EEP/IAS 118
Andrew Dustan
Section Handout 3

Assumptions for Meaningful Regression

1. Big picture

We started this course by learning to mechanically perform OLS, which allowed us to _____

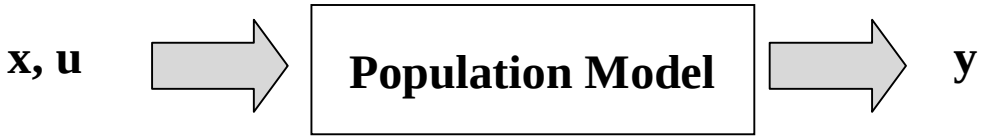
_____ (check Handout 1 if unsure). However, there was a big warning to go along

with OLS: OLS *cannot*, by itself, tell us whether _____.
Think of it this way. Learning to do OLS is like learning any other skill, such as skydiving. Just because someone knows how to skydive doesn't mean it's always a good idea. In particular, skydiving might be a good idea only when a) the skies are clear of thunderstorms, b) the airplane is free of mechanical problems, and c) your backpack actually has a working parachute in it. You can never be 100% *sure* that a, b, and c are true on any given day, but you should do your best to find out beforehand.

It's the same when doing OLS. We know how to do it, but when is it a good idea? When can we use the results of a regression to make statements about the marginal effect of x on y? **Before we even think of making these statements, we have to think about whether the assumptions of OLS are satisfied. They usually aren't.** Just like skydiving, we'll never be 100% sure they're satisfied, but we owe it to ourselves to do some due diligence beforehand. Many economists just make the OLS assumptions without thinking about whether they'll hold or not. This leads to wrong conclusions. If econometrics were skydiving, we would have a lot of dead econometricians on our hands.

2. From true population model to OLS on a sample

The population model tells us how, in real life, the dependent variable (y) is determined, both by observed (x) and unobserved (u) variables:



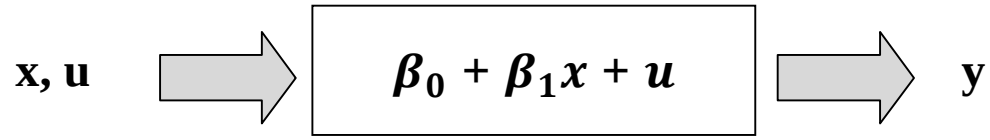
We never see this model. If we did, we wouldn't have to do a regression. The goal of regression is to get a guess of what the population model is. Here is a table of true values in the population and the guesses that we get for them using OLS on a sample drawn from the population:

POPULATION VALUE (true but unobserved)	SAMPLE VALUE (our best guess)
β_0, β_1	
σ^2	
u	
$E(y\|x)$	

3. Assumptions for OLS

Here are the assumptions necessary in order for OLS regressions on a sample of data to give us guesses about the population model that are, on average, correct. ("On average correct" = "unbiased.")

SLR.1: y is linear in parameters: $y = \beta_0 + \beta_1 x + u$
This puts a lot of structure on our population model "machine":



Draw and label one example of a population model where SLR1 holds and one where it does not:

HOLDS **FAILS**

Note that we can remedy some of these problems by transforming the x and y variables, either taking logs or using a polynomial in multiple regression. So this assumption isn't as bad as it seems!

SLR.2: $\{(x_i, y_i), i = 1, \dots, n\}$ is a random sample from the population.
This just says that the sample has to represent the population it comes from. In practice, the way that researchers collect data often violates this assumption.

Think of one example of how SLR.2 could be violated in real research. Fill the rest with in-section examples:

SLR.3: Not all sampled x values $\{x_i : i = 1, \dots, n\}$ are the same.
If we don't have any variation in x, then we can't mechanically perform OLS. What would be the point even if we could? We wouldn't have any way to see how y changed when x changed.

SLR.4: No matter what the value of the observed variable (x), we expect the unobserved variable (u) to be zero: $E(u|x) = 0$.
This is probably the biggest assumption and the hardest to convince people holds in a research project. It says that the observable variable (x) is totally unrelated to the unobservable things (u) that affect our outcome (y). We've talked about many cases where this is blatantly untrue (like a regression of wage on education, or crop yield on the amount of fertilizer used.)

SLR.5: The "error term (u)" has the same variance for any value of the explanatory (x) variable: $Var(u|x) = \sigma^2$.

This is *not* necessary in order for $\hat{\beta}$ to be unbiased, but rather for s^2 to be an unbiased estimator of σ^2. We'll talk more about this next time.

4. Practice: finding violation of assumptions

Reading this example of a research project, decide which assumptions (SLR.1-5) are likely to be violated and which aren't.

A recent graduate of Stanford has been hired by a county health board to investigate whether emissions from an industrial plant are having a negative health impact on county residents. The current hypothesis is that residents very near the plant are inhaling a lot of toxic emissions that result in severe acute respiratory illnesses, but that the particles disperse rapidly as they move farther from the source. This is particularly worrisome for the county because several retirement communities are near the plant and their residents are getting very angry with the county for doing nothing about the perceived problem.

The researcher obtains Census data to form a random sample of households within the county. First, he records the household's distance from the plant. Then, he drives to each household and, as a rough measure of health status, asks the person at the door how many days he/she has been in the hospital with respiratory issues in the past two months. The proposed population model is:

$$days\ in\ hospital = \beta_0 + \beta_1(miles\ from\ plant) + u$$

The researcher estimates the model and finds the following:

$$\widehat{days\ in\ hospital} = 3.6 - 0.6(miles\ from\ plant)$$

He concludes that the plant is having negative health impacts on residents. The county shuts down the plant, which lays off all of its workers.

Assumption	Violated? (X one)			Why?
SLR.1	Almost Surely	Maybe/ Unclear	Probably Not	
SLR.2	Almost Surely	Maybe/ Unclear	Probably Not	
SLR.3	Almost Surely	Maybe/ Unclear	Probably Not	
SLR.4	Almost Surely	Maybe/ Unclear	Probably Not	
(SLR.5)	Almost Surely	Maybe/ Unclear	Probably Not	