14.771 Development Economics: Microeconomic Issues and Policy Models
Fall 2008

14:771: Recitation Handout #2
Differences-in-Differences and Clustering

Differences-in-Differences

There are many instances where a policy change occurs and has different impacts on different regions or individuals. Suppose we are trying to estimate the effect of this policy. There are two naïve ways to get at this. First, we could look solely in the cross-section: compare those who receive the program and those who do not. And second, look at the change in outcomes over time for those who receive the program. In the first case, our estimates would be biased by anything that may distinguish the receivers from the non-receivers. In the second case, our estimates would be biased by any time effects. A commonly used method to try and get around these problems is to use panel data, and in particular, differences-in-differences.

In the simplest case, we have a group of observations that should be impacted by the policy change - the "treatment group" - and a group of observations that should not be impacted by the policy - the "control group". We also need data on outcomes before and after the policy change. The important assumption that DD makes is the *no differential trends assumption*. In terms of our regression equation, this means that we assume that time and group effects enter in an additive linear manner (we can make this conditional on other covariates). Intuitively, this assumption states that while the treatment and control groups may have different outcome values, these outcomes would have followed the same trend in the absence of the policy change. That is, a graph of the outcomes for both groups (in a policy change free world) should be parallel. Think of some outcome, Y, in the absence of the program, we can then write

$$E\left[Y \mid treatment,\ post\right] = \alpha + \beta * post + \gamma * treated \tag{1}$$

Now, assume that the impact of the policy also enters linearly. How can we expand our equation to capture this impact?

$$E\left[Y \mid treatment,\ post\right] = \alpha + \beta * post + \gamma * treated + \delta post * treated \tag{2}$$

What should δ equal if there is no policy change? If there is a policy change (and it has a significant impact)?

Now we can see why before and after comparisons or treatment-control comparisons in the cross section will not work:

$$E\left[Y \mid post = 1, treatment = 1\right] - E\left[Y \mid post = 0, treatment = 1\right] = \beta + \delta \tag{3}$$
$$E\left[Y \mid post = 1, treatment = 1\right] - E\left[Y \mid post = 1, treatment = 0\right] = \gamma + \delta \tag{4}$$

We want to be able to difference out either β or γ. How does the outcome for the control group change over time?

$$E\left[Y \mid post = 1, treatment = 0\right] - E\left[Y \mid post = 0, treatment = 0\right] = \beta$$

so then we see that

$$E\left(Y_{post,treated}\right) - E\left(Y_{pre,treated}\right) - \left[E\left(Y_{post,control}\right) - E\left(Y_{pre,control}\right)\right] = \delta$$

which is our program impact. But - how strong is our additive linear assumption? Can you think of some reasons why it might be violated?

Problems with DD

- Need to assume that the timing of the program is exogenous: did not occur at this particular time or place because of the outcome measure.
- Dependent on the functional form (additive linearity).
- Need to cluster at a level that allows for serial correlation, in particular if you have a long time series and few policy changes. Improper clustering can lead to overrejection of the null. If you don't have very many groups, accounting for this is difficult (more on this later).
- Tells us the impact of a particular policy - that is, it gives us a reduced form impact, when we may be interested in something else (i.e. in the example of Esther's INPRES paper, the DD tells us the impact of the INPRES program on wages, not the impact of more schooling on wages).

Checks of DD Strategy

As we'll see in the clustering section, DDs are particularly prone to false positives - and hopefully you were able to think of many reasons why the DD identification assumptions could be violated. As such, it's very, very important to do specification checks when running DDs. I also find that the no differential trends assumption is best conveyed graphically - if you're reading a paper that runs DDs and never graphs the data, be skeptical - especially if they don't run careful specification checks.

- Use data for prior periods to redo the DD, i.e. compare periods 0 and -1. If you find an effect, maybe your estimate of period 0 versus 1 is spurious too. Always plot your data over time! You can plot after partialling out covariates if you think that you need to control for other factors before the DD assumptions are satisfied (but if you need to control for everything but the kitchen sink, you might worry...)
- Use alternative control group - your results should be similar if both control groups are good.
- Use outcomes that shouldn't be affected by policy change - significant results indicate that you may have problems
- See if treatment and control groups differ by covariates - much more likely to convince people if the groups are similar

Estimation of Difference-in-Difference Models

Estimation of DDs using regression is very straightforward - this is because the coefficient on a dummy variable (X) in a linear regression just gives the difference in means of the outcome: $\beta_x = E[Y \mid X = 1] - E[Y \mid X = 0]$. Here I'll go through how this can be done in two steps (useful for understanding Bleakley (2006)) and how we can take care of all estimation in one step. Using the difference in means property, we can run two separate regressions - one for the pre period and one for the post period to measure the difference in the outcome variable between treatment and control groups:

$$\begin{aligned} Y_{pre} &= \alpha_{pre} + \lambda_{pre} treat_{pre} + \varepsilon_{pre} \\ Y_{post} &= \alpha_{post} + \lambda_{post} treat_{post} + \varepsilon_{post} \end{aligned}$$

from (2) we see that $\text{plim}\hat{\lambda}_{pre} = \gamma$ and $\text{plim}\hat{\lambda}_{post} = \gamma + \delta$, so we can save these coefficients and run the regression:

$$\hat{\lambda}_i = \alpha + \delta post_i + \eta_i$$

which will give us the effect we're interested in. With two observations in the second step, this seems silly - can you think of a case where we'd have more than two observations? Now - what do you think about the standard errors for this procedure - are they likely to be correct? Why not?

Instead of running a two-step estimator, we can simply run one regression, lifted from (2). We regress

$$Y = \alpha + \beta * post + \gamma * treated + \delta post * treated + \Gamma X$$

(where X are other controls) and just read the coefficient off the interaction between the treatment and post period indicators.

Changes-in-Changes and Quantile Difference-in-Difference

One way of getting around the issue of functional dependence in a DD model is to use the Athey-Imbens Changes-in-Changes model (Econometrica, 2006). It is a lot less restrictive, since it is nonparametric. It allows for the effects of both time and treatment to differ systematically over individuals, and gives you the entire counterfactual distribution of effects of treatment on the treated group as well as the distribution of effects of treatment on the control group. Of course, it's not nearly as straightforward to estimate or interpret, but you should have a look if you're interested.

DD as an IV

All we have said above relates to the "reduced form" effect of a policy on an outcome. What if we want to have a treatment effect measure (as in Esther's paper). Please think carefully in the problem set about the added restriction one needs to impose to use a DD as an instrument. Do the DD conditions still need to hold?

Clustered Error Structures

Non-spherical error structure

OLS has minimum variance among linear estimators in one particular case: when the error term is homoskedastic and not serially correlated across units. When this is not the case, GLS has a better performance. However, GLS is almost never used by applied economists. Why is that?

Instead, most of the literature uses OLS (or IV) to obtain an unbiased estimate of the coefficient and then correct the standard errors. How bad will this correction be?

Let's think of the ratio of true standard errors to the OLS homoskedastic version when we have some correlation within groups. Specifically, let's think of the "Moulton" model. Here, groups are indexed by $g = 1, ...G$, and we assume that errors are correlated within group, but not across group. That is:

$$E\left[\varepsilon_{ig}\varepsilon_{jg}\right] = \rho\sigma_{\varepsilon}^{2}$$

where we call ρ the intraclass/intracluster correlation coefficient. In the special case where regressors are fixed at the group level (i.e. invariant within groups) and groups of equal size (n) we can do some math and see that

$$\begin{aligned}\frac{Var_{\text{cluster}}\left(\hat{\beta}\right)}{Var_{\text{homo}}(\hat{\beta})} &= \frac{\left(X'X\right)^{-1}X'\Omega X\left(X'X\right)^{-1}}{\hat{\sigma}_{\varepsilon}^{2}\left(X'X\right)^{-1}} \\ &= 1+\left(n-1\right)\rho\end{aligned}$$

We can generalize this to allow variable group sizes of n_g and independent variables that vary within groups (say individual income, when the group is the state), then we get that

$$\frac{Var_{\text{cluster}}\left(\hat{\beta}\right)}{Var_{\text{homo}}(\hat{\beta})} = 1+\left[\frac{V\left(n_g\right)}{\bar{n}}+\bar{n}-1\right]\rho_x\rho$$

where ρ_x is the intraclass correlation of the x_{ig}:

$$\rho_x = \frac{\Sigma_g\Sigma_{i\neq k}\left(x_{ig}-\bar{x}\right)\left(x_{kg}-\bar{x}\right)}{V\left(x_{ig}\right)\Sigma_g n_g\left(n_g-1\right)}$$

Thus, our unadjusted standard errors will be worse when we have more correlation across groups, a larger average group size, and more variable group sizes.

Ways to control for this problem

If we have only heteroskedasticity, we should use White's standard errors:

$$\left(X'X\right)^{-1}\left(\sum e_i^2 x_i x_i'\right)\left(X'X\right)^{-1}$$

This is what STATA calls ROBUST standard errors. We still assume that the observations are uncorrelated but they do not have the same variance. In general, this should be larger than the simple OLS standard errors. As a rule of thumb, you should always specify ", robust" after regressions (unless you're implementing another standard error correction, like clustering) - we don't think homoskedasticity is a particularly realistic assumption for most applications. This is particularly true if you're running a linear probability model, in which case there's no way that the errors can be homoskedastic. Also note - when you specify the "robust" option, STATA implements White standard errors - however, there are more conservative heteroskedasticity corrections - HC2 (White + a degrees of freedom correction) and HC3 (approximates a jackknife estimator), that you can also implement.

Second, if we have a good idea of the error structure, we can do FGLS. This should provide a more efficient estimate of the coefficients of interest. But how can we know the structure of the error term? If we have a time series process, we may want to assume an AR(1) process. If we have spatially correlated elements, we may know something about how the variance is correlated across geographical areas. And sometimes theory provides guidance.

However, in general, we do not have that type of info (and if your results hang on doing FGLS with an assumed error structure, your results aren't very strong). Thus, we must use a more flexible way to estimate the standard errors. Clustering assumes that there is correlation across the units within the same cluster but not across them. The nice thing here is that the correlations within the cluster can take any arbitrary form. The exact formula (for G clusters) is given by:

$$(X'X)^{-1}\left(\sum_{g=1}^{G}\hat{u}_g'\hat{u}_g\right)(X'X)^{-1}$$

where

$$\hat{u}_j=\sum_{i\in \text{cluster}_j}\hat{e}_{ij}x_{ij}$$

When will the clustered standard errors be smaller than the OLS or robust ones? When the inner-cluster correlation is negative. What does this mean? Does that seem reasonable to you?
In general, you should try to cluster to include all the variation that might be correlated. For example, in most DD, you do not want to cluster by State*Time because you want to allow for serial correlation within a state. Bertrand, Duflo and Mullainathan discuss the case of DD and standard errors more carefully in their paper. (This, by the way, is a great paper to read - they made a big contribution to applied econometrics by thinking carefully and making excellent use of Monte Carlo simulations.) However, BDM (and subsequent studies) show that the cluster correction does very poorly when the number of clusters is small as they are based on asymptotics where we send $G\rightarrow\infty$ and the small sample properties are poor (indeed, for small G, the estimated standard errors will often be smaller than the vanilla homoskedastic estimates!).

When you have $G\leq 25$, you should worry about using the cluster adjustment above. So what to do? There are a few options. Probably the most general solution is suggested by Cameron, Gelbach, and Miller (2008) who show that some forms of block bootstrap perform

well with small G (incidentally, the form used by BDM did not do so well! See their paper for more detail). You may also want to have a look at the estimator suggested by Donald and Lang (2007), although it comes with some restrictions. The real drawback here is that the power properties of these corrections appear to be very poor - which means you aren't very likely to find a significant effect even if there truly is one.

Finally, a recent paper from Cameron, Gelbach, and Miller (2006) provide new tools to compute multi-way clusters - you can see the paper for more detail.