14.661: Recitation 2

Chris Walters

September 24, 2010

1 Differences-in-differences

1.1 Basic Structure

Differences-in-differences (or "DD" for short) is a technique for estimating the effect of some treatment on some outcome. In Eissa and Liebman (1996), the treatment is the expansion of the EITC, and the outcome is employment. The basic DD research design has 2 groups (treatment and control) and two time periods (pre and post). The treatment group receives the treatment in the post period. Let's introduce some notation. If Y_i is our outcome, then let

$$Y_{0i} = i\text{'s outcome if he or she does not receive the treatment}$$

$$Y_{1i} = i\text{'s outcome if he or she receives the treatment}$$

As researchers, what we ultimately want is a plausible estimate of

$$\delta \equiv E\left[Y_{1i}|treatment\ group, post\right] - E\left[Y_{0i}|treatment\ group, post\right]$$

That is, we want to know is how the treatment affected the treatment group in the post period. How much higher was labor force participation for single women with kids than it would've been without the expansion of the EITC?

The first term is easy to estimate – it is just the average Y for the treatment group in the post period. The problem is that it's not clear how to estimate $E\left[Y_{0i}|treatment\ group, post\right]$. We cannot observe what LFP would've been for single women with kids without the expansion, because the expansion happened. We therefore have to estimate it somehow. If we were running a randomized trial, we could simply estimate it from the control group, but we often can't do this. In DD estimation, we select a non-randomized control group. In Eissa and Liebman (1996), the control group is single women without kids. This group does not receive the treatment.

Given such a control group, we might think to estimate δ as follows:

$$\hat{\delta}_1 = \hat{E}\left[Y_i|treatment\ group, post\right] - \hat{E}\left[Y_i|control\ group, post\right]$$

Here our estimate of $E\left[Y_{0i}|treatment\ group, post\right]$ is the control mean in the post period. The trouble with this is that if there are any differences between our treatment and control groups, this estimator will attribute them to the treatment. This estimator would work in a randomized trial, but if we have selected our control group in a nonexperimental setting then we might be worried.

Another estimator we might try is to simply throw out the control group and look at the change over time for the treated:

$$\hat{\delta}_2 = \hat{E}\left[Y_i|treatment\ group, post\right] - \hat{E}\left[Y_i|treatment\ group, pre\right]$$

Now, our estimate of $E\left[Y_{0i}|treatment\ group, post\right]$ is the mean for the treatment group in the pre-period. The trouble with this is that if there is anything else causing Y to change over time, this estimator will incorrectly attribute the change to the treatment.

The DD estimator tries to solve the problems with these naive estimators by comparing CHANGES over time in the treatment and control groups. Formally, the DD estimator of δ is

$$\begin{aligned}\hat{\delta}_{DD} = &\left(\hat{E}\left[Y_i|treatment\ group, post\right] - \hat{E}\left[Y_i|treatment\ group, pre\right]\right) - \\ &\left(\hat{E}\left[Y_i|control\ group, post\right] - \hat{E}\left[Y_i|control\ group, pre\right]\right)\end{aligned}$$

You can think about this as correcting our second estimator above by removing other things that cause Y to change over time, where the effect of these things is estimated from the change in the control group. This interpretation makes the identification assumption clear: In the absence of the treatment, the treatment and control groups would move in parallel over time, so changes in the control group capture any time effects that would've affected the treatment group. This assumption is called NO DIFFERENTIAL TRENDS. DD attributes any difference in the changes over time to the treatment, so there can't be anything else causing such differences if the estimator is to be unbiased.

Note that by simply rearranging terms, we can rewrite the DD estimator as

$$\begin{aligned}\hat{\delta}_{DD} = &\left(\hat{E}\left[Y_i|treatment\ group, post\right] - \hat{E}\left[Y_i|control\ group, post\right]\right) - \\ &\left(\hat{E}\left[Y_i|treatment\ group, pre\right] - \hat{E}\left[Y_i|control\ group, pre\right]\right)\end{aligned}$$

This can be interpreted as correcting our first estimator from above by removing the noncomparability between treatment and control, where this noncomparability is estimated from the pre period. This makes it clear that for the DD strategy to work, any confounding differences between treatment and control must be constant over time. This is another way of stating the no differential trends assumption.

1.2 DD as Regression

It is clear from above that it is easy to compute a DD – we just need to compute 4 means. But getting a standard error could be a little complicated if there are covariances between these estimates. It is therefore often easier to compute the DD estimator from a regression. Consider the regression:

$$Y_{it} = \alpha + \beta Treat_i + \gamma Post_t + \delta Treat_i \cdot Post_t + \epsilon_{it}$$

How is this the same as the DD? Consider the coefficients associated with each possible combination of $Treat$ and $Post$:

	$E\left[Y_i\right]$
Treatment group, post	$\alpha + \beta + \gamma + \delta$
Control group, post	$\alpha + \gamma$
Treatment group, pre	$\alpha + \beta$
Control group, pre	α

Then the DD estimator is

$$\hat{\delta}_{DD} = \left(\left(\alpha + \beta + \gamma + \delta\right) - \left(\alpha + \gamma\right)\right) - \left(\left(\alpha + \beta\right) - \alpha\right)$$

$$= \beta + \delta - \beta$$

$$= \delta$$

so δ from this regression gives us the DD estimate. Note furthermore that in this regression setup, we could add controls for any variables that we think might be causing differential trends between treatment and control. The identifying assumption then becomes that there are no differential trends besides the ones we've controlled for.

1.3 Other DD Issues: Pre-periods, DDD, and inference

As we've said repeatedly, the identifying assumption required for DD to work is that in the absence of the treatment, the treated and control groups would move in parallel. In many cases this is not obvious. In the Eissa and Liebman paper, should we expect the labor force participation of women with and without children to move in parallel?

Fortunately, in cases where we have multiple periods, we can perform a check of the identifying assumption: We can see whether the treatment and control groups moved in parallel in periods before the treatment happened. This is like running a DD in time periods without a treatment. If you get a non-zero estimate, the trends are not parallel in the pre-period and this should make you worried about your main DD estimate. In general, you should graph the pre-trend whenever possible. In fact, the most convincing DD evidence is often graphical – if an author doing DD doesn't graph the data, be skeptical!

In the previous paragraph I argued that differential trends in the pre-period should make you skeptical about a DD estimate. Another way to respond to this finding is to try to correct for these differential trends by removing them from your DD estimate. That is, you run a DD in the pre-period, and subtract this from your DD estimate. This is called "differences in differences in differences" (DDD). If there 2 periods prior to the treatment, the DDD estimator would be

$$\hat{\delta}_{DDD} = \left[\left(\hat{E}\left[Y_i|treatment\ group, post\right] - \hat{E}\left[Y_i|control\ group, post\right]\right) - \left(\hat{E}\left[Y_i|treatment\ group, pre2\right] - \hat{E}\left[Y_i|control\ group, pre2\right]\right)\right] - \left[\left(\hat{E}\left[Y_i|treatment\ group, pre2\right] - \hat{E}\left[Y_i|control\ group, pre2\right]\right) - \left(\hat{E}\left[Y_i|treatment\ group, pre1\right] - \hat{E}\left[Y_i|control\ group, pre1\right]\right)\right]$$

You can also do DDD if you have additional groups that you think might capture differential trends between treatment and control. For example, in Eissa and Liebman, one might compare changes among highly educated women with and without kids, and take the difference in trends for this group out of the DD estimator for low educated women. Some people don't find DDD very convincing – if there are different levels and trends, maybe we should just conclude that the controls are a really bad comparison group!

In addition, in many DD analyses there are big issues with getting the right standard errors. A famous paper about this is Bertrand, Duflo and Mullainathan (2004). They show that OLS standard errors are badly biased downward in DD analyses that use US states as comparison groups, and that one must cluster at the state level to get reasonable standard errors. More generally, in many DDs it is appropriate to cluster your standard errors at the group level, in which case you effectively have many fewer observations than you thought. You will learn more about clustering in econometrics, and possibly in a future recitation in this class.

Finally, one under-discussed point is that DD analyses depend heavily on functional form assumptions. For example, no differential trends in logs IMPLIES differential trends in levels! One should be skeptical of DDs that are highly sensitive to specification of the dependent variable.

2 Panel Data Methods

In many data sets, we get to see repeated observations on the same units (people, firms, countries, etc.) over time. This is called "panel data." Formally, our data include $i = 1...N$ units and $t = 1...T$ time periods. Suppose we want to estimate a model like the following:

$$y_{it} = \alpha + \beta x_{it} + \epsilon_{it}$$

In such models we often decompose the error term ϵ_{it} into a permanent individual-specific component θ_i and an idiosyncratic error term η_{it}:

$$y_{it} = \alpha + \beta x_{it} + \theta_i + \eta_{it}$$

with

$$Cov(\alpha_i, \eta_{it}) = 0$$

$$Cov\left(\eta_{it}, \eta_{js}\right) = 0 \ \forall i \neq j, \ t \neq s$$

In addition let's suppose that

$$Cov(\eta_{it}, x_{it}) = 0$$

In this case the only bias we are worried about comes from θ_i; for now, we are assuming that the only potential omitted variables are things that are fixed over time. There are two standard ways to proceed from here.

2.1 Random Effects

Suppose we are comfortable making the assumption that

$$Cov\left(x_{it}, \theta_i\right) = 0$$

In this case, we have

$$Cov\left(x_{it}, \theta_i + \eta_{it}\right) = 0$$

so there is no correlation between our right-hand side variable and the composite error term. Then we know that OLS will be consistent! However, given the nature of our data, we can actually do even better than OLS. Note that

$$Cov\left(\theta_i + \eta_{it}, \theta_i + \eta_{is}\right) = Var\left(\theta_i\right),$$

so we have autocorrelation in the unobserved part of the model. Furthermore, we know the structure of this correlation – there is a common covariance between the error terms for observations on the same individual, and no other autocorrelation. In situations with a known non-spherical error structure, the most efficient estimator is Generalized Least Squares (GLS). For this panel model, GLS is called "Random Effects." You'll learn how to do GLS in econometrics.

2.2 Fixed Effects

In most cases, we won't be comfortable making the assumption that $Cov(x_{it}, \theta_i) = 0$. Instead, we view θ_i as a potential source of omitted variable bias. Fortunately, we can deal with this by simply controlling for θ_i! We can do this by directly including a vector of person-dummies in our regression:

$$y_{it} = \beta x_{it} + \sum_{j=1}^{N} \theta_j D_{ij} + \eta_{it}$$

Here D_{ij} is a dummy variable that is one if $i = j$ and zero else; each person gets their own dummy variable (note that I've now excluded the constant). We can just run this with OLS, knowing that including the person-dummies has eliminated any bias due to permanent unobserved characteristics. This procedure is called Fixed Effects (FE).

It is worth thinking more about how to interpret Fixed Effects estimates. Recall from the Frisch-Waugh theorem that we can obtain our estimates by first partialling out the person-dummies and then regressing y on the resulting residuals. For once, it will be easier to use matrices. Let's order the data with our T observations on person 1 first, followed by our T observations on person 2, etc. Let $X_{NT\times 1}$ be the vector containing the x_{it}. Then the coefficient vector from regressing X on the person-dummies is

$$(D'D)^{-1}D'X$$

and the residuals are given by

$$\tilde{X} = X - D\left(D'D\right)^{-1} D'X$$

where

$$D_{NT\times N} = \begin{bmatrix} 1_T & 0_T & \cdots & 0_T \\ 0_T & 1_T & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0_T & \cdots & 0_T & 1_T \end{bmatrix}$$

is our matrix of person dummies. Here 1_T is a column-vector of T 1's. Writing this out yields

$$\tilde{X} = X - D\left(\begin{bmatrix} 1_T' & 0_T' & \cdots & 0_T' \\ 0_T' & 1_T' & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0_T' & \cdots & 0_T' & 1_T' \end{bmatrix} \cdot \begin{bmatrix} 1_T & 0_T & \cdots & 0_T \\ 0_T & 1_T & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0_T & \cdots & 0_T & 1_T \end{bmatrix}\right)^{-1} D'X$$

$$= X - D\begin{bmatrix} T & 0 & \cdots & 0 \\ 0 & T & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & T \end{bmatrix}^{-1} D'X$$

$$= X - D\begin{bmatrix} \frac{1}{T} & 0 & \cdots & 0 \\ 0 & \frac{1}{T} & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{1}{T} \end{bmatrix} D'X$$

$$= X - \begin{bmatrix} 1_T & 0_T & \cdots & 0_T \\ 0_T & 1_T & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0_T & \cdots & 0_T & 1_T \end{bmatrix} \begin{bmatrix} \frac{1}{T} & 0 & \cdots & 0 \\ 0 & \frac{1}{T} & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{1}{T} \end{bmatrix} D'X$$

$$= X - \begin{bmatrix} \frac{1_T}{T} & 0_T & \cdots & 0_T \\ 0_T & \frac{1_T}{T} & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0_T & \cdots & 0_T & \frac{1_T}{T} \end{bmatrix} \begin{bmatrix} 1'_T & 0'_T & \cdots & 0'_T \\ 0'_T & 1'_T & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0'_T & \cdots & 0'_T & 1'_T \end{bmatrix} X$$

$$= X - \begin{bmatrix} \frac{1_{T\times T}}{T} & 0 & \cdots & 0 \\ 0 & \frac{1_{T\times T}}{T} & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{1_{T\times T}}{T} \end{bmatrix} \begin{bmatrix} X_{1_{T\times 1}} \\ X_{2_{T\times 1}} \\ \vdots \\ X_{N_{T\times 1}} \end{bmatrix}$$

$$= X - \bar{X}$$

where $\bar{X}$ is a matrix where each individual observation has been replaced with the mean for the relevant person. Then fixed effects is equivalent to estimating the regression

$$y_{it} - \bar{y}_i = \beta\left(x_{it} - \bar{x}_i\right) + \omega_{it}$$

That is, fixed effects estimates the model using deviations from person-means. This is called the "within" model because it uses only variation within persons and does not use the variation in mean x's and mean y's across people. Other things to know about fixed effects:

1. Fixed effects cannot be used to estimate the coefficients on time-invariant variables – there is no variation left in such variables once we take out the individual-specific means
2. With only 2 time periods, fixed effects is equivalent to first differences (with no constant; fixed effects with a time dummy is equivalent to first differences with a constant)
3. Fixed effects and differencing can make measurement error a lot worse. We will see this in a future recitation.

14.661 Labor Economics I
Fall 2010