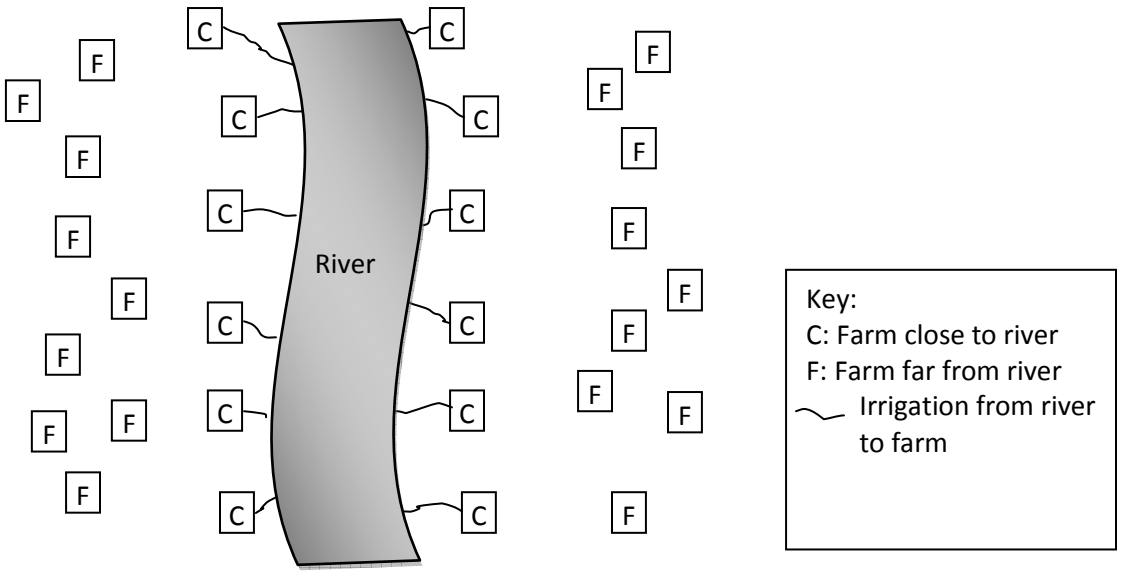
1. Difference-in-Differences

Remember that one of the main goals of econometrics is to estimate the *causal* effect of x on y. Omitted variables bias makes this a very difficult task, and in most of the regressions we have carried out this year you can make convincing arguments that omitted variables prevent us from making credible causal claims.

In this part of the course, we're looking at some methods that bring us much closer to making credible claims of causality. The first of these methods is called difference-in-differences. We went through the theory in class, so let's do an example that shows how difference-in-differences works and why it's better than a cross-sectional regression:

Example:
The World Bank used to think that big infrastructure projects were the key to development in poor countries. One such project was the construction of many irrigation projects. One form this could take is diverting water from a river to nearby farmland. Suppose you were asked to evaluate whether a particular irrigation project of this type had been successful in increasing farmers' crop yields. How would you do it?

Attempt 1: Cross-sectional regression
Say the World Bank does the project and then collects on 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 and so receive no benefit from the project. Here is an illustrative diagram of what you have:



Which cross-sectional regression could you use to estimate the effect of the project?:

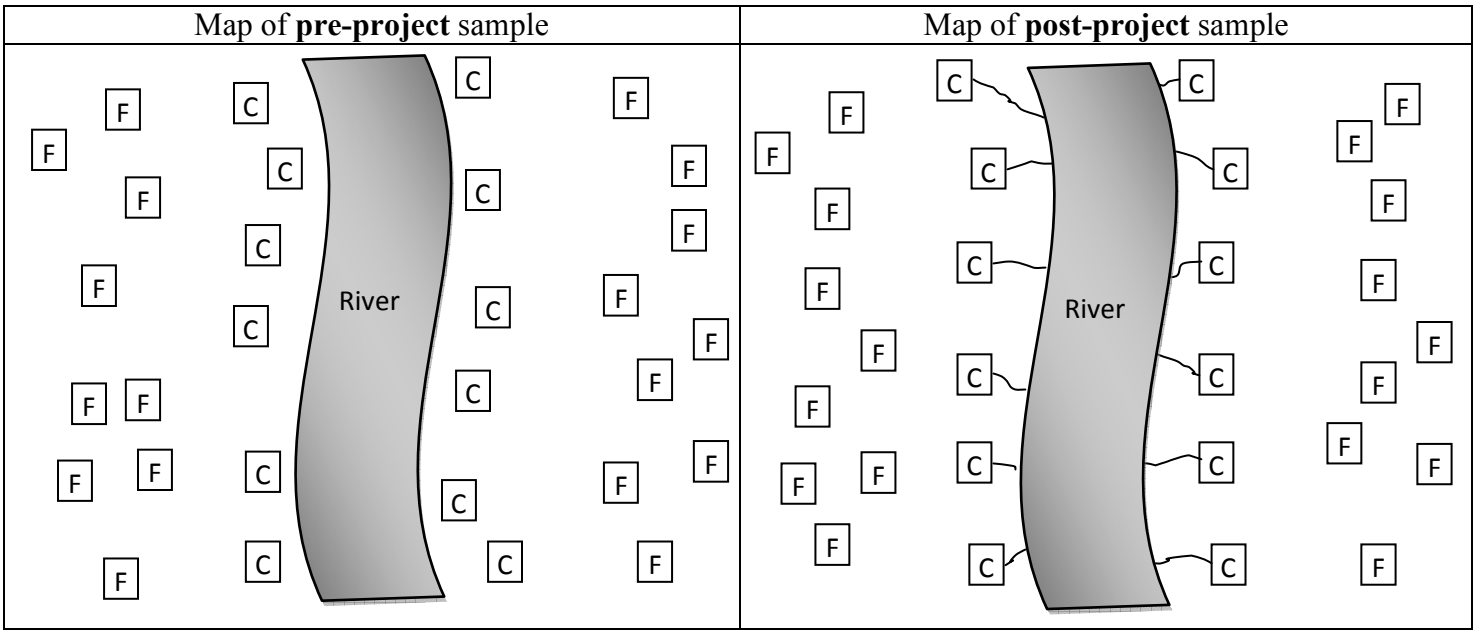
$$yield = \beta_0 + \beta_1 irrigation + u$$

Is it a good idea to run this regression? Why?

Probably not. There are differences between the farms close and far from the river that affect their yields. Thus β_1 will give the effect of irrigation *plus* all of these other factors.

Attempt 2: Difference-in-differences regression

What if the World Bank had collected *two* waves of data, one *before* the project was built and one *after* it was built? Here's what you would have:



How can we change our regression from Attempt 1 to get an estimate of the causal effect of the project?

$$yield = \beta_0 + \beta_1 close + \beta_2 post + \beta_3 (close * post) + u$$

Here we have four parameters. What is the corresponding expected yield for "F" (far) and "C" (close) farms in each period (pre-project and post-project)?

		Control for fixed differences in average yields between close and far farms →		
		Far	**Close**	
Control for time trend in yields that affects all farms ↓	**Pre-project**	β_0	$\beta_0 + \beta_1$	
	Post-project	$\beta_0 + \beta_2$	*(Only farms here actually have irrigation)* $\beta_0 + \beta_1 + \beta_2 + \beta_3$	**Difference**
	Difference	$(\beta_0 + \beta_2) - \beta_0 = \beta_2$	$(\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 - \beta_1) = \beta_2 + \beta_3$	β_3

That last box is the difference-in-differences estimate, and it gives the effect of *actually getting the irrigation*, after controlling for the fixed differences in average yields between the close and far farms and a time trend that affected close and far farms equally. Basically, what we are doing is controlling for all cross-sectional differences and the evolution of yields over time, so that (hopefully) all that is left is the one thing that varied systematically over *both* the cross-sectional *and* time dimensions: the irrigation project.

How do you do this in Stata?

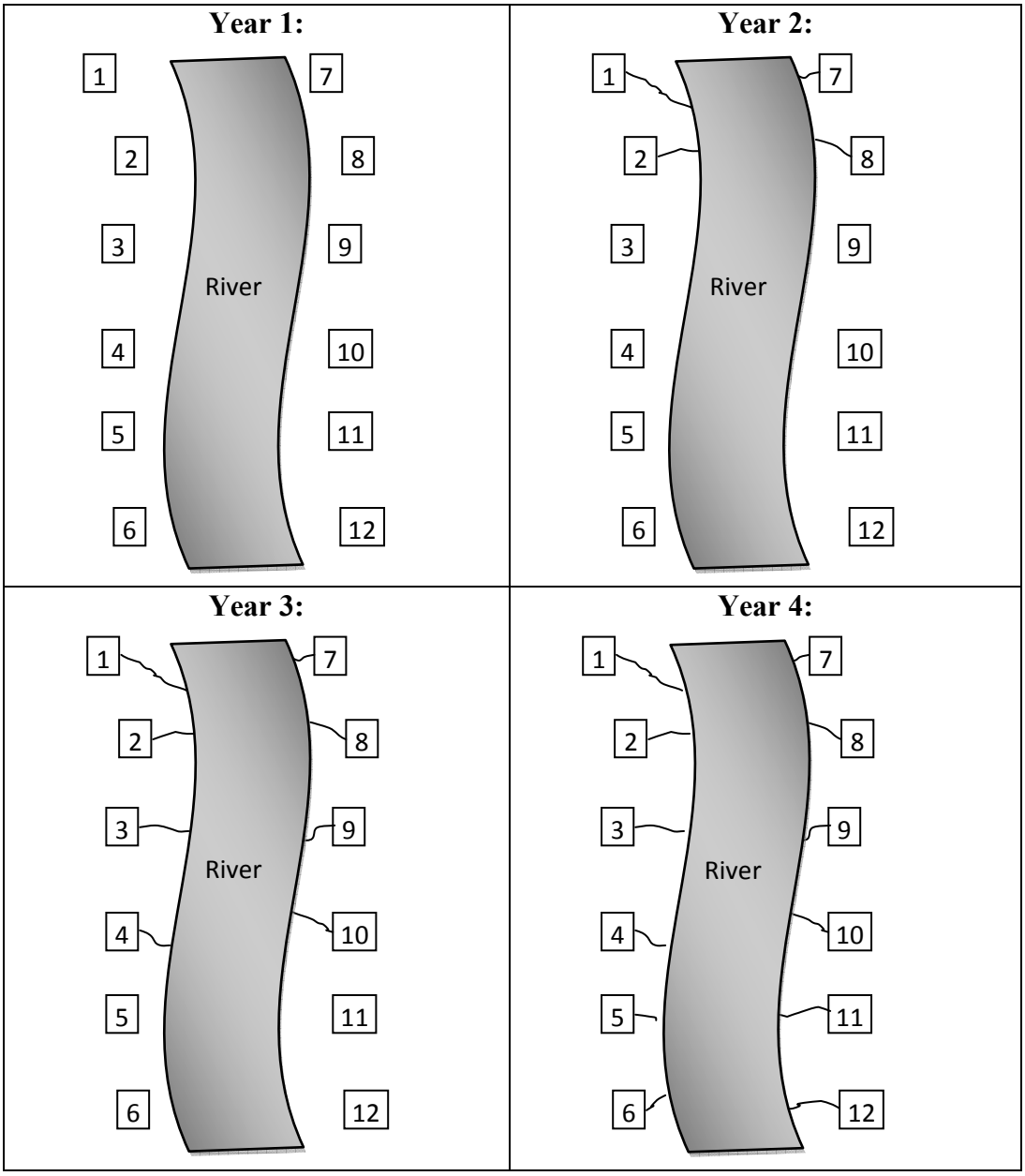
It's pretty easy. You just create the interaction term, in this case $(close * post)$ and run the regression on the two dummy variables and this new interaction term. The nice thing is that the difference-in-differences estimate you get (the coefficient on the interaction term) has the standard error and t-statistic estimated for you.

2. Panel Data & Fixed Effects Regression

"Panel data" is data that follows some cross-sectional units (people, cities, firms) for 2+ time periods. "Fixed effects" regression is a useful technique that takes advantage of both the cross-sectional (across units) and time dimensions of panel data. To illustrate, let's stick with the idea of irrigation and crop yields.

Example:

Say that the World Bank wants to you evaluate another irrigation project. This one had a limited budget, so they only completed part of the project each year. Luckily, they surveyed the same sample of beneficiary households every year until the project's completion. Here is what the data look like:



This is a great situation for us. We can use fixed effects regression to find the effect of irrigation, while controlling for *all* time-invariant characteristics of the farms—farmer ability, soil quality, land slope, etc.—that affect yield.

We have T=4 years (*year1, year2, year3, and year4*). We also have some number of farms, 12 in the diagram but it could be many more than that (*n* of them).

Let's think back to lecture and write out the fixed effects regression. Remember: we need to control for both the year effects and the farm-specific effects on yields:

$$yield_{it} = \beta_0 + \delta_2 year2_t + \delta_3 year3_t + \delta_4 year4_t + \beta_1 irrigation_{it} + farm_i + u_{it}$$

Note about the subscripts (i, t, it):

- If a variable has the same value for every farm in a particular year, it has a **t** subscript.
- If a variable has the same value in every year for a particular farm, it has an **i** subscript.
- If a variable can take different values between farms in a particular year, and can also take different values across years for the same farmer, it has an **it** subscript.

Note about the $farm_i$ term:

This term can be thought of as shorthand for the *n* dummy variables representing each of the *n* farms. It gets tiring to write out dummy variable terms for dozens or hundreds of farms, so we risk the confusing notation to save ourselves the work.

The fixed effects estimator controls for fixed differences in yields across farms as well as general time trends in yields. For the estimated coefficient on $irrigation_{it}$ to be unbiased, we need *changes* in $irrigation_{it}$ to be uncorrelated with *changes* in unobserved farm-level variables that affect yield (farmer ability, soil quality, etc.). This is much less daunting than requiring that $irrigation_{it}$ be uncorrelated with the *level* of unobserved farm-specific variables that affect yield. Think about it!

Tips for doing this in Stata:

Your data would look like this (hint: this is what your data will look like in the problem set):

farm	year	yield	irrigation	other *x*'es
1	1	6.1	0	#
1	2	11.2	1	#
1	3	10.8	1	#
1	4	12.1	1	#
…	…	…	…	…
12	1	4.6	0	#
12	2	4.2	0	#
12	3	5.2	0	#
12	4	9.0	1	#

The main challenge you have is turning the *farm* and *year* variables into dummy variables, one for each year or farm. Here's how you do it:

```
tab year, gen(y);
tab farm, gen(f);
```

These commands will make two sets of dummy variables: {y1, y2, y3, y4} and {f1, f2, f3,…, f12}. You can change "y" and "f" to whatever you want—it's just the root for the variable names you generate.

Once you have all of those dummy variables generated, how do you run the fixed effects regression without having to type out all of those variables? Do this:

```
regress yield y2-y4 irrigation f2-f12;
```

The dash just tells Stata to include all the variables between *y2* and *y4*, or *f1* and *f12*, respectively.

Or, since we usually don't care about the $farm_i$ coefficient terms and they can make the regression output really ugly if we have a lot of farms, we can do an identical regression with this command:

```
xtreg yield y2-y4 irrigation, fe i(farm);
```

The i(farm) part allows you to take out the *f1-f12* variables, because it automatically generates a dummy variable for each value of *farm* and includes them in the regression. It just doesn't tell you the estimated coefficients for them. And it will report a different constant term ($\hat{\beta}_0$), but this is actually a technical rather than deep issue.