STAT 213
Two-Way ANOVA I

Colin Reimer Dawson

Oberlin College

May 4, 2020

Outline

Outline

Two-Way ANOVA: Additive Model

FIT: Estimating Parameters

ASSESS: Variance Decomposition

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.

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:

$$\begin{aligned} \texttt{Height}_i &= \mu + \alpha_{\texttt{Acid}_i} + \varepsilon_i \\ \texttt{Acid}_i &\in \{\texttt{water}, \texttt{moderate}, \texttt{strong}\} \end{aligned}$$

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:

$$\begin{aligned} \texttt{Height}_i &= \mu + \alpha_{\texttt{Acid}_i} + \varepsilon_i \\ \texttt{Acid}_i &\in \{\texttt{water}, \texttt{moderate}, \texttt{strong}\} \end{aligned}$$

Any concerns about the ANOVA/regression conditions?

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:

$$\texttt{Height}_i = \mu + \alpha_{\texttt{Acid}_i} + \varepsilon_i$$
$$\texttt{Acid}_i \in \{\texttt{water}, \texttt{moderate}, \texttt{strong}\}$$

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

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

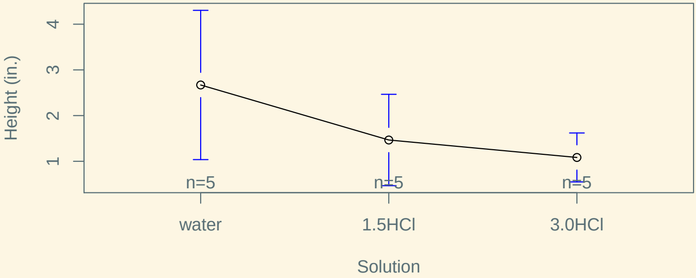
Alfalfa Data

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.

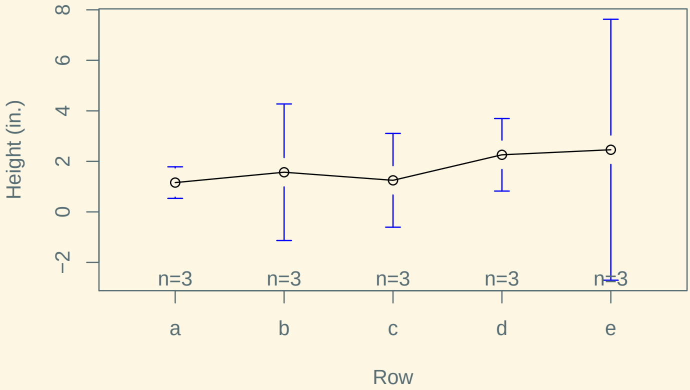
Means Plots

```
library(Stat2Data); library(mosaic); library(gplots)
data(Alfalfa)
AlfalfaOrdered <- Alfalfa %>%
    mutate( # Reorder Acid to be in the natural ordering
        Acid = factor(Acid, levels = c("water", "1.5HCl", "3.0HCl")),
        Row  = factor(Row)) # Make sure it's treated as categorical)
plotmeans(Ht4 ~ Acid,
          data = AlfalfaOrdered, xlab = "Solution", ylab = "Height (in.)")
```



Means Plots

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



The One-way ANOVA Population Model (X categorical)

$$Y_i = f(X_i) + \varepsilon_i$$
$$Y = \mu + \alpha_{X_i} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$$

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

The One-way ANOVA Population Model (X categorical)

$$Y_i = f(X_i) + \varepsilon_i$$
$$Y = \mu + \alpha_{X_i} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2_\varepsilon)$$

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

The Two-way ANOVA Additive Model (A, B categorical)

$$Y_i = f(A_i, B_i) + \varepsilon_i$$
$$Y_i = \mu + \alpha_{A_i} + \beta_{B_i} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2_\varepsilon)$$

One α_A for each level of A ("row" deviation from overall mean)
One β_B for each level of B ("column" deviation from overall mean)

Concretely: Alfalfa Sprouts

$$\widehat{\texttt{Height}}_i = \begin{cases} \mu + \alpha_{\texttt{Water}} + \beta_a & \text{if } \texttt{Acid}_i = \texttt{"Water"} \text{ and } \texttt{Row}_i = \texttt{"a"} \\ \mu + \alpha_{\texttt{Water}} + \beta_b & \text{if } \texttt{Acid}_i = \texttt{"Water"} \text{ and } \texttt{Row}_i = \texttt{"b"} \\ \dots & \dots \\ \mu + \alpha_{\texttt{Water}} + \beta_e & \text{if } \texttt{Acid}_i = \texttt{"Water"} \text{ and } \texttt{Row}_i = \texttt{"e"} \\ \dots & \dots \\ \mu + \alpha_{\texttt{HCl1.5}} + \beta_a & \text{if } \texttt{Acid}_i = \texttt{"HCl1.5"} \text{ and } \texttt{Row}_i = \texttt{"a"} \\ \dots & \dots \\ \mu + \alpha_{\texttt{HCl1.5}} + \beta_e & \text{if } \texttt{Acid}_i = \texttt{"HCl1.5"} \text{ and } \texttt{Row}_i = \texttt{"e"} \\ \mu + \alpha_{\texttt{HCl3.0}} + \beta_a & \text{if } \texttt{Acid}_i = \texttt{"HCl3.0"} \text{ and } \texttt{Row}_i = \texttt{"a"} \\ \dots & \dots \\ \mu + \alpha_{\texttt{HCl3.0}} + \beta_e & \text{if } \texttt{Acid}_i = \texttt{"HCl3.0"} \text{ and } \texttt{Row}_i = \texttt{"e"} \end{cases}$$

Outline

Two-Way ANOVA: Additive Model

FIT: Estimating Parameters

ASSESS: Variance Decomposition

FIT: Parameter Estimation

- Population model:

$$y_{a,b,i} = \mu + \alpha_a + \beta_b + \varepsilon_{a,b,i}$$

where, for simplicity we let $i = 1, \ldots, N_{a,b}$ index which of the $N_{a,b}$ observations that have $A = a$ and $B = b$ we are referring to (in the alfalfa case, $N_{a,b} = 1$ for all a, b)

FIT: Parameter Estimation

- Population model:

$$y_{a,b,i} = \mu + \alpha_a + \beta_b + \varepsilon_{a,b,i}$$

  where, for simplicity we let $i = 1, \ldots, N_{a,b}$ index which of the $N_{a,b}$ observations that have $A = a$ and $B = b$ we are referring to (in the alfalfa case, $N_{a,b} = 1$ for all a, b)
- When sample sizes are equal, estimate terms by

$$\begin{aligned}
\hat{\mu} &= \bar{\bar{Y}} \text{ (“grand” mean)} \\
\hat{\alpha}_a &= \bar{Y}_{a\cdot} - \bar{\bar{Y}} \text{ (“row” deviation)} \\
\hat{\beta}_b &= \bar{Y}_{\cdot b} - \bar{\bar{Y}} \text{ (“column” deviation)} \\
\hat{Y}_{a,b,i} &= \hat{\mu} + \hat{\alpha}_a + \hat{\beta}_b \text{ (predicted value)} \\
\hat{\varepsilon}_{a,b,i} &= Y_{a,b,i} - \hat{Y}_{a,b,i} \text{ (residual)}
\end{aligned}$$

Practice: 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

Find: $\hat{\mu}, \hat{\alpha}_{\texttt{Water}}, \hat{\alpha}_{\texttt{moderate}}, \hat{\alpha}_{\texttt{strong}}, \hat{\beta}_a, \ldots, \hat{\beta}_e$

Outline

Sums of Squares

$$Y_{a,b,i} = \hat{\mu} + \hat{\alpha}_a + \hat{\beta}_b + \varepsilon_{a,b,i}$$

$$(Y_{a,b,i} - \hat{\mu})^2 = (\hat{\alpha}_a + \hat{\beta}_b + \varepsilon_{a,b,i})^2$$

$$SS_A = \sum_{a=1}^{J} \sum_{b=1}^{K} \sum_{i=1}^{N_{A,B}} \hat{\alpha}_a^2 = \sum_{a=1}^{J} N_{a.} \hat{\alpha}_a^2$$

$$SS_B = \sum_{a=1}^{J} \sum_{b=1}^{K} \sum_{i=1}^{N_{A,B}} \hat{\beta}_b^2 = \sum_{b=1}^{K} N_{.b} \hat{\beta}_b^2$$

$$SS_{Error} = \sum_{a=1}^{J} \sum_{b=1}^{K} \sum_{i=1}^{N_{a,b}} \hat{\varepsilon}_{a,b,i}^2 \text{ (doesn't simplify)}$$

$$SS_{Total} = \sum_{a=1}^{J} \sum_{b=1}^{K} \sum_{i=1}^{N_{a,b}} (Y_{a,b,i} - \hat{\mu})^2 = SS_A + SS_B + SS_{Error}$$

Alfalfa: Sums of Squares

Treatment	Row	i	Height	$\hat{\mu}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\varepsilon}$
water	a	1	1.45	1.74	0.93	-0.58	
water	b	1	2.79	1.74	0.93	-0.17	
water	c	1	2.79	1.74	0.93	-0.49	
water	d	1	2.79	1.74	0.93	0.52	
water	e	1	4.85	1.74	0.93	0.72	
moderate	a	1	1.00	1.74	-0.27	-0.58	
...	...	...	...	...	...	...	
moderate	e	1	1.46	1.74	-0.27	0.72	
strong	a	1	1.03	1.74	-0.67	-0.58	
...	...	...	...	...	...	...	
strong	e	1	1.07	1.74	-0.67	0.72	
					$SS_A = \sum \hat{\alpha}^2$	$SS_B = \sum \hat{\beta}^2$	$SS_E = \sum \hat{\varepsilon}^2$