Stat 414 – Day 6
Applet Demonstration – Adjusted Associations

The file saldata.txt contains data on 24 college graduates, including their starting salary (in thousands of dollars), how many semesters they spent in college, and their major. (This isn't real data but is based on real data :)).

(a) Use this link to open the Multiple Variables applet and press the **Use Data** button to load in the salary data. Drag *salary* to the Response box and *semesters* to the Explanatory box. Summarize the nature of the association. Is it what you expected?

(b) Check the **Show equation** and **Show residuals** boxes. Everything look ok?

(c) Check the **R-squared** box and report the value.

(d) Remove *semesters* from the Explanatory box and move *major* to the **Subset by** box. Also check the **Show descriptive** box. Are the validity conditions for an "analysis of variance" met? How much variation in salaries is explained by major in this dataset?

- Write out the estimated model equation using *indicator coding* with chemistry as the reference group.

- Write out the estimated model equation using effect coding. (*Hint*: Need more info…)

(e) Drag *major* to the explanatory variable box. Check the box for **Statistical model** and confirm your answers for both types of coding. Also, what interesting feature do you notice about the standard errors of the slope coefficients? Is this a coincidence?

(f) Drag *semesters* to the explanatory box below the *major* variable. This will show the graph of *salary vs semesters* but color-coding the dots by major. How would you describe the relationship between salary and semesters for students *within the same major*?

(g) Report and interpret the slope of semester in the model with both variables. How has your interpretation changed from (a)? Why?

(h) Is the R^2 value for this model equal to the sum of the other two R^2 values? Is this what you expected?

(i) Sketch the scatterplot you expect to see if I graph the mean number of semesters and the number salary for each major (i.e., using major as the observational unit). Would this have a positive or negative association?

(j) Check the **ANOVA table** box. Explain what the SS values represent. How do these values relate to what's in the pie chart?

Source	df	SS	MS	F-stat	p-value
model	3	1643.21	547.74	63.78	< 0.0001
semesters	1	352.13	352.13	41.01	< 0.0001
major	2	1043.15	521.57	60.74	< 0.0001
Error	20	171.75	8.59		
Total	23	1814.96			

(k) What does R do instead?

```
> summary(aov(saldata$salary ~ saldat
                  Df Sum Sq Mean Sq F
saldata$major      2 1291.1   645.5
saldata$semesters  1  352.1   352.1
Residuals         20  171.7     8.6
```

(l) Suppose I tell you I know the salary of a business major and I'm going to randomly select another business major. Do you think you have a pretty good prediction of the second student's salary?

Recall: Monday we suggested this formula for the ICC

$$ICC = \frac{MSbetween - MSwithin}{MSbetween + (n-1)MSwithin}$$

(m) Compute this value from the ANOVA table with just major

Source	df	SS	MS	F-stat	p-value
major	2	1291.08	645.54	25.88	< 0.0001
Error	21	523.88	24.95		
Total	23	1814.96			

What percentage of the variation in salaries is due to major?