STA305/1004 - Class 17

March 7, 2016

Today's class

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

ANOVA Demonstration

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

2 purple 2 red 1 pink

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	

-

Smarties Data from 3 boxes

```
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)
```

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

Estimating Treatment Effects in ANOVA using Regression

dummy variable coding is one type of coding in for qualitative variables in regression.

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

X_{i1} = 1, if colour is pink and X_{i1}=0 otherwise.
X_{i2} = 1, if colour is purple and X_{i2}=0 otherwise.
X_{i3} = 1, … etc.

\mu and \tau_i will depend on how X_{ij} and defined \epsilon_{ij} = within “treatment” (colour) error/variation.

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
```

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
```

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

basically, we have 4 treatments, but we constraint one of them to be all “zeroes” (for regression)

$E(y_{i1})=\mu_1=\mu$, $E(\epsilon_{ij})=0$
$E(y_{i2})=\mu_2=\mu+\tau_1 \rightarrow \tau_1=\mu_2-\mu_1$
$E(y_{i3})=\mu_3=\mu+\tau_2 \rightarrow \tau_2=\mu_3-\mu_1$
$E(y_{i4})=\mu_4=\mu+\tau_3 \rightarrow \tau_3=\mu_4-\mu_1$

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

Least squares estimators
$\hat{\mu} = \bar{y_{\cdot 1}} = \hat{\mu_1}$
$\hat{\tau_1}=\bar{y_{\cdot 2}}-\bar{y_{\cdot 1}}$
$\hat{\tau_2}=\bar{y_{\cdot 3}}-\bar{y_{\cdot 1}}$
$\hat{\tau_3}=\bar{y_{\cdot 4}}-\bar{y_{\cdot 1}}$

```
##
## Call:
## lm(formula = count ~ colour)
##
## Coefficients:
## (Intercept)      colour2      colour3      colour4
##   2.667e+00   -3.333e-01    4.710e-16    1.000e+00
```

Deviation coding

compare mean count for a given colour to the mean of all the colours.

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

```r
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
```

X_{i1} = 1 green,
= 0 o.w.
= -1 yellow
X_{i2} = 1 pink,
= 0 o.w.
= -1 yellow
X_{i3} = 1 purple,
= 0 o.w.
= -1 yellow

E(y_{i1}) = \mu_1 = \mu + \tau_1
E(y_{i2}) = \mu_2 = \mu + \tau_2
E(y_{i3}) = \mu_3 = \mu + \tau_3
E(y_{i4}) = \mu_4
= \mu - \tau_1 - \tau_2 - \tau_3

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

```
##
## Call:
## lm(formula = count ~ colour)
##
## Coefficients:
## (Intercept)      colour1      colour2      colour3
##      2.8333      -0.1667      -0.5000      -0.1667
```

$(\mu_1+\mu_2+\mu_3+\mu_4)/4 = \ldots$