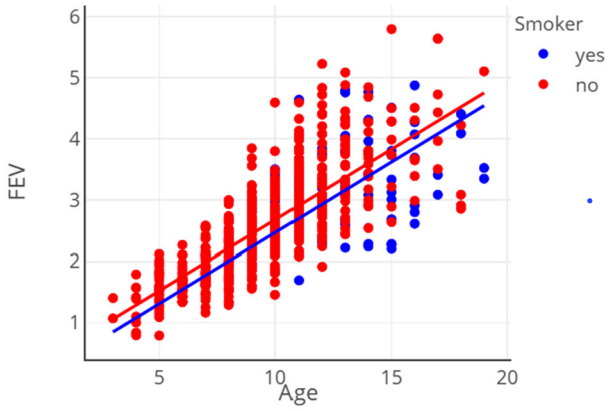
Stat 414 - Day 10
Interactions

Example 1: Forced expiratory volume revisited Recall the data collected on 654 youths (3-19 yrs) in the area of East Boston during the middle to late 1970s. FEV (measured in liters) is a measure of the strength of a person's lengths – the maximum volume of air a person can blow out in the first second; higher numbers are better/heathier lungs). Previously we found that smokers, after adjusting for age, tended to have lower FEV values than non-smokers.

Results



☑ Show equation

yes: *predicted FEV = 0.1584 + 0.2306 × Age*
no: *predicted FEV = 0.3674 + 0.2306 × Age*

Statistical model: ☑ Effect coding

Term	Coeff	SE	t-stat	p-value
Intercept	0.2629	0.1013	2.59	0.0097
Smoker				
yes	-0.1045	0.0404	-2.59	0.0099
no	0.1045			
Age	0.2306	0.0082	28.18	< 0.0001

(a) How do we interpret the intercept, coefficient of smoker, and coefficient of age in this model?

.23 = predicted per year increase in FEV after adjusting for smoker status

-.1045 = predict smokers' avg FEV is .1045 below average comparing people of same age

(b) What pattern do you see in the residuals that might suggest there is a better model out there?

non smokers have a larger slope than nonsmokers

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.

(c) 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.

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

(0, 1) × Age = 0, ..., 0, Age, ... Age
(1, −1) × Age = −Age, ... −Age, Age, ..., Age

Add the interaction to the model

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

FE

(f) How do we interpret the intercept? How do we interpret the coefficient of Age?

(g) 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?)

(h) Is this model valid?

(i) Are you surprised there is some multicollinearity? What could we do about it?

(j) BTW, 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.
(k) Did we improve the multicollinearity?

(l) How do we interpret the intercept, coefficient of age, and coefficient of smoker in this model? (Hint: Can you make the interaction go away in order to interpret the main effect?)

(m) Did adding the interaction help our unequal variance problem? Could it have?

Computer Problem 10: Computer problem 10: The file HarrisBank.txt contains data on 93 employees of Harris Bank Chicago in 1977 (being investigated for discrimination). Variables include beginning salaries in dollars, years of schooling at time of hire, and number of months of previous work experience.
(a) Fit a model to predict sal77 from education (quantitative) and experience (quantitative) and the interaction. Write out the full regression model.

Interpreting a quantitative x quantitative interaction is not as easy. The main idea is to show how the association between y and x_1 changes depending on the value of x_2.
Interpret the sign of the coefficient:
(b) What do the signs of the coefficients seem to tell you?

Express the slope of one variable in terms of the other variable:
(c) What is the slope of education in the full regression model?

Give a brief table of values
(d) What is the equation for predicting salary from experience when education = 8?
(e) What is the equation for predicting salary from experience when education = 16?

Graphs are often the best bet.
(f) Pick and include your favorite of these graphs. Interpret the interaction in context using your graph. (Be clear whether you are looking at a graph of the data or a graph of the model!

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).

Example 2: Beach data revisited Recall our Beach data. We allowed the Beaches to have random intercepts.
(a) Do you see any problems with the model?

To allow the slopes to vary across the Level 2 units in the model equation, we add a j index to the slope too. Level 1 equation: $y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \epsilon_{ij}$
where $\beta_{0j} = \beta_{00} + u_{0j}$ and $u_{0j} \sim N(0, \tau_0^2)$ and …
(c) Write out the second level 2 equation.

(d) Now create the composite equation.

(e) Give the expression for beach j's intercept. Give the expression for beach j's slope.

(f) What assumptions are we making about the distribution of the random slopes?

Fit the random slopes (or "random coefficients") model:

```
model2 = lmer(Richness ~ NAP + (1 + NAP | Beach), data = rikzdata, REML = FALSE)
# you get a warning (not an error) and can ignore it
```

(g) How many parameters does this add to the model? (What if beach was a fixed effect?) What do these new parameter estimate(s) tell you?

(h) Based on this model, what is the equation for Beach 1? What is the equation for Beach 9?

(i) Are the differences in the slopes statistically significant? (State hypotheses, df, test statistic, p-value, conclusion in context.)

Notes:

- In a random slopes model, be careful with the interpretation of the intercept variance and the intercept-by-slope covariance, they assume x = 0. Another reason why it's always good practice to center your explanatory variables so the intercept is meaningful.
- As you can see, randomly slopes can improve the complexity of the model pretty quickly, so you should have empirical, or better yet theoretical, reasons for doing so.