimates vs numbers
plot(sqrt(num), ranf, xaxt="n", pch=c(1,3)[as.numeric(privpub)],
     xlab="# in class (square root scale)",
     ylab="Estimate of class effect")
lines(lowess(sqrt(num[privpub=="private"]),
             ranf[privpub=="private"], f=1.1), lty=2)
```

```
lines(lowess(sqrt(num[privpub=="public"]),
             ranf[privpub=="public"], f=1.1), lty=3)
axis(1, at=sqrt(numlabs), labels=paste(numlabs))
res <- residuals(science1.lmer)
vars <- tapply(res, INDEX=list(flist), FUN=var)*(num-1)/(num-2)
## Panel B: Within class variance estimates vs numbers
plot(sqrt(num), vars, pch=c(1,3)[unclass(privpub)])
lines(lowess(sqrt(num[privpub=="private"]),
             as.vector(vars)[privpub=="private"], f=1.1), lty=2)
lines(lowess(sqrt(num[privpub=="public"]),
             as.vector(vars)[privpub=="public"], f=1.1), lty=3)
## Panel C: Normal probability plot of site effects
qqnorm(ranf, ylab="Ordered site effects", main="")
## Panel D: Normal probability plot of residuals
qqnorm(res, ylab="Ordered w/i class residuals", main="")
```

Panels A shows no clear evidence of a trend. Panel B perhaps suggests that variances may be larger for the small number of classes that had more than about 30 students. Panels C and D show distributions that seem acceptably close to normal. The interpretation of panel C is complicated by the fact that the different effects are estimated with different accuracies.

10.3.2 Instructive, though faulty, analyses

Ignoring class as the random effect

It is important that the specification of random effects be correct. It is enlightening to do an analysis that is not quite correct, and investigate the scope that it offers for misinterpretation. We fit `school`, ignoring `class`, as a random effect. The estimates of the fixed effects change little.

```
> science2.lmer <- lmer(like ~ sex + PrivPub + (1 | school),
+                       data = science, na.action=na.exclude)
> science2.lmer
. . . .
Fixed effects:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.738      0.163   29.00   <2e-16
sexm            0.197      0.101    1.96    0.051
PrivPubpublic   0.417      0.185    2.25    0.030
```

This analysis suggests, wrongly, that the between schools component of variance is substantial. The estimated variance components are[4]

```
Between schools  0.166
Between students 3.219
```

This is misleading. From our earlier investigation, it is clear that the difference is between classes, not between schools!

[4] `print(VarCorr(science2.lmer), comp="Variance", digits=3)`
`## The component of variance that is labeled 'Residual' is`
`## the estimate of the within class variance.`

Ignoring the random structure in the data

Here is the result from a standard regression (linear model) analysis, with `sex` and `PrivPub` as fixed effects:

```
> ## Faulty analysis, using lm
> science.lm <- lm(like ~ sex + PrivPub, data=science)
> summary(science.lm)$coef

              Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.740     0.0996   47.62 0.000000
sexm             0.151     0.0986    1.53 0.126064
PrivPubpublic    0.395     0.1051    3.76 0.000178
```

Do not believe this analysis! The SEs are too small, and the number of degrees of freedom for the comparison between public and private schools is much too large. The contrast is more borderline than this analysis suggests.

10.3.3 Predictive accuracy

The variance of a prediction of the average for a new class of n pupils, sampled in the same way as existing classes, is $0.32 + 3.05/n$. If classes were of equal size, we could derive an equivalent empirical estimate of predictive accuracy by using a resampling method with the class means. With unequal class sizes, use of the class means in this way will be a rough approximation. There were 60 classes. If the training/test set methodology is used, the 60 class means would be divided between a training set and a test set.

An empirical estimate of the within class variation can be derived by applying a resampling method (cross-validation, or the bootstrap) to data for each individual class. The variance estimates from the different classes would then be pooled.

The issues here are important. Data do commonly have a hierarchical variance structure comparable with that for the attitudes to science data. As with the Antiguan corn yield data, the population to which results are to be generalized determines what estimate of predictive accuracy is needed. There are some generalizations, e.g. to another state, that the data cannot support.

10.4 A Multi-level Experimental Design

The data in `kiwishade` are from a designed experiment that compared different kiwifruit shading treatments. [These data relate to Snelgar et al. (1992). Maindonald (1992) gives the data in Table 10.2, together with a diagram of the field layout that is similar to Figure 10.5. The two papers have different shorthands (e.g. Sept–Nov versus Aug–Dec) for describing the time periods for which the shading was applied.] Figure 10.5 shows the layout. In summary:

Note also:

- This is a balanced design with 4 vines per plot, 4 plots per block, and three blocks.
- There are three levels of variation that will be assumed random – between vines within plots, between plots within blocks, and between blocks.

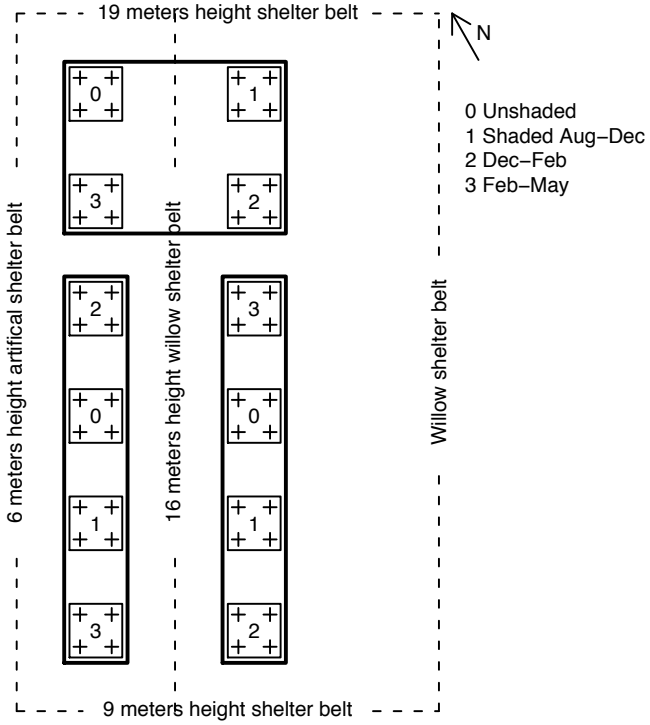

Figure 10.5: The field layout for the kiwifruit shading trial.

- The experimental unit is a plot; this is the level at which the treatment was applied. We have four results (vine yields) per plot.
- Within blocks, treatments were randomly allocated the four plots.

The northernmost plots were grouped together because they were similarly affected by shading from the sun in the north. For the remaining two blocks, shelter effects, whether from the west or from the east, were thought more important. Table 10.2 displays the data.

The `aov()` function allows calculation of an analysis of variance table that accounts for the multiple levels of variation, and allows the use of variation at the plot level to compare treatments. Variance components can be derived, should they be required, from the information in the analysis of variance table. The section will conclude by demonstrating the use of `lmer()` to calculate the variance components directly, and provide information equivalent to that from the use of `aov()`.

The model is

$$\text{yield} = \text{overall mean} + \underset{\text{(random)}}{\text{block effect}} + \underset{\text{(fixed)}}{\text{shade effect}} + \underset{\text{(random)}}{\text{plot effect}} + \underset{\text{(random).}}{\text{vine effect}}$$

We characterize the design thus:

Fixed Effect: shade (treatment).
Random effect: vine (nested) in plot in block, or block/plot/vine.

Table 10.2: *Data from the kiwifruit shading trial. The level names for the factor* `shade` *are mnemonics for the time during which shading was applied. Thus (*`none`*) implies no shading,* `Aug2Dec` *means "August to December", and similarly for* `Dec2Feb` *and* `Feb2May`. The final four columns give yield measurements in kilograms.

Block	Shade	Vine1	Vine2	Vine3	Vine4
east	none	100.74	98.05	97.00	100.31
east	Aug2Dec	101.89	108.32	103.14	108.87
east	Dec2Feb	94.91	93.94	81.43	85.40
east	Feb2May	96.35	97.15	97.57	92.45
north	none	100.30	106.67	108.02	101.11
north	Aug2Dec	104.74	104.02	102.52	103.10
north	Dec2Feb	94.05	94.76	93.81	92.17
north	Feb2May	91.82	90.29	93.45	92.58
west	none	93.42	100.68	103.49	92.64
west	Aug2Dec	97.42	97.60	101.41	105.77
west	Dec2Feb	84.12	87.06	90.75	86.65
west	Feb2May	90.31	89.06	94.99	87.27

Although block is included as a random effect, the estimate of the block component of variance has limited usefulness. On the one hand, the common location of the three blocks has exposed them to similar soil and general climate effects. On the other hand, their different orientations (N, W and E) to sun and prevailing wind will lead to systematic differences. At best, the estimate of the block component of variance will give only the vaguest of hints on the likely accuracy of predictions for other blocks.

There is some limited ability to generalize to other plots and other vines. When horticulturalists apply these treatments in their own orchards, there will be different vines, plots and blocks. Thus, vines and plots are treated as random effects. A horticulturalist will however reproduce, as far as possible, the same shade treatments as were used in the scientific trial, and these are taken as fixed effects. There is however a caveat. In the different climate, soil and other conditions of a different site, these effects may well be different.

10.4.1 The anova table

The model formula that is supplied to `aov()` is an extension of an `lm()` style of model formula that includes an `Error()` term. Observe that each different plot within a block has a different shading treatment, so that the `block:shade` combination can be used to identify plots. Thus the two error terms that need to be specified can be written `block` and `block:shade`. There is a shorthand way to specify both of these together. Write `block/shade`, which expands into `block+block:shade`. Suitable code for the calculation is:

```
## Analysis of variance: data frame kiwishade (DAAG)
kiwishade.aov <- aov(yield ~ shade + Error(block/shade),
                     data=kiwishade)
```

Table 10.3: *Mean squares in the analysis of variance table. The final column gives expected values of mean squares, as functions of the variance components.*

	Df	Sum of Sq	Mean sq	E[Mean sq]
`block` level	2	172.35	86.17	$16\sigma_B^2 + 4\sigma_P^2 + \sigma_V^2$
`block.plt` level				
`shade`	3	1394.51	464.84	$4\sigma_P^2 + \sigma_V^2$ + treatment component
`residual`	6	125.57	20.93	$4\sigma_P^2 + \sigma_V^2$
`block.plt.vines` level	36	438.58	12.18	σ_V^2

The output is:

```
> summary(kiwishade.aov)

Error: block
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals  2 172.348  86.174

Error: block:shade
          Df  Sum Sq Mean Sq F value   Pr(>F)
shade      3 1394.51  464.84  22.211 0.001194
Residuals  6  125.57   20.93

Error: Within
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 36 438.58   12.18
```

Notice the use of `summary()` to give the information that is required. The function `anova()` is, in this context, unhelpful.

Table 10.3 structures the output, with a view to making it easier to follow. The final column will be discussed later, in Subsection 10.4.2.

10.4.2 Expected values of mean squares

The final column of Table 10.3 shows how the variance components combine to give the expected values of mean squares in the analysis of variance table. In this example, calculation of variance components is not necessary for purposes of assessing the effects of treatments. We compare the `shade` mean square with the `residual` mean square for the `block.plt` level. The ratio is

$$F\text{-ratio} = \frac{464.84}{20.93} = 22.2, \text{ on 3 and 6 d.f. } (p = 0.0024).$$

As this is a balanced design, the treatment estimates can be obtained using `model.tables()`:

```
> model.tables(kiwishade.aov, type="means")
. . . .
```

Table 10.4: *Plot means, and differences of yields for individual vines from the plot mean.*

block	shade	Mean	vine1	vine2	vine3	vine4
east	none	99.03	1.285	−2.025	−0.975	1.715
east	Aug2Dec	105.55	3.315	−2.415	2.765	−3.665
east	Dec2Feb	88.92	−3.520	−7.490	5.020	5.990
east	Feb2May	95.88	−3.430	1.690	1.270	0.470
north	none	104.03	−2.915	3.995	2.645	−3.725
north	Aug2Dec	103.59	−0.495	−1.075	0.425	1.145
north	Dec2Feb	93.70	−1.528	0.112	1.062	0.352
north	Feb2May	92.03	0.545	1.415	−1.745	−0.215
west	none	97.56	−4.918	5.932	3.123	−4.138
west	Aug2Dec	100.55	5.220	0.860	−2.950	−3.130
west	Dec2Feb	87.15	−0.495	3.605	−0.085	−3.025
west	Feb2May	90.41	−3.138	4.582	−1.347	−0.097
	Grand mean = 96.53					

```
shade
   none Aug2Dec Dec2Feb Feb2May
 100.20  103.23   89.92   92.77
```

The footnote gives an alternative way to calculate these means.[5]

Treatment differences are estimated within blocks, so that σ_B^2 does not contribute to the standard error of the differences (SED) between means. The SED is, accordingly, $\sqrt{2}\times$ the square root of the residual mean square divided by the sample size: $\sqrt{2} \times \sqrt{(20.93/12)} = 1.87$. The sample size is 12, since each treatment comparison is based on differences between two different sets of 12 vines.

The next subsection will demonstrate calculation of the sums of squares in the analysis of variance table, based on a breakdown of the observed yields into components that closely reflect the different terms in the model. Readers who do not at this point wish to study Subsection 10.4.3 in detail may nevertheless find it helpful to examine Figures 10.6, taking on trust the scalings used for the effects that they present.

10.4.3 The analysis of variance sums of squares breakdown*

This subsection shows how to calculate the sums of squares and mean squares. These details are not essential to what follows, but do give useful insight. The breakdown extends the approach used in the simpler example of Subsection 10.1 and 10.2.

[5]
```
## Calculate treatment means
with(kiwishade, sapply(split(yield, shade), mean))
```

Table 10.5: *Each plot mean is expressed as the sum of overall mean (= 96.53), block effect, shade effect, and residual for the* `block:shade` *combination (or* `plt`*).*

`block`	`shade`	Mean	Block effect	Shade effect	`block:shade` residual
east	none	99.02	0.812	3.670	−1.990
east	Aug2Dec	105.56	0.812	6.701	1.509
east	Dec2Feb	88.92	0.812	−6.612	−1.813
east	Feb2May	95.88	0.812	−3.759	2.294
north	none	104.02	1.805	3.670	2.017
north	Aug2Dec	103.60	1.805	6.701	−1.444
north	Dec2Feb	93.70	1.805	−6.612	1.971
north	Feb2May	92.04	1.805	−3.759	−2.545
west	none	97.56	−2.618	3.670	−0.027
west	Aug2Dec	100.55	−2.618	6.701	−0.066
west	Dec2Feb	87.15	−2.618	−6.612	−0.158
west	Feb2May	90.41	−2.618	−3.759	0.251
			square, add, multiply by 4, divide by df=2, to give ms	square, add, multiply by 4, divide by df=3, to give ms	square, add, multiply by 4, divide by df=6, to give ms

For each plot, we calculate a mean, and differences from the mean. See Table 10.4.[6] Note that whereas we started with 48 observations we have only 12 means. Now we break the means down into overall mean, plus block effect (the average of differences, for that block, from the overall mean), plus treatment effect (the average of the difference, for that treatment, from the overall mean), plus residual.

Table 10.5 uses the information from Table 10.4 to express each plot mean as the sum of a block effect, a shade effect and a residual for the `block:shade` combination. The notes in the last row of each column show how to determine the mean squares in Table 10.3. Moreover, we can scale the values in the various columns so that their standard deviation is the square root of the error mean square, and plot them. Figure 10.6 plots all this information. It shows the individual values, together with one standard deviation limits either side of zero. The chief purpose of these plots is to show the much greater variation at these levels than at the `plt` and vine level.

The estimate of between plot variance in Table 10.3 was 20.93. While larger than the between vine mean square of 12.18, it is not so much larger that the evidence from Figure 10.6 can be more than suggestive. Variation between treatments does appear much greater

[6]
```
## For each plot, calculate mean, and differences from the mean
vine <- paste("vine", rep(1:4, 12), sep="")
vine1rows <- seq(from=1, to=45, by=4)
kiwivines <- unstack(kiwishade, yield ~ vine)
kiwimeans <- apply(kiwivines, 1, mean)
kiwivines <- cbind(kiwishade[vine1rows,  c("block","shade")],
                   Mean=kiwimeans, kiwivines-kiwimeans)
kiwivines <- with(kiwivines, kiwivines[order(block, shade), ])
mean(kiwimeans)      # the grand mean
```

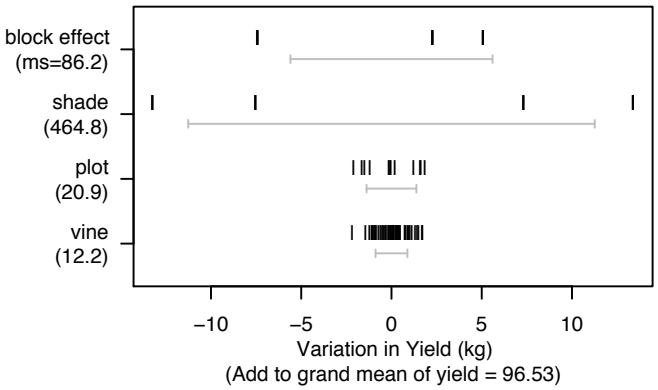

Figure 10.6: Variation at the different levels, for the kiwifruit shading data. The individual data values are given, together with one standard deviation limits either side of zero.

than can be explained from variation between plots, and the same is true for variation between blocks.

We now explain the scaling of effects in Figure 10.6. Consider first the 48 residuals at the vine level. Because 12 degrees of freedom were expended when the 12 plot means were subtracted, the 48 residuals share 36 degrees of freedom and are positively correlated. To enable the residuals to present the appearance of uncorrelated values with the correct variance, we scale the 48 residuals so that the average of their squares is the between vine estimate of variance σ_V^2; this requires multiplication of each residual by $\sqrt{(48/36)}$.

At the level of plot means, we have 6 degrees of freedom to share between 12 plot effects. In addition, we need to undo the increase in precision that results from taking means of four values. Thus, we multiply each plot effect by $\sqrt{(12/6)} \times \sqrt{(4)}$. If the between plot component of variance is zero, the expected value of the average of the square of these scaled effects will be σ_V^2. If the between plot component of variance is greater than zero, the expected value of the average of these squared effects will be greater than σ_V^2.

In moving from plot effects to treatment effects, we have a factor of $\sqrt{(4/3)}$ that arises from the sharing of 3 degrees of freedom between 4 effects, further multiplied by $\sqrt{(12)}$ because each treatment mean is an average of 12 vines. For block effects, we have a multiplier of $\sqrt{(3/2)}$ that arises from the sharing of 2 degrees of freedom between 3 effects, further multiplied by $\sqrt{(16)}$ because each block mean is an average of 16 vines. Effects that are scaled in this way allow visual comparison, as in Figure 10.6.

10.4.4 The variance components

The mean squares in an analysis of variance table for data from a balanced multi-level model can be broken down further, into variance components. The variance components analysis gives more detail about model parameters. Importantly, it provides information that will help design another experiment. Here is the breakdown for the kiwifruit shading data:

- Variation between vines in a plot is made up of one source of variation only. Denote this variance by σ_V^2.
- Variation between vines in different plots is partly a result of variation between vines, and partly a result of additional variation between plots. In fact, if σ_P^2 is the (additional) component of the variation that is due to variation between plots, the expected mean square equals

$$4\sigma_P^2 + \sigma_V^2.$$

  (NB: the 4 comes from 4 vines per plot.)
- Variation between treatments is

$$4\sigma_P^2 + \sigma_V^2 + T$$

  where T (> 0) is due to variation between treatments.
- Variation between vines in different blocks is partly a result of variation between vines, partly a result of additional variation between plots, and partly a result of additional variation between blocks. If σ_B^2 is the (additional) component of variation that is due to differences between blocks, the expected value of the mean square is

$$16\sigma_B^2 + 4\sigma_P^2 + \sigma_V^2$$

  (16 vines per block; 4 vines per plot).

We do not need estimates of the variance components in order to do the analysis of variance. The variance components are helpful for designing another experiment. We calculate the estimates thus:

$$\begin{aligned} \widehat{\sigma}_V^2 &= 12.18, \\ 4\widehat{\sigma}_P^2 + \widehat{\sigma}_V^2 &= 20.93, \text{ i.e. } 4\widehat{\sigma}_P^2 + 12.18 = 20.93. \end{aligned}$$

This gives the estimate $\widehat{\sigma}_P^2 = 2.19$. We can also estimate $\widehat{\sigma}_B^2 = 4.08$.

We are now in a position to work out how much the precision would change if we had 8 (or, say, 10) vines per plot. With n vines per plot, the variance of the plot mean is

$$(n\widehat{\sigma}_P^2 + \widehat{\sigma}_V^2)/n = \widehat{\sigma}_P^2 + \widehat{\sigma}_V^2/n = 2.19 + 12.18/n.$$

We could also ask how much of the variance, for an individual vine, is explained by vine to vine differences. This depends on how much we stretch the other sources of variation. If the comparison is with vines that may be in different plots, the proportion is $12.18/(12.18 + 2.19)$. If we are talking about different blocks, the proportion is $12.18/(12.18 + 2.19 + 4.08)$.

10.4.5 The mixed model analysis

For a mixed model analysis, we specify that treatment (`shade`) is a fixed effect, that `block` and `plot` are random effects, and that `plot` is nested in `block`. The software works out for itself that the remaining part of the variation is associated with differences between vines.

For using `lmer()`, the command is

```
kiwishade.lmer <- lmer(yield ~ shade + (1|block) + (1|block:plot),
                       data=kiwishade)
  # block:shade is an alternative to block:plot
```

The following agree with results from the preceding section:

```
> print(kiwishade.lmer, ranef.comp="Variance", digits=3)
. . . .
Random effects:
 Groups     Name        Variance
 block:plot (Intercept)  2.19
 block      (Intercept)  4.08
 Residual               12.18
Number of obs: 48, groups:  block:plot, 12; block, 3
. . . .
```

Residuals and estimated effects

In this hierarchical model there are three levels of variation: level 1 is between blocks, level 2 is between plots, and level 3 is between vines. The function `fitted()` adjusts for all levels of random variation except between individual vines, i.e. fitted values are at level 2. Unfortunately, `lmer()`, which was designed for use with crossed as well as hierarchical designs, does not recognize the notion of levels. The function `ranef()` can however be used to extract the relevant random effect estimates.

Figure 10.7A plots residuals after accounting for plot and block effects.[7] Figure 10.7B is a normal probability plot that shows the plot effects. The locations of the four plots that suggest departure from normality are printed in the top left of the panel.[8] The plot effects are however estimates from a calculation that involves the estimation of a number of parameters. Before deciding that normality assumptions are in doubt, it is necessary to examine normal probability plots from data that have been simulated according to the normality and other model assumptions. Figure 10.7C shows overlaid normal probability plots from two such simulations. As the present interest is in the normality of the effects, not in variation in standard deviation (this would lead, in Figure 10.7C, to wide variation in aspect ratio), the effects are in each case standardized.[9] It is the plot effects that are immedi-

[7]
```
## Simplified version of plot
xyplot(residuals(kiwishade.lmer) ~ fitted(kiwishade.lmer)|block, data=kiwishade,
               groups=shade, layout=c(3,1), par.strip.text=list(cex=1.0),
               xlab="Fitted values (Treatment + block + plot effects)",
               ylab="Residuals", pch=1:4, grid=TRUE, aspect=1,
               scales=list(x=list(alternating=FALSE), tck=0.5),
               key=list(space="top", points=list(pch=1:4),
                        text=list(labels=levels(kiwishade$shade)),columns=4))
```

[8]
```
## Simplified version of graph that shows the plot effects
ploteff <- ranef(kiwishade.lmer, drop=TRUE)[[1]]
qqmath(ploteff, xlab="Normal quantiles", ylab="Plot effect estimates",
       aspect=1, scales=list(tck=0.5))
```

[9]
```
## Overlaid normal probability plots of 2 sets of simulated effects
## To do more simulations, change nsim as required, and re-execute
simvals <- simulate(kiwishade.lmer, nsim=2)
simeff <- apply(simvals, 2, function(y) scale(ranef(refit(kiwishade.lmer, y),
                                                    drop=TRUE)[[1]]))
simeff <- data.frame(v1=simeff[,1], v2=simeff[,2])
qqmath(~ v1+v2, data=simeff, xlab="Normal quantiles",
       ylab="Simulated plot effects\n(2 sets, standardized)",
              scales=list(tck=0.5), aspect=1)
```

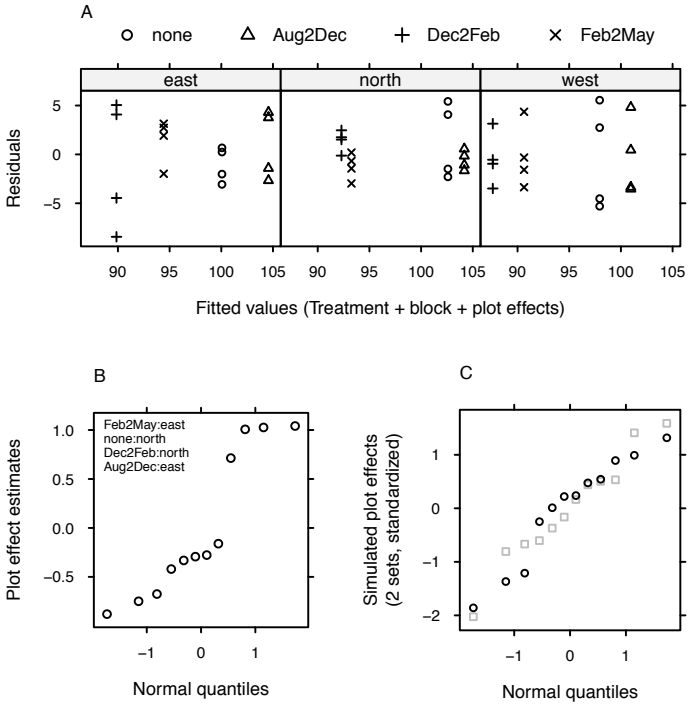

Figure 10.7: Panel A shows standardized residuals after fitting block and plot effects, plotted against fitted values. There are 12 distinct fitted values, one for each plot. Panel B is a normal probability plot that shows the plot effects. The names in the top left hand corner identify the plots with the largest residuals. Panel C shows overlaid normal probability plots of plot effects from two simulations.

ately relevant to assessing the validity of assumptions that underly statistical comparisons between treatment means, not the residuals. The plot effect estimates seems clearly inconsistent with the assumption of normal plot effects. Remember however that each treatment mean is an average over three plots. This averaging will take the sampling distribution of the treatment means closer to normality.

It may be relevant to Figure 10.7B to note that the treatment means are, in order,

```
Dec2Feb Feb2May  none    Aug2Dec
89.92   92.77    100.20  103.23
```

Notice that the plot-specific effects go in opposite directions, relative to the overall treatment means, in the `east` and `north` blocks.

10.4.6 Predictive accuracy

We have data for one location on one site only. We thus cannot estimate a between location component of variance for different locations on the current site, let alone a between site component of variance. Use of resampling methods will not help; the limitation is inherent in the experimental design.

Where data are available from multiple sites, the site to site component of variance will

almost inevitably be greater than zero. Given adequate data, the estimate of this component of variance will then also be greater than zero, even in the presence of explanatory variable adjustments that attempt to adjust for differences in rainfall, temperature, soil type, etc. (Treatment differences are often, but by no means inevitably, more nearly consistent across sites than are the treatment means themselves.)

Where two (or more) experimenters use different sites, differences in results are to be expected. Such different results have sometimes led to acriminious exchanges, with each convinced that there are flaws in the other's experimental work. Rather, assuming that both experiments were conducted with proper care, the implication is that both sets of results should be incorporated into an analysis that accounts for site to site variation. Better still, plan the experiment from the beginning as a multi-site experiment!

10.5 Within and Between Subject Effects

The data frame `tinting` is from an experiment that aimed to model the effects of the tinting of car windows on visual performance. (For more information, see Burns et al., 1999.) Interest is focused on effects on side window vision, and hence on visual recognition tasks that would be performed when looking through side windows.

The variables are `csoa` (critical stimulus onset asynchrony, i.e., the time in milliseconds required to recognize an alphanumeric target), `it` (inspection time, i.e., the time required for a simple discrimination task) and `age`, while `tint` (three levels) and `target` (two levels) are ordered factors. The variable `sex` is coded `f` for females and `m` for males, while the variable `agegp` is coded `Younger` for young people (all in their 20s) and `Older` for older participants (all in their 70s).

Data were collected in two sessions, with half the individuals undertaking the `csoa` task in the first session and the `it` task in the second session, and the other half doing these two types of task in the reverse order. Within each session, the order of presentation of the two levels of target contrast was balanced over participants. For each level of target contrast the levels of `tint` were in the order `no` (100% VLT = visible light transmittance), `lo` (81.3% VLT = visible light transmittance) and `hi` (35% VLT = visible light transmittance). Each participant repeated the combination of high contrast with no tinting (100% VLT) at the end of the session. Comparison with the same task from earlier in the session thus allows a check on any change in performance through the session.

We have two levels of variation – within individuals (who were each tested on each combination of `tint` and `target`), and between individuals. Thus we need to specify `id` (identifying the individual) as a random effect. Plots of the data make it clear that, to have variances that are approximately homogeneous, we need to work with `log(csoa)` and `log(it)`. Here, we describe the analysis for `log(it)`.

Model fitting criteria

The function `lmer()` allows use of one of two criteria: restricted (or residual) maximum likelihood (REML), which is the default, and maximum likelihood (ML). The parameter estimates from the REML method are generally preferred to those from ML, as more nearly unbiased. Comparison of models using `anova()` relies on maximum likelihood theory, and

the models should be fitted using ML.

10.5.1 Model selection

A good principle is to limit initial comparisons between models to several alternative models within a hierarchy of increasing complexity. For example, consider main effects only, main effects plus all first order interactions, main effects plus all first and second order interactions, as far on up this hierarchy as seems reasonable. This makes for conceptual clarity, and simplifies inference. (Where a model has been selected from a large number of candidate models, selection effects come into play, and inference must account for this.)

Here, three models will be considered:

1. All possible interactions (this is likely to be more complex than is needed):
```
## Change initial letters of levels of tinting$agegp to upper case
library(R.utils)
levels(tinting$agegp) <- capitalize(levels(tinting$agegp))
## Fit all interactions: data frame tinting (DAAG)
it3.lmer <- lmer(log(it) ~ tint*target*agegp*sex + (1 | id),
                 data=tinting, REML=FALSE)
```
2. All two-factor interactions (this is a reasonable guess; two-factor interactions may be all we need):
```
it2.lmer <- lmer(log(it) ~ (tint+target+agegp+sex)^2 + (1 | id),
                 data=tinting, REML=FALSE)
```
3. Main effects only (this is a very simple model):
```
it1.lmer <- lmer(log(it)~(tint+target+agegp+sex) + (1 | id),
                 data=tinting, REML=FALSE)
```

The use of `REML=FALSE`; is advisable for the `anova()` (likelihood ratio) comparison that now follows:

```
> anova(it1.lmer, it2.lmer, it3.lmer)
Data: tinting
Models:
it1.lmer: log(it) ~ (tint + target + agegp + sex) + (1 | id)
it2.lmer: log(it) ~ (tint + target + agegp + sex)^2 + (1 | id)
it3.lmer: log(it) ~ tint * target * agegp * sex + (1 | id)
         Df  AIC  BIC logLik Chisq Chi Df Pr(>Chisq)
it1.lmer  7 -0.9 21.6    7.4
it2.lmer 16 -5.7 45.5   18.9 22.88      9     0.0065
it3.lmer 25  6.1 86.2   21.9  6.11      9     0.7288
```

Notice that `Df` is now used for degrees of freedom, where `DF` was used in connection with `summary.aov()`. earlier.

The p-value for comparing model 1 with model 2 is 0.73, while that for comparing model 2 with model 3 is 0.0065. This suggests that the model that limits attention to two-factor interactions is adequate. (Note also that the AIC statistic favors model 2. The BIC statistic, which is an alternative to AIC, favors the model that has main effects only.

Hastie et al. (2009, p. 235) suggest, albeit in reference to models with i.i.d. errors, that BIC's penalty for model complexity can be unduly severe when the number of residual degrees of freedom is small. (Note also that the different standard errors are based on vari-

ance component information at different levels of the design, so that the critique in Vaida and Blanchard (2005) perhaps makes the use of either of these statistics problematic. See Spiegelhalter et al. (2002) for various alternatives to AIC and BIC that may be better suited to use with models with "complex" error structures. Our advice is to use all such statistics with caution, and to consider carefully implications that may arise from the intended use of model results.)

The analysis of variance table indicated that main effects together with two-factor interactions were enough to explain the outcome. Interaction plots, looking at the effects of factors two at a time, are therefore an effective visual summary of the analysis results. In the table of coefficients that appears below, the highest *t*-statistics for interaction terms are associated with `tint.L:agegpOlder`, `targethicon:agegpOlder`, `tint.L:targethicon` and `tint.L:sexm`. It makes sense to look first at those plots where the interaction effects are clearest, i.e. where the *t*-statistics are largest. The plots may be based on either observed data or fitted values, at the analyst's discretion.[10]

10.5.2 Estimates of model parameters

For exploration of parameter estimates in the model that includes all two-factor interactions, we re-fit the model used for `it2.lmer`, but now setting `REML=TRUE` (restricted maximum likelihood estimation), and examine the estimated effects. The parameter estimates that come from the `REML` analysis are in general preferable, because they avoid or reduce the biases of maximum likelihood estimates. (See, e.g., Diggle et al. (2002). The difference from likelihood can however be of little consequence.)

```
> it2.reml <- update(it2.lmer, REML=TRUE)
> print(coef(summary(it2.reml)), digits=2)
                      Estimate Std. Error t value    DF
(Intercept)             3.6191      0.130   27.82   145
tint.L                  0.1609      0.044    3.64   145
tint.Q                  0.0210      0.045    0.46   145
targethicon            -0.1181      0.042   -2.79   145
agegpolder              0.4712      0.233    2.02    22
sexm                    0.0821      0.233    0.35    22
tint.L:targethicon     -0.0919      0.046   -2.00   145
tint.Q:targethicon     -0.0072      0.048   -0.15   145
tint.L:agegpolder       0.1308      0.049    2.66   145
tint.Q:agegpolder       0.0697      0.052    1.34   145
tint.L:sexm            -0.0979      0.049   -1.99   145
tint.Q:sexm             0.0054      0.052    0.10   145
targethicon:agegpolder -0.1389      0.058   -2.38   145
targethicon:sexm        0.0779      0.058    1.33   145
agegpolder:sexm         0.3316      0.326    1.02    22
```

[10] `## Code that gives the first four such plots, for the observed data`
`interaction.plot(tinting$tint, tinting$agegp, log(tinting$it))`
`interaction.plot(tinting$target, tinting$sex, log(tinting$it))`
`interaction.plot(tinting$tint, tinting$target, log(tinting$it))`
`interaction.plot(tinting$tint, tinting$sex, log(tinting$it))`

```
>  # NB: The final column, giving degrees of freedom, is not in the
>  # summary output. It is our addition.
```

Because `tint` is an ordered factor with three levels, its effect is split up into two parts. The first, which always carries a `.L` (linear) label, checks if there is a linear change across levels. The second part is labeled `.Q` (quadratic), and as `tint` has only three levels, accounts for all the remaining sum of squares that is due to `tint`. A comparable partitioning of the effect of `tint` carries across to interaction terms also.

The *t*-statistics for interactions involving `tint.Q` are 0.46, -0.15, 1.34 and 1.10. The output can be simplified by omitting these interactions.

None of the main effects and interactions involving `agegp` and `sex` are significant at the conventional 5% level, though `agegp` comes close. On the other hand, the interaction terms (`tint.L:agegpOlder`, `targethicon:agegpOlder`, `tint.L:targethicon` and `tint.L:sexm`) that are statistically significant stand out much less clearly in Figures 2.12A and 2.12B.

This may seem inconsistent with Figures 2.12A and 2.12B, where it is the older males who seem to have the longer times. To resolve this apparent inconsistency, observe that

- Comparisons that relate to `agegp` and `sex` are made relative to variation between individuals. Standard errors for such comparisons, in the output, are in the range 0.23 - 0.32, in each case with 22 degrees of freedom. (There are 9 younger and 4 older females, against 4 younger and 9 older males.[11])
- Comparisons between levels of `tint` or `target` are made several times for each of the 26 individuals, and are relatively consistent from one individual to another. Standard errors for these comparisons are small – in the range 0.042 - 0.058.

Statistical variation cannot be convincingly ruled out as the explanation for the effects that stand out most strongly in the graphs. The graphs are not designed to highlight the consistency with which individuals respond to differences between levels of tinting and target contrast.

10.6 A Generalized Linear Mixed Model

Consider again the `moths` data of Subsection 8.4.2. The analysis in Subsection 8.4.2 assumed a quasipoisson error, which uses a constant multiplier for the Poisson variance. It may be better to assume a random between transects error that is additive on the scale of the linear predictor. The model incorporates a term that allows for normally distributed random variation, additional to the poisson variation at each observation. Technically, this is an example of the use of "observation level random effects".

The attempt to fit a model that uses the default log link generates (lme4_1.1-7), if data for the habitat `Bank` is included, a warning that the model is nearly unidentifiable. This problem is avoided if a square root link is used.

The code is:

```
moths$transect <- 1:41  # Each row is from a different transect
moths$habitat <- relevel(moths$habitat, ref="Lowerside")
```

[11] `subs <- with(tinting, match(unique(id), id)); with(tinting, table(sex[subs], agegp[subs]))`

```
A.glmer <-  glmer(A~habitat+sqrt(meters)+(1|transect),
                family=poisson(link=sqrt), data=moths)
print(summary(A.glmer), show.resid=FALSE, correlation=FALSE)
```

Output is:

```
. . . .
     AIC      BIC   logLik deviance df.resid
     213      230      -96      193       31

Random effects:
 Groups   Name        Variance Std.Dev.
 transect (Intercept) 0.319    0.564
Number of obs: 41, groups:  transect, 41

Fixed effects:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.7322     0.3513    4.93  8.2e-07
habitatBank       -2.0415     0.9377   -2.18    0.029
habitatDisturbed  -1.0359     0.4071   -2.54    0.011
habitatNEsoak     -0.7319     0.4323   -1.69    0.090
habitatNWsoak      2.6787     0.5101    5.25  1.5e-07
habitatSEsoak      0.1178     0.3923    0.30    0.764
habitatSWsoak      0.3900     0.5260    0.74    0.458
habitatUpperside  -0.3135     0.7549   -0.42    0.678
sqrt(meters)       0.0675     0.0631    1.07    0.285
```

The Poisson component of the variance, on the square root scale of the linear predictor, is 0.25. The observation level random effect, labelled `transect` in the above output, increases this by 0.319 to 0.569, i.e., by a factor of 2.28. Compare this with the increase by a factor of 2.7 for the quasipoisson model.

Now compare, between the quasipoisson model and the observation level random effects model, predicted values for habitats and standard errors of difference from `Lowerside`:[12]

```
           fit-glm glm-SE fit=glmer glmer-SE
Lowerside     2.13   0.00      1.99     0.00
Bank          0.00   0.86      0.00     0.95
Disturbed     1.07   0.41      0.86     0.40
NEsoak        1.53   0.43      1.42     0.41
NWsoak        4.86   0.54      4.72     0.51
SEsoak        2.30   0.41      2.18     0.39
SWsoak        2.58   0.54      2.53     0.51
Upperside     2.37   0.45      2.34     0.43
```

[12]
```
A1quasi.glm <- glm(A~habitat, data=moths, family=quasipoisson(link=sqrt))
A1.glmer <- glmer(A~habitat+(1|transect), data=moths, family=poisson(link=sqrt))
Cglm <- coef(summary(A1quasi.glm))
Cglmer <- coef(summary(A1.glmer))
fitboth <- cbind("fit-glm"=Cglm[1,1]+c(0, Cglm[-1,1]), "glm-SE"=c(0, Cglm[-1,2]),
                 "fit=glmer"=Cglmer[1,1]+c(0, Cglmer[-1,1]), "glmer-SE"=c(0, Cglmer[-1,2]))
rownames(fitboth)[-1] <- substring(rownames(fitboth)[-1],8)
rownames(fitboth)[1] <- "Lowerside"
round(fitboth, 2)  # NB, all SEs are for the difference from 'Lowerside'
```

Observe that the standard errors for comparisons with `Lowerside` are similar for the two models. The fitted values in the the observation level random effects model are pulled in towards zero, relative to the quasipoisson model.

It is left as an exercise for the reader to compare the plots of residuals versus fitted values between the two models.

Mixed models with a binomial error and logit link

On a logit scale, the binomial contribution to the error increases as the expected value moves away from 0.5. (One the scale of the response, however, error decreases as the expected value moves away from 0.5). Thus, relative to a quasibinomial model, the SED will be reduced for more extreme comparisons, and increased for less extreme comparisons.

10.7 Repeated Measures in Time

Whenever we make repeated measurements on a treatment unit we are, technically, working with repeated measures. In this sense, both the kiwifruit shading data and the window tinting data sets were examples of repeated measures data sets. Here, our interest is in generalizing the multi-level modeling approach to handle models for longitudinal data, i.e., data where the measurements were repeated at different times.

In the kiwifruit shading experiment, we gathered data from all vines in a plot at the one time. In principle, we might have taken data from different vines at different points in time. For each plot, there would then be data at each of four time points.

There is a close link between a wide class of repeated measures models and time series models. In the time series context, there is usually just one realization of the series, which may however be observed at a large number of time points. In the repeated measures context, there may be a large number of realizations of a series that is typically quite short.

Perhaps the simplest case is where there is no apparent trend with time. Thus consider data from a clinical trial of a drug (progabide) used to control epileptic fits. (For an analysis of data from the study to which this example refers, see Thall and Vail, 1990.) The analysis assumes that patients were randomly assigned to the two treatments – placebo and progabide. After an eight-week run-in period, data were collected, both for the placebo group and for the progabide group, in each of four successive two-week periods. The outcome variable was the number of epileptic fits over that time.

One way to do the analysis is to work with the total number of fits over the four weeks, perhaps adjusted by subtracting the baseline value. It is possible that we might obtain extra sensitivity by taking account of the correlation structure between the four sets of fortnightly results, taking a weighted sum rather than a mean.

Where there is a trend with time, working with a mean over all times will not usually make sense. Any or all of the following can occur, both for the overall pattern of change and for the pattern of difference between one profile and another.

1. There is no trend with time.
2. The pattern with time may follow a simple form, e.g., a line or a quadratic curve.
3. A general form of smooth curve, e.g., a curve fitted using splines, may be required to

account for the pattern of change with time.

The theory of repeated measures modeling

For the moment, profiles (or subjects) are assumed independent. The analysis must allow for dependencies between the results of any one subject at different times. For a balanced design, we will assume n subjects ($i = 1, 2, \ldots, n$) and p times ($j = 1, 2, \ldots, p$), though perhaps with missing responses (gaps) for some subjects at some times. The plot of response versus time for any one subject is that subject's *profile*.

A key idea is that there are (at least) two levels of variability – between subjects and within subjects. In addition, there is measurement error.

Repeating the same measurement on the same subject at the same time will not give exactly the same result. The between subjects component of variation is never observable separately from sources of variation that operate "within subjects". In any data that we collect, measurements are always affected by "within subjects" variability, plus measurement error. Thus the simplest model that is commonly used has a between subjects variance component denoted by ν^2, while there is a within subjects variance at any individual time point that is denoted by σ^2. The measurement error may be bundled in as part of σ^2. The variance of the response for one subject at a particular time point is $\nu^2 + \sigma^2$.

In the special case just considered, the variance of the difference between two time points for one subject is $2\sigma^2$. Comparisons "within subjects" are more accurate than comparisons "between subjects".

**Correlation structure*

The time dependence of the data has implications for the correlation structure. The simple model just described takes no account of this structure. Often, the separation of points in time is reflected in a correlation between time points that decreases as the time separation increases. The variance for differences between times may increase as points move further apart in time.

We have seen that correlation structure is also a key issue in time series analysis. A limitation, relative to repeated measures, is that in time series analysis the structure must typically be estimated from just one series, by assuming that the series is in some sense part of a repeating pattern. In repeated measures there may be many realizations, allowing a relatively accurate estimate of the correlation structure. By contrast with time series, the shortness of the series has no effect on our ability to estimate the correlation structure. Multiple realizations are preferable to a single long series.

While we are typically better placed than in time series analysis to estimate the correlation structure there is, for most of the inferences that we commonly wish to make, less need to know the correlation structure. Typically our interest is in the consistency of patterns between individuals. For example we may want to know: "Do patients on treatment A improve at a greater rate than patients on treatment B?"

There is a broad distinction between approaches that model the profiles, and approaches that focus more directly on modeling the correlation structure. Direct modeling of the profiles leads to random coefficient models, which allow each individual to follow their own profile. Variation between profiles may largely account for the sequential correlation structure. Direct modeling of the correlation is most effective when there are no evident systematic differences between profiles.

For further discussion of repeated measures modeling, see Diggle et al. (2002); Pinheiro and Bates (2000). The Pinheiro and Bates book is based around the S-PLUS version of the *nlme* package.

Different approaches to repeated measures analysis

Traditionally, repeated measures models have been analyzed in many different ways. Here is a summary of methods that have been used:

- A simple way to analyze repeated measures data is to form one or more summary statistics for each subject, and then use these summary statistics for further analysis.
- When the variance is the same at all times and the correlation between results is the same for all pairs of times, data can in principle be analyzed using an analysis of variance model. This allows for two components of variance, (1) between subjects and (2) between times. An implication of this model is that the variance of the difference is the same for all pairs of time points, an assumption that is, in general, unrealistic.
- Various adjustments adapt the analysis of variance approach to allow for the possibility that the variance of time differences are not all equal. These should be avoided now that there are good alternatives to the analysis of variance approach.
- Multivariate comparisons accommodate all possible patterns of correlations between time points. This approach accommodates the time series structure, but does not take advantage of it to find an economical parameterization of the correlation structure.
- Repeated measures models aim to reflect the sequential structure, in the fixed effects, in the random effects, and in the correlation structure. They do this in two ways: by modeling the overall pattern of difference between different profiles, and by direct modeling of the correlation structure. This modeling approach often allows insights that are hard to gain from approaches that ignore or do not take advantage of the sequential structure.

10.7.1 Example – random variation between profiles

The data frame `humanpower1` has data from investigations (Bussolari, 1987; Nadel and Bussolari, 1988) designed to assess the feasibility of a proposed 119 kilometer human powered flight from the island of Crete – in the initial phase of the *Daedalus* project. After an initial 5-minute warm-up period and 5-minute recovery period, the power requirements from the athletes were increased, at two-minute intervals, in steps of around 30 watts. Figure 10.8 gives a visual summary of the data.[13]

We leave it as an exercise to verify, using a fixed effects analysis such as was described in Section 7.3, that separate lines are required for the different athletes, and that there is no

[13]
```
## Plot points and fitted lines (panel A)
library(lattice)
xyplot(o2 ~ wattsPerKg, groups=id, data=humanpower1,
        panel=function(x,y,subscripts,groups,...){
           yhat <- fitted(lm(y ~ groups*x))
           panel.superpose(x, y, subscripts, groups, pch=1:5)
           panel.superpose(x, yhat, subscripts, groups, type="l")
        },
        xlab="Watts per kilogram",
        ylab=expression("Oxygen intake ("*ml.min^{-1}*.kg^{-1}*")"))
```

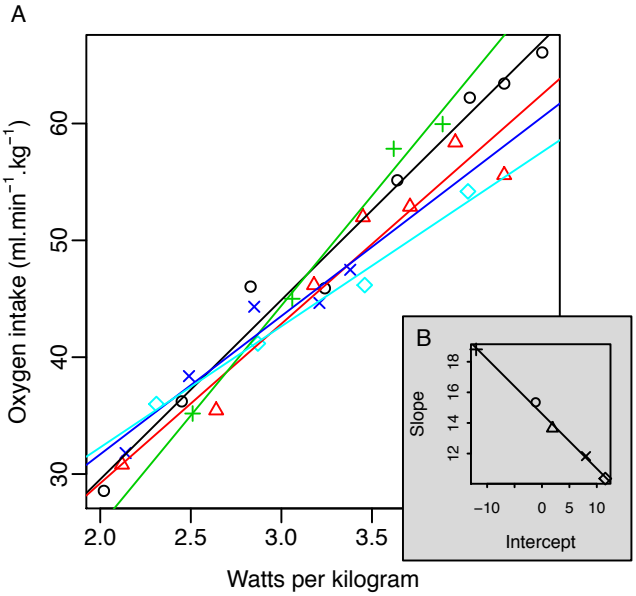

Figure 10.8: Panel A shows oxygen intake, plotted against power output, for each of five athletes who participated in investigations designed to assess the feasibility of a proposed *Daedalus* 119 kilometer human powered flight. Panel B plots the slopes of these separate lines against the intercepts. A fitted line, with a slope of 2.77, has been added.

case for anything more complicated than straight lines. The separate lines fan out at the upper extreme of power output, consistent with predictions from a random slopes model.

Separate lines for different athletes

The model is

$$y_{ij} = \alpha + \beta x_{ij} + a + b x_{ij} + e_{ij}$$

where i refers to individual, and j to observation j for that individual, α and β are fixed, a and b have a joint bivariate normal distribution, each with mean 0, independently of the e_{ij} which are i.i.d. normal. Each point in Figure 10.8B is a realization of an $(\alpha + a, \beta + b)$ pair.

The following is the code that handles the calculations:

```
## Calculate intercepts and slopes; plot Slopes vs Intercepts
## Uses the function lmList() from the lme4 package
library(lme4)
hp.lmList <- lmList(o2 ~ wattsPerKg | id, data=humanpower1)
coefs <- coef(hp.lmList)
names(coefs) <- c("Intercept", "Slope")
plot(Slope ~ Intercept, data=coefs)
abline(lm(Slope~Intercept, data=coefs))
```

Note the formula `o2 ~ wattsPerKg | id` that is given as argument to the function `lmList()`. For each different level of the factor `id`, there is a regression of `o2`

on wattsPerKg. Summary information from the calculations is stored in the object hp.lmList:

A random coefficients model

Two possible reasons for modeling the variation between slopes as a random effectare:

- There may be an interest in generalizing to further athletes, selected in a similar way — what range of responses is it reasonable to expect?
- The fitted lines from the random slopes model may be a better guide to performance than the fitted "fixed" lines for individual athletes. The range of the slopes for the fixed lines will on average exaggerate somewhat the difference between the smallest and largest slope, an effect which the random effects analysis corrects.

Here, the major reason for working with these data is that they demonstrate a relatively simple application of a random effects model. Depending on how results will be used a random coefficients analysis may well, for these data, be overkill!

The model that will now be fitted allows, for each different athlete, a random slope (for wattsPerKg) and random intercept. We expect the correlation between the realizations of the random intercept and the random slope to be close to 1. As it will turn out, this will not create any undue difficulty. Calculations proceed thus:

```
> hp.lmer <- lmer(o2 ~ wattsPerKg + (wattsPerKg | id),
+                 data=humanpower1)
> print(summary(hp.lmer), digits=3)
Linear mixed model fit by REML ['lmerMod']
Formula: o2 ~ wattsPerKg + (wattsPerKg | id)
   Data: humanpower1
. . . .
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 id       (Intercept) 50.73    7.12
          wattsPerKg   7.15    2.67     -1.00
 Residual              4.13    2.03
Number of obs: 28, groups:  id, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept)     2.09       3.78    0.55
wattsPerKg     13.91       1.36   10.23

Correlation of Fixed Effects:
           (Intr)
wattsPerKg -0.992
```

The predicted lines from this random lines model are shown as dashed lines in Figure 10.9A. These are the BLUPs that were discussed earlier in this chapter.

```
hat <- fitted(hp.lmer)
lmhat <- with(humanpower1, fitted(lm(o2 ~ id*wattsPerKg)))
```

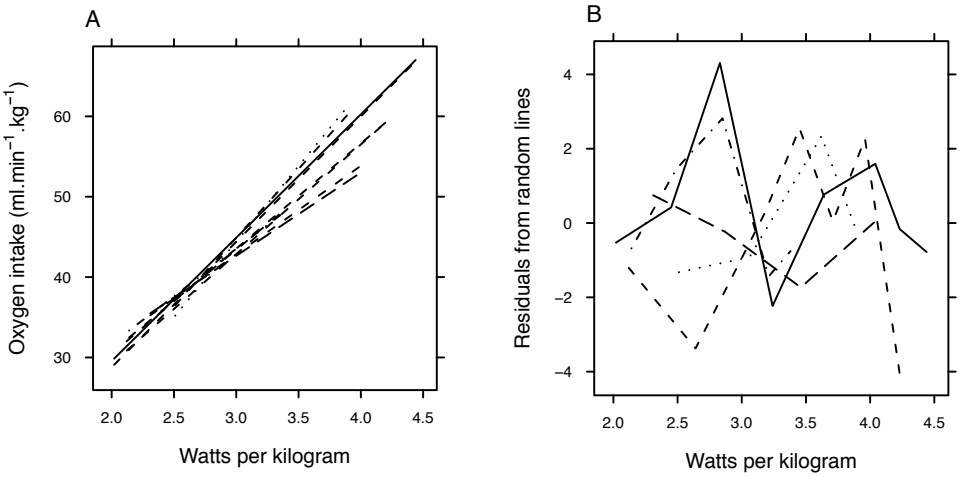

Figure 10.9: In panel A lines have been fitted for each individual athlete, as in Figure 10.8. Also shown, as dashed lines, are the fitted lines from the random lines model. Panel B shows the profiles of residuals from the random lines.

```
panelfun <-
  function(x, y, subscripts, groups, ...){
           panel.superpose(x, hat, subscripts, groups, type="l",lty=2)
           panel.superpose(x, lmhat, subscripts, groups, type="l",lty=1)
}
xyplot(o2 ~ wattsPerKg, groups=id, data=humanpower1, panel=panelfun,
       xlab="Watts",
       ylab=expression("Oxygen intake ("*ml.min^{-1}*"."*kg^{-1}*")"))
```

Figure 10.9B is a plot of residuals, with the points for each individual athlete connected with broken lines.[14]There is nothing in these residual profiles that obviously calls for attention. For example, none of the athletes shows exceptionally large departures, at one or more points, from the linear trend.

The standard errors relate to the accuracy of prediction of the mean response line for the population from which the athletes were sampled. The slopes are drawn from a distribution with estimated mean 13.9 and standard error $\sqrt{1.36^2 + 2.67^2} = 3.0$. This standard deviation may be compared with the standard deviation (= 3.28) of the five slopes that were fitted to the initial fixed effects model.[15]

Standard errors for between athletes components of variation relate to the particular population from which the five athletes were sampled. Almost certainly, the pattern of variation would be different for five people who were drawn at random from a population of recreational sportspeople.

[14]
```
## Plot of residuals
xyplot(resid(hp.lmer) ~ wattsPerKg, groups=id, type="b", data=humanpower1)
```
[15]
```
## Derive the sd from the data frame coefs that was calculated above
sd(coefs$Slope)
```

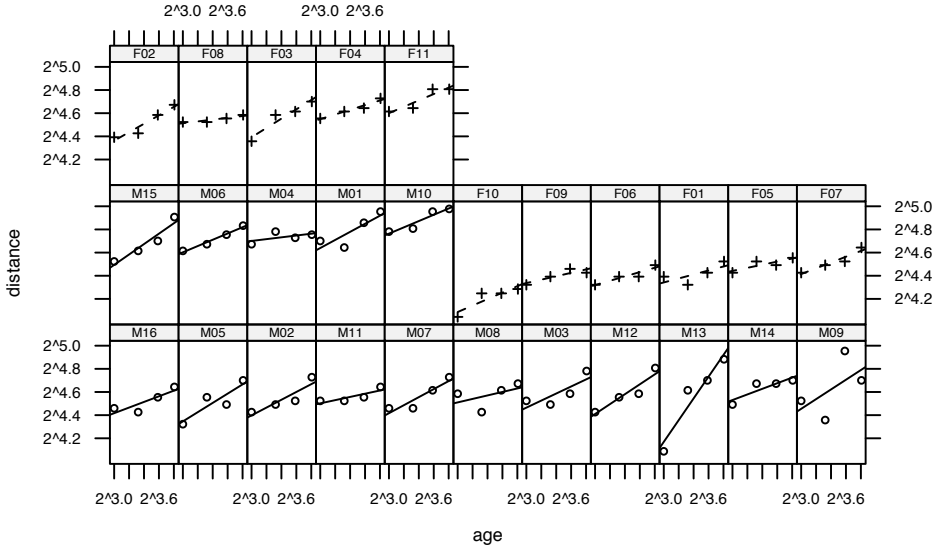

Figure 10.10: Distance between two positions on the skull on a scale of $\log_2$, plotted against age, for each of 27 children.

In this example, the mean response pattern was assumed linear, with random changes, for each individual athlete, in the slope. More generally, the mean response pattern will be nonlinear, and random departures from this pattern may be non-linear.

10.7.2 Orthodontic measurements on chlldren

The `Orthodont` data frame (*MEMSS* package) has measurements on the distance between two positions on the skull, taken every two years from age 8 until age 14, on 16 males and 11 females. Is there a difference in the pattern of growth between males and females?

Preliminary data exploration

Figure 10.10 shows the pattern of change for each of the 25 individuals. Lines have been added; overall the pattern of growth seems close to linear.[16]

A good summary of these data are the intercepts and slopes, as in Figure 10.11. We calculate these both with untransformed distances (panel A) and with distances on a logarithmic scale (Panel B): Here is the code:

```
## Use lmList() to find the slopes
ab <- coef(lmList(distance ~ age | Subject, Orthodont))
names(ab) <- c("a", "b")
## Obtain the intercept at x=mean(x)
## (For each subject, this is independent of the slope)
```

[16]
```
## Plot showing pattern of change for each of the 25 individuals
library(MEMSS)
xyplot(distance ~ age | Subject, groups=Sex, data=Orthodont,
       scales=list(y=list(log=2)), type=c("p","r"), layout=c(11,3))
```

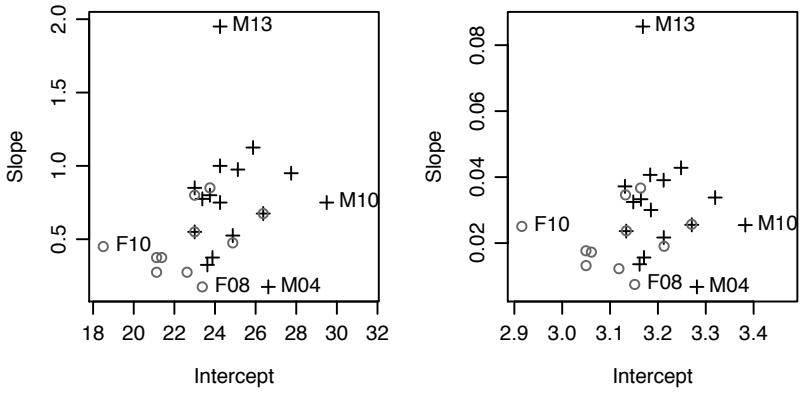

Figure 10.11: Slopes of profiles, plotted against intercepts at age = 11. Females are shown with open circles, and males with +'s. Panel A is for distances, and Panel B is for logarithms of distances.

```
ab$ybar <- ab$a + ab$b*11  # mean age is 11, for each subject.
sex <- substring(rownames(ab), 1 ,1)
plot(ab[, 3], ab[, 2], col=c(F="gray40", M="black")[sex],
     pch=c(F=1, M=3)[sex], xlab="Intercept", ylab="Slope")
extremes <- ab$ybar %in% range(ab$ybar) |
            ab$b %in% range(ab$b[sex=="M"]) |
            ab$b %in% range(ab$b[sex=="F"])
text(ab[extremes, 3], ab[extremes, 2], rownames(ab)[extremes], pos=4, xpd=TRUE)
## The following makes clear M13's difference from other points
qqnorm(ab$b)
Orthodont$logdist <- log(Orthodont$distance)
## Now repeat, with logdist replacing distance
```

The intercepts for the males are clearly different from the intercepts for females, as can be verified by a *t*-test. One slope appears an outlier from the main body of the data. Hence, we omit the largest (M13) and (to make the comparison fair) the smallest (M04) values from the sample of male slopes, before doing a *t*-test. On the argument that the interest is in relative changes, we will work with logarithms of distances.[17]

The output is

```
        Two Sample t-test

data:  b[sex == "F"] and b[sex == "M" & !extreme.males]
t = -2.32, df = 23, p-value = 0.02957
```

[17]
```
## Compare males slopes with female slopes
Orthodont$logdist <- log(Orthodont$distance)
ablog <- coef(lmList(logdist ~ age | Subject, Orthodont))
names(ablog) <- c("a", "b")
## Obtain the intercept at mean age (= 11), for each subject
## (For each subject, this is independent of the slope)
ablog$ybar <- with(ablog, a + b*11)
extreme.males <- rownames(ablog) %in% c("M04","M13")
sex <- substring(rownames(ab), 1, 1)
with(ablog,
  t.test(b[sex=="F"], b[sex=="M" & !extreme.males], var.equal=TRUE))
  # Specify var.equal=TRUE, to allow comparison with anova output
```

```
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.016053 -0.000919
sample estimates:
mean of x mean of y
   0.0211    0.0296
```

The higher average slope for males is greater than can comfortably be attributed to statistical error.

A random coefficients model

Now consider a random coefficients model. The model will allow different slopes for males and females, with the slope for individual children varying randomly about the slope for their sex. We will omit the same two males as before:

```
> keep <- !(Orthodont$Subject%in%c("M04","M13"))
> orthdiff.lmer <- lmer(logdist ~ Sex * I(age-11) + (I(age-11) | Subject),
+                       data=Orthodont, subset=keep, REML=FALSE)
> print(summary(orthdiff.lmer), digits=3)
Linear mixed model fit by maximum likelihood  ['lmerMod']
. . . .
Scaled residuals:
   Min     1Q Median     3Q    Max
-3.370 -0.482  0.004  0.534  3.993

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 5.79e-03 0.076124
          I(age - 11) 7.71e-07 0.000878 1.00
 Residual             2.31e-03 0.048109
Number of obs: 100, groups:  Subject, 25

Fixed effects:
                    Estimate Std. Error t value
(Intercept)          3.11451    0.02407   129.4
SexMale              0.09443    0.03217     2.9
I(age - 11)          0.02115    0.00325     6.5
SexMale:I(age - 11)  0.00849    0.00435     2.0

Correlation of Fixed Effects:
            (Intr) SexMal I(-11)
SexMale     -0.748
I(age - 11)  0.078 -0.058
SxMl:I(-11) -0.058  0.078 -0.748
```

The t-statistic for comparing the 14 male slopes with the 11 female slopes is 2.0, with 23 (14-1+11-1) degrees of freedom. A formal significance test gives a p-value of 0.057.[18]

[18] `2*pt(-2, 23)`. Here this is effectively, on the assumption of equal variances for the two sexes, an exact test that treats the slopes as summary statistics.

An alternative, in general (but not here) more trustworthy, is a likelihood ratio test:

```
> orthsame.lmer <- lmer(logdist ~ Sex + I(age-11) + (I(age-11) | Subject),
+                       data=Orthodont, subset=keep, REML=FALSE)
> print(anova(orthsame.lmer, orthdiff.lmer)[2, "Pr(>Chisq)"], digits=3)
[1] 0.054
```

There is a weak suggestion, not quite at the commonly used 5% level of statistical significance, that the slopes are different.

The estimates of fixed effects from the REML model are in general, because less biased, preferable to those from the full maximum likelihood (ML) model. Here are the estimates for the REML model, with different slopes for the two sexes:

```
> orthdiffr.lmer <- update(orthdiff.lmer, REML=TRUE)
> print(summary(orthdiffr.lmer), digits=3)
Linear mixed model fit by REML ['lmerMod']
Formula: logdist ~ Sex * I(age - 11) + (I(age - 11) | Subject)
   Data: Orthodont
 Subset: keep
. . . .
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 6.33e-03 0.079581
          I(age - 11) 8.42e-07 0.000918 1.00
 Residual             2.38e-03 0.048764
Number of obs: 100, groups:  Subject, 25

Fixed effects:
                    Estimate Std. Error t value
(Intercept)          3.11451    0.02510   124.1
SexMale              0.09443    0.03354     2.8
I(age - 11)          0.02115    0.00330     6.4
SexMale:I(age - 11)  0.00849    0.00441     1.9

Correlation of Fixed Effects:
            (Intr) SexMal I(-11)
SexMale     -0.748
I(age - 11)  0.080 -0.060
SxMl:I(-11) -0.060  0.080 -0.748
```

The estimate 8.42×10^{-7} of the slope component of variance is for all practical purposes zero. The variation in the slope of lines is entirely explained by variation of individual points about lines, within and between subjects of the same sex. The use of 23 degrees of freedom for the *t*-test for comparing slopes may thus be overly conservative.

Correlation between successive times

We can calculate the autocorrelations across each subject separately, and check the distribution. The interest is in whether any autocorrelation is consistent across subjects.

```
> res <- resid(orthdiffr.lmer)
> Subject <- factor(Orthodont$Subject[keep])
> orth.acf <- t(sapply(split(res, Subject),
+                      function(x)acf(x, lag=4, plot=FALSE)$acf))
> ## Calculate respective proportions of Subjects for which
> ## autocorrelations at lags 1, 2 and 3 are greater than zero.
> apply(orth.acf[,-1], 2, function(x)sum(x>0)/length(x))
[1] 0.20 0.24 0.40
```

Thus a test for a zero lag 1 autocorrelation has $p = 0.20$. The suggestion of non-zero autocorrelation is very weakly supported.

**The variance for the difference in slopes*

This can be calculated from the components of variance information. The sum of squares about the mean, for one line, is $\sum(x-\bar{x})^2 = 20$. The sum of the two components of variance for an individual line is then: $1.19\times10^{-12} + 0.002383/20 = 0.00011915$. The standard error of the difference in slopes is then:

$$\sqrt{0.00011915(1/14+1/11)} = 0.00440$$

Compare this with the value given against the fixed effect `SexMale:I(age - 11)` in the output above. The numbers are, to within rounding error, the same. Degrees of freedom for the comparison are 23 as for the *t*-test.

10.8 Further Notes on Multi-level and Other Models with Correlated Errors

10.8.1 Different sources of variance – complication or focus of interest?

In the discussion of multi-level models, the main interest was in the parameter estimates. The different sources of variance, were a complication. In other applications, the variances may be the focus of interest. Many animal and plant breeding trials are of this type. The aim may be to design a breeding program that will lead to an improved variety or breed. Where there is substantial genetic variability, breeding experiments have a good chance of creating improved varieties.

Investigations into the genetic component of human intelligence have generated fierce debate. Most such studies have used data from identical twins who have been adopted out to different homes, comparing them with non-identical twins and with sibs who have been similarly adopted out. The adopting homes rarely span a large part of a range from extreme social deprivation to social privilege, so that results from such studies may have little or no relevance to investigation of the effects of extreme social deprivation. The discussion in Leavitt and Dubner (2005, chapter 5) sheds interesting light on these these effects.

There has not been, until recently, proper allowance for the substantial effects that arise from simultaneous or sequential occupancy of the maternal womb (Bartholemew, 2004; Daniels et al., 1997). Simple forms of components of variance model are unable to account for the Flynn effect (Bartholemew, 2004, pp. 138-140), by which measured IQs in many parts of the world have in recent times increased by about 15 IQ points per generation. The

simple model, on which assessments of proportion of variance that is genetic have been based, seems too simplistic to give useful insight.

We have used an analysis of data from a field experimental design to demonstrate the calculation and use of components of variance. Other contexts for multi-level models are the analysis of data from designed surveys, and general regression models in which the "error" term is made up of several components. In all these cases, errors are no longer independently and identically distributed.

10.8.2 Predictions from models with a complex error structure

Here, "complex" refers to models that assume something other than an i.i.d. error structure. Most of the models considered in this chapter can be used for different predictive purposes, and give standard errors for predicted values that differ according to the intended purpose. Accurate modeling of the structure of variation allows, as for the Antiguan corn yield data in Section 10.1), these different inferential uses.

As has been noted, shortcuts are sometimes possible. Thus for using the kiwifruit shading data to predict yields at any level other than the individual vine, there is no loss of information from basing the analysis on plot means.

Consequences from assuming an overly simplistic error structure

In at least some statistical application areas, analyses that assume an overly simplistic error structure (usually, an i.i.d. model) are relatively common in the literature. Inferences may be misleading, or not, depending on how results are used. Where there are multiple levels of variation, all variation that contributes to the sampling error of fixed effects must be modeled correctly. Otherwise, the standard errors of model parameters that appear in computer output will almost inevitably be wrong, and should be ignored.

In data that have appropriate balance, predicted values will ordinarily be unbiased, even if the error structure is not modeled appropriately. The standard errors will almost certainly be wrong, usually optimistic. A good understanding of the structure of variation is typically required in order to make such limited inferences as are available when an overly simplistic error structure is assumed!

10.8.3 An historical perspective on multi-level models

Multi-level models are not new. The inventor of the analysis of variance was R.A. Fisher. Although he would not have described it that way, many of the analysis of variance calculations that he demonstrated were analyses of specific forms of multi-level model. Data have a structure that is imposed by the experimental design. The particular characteristic of the experimental design models that Fisher used was that the analysis could be handled using analysis of variance methods. The variance estimates that are needed for different comparisons may be taken from different lines of the analysis of variance table. This circumvents the need to estimate the variances of the random effects that appear in a fully general analysis.

Until the modern computing era, multi-level data whose structure did not follow one of

the standard designs and thus did not fit the analysis of variance framework required some form of approximate analysis. Such approximate analyses, if they were possible at all, often demanded a high level of skill.

Statistical analysts who used Fisher's experimental designs and methods of analysis followed Fisher's rules, and those of his successors, for the calculations. Each different design had its own recipe. After working through a few such analyses, some of those who followed Fisher's methods began to feel that they understood the rationale fairly well at an intuitive level. Books appeared that gave instructions on how to do the analyses. The most comprehensive of these is Cochran and Cox (1957).

The Genstat system (Payne et al., 1997) was the first of the major systems to implement general methods for the analysis of multi-level models that had a suitable "balance". Its coherent and highly structured approach to the analysis of data from suitably balanced designs takes advantage of the balance to simplify and structure the output.

General purpose software for use with unbalanced data, in the style of the *lme4* package, made its appearance relatively recently. The analyses that resulted from earlier ad hoc approaches were in general less insightful and informative than the more adequate analyses that are available within a multi-level modeling framework.

Regression models are another starting point for consideration of multi-level models. Both the fixed effects parts of the model have a structure, thus moving beyond the models with a single random (or "error") term that have been the stock in trade of courses on regression modeling. Even now, most regression texts give scant recognition of the implications of structure in the random part of the model. Yet data commonly do have structure – students within classes within institutions, nursing staff within hospitals within regions, managers within local organizations within regional groupings, and so on.

As has been noted, models have not always been written down. Once theoretical statisticians did start to write down models, there was a preoccupation with models that had a single error term. Theoretical development, where the discussion centered around models, was disconnected from the practical analysis of experimental designs, where most analysts were content to follow Cochran and Cox and avoid formal mathematical description of the models that underpinned their analyses.

Observational data that have a multilevel structure, which is typically unbalanced, can nowadays be analyzed just as easily as experimental data. It is no longer necessary to look up Cochran and Cox to find how to do an analysis. There are often acceptable alternatives to Cochran and Cox style experimental designs.

Problems of interpretation and meaningfulness remain, for observational data, as difficult as ever. The power of modern software can become a trap. There may be inadequate care in the design of data collection, in the expectation that computer software will take care of any problems. The result may be data whose results are hard to interpret or cannot be interpreted at all, or that make poor use of resources.

10.8.4 Meta-analysis

Meta-analysis is a name for analyses that bring together into a single analysis framework data from, for example, multiple agricultural trials, or from multiple clinical trials, or from multiple psychological laboratories. Multi-level modeling, and extensions of multi-level modeling such as repeated measures analysis, make it possible to do analyses that take

proper account of site to site or center to center or study to study variation. If treatment or other effects are consistent relative to all identifiable major sources of variation, the result can be compelling for practical application.

Meta-analysis is uncontroversial when data are from a carefully planned multi-location trial. More controversial is the bringing together into one analysis of data from quite separate investigations. There may however be little choice; the alternative may be an informal and perhaps unconvincing qualitative evaluation of the total body of evidence. Clearly such analyses challenge the critical acumen of the analyst. A wide range of methodologies have been developed to handle the problems that may arise. Gaver et al. (1992) is a useful summary. Turner et al. (2009) is an interesting and comprehensive state-of-the-art account.

10.8.5 Functional data analysis

Much of the art of repeated measures modeling lies in finding suitable representations, requiring a small number of parameters, both of the individual profiles and of variation between those profiles. Spline curves are widely used in this context. Chapter 12 will discuss the use of principal components to give a low-dimensional representation of multivariate data. A similar methodology can be used to find representations of curves in terms of a small number of basis functions. Further details are in Ramsay and Silverman (2002).

10.8.6 Error structure in explanatory variables

This chapter has discussed error structure in response variables. There may also be a structure to error in explanatory variables. Studies of the health effects of dietary components, such as were described in Subsection 6.8, provide an interesting and important example, with major implications for the design of such studies.

10.9 Recap

Multi-level models account for multiple levels of random variation. The random part of the model possesses structure; it is a sum of distinct random components.

In making predictions based on multi-level models, it is necessary to identify precisely the population to which the predictions will apply.

The art in setting up an analysis for these models is in getting the description of the model correct. Specifically it is necessary to

- identify which are fixed and which random effects,
- correctly specify the nesting of the random effects.

In repeated measures designs, it is necessary to specify or otherwise model the pattern of correlation within profiles.

A further generalization is to the modeling of random coefficients, for example, regression lines that vary between different subsets of the data.

Skill and care may be needed to get output into a form that directly addresses the questions that are of interest. Finally, output must be interpreted. Multi-level analyses often require high levels of professional skill.

10.10 Further Reading

Fisher (1935) is a non-mathematical account that takes the reader step by step through the analysis of important types of experimental design. It is useful background for reading more modern accounts. Williams et al. (2002) is similarly example-based, with an emphasis on tree breeding. See also Cox (1958); Cox and Reid (2000). Cox and Reid is an authoritative up to date account of the area, with a more practical focus than its title might seem to imply. On multi-level and repeated measures models see Gelman and Hill (2007); Snijders and Bosker (1999); Diggle et al. (2002); Goldstein (1995); Pinheiro and Bates (2000); Venables and Ripley (2002).

Talbot (1984) is an interesting example of the use of multi-level modeling, with important agricultural and economic implications. It summarizes a large amount of information that is of importance to farmers, on yields for many different crops in the UK, including assessments both of center to center and of year to year variation.

The relevant chapters in Payne et al. (1997), while directed to users of the Genstat system, have helpful commentary on the use of the methodology and on the interpretation of results. Pinheiro and Bates (2000) describes the use of the *nlme* package for handling multi-level analyses.

On meta-analysis see Chalmers and Altman (1995); Gaver et al. (1992); Turner et al. (2009).

References for further reading

Analysis of variance with multiple error terms

Cochran, W.G. and Cox, G.M. 1957. *Experimental Designs*, 2nd edn.

Cox, D.R. 1958. *Planning of Experiments.*

Cox, D.R. and Reid, N. 2000. *Theory of the Design of Experiments.*

Fisher, R.A. 1935 (7th edn. 1960). *The Design of Experiments.*

Payne, R.W., Lane, P.W., Digby, P.G.N., Harding, S.A., Leech, P.K., Morgan, G.W., Todd, A.D., Thompson, R., Tunnicliffe Wilson, G., Welham, S.J. and White, R.P. 1997. *Genstat 5 Release 3 Reference Manual*.

Williams, E.R., Matheson, A.C. and Harwood, C.E. 2002. *Experimental Design and Analysis for Use in Tree Improvement*, revised edn.

Multi-level models and repeated measures

Diggle, P.J., Heagerty, P.J., Liang, K.-Y. and Zeger, S.L. 2002. *Analysis of Longitudinal Data*, 2nd edn.

Gelman, A. and Hill, J., 2007. Data analysis using regression and multilevel/hierarchical models.

Goldstein, H. 1995. *Multi-level Statistical Models.*

Payne, R.W., Lane, P.W., Digby, P.G.N., Harding, S.A., Leech, P.K., Morgan, G.W., Todd, A.D., Thompson, R., Tunnicliffe Wilson, G., Welham, S.J. and White, R.P. 1993. *Genstat 5 Release 3 Reference Manual*.

Pinheiro, J.C. and Bates, D.M. 2000. *Mixed Effects Models in* S *and* S-PLUS.

Snijders, T.A.B. and Bosker, R.J. 1999. *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modelling*.

Talbot, M. 1984. Yield variability of crop varieties in the UK. *Journal of the Agricultural Society of Cambridge* 102: 315–321. *JRSS* A 172: 21-47.

Venables, W.N. and Ripley, B.D. 2002. *Modern Applied Statistics with* S, 4th edn.

Meta-analysis

Chalmers, I. and Altman, D.G. 1995. *Systematic Reviews*.

Gaver, D.P., Draper, D.P., Goel, K.P., Greenhouse, J.B., Hedges, L.V., Morris, C.N. and Waternaux, C. 1992. *Combining Information: Statistical Issues and Opportunities for Research*.

Turner et al. 2009. Bias modelling in evidence synthesis.

10.11 Exercises

1. Repeat the calculations of Subsection 10.4.5, but omitting results from two vines at random. Here is code that will handle the calculation:

```
n.omit <- 2
take <- rep(TRUE, 48)
take[sample(1:48,2)] <- FALSE
kiwishade.lmer <- lmer(yield ~ shade + (1|block) + (1|block:plot),
                       data = kiwishade,subset=take)
vcov <- VarCorr(kiwishade.lmer)
print(vcov, comp="Variance")
```

   Repeat this calculation five times, for each of `n.omit` = 2, 4, 6, 8, 10, 12 and 14. Plot (i) the plot component of variance and (ii) the vine component of variance, against number of points omitted. Based on these results, for what value of `n.omit` does the loss of vines begin to compromise results? Which of the two components of variance estimates is more damaged by the loss of observations? Comment on why this is to be expected.

2. Repeat the previous exercise, but now omitting 1, 2, 3, 4 complete plots at random.

3. The data set `Gun` (*MEMSS* package) reports on the numbers of rounds fired per minute, by each of nine teams of gunners, each tested twice using each of two methods. In the nine teams, three were made of men with slight build, three with average, and three with heavy build. Is there a detectable difference, in number of rounds fired, between build type or between firing methods? For improving the precision of results, which would be better – to double the number of teams, or to double the number of occasions (from 2 to 4) on which each team tests each method?

4. *The data set `ergoStool` (*MEMSS* package) has data on the amount of effort needed to get up from a stool, for each of nine individuals who each tried four different types of stool. Analyze the data both using `aov()` and using `lme()`, and reconcile the two sets of output. Was there any clear winner among the types of stool, if the aim is to keep effort to a minimum?

5. *In the data set `MathAchieve` (*MEMSS* package), the factors `Minority` (levels `yes` and `no`) and `sex`, and the variable `SES` (socio-economic status) are clearly fixed effects. Discuss how the decision whether to treat `School` as a fixed or as a random effect might depend on the purpose of the study? Carry out an analysis that treats `School` as a random effect. Are differences between

schools greater than can be explained by within school variation?

6. *The data frame `sorption` (*DAAG*) includes columns `ct` (concentration-time sum), `Cultivar` (apple cultivar), `Dose` (injected dose of methyl bromide), and `rep` (replicate number, within `Cultivar` and `year`). Fit a model that allows the slope of the regression of `ct` on `Dose` to be different for different cultivars and years. and to vary randomly with replicate. Consider the two models:

```
cult.lmer <- lmer(ct ~ Cultivar + Dose + factor(year) +
                  (-1 + Dose | gp), data = sorption,
                  REML=TRUE)
cultdose.lmer <- lmer(ct ~ Cultivar/Dose + factor(year) +
                      (-1 + Dose | gp), data = sorption,
                      REML=TRUE)
```

Explain (i) the role of the each of the terms in these models, and (ii) how the two models differ. Which model seems preferable? Write a brief commentary on the output from the preferred model.

[NB: The factor `gp`, which has a different level for each different combination of `Cultivar`, `year` and replicate, associates a different random effect with each such combination.]