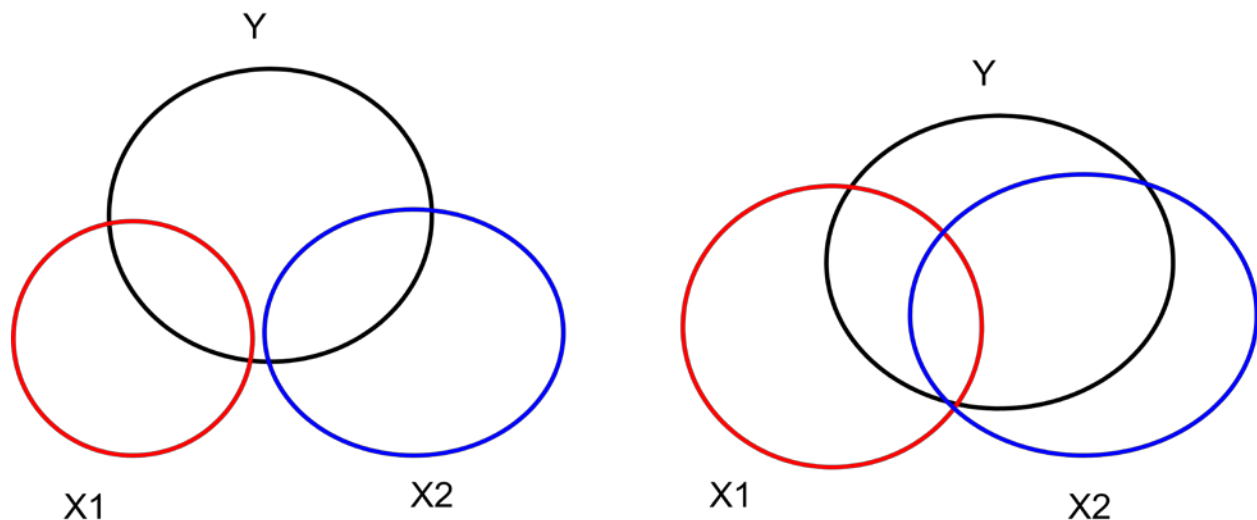


LECTURE 32: BUILDING A MULTIVARIABLE REGRESSION

I. Multicollinearity

- a. If two or more of our explanatory variables are highly correlated, we have multicollinearity.
 - i. It doesn't require perfect correlation—if there's perfect correlation than the program will drop one of the variables.
 - ii. Example: The dummy variable trap, when you have a number of dummies equal to the number of categories. Having both “Male?” and “Female?” will lead to perfect correlation.
- b. Multicollinearity is a problem because the regression will try to get two variables to do the same job. It can easily render both variables insignificant by producing large standard errors.



- i. On the right, there is a portion of variation in Y that can be attributed to either X.
 - c. You can notice multicollinearity if:
 - i. The F-test suggests your model as a whole is strong but none of your variables are significant.
 - ii. The regression coefficients change a lot when a variable is added or deleted.
 - d. The easiest way to correct for multicollinearity is to drop variables.
- ### II. 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*.
 - 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.

III. 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 are looking at the variable Births Per Capita. For every person in a country, how many additional people are born?
- c. 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.
- d. 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.
- e. Suppose you multiply Births Per Capita by 1,000.
 - i. That means for every 1,000 people, $(0.02)(1,000)=20$ people are born on average.
 - ii. In your regression, your beta-value will be 1/1,000 of what it was: it will now be 0.5.
- f. Why does the average increase but the beta-value fall?
 - i. 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.
 - ii. Because the average increased, the beta must decrease to adjust (since the average times the beta must result in the same product).