Blocking, Analysis of Covariance (ANCOVA), & Mixed Models

“99 percent of all statistics only tell 49 percent of the story”
Ron DeLegge II (Economist)
Experimental Design Vocabulary

- **Experimental unit** — Individual research subject (rows in your data)

- **Treatment** — predictor variable (e.g. variety, fertilization, irrigation, etc.)

- **Treatment level** — groups within treatments (e.g. A,B,C or Control, 1xN, 2xN)

- **Blocking** — the arranging of experimental units in groups (blocks) that are similar to one another

- **Nuisance factor** — those that may affect the measured result, but are not of primary interest (ALL EXPERIMENTS HAVE THEM)
  
  Eg: the time of day you collected your data, the person who scored your treatment
Experimental Blocking — why do we do this?

• Blocking can be used to reduce or eliminate the contribution to experimental error contributed by nuisance factors by creating homogeneous groups (blocks) in which the nuisance factors are held constant and the factor of interest is allowed to vary

• Example from environmental science:

  Often we have environmental gradients in our sites that cause variation or effect our results, so to account for this we can place our blocks in such a way so that we minimize environmental heterogeneity (difference) Therefore our results are more likely due to your treatment effect
Experimental Blocking

• The general rule is:

"Block what you can; randomize what you cannot."

Blocking is used to remove the effects of a few of the most important nuisance variables

Randomization is then used to reduce the contaminating effects of the remaining nuisance variables

Blocking in R:

```r
anova(lm(YIELD~VARIETY+BLOCK))
aov(lm(YIELD~VARIETY+BLOCK))
```

NOTE: BLOCK needs to be a factor variable
Completely Randomized Design (CRD)

- Experimental units (our plots) are assigned to blocks at random
- Treatment levels are typically equally represented

**Example:**
From our lentil dataset - Varieties are planted randomly in blocks throughout plots
Randomized Block Design (RBD)

- Experimental units are assigned to blocks, then randomly to treatment levels
- The representation of treatment levels in each block are not necessarily equal

Example:
People split by medical history, then given a drug treatment
Incomplete Block Design

- Experimental units assigned to blocks, then randomly to treatment levels
- The representation of treatment levels in each block are NOT equal

Example:
In a harvesting study, when the size of available forest is not big enough to accommodate all thinning treatments
Randomized Complete Block Design (RCBD)

- Experimental units assigned to blocks, then randomly to treatment levels
- Treatment levels are equally represented in each block

Example:
Sites arranged on environmental gradient, with all treatments in each
Randomized Complete Block Design (RCBD)

Block 1  Block 2  Block 3  Block 4
A  B  B  C
B  A  C  B
C  C  A  A

Not so good.
Blocks are running perpendicular to the environmental gradient. Therefore, A and C are biased.
Randomized Complete Block Design (RCBD)

![Diagram of RCBD]

**Block 1**
- A
- B
- C

**Block 2**
- B
- A
- C

**Block 3**
- B
- C
- A

**Block 4**
- C
- A
- B

**Much better.**
Blocks represent similar environmental conditions.
No biased treatments.

The idea for any blocking is that we want to keep within block variation as minimal as possible.
Blocking by observer

You observe these (Block 1).

Assistant observes these (Block 2).

You observe these (Block 1).

Assistant observes these (Block 2).

Example:
Comparing the days for each stage of budflush between 2 environments
Experimental Blocking: Types

Complete Block Design (CBD) – Analyze with Multi-Way ANOVA

Randomized Block Design (RBD) – Analyze with Multi-Way ANOVA – However weaker power to detect differences

Incomplete Block Design (IBD) – Must create a clever algorithm to design how you are going to “combine treatment levels” – but even if you create an algorithm it is often difficult to actually make trial fit (e.g. hard to plant) Alpha design or Lattice design – RENR 580

Randomized Complete Block Design (RCBD) – BEST CHOICE – Analyze with Multi-Way ANOVA

Blocking by Observer – treat it like a CBD or RBD
But make sure you observations are randomized – otherwise you need to use a Lattice or Alpha design similar to an incomplete randomized design
Analysis of Covariance (ANCOVA)

What you use if your blocking variable is a continuous variable

- Use when the nuisance factor is too small to be blocked out
- Add the covariate to the model to increase the power of the test

**ANCOVA in R:**

```r
anova(lm(YIELD~VARIETY+BLOCK+COVARIATE))
aov(lm(YIELD~VARIETY+BLOCK+COVARIATE))
```

**NOTE:** BLOCK needs to be a factor variable
But... COVARIATE needs to be a continuous numeric variable
ANCOVA Assumptions

1. The experimental errors of your data are normally distributed

2. Equal variances between treatments
   - Homogeneity of variances
   - Homoscedasticity

3. Independence of samples
   - Each sample is randomly selected and independent
ANCOVA Assumptions

4. The errors of the covariate are normally distributed around the regression

5. The variances of the covariate are equal for each treatment

6. Homogeneity of the regression coefficients
   The slopes of each of the regression lines, describing the relationship between the dependent variable and the covariate for each treatment level, are equal

Analysis of ANCOVA appears to be robust to this assumption
Comparing results from ANOVA and ANCOVA

• Notice that when we include the COVARIATE our sum of squares and mean sum of squares for VARIETY (signal) stays the same

• But the inclusion of NITROGEN accounts for part of the error!

• This lowers the residual sum of squares and mean sum of squares and thus increasing the F-ratio and decreasing the corresponding p-value

MORE POWERFUL AT DETECTING DIFFERENCES

\[ F = \frac{\text{variance between}}{\text{variance within}} \]

\[ F = \frac{MS_{\text{Treatment}}}{MS_{\text{ERROR}}} \]
Mixed Models

Don’t use sum of squares approach (e.g. ANOVA, ANCOVA) to find differences

But rather these models guess at the parameters and compare the errors by an iterative process to see what gets worse when the generated parameters are varied

Mixed Model to Estimate Means

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>ERROR</td>
</tr>
<tr>
<td>724</td>
<td>580</td>
<td>562</td>
<td>256</td>
</tr>
<tr>
<td>722</td>
<td>580</td>
<td>562</td>
<td>257</td>
</tr>
<tr>
<td>728</td>
<td>580</td>
<td>562</td>
<td>254</td>
</tr>
</tbody>
</table>

When I change my parameter guess does my error get better or worse?
Mixed Models

• **Fixed Effect** – treatments that are control (fixed) by the experiment

• **Random Effect** – effects where you do not care about their significance such as a block treatment meant to account for noise

Mixed Models in R:

```r
install.packages("lme4")
library(lme4)

MyModel = lmer(YIELD~VARIETY+(1|BLOCK))
summary(MyModel)
anova(MyModel)
1-pf(651.22,2,10) # Calculate your p-value
```

1 degrees of freedom

Look at taking RENR 580 for a more in-depth study of Mixed Models