Section 5: Regression

2/18/20 & 2/19/20

POL 144A: Eastern European Democratization

Isaac Hale

Winter 2020

UCDAVIS
UNIVERSITY OF CALIFORNIA

Outline

1. Regression Crash Course

2. Regression in Excel

What Is Regression Analysis?

- A statistical technique for estimating the relationship between:
  - A dependent variable (“Y”)
  - One or more independent variables (“X”)

- It is not a coincidence that we call these Y and X – you should think of the dependent variable as the y-axis of a graph

- While there are many kinds of regression, we will be using a simple type of linear regression called *Ordinary Least Squares* (“OLS”) in this class

Regression Basics

- When we get a regression, we will get a series of numbers

- Here are the ones you should pay attention to:
  - Coefficients for each of our independent variables (our X's)
  - P-values for each independent variable
  - An "intercept"
  - Number of observations
  - R-squared

- Remember the equation for a line? (think high school math!)

- Y= MX + b

Regression Basics

- The equation for a simple regression is very similar:
  - $Y = a + B_1X_1 + E$
- Y is our dependent variable, “a” is the intercept, “X_1” is our first independent variable, “B_1” is the coefficient for our independent, and “E” is our error term
  - Think of the coefficient like the slope in our line equation (“m”)
  - Why an error term? Regression is an *estimate*, not reality
- If we have multiple independent variables, our regression might look like this:
- $Y = a + B_1X_1 + B_2X_2 + B_3X_3 + E$

A Simple Regression: Theory First!

- Let’s do an imaginary regression together, with made-up data

- Let’s imagine that we’re looking at data on literacy and K-12 education spending for several Eastern European countries across many years

- How might we expect education spending and literacy are related?

- Hypothesis: more education spending *causes* higher literacy

- NOTE: unlike with correlations, regression assumes that X is *causing* Y.
  - You, the researcher, must justify this!

A Simple Regression: Interpretation

- IMPORTANT: how exactly is each variable measured?
  - Let’s say our dependent variable, literacy, is the percent of the population that is literate
  - Let’s imagine that our independent variable, education spending, is millions of dollars spent by the country on K-12 education

- Let’s imagine we run the regression, and we get results like this:
  - Intercept = 50
  - Education Spending B_1 = 1

- What does this mean?
  - When education spending is *zero*, we expect literacy to be 50%
  - For each *million* dollars spent on K-12 education, literacy increases 1%

A Simple Regression: Interpretation

- What would happen if our “X” variable were thousands of dollars, not millions?

- Our Education Spending B_1 would be .001, and our intercept would be unchanged at 50

- What does this mean?
  - When education spending is *zero*, we expect literacy to be 50%
  - For each *thousand* dollars spent on K-12 education, literacy increases .001%

- This is why knowing the measurement of our variables is *critical* for interpreting a regression!

Another Simple Regression Example

- Let’s try another!
- Let’s imagine that we’re looking at data on life expectancy and poverty in Eastern Europe
- What might we hypothesize?
- Hypothesis: higher levels of *poverty* cause lower *life expectancy*
  - What is our dependent variable (Y)?
  - What is our independent variable (X)?

A Simple Regression: Interpretation

- MEASUREMENT
  - Let’s say life expectancy is *years*
  - Let’s say poverty is *percent of the country’s population*

- Let’s imagine we run the regression, and we get results like this:
  - Intercept = 80
  - Poverty Level B_1 = -0.5

- What does this mean?
  - When poverty is *zero*, we expect life expectancy to be 80
  - For each *percent* higher poverty in a country, life expectancy decreases by half a year

P-Values & “Stars”

- Something else our model will tell us is the p-value for each coefficient

- Basically, the p-value tells us if we can have confidence that the effect of our independent variables is *significant*

- If a coefficient has a p-value of 0.05 or **lower**, we generally say the variable is significant
  - This means we are 95% confident that the effect of the variable is distinct from zero

- This is the same things the “stars” in the regression tables from the readings were telling us

Excel Time!

- Let’s see how to do this in Excel!

- You will be doing your own regression for the homework for next week

- If you get lost, here is a YouTube link to walk you through how to do a regression in Excel: https://youtu.be/0lpfmFnlDHI