

Multilevel Structures

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

April 17, 2008

Data Description

- We consider a subset of a larger data set on corn grown on the island Antigua.
- The response variable we consider is the harvest weight (`harvwt`) per plot (units unknown).
- There are eight sites with eight separate plots within each site where the corn is grown under the same treatment conditions.
- We can ask if the site has an effect on the harvest weight.
- In a standard regression framework, we could analyze the data as a one-way ANOVA with eight fixed parameters for the expected values (an intercept which is the mean of a reference group and seven differences in means between the other groups and the reference) and a single plot-level source of error.
- In a *multilevel model*, we can have covariates and error associated with the plot level and separate covariates and error associated with the site level.

- Standard ANOVA model:

$$y_i = \beta_1 + \beta_2 \cdot 1(\text{site } 2) + \cdots + \beta_8 \cdot 1(\text{site } 8) + e_i$$

where $i = 1, \dots, 64$ indexes the observation.

- ▶ $e_i \sim \text{iid } N(0, \sigma^2)$;
 - ▶ $\beta_j, j = 1, \dots, 8$ and σ^2 are fixed parameters.
- In a multilevel model, we may have

$$y_i = \alpha_{j[i]} + e_i$$

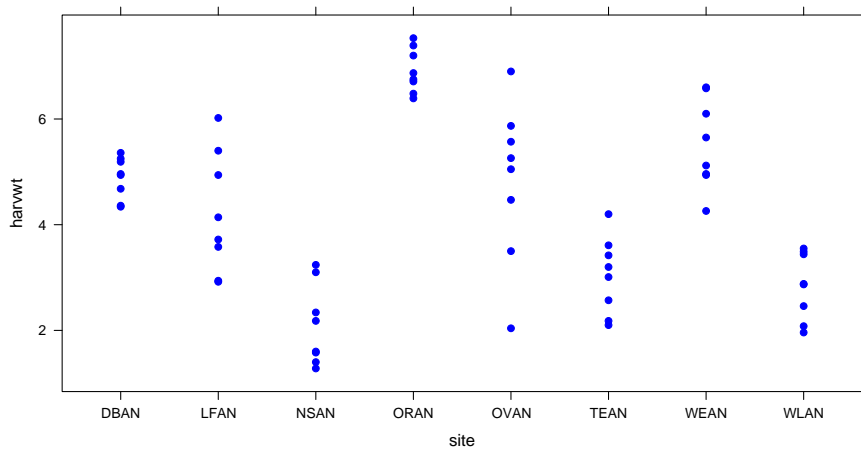
where $i = 1, \dots, 64$ indexes the observation and $j[i] = 1, \dots, 8$ indicates which of the eight sites contains the i th observation.

- ▶ $\alpha_j, j = 1, \dots, 8, \sim N(\mu_\alpha, \sigma_\alpha^2)$ are *random effects* for the sites;
 - ▶ $e_i \sim \text{iid } N(0, \sigma^2)$;
 - ▶ $\mu_\alpha, \sigma_\alpha$, and σ^2 are fixed and unknown.
- Notice here that we have a regression model for the response, and also a regression model for the coefficients of the first regression model.
 - Multilevel models include *sources of variation at more than one level*.

```
> corn = read.table("corn.txt", header = T)
> summary(corn)
```

	site	block	ears	harvwt
DBAN	: 8	I :16	Min. :13.00	Min. :1.280
LFAN	: 8	II :16	1st Qu.:37.75	1st Qu.:2.935
NSAN	: 8	III:16	Median :43.00	Median :4.300
ORAN	: 8	IV :16	Mean :41.22	Mean :4.292
OVAN	: 8		3rd Qu.:46.00	3rd Qu.:5.442
TEAN	: 8		Max. :58.00	Max. :7.530
(Other)	:16			

Plot of Data



Standard Regression Model

```
> corn.lm = lm(harvwt ~ site, data = corn)
> display(corn.lm)
```

```
lm(formula = harvwt ~ site, data = corn)
```

```
      coef.est coef.se
```

```
(Intercept)  4.89    0.31
```

```
siteLFAN     -0.68    0.44
```

```
siteNSAN     -2.79    0.44
```

```
siteORAN      2.03    0.44
```

```
siteOVAN     -0.05    0.44
```

```
siteTEAN     -1.85    0.44
```

```
siteWEAN      0.64    0.44
```

```
siteWLAN     -2.04    0.44
```

```
---
```

```
n = 64, k = 8
```

```
residual sd = 0.87, R-Squared = 0.77
```

Multilevel Model

```
> corn.lmer = lmer(harvwt ~ (1 | site), data = corn)
> display(corn.lmer)
```

```
lmer(formula = harvwt ~ (1 | site), data = corn)
      coef.est coef.se
(Intercept) 4.29    0.56
```

Error terms:

Groups	Name	Std.Dev.
site		1.55
Residual		0.87

number of obs: 64, groups: site, 8

AIC = 192.9, DIC = 190.3

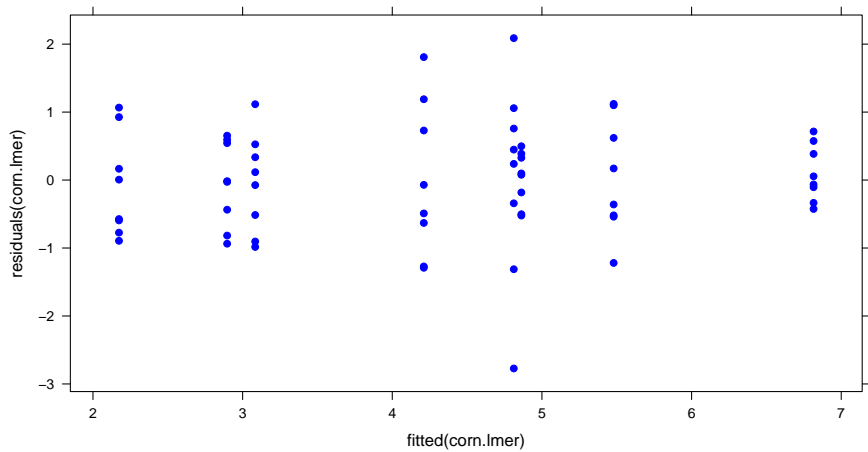
deviance = 189.6

- Discuss the different parameter estimates on the board.

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`.
- 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.

Residual Plot



Compare Fitted Values

```
> means = with(corn, sapply(split(harvwt, site), mean))
> fitted.lm = with(corn, sapply(split(fitted(corn.lm), site),
+   mean))
> fitted.lmer = with(corn, sapply(split(fitted(corn.lmer), site),
+   mean))
> signif(rbind(means, fitted.lm, fitted.lmer), 3)
```

	DBAN	LFAN	NSAN	ORAN	OVAN	TEAN	WEAN	WLAN
means	4.88	4.21	2.09	6.92	4.83	3.04	5.53	2.84
fitted.lm	4.88	4.21	2.09	6.92	4.83	3.04	5.53	2.84
fitted.lmer	4.86	4.21	2.17	6.82	4.81	3.08	5.48	2.90

- The overall mean is 4.29.
- The multilevel model *shrinks the estimates* toward the overall mean.

- Multilevel models are often used in these situations.
 - ▶ *Repeated measures* — when a single individual is measured multiple times, it is often appropriate to model two levels of variation, *one for individuals* and *one for measurements*.
 - ▶ *Split-plot designs* — in agricultural or ecological studies, it is often the case that sites are broken into plots and possibly subplots. Variables can be measured at the site, plot, subplot, or individual measurement level.
 - ▶ Multilevel models are also appropriate for *non-nested* variables. For example, measurements could be clustered by year and by site if a single site is measured over multiple years.

Summary of Classical Regression

- Prediction for continuous and discrete outcomes;
- Fitting nonlinear relationships using transformations;
- Inclusion of categorical predictors with indicator random variables;
- Modeling interactions;
- Causal inference.

Motivations for Multilevel Models

- Accounting for both individual and group level variation in estimating group-level effects.
- Modeling individual level regression coefficients.
- Estimation of effects for subgroups.

When is it worth fitting multilevel models?

- If the group size is small, there may not be much data to estimate random effects and there is little to gain.
- The complexity of multilevel models is greater than classical regression. The added complexity is often worthwhile, but perhaps not when there are only a small number (say less than five) individuals in a group.