Stat 414 – Day 5
Adjusted associations, Intraclass correlation

Last Time: Categorical variable with *k* categories
- Adds *k* – 1 *terms* to the model
  - Indicator/dummy/one-hot encoding vs. effect/sum to zero coding
- With just one variable is equivalent to one-way ANOVA (equal variance, normality)
  - Use partial F-test to assess the significance of the variable
- Reminders: $R^2 = corr(y, \hat{y}) = 1 - SSError/SSTotal$; $R^2_{adj} = 1 - MSError/MSTotal$

On HW 2, we saw that the coefficient of DML changed slightly when fMONTH was added to the model, and the ordering of the months differed when DML was in the model. With multiple regression, we *always* have to interpret the slope coefficients conditional on the other variables in the model (e.g., the "effect" of DML after adjusting for MONTH). But what does that mean?

Example 1: 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) Load the data into the Multiple Variables applet. Drag *salary* to the Response box and *semesters* to the Explanatory box. Summarize the nature of the association. Is it what you expected? It looks like overall increasing semesters corresponds to increasing salary

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

(c) Check and interpret the R-squared box.
33.1% of the variation in salaries is explained by the number of semesters

(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?
71% of the variation in salaries is explained by major. Validity conditions look fine (including normality of salaries in each major). Means are 56.38, 49.00, 38.50

- Write out the estimated model equation using *indicator coding* with chemistry as the reference group.
  predicted salary = 56.38 -17.87(business) - 7.37(physicis)

- Write out the estimated model equation using effect coding. (*Hint*: Need more info…)
  overall mean = 47.96
  predicted salary = 47.96 - 9.46(business) + 1.04 (physics)

(e) Drag *major* to the explanatory variable box. Check the box for **Statistical model** and confirm your answers for both types of coding.

Prediction: How much variation in salaries will be explained by a model that includes both of these variables?
more than 77% but less than 71% + 33 %

(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*?

each has a negative assocaition

(g) Suppose I want to "subtract off" the "major effects" to see the relationship between salary and semesters without the noise of some majors tend to spend more semesters in school then others. In other words, I want to look at the relationship between salary and semesters as if everyone was in the same major. How will I adjust the salaries of business students to put them more on par with the physics students?

move down the salaries of chemistry majors and move up the salaries of business majors

(h) How will I adjust the semesters of business students to put them more on par with the physics students?

move business majors to the right and chemistry majors to the left

(i) Check the Adjust *y* values box and the Adjust *x* values box to confirm. Check the Show equation box. Report

Original slope between salary and semesters: 1.156

Major-adjusted slope between salary and semesters: -2.169

Key Idea: An *added-variable plot* is a useful graphical tool to display the adjusted association. Most software packages construct the plot by regressing y on $x_1, \ldots, x_q$ and regressing x_{q+1} on $x_1, \ldots, x_q$ and plotting these residuals against each other. The slope of this line matches the slope matches the slope of x_{q+1} in the multiple regression model!

(j) What will this graph look like if *salary* is perfectly explained by *major*? What if *semesters* is perfectly explained by *major*?

flat horizontal line at "y" (residuals of salaries vs. major) = 0

veritical line at "x" (residuals of semesters vs. major) = 0

(k) Complete these statements

Between majors, the association between semesters and salary is positive

Within majors, the association between semesters and salary is negative

(l) Change the order of the explanatory variables so *major* comes after *semester*. This shows the original 'scatterplot' between *salary* and *major* color-coded by *semester*. How will the salaries change if I adjust for the positive association between salary and semesters? How to adjust the categorical predictor variable is a little trickier. If there were just two categories, we would regress the dummy variable on semesters, and plot these residuals on the horizontal axis. A *leverage plot* finds the residuals from regressing y on $x_1, \ldots, x_q$, storing those residuals, and then regressing each of the dummy variables on $x_1, \ldots, x_q$ and then

move down the white/blue dots and move up the dark red dots.

regressing the first residuals on this matrix of residuals, and then plots the first set of residuals against these fitted values. The main advantage is you get one plot, and you look for the same features – is there a strong association between the adjusted variables? Also look for outliers, influential observations etc. (before just blindly adding major to the model).

(m) Use the applet to adjust the values and describe what you see!

a strong negative association, telling us that major will be useful to add to a model of sal vs. sem

(n) What is the R^2 value for this model? How does it compare to your earlier prediction?

.905

(o) 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			

how many variation is explained by the model (major + sem)
<<contributed of semesters after adjusting for major
<<contributed of major after adjusting for semesters

The pie chart starts with one variable (e.g., SSprev = 600.1) is variation explained by semester (alone) and the blue slice is the additional variation explained by adding major to this model (1043)

(p) 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?

Probably some "within group" agreement

Definition: The *intraclass correlation coefficient (ICC)* is a measure of how correlated two responses are from individuals in the same "class." It measures the degree of "sameness" of individuals in the same group vs. across groups. The most traditional application is as a measure of "reliability" of repeat observations.

One way to measure ICC also compares the between group variation to the within group variation. If the observations within a group are more similar than observations between groups, we don't expect these two "variations" to be the same.

$$possible\ ICC = \frac{(n-1)SSbetween - SSwithin}{(n-1)(SSbetween + SSwithin)}$$

(q) Compute and interpret this value from the ANOVA table with just major

Source	df	SS	MS
major	2	1291.08	645.54
Error	21	523.88	24.95
Total	23	1814.96	

n = 8 in each major

$$\frac{7(1291.08) - 523.88}{7(1291.08 + 523.8)} = 0.67$$

(r) How do I estimate the correlation of the salaries of two students in the same major?

We could make all the possible pairs of students in the same major, repeat this for each major, and then find the correlation coefficient for these two columns... we will get the above value!

Another way to calculate ICC takes the "degrees of freedom" into account.

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

($MSbetween$ is an unbiased estimator of $n\sigma_g^2 + \sigma_\epsilon^2$ and $MSerror$ is unbiased estimator of σ_ϵ^2.)

(s) Compute this value.

645.54 - 24.95)/(645.54 + 7*24.95) = 0.76

In R, you can use

- ICC package ICC:ICCbare(y=saldata$salary, x = saldata$major)
- Multilevel package multilevel::ICC1(aov(saldata$salary ~ saldata$major))

Example 2: Caffeine is widely used as a stimulant – but are there other ways to get the same effects, with little to no downside? To begin to answer this question, a study compared the effects of caffeine with theobromine, which is the active chemical naturally found in chocolate and is an alkaloid with a similar molecular structure and effects on people as caffeine (Scott & Chen, 1944). To measure the effects of these two different chemicals, the researchers trained subjects to tap their fingers in such a way that the rate could be measured. After learning/practicing this type of finger tapping, participants took either took a caffeine pill (200 mg), a theobromine pill (200 mg), or a placebo, and then their finger tapping rate was measured two hours later.

Consider a one-way ANOVA on the stimulants:

```
summary(aov(Taps ~ Stimulant))
```

(a) Is the difference among the stimulants statistically significant? (Be clear how you are deciding.)

No, F = .675 < 1 and p-value = 0.533 > 0.05

(b) Is there significant person to person variation?

```
#using effect coding
summary(modelB <- lm(Taps ~ participant, contrasts = list(participant = contr.sum)))
```

Yes, F = 12.13 (df = 3, 8) and p-value = .002399

(c) Is it reasonable to consider the observations in this study independent from each other? What might the variance-covariance matrix of the residuals look like?

Repeated observations on same person (even under different conditions) are correlated

```
anova(modelB)
```

(d) Calculate and interpret the intraclass correlation for the subjects in the stimulant study.

> (1826 - 150.5)/(1826 + 2*(150.5)) = 0.7877292; this is the correlation between any two observations for the same subject (subjects that are high /above average on one treatment tend to be high on the other treatment and vice versa)

Notes:

- The Pearson correlation coefficient measures the strength of the linear association between two variables. Whereas the intraclass correlation coefficient measures the amount of agreement of pairs of observations in the same group.