STAT 215
Multifactor ANOVA I

Colin Reimer Dawson

Oberlin College

November 28, 2017

Outline

Alfalfa sprouts (Ex. 6.25)

Some students were interested in the effect of acidic environments on plant growth. They planted alfalfa seeds in fifteen cups and randomly chose five to get plain `water`, five to get a `moderate` amount of acid and five to get a `stronger` acid solution. The cups were arranged indoors near a window in five rows of three with one cup from each `Acid` level in each row (with row a nearest the window, and row e farthest away). The response variable was average `Height` of the alfalfa sprouts after four days.

A model:

$$\texttt{Acid} = \mu + \alpha_k + \varepsilon, \qquad k = \texttt{water}, \texttt{moderate}, \texttt{strong}$$

Any concerns about the ANOVA/regression conditions? The residuals might not be independent within rows!

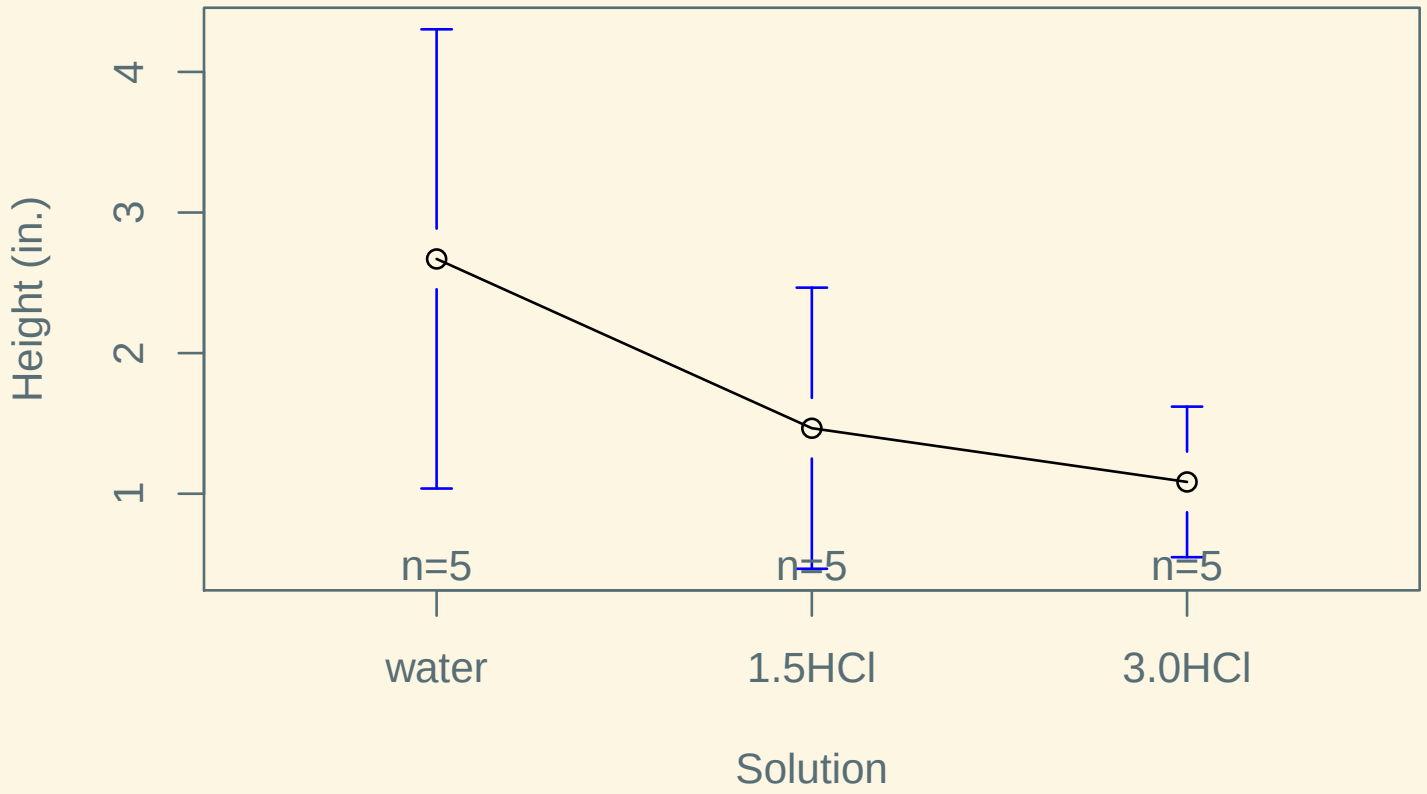
Alfalfa Data

Treatment/Row	a	b	c	d	e	Trt. mean
water	1.45	2.79	1.93	2.33	4.85	2.67
moderate acid	1.00	0.70	1.37	2.80	1.46	1.47
strong acid	1.03	1.22	0.45	1.65	1.07	1.08
Row mean	1.16	1.57	1.25	2.26	2.46	1.74

Since each treatment is applied to each row, we can include row as a predictor.

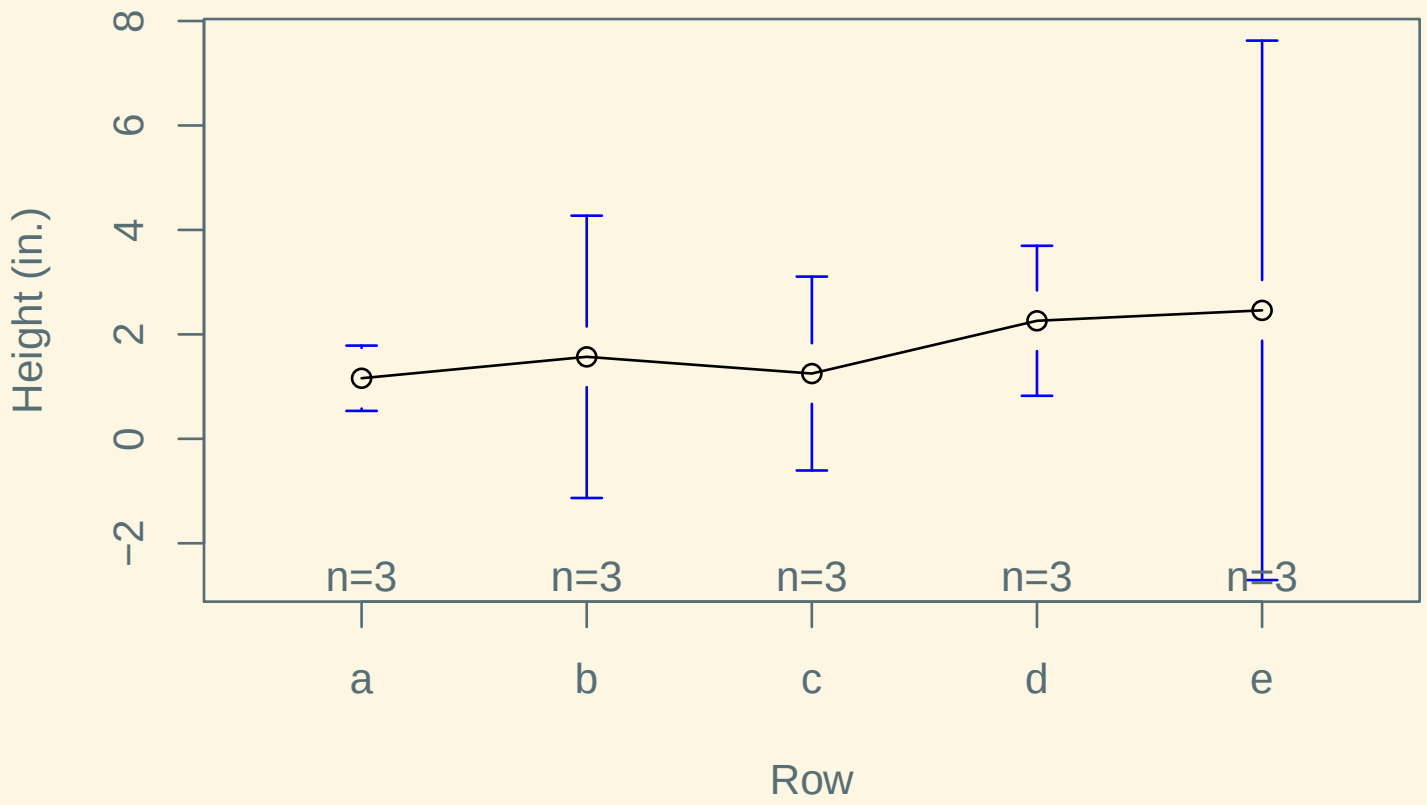
Means Plots

```
library("Stat2Data"); library("mosaic"); library("gplots")
data("Alfalfa")
plotmeans(Ht4 ~ factor(Acid, levels = c("water", "1.5HCl", "3.0HCl")),
          data = Alfalfa, xlab = "Solution", ylab = "Height (in.)")
```



Means Plots

```
plotmeans(Ht4 ~ factor(Row), data = Alfalfa,
          xlab = "Row", ylab = "Height (in.)")
```



The One-way ANOVA Population Model (X categorical)

$$Y = f(X) + \varepsilon$$
$$Y = \mu + \alpha_k + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2_\varepsilon)$$

One α_k for each level of X: group deviation from overall mean

The Two-way ANOVA Additive Model (X_A, X_B categorical)

$$Y = f(X) + \varepsilon$$
$$Y = \mu + \alpha_j + \beta_k + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2_\varepsilon)$$

One α_j for each level of X_A

One β_k for each level of X_B

FIT: Parameter Estimation

- Population model:

$$y_{j,k,i} = \mu + \alpha_j + \beta_k + \varepsilon_{j,k,i}$$

- Estimate terms by

$$\begin{aligned}
\hat{\mu} &= \bar{\bar{y}} \\
\hat{\alpha}_j &= \bar{y}_j - \bar{\bar{y}} \\
\hat{\beta}_k &= \bar{y}_k - \bar{\bar{y}} \\
\hat{y}_{j,k,i} &= \hat{\mu} + \hat{\alpha}_j + \hat{\beta}_k \\
\hat{\varepsilon}_{j,k,i} &= y_{j,k,i} - \hat{y}_{j,k,i}
\end{aligned}$$

Sums of Squares

$$y_{j,k,i} = \hat{\mu} + \hat{\alpha}_j + \hat{\beta}_k + \varepsilon_{j,k,i}$$

$$(y_{j,k,i} - \hat{\mu})^2 = (\hat{\alpha}_j + \hat{\beta}_k + \varepsilon_{j,k,i})^2$$

$$SS_A = \sum_j \sum_k \sum_{i=1}^{n_{j,k}} \hat{\alpha}_j^2$$

$$SS_B = \sum_j \sum_k \sum_{i=1}^{n_{j,k}} \hat{\beta}_k^2$$

$$SS_{Error} = \sum_j \sum_k \sum_{i=1}^{n_{j,k}} \hat{\varepsilon}_{j,k,i}^2$$

Note: $SS_{Total} = SS_A + SS_B + SS_{Error}$, since cross terms are all zero.

The Two-Way ANOVA Table

Source	df	SS	MS	F	P
Factor A	$J-1$				
Factor B	$K-1$				
Residuals	$N-J-K+1$			—	—
Total	$N-1$			—	—

Pair "Quiz": Factor A has $J = 3$ levels, factor B has $K = 5$ levels, with one observation per cell. How many degrees of freedom in each row of the table above?

Two-Way ANOVA Table

```
library("mosaic"); library("Stat2Data")
data("Alfalfa")
two.way.model <- aov(Ht4 ~ Acid + Row, data = Alfalfa)
summary(two.way.model)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Acid          2  6.852   3.426   4.513 0.0487 *
Row           4  4.183   1.046   1.378 0.3235
Residuals     8  6.072   0.759
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Caution: The F tests here amount to sequential *nested* F-tests, so order matters if there is any collinearity (here there is none, since the design is perfectly balanced)

Getting Means

```
## Note: this only works if you used aov(), not lm()
model.tables(two.way.model, type = "means")


     Tables of means
     Grand mean

     1.74

      Acid
     Acid
     1.5HCl 3.0HCl  water
      1.466  1.084  2.670

      Row
     Row
        a    b    c    d    e
     1.16 1.57 1.25 2.26 2.46
```

Getting "Effects" (αs and βs)

```
## Note: this only works if you used aov(), not lm()
model.tables(two.way.model, type = "effects")

     Tables of effects

      Acid
     Acid
     1.5HCl 3.0HCl  water
     -0.274 -0.656  0.930

      Row
     Row
         a     b     c     d     e
     -0.58 -0.17 -0.49  0.52  0.72

## Notice that the alphas and betas each sum to zero
```

Post-Hoc Pairwise Comparisons

```
TukeyHSD(two.way.model)


  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Ht4 ~ Acid + Row, data = Alfalfa)

$Acid
                 diff         lwr      upr     p adj
3.0HCl-1.5HCl -0.382 -1.95650626 1.192506 0.7739299
water-1.5HCl   1.204 -0.37050626 2.778506 0.1338368
water-3.0HCl   1.586  0.01149374 3.160506 0.0484908

$Row
     diff      lwr     upr     p adj
b-a  0.41 -2.04758 2.86758 0.9750089
c-a  0.09 -2.36758 2.54758 0.9999282
d-a  1.10 -1.35758 3.55758 0.5642564
e-a  1.30 -1.15758 3.75758 0.4211177
c-b -0.32 -2.77758 2.13758 0.9899007
d-b  0.69 -1.76758 3.14758 0.8613573
e-b  0.89 -1.56758 3.34758 0.7251160
d-c  1.01 -1.44758 3.46758 0.6333208
e-c  1.21 -1.24758 3.66758 0.4830625
e-d  0.20 -2.25758 2.65758 0.9983249
```

Additive vs. Interaction Model

The Two-way ANOVA Additive Model (X_A, X_B categorical)

$$Y = f(X) + \varepsilon$$

$$Y = \mu + \alpha_j + \beta_k + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2_\varepsilon)$$

One α_j for each level of X_A

One β_k for each level of X_B

Assumes the "effect" of Factor A is the same at each level of Factor B (like parallel lines models in regression).

Interaction Model

The Two-way ANOVA Interaction Model (X_A, X_B categorical)

$$Y = f(X) + \varepsilon$$

$$Y = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2_\varepsilon)$$

One α_j for each level of X_A
One β_k for each level of X_B

One γ_{jk} for each *combination* of X_A and X_B

- Predicted "effect" of level j of factor A, when at level k of factor B: $\alpha_j + \gamma_{jk}$.
- Predicted "effect" of level k of factor B, when at level j of factor A: $\beta_k + \gamma_{jk}$.
- "Effects" are modulated by the interaction term, γ_{jk}: a "difference of differences".

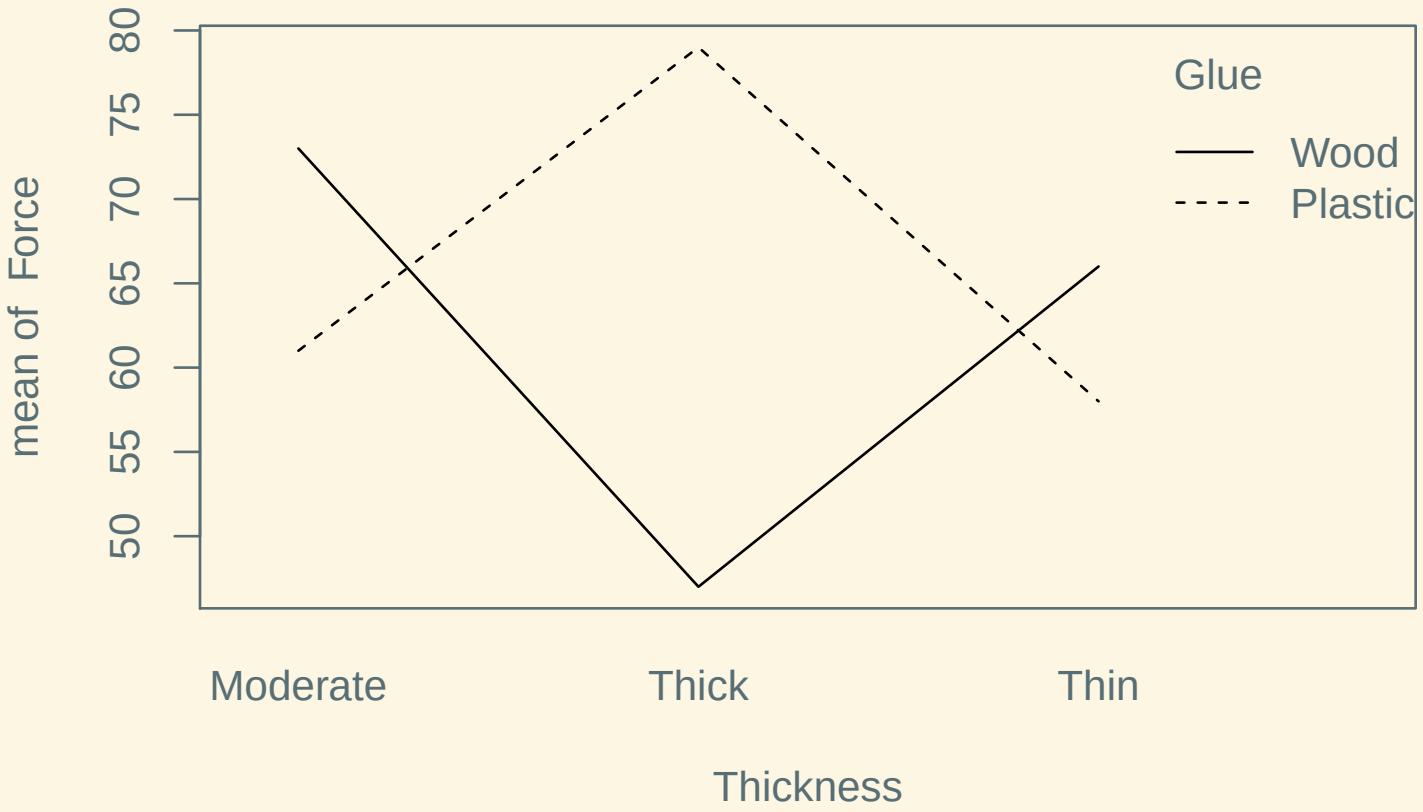
Example: Glue!

An experiment recorded the amount of force in Newtons (the response) that it takes to separate two pieces of plastic that have been glued together, for three different thicknesses of material (thin, moderate, thick), and two types of glue (wood vs. plastic). There are two cases at each combination of factors. The data is below usual one-row-per-case form.

Thickness/Glue	Plastic	Wood	**Mean**
Thin	$52, 64$	$72, 60$	62
Moderate	$67, 55$	$78, 68$	67
Thick	$86, 72$	$43, 51$	63
Mean	66	62	64

Glue!

```
library("mosaic")
GlueData <- read.file("http://colinreimerdawson.com/data/glue.csv")
### Plot the means
with(GlueData, interaction.plot(Thickness, Glue, Force))
```



Glue!

We can write down a two-way ANOVA model with an interaction as follows:

$$Y = \mu + \alpha_j + \beta_k + \gamma_{j,k} + \varepsilon$$

How do we interpret each coefficient?

Fitting the Model

Demo

Degrees of Freedom

With J levels of factor A and K levels of factor B:

- J different αs, but $J-1$ degrees of freedom (they must sum to zero)
- K different βs, but $K-1$ degrees of freedom (they must sum to zero)
- JK different γs, but only $(J-1)(K-1)$ df (must sum to zero at *each* J and *each* K)

The interaction model has enough flexibility to fit *any* pattern of cell means. Need more than one obs. per cell to estimate fit / to do hypothesis tests.