David Youngberg
BSAD 210—Montgomery College

LECTURE 28: UNDERSTANDING REGRESSIONS II

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. 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).