Part VIII
Multi-level Models

For background, see SPDM: Sections 14-15.

1 Description and Display of the Data

1.1 Description

This laboratory will work with data on corn yields from the Caribbean islands of Antigua and St Vincent. Data are yields from packages on eight sites on the Caribbean island of Antigua. They are a summarized version of a subset of data given in Andrews and Herzberg 1985, pp.$\tilde{3}$39-353. The data frames `ant111b` and `vince111b` hold yields for the standard treatment, here identified as `111`, for sites on Antigua and St Vincent respectively. Additionally, there will be some use of the more extensive data in the data frame `antigua`. All three data frames are included in recent versions (≥ 0.84) of the *DAAG* package.

The data frame `ant111b` has data for n=4 packages of land at each of eight sites, while `vince111b` data for four packages at each of nine sites. As will be described below, two possible predictions are:

(a) Predictions for new packages of land in one of the existing sites.

(b) Predictions for new packages in a new site.

The accuracies for the second type of prediction may be much less accurate than for the first type. A major purpose of this laboratory is to show how such differences in accuracy can be modeled.

1.2 Display

We begin by examining plots, for the treatment `111`, from the combined data for the two islands. This information for the separate islands is summarized in the datasets `ant111b` and `vince111b` in the *DAAG* package.

A first step is to combine common columns of `ant111b` and `vince111b` into the single data frame `corn111b`.

```
> library(lattice)
> library(DAAG)
> corn111b <- rbind(ant111b[, -8], vince111b)
> corn111b$island <- c("Antigua", "StVincent")[corn111b$island]
```

- The following plot uses different panels for the two islands:

  ```
  > corn.strip1 <- stripplot(site ~ harvwt | island, data = corn111b,
  +     xlab = "Harvest weight")
  ```

- The following plot uses different panels for the two islands, but allows separate (”free” = no relation) vertical scales for the two plots.

  ```
  > corn.strip2 <- stripplot(site ~ harvwt | island, data = corn111b,
  +     xlab = "Harvest weight", scale = list(y = list(relation = "free")))
  ```

- The following uses a single panel, but uses different colours (or, on a black and white device, different symbols) to distinguish the two islands. Notice the use of `auto.key` to generate an automatic key:

  ```
  > corn.strip3 <- stripplot(site ~ harvwt, data = corn111b, groups = island,
  +     xlab = "Harvest weight", auto.key = list(columns = 2))
  ```

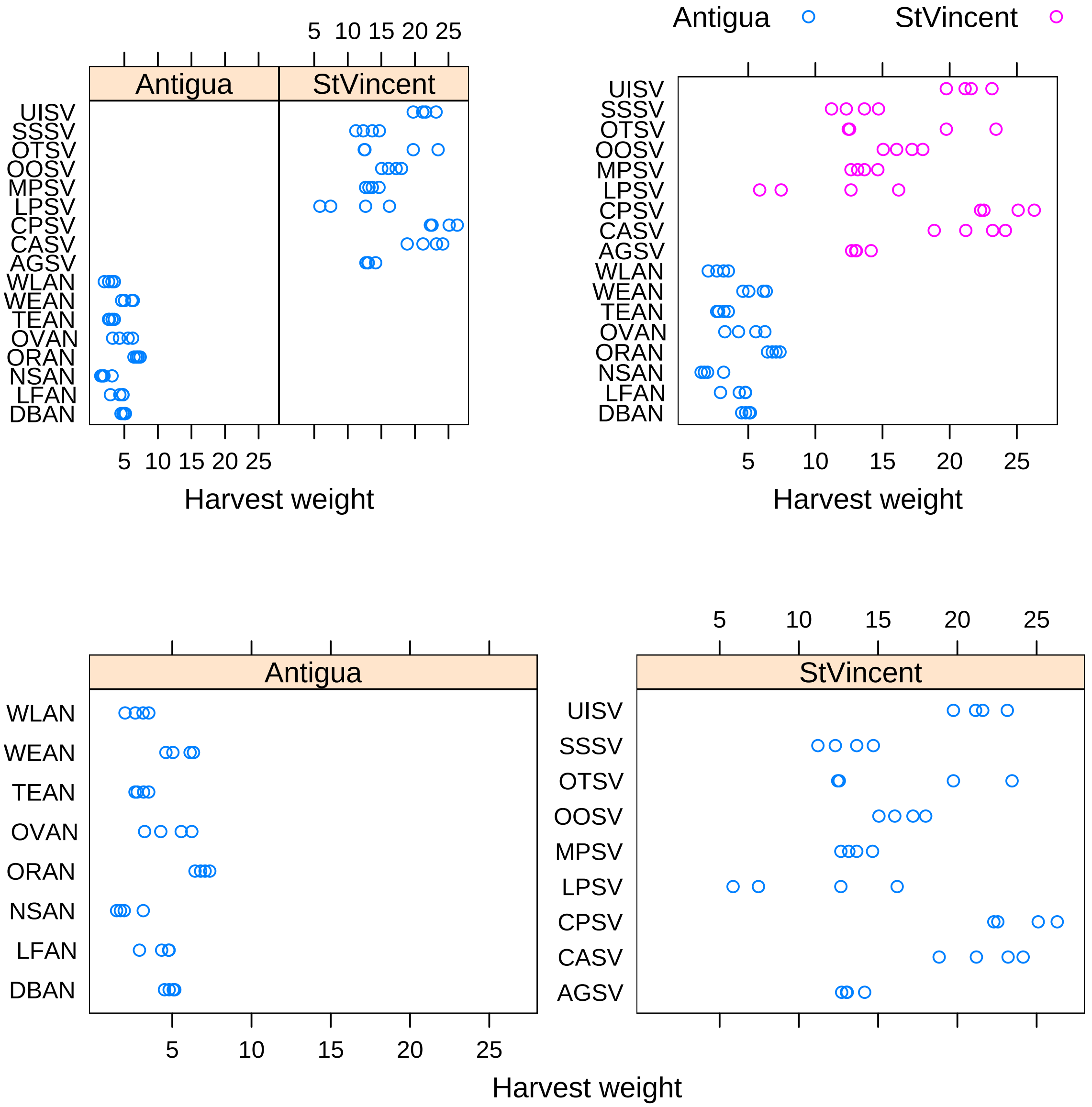

Figure 3: Yields for the four packages of corn on sites on the islands of Antigua and St Vincent.

Next, we will obtain package means for the Antiguan data, for all treatments.

```
> with(antigua, antp <<- aggregate(harvwt, by = list(site = site,
+     package = block, trt = trt), FUN = mean))
> names(antp)[4] <- "harvwt"
```

Notice the use of the version `<<-` of the assignment symbol to ensure that assignment takes place in the workspace.

Now plot mean harvest weights for each treatment, by site:

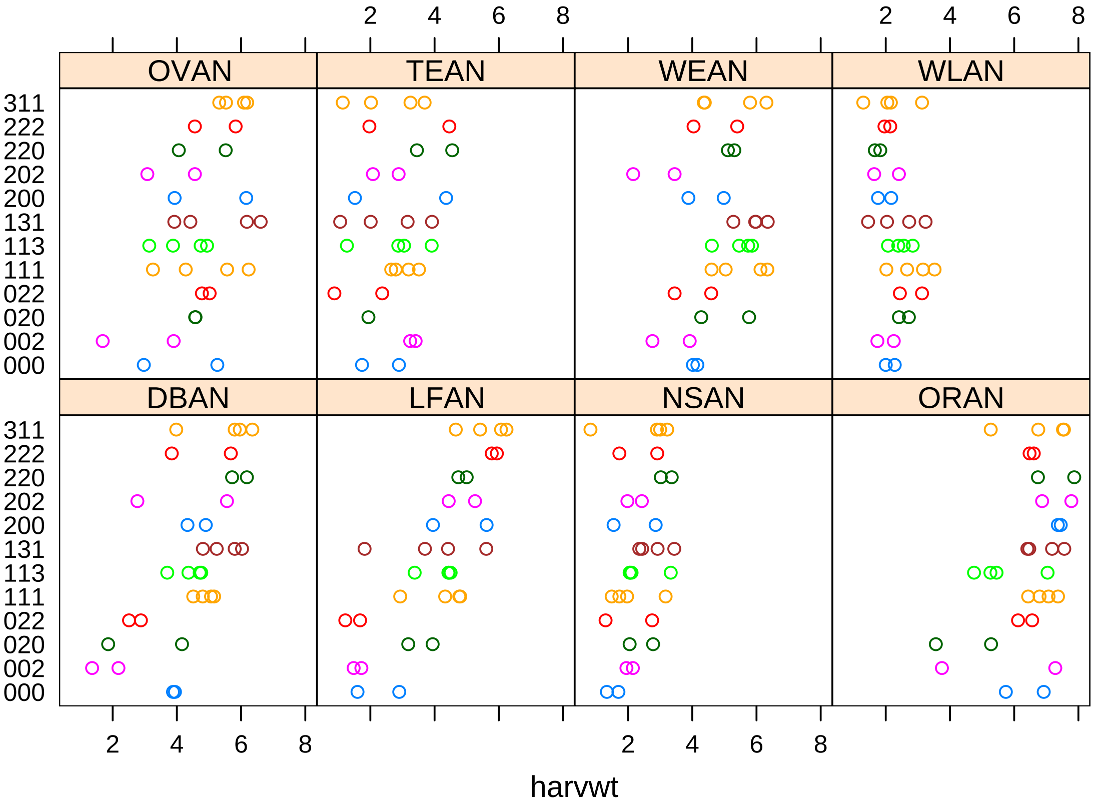

Figure 4: Yields for the four packages of corn on each of eight sites on the island of Antigua.

Questions and Exercises

(a) Which set of sites (Antigua or St Vincent) shows the largest yields?

(b) Create a plot that compares the logarithms of the yields, within and between sites on the two islands. From this plot, what, if anything, can you say about the differet variabilities in yield, within and between sites on each island?

2 Multi-level Modeling – the Antiguan Data

***Analysis using *lme*:** The modeling command takes the form:

```
> library(nlme)
> ant111b.lme <- lme(fixed = harvwt ~ 1, random = ~1 | site, data = ant111b)
```

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

```
> options(digits = 4)
> ant111b.lme
```

```
Linear mixed-effects model fit by REML
  Data: ant111b
  Log-restricted-likelihood: -47.21
  Fixed: harvwt ~ 1
(Intercept)
      4.292

Random effects:
 Formula: ~1 | site
        (Intercept) Residual
StdDev:       1.539     0.76

Number of Observations: 32
Number of Groups: 8
```

Notice that *lme* gives, not the components of variance, but the standard deviations (`StdDev`) which are their square roots. Observe that, according to *lme*, $\widehat{\sigma_B^2} = 0.76^2 = 0.578$, and $\widehat{\sigma_L^2} = 1.539^2 = 2.369$. The variance for an individual package is $\widehat{\sigma_B^2} + \widehat{\sigma_L^2}$.

Those who are familiar with an analysis of variance table for such data should note that *lme* does not give the mean square at any level higher than level 0, not even in this balanced case.

Note that the yields are not independent between different packages on the same site, in the population that has packages from multiple sites. (Conditional on coming from one particular site, package yields are however independent.)

The take-home message from this analysis is:

- For prediction for a new package at one of the existing sites, the standard error is 0.76
- For prediction for a new package at a new site, the standard error is $\sqrt{1.539^2 + .76^2} = 1.72$
- For prediction of the mean of n packages at a new site, the standard error is $\sqrt{1.539^2 + 0.76^2/n}$. This is NOT inversely proportional to n, as would happen if the yields were independent within sites.

Where there are multiple levels of variation, the predictive accuracy can be dramatically different, depending on what is to be predicted. Similar issues are arise in repeated measures contexts, and in time series. Repeated measures data has multiple profiles, i.e., many small time series.

2.1 Simulation

The following function simulates results from a multilevel model for the case where there are `npackages` packages at each of `nplots` plots.

```
> "simMlevel" <- function(nsites = 8, npackages = 4, mu = 4, sigmaL = 1.54,
+     sigmaB = 0.76) {
+     facSites <- factor(1:nsites)
+     facPackages <- factor(1:npackages)
+     dframe <- expand.grid(facPackages = facPackages, facSites = facSites)
+     nall <- nsites * npackages
+     siteEffects <- rnorm(nsites, 0, sigmaL)
+     err <- rnorm(nall, 0, sigmaB) + siteEffects[unclass(dframe$facSites)]
+     dframe$yield <- mu + err
+     dframe
+ }
```

The default arguments are `sigmaB` = 0.76 and `sigma` = 1.54, as for the Antiguan data.

2.2 Questions and Exercises

(a) Repeat the analysis

(a) for the Antiguan data, now using a logarithmic scale.

(b) for the St Vincent data, using a logarithmic scale.

(b) Overlay plots, for each of the two islands, that show how the variance of the mean can be expected to change with the number of packages n.

(c) Are there evident differences between islands in the contributions of the two components of variance? What are the practical implications that flow from such differences as you may observe?

(d) Use the function `simMlevel()` to simulate a new set of data, using the default arguments. Analyse the simulated data. Repeat this exercise 25 or more times. How closely do you reproduce the values of `sigmaL`=1.54 and `sigmaB`=0.76 that were used for the simulation?

3 Multi-level Modeling – Attitudes to Science Data

These data are from in the *DAAG* package for R. The data 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.

There are three variance components:

```
Between schools 0.00105
Between classes 0.318
Between students 3.05
```

The between schools component can be neglected. The variance for a class mean is $0.318 + 3.05/n$, where n is the size of the class. The two contributions are about equal when n =10.

4 *Additional Calculations

We return again to the corn yield data.

Is variability between packages similar at all sites?:

```
> if (dev.cur() == 2) invisible(dev.set(3))
> vars <- sapply(split(ant111b$harvwt, ant111b$site), var)
> vars <- vars/mean(vars)
> qqplot(qchisq(ppoints(vars), 3), 3 * vars)
```

Does variation within sites follow a normal distribution?:

```
> qqnorm(residuals(ant111b.lme))
```

What is the pattern of variation between sites?

```
> locmean <- sapply(split(log(ant111b$harvwt), ant111b$site), mean)
> qqnorm(locmean)
```

The distribution seems remarkably close to normal.

Fitted values and residuals in *lme*: By default fitted values account for all random effects, except those at level 0. In the example under discussion `fitted(ant111b.lme)` calculates fitted values at level 1, which can be regarded as estimates of the site means. They are not however the site means, as the graph given by the following calculation demonstrates:

```
> hat.lm <- fitted(lm(harvwt ~ site, data = ant111b))
> hat.lme <- fitted(ant111b.lme)
> plot(hat.lme ~ hat.lm, xlab = "Site means", ylab = "Fitted values (BLUPS) from lme")
> abline(0, 1, col = "red")
```

The fitted values are known as BLUPs (Best Linear Unbiased Predictors). Relative to the site means, they are pulled in toward the overall mean. The most extreme site means will on average, because of random variation, be more extreme than the corresponding "true" means for those sites. There is a theoretical result that gives the factor by which they should be shrunk in towards the true mean.

Residuals are by default the residuals from the package means, i.e., they are residuals from the fitted values at the highest level available. To get fitted values and residuals at level 0, enter:

```
> hat0.lme <- fitted(ant111b.lme, level = 0)
> res0.lme <- resid(ant111b.lme, level = 0)
> plot(res0.lme, ant111b$harvwt - hat0.lme)
```

5 Notes – Other Models for Data with Complex Errors

Time series are another important special case. A first step is, often, to subtract off any trend, and base further analysis on residuals about this trend. Observations that are close together in time are typically more closely correlated than observations that are widely separated in time.

The variances of the mean of n observations with variance σ^2 will, assuming that positive correlation between neighbouring observations makes the major contribution to the correlation structure, be greater than $\frac{\sigma^2}{n}$.

Here is a simple way to generate data that are sequentially correlated. The autocorrelation plot shows how the estimated correlation changes as observations move further apart.

```
> y <- rnorm(200)
> y1 <- y[-1] + 0.5 * y[-length(y)]
> acf(y1)
```

Of course the multiplier in `y1 <- y[-1] + 0.5*y[-1000]` can be any number at all, and more complex correlation structures can be generated by incorporating further lags of `y`.