STA305/1004 - Class 14

March 6, 2017

Today's class

- ANOVA demonstrtation
- Estimating Treatment Effects in ANOVA using Regression
- Coding Qualitative Predictors in Regression Models

ANOVA Demonstration

Figure 1:

- Count the total number of each colour (e.g., yellow, purple, pink, green).
- Eat the Smarties.

ANOVA Data Setup

- How should the data be setup?

Box	Colour	Count
1	Green	
1	Pink	
1	Purple	
1	Yellow	
2	Green	
2	Pink	
2	Purple	
2	Yellow	
3	Green	
3	Pink	
3	Purple	
3	Yellow	
4	Green	
4	Pink	
4	Purple	
4	Yellow	
5	Green	
5	Pink	
5	Purple	
5	Yellow	

-

Figure 2:

Smarties Data from 3 boxes

Colour 4 = $\bar{y}_{4.} - \bar{y}_{1.} = 3.7 - 2.7 = 1$

```
count <- c(4,3,4,3,1,4,2,5,1,1,2,4)
colour <- as.factor(c(rep("Yellow",3),rep("Purple",3),
                      rep("Green",3),rep("Pink",3)))
#Get means for each flavour
sapply(split(count,colour),mean)
```

$\bar{y}_{..}$ = ave

```
   Green     Pink   Purple   Yellow
2.666667 2.333333 2.666667 3.666667
```

Intercept = $\bar{y}_{1.} = 2.7$ (green)

Colour 2 = $\bar{y}_{2.} - \bar{y}_{1.} = 2.3 - 2.7 = -0.4$

Colour 3 = $\bar{y}_{3.} - \bar{y}_{1.} = 2.7 - 2.7 = 0$

Estimating Treatment Effects in ANOVA using Regression

X_{ij} = not uniquely defined.

ϵ_{ij} = represents the error / variation of # Smarties within each Colour.

- y_{ij} is the j^{th} observation under the i^{th} treatment.
- The model for smarties $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, $\epsilon_{ij} \sim N(0, \sigma^2)$ can be written in terms of the dummy variables X_1, X_2, X_3 as:

$$y_{ij} = \mu + \tau_1 X_{i1} + \tau_2 X_{i2} + \tau_3 X_{i3} + \epsilon_{ij}.$$

- What is $y_{ij}, \mu, \tau_i, X_{ij}, \epsilon_{ij}$?

Smarties: Unit of analysis: box

Treatment: Colour = Green, pink, purple, yellow

y_{ij} = # Smarties in box j of Colour i.

The ANOVA Table

```r
#ANOVA table
anova(lm(count~colour))
```

```
Analysis of Variance Table

Response: count
          Df Sum Sq Mean Sq F value Pr(>F)
colour     3  3.000  1.0000  0.4286 0.7381
Residuals  8 18.667  2.3333
```

does not require prob.

Requires Probability.

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

$H_1: \mu_i \neq \mu_j$

$SS_{Colour} + SS_{Resid} = SS_{Total}$

$3.000 + 18.667 = 3 + 18.667$

Dummy coding

- Dummy coding compares each level to the reference level. The intercept is the mean of the reference group.
- Dummy coding is the default in R and the most common coding scheme. It compares each level of the categorical variable to a fixed reference level.

```
contrasts(colour) <- contr.treatment(4)  #Treatment contrast
contrasts(colour) # print dummy coding
```

$x_{1i}\, x_{2i}\, x_{3i}$ — the columns define x_{1i}, x_{2i}, x_{3i}

```
        2 3 4
Green   0 0 0
Pink    1 0 0
Purple  0 1 0
Yellow  0 0 1
```

this is $\underline{0}$ so that $(X^TX)^{-1}$ exists.

$$x_{1j} = \begin{cases} 1 & \text{if Smartie in box}_j \text{ is pink} \\ 0 & \text{o.w} \end{cases}$$

$$x_{2j} = \begin{cases} 1 & \text{if Smartie in box}_j \text{ is purple} \\ 0 & \text{o.w} \end{cases}$$

$$x_{3j} = \begin{cases} 1 & \text{if Smartie is yellow} \\ 0 & \text{o.w} \end{cases}$$

```
lm(count~colour)
```

```
Call:
lm(formula = count ~ colour)

Coefficients:
(Intercept)      colour2      colour3      colour4
  2.667e+00   -3.333e-01    4.710e-16    1.000e+00
```

$\bar{y}_{1\cdot}$ $\quad -3.3 \times 10^{-1} \quad 0$

least Squares(LS) estimates.

$$y_{ij} = \mu + \tau_1 x_{1j} + \tau_2 x_{2j} + \tau_3 x_{3j} + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

If the Colour is green (this is the ref. Colour).

$$E(y_{1j}) = \mu = \mu_1$$

If Colour is pink

$$E(y_{2j}) = \mu + \tau_1 = \mu_2 \Rightarrow \tau_1 = \mu_2 - \mu_1$$

If Colour is purple

$$E(y_{3j}) = \mu + \tau_2 = \mu_3 \Rightarrow \tau_2 = \mu_3 - \mu_1$$

$$E(y_{4j}) = \mu + \tau_3 = \mu_4 \Rightarrow \tau_3 = \mu_4 - \mu_1$$

The LS Estimates:

$$\hat{\mu} = \bar{y}_{1\cdot} = \hat{\mu}_1$$

$$\hat{\tau}_1 = \bar{y}_{2\cdot} - \bar{y}_{1\cdot}$$

$$\hat{\tau}_2 = \bar{y}_{3\cdot} - \bar{y}_{1\cdot}$$

$$\hat{\tau}_3 = \bar{y}_{4\cdot} - \bar{y}_{1\cdot}$$

$$y_{1\cdot} = \sum_{j=1}^{n_1} y_{ij}$$

$$\bar{y}_{1\cdot} = \frac{\sum_{j=1}^{n_1} y_{ij}}{n_1}$$

Deviation coding

- This coding system compares the mean of the dependent variable for a given level to the overall mean of the dependent variable.

```
contrasts(colour) <- contr.sum(4)  # Deviation contrast
contrasts(colour) # print deviation coding
```

```
       [,1] [,2] [,3]
Green     1    0    0
Pink      0    1    0
Purple    0    0    1
Yellow   -1   -1   -1
```

Compares the mean dependent. Variable for given Colour to overall mean.

$$y_{ij} = \mu + \tau_1 x_{1j} + \tau_2 x_{2j} + \tau_3 x_{3j} + \varepsilon_{ij}$$

```
lm(count~colour)
```

$$x_{1j} = \begin{cases} 1 & \text{green} \\ 0 & \text{ow} \\ -1 & \text{yellow} \end{cases}, \quad x_{2j} = \begin{cases} 1 & \text{pink} \\ 0 & \text{ow} \\ -1 & \text{yellow} \end{cases}$$

$$x_{3j} = \begin{cases} 1 & \text{purple} \\ 0 & \text{ow} \\ -1 & \text{yellow} \end{cases}$$

```
Call:
lm(formula = count ~ colour)

Coefficients:
(Intercept)      colour1      colour2      colour3
     2.8333      -0.1667      -0.5000      -0.1667
```

$$E(y_{1j}) = \mu_1 = \mu + \tau_1$$

$$E(y_{2j}) = \mu_2 = \mu + \tau_2$$

$$E(y_{3j}) = \mu_3 = \mu + \tau_3$$

$$E(y_{4j}) = \mu_4 = \mu - \tau_1 - \tau_2 - \tau_3$$

$$\mu = \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4} = \frac{\mu + \tau_1 + \mu + \tau_2 + \mu + \tau_3 + \mu - \tau_1 - \tau_2 - \tau_3}{4}$$

$$\mu_1 - \left(\frac{\mu_1+\mu_2+\mu_3+\mu_4}{4}\right) = \mu+\tau_1-\mu = \tau_1$$

$$\mu_2 - \left(\frac{\mu_1+\mu_2+\mu_3+\mu_4}{4}\right) = \mu+\tau_2-\mu = \tau_2$$

$$\mu_3 \Rightarrow \left(\frac{\mu_1+\mu_2+\mu_3+\mu_4}{4}\right) = \tau_3$$

LS Estimates:

$$\bar{y}_{..} = \frac{2.7 + 2.3 + 2.7 + 3.7}{4}$$

$$= 2.85$$

$$\bar{y}_{1.} - \bar{y}_{..} = \hat{\tau}_1 = 2.7 - 2.85 = 0.15$$

$$\bar{y}_{2.} - \bar{y}_{..} = \hat{\tau}_2 = 2.3 - 2.85 = -0.55$$

$$\bar{y}_{3.} - \bar{y}_{..} = \hat{\tau}_3 = 2.7 - 2.85 = -0.15$$