David Youngberg
BSAD 210—Montgomery College

LECTURE 31: INTERPRETING MULTIVARIABLE REGRESSIONS

I. More Output from Excel
   a. *Regression Sum of Squares (SSR)*—the squared difference between the average and the predicted value of the dependent variable. This difference is taken for each observation and then added together.
   b. *Error Sum of Squares (SSE)*—We've discussed this before.
   c. *Total Sum of Squares (SST)*—SSE + SSR
   d. R^2—SSR/SST, or the percent of deviation that our regression explains. There is no threshold for a "good" R^2.
      i. This is sometimes also called the "coefficient of determination."
   e. *Adjusted R^2*—The R^2 value adjusted for the number of explanatory variables.
      i. A weakness of R^2 is that it adding additional explanatory variables causes it to increase, regardless of the quality of explanatory variables. This is a problem because having many explanations for something is the same as having few.
      ii. Adjusted R^2 penalizes the researcher for adding explanations, especially if it's large relative to the number of observations. The equation is:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

      Where n is the number of observations and k is the number of explanatory variables, excluding the intercept.
   f. *F-statistic*—The ratio between the explained and unexplained variance. It's equation is:

$$F = \frac{SSR}{SSE/n-2}$$

      i. Higher values of F indicate a model with more explanatory power. Because the shape of the F distribution is known (its exact shape changes based on the number of observations and number of explanatory variables), it is possible to determine critical values.

ii. The *p-value* does this work for you: if it is smaller than, say, 0.05, the model is significant with 95% confidence. If it is smaller than 0.01, the model is significant with 99% confidence.

g. *p-value*—each variable will give you a p-value, which is a summary of significance (it will also report the t-value but significance depends on that and on the degrees of freedom). The smaller the value, the better the threshold you achieve (if less than 0.05, it's significant to the 5% level; less than 0.01 and it's significant to the 1% level).

II. Word of Caution

a. Be wary of predicting values outside the range of your data.
   i. For example, suppose you're using age to predict height (as we did last class). Suppose the line of best fit is $HEIGHT_i = 80 + 5.6\ AGE_i + \varepsilon_i$. If you predicted the height of someone with an age of 50, you'd get 360 inches, or 30 feet tall. That doesn't make sense.
   ii. You got this result because the data for age ranged from 4 to 12. If people really did just keep growing at the same rate, your analysis would be spot on. But in reality they typically stop growing in their mid-to-late teens.

b. Recall the key thing to understand about regressions is that they are making a causal claim.
   i. You are claiming your Xs cause Y. Not the other way around.
   ii. Thus when you change one X, Y changes by β. The only way Y changes in your model is if X changes independently (hence the name, independent variable).

c. In a multivariable regression, changing one variable means you're holding other variables constant.