Warm-Up

Consider the two panel data regressions below, where i indexes individuals and t indexes time in months:

$$y_{it} = \beta_0 + \beta_1 x_{1,it} + ... + \beta_k x_{k,it} + u_{it} \tag{1}$$
$$y_{it} = \beta_0 + \beta_1 x_{1,it} + ... + \beta_k x_{k,it} + a_i + u_{it} \tag{2}$$

What kind of regression is model (2)?

A fixed effect regression.

What are the MLR.4 assumptions for each model?

For (1): $\mathbb{E}[u|x_1, ..., x_k] = 0$. For (2): $\mathbb{E}[u|x_1, ..., x_k, a_i] = 0$

We've talked about how OVB is a violation of the MLR.4 assumption. What kind of omitted variable bias is mitigated by using model (2) instead of model (1)? [Why is model (2) *better* than model (1)]

Any omitted variable that is constant (or relatively constant) over time for a unit i will bias (1), but will not bias (2) because the fixed effect will capture any effect they have.

1 Impact Evaluation: Randomized Controlled Trial (RCT)

Impact evaluation is a collection of estimation strategies to determine the causal effect of a program, a policy change, etc. Econometricians are mainly concerned with revealing causal relationships, which is actually very difficult and often requires more than a little creativity. Omitted variable bias, in particular, makes this task a hard one and has prevented us from making any credible causal claims so far this semester.

If we're trying to calculate the effect of a policy intervention or program, there's one thing we need if we're going to estimate the impact credibly:

- A counterfactual or comparison group: Imagine what would have happened if the policy intervention had in fact *not* taken place—this is the counterfactual world. Hypothetically speaking, if we could observe both what happened in the real world and the counterfactual world, we could perfectly estimate the causal effect of the intervention. We would know what each person's wage would be with and without a job training program, so we could easily see how the training affected each person's wage.

The most obvious issue with this is: We can't observe this counterfactual world (and time travel is impossible). The easiest way to get around this problem is to:

- Randomly assign the "treatment" or policy intervention to units, and then compare them. *On average*, the control group will look exactly like the treatment group along any characteristics you measure because you *made sure* that each unit's status as either treatment or control was *not* related to any of their characteristics by randomly deciding.

Even though RCT is the benchmark impact evaluation strategy because you are in control of how the counterfactual group is chosen (and it's chosen in the best statistical way possible), there's still the possibility that your randomization went horribly awry (or that the randomization wasn't actually implemented, i.e.

administrators gave the treatment to who they wanted). Remember that the key assumption for an RCT to measure a causal effect of a treatment is:

- We assume that our random assignment of treatment "worked" in the sense that the average outcomes of our treated units *would have* been the same as the control units *if not for the treatment*—simply because of the way we randomly picked people for treatment. Compare this to a medical study in which anyone is allowed to take a new drug to lower blood pressure. *Only* people with high blood pressure will choose to take the drug, which means who ended up being treated with the drug was dependent on a person's characteristic: their blood pressure level. Then if you compared people who took the drug to those you didn't, you might falsely conclude that the drug increased blood pressure, when it was really only high blood pressure people choosing to be treated. By randomly assigning who gets the drug and who gets a placebo, researchers make sure that the treatment will have nothing to do with any other characteristics, and only reflect the actual impact of the drug on blood pressure.

If we look at the regression model used with RCTs, this key assumption is simply our MLR.4:

$$outcome_i = \beta_0 + \beta_1 Treated_i + u_i$$

We assume that the treatment was random and didn't actually have anything to do with other characteristics (omitted variables in u_i): $\mathbb{E}[u_i|Treated_i] = 0$.

And though we cannot actually verify this (which is why we have to *assume* it's true), we can provide support for this assumption by:

- We compare the averages of observable characteristics (that we don't think would have been impacted by the treatment) between the two groups, e.g. compare average household size in the treated units to average household size in the control units. Then we use t-tests to determine if any of these averages are statistically different across the groups.[1]

2 Impact Evaluation: Difference-in-Difference

However, randomizing treatment by a program or policy is often impossible because the intervention has actually already happened, or it's unethical or otherwise infeasible to withhold the treatment or intervention from a group in order to have a control.

If we haven't been able to do an RCT, and haven't been able to assign the treatment (or program, or policy, etc.) randomly ourselves, then we have to assume that the treatment was *not* randomly assigned—that it depends on either observable or unobservable characteristics of the people, firms, cities under consideration. In this case, there are important differences between our treated and untreated units that we cannot control for in a regression. Leaving these variables out in u_i will cause OVB, which totally foils our attempts at estimating the causal effect!

Difference-in-differences is a way of getting around a non-random assignment of a program or policy.

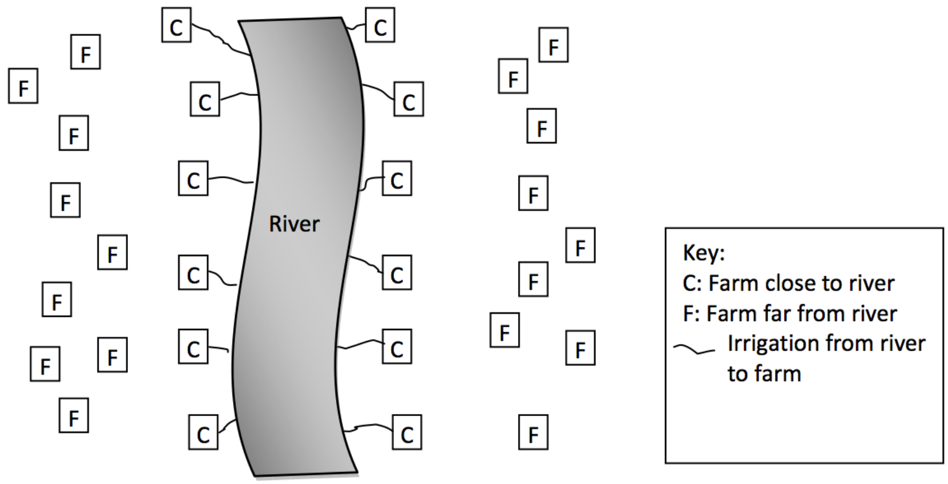
[1] Remember Type I error? If we're comparing the means of 100 variables across treatment and control, t-tests at the 95% level still have a 5% chance of being falsely rejected. So we should expect *some* significant differences between treatment and control, even when the randomization worked.

Example. Effect of an irrigation project on agricultural yields

The World Bank used to think that big infrastructure projects were the key to development in poor countries. For example, building irrigation passages to divert water from a river to nearby farms. Suppose you were asked to evaluate whether this particular irrigation project successfully increased farmer's yields.

Attempt 1: Cross sectional regression

First suppose that the World Bank does the project and then collects a season's worth of data on crop yields (metric tons per hectare) for farms in the area, both those close enough to the river to get irrigation and those too far away to be irrigated. So the snapshot of data looks something like this (plus yield data):



What regression would you estimate here to find the effect of getting irrigation?

$$yield_i = \beta_0 + \beta_1 irrigation_i + u_i$$

where *irrigation* is a dummy variable equal to 1 if the farmer got an irrigation passage and 0 if not.

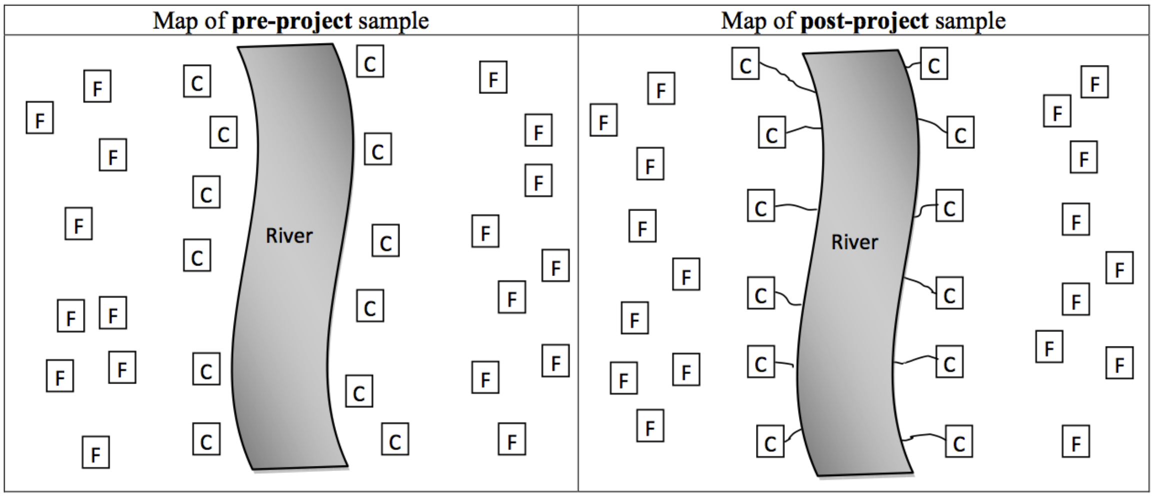
What is the counterfactual here, i.e. what's the comparison group? What's wrong with this strategy?

Because variable *irrigation* is a dummy indicating whether a farm got irrigation from the river, we're comparing irrigated farms to non-irrigated farms. Our assumption is that the non-irrigated farms are a good counterfactual for the farms that got the irrigation, i.e. the only real difference between them is the actual irrigation. This shouldn't feel right: the C farms are the ones that got the irrigation and they're all farms that are close to the river! There are probably a lot of things that vary between C and F farms besides irrigation, and the *irrigation* variable is going to pick *all* of those up, not just the project effect!

Another way to think about it is that distance to the river is an omitted variable in this regression that's correlated with both irrigation and yields (maybe the soil or other land characteristics vary by distance to the river), so there will be OVB for β_1.

Attempt 2: Difference in difference regression

Instead, supposed that the World Bank collects two waves of data, one before the project was started and one after it was completed. Here's what you now have:



What regression would you estimate here to find the effect of getting irrigation?

$$yield_i = \beta_0 + \beta_1 irrigation_i + \beta_2 post_i + \beta_3 (irrigation * post)_i + u_i$$

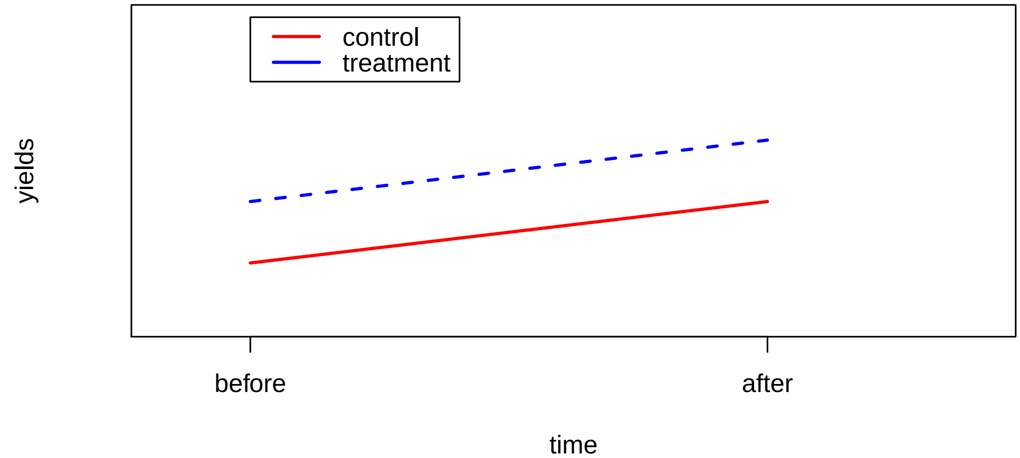
This is better than Attempt 1 because it accounts for differences (some of which we can't observe) between the C and F farms, getting around the fact that the irrigation was not randomly assigned across farms.

What is the counterfactual here?

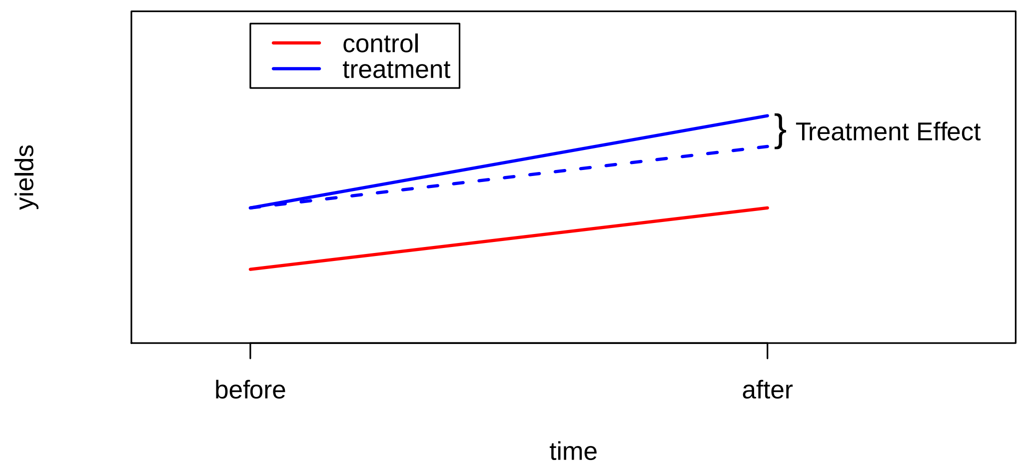
The counterfactual here is much trickier, but you can think of it this way: instead of there being a control *group* of farms that we can point to as what would've happened to the treatment farms had the irrigation program not occurred, we have a a control *change*. We have an idea of what the *change* in yield for treatment farmers between the pre-project sample and the post-project sample should be—about the same as the *change* in yield for control farmers over the same period.

We can also draw a picture to understand the diff in diff assumptions and strategy:

First, let's think about what might have happened to the close and far farms' yields if the World Bank hadn't done anything:



This picture demonstrates the key assumption: if not for the irrigation program, the close farms would have had the same *change* in yields over time as the far farms. Now we can think about what actually happened (the irrigation passages were dug, and the close farms were treated by the program), and you should see how the difference-in-differences strategy finds the treatment effect—given that the key assumption holds.



Basically, the diff-in-diff strategy is to conclude that any difference in the slope of these two lines is due to the treatment (because we are assuming that the slopes *would* have been the same without the program).

So, how do we interpret β_3? It's the average treatment effect! It's the additional difference between irrigated farms and non-irrigated farms after the irrigations passages have been dug—in other words, the estimated effect of improved irrigation on farmer yields under our key assumption.