Unit 3 Assignment (MATH ACT Scores)

Ready to discuss at Live Session #3

# The Data

Two Way ANOVA is a special case of multiple regression when there are exactly 2 categorical explanatory variables. This data set is a random sample of ACT score from a group of high school students. Below is a summary

Data set on 861 ACT Assessment Mathematics Usage Test scores from 1987. The test was given to a sample of high school seniors who met one of three profiles of high school mathematics course work:

1. Algebra I only
2. Two Algebra courses and Geometry
3. Two Algebra courses, Geometry, Trigonometry, Advanced Mathematics and Beginning Calculus

Unlike multiple regression discussed last week, 2-way ANOVA tends to focus on hypothesis testing rather than prediction. The question we are trying to ask ourselves for this data set is if there are changes in the overall performance of high school student ACT score based on their gender, their mathematical background, or if a certain combination of the two produces better scores than others.

# Discussion #1

The data set is provided in MathACT.csv along with some SAS code. The convenient thing about two-way ANOVA (once decided that’s what we have) is that there is no need for model selection-type things as we know exactly what predictors we have to work with and that there are only two. This also helps as we do not have to worry as much about multicollinearity issues that we discussed before. The SAS code has 3 main components:

1. Read in the data and create some summary statistics tables
2. Create a profile plot of the summary statistics.
3. Run a two-way ANOVA model.   
   *The plot is a little bit complicated to build in SAS but it’s very helpful and you can use it as a template to plot other data in a similar fashion so it will help to play around with that on your own.*

There are a few times throughout the videos where the discussion of additive and nonadditive models (there is a good section in the book on this). The best way to visually investigate whether which one is more appropriate is through a simple visualization of the mean ACT scores by the different groups of the factors.

## Question 1

Using the plot as a guide and your knowledge from the videos and text, what type of model (additive or nonadditive) do you think most appropriately fits this data set? Be prepared to explain your choice.

**Since the lines are not quite parallel, I think that a nonadditive model most appropriately fits this data set.**

## Question 2

My SAS code fits a nonadditive model. You can tell it is nonadditive because my model statement has a Sex\*Background term. Examine the various F-tests tables, specifically the one that has “Type III SS” in the tables. What are each of the three tests telling us? How do these tests correspond back to our visualization plot?

**Type I SS evaluates the variables in order, which may mistakenly identify the most important variable. Type III SS does not evaluate the variables sequentially, so order doesn’t matter. In comparing MatchACT and MatchACT\_2, we can see a much more pronounced difference in the effect on males versus females.**

\*\*\*To help you answer the last question I’ve provided an additional data set MatchACT\_2.csv. Run the code again on this new data set and compare the differences between the resulting F-tables. This should help you see what’s going on.

# Discussion #2

In addition to the general F-tests, I’ve included some additional options with the lsmeans and estimate statement. These options are used to write specific contrasts to conduct hypothesis tests that are a little more specific than what F-tests tell us. Take a look at the output and find the Estimates table. There are three different tests that I asked SAS to do. For the first one, I tell you exactly what I’m testing for, I want you to see if you can decipher what the other two are testing for. Besides the estimate table and the SAS code, there is another table within the output (LSMean tables) that could possibly help so be sure to take a look at the other output results (note: I’m referring here to an actual number table, not some of the graphics, the graphics can be a little confusing). You only need to use the first data set for this discussion.

**I think the “What do you think” row is C vs. A, but I’m not sure about “What do you think2?” It doesn’t seem to show up in the least squares means tables…is it gender — male vs. female?**