NYU Department of Politics - Quant I
Fall 2013 - Prof. Patrick Egan

12 Lecture 12

12.1 Associations between and among variables

- Until now, we've focused on description and inference regarding one variable, and at times we've considered description and inference regarding comparisons of two variables drawn from different units.

- Now we turn the page to a task that political scientists spend a lot of time doing: considering relationships (or associations) between variables drawn from the *same* units.

- We'll start by considering the relationship between two variables, but quickly move on to considering relationships among many variables.

- First some very familiar terminology:

  - When we talk about a bivariate relationship, we typically refer to the two variables as X and Y.
  - We, of course, usually choose these labels with a causal model in mind: specifically, that X causes Y. But that assumption is UNNECESSARY for what we'll be discussing today. In this vein, X is known as the "independent variable," and Y is known as the "dependent variable."

- We'll proceed in several steps (list on board):

  1. DISPLAYING the relationship
     (a) Cross(tabulations)
     (b) Scatterplot
     (c) Boxplot
     (d) "Binning out" X

2. SUMMARIZING the relationship NONPARAMETRICALLY with central tendencies of Y by values of X:
    (a) TABLE: Summary statistics of Y for values of X
    (b) FIGURES (generally appropriate when X, Y or both are interval-level or higher)
        i. bar chart (more values of X, Y is interval-level)
        ii. scatterplot with smoother, indicating the central tendency of Y by values of X
3. SUMMARIZING the relationship PARAMETRICALLY, that is saying how closely the relationship approximates a perfectly linear relationship
    (a) Correlation
    (b) Bivariate Regression
4. Making INFERENCES about the nature of the relationship in a population from the relationship in a sample
    (a) Non-parametric: Pearson's chi-squared
    (b) Parametric: Correlation
    (c) Parametric: Linear regression

- (Walk through Parts 1 and 2 with handout)

- (Before part 3) To simplify this task, we will often need to resort to *models* that describe the theoretical relationship between the variables. As usual, we face a tradeoff between parsimony and precision. What these models buy us is parsimony: the assumptions we make with models allow us to summarize relationships between and among variables with just a few numbers. But what we pay for with models is that we lose some detail about the world. And if our theoretical model is incorrect, our descriptions, inferences and predictions about relationships between and among variables will be wrong.

- It's worth noting that (so far) our development of statistical tools for making inferences about univariate data has been remarkably free of assumptions. In fact, we can develop a list of these assumptions–and it's a short list:
    - in making inferences about the population mean, μ, with *large* samples,

        - **identicality** is necessary for our estimator, $\overline{Y}$, to be an unbiased estimator of μ.
        - **independence** is necessary in order for us to say that the variance of $\overline{Y}$ is equal to $\frac{\sigma^2}{n}$.
            - both of these assumptions are met when we have a **random sample**.
    - in making inferences about the population mean, μ, with *small* samples, we need an additional assumption:
        - the distribution of the underlying population is **Normal**.
            - although the tools we've learned are robust under moderate departures from this assumption.
    - finally, in making inferences about the differences between two population means, we need additional assumptions:
        - in large samples:
            - the two samples are drawn **independently**
        - in small samples:
            - the two samples are drawn **independently**, they have the same variance, and the underlying populations are Normal (although, again, these tools are robust under moderate departures from this last assumption).

- Of course, there are lots of instances when these assumptions don't hold, and statisticians spend a lot of time thinking about how to revise their tools to account for these cases. But still, it's a remarkably short list. It will get a lot longer as we move to bivariate and multivariate analysis.

- To begin thinking together about parametric ways to describe a bivariate relationship, let's revisit the notion of correlation. You'll recall the population correlation coefficient, ρ, which is equal to
$$\rho = \frac{COV(Y_1, Y_2)}{\sigma_1 \sigma_2} = \frac{E\left[(Y_1 - \mu_1)(Y_2 - \mu_2)\right]}{\sigma_1 \sigma_2}.$$

- ρ is, of course, is a theoretical quantity. Like μ or σ^2, we never actually observe it. But the maximum-likelihood estimator of ρ is the sample correlation coefficient, r :

$$r = \frac{\sum_i \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_i \left(X_i - \overline{X}\right)^2 \sum_i \left(Y_i - \overline{Y}\right)^2}}.$$

13 Lecture 14

- Rewrite sequence of four tasks on board:

  1. DISPLAYING bivariate relationships
  2. SUMMARIZING bivariate relationships NONPARAMETRICALLY
  3. SUMMARIZING bivariate relationships PARAMETRICALLY–that is, the extent to which they approximate a linear relationship
  4. MAKING INFERENCES about the nature of this relationship from a sample to a population

- – Now picking up at section 3:

- recall the correlation coefficient, r :

$$r = \frac{\sum_i \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_i \left(X_i - \overline{X}\right)^2 \sum_i \left(Y_i - \overline{Y}\right)^2}}.$$

  1. – You'll recall from last lecture the strengths and drawbacks of using r as a measure of the association between two variables. Remember that like all statistics, with r we are gaining parsimony while losing details.

- – Because the quantities that appear in r are used often in regression analysis, they have special symbols:

$$\begin{aligned} S_{xy} &= \sum_i \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) \\ S_{xx} &= \sum_i \left(X_i - \overline{X}\right)^2 = \sum_i \left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right) \\ S_{yy} &= \sum_i \left(Y_i - \overline{Y}\right)^2 \end{aligned}$$

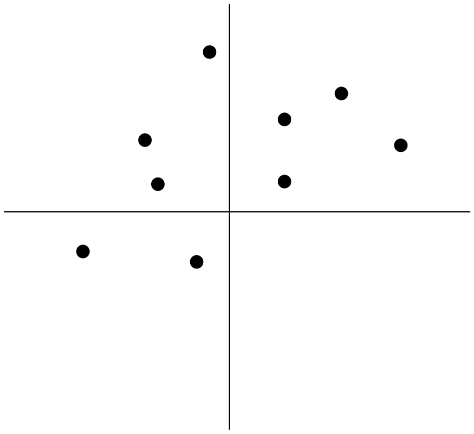
So how can we write r?

$$r = \frac{\sum_i \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_i \left(X_i - \overline{X}\right)^2 \sum_i \left(Y_i - \overline{Y}\right)^2}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

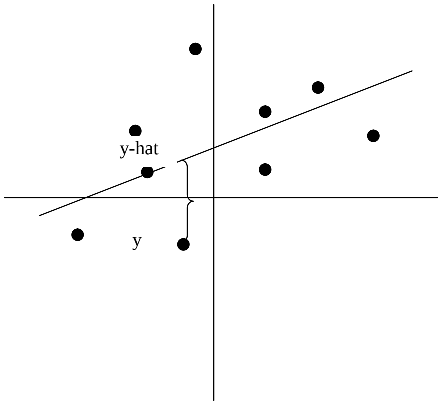
- now, (finally) linear regression.
- what if we wanted to put some more meat on the extent to which a relationship between x and y is linear? In particular, what if we wanted to say something about the line that best represented the relationship in linear terms? Well, we'd start by specifying a generic formula for that line, which convention has led us to write as follows:

$$y = \beta_0 + \beta_1 x$$

  where β_0 is the intercept and β_1 the slope.
- Draw on board:



-
- There are many such lines we might choose to represent this relationship in linear terms. And even the best line can't hit every point on the nose; it will certainly make mistakes. We define the mistake our line makes for any particular observation i as the observation y_i minus the "fitted value" for that observation, which we'll write as $\widehat{y}_i$:

- 

- Another term for these mistakes is residuals. We for any observation i we write the residual associated with that observation as $\widehat{u}_i = y_i - \widehat{y}_i$.

- An obvious criterion to use for picking the best line would be to minimize the mistakes it makes as it proceeds through the xy plane. That is, we could minimize $|y_i - \widehat{y}_i|$ as much as possible by minimizing $\sum_i |y_i - \widehat{y}_i|$.

- Well (as you will or have already learned), the absolute value function is a lousy one to work with mathematically. It has annoying properties that do not lend itself to easy manipulation. A more useful and easier sum to minimize is

$$SSR = \sum_i (y_i - \widehat{y}_i)^2,$$

  the **sum of squared residuals (SSR).** Note that the square of the residual has the nice property of becoming bigger as the magnitude of the residual increases. (It has the less felicitous property of counting bigger distances as greater than smaller distances: for example if we double a distance of 3 to 6, the corresponding squares are 9 and 36: a quadrupling. Thus whatever method we pick that minimizes SSR is going to work harder to minimize big deviations from the line than it probably should. But we're getting ahead of ourselves.)

- Now let's go back to our formula for a line. Let's do two things:

  - rewrite it with hats to be clear that we are generating estimates rather than saying anything specific (yet) about population values, and

- add subscripts for x and y:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

- Now let's consider the residual again, $\widehat{u}_i$, and substitute:

$$\begin{aligned}
\widehat{u}_i &= y_i - \widehat{y}_i \\
&= y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right) \\
(\widehat{u}_i)^2 &= \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right]^2 \\
\sum_i (\widehat{u}_i)^2 &= \sum_i \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right]^2
\end{aligned}$$

- So the $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that minimize this quantity are the intercept and the slope of what is known as the **least squares** line–the line that best represents the relationship between x and y. We can use the tools of calculus to find $\widehat{\beta}_0$ and $\widehat{\beta}_1$ by taking the partial derivative of SSR with respect to each of these quantities and setting these derivatives equal to zero:

$$\begin{aligned}
\frac{\partial SSR}{\partial \widehat{\beta}_0} &= \frac{\partial}{\partial \widehat{\beta}_0} \sum_i \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right]^2 \\
&= -2 \sum_i y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right) \\
&= -2 \left(\sum_i y_i - n\widehat{\beta}_0 - \widehat{\beta}_1 \sum_i x_i\right) = 0
\end{aligned}$$

$$\begin{aligned}
\frac{\partial SSR}{\partial \widehat{\beta}_1} &= \frac{\partial}{\partial \widehat{\beta}_1} \sum_i \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right]^2 \\
&= -2 \sum_i \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right] x_i \\
&= -2 \left(\sum_i x_i y_i - \widehat{\beta}_0 \sum_i x_i - \widehat{\beta}_1 \sum_i x_i^2\right) = 0
\end{aligned}$$

- These lead to the **normal equations**

$$\begin{aligned}
n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_i x_i &= \sum_i y_i \\
\widehat{\beta}_0 \sum_i x_i + \widehat{\beta}_1 \sum_i x_i^2 &= \sum_i x_i y_i
\end{aligned}$$

- Rewrite in matrix form:
$$\begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix} \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix}$$

- So:
$$\begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix} = \begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix}$$

- The inverse of a 2x2 matrix $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is $\mathbf{A^{-1}} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

- So
$$\begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}^{-1} = \frac{1}{n \sum_i x_i^2 - \left(\sum_i x_i\right)^2} \begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix}$$

- And thus
$$\begin{aligned} \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix} &= \frac{1}{n \sum_i x_i^2 - \left(\sum_i x_i\right)^2} \begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix} \begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix}, \text{ so} \\ \widehat{\beta}_0 &= \frac{\sum_i x_i^2 \sum_i y_i - \sum_i x_i \sum_i x_i y_i}{n \sum_i x_i^2 - \left(\sum_i x_i\right)^2}, \widehat{\beta}_1 = \frac{n \sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n \sum_i x_i^2 - \left(\sum_i x_i\right)^2}. \end{aligned}$$

  We can simplify these as follows :

$$\begin{aligned} \widehat{\beta}_0 &= \frac{n\overline{y} \sum_i x_i^2 - n\overline{x} \sum_i x_i y_i}{n \sum_i x_i^2 - \left(n\overline{x}\right)^2} = \frac{\overline{y} \sum_i x_i^2 - \overline{x} \sum_i x_i y_i}{\sum_i x_i^2 - n\left(\overline{x}\right)^2} \text{ and similarly} \\ \widehat{\beta}_1 &= \frac{n \sum_i x_i y_i - n^2 \overline{xy}}{n \sum_i x_i^2 - n^2 \left(\overline{x}\right)^2} = \frac{\sum_i x_i y_i - n\overline{xy}}{\sum_i x_i^2 - n\left(\overline{x}\right)^2}. \end{aligned}$$

- Simplify further by substituting the symbols we learned earlier with a little manipulation:

$$\begin{aligned} S_{xx} &= \sum_i \left(X_i - \overline{X}\right)^2 = \sum_i X_i^2 + \sum_i \overline{X}^2 - \sum_i 2X_i\overline{X} \\ &= \sum_i X_i^2 - n\overline{X}^2; \\ S_{yy} &= \sum_i Y_i^2 - n\overline{Y}^2 \\ S_{xy} &= \sum_i \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) = \sum_i X_i Y_i - \sum_i X_i \overline{Y} - \sum_i \overline{X} Y_i + \sum_i \overline{XY} \\ &= \sum_i X_i Y_i - n \sum_i \overline{XY} \end{aligned}$$

- So:

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{S_{xy}}{S_{xx}}.\text{ Note that because } \frac{cov(x,y)}{var(x)} = \frac{\frac{S_{xy}}{n}}{\frac{S_{xx}}{n}}, \\
\widehat{\beta}_1 &= \frac{cov(x,y)}{var(x)}.
\end{aligned}$$

- We can write $\widehat{\beta}_0$ more simply if note that we can rewrite

$$\begin{aligned}
-2\left(\sum_i y_i - n\widehat{\beta}_0 - \widehat{\beta}_1 \sum_i x_i\right) &= 0 \\
\sum_i y_i - n\widehat{\beta}_0 - \widehat{\beta}_1 \sum_i x_i &= 0 \\
n\overline{y} - n\widehat{\beta}_0 - n\widehat{\beta}_1\overline{x} &= 0 \\
\widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1\overline{x},\text{ and so} \\
\widehat{\beta}_0 &= \overline{y} - \frac{S_{xy}}{S_{xx}}\overline{x}.
\end{aligned}$$

- I will spare you a proof that we have satisfied the second-order condition for a minimum!

- Let's do a quick example. ("Handout on the Math of the Least-Squares Line.")

- Note a bit of terminology that gets thrown around:

  - We "regress y on x"
  - x is a "regressor"
  - x is on the "right hand side" of the regression equation

- Note that we NOT talking about making inferences (yet). Everything I've shown you thus far are simply mathematical properties that follow from the desired criterion of fitting a least squares line to data. Although this line tells us how well the data approximates a line, we have so far made no assumptions about the distribution of x and y in the underlying population.

- We will continue in this vein for a little while longer by discussing additional mathematical properties of the least squares line. Here are a few:

1. $\widehat{\beta}_1 = \frac{\Delta \widehat{y}}{\Delta x}$. If we start with our fitted line,

$$\begin{aligned}
\widehat{y} &= \widehat{\beta}_0 + \widehat{\beta}_1 x \quad \text{and take its derivative with respect to } x, \\
\frac{d\widehat{y}}{dx} &= \widehat{\beta}_1, \text{ we see (obviously) that:} \\
&\quad \text{a one-unit change in } x \text{ is associated with a change of } \widehat{\beta}_1 \text{ units of } \widehat{y}\text{, or} \\
\widehat{\beta}_1 &= \frac{\Delta \widehat{y}}{\Delta x}.
\end{aligned}$$

   Note that I'm saying associated with, not "causes." Don't get too consumed by this yet; we'll talk about this a lot more soon.

2. $\sum_i \widehat{u}_i = 0; \overline{\widehat{u}} = 0.$

   – The sum of the residuals, $\sum_i \widehat{u}_i$, equals zero. We see this immediately by noting that

$$\begin{aligned}
\widehat{u}_i &= y_i - \widehat{y}_i = y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right), \\
\sum_i \widehat{u}_i &= \sum_i y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)
\end{aligned}$$

   and noting that the formula for our regression line satisfies the F.O.C. associated with $\widehat{\beta}_0$:

$$\begin{aligned}
\frac{\partial SSR}{\partial \widehat{\beta}_0} &= \frac{\partial}{\partial \widehat{\beta}_0} \sum_i \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right]^2 \\
&= -2 \sum_i y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right) = 0 \\
&= \sum_i y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right) = 0, \text{ so} \\
\sum_i \widehat{u}_i &= 0
\end{aligned}$$

   – Note that this is just a property of the mechanics of fitting a line. We say that $\sum_i \widehat{u}_i = 0$ "by construction." This property is always the case, and it tells us nothing about our data or the relationship between x and y.

- Note that this also means that

$$\begin{aligned}\textstyle\sum_i \widehat{u}_i &= n\overline{\widehat{u}} = 0, \text{ and so}\\ \overline{\widehat{u}} &= 0.\end{aligned}$$

- That is, the sample mean of the residuals is zero.

3. $cov(x, \widehat{u}) = 0$.

   - By construction, the sample covariance between the regressors and the residuals is zero. his follows from the F.O.C. associated with $\widehat{\beta}_1$:

$$\begin{aligned}\frac{\partial SSR}{\partial \widehat{\beta}_1} &= \frac{\partial}{\partial \widehat{\beta}_1}\textstyle\sum_i \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right]^2\\ &= -2\textstyle\sum_i \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right] x_i = 0\\ \textstyle\sum_i \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right] x_i &= 0\end{aligned}$$

   - But since

$$y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right) = \widehat{u}_i,$$

   We can substitute and write

$$\textstyle\sum_i \widehat{u}_i x_i = 0.$$

   - Since the sample covariance of x and $\widehat{u}$ is

$$\begin{aligned}cov(x, \widehat{u}) &= \frac{\sum_i \left(\widehat{u}_i - \overline{\widehat{u}}\right)(x_i - \overline{x})}{n}, \text{ it follows that}\\ &= \frac{\sum_i (\widehat{u}_i)(x_i - \overline{x})}{n}\\ &= \frac{\sum_i \widehat{u}_i x_i}{n} - \frac{\sum_i \widehat{u}_i \overline{x}}{n}\\ &= 0 - \frac{\overline{x}\sum_i \widehat{u}_i}{n} = 0 - 0 = 0.\end{aligned}$$

4. $cov\left(\widehat{y}_i, \widehat{u}_i\right) = 0$. [LEAVE AS EXERCISE.]

   - The sample covariance between the fitted values and the residuals is zero.

$$
\begin{aligned}
cov\left(\widehat{y}_i, \widehat{u}_i\right) &= \frac{\sum_i \left(y_i - \overline{y}\right)\left(\widehat{u}_i - \overline{\widehat{u}}\right)}{n} \\
&= \frac{\sum_i \left(y_i - \overline{y}\right)\left(\widehat{u}_i\right)}{n} \\
&= \frac{\sum_i \widehat{u}_i y_i}{n} - \frac{\sum_i \widehat{u}_i \overline{y}}{n} \\
&= \frac{\sum_i \widehat{u}_i y_i}{n} - \frac{\overline{y} \sum_i \widehat{u}_i}{n}
\end{aligned}
$$

- What is $\sum_i \widehat{u}_i y_i$? It's

$$\sum_i \widehat{u}_i y_i = \sum_i \left(y_i - \widehat{y}_i\right) y_i$$

5. The point $\left(\overline{x}, \overline{y}\right)$ is always on the regression line.

   - Show this by substituting $\overline{x}$ for x, and $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$ in the formula for the regression line:

$$
\begin{aligned}
\widehat{y} &= \widehat{\beta}_0 + \widehat{\beta}_1 x \\
\widehat{y}\left(\overline{x}\right) &= \overline{y} - \widehat{\beta}_1 \overline{x} + \widehat{\beta}_1 \overline{x} \\
\widehat{y}\left(\overline{x}\right) &= \overline{y}.
\end{aligned}
$$

6. $\frac{\sum_i \widehat{y}_i}{n} = \frac{\sum_i y_i}{n}$.

   - By construction, the sample average of the fitted values is equal to the sample average of the observed y's, that is $\frac{\sum_i \widehat{y}_i}{n} = \frac{\sum_i y_i}{n}$. To show this, note that

$$
\begin{aligned}
y_i &= \widehat{y}_i + \widehat{u}_i, \text{ and so} \\
\sum_i y_i &= \sum_i \widehat{y}_i + \sum_i \widehat{u}_i \\
\sum_i y_i &= \sum_i \widehat{y}_i + 0 \\
\frac{\sum_i y_i}{n} &= \frac{\sum_i \widehat{y}_i}{n}.
\end{aligned}
$$

- We can view the process of fitting a least squares line as decomposing each y_i into two parts: $\widehat{y}_i$ and $\widehat{u}_i$. These values are by construction uncorrelated in the sample.

- Define the total sum of squares (SST) as

$$SST = \sum_i (y_i - \overline{y})^2 .$$

- Define the explained sum of squares (SSE) as

$$SSE = \sum_i (\widehat{y}_i - \overline{y})^2$$

- Recall that SSR is

$$SSR = \sum_i \widehat{u}_i^2$$

- It can be shown that $SST = SSE + SSR$. (From proof on p. 39 of Wooldridge, but it is complete here.)
- Start with the identity

$$\begin{aligned} SST &= \sum_i (y_i - \overline{y})^2 = \sum_i (y_i - \widehat{y}_i + \widehat{y}_i - \overline{y})^2 \\ &= \sum_i (\widehat{u}_i + \widehat{y}_i - \overline{y})^2 \\ &= \sum_i (\widehat{u}_i)^2 + \sum_i (\widehat{y}_i - \overline{y})^2 + 2\sum_i \widehat{u}_i (\widehat{y}_i - \overline{y}) \\ &= SSR + SSE + 2\sum_i (\widehat{y}_i - \overline{y}) \widehat{u}_i \end{aligned}$$

- Since

$$\begin{aligned} 0 &= cov(\widehat{y}_i, \widehat{u}_i) = \frac{\sum_i (y_i - \overline{y}) \left(\widehat{u}_i - \overline{\widehat{u}}\right)}{n} \\ &= \frac{\sum_i (y_i - \overline{y}) (\widehat{u}_i)}{n} = \sum_i (y_i - \overline{y}) (\widehat{u}_i) , \end{aligned}$$

we can write

$$SST = SSR + SSE.$$

13.1 More on Sums of Squares

- Last time we defined the three quantities

$$SST = \sum_i (y_i - \overline{y})^2 .$$

  -

$$SSE = \sum_i (\widehat{y}_i - \overline{y})^2$$

$$SSR = \sum_i \widehat{u}_i^2$$

- And showed that

$$SST = SSE + SSR.$$

- Let's think about these quantities in a bit more detail:
  - What is SST? It's the sample variance of Y - the extent to which it varies about its mean.
  - If the relationship between x and y could be perfectly described by a line, the fitted values (the y-hats) would would be the same distance from the mean as the observed values every time.
  - In that case $SSE = SST$, and the ratio $\frac{SSE}{SST} = 1$.
  - This is never the case, but we can exploit this fact to construct a measure of the "goodness of fit" of the regression line to the data. Call the ratio $\frac{SSE}{SST} = R^2$.
    - ∗ It always ranges between zero and one.
    - ∗ When reporting R^2, we typically report it to two decimal places.
    - ∗ R^2 is the proportion of the sample variation in y that is explained by x.
    - ∗ Also note that $R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$.
    - ∗ Where did R^2 get its name? Because it is also the case that in the bivariate context, $R^2 = (r)^2$.
      - · You'll show that in an exercise on this week's homework.

13.2 Least-Squares Regression is Invariant to Change in Units of Measurement

- What happens when we change the units in which the IV is measured (typically by multiplying these values by some constant, c)?

  - Recall that $\widehat{\beta}_1$ = the change in $\widehat{y}$ associated with a one-unit change in x.
  - So the change in $\widehat{y}$ associated with a one-unit change in cx should be $\frac{\widehat{\beta}_1}{c}$. And indeed it is:

- So when the IV is multiplied by c,

  - $\widehat{\beta}_1$ is divided by c.
  - $\widehat{\beta}_0$ does not change (it is the y-intercept, and $cx = 0$ when $x = 0$)
  - R^2 does not change.

- When the DV is multiplied by c,

  - (Again) recall that $\widehat{\beta}_1$ = the change in $\widehat{y}$ associated with a one-unit change in x.
  - So the change in $c\widehat{y}$ associated with a one-unit change in x is $c\widehat{\beta}_1$.
  - When $x = 0$, the intercept is $c\widehat{y}$.
  - So $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are multiplied by c when the DV is multiplied by c.
  - R^2 does not change.

- All of these will be more exciting exercises on this week's homework, too.

14 Lecture 15

14.1 Moving from Description to Inference

- As I've said (over and over), all we've done so far is talked about the regression line $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$ as a description of how well the relationship between x and y can be approximated by a linear relationship. In this context, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are simply descriptive statistics, like the empirical mean or empirical variance of the observed values of a variable.

- Now we move to using the formula for the least squares line to make **inferences** about an underlying population from the sample under analysis. Just as we used the statistic $\overline{Y}$ to make inferences about the parameter μ, we will now use the statistics $\widehat{\beta}_1 = \frac{cov(x,y)}{var(x)}$ and $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$ to make inferences about theparameters β_0 and β_1.

- As usual, we would like to find unbiased, relatively low-variance estimators of these parameters. As it turns out, the formulae for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that we developed earlier generate unbiased estimates of the parameters β_0 and β_1 as long as certain assumptions hold. We will now develop those assumptions. Later, we will introduce the additional assumptions necessary to show that $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are not only unbiased, but that have smaller variances than any other possible linear unbiased estimator of β_0 and β_1.

- **Assumption 1: The relationship between x and y in the population is linear in its parameters, and it is probabilistic.**

  - We begin with the assumption of a linear relationship between x and y. This not only rules out curvilinear relationships (draw). But (perhaps more importantly), it rules out relationships in which there are diminishing returns to x (draw) or in which the effects of x are smaller at its extreme values (draw sigmoidal function). (Ask for examples of these relationships.)

  - In many cases, these alternate functional forms describe the relationship between x and y with much more verisimilitude than a linear functional form. As we will see, there are all kinds of ways to account for these nonlinearities; but for now we are stuck in linearities: we assume that the change in y associated with a one-unit change in x is the same across the entire range of x.

  - Furthermore, we assume that the linear relationship between x and y is not deterministic – that is, it is not always the case that a particular $y_i = \beta_0 + \beta_1 x_i$. Rather, we assume that the linear relationship is **probabilistic**, and therefore write the complete population model as

$$y = \beta_0 + \beta_1 x + u,$$

and we will say that the relationship between any x_i and y_i in the population may be written

$$y_i = \beta_0 + \beta_1 x_i + u_i.$$

Here, u stands for "unexplained," and in this class, we will call u the **error**. For a unit i in the population, u_i is the amount by which the sum $\beta_0 + \beta_1 x_i$ comes in less than y_i. It is thus the variation in y which is unexplained by x.

- We view y, x and u as random variables when stating this model.
- Note the difference between u_i and $\widehat{u}_i$:
  - u_i is the *error* associated with a hypothetical unit in the population. It is never observed.
  - $\widehat{u}_i$ is the *residual* associated with an observed unit in our sample once we've estimated a regression line.
- Note that I have removed the "hats," because we are now talking about β_0 and β_1 as parameters of interest. β_0 and β_1 are unknown and never directly observed.
- **Assumption 2: we have a random sample of (x, y) pairs from the population, making them i.i.d.**
  - Although this assumption can often fail to hold in practice, let's note it for now and move on. Note that this is no stronger of an assumption than we needed in order to say that univariate estimators (such as Y-bar) are unbiased.
- **Assumption 3: The variance of our observed x$'s$ is nonzero.**
  - Happily, this is a weak assumption and can be easily confirmed by examining our sample.
- **Assumption 4: The error u has the expected value of zero, no matter what the value of x. That is, $E(u|x) = 0$.**
  - This is known as the assumption of "zero conditional mean." It is a statement about the unexplained factors contained in u_i, and it asserts that these other factors are unrelated

to x_i in that, given a value of x_i, the mean of the distribution of these other factors equals zero.

- This is a very strong assumption. It means that these other factors are uncorrelated with x. It practical terms, it means that there are no confounds that could render the relationship between x and y spurious.
  - Example from $y =$ income and $x =$ education.
  - When we write the population model

$$\text{income} = \beta_0 + \beta_1 \text{education} + u,$$

  - we are assuming that

$$\begin{aligned} E(u|\text{education}) &= 0, \text{ or more to he point} \\ cov\,(\text{education}, u) &= 0. \end{aligned}$$

  - But can we think about a factor that winds up in u (that is, the factor helps to explain income) but is correlated with education? Three prominent examples include: parents' education, ability, motivation. To the extent that these factors explain y and are correlated with x, Assumption 4 does not hold.
  - We'll talk about this in detail when we move to multivariate regression.
  - BTW, can we test this assumption by seeing if $corr(x_i, \widehat{u}_i) = 0$?
    - No: the assumption is about errors in the population, not the residuals in our sample. (What conclusion can we draw if we observe $corr(x_i, \widehat{u}_i) = 0$? (Nothing! This is by construction in OLS!)
- The assumption $E(u|x) = 0$ means that in the statistical derivations we are about to do, we will treat x as fixed. Technically, this isn't true as typically we are working with a random sample of (x, y) pairs. But nothing is lost in the derivations by treating the x as nonrandom. This allows us to write that $E(x) = x$.

14.2 The Unbiasedness of the OLS estimator for β_1

- These four assumptions together allow us to say that $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that we developed earlier generate unbiased estimates of the parameters β_0 and β_1. In this context, we call $\widehat{\beta}_0$ and $\widehat{\beta}_1$ the **ordinary least squares (OLS)** estimators of β_0 and β_1.

- Recall the formula for $\widehat{\beta}_1$:

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$$

- In order for $\widehat{\beta}_1$ to be defined, we need Assumption 3: $var(x) > 0$.

- Note that we can rewrite

$$\begin{aligned}\sum (x_i - \overline{x})(y_i - \overline{y}) &= \sum (x_i - \overline{x}) y_i - \sum (x_i - \overline{x}) \overline{y} \\ &= \sum (x_i - \overline{x}) y_i - \left[\sum x_i \overline{y} - \sum \overline{xy}\right] \\ &= \sum (x_i - \overline{x}) y_i - (n\overline{xy} - n\overline{xy}) \\ &= \sum (x_i - \overline{x}) y_i\end{aligned}$$

- So

$$\begin{aligned}\widehat{\beta}_1 &= \frac{\sum (x_i - \overline{x}) y_i}{\sum (x_i - \overline{x})^2} \\ &= \frac{\sum (x_i - \overline{x})(\beta_0 + \beta_1 x_i + u_i)}{SST_x} \\ &= \frac{\beta_0 \sum (x_i - \overline{x}) + \beta_1 \sum (x_i - \overline{x}) x_i + \sum (x_i - \overline{x}) u_i}{SST_x}\end{aligned}$$

- Note that $\sum(x_i - \overline{x}) = 0$, and

$$\begin{aligned}
\sum(x_i - \overline{x})\, x_i &= \sum\left(x_i^2 - \overline{x}x_i\right) \\
&= \sum x_i^2 - \overline{x}\sum x_i \\
&= \sum x_i^2 - n\left(\overline{x}\right)^2 \\
&= \sum x_i^2 - 2n\left(\overline{x}\right)^2 + n\left(\overline{x}\right)^2 \\
&= \sum x_i^2 - 2\overline{x}\sum x_i + \sum\left(\overline{x}\right)^2 \text{ [since } n\overline{x} = \sum x_i, \text{ and } n\left(\overline{x}\right)^2 = \sum\left(\overline{x}\right)^2 \text{]} \\
&= \sum x_i^2 - 2\overline{x}x_i + \left(\overline{x}\right)^2 \\
&= \sum\left(x_i - \overline{x}\right)^2 = SST_x.
\end{aligned}$$

- So:

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\beta_1 SST_x + \sum(x_i - \overline{x})\, u_i}{SST_x} \\
&= \beta_1 + \frac{\sum(x_i - \overline{x})\, u_i}{SST_x}
\end{aligned}$$

- Now let's find the expected value of $\widehat{\beta}_1$:

$$\begin{aligned}
E\left(\widehat{\beta}_1\right) &= E\left[\beta_1 + \frac{\sum(x_i - \overline{x})\, u_i}{SST_x}\right] \\
\text{Employing Assumptions 2 and 4, we can proceed} &: \\
&= \beta_1 + \frac{1}{SST_x} E\left[\sum(x_i - \overline{x})\, u_i\right] \\
&= \beta_1 + \frac{1}{SST_x} \sum(x_i - \overline{x})\, E\left(u_i\right) \\
&= \beta_1 + \frac{1}{SST_x} \sum(x_i - \overline{x})\, 0 \\
&= \beta_1.
\end{aligned}$$

- And for $\widehat{\beta}_0$:

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}$$

$$\begin{aligned}
\text{Since } y_i &= \beta_0 + \beta_1 x_i + u_i, \text{ then } \overline{y} = \beta_0 + \beta_1 \overline{x} + \overline{u}, \text{ substituting:} \\
\widehat{\beta}_0 &= \beta_0 + \beta_1 \overline{x} + \overline{u} - \widehat{\beta}_1 \overline{x} \\
\widehat{\beta}_0 &= \beta_0 + \left(\beta_1 - \widehat{\beta}_1\right) \overline{x} + \overline{u} \\
E\left(\widehat{\beta}_0\right) &= E\left[\beta_0 + \left(\beta_1 - \widehat{\beta}_1\right) \overline{x} + \overline{u}\right] \text{ Again using Assumptions 2 and 4, we proceed:} \\
&= \beta_0 + \left[\beta_1 - E\left(\widehat{\beta}_1\right)\right] \overline{x} + E\left(\overline{u}\right) \\
&= \beta_0 + \left[\beta_1 - \beta_1\right] \overline{x} + E\left(\overline{u}\right) \\
&= \beta_0
\end{aligned}$$

14.3 The variances of the OLS estimators

- Last time, we showed how four assumptions regarding x and y...

  1. Relationship is linear in its parameters
  2. x,y drawn from random sample, making them i.i.d.
  3. variance of x is nonzero
  4. $E(u|x) = 0$

- ..resulted in the least-squares solutions for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ being unbiased estimators for β_0 and β_1.

- $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are statistics, just like (say) $\overline{Y}$. Now, just as we did when making inferences about univariate distributions, we can say something about not only the unbiasedness of our estimates but also how far we can expect them to be away from the true parameter on average. In the univariate context, we were interested in describing the sampling distribution of the statistic X-bar.

- We will do the same thing here and consider the sampling distribution of $\widehat{\beta}_0$ and $\widehat{\beta}_1$. We, of course, already know the means of these sampling distributions: they are β_0 and β_1. Let's now compute the variances of $\widehat{\beta}_0$ and $\widehat{\beta}_1$. To compute these , we make a fifth assumption [add to list under separate heading]:

- **Assumption 5:** $VAR(u|x) = \sigma^2$. That is, the error u has the same variance no matter what the value of x. This is also known as homoskedasticity. When the assumption is violated, we have "heteroskedasticity."

- Draw pictures on board: Wooldridge p. 54 and p.55 – to illustrate homoskedasticity and heteroskedasticity.

  - Note difference between $VAR(u|x) = \sigma^2$ and $E(u|x) = 0$.

  - We did not need Assumption 5 to establish the unbiasedness of the OLS estimators.

- Note that this now allows us to write:

$$\begin{aligned}
y &= \beta_0 + \beta_1 x + u \text{ [the model]} \\
E(y|x) &= E\left[(\beta_0 + \beta_1 x + u)\,|x\right] \text{ [taking expectations conditional on x]} \\
&= E\left(\beta_0|x\right) + E(\beta_1 x|x) + E(u|x) \\
E(y|x) &= \beta_0 + \beta_1 x + 0 \text{ [thanks to assumption 4]}
\end{aligned}$$

- and

$$\begin{aligned}
VAR(y|x) &= VAR\left[(\beta_0 + \beta_1 x + u)\,|x\right] \\
&= 0 + 0 + VAR(u|x) \\
&= \sigma^2. \text{ [by assumption 5]}
\end{aligned}$$

- What is σ^2?

  - Ask class.

  - Note that it is not $VAR(y)$. It is $VAR(y|x)$.

  - Note that it is a parameter-not an estimate. So it's something in the population, not the sample.

  - σ^2 is a measure of the extent to which unexplained factors are affecting the value of y.

    - Assumption 4 means that we assume that these factors are unrelated to x.

    - Assumption 5 means that these factors are constant regardless of the value of x.
    - When σ^2 is bigger, it is the case that other factors explain a great deal of the variation in y in addition to x.
    - When σ^2 is smaller, it is the case that x is explaining a great deal of the variation in y on its own.

- Later, we'll show how to estimate the variances of the OLS estimators when this assumption is violated.

- Now, we're ready to derive the sampling variances of the OLS estimators. It is the case that:

$$\begin{aligned} VAR\left(\widehat{\beta}_1\right) &= \frac{\sigma^2}{SST_x}; \\ VAR\left(\widehat{\beta}_0\right) &= \frac{\sigma^2 \frac{\sum_i x_i^2}{n}}{SST_x}. \end{aligned}$$

- Proof:

  - Recall that in the proof establishing the unbiasedness of $\widehat{\beta}_1$, we wrote

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum (x_i - \overline{x})\, u_i}{SST_x}.$$

  - We then used the conditional mean assumption in order to say that the value of the last term of this expression is zero.

  - Now let's consider

$$\begin{aligned} VAR\left(\widehat{\beta}_1\right) &= VAR\left[\beta_1 + \frac{\sum (x_i - \overline{x})\, u_i}{SST_x}\right]. \\ &= 0 + \frac{1}{(SST_x)^2} VAR\left[\sum (x_i - \overline{x})\, u_i\right] \quad \text{[fixed X means we can treat as constant]} \\ &= \frac{1}{(SST_x)^2} \sum (x_i - \overline{x})^2\, VAR(u_i) \quad \text{[again]} \\ &= \frac{1}{(SST_x)^2} \sum (x_i - \overline{x})^2\, \sigma^2 \quad \text{[by Assumption 5]} \\ &= \frac{SST_x}{(SST_x)^2} \sigma^2 = \frac{\sigma^2}{SST_x}. \end{aligned}$$

- I'll leave the proof of the variance of $\widehat{\beta}_0$ as an exercise.

- Let's have a look at $VAR\left(\widehat{\beta}_1\right)$. We'd obviously like this to be as (what?) small as possible. What are the two quantities that make it small?

  - σ^2 : as gets smaller, $VAR\left(\widehat{\beta}_1\right)$ gets smaller.

  - SST_x : as gets bigger, $VAR\left(\widehat{\beta}_1\right)$ gets smaller.

    * Now let's take a closer look at SST_x :

$$\begin{aligned} SST_x &= \sum (x_i - \overline{x})^2 \\ \frac{SST_x}{n} &= \frac{\sum (x_i - \overline{x})^2}{n} \\ \frac{SST_x}{n} &= var(x) \\ SST_x &= n \cdot var(x) \end{aligned}$$

    * So

$$VAR\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{SST_x} = \frac{\sigma^2}{n \cdot var(x)}.$$

- Of these three properties, what can we really do anything about?

- In most cases, only n:

  - σ^2 is a parameter; it's declines only to the extent that x explains y well.

  - $var(x)$ is the empirical variance of x in our sample. In a random sample, it will of course look like the variance of x in the population. Not a whole lot we can do about that, either.

  - So what does this say about the relationship between sample size and the variance of $\widehat{\beta}_1$?

- Remember $VAR(\overline{Y}) = \frac{\sigma^2}{n}$? Compare to $\frac{\sigma^2}{n \cdot var(x)}$. Again, we have a ratio of something we can't control (the variance of y) over something we can (n).

15 Lecture 16

15.1 Estimating the error variance

- You'll recall that in the univariate context, we encountered a roadblock when we wrote $VAR(\overline{Y}) = \frac{\sigma_Y^2}{n}$. That is that we rarely know σ_Y^2. Well, when we write $VAR\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{SST_x}$, we have the same problem. We rarely have reason to know σ^2 in the OLS context, either.

- What did we do in the univariate case? We estimated σ_Y^2 with $S_U^2 = \frac{\sum_i (y_i - \overline{y})^2}{n-1}$. You'll recall that this was the empirical variance of y adjusted for the number of degrees of freedom (one) used in generating the estimate.

- Well, we'll do a similar thing here. We will estimate σ^2 with

$$\widehat{\sigma}^2 = \frac{\sum \widehat{u}_i^2}{(n-2)} = \frac{SSR}{(n-2)}.$$

- A proof that $E(\widehat{\sigma}^2) = \sigma^2$ may be found on p.57 of Wooldridge. The intuition here is that we have the variance of the residuals, again adjusted by the number of degrees of freedom (two) – since we've already generated estimates $\left(\widehat{\beta}_0 \text{ and } \widehat{\beta}_1\right)$ of two parameters using the two first order conditions for deriving the OLS estimators, which required that:

$$\sum \widehat{u}_i = 0 \text{ and } \sum \widehat{u}_i x_i = 0.$$

- The way to think about this (or any degrees of freedom scenario) is: how many pieces of data are free to vary once we've made our estimate? Here, if we know $n-2$ of the residuals, we can always calculate the other two residuals via the formulas above. They are not free to vary. We therefore lose two degrees of freedom, resulting in a total of $n-2$ degrees of freedom in our estimate of σ^2.

- Thus our unbiased estimators of $VAR\left(\widehat{\beta}_1\right)$ and $VAR\left(\widehat{\beta}_0\right)$ are:

$$\begin{aligned}\widehat{VAR\left(\widehat{\beta}_1\right)} &= \frac{\widehat{\sigma}^2}{SST_x} = \frac{\frac{SSR}{(n-2)}}{SST_x} \\ \widehat{VAR\left(\widehat{\beta}_0\right)} &= \frac{\widehat{\sigma}^2 \frac{\sum_i x_i^2}{n}}{SST_x} = \frac{\frac{SSR}{(n-2)} \frac{\sum_i x_i^2}{n}}{SST_x}.\end{aligned}$$

- $\widehat{\sigma}^2$, our estimate of σ^2, plays another important role, because

$$\sqrt{\widehat{\sigma}^2} = \widehat{\sigma} \xrightarrow{p} \sigma.$$

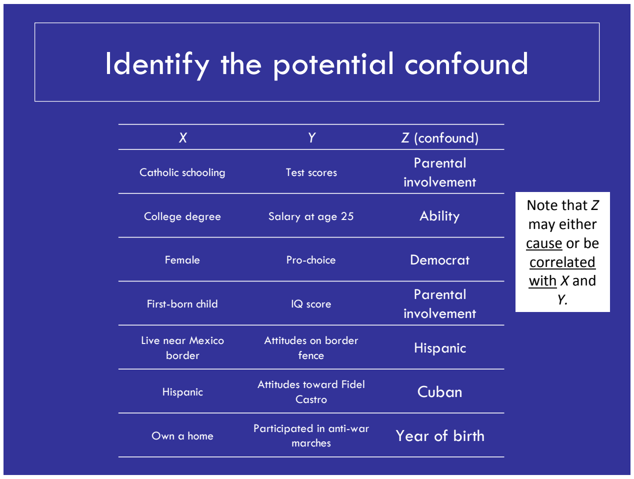
- Thus $\widehat{\sigma}$ is an interesting quantity in and of itself. It is expressed in units of y, which means that it tells us:
  - empirically, how far off the typical fitted value of y is away from the observed value; and
  - theoretically, the extent to which unexplained factors are affecting the value of y.
- It is a very informative statistic that gets much less attention than it deserves.
- Terminology:
  - Wooldridge calls $\widehat{\sigma}$ the Standard Error of the Regression (SER).
  - In Stata's regression output, $\widehat{\sigma}$ is displayed as "Root MSE," which stands for the root of the mean squared error of the regression.
  - I call $\widehat{\sigma}$ the standard error of the estimate, or SEE.
  - And sometimes you'll just see it displayed as $\widehat{\sigma}$.
- [NEXT YEAR: RELATIONSHIP BETWEEN R^2 AND $\widehat{\sigma}$.]

15.2 Hypothesis tests about β_1

- For now, we'll hold off on a discussion of how to conduct hypothesis tests on β_1. It will be more efficient to turn to it once we encounter multiple regression in the next lecture.

15.3 Controlling for a variable

- We are about to move on to multivariate regression.
- But before we do that, let's motivate the notion of controlling for a variable, and noticing how this does and does not compare to multiple regression.
- As we conduct research on political phenomena, we are often interested in what is known as the *ceteris paribus*–that is, the "all things being equal"–relationship between X and Y. [Draw diagram on board.]
  - That is, we are interested in the (often counterfactual case) of what the relationship between X and Y would look like if all other aspects of our units were the same.
    - We often call those other aspects variables Z.
- [NEXT YEAR: WHY IS THIS A PROBLEM? BECAUSE IF Z IS CORRELATED WITH BOTH X AND Y, THEN THE BIVARIATE RELATIONSHIP BETWEEN X AND Y MAY LEAD US TO IMPROPER CONCLUSIONS ABOUT THE CETERIS PARIBUS RELATIONSHIP BETWEEN X AND Y.
  - MAYBE INCLUDE EXAMPLES WITH CORRELATIONS?
  - Sometimes we do this because we are interested in the effect of X on Y, and we want to be sure that it is not due to Z.
  - But often, we're simply interested in the relationship between X and Y, holding everything else constant.
- Let's get specific about the terminology used here:
  - In this context, Z is called the potential **confound.**
  - If Z confounds the relationship between X and Y, it *renders the relationship spurious*.
    - That is, it leads us to improper conclusions about the *ceteris paribus*–that is, the "all things being equal"–relationship between X and Y.
  - Let's think a bit about potential confounds that may render a relationship spurious:



Identify the potential confound

X	Y	Z (confound)
Catholic schooling	Test scores	Parental involvement
College degree	Salary at age 25	Ability
Female	Pro-choice	Democrat
First-born child	IQ score	Parental involvement
Live near Mexico border	Attitudes on border fence	Hispanic
Hispanic	Attitudes toward Fidel Castro	Cuban
Own a home	Participated in anti-war marches	Year of birth

Note that *Z* may either cause or be correlated with *X* and *Y*.

- To determine whether Z renders the relationship between X and Y spurious, we:
  - "control for Z"
  - "condition on Z"
  - "hold Z constant."
- All three of these phrases typically mean the same thing.
- But there are several different ways to do this. Ideally, we would do exactly what "holding Z constant" suggests: divide our units by categories of Z and examine the relationship between X and Y within each category of Z.
  - If the relationship persists after controlling for Z, we say that it is not spurious.
  - If it no longer persists, we say that Z is a confound rendering the relationship between X and Y spurious.
- In practice, we usually do something much less careful.
- Handout: controlling for a variable.

Lecture 17

Although controlling for a variable by adding Z as an additive term in a multiple regression seems overly simple, it can still provide us with unbiased estimates of the ceteris paribus relationship between X and Y.

- - To see this, let's first analyze what happens when we don't control for Z:

- Assume that the true model is

$$y = \beta_0 + \beta_1 x + \beta_2 z + \nu$$

  where u is an error term such that $cor\left(u|x,z\right) = 0$. Regressing y on x1 and x2 will yield unbiased, consistent estimates of β.

- Notice that we're making a big assumption here: no interaction between x and z, and z enters into the DGP in a linear fashion.

- But if instead we regress y only on x1, obtaining the equation

$$y = \beta_0 + \beta_1 x + u,$$

  then what we are really doing is moving $\beta_2 x_2$ to the error term, ν :

$$\begin{aligned} y &= \beta_0 + \beta_1 x + (\beta_2 z + \nu), \\ \text{where } u &= (\beta_2 z + \nu). \end{aligned}$$

- You'll recall that in the bivariate case,

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum \left(x_i - \overline{x}\right) u_i}{SST_x},$$

- When then rely on the assumption that the covariance of x and u is zero to make the final term dissappear, and thus say that $E\left(\widehat{\beta}_1\right) = \beta_1$. But now consider

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum \left(x_i - \overline{x}\right)\left(\beta_2 z_i + \nu\right)}{SST_x}.$$

- Taking expectations, we now have

$$\begin{aligned} E\left(\widehat{\beta}_1\right) &= E\left(\beta_1\right) + E\left[\frac{\sum\left(x_i - \overline{x}\right)\left(\beta_2 z_i + \nu\right)}{SST_{x_1}}\right] \\ &= \beta_1 + \frac{\sum\left(x_{1i} - \overline{x}\right) E\left[\left(\beta_2 z_i + \nu\right)\right]}{SST_x} \\ &= \beta_1 + \beta_2\left[z_i \frac{\sum\left(x_i - \overline{x}\right)}{SST_x}\right]. \end{aligned}$$

- It turns out that $z_i \frac{\sum(x_i - \overline{x})}{SST_x} = \frac{cov(x,z)}{var(x)}$, which is the slope coefficient we would obtain if we regressed z on x!

- What if we wanted to say something about the sign of the bias? Well, note that $sign\left[\frac{cov(x,z)}{var(x)}\right] = sign\left[cov(x,z)\right]$. So if we omit x2 from our equation, we can now say that its sign is

$$sign\left[cov(x,z) \times \beta_2\right]$$

- What does this mean in practice? Consider a regression in which you model feelings toward Barack Obama as a function of Democratic Party identification. You omit a dummy variable for whether an individual is African-American. In what direction is your estimate of β_1 almost assuredly biased?

- What happens if $cov(x,z) = 0$? What happens if $\beta_2 = 0$?
  - That's right: when a variable is omitted, TWO problems must be present in order for it to cause bias:
    1. it is correlated with one or more x's in your model.
    2. its partial effect on y is not zero.
  - Why, then, do we love randomly assigning individuals to x? Because by construction, $cov(x,z)$ (for any omitted z you can think of) is zero, making $\widehat{\beta}_1$unbiased.

- This is a nice simple example, but it gets more complicated in a multivariate context. You'll see that next time.

- [That's because the term $\beta_2 \left[x_2 \frac{\sum(x_{1i} - \overline{x_1})}{SST_{x_1}} \right]$ becomes $\beta_2 \left[\left(\frac{1}{N} X'X \right)^{-1} \left(\frac{1}{N} X'x_2 \right) \right]$, which takes into account the extent to which the omitted variable (x_2) is collinear with all the included x's in the model. In practice, the sign of this bias is hard to consider in such a back-of-the-envelope fashion.]

- Take-home-point: if you leave out a variable that is BOTH correlated with included x's and has a separate effect on y, your estimates will suffer from omitted variable bias.