Random Effects in R

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

March 15, 2007

The Big Picture

- The completely randomized design with a *random effect* assumes the following model.

$$y_{ij} = \mu + a_i + e_{ij}, \quad j = 1, \ldots, n_i, i = 1, \ldots, k,$$

  where $a_i \sim \text{iid} N(0, \sigma_a^2)$ and $e_{ij} \sim \text{iid} N(0, \sigma_e^2)$ with $\{a_i\}$ independent of $\{e_{ij}\}$.
- If the data is perfectly balanced with equal numbers of individuals in each group, ($n_i = n$ for all i), it is possible to make modifications to a fixed effects ANOVA analysis by computing a different expected mean square (EMS) for inference.
- When the data is not balanced, this approach fails — sometimes approximations work, sometimes not.
- An alternative regression approach is appropriate even when the design is imbalanced, but the scope of inference is then different.
- I will first show the regression approach (described in Chapter 10 of your book).

Correlation

- One consequence of the random effects model is that observations can be *correlated*.
- In the model we study here, observations in different groups are independent just as in the fixed effect model, because they do not share any random variables.
- For example, the first observations from groups one, y_{11}, depends a_1 and e_{11} while y_{21} depends on a_2 and e_{21}.
- In contrast, observations in the same group are correlated.
- y_{11} and y_{12} each depend on a_1.

Variance

- Recall from Statistics 571 the following facts about variances of independent random variables X and Y and constant c.
  1. $\mathrm{V}(X+Y) = \mathrm{V}(X-Y) = \mathrm{V}(X) + \mathrm{V}(Y)$ (if independent).
  2. $\mathrm{V}(cX) = c^2\mathrm{V}(X)$.
- With this reminder, we can recognize that σ_a measures the size of a typical difference between a random effect and the grand mean.
- A typical difference between two regression effects is $\sqrt{\mathrm{V}(a_i - a_j)} = \sqrt{\mathrm{V}(a_i) + \mathrm{V}(a_j)} = \sqrt{2}\sigma_a$.

Linear Mixed Effects Models using lmer

- The most recently developed R package for fitting linear models with random effects is in the library `lme4`.
- The function to use instead of `lm` is named `lmer`.
- Mac users will need a recent operating system, OS X 10.4.4 or later and R 2.4.1 or later to use this library.
- Users of older Macs can use the older package `nlme` and the function `lme`.
- `lme` cannot fit as rich a class of random effects models as `lmer` (for example, random effects cannot be nested and you cannot use generalized linear models), but it will suffice for much of what we do in the course.
- A model formula with a random effect in `lmer` differs from `lm` by including a term of the form `(a | b)` where `a` is a model matrix (often the intercept 1) for the scope of the random effect and `b` is the group to which the random effect applies.

Data Description

- We will consider the data set `ant111b` from the textbook.
- This data set is a subset of a much larger data set on corn yields on different islands in the Carribean.
- In the subset of data, two possibly interesting response variables are harvest weight (`harvwt`) and the number of ears (`ears`).
- Both `harvwt` and `ears` are measured by plot, but the units are unspecified.
- There are four plots within each site. The variable `plot` is only meaningful nested within a site.
- We will consider the site as a random effect as there are many possible sites on each island.
- In the subset of the data we have, there is only a single treatment and all sites are on the same island.

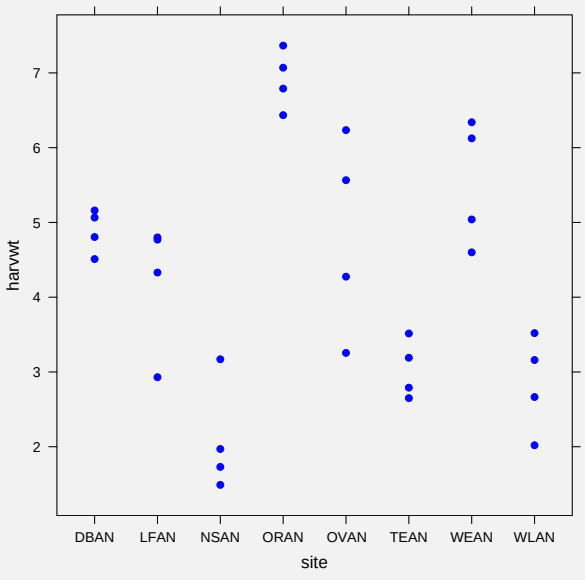
Data

```
> library(DAAG)
> str(ant111b)
'data.frame': 32 obs. of  9 variables:
 $ site  : Factor w/ 8 levels "DBAN","LFAN",..: 1 2 3 4 5 6 7 8 1 2 ..
 $ parcel: Factor w/ 4 levels "I","II","III",..: 1 1 1 1 1 1 1 1 2 2 .
 $ code  : num  58 58 58 58 58 58 58 58 58 58 ...
 $ island: num  1 1 1 1 1 1 1 1 1 1 ...
 $ id    : num    3  40 186 256 220 ...
 $ plot  : num  3 4 5.5 4.5 3.5 5 7 7 15.5 15 ...
 $ trt   : num  111 111 111 111 111 111 111 111 111 111 ...
 $ ears  : num  43.5 40.5 20 42.5 31.5 32.5 43.5 50 46 46.5 ...
 $ harvwt: num  5.16 2.93 1.73 6.79 3.25 ...
```


Plot of Data

```
> attach(ant111b)
> library(lattice)
> fig1 = xyplot(harvwt ~ site, pch = 16, col = "blue")
> print(fig1)
```

- Notice that there is considerable variation among sites.
- Variation within sites (from plot to plot) appears smaller.



Fitting a Random Effects Model

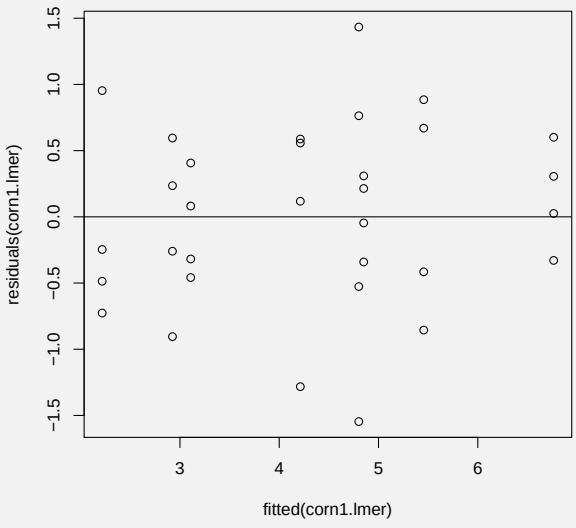
```
> library(lme4)
> corn1.lmer = lmer(harvwt ~ 1 + (1 | site))
> summary(corn1.lmer)

Linear mixed-effects model fit by REML
Formula: harvwt ~ 1 + (1 | site)
   AIC   BIC logLik MLdeviance REMLdeviance
 98.42 101.3 -47.21      95.08        94.42
Random effects:
 Groups   Name        Variance Std.Dev.
 site     (Intercept) 2.36773  1.53874
 Residual             0.57754  0.75996
number of obs: 32, groups: site, 8

Fixed effects:
            Estimate Std. Error t value
(Intercept)   4.2917     0.5604   7.659

> sig2site = as.vector(VarCorr(corn1.lmer)$site)
> sigma = attributes(summary(corn1.lmer))$sigma
```

- The first 1 is the fixed effect.
- The term `(1 | site)` means that there is a random effect for each `site` and this effect is nested within the intercept (the whole model).
- There are two sources of random variation, one for site and one for parcel within site.
- The estimated variance components are $\sigma_a^2 = 2.36773$ and $\sigma_e^2 = 0.57754$.

Residual Plot

```
> plot(fitted(corn1.lmer), residuals(corn1.lmer))
> abline(h = 0)
```

- There are no bad patterns in the residual plot.



Fitted Values

```
> means = sapply(split(harvwt, site), mean)
> siteFit = sapply(split(fitted(corn1.lmer), site), mean)
> data.frame(mean = means, fitted = siteFit)

        mean   fitted
DBAN 4.88500 4.850901
LFAN 4.20750 4.212341
NSAN 2.09000 2.216545
ORAN 6.91500 6.764226
OVAN 4.83250 4.801418
TEAN 3.03625 3.108409
WEAN 5.52625 5.455295
WLAN 2.84125 2.924616
```

- Notice that the fitted values are not just the sample means.
- These values are *shrinkage estimates* that are closer to the grand mean than the sample means.

Estimation

- We can use the two different variance components to estimate the error in different kinds of predictions.
- A prediction of the harvest weight of corn from a new parcel in one of the existing sites (assuming other factors were equal) would have standard deviation σ_e.
- A prediction of the harvest weight of corn from a new parcel in a new site would have standard deviation $\sqrt{\sigma_a^2 + \sigma_e^2}$.
- An estimate of the mean harvest weight of m parcels in an existing site would have standard deviation $\sigma_e/\sqrt{m}$.
- An estimate of the mean harvest weight of m parcels at a new site would have standard deviation $\sqrt{\sigma_a^2 + \sigma_e^2/m}$.

Comparison to ANOVA Approach

```
> corn1.aov = aov(harvwt ~ Error(site))
> summary(corn1.aov)

Error: site
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  7 70.339  10.048

Error: Within
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals 24 13.8609  0.5775
```

- For a *balanced* design, this analysis is acceptable.
- For this specific design, the expected mean squares (EMS) are σ_e^2 for MSErr and $\sigma_e^2 + n\sigma_a^2$ for MSTrt.
- We need to do algebra to find estimates of the variance components.
- $\sigma_e^2 = \text{MSErr}$ and $\sigma_a^2 = \frac{\text{MSTrt} - \text{MSErr}}{4}$.
- Here $n = 4$.
- Anyone see a potential problem?