David Youngberg
BSAD 210—Montgomery College

LECTURE 19: UNDERSTANDING REGRESSIONS

I. Adjusting for population
   a. Sometimes you get raw numbers for data and those numbers aren't useful in that form.
   b. A common correction is to adjust for population, or *per capita*.[1]
      i. For example, you can't use GDP to see which people are wealthier. China has the world's second highest GDP but its people are not the second wealthiest in the world. Its GDP is high because, in part, its population is high.
      ii. Divide the GDP for a country by the total population of that country. This gives you GDP per capita.
   c. Any variable that should be directly influenced by population should be adjusted for population; values like latitude and percent forest cover shouldn't be adjusted for population.

II. Scalars
   a. A *scalar* is a constant value you can use to simplify regressions interpretation.
      i. If you multiply an independent variable by a scalar, the beta-value will change, but the statistical significance will not. Other betas won't change either.
      ii. Thus you can use scalars to aid interpretation.
   b. Suppose you're interested in murders in various states. Total number of murders isn't good enough—large states will have more murders than small states—so you want to adjust for population.
      i. Murders per capita is a good start, but it's an awkward number. In 2012, the Alabama's murders per capita was 0.000071.
      ii. Why so small? Because this is murders *per person*. A rate of 0.5 would mean half the population is being murdered!
   c. This is why rare events have a scalar. The values are multiplied by 1,000 (births) or 100,000 (crime) to make the values readable. Alabama's murder rate is 7.1 murders per 100,000 people.
   d. Imagine you didn't do this and you ran a regression with murders predicting unemployment (perhaps because if a state gets more dangerous, it will be hard to do business and to shop so the unemployment rate will go up). You'd get:

[1] Capita is Latin for head. It's where we get the word, capital, or head of government, from.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	6.05414313	0.473254678	12.792569	3.0974E-17	5.103102266	7.005183995	5.103102266	7.005183995
Murder and nonnegli	29633.62731	9408.043501	3.1498183	0.002783092	10727.45641	48539.7982	10727.45641	48539.7982

i. First, note it's statistically significant.
ii. Second, look at the coefficient. For every additional murder per person, the unemployment rate goes up by 29,633.6 percentage points. That's hard to wrap your mind around.
iii. So let's do the same thing, but with murders per person now murders per 100,000 people.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	6.05414313	0.473254678	12.792569	3.0974E-17	5.103102266	7.005183995	5.103102266	7.005183995
Murder and nonnegli	0.296336273	0.094080435	3.1498183	0.002783092	0.107274564	0.485397982	0.107274564	0.485397982

iv. Note the P-value is exactly the same but the coefficient is much easier to interpret. For every additional murder per 100,000 people, the unemployment rate increases by 0.29 percentage points.

e. Mathematically, this is what's happening:

$$UNEMPLOY_i = \beta_0 + \beta_1 MURDERS_i + \varepsilon_i$$

$$UNEMPLOY_i = \beta_0 + \left(\frac{100{,}000}{100{,}000}\right)\beta_1 MURDERS_i + \varepsilon_i$$

$$UNEMPLOY_i = \beta_0 + \left(\frac{1}{100{,}000}\right)\beta_1 MURDERS_i(100{,}000) + \varepsilon_i$$

$$UNEMPLOY_i = \beta_0 + \left(\frac{\beta_1}{100{,}000}\right) MURDERS_i(100{,}000) + \varepsilon_i$$

i. $MURDERS_i(100{,}000)$ is your new variable so β_1 *must* be divided by 100,000. It's how the equation balances.
ii. If instead you decreased the independent variable (say, you changed watts used per person to kilowatts used per person), β would increase.

f. And if you change the independent variable (perhaps unemployment causes murders):

$$MURDERS_i = \beta_0 + \beta_1 UNEMPLOY_i + \varepsilon_i$$

$$\left(\frac{100{,}000}{100{,}000}\right) MURDERS_i = \beta_0 + \beta_1 UNEMPLOY_i + \varepsilon_i$$

$$MURDERS_i(100{,}000) = (100{,}000)(\beta_0 + \beta_1 UNEMPLOY_i + \varepsilon_i)$$

$$MURDERS_i(100{,}000) = (100{,}000)\beta_0 + (100{,}000)\beta_1 UNEMPLOY_i + (100{,}000)\varepsilon_i$$

i. Each β adjusts to the same degree and in the same direction as how the independent variable was adjusted.

III. Predicting & Interpreting

a. There are two kinds of prediction you can make with a regression line.
   i. What the dependent variable would be if you put in various values for the explanatory variables.
   ii. How much you can expect the dependent variable to change if you change one explanatory variable.

b. Consider the regression line we did earlier:

$$PRICE = 79 + 237(SQFT) - 23{,}792(MILES)$$

c. Suppose we wanted a house that was in bad shape so we could buy it, renovate it, and sell it for a profit (called "flipping"). Such a house would be going for less than what we would expect given its location and square footage. If we have both variables, we can predict what the value of the house would be if it was in typical shape.
   i. Imagine a house had 2,000 square feet and was 3 miles from the city center. We can expect that house to go for:

$$PRICE = 79 + 237(2{,}000) - 23{,}792(3)$$
$$PRICE = 79 + 474{,}000 - 71{,}376$$
$$PRICE = 402{,}703$$

   ii. So if we see a price for $200,000 or $300,000, we can infer either (a) the seller isn't asking enough for that house or (b) the house is in really terrible shape. Either way, it's a candidate to invest in.

d. Suppose you wanted to know how much cheaper a house could be if it was farther from the city. Since we're changing an explanatory variable (in this case, increasing the MILES variable), we only need to look at the coefficient for MILES to answer the question; the rest of the equation doesn't matter.
    i. Here's the proof. Suppose $MILES_N$ is the miles from city center we are considering now and $MILES_B$ is the miles from city center we were considering. $PRICE_N$ and $PRICE_B$ are the price now and the price before, respectively. Thus we have two equations:

$$PRICE_N = 79 + 237(SQFT) - 23{,}792(MILES_N)$$
$$PRICE_B = 79 + 237(SQFT) - 23{,}792(MILES_B)$$

    ii. We are curious how much the price changed. We want to know what this is:

$$PRICE_N - PRICE_B = ?$$

    iii. Let's put in the equations from before and do some algebra.

$$[79 + 237(SQFT) - 23{,}792(MILES_N)] - [79 + 237(SQFT) - 23{,}792(MILES_B)]$$
$$79 + 237(SQFT) - 23{,}792(MILES_N) - 79 - 237(SQFT) + 23{,}792(MILES_B)$$
$$79 - 79 + 237(SQFT) - 237(SQFT) - 23{,}792(MILES_N) + 23{,}792(MILES_B)$$
$$-23{,}792(MILES_N) + 23{,}792(MILES_B)$$
$$-23{,}792(MILES_N - MILES_B)$$
$$-23{,}792(\Delta MILES)$$

    iv. The Δ symbol is delta and stands for change. Getting one mile farther from the city center drops the price by $23,792.
    v. Going an additional 1.5 miles farther out drops the price by $35,688 ($23,792 times 1.5).
    vi. Going an additional 2.7 miles farther out drops the price by $64,238.4 ($23,792 times 2.7).
    vii. Note all of this keeps the size of the house the same. All other coefficients don't matter because all other variables are held constant.

IV. Percentage Points

a. Sometimes the explanatory 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).
b. We shouldn't think of increasing or decreasing in terms of percent. We should think in terms of increasing or decreasing in terms of percentage points.
    i. Imagine unemployment is 6%. If it increases to 9%, how much did it increase?
    ii. If you said "3%", you'd be wrong because a 3% increase would be 6.18% (3% of 6 is 0.18). This is a 50% increase.
    iii. Instead, we can say it increased by three percentage points.
c. 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. If the percent of people who own their home increases from 35% to 35% (0.35 to 0.36), then it increased by one percentage point, ***not 1%***.
    iii. A one percentage point increase in homeownership increases weekly sales by $3 thousand (300 times 0.01 results in 3, or $3,000).
d. Consider the following hypothetical regression (all variables are statistically significant):

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

    i. This regression uses a dummy variable as the dependent variable. Thus, all coefficients involve moving the dependent variable closer to either zero or one.
    ii. For any predicted value, you'll get a number between zero and one; this should be interpreted as a percent chance. 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.

iii. 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%.
iv. 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%.

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

$$\%HAPPY = 1.3 - 0.07(PRICE) - 0.05(TIME) + 0.16(CHEESE)$$

i. This regression predicts customer satisfaction for pizza delivery. Lower prices, faster deliveries, and more cheese 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?

V. 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.
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.

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).