David Youngberg
BSAD 210—Montgomery College

LECTURE 19: INTERPRETING REGRESSIONS

I. The Punchline
   a. Recall the point of doing regression boils down to figuring out beta, or the magnitude of the effect of X on Y (if there is an effect). If people could take away just one idea from any regression, it's beta.
   b. Thus it's worth including a "punchline,"[1] or a single sentence that interprets beta into everyday language.
   c. Before I give you the general format of a punchline, recall this equation:

$$\Delta Y = \beta_k * (\Delta X_k)$$

      i. This equation will guide us as we explore the punchline. **BURN IT INTO YOUR BRAIN (AGAIN)!**
   d. The structure of the punchline generally follows this format:

For every additional _____ _____ _____, *(amount)* *(unit)* *(x variable)* _____ increases/decreases by ____ _____. *(y variable)* *(choose one)* *(beta)* *(unit)*

   e. Examples from previous classes:
      i. For every additional point of difficulty a professor is rated, their quality rating decreases by 0.86 points.
      ii. For every additional mile from the city center, that home's price decreases by about $23,800.
      iii. For every additional 10 square feet a house is, that home's price increases by about $2,370.
      iv. For every additional pallet added to a shipment, that shipment's time to load increases by 2.3 minutes.

II. Percents and Percentage Points
   a. Sometimes the independent or dependent variable is a value from zero to one. That could be a dummy variable or simply a percent (such as the percent of a population that's below the poverty line).

[1] "Punchline" is how a professor of mine from grad school described what we're about to do. This is a useful way to think about it because it's a takeaway sentence, but it is not an official term.

b. We shouldn't think of increasing or decreasing in terms of percent because percents are nonlinear.
    i. Going from 3% to 4% is a 33% increase; going from 4% to 5% is a 25% increase; going from 5% to 6% is a 20% increase.
c. Thus, we should think in terms of increasing or decreasing in terms of *percentage points*.
    i. Each step of going from 3% to 4% to 5% to 6% is a one *percentage point* increase.
d. Also keep in mind that percents are expressed between 0 and 1. If an observation in Excel is 42%, the number is not 42; it's 0.42.
    i. Thus, a one percentage point increase isn't an increase of 1; it's an increase of 0.01.
    ii. Again, keep in mind this equation:

$$\Delta Y = \beta_k * (\Delta X_k)$$

    iii. If ΔX is 0.01, then ΔY will not be β but rather β/100.
e. Consider the following hypothetical regression (all variables are statistically significant):

$$SALES\ (K) = 50 - 9.4(RIVALS) + 300(\%HOMEOWNERS)$$

    i. This regression predicts weekly sales (in thousands) of a home improvement store location based on the number of rivals in a 20 mile radius and the percent of people in that radius who own their own home.
    ii. A one percentage point increase in homeownership increases weekly sales by $3 thousand (300 times 0.01 results in 3, or $3,000).

III. Dummy Variables and Limits of the Data
a. You can have a dummy variable be the y-variable. When you do, predicted y becomes the chance that the observation will be "yes."
b. Because it's a percent chance of something, beta values have to be thought of in terms of percentage points.
c. Consider the following hypothetical regression (all variables are statistically significant):

$$EMPLOYED? = 0.6 + 0.08(GPA) - 0.12(FEMALE?)$$

i. For example, the predicted value of a male with a 3.0 GPA is 0.84. Such a person has an 84% chance of being employed.
ii. Changing the explanatory variable changes the percentage points. If that same 3.0 student was female, the chance of being employed would fall by 12 percentage points (to 72%).
iii. If you increase a person's GPA by 1 point, the chance that person is employed rises by 8 percentage points. It *does not* rise by 8%.

d. Consider the following hypothetical regression (all variables are statistically significant):

$$\%HAPPY = 1.3 - 0.07(PRICE) - 0.05(TIME)$$

i. This regression predicts customer satisfaction for pizza delivery. Lower prices and faster deliveries make for happier customers.
ii. Like the previous example, it's bounded between zero and one (or 0% and 100%). But notice that the constant is greater than one! How is that possible?
iii. It's because the regression is formed from the data and the data is filled with values that will be, at minimum, pretty large and all will cause the dependent variable to go lower. No pizza will cost zero (or just a few) dollars, no pizza will be delivered in zero (or just a few) minutes, thus the intercept has to be pretty high to compensate for the large negatives that will inevitably happen.
iv. Similarly, if the independent variables were only values that cause %HAPPY to increase, the intercept would likely be negative to compensate.

IV. Forecasting
a. We can use time as an X variable, using time to predict some other value.
b. Open **Data Set 6**. There you will find, under the Real GDP tab, quarterly Gross Domestic Product, in billions, adjusted for inflation and season.[2]
c. Each observation maps to a year and quarter, with quarters indicated with .00 (1st quarter), .25 (2nd quarter), .50 (3rd quarter), and .75 (4th quarter).
i. For example, 2022.50 would be the 3rd quarter in 2022.

[2] Adjusting for inflation means the numbers take into account how much prices are changing. Adjusting for season means the numbers take into account how economic activity fluctuates over the course of the year.

d. By using year and quarter (which we'll call YQ) to predict GDP, we can create a trend line. We should get:

$$GDP = -858388 + 435.2 * YQ$$

   i. Note YQ is statistically significant.
   ii. For every additional year that passes, real GDP should increase by about $435.2 billion.
   iii. For every quarter that passes, real GDP should increase by about $108.8 billion (=$435.2/4).
   iv. If we wanted to predict what real GDP would be in the second quarter of 2025, we would put 2025.25 into our equation and get:

$$GDP = -858388 + 435.2 * 2025.25 = 23{,}000.8$$

      1. Or about $23 trillion.

e. Note we would get a different line if we chose a different data. Suppose, for example, we considered only the observations set after 2020 to be relevant. Then the equation would be:

$$GDP = -967498 + 489.2 * YQ$$

   i. Note YQ is still statistically significant.
   ii. Now the projected GDP is about $23.3 trillion.

V. Word of Caution

a. Be wary of predicting values well 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 = 80 + 5.6 AGE. 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.
   iii. Similarly, the %HAPPY regression starts with 1.3 because variables like PRICE and TIME will always be far greater than one. CHEESE, in contrast, will probably not be greater than 2.
   iv. Or consider the time series forecasting. Forecasting a few years in advance is reasonable. Forecasting fifty years in advance is not.

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. 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).