David Youngberg

Acct/Busa/Econ 222—Bethany College

**Lecture 09: Regression Organization & Interpretation**

1. Causality
	1. Recall the key thing to understand about regressions is that they are making a causal claim.
		1. You are claiming your Xs cause Y. Not the other way around.
		2. 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).
		3. It’s also important to argue, in your paper, why the causality is set up as it is. Does internet use cause income or does income cause internet use?
	2. In a multivariable regression, changing one variable means you’re holding other variables constant.
	3. Consider the following generic regression:

$$Y=β\_{0}+β\_{1}X\_{1}+β\_{2}X\_{2}+ε$$

* 1. Suppose we want to see how Y changes when X2 changes. Let superscripts “1” and “2” denote which variable is before (1) and which is after (2). Note since we are only changing X2, most things do not change.

$$Y^{1}-Y^{2}=\left(β\_{0}+β\_{1}X\_{1}+β\_{2}X\_{2}^{1}+ε\right)-(β\_{0}+β\_{1}X\_{1}+β\_{2}X\_{2}^{2}+ε)$$

$$Y^{1}-Y^{2}=β\_{0}-β\_{0}+β\_{1}X\_{1}-β\_{1}X\_{1}+β\_{2}X\_{2}^{1}-β\_{2}X\_{2}^{2}+ε-ε$$

$$Y^{1}-Y^{2}=β\_{2}X\_{2}^{1}-β\_{2}X\_{2}^{2}$$

$$Y^{1}-Y^{2}=β\_{2}(X\_{2}^{1}-X\_{2}^{2})$$

$$∆Y=β\_{2}(∆X\_{2})$$

* 1. The impact of X is determined by its coefficient; nothing else.
	2. Why am I making such a big deal about this? Because every empirical analysis should have a “punch line:” a single sentence which summarizes the result of the regression.
1. Adjusting for population
	1. Sometimes you get raw numbers for data and those numbers aren’t useful in that form.
	2. A common correction is to adjust for population, or *per capita*.
		1. 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.
		2. Divide the GDP for a country by the total population of that country. This gives you GDP per capita.
	3. 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.
2. Scalars
	1. A *scalar* is a constant value you can use to simplify regressions interpretation.
		1. 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.
		2. Thus you can use scalars to aid interpretation.
	2. Suppose you are looking at the variable Births Per Capita. For every person in a country, how many additional people are born?
	3. Unless the population is having a lot of sex, this value will way less than 1. Births Per Capita might be 0.02. That’s an awkward number to work with.
	4. If you use it as an explanatory (independent) variable, your beta-value will large to balance out the very small averages. Suppose you are using Births Per Capita to predict population density and your beta-value is 500.
	5. Suppose you multiply Births Per Capita by 1,000.
		1. That means for every 1,000 people, (0.02)(1,000)=20 people are born on average.
		2. In your regression, your beta-value will be 1/1,000 of what it was: it will now be 0.5.
	6. Why does the average increase but the beta-value fall?
		1. The average increased because you have to adjust for the fact that you are looking at 1,000 people at a time rather than one.
		2. Because the average increased, the beta must decrease to adjust (since the average times the beta must result in the same product).
3. Natural log
	1. Suppose you don’t think the relationship between variables is linear. For example, income should matter a lot for life expectancy when you are poor but its effect should taper off as you become wealthier.
		1. Natural log is a good way to approach this problem.
	2. What is natural log? It is a logarithm with a base of e (the number).
		1. We use it when the raw data have a strong positive skew.
		2. We sometimes use it in econometrics to help combat heteroscedasticity.
		3. We use it when we want the independent variable to have a percent change influence on the dependent variable (see below).
	3. We can transform any independent variable into a natural log version.
	4. Natural logs change how we interpret the regression. Now we think in terms of percent increases rather than raw numbers.
	5. If just X is logarithmic…
		1. Add one to the percent increase in X (in decimal form) and then multiply the result by β.

$$∆Y=(1+\%∆X)β$$

* + 1. Suppose β=0.2 for a logarithmic variable, X. If X increases by 1%, then: (1.01)0.2 = 0.202; Y increases by 0.202.
	1. If just Y is logarithmic…
		1. Take e (the number) to the power of β and then subtract one to find (in decimal form) the percent change of Y for every one unit increase in X. If you want X to change by more than 1, multiply that change in X by β first before adjusting with e.

$$\%∆Y=e^{β(∆X)}-1$$

* + 1. Suppose β=0.2 and the dependent variable, Y, is logarithmic. If X increased by 1 then: e0.2(1) – 1 = 1.221 – 1 = 0.221; Y increases by 22.1% for each 1 unit increase in X.
	1. If both X and Y are logarithmic…
		1. Suppose X increases by 1%; how much will Y increase?
		2. Add one to the decimal form of 1%, or 1.01. Take this to the power of β and then subtract 1 from the result. This is the percent change (in decimal form) of Y.

$$\%∆Y=\left(1+\%∆X\right)^{β}-1$$

* + 1. Suppose β=0.2 for a logarithmic variable, X. If X increases by 1%, then: (1.01)0.2 – 1 = 1.00199 – 1 = 0.00199; Y increases by 0.199%.

**Lab Section**

1. Using Math in Excel
	1. Let X be the cell you wish to manipulate.
	2. For scalars, use “=X\*a” where “a” is a scalar
	3. For population adjustment, use “=X/Y” where Y is the cell with the appropriate population
	4. For natural log, use “=LN(X)”
	5. For e, use “=EXP(X)”
	6. When you run your regression, use the manipulated column instead of the original one.
	7. Sometimes, when you’ve changed a column, you don’t want the original one around anymore. But if you remove it, you will lose your references. But if you copy and “Past Values” (HOME>>Paste drop down) you will remove the need to keep your references.
2. Homework

**#1:** In the class dataset, find one variable that should be adjusted for population and then adjust it for population (make it per capita). ***Report*** the variable you adjusted and two averages: per capita and the original (raw) value.

**#2:** Let’s practice interpreting results with the natural log.

* + 1. Suppose you’re estimating the following and β1 = 0.3.

$$Y=β\_{0}+β\_{1}LN(X\_{1})+β\_{2}X\_{2}+ε$$

Interpret the results for X1 (for every 1% increase in X1, how does Y change?).

* + 1. Now suppose you’re estimating the following and β1 = 0.3.

$$LN(Y)=β\_{0}+β\_{1}X\_{1}+β\_{2}X\_{2}+ε$$

Interpret the results for X1 (for every 1 point increase in X1, how does Y change?).

* + 1. Now suppose you’re estimating the following and β1 = 0.3.

$$LN(Y)=β\_{0}+β\_{1}LN(X\_{1})+β\_{2}X\_{2}+ε$$

Interpret the results for X1 (for every 1% increase in X1, how does Y change?).

**#3:** Suppose you’re curious if colder climates are healthier than warmer ones and you want to hold education constant. Run a regression using total children out of school (childoutall) and the ***absolute value*** of latitude (latitude) as independent variables and deaths per 1,00 people (deathsper1000) as the dependent variable. ***Present the regression output and do the following***:

* + 1. Determine which variables, if any, are significant. Interpret the significant coefficients. (For such-and-such change in X, what is the change in Y?)
		2. Explain why it is a good idea to take the absolute value of latitude given what you’re curious about.

**#4:** Suppose someone points out from the previous question you didn’t adjust for population for the number of children out of school. To do this you must first find the population of children who should be in school. Assuming all children between the ages of 5 and 14 should be in school, find the average number of out of school children per child who should be in school. (You are effectively finding the absent rate, in decimal form.) ***You only need to report the old average (without adjusting for population) and the new average (adjusting for population).***

**#5:** Run the regression from the third question again, this time using “percent absent” instead of total absent. ***Present the regression output*** and do the following:

* + 1. Determine the following which variables, if any, are significant. Interpret the significant coefficients. (For such-and-such change in X, what is the change in Y?)
		2. If you used number of students absent per 1,000 students rather than per 1 student, numerically explain how the coefficient would change.