Stat 414 - Day 7
Adjusted associations and Interactions

Last Time:

- Testing the significance of regression coefficients
  - t-tests for individual terms, after adjusting for all other terms in the model
  - partial F-tests for groups of terms, after adjusting for all other terms in the model
    - Know what different "anova" commands, output are doing
  - overall F-test for all slope coefficients at once

Example 1: Forced expiratory volume and smoking

Data were collected on 654 youths in the area of East Boston during the middle to late 1970s. The youth in the study were of ages 3 to 19 years, an age period during which much physical development, such as increase in lung capacity, takes place. The objective was to analyze the relationship between smoking status, and forced expiratory volume (FEV, measured in liters). (FEV is a measure of strength of a person's lungs – the maximum volume of air a person can blow out in the first second; higher numbers are better/heathier lungs)
Load in the data and explore the "smoker" variable

```
contrasts(factor(FEVdata$Smoker)) #see how R will code the variable
iscamsummary(FEVdata$FEV, FEVdata$Smoker)
```

(a) How will the Smoker variable be coded?

(b) You have enough information to write out the fitted regression model using indicator coding.

(c) Fit a linear regression model to predict FEV from the smoker variable and confirm your equation.

```
model1 = lm(FEV ~ Smoker, data=FEVdata); summary(model1)
```

(d) Is the smoker variable statistically significant? How are you deciding?

Note: The previous test is equivalent to a one-way ANOVA or two-sample (pooled) t-test.

(e) Is the model valid?

Note: The third graph (Scale-Location) is also used to explore the equal variance assumption. It plots the square root of the (abolute value of standardized) residuals. Like residuals vs. fitted values, you don't want to see changes in the variation at different fitted values. You would also like the red smoother to be approximately horizontal. Otherwise it tells you that the average magnitude of the residuals is changing with the fitted values (e.g., like a linear relationship…)

Sometimes patterns in the residual plots reveal more than nonlinearity or unequal variance. For example, you can see whether a pattern in the residuals appears to be related to another variable not currently in the model.

(f) What do you learn from a graph of residuals vs. age?

(g) Add Age to the model. Has the coefficient of Smoker changed? Is Smoker still significant? How do we now interpret the coefficient of Smoker?

(h) What does this model look like?

Including the binary variable allows the intercepts to differ, but we are still assuming the slopes are the same. Produce a graph to decide whether there is evidence that the relationship between FEV and age differs for the smokers and nonsmokers. (See RMarkdown)

(i) What do you learn?

Definition: A quantitative variable and a categorical variable $interact$ if the slopes of the regression lines differ. (After all, it's the slope that tells us about the association between the two variables, so this says the association between the response and the quantitative variable depends on the category of the categorical variable.) To include an interaction between x_1 and x_2 in the model, we literally multiply x_1 and x_2 together and add this variable to the model.

(j) What does it mean to multiply Smoker and Age (one categorical and one quantitative variable) together?

Add the interaction to the model

```
#You can make R do the multiplication for you by including Smoker:Age with a colon
to signify an interaction
model3 = lm(FEV ~ Smoker + Age + Smoker:Age, data = FEVdata)
```

(k) Write out the full equation and then write out the equation (FEV vs. age) for the smokers and the non-smokers.

(l) What does the sign of the interaction term tell you? (Note: Another way to interpret this interaction - what is the slope of age in the full equation?)

(m) Is this model valid?

(n) What about multicollinearity?

(o) By the way, why are VIF values more informative than looking at pairwise correlation coefficients among the explanatory variables?

Because an interaction is a "product," centering the quantitative variable might help with the multicollinearity.

(p) After creating and using Age.c: did we improve the multicollinearity?

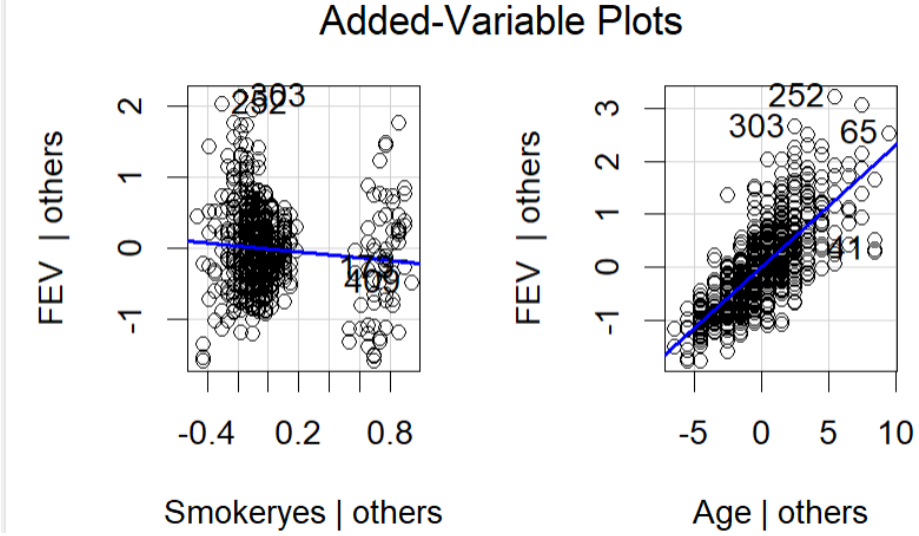
(q) How do we interpret the coefficient of smoker in this model? (*Hint*: Can you make the interaction go away in order to interpret the main effect?)

Notes:

- Indicator variables change intercepts; Interaction terms change slopes.
- *Always* best to use graphs to help illustrate an interaction.
- Centering to remove multicollinearity doesn't work on all pairs of variables, just "products" like quadratic and interaction.
- When center with interaction, the interpretation of the "main effect" is about the change in response when the other variable is at its mean (to "zero out" the interaction).

In this case, the *adjusted association* is much different from the *unadjusted association.* You can get a graph of the adjusted association, for example an *added variable plot*.

```
car::avPlots(model2)
```



The graph on the left shows the adjusted association between smoking status and FEV after adjusting for age. Meaning, this is the predicted difference in FEVs between smokers and non-smokers *of the same age*. (There are just the two smoking values, but the points have been "jittered" to better show the distributions.)

The graph on the right shows the adjusted association between FEV and age after adjusting for smoking status, meaning we are assuming the association is the same for smokers as for non-smokers and this is a graph of that "common" association. The steepness of this line compared to the remaining unexplained variation corresponds to the statistical significance of that adjusted association.

If these adjusted associations are not significant, then that says that variable would not be worth adding to the model that doesn't include it.

Practice: in-state tuition vs. pct faculty with PhD, coloring by public vs. private school (1995 USNews & World Reports)