PSC 504: Differences-in-differeces estimators

Matthew Blackwell

3/22/2013

Basic differences-in-differences model

Setup

- The basic idea behind a differences-in-differences model (shorthand: diff-in-diff, DID, or just DD) is that some group or groups are treated at a given point in time and we can compare the change before and after this intervention against the same change in other, untreated group.
- It seems ideal that we have before and after measurements for the treated group, but it might be the case that the treatment time is correlated with secular trends in the outcome. The role of the control group is to identify the secular trend in the outcome, so that we can separate secular trends from the causal effect of the treatment.
- The classic example of this is the Card and Krueger estimates of the effect of a change in the minimum wage in New Jersey on employment using the changes in employment in Pennsylvania, which did not undergo a change in minimum wage laws. On the syllabus, we have the Lyall paper that looks at the effect of artillery shelling on insurgent attacks, using villages that were not shelled as a control group.

Identification

- Again, let's have Y_{it} be the outcome under control for unit i in time period t. For now, we'll have two time periods, pre-treatment $t = 0$ and post-treatment $t = 1$. Of course, we have a treatment indicator: $A_{it} = 1$ for those units who received the treatment in time t. A special note in this literature is that no one is treated in the first period, so $A_{i0} = 0$ for all i.
- The typical way we motivate the DID estimator is using a linear parametric model, similar to how we justified fixed-effects models last week. So, first we will ignore any potential outcomes and then come to them later.
- The specific model we will assume is this:

$$Y_{it} = \delta_t + \tau A_{it} + \alpha_i + \eta_{it}$$

- Here we have a period effect, δ_t and a unit effect α_i, and a transitory shock, η_{it}, which has mean zero. Without further assumptions, τ is not identified, because it might be correlated with the transitory shocks. Thus, one identifying assumption, might be that the treatment is independent of the idiosyncratic error:

$$\Pr[A_{i1} = 1|\eta_{i0}, \eta_{i1}] = \Pr[A_{i1} = 1]$$

- This implies that functions of the errors are also independent of the treatment, so that $\eta_{i1} - \eta_{i0}$ is independent of A_{it}. This means that the any trends in the outcome are uncorrelated with the treatment. This is important because it means that both the treatment and control group would have the same trends in the outcome.

- With this assumption, we can rewrite the above model as the following:

$$Y_{it} = \mu + \delta t + \gamma A_{i1} + \tau A_{it} + \varepsilon_{it}$$

- The parameters are the following:

$$\begin{aligned}
\varepsilon_{it} &= \alpha_i - E[\alpha_i|A_{i1}] + \eta_{it} \\
\delta &= (\delta_1 - \delta_0) \\
\mu &= E[\alpha_i|A_{i1} = 0] + \delta_0 \\
\gamma &= E[\alpha_i|A_{i1} = 1] - E[\alpha_i|A_{i1} = 0]
\end{aligned}$$

- The first of these is just a new error, the second is the time trend, the third is the initial mean for the control group and the last is the difference between the treatment and control groups in terms of their individual effects.

- Using the above assumption, we can show that the treatment is independent of the error in this model:

$$\begin{aligned}
E[\varepsilon_{it}|A_{i1}, A_{i0}] &= E[\varepsilon_{it}|A_{i1}] \\
&= E[(\alpha_i - E[\alpha_i|A_{i1}] + \eta_{it})|A_{i1}] \\
&= E[\alpha_i|A_{i1}] - E[E[\alpha_i|A_{i1}]|A_{i1}] + E[\eta_{it}|A_{i1}] \\
&= E[\eta_{it}|A_{i1}] \\
&= E[\eta_{it}] = 0
\end{aligned}$$

- Note that we have this even though we have made no assumptions on the distribution of the unit-specific effects and their relation to the treatment.

- Now, we can investigate how two differences here. First, the time trend for the untreated:

$$E[Y_{i1}|A_{i1} = 0] - E[Y_{i0}|A_{i1} = 0] = \delta$$

- And now the trend for the treated group:

$$E[Y_{i1}|A_{i1} = 1] - E[Y_{i0}|A_{i1} = 1] = \delta + \tau$$

- This motivates the differences-in-differences estimator as the difference between these two differences. We can estimate each of these CEFs from the data and compute their sample versions to get an estimate of τ.

Estimation

- For the two period, binary treatment case, a regression of the outcome on time (pre-treatment, post-treatment), treated group, and their interaction can estimate τ. As indicated by the above model, $\hat{\tau}$ would be the coefficient on the interaction between time and the treatment.
- Note that, for this, we only need to two cross-sections, one from before the treatment and one from after in the groups.
- If we have panel data, then we can estimate this a different, more direct way. Note that:

$$\tau = E[Y_{i1} - Y_{i0}|A_{i1} = 1] - E[Y_{i1} - Y_{i0}|A_{i1} = 0]$$

- Thus, in the panel data case, we can estimate the effect by regressing the change for each unit, $Y_{i1} - Y_{i0}$, on the treatment.

Threats to identification

- Obviously, the treatment needs to be independent of the idiosyncratic shocks so that the variation of the outcome is the same for the treated and control groups, but this might not be plausible.
- One example from economics is Ashenfelter's dip, which is a empirical finding that people who enroll in job training programs see their earnings decline prior to that training. In the Lyall paper, it might be the case that insurgent attacks might be falling in places where there is shelling because rebels attacked in those areas and have moved on.
- Thus, the independence of the treatment and idiosyncratic shocks might only hold conditional on covariates.

Regression DD

- When the key assumption only holds conditional on covariates, then we have to control for those covariates in some way. The usual way to do this is with a regression DID, which includes covariates in a linear, additive manner:

$$Y_{it} = \mu + X_i'\beta_t + \delta t + \gamma A_{i1} + \tau A_{it} + \varepsilon_{it}$$

- If we have repeated observations, we can take the differences between $t = 0$ and $t = 1$:

$$Y_{i1} - Y_{i0} = \delta + X_i'\beta + \tau(A_{i1} - A_{i0}) + (\varepsilon_{i1} - \varepsilon_{i0})$$

- Here, we have $\beta = \beta_1 - \beta_0$. Further note that because everyone is untreated in the first period, $A_{i1} - A_{i0} = A_{i1}$.
- Thus, if we have repeated observations on the same units, we can simply regress the change in Y_i on the treatment and the covariates.
- You'll note from our earlier work that a regression like this is going to depend on the constant effects assumption (in addition to the linearity of the effect ofX_i). We may want to generalize this.

Serial correlation and placebo tests

- Bertrand, Duflo, and Mullainathan (2004) shows that many DID applications suffer from issues of serial correlation in the dependent variable.
- Serial correlation doesn't necessarily affect the consistency of the estimator, but rather the calculation of the standard errors.
- They use placebo tests to show that serial correlation can lead to a random, no-effect intervention being significant 45% of the time at the 5% level.
- What is a placebo test? It's a very useful way to assess if there is something questionable going on in a DID application. In Bertrand, Duflo, and Mullainathan (2004), they took data on female wages from the Current Population Survey from 1979 until 1999. Then, they randomly drew a year between 1985 and 1995 and then randomly drew half of the states in the U.S. to receive this treatment.
- Obviously, since they created these interventions randomly, they have no effect. Thus, in this case, they should only register as significant 5% of the time at the 0.05 level. What they found is that these placebo interventions could be significant up to 50% of the time. Pretty terrible.
- Why do they find this? It's the fact that DID time-series tend to be very serially correlated and that most DID applications use long pre-treatment and post-treatment series. That is, we have been looking at a two-period case, but you can imagine adding in many periods before and after the treatment.
- How can we correct for this? Two ways: a block bootstrap (especially if there are many units). Or, simply aggregate across the pre-intervention and across the post-intervention time periods.

Nonparametric identification

- The last section described how the DID model is identified parametrically—using a model for the outcome. Now, we want to relax that and see how we can identify effects without a model.
- Let $Y_{it}(a)$ be the potential outcome under treatment a at time t. Again, the individual causal effect is just $Y_{it}(1) - Y_{it}(0)$. Because no one is treated at time $t = 0$, we have $A_{i1} = A_i$ with $Y_{i0}(0) = Y_{i0}$ and $Y_{i1} = A_i Y_{i1}(1) + (1 - A_i)Y_{i1}(0)$.
- We'll focus on two estimands, the ATT, $\tau_{ATT} = E[Y_{it}(1) - Y_{it}(0)|A_i = 1]$ and the conditional ATT: $\tau_{ATT}(x) = E[Y_{it}(1) - Y_{it}(0)|X_i = x, A_i = 1]$.
- Let's make the crucial identifying assumption of a DID model:

$$E[Y_{i1}(0) - Y_{i0}(0)|X_i, A_i = 1] = E[Y_{i1}(0) - Y_{i0}(0)|X_i, A_i = 0]$$

- What does this assumption say? It says that the potential trend under control is the same for the control and treated groups, conditional on covariates. Of, equivalently, the treated group would have seen a similar trend to the control group if that had been control instead. This is often called the "parallel trends" assumption.
- Note that, if the two groups have the same mean potential outcome under control in the first period, $E[Y_{i0}(0)|X_i, A_i = 1] = E[Y_{i0}(0)|X_i, A_i = 0]$, then this assumption just becomes regular ignorability: $E[Y_{i1}(1)|X_i, A_i = 1] = E[Y_{i1}(1)|X_i, A_i = 0]$.

- We can show that this is the key assumption for identifying the DID approach:

$$\begin{aligned}
E[Y_{i1}(1) - Y_{i1}(0)|X_i, A_i = 1] =& E[Y_{i1}(1) - Y_{i0}(0) + Y_{i0}(0) - Y_{i1}(0)|X_i, A_i = 1] \\
=& (E[Y_{i1}(1)|X_i, A_i = 1] - E[Y_{i0}(0)|X_i, A_i = 1]) \\
& - (E[Y_{i1}(0) - Y_{i0}(0)|X_i, A_i = 1]) \\
=& (E[Y_{i1}(1)|X_i, A_i = 1] - E[Y_{i0}|X_i, A_i = 1]) \\
& - (E[Y_{i1}(0) - Y_{i0}(0)|X_i, A_i = 0]) \\
=& (E[Y_{i1}|X_i, A_i = 1] - E[Y_{i0}|X_i, A_i = 1]) \\
& - (E[Y_{i1}(0)|X_i, A_i = 0] - E[Y_{i0}(0)|X_i, A_i = 0]) \\
=& (E[Y_{i1}|X_i, A_i = 1] - E[Y_{i0}|X_i, A_i = 1]) \\
& - (E[Y_{i1}|X_i, A_i = 0] - E[Y_{i0}|X_i, A_i = 0])
\end{aligned}$$

- This is just the DID estimator that we saw above. Thus, we could estimate each of these values non-parametrically, but we are going to run into the curse of dimensionality if X_i has many dimensions or has continuous components (sound familiar?). Here we have four CEFs to estimate per level of X_i. If we have repeated observations, we can take differences and estimate only two CEFs.
- Note what is powerful here: we didn't have to make any kind of ignorability assumption, either conditional on a unit-specific effect or not. All we had to do was assume there are parallel trends. Of course this assumption will be more plausible when the treated and control group are similar.
- Also, it's important to note that the parallel trends assumption may not hold for transformations of the data. So if it holds on the original data, it will not hold on the logged data and vice versa.
- In order to make progress with assuming a model for the relationship between X_i and Y_{it}, we can instead take a weighting approach, first described by Abadie (2005).

Semiparametric estimation with repeated outcomes

- With repeated outcomes, we can take a weighting approach to estimating the effect. This is similar to the weighting approach taken with selection on the observables:

$$E[Y_{i1}(1) - Y_{i1}(0)|X_i, A_i = 1] = E\left[\frac{A_i(Y_{i1} - Y_{i0})}{\Pr[A_i = 1|X_i]} - \frac{(1 - A_i)(Y_{i1} - Y_{i0})}{1 - \Pr[A_i = 1|X_i]}\Big|X_i\right]$$

- The tradeoff here is that we have to estimate the propensity score to estimate these weights for each unit:

$$\rho_0 = \frac{A_i - \Pr[A_i = 1|X_i]}{\Pr[A_i = 1|X_i](1 - \Pr[A_i = 1|X_i])}$$

- We can see how this works with the following proof:

$$\begin{aligned}
E[\rho_0(Y_{i1} - Y_{i0})|X_i] =& E[\rho_0(Y_{i1} - Y_{i0})|X_i, A_i = 1] \Pr[A_i = 1|X_i] \\
& + E[\rho_0(Y_{i1} - Y_{i0})|X_i, A_i = 0] \Pr[A_i = 0|X_i] \\
=& E[Y_{i1} - Y_{i0}|X_i, A_i = 1] - E[Y_{i1} - Y_{i0}|X_i, A_i = 0]
\end{aligned}$$

Robustness checks

Lags and Leads

- One thing we would like to do is perhaps check that the intervention really does occur before its effect. We know that if A_{it} causes Y_{it}, and not the other way around, then current and lagged values of A_{it} should have an effect on Y_{it}, but future values of A_{it} should not.
- Thus, one robustness check is to include lags and leads of the treatment in a regression DID and see if the effects follow this pattern. If the leads of the treatment matter, this might be problematic for the estimation. It might mean that the parallel trends assumption is violated.
- Also note that the lagged values might substantively interesting, but it isn't clear that they have any causal interpretation, as we saw last time with the fixed-effects models.

Time trends

- If we have more than two periods, we can add unit-specific linear trends to the regression DID model can help to assess whether or not the parallel trends assumption is problematic. That is, it allows each state to have its own trend, which can be estimated from the pre-treatment data. Thus,