ARE 213 **Applied Econometrics**

UC Berkeley Department of Agricultural and Resource Economics

Selection On Unobservables Designs:

Part 2, Differences-in-Differences

The most common research design for policy analysis with panel data is the differences-in-differences model. In its simplest incarnation, the diffs-in-diffs model entails identifying two cross-sectional units (states, cities, countries, etc.), one of which was exposed to a policy change (or some other treatment) and the other of which was not. With longitudinal data, we collect information on the two units both before the policy change and after the policy change. To estimate the effect of the policy on a given outcome, we simply compare the change in the outcome for the treated unit to the change in the outcome for the control unit.

1 Differences-in-Differences

Suppose that we observe two states, $s = 0$ and $s = 1$, one of which is affected by a policy change and the other of which was not. Further suppose that we observe these states for two time periods, $t = 0$ (pre-policy change) and $t = 1$ (post-policy change). Formally, for some outcome Y_{ist} that we observe at the individual level, the differences-in-differences estimator is

$$(\overline{Y}_{11} - \overline{Y}_{10}) - (\overline{Y}_{01} - \overline{Y}_{00})$$

where $\overline{Y}_{st} = \frac{1}{N_{st}} \sum_i Y_{ist}$. To examine the strengths and weaknesses of this estimator, write $Y_{ist} = \overline{Y} + \tau D_{st} + \gamma_s + \delta_t + \varepsilon_{st} + u_{ist}$. Note that the inclusion of ε_{st} guarantees that $\overline{u}_{st} = 0$.

$$(\overline{Y}_{11} - \overline{Y}_{10}) - (\overline{Y}_{01} - \overline{Y}_{00}) =$$

$$[(\overline{Y} + \tau + \gamma_1 + \delta_1 + \varepsilon_{11}) - (\overline{Y} + \gamma_1 + \delta_0 + \varepsilon_{10})] - [(\overline{Y} + \gamma_0 + \delta_1 + \varepsilon_{01}) - (\overline{Y} + \gamma_0 + \delta_0 + \varepsilon_{00})] =$$

$$(\tau + \delta_1 - \delta_0 + \varepsilon_{11} - \varepsilon_{10}) - (\delta_1 - \delta_0 + \varepsilon_{01} - \varepsilon_{00}) =$$

$$\tau + (\varepsilon_{11} - \varepsilon_{10}) - (\varepsilon_{01} - \varepsilon_{00}) \qquad \mathbb{E}\left[\varepsilon_{11} - \varepsilon_{10}\right] = \mathbb{E}\left[\varepsilon_{01} - \varepsilon_{00}\right]$$

parallel trends assumption ⇒ unbiasedness
⇏ consistency

The key assumption for identifying τ will therefore be $E[\varepsilon_{11} - \varepsilon_{10}] = E[\varepsilon_{01} - \varepsilon_{00}]$. In other words, the outcomes for the two states must have similar trajectories over the two time periods absent any treatment effect. Any factor that is specific to state s but does not change over time, or changes over time but changes in equal amount for both states, is netted out in the diffs-in-diffs estimator.

It is important to note, however, that the condition above only guarantees that we will identify τ *in expectation*. Because we only observe a single observation for each of the ε_{st} terms in the expression above, there is no guarantee that the noise from ε_{st} will not swamp our estimate of the treatment effect, τ – we cannot appeal to the law of large numbers as we do when we have N independent observations. This is an issue that we will return to shortly.

The diffs-in-diffs estimator can also be easily implemented within a regression framework. Consider running the regression:

$$Y_{ist} = \alpha + \tau D_{st} + \gamma \mathbf{1}(s = 1) + \delta \mathbf{1}(t = 1) + \varepsilon_{st} + u_{ist}$$

* cluster SE at level of treatment

if we have a reason to think spacial correllation is important ⇒ theres a fancy way to include that

In other words, simply regress Y on a treatment indicator, a state dummy, and a time dummy. The state dummy controls for between-state differences in Y that are constant over time, and the time dummy controls for between-time period differences in Y that are identical across states. Identification of τ again comes from the assumption that ε_{st} is uncorrelated with the treatment indicator (which is equal to the interaction between the state dummy and the time dummy) conditional on the state dummy and the time dummy. Note that in the regression format, it is easy to control for individual-level covariates. You can also see the standard errors issue in this framework. If we use the typical OLS standard errors that assume independence across all observations, we are effectively claiming that the only error in our estimator is sampling error that arises because we do not observe the entire population of each state. However, if $\sigma_\varepsilon^2 \neq 0$, i.e. there are state-specific shocks that vary over time, then this independence assumption is violated, and our standard errors will be wrong.

2 Triple Differences

A diffs-in-diffs research design can sometimes be made more compelling by adding another layer of differencing to the estimator, resulting in a triple-diffs estimator. For example, consider a policy change in state 1 in time period 1 that only affects persons 65 years and older. In that case, we might use individuals aged 55-64 as an additional "control" group. In practice, we would implement this with a triple differences estimator. Let $\overline{Y}_{sta}$ be defined as above, but with $a = 0$ signifying persons of age 55-64 and $a = 1$ signifying persons of age 65 and older. Then the triple differences estimator is:

$$[(\overline{Y}_{111} - \overline{Y}_{110}) - (\overline{Y}_{101} - \overline{Y}_{100})] - [(\overline{Y}_{011} - \overline{Y}_{010}) - (\overline{Y}_{001} - \overline{Y}_{000})]$$

In other words, we compare the evolution of the gap between 65+ year olds and 55-64 year olds in the treated state to the evolution of the gap between 65+ year olds and 55-64 year olds in the control state. The advantage of this triple-diffs structure is that it allows us to relax our assumptions on ε_{st}. We no longer need to assume that outcomes for both states would evolve similarly in expectation – we now need only assume that, to the extent that outcomes evolve differently in state $s = 1$ than state $s = 0$, the differences affect age groups $a = 1$ and $a = 0$ similarly.

We can easily implement this triple-diffs estimator within the regression framework. The key is to put in an indicator for every main effect or interaction up to, but not including, the level at which the treatment varies. Thus we include main effects for age, state, and time, as well as all possible two-way interactions between each of those indicators. The regression looks like:

$$Y_{ista} = \alpha + \tau D_{sta} + \gamma_1 \mathbf{1}(s=1) + \gamma_2 \mathbf{1}(t=1) + \gamma_3 \mathbf{1}(a=1) + \gamma_4 \mathbf{1}(s=1)\mathbf{1}(t=1)$$

$$+\gamma_5 \mathbf{1}(s=1)\mathbf{1}(a=1) + \gamma_6 \mathbf{1}(t=1)\mathbf{1}(a=1) + \varepsilon_{sta} + u_{ista}$$

3 Applications: Card (1990), Card & Krueger (1994), Kellogg & Wolff (2008)

Two canonical examples of diffs-in-diffs papers are Card's (1990) study of the Mariel Boatlift and Card and Kruger's (1994) study of the minimum wage increase in New Jersey.[1] The Mariel Boatlift occurred from May to September of 1980 when Cuba allowed any citizen wishing to emigrate to the United States free passage from the port of Mariel. Approximately 125,000 Cuban immigrants arrived in Miami during this time period, increasing the local labor force by about 7%.

Card examines wage and employment outcomes for various groups of natives, particularly blacks and lower-skilled workers – the latter group is more likely to be in direct competition with the newly arrived immigrants (who were relatively low-skilled). He compares the evolution of these outcomes over the 1979 to 1981 period in Miami to their evolution in four comparison cities: Atlanta, Los Angeles, Houston, and Tampa-St. Petersburg. For blacks, the difference in log wages between Miami and comparison cities changes from -0.15 in 1979 to -0.11 in 1981, so the diffs-in-diffs estimate for log wages is 0.04 (with a standard

[1] The term "differences-in-differences" is thrown around often, but to my knowledge there is no formal definition for what classifies as a "diffs-in-diffs" paper. Arguably many panel data papers that control for both individual-specific effects and aggregate time effects are using some form of double differencing estimator, but that doesn't mean that we'd necessarily refer to them as diffs-in-diffs papers. In my mind, a diffs-in-diffs paper generally uses some sort of variation in the treatment that occurs at an aggregated level, e.g. the city or state level. We therefore tend not to worry so much about individuals selecting into the treatment (it's unlikely that most will move in response to just one shock), but rather we worry that the treatment was implemented in one area rather than another for some non-random reason (e.g., legislative endogeneity).

error of about the same size). The difference in the employment-to-population ratio between Miami and comparison cities changes from 0.00 in 1979 to 0.02 in 1981, so the diffs-in-diffs estimate for employment is 0.02. Estimates for unemployment rates and low-skilled blacks show similar patterns. Overall, there is no evidence that immigrants harm natives' labor market outcomes.[2]

Card and Krueger (1994) study the impact of a 19% increase in the New Jersey minimum wage in 1992. They survey fast-food restaurants before and after the change on both sides of the New Jersey-Pennsylvania border. Examining average employment per store in the diffs-in-diffs framework they find:

$$(Emp_{NJ,1} - Emp_{NJ,0}) - (Emp_{PA,1} - Emp_{PA,0}) = (21.03 - 20.44) - (21.17 - 23.33) = 2.75$$

Thus we see that, if anything, the minimum wage increase appears to have *raised* employment (though the increase is not statistically significant). Because the findings run counter to the predictions of economic theory, they have been heavily scrutinized and criticized, though they survive a battery of robustness checks and have been replicated in other settings. One arguably valid criticism, however, concerns the standard errors. As we saw earlier, if the ε_{st} terms have non-zero variance, then we cannot accurately compute the standard errors because we do not have enough observations to compute the variance for the state-by-time level shocks. In essence, despite the fact that the Card (1990) and Card and Krueger (1994) studies contain hundreds or even thousands of observations, both are essentially case studies in that the treatments vary only at the city or state level and the studies contain only a few cities or states.[3] We will return to this issue shortly.

Kellogg and Wolff (2008) provide a nice example of a triple differences research design. Their interest is in estimating the effect of Daylight Savings Time (DST) on electricity usage. DST may reduce energy usage because, for example, it aligns the hours at which people are

[2] Interestingly, for Americans of Cuban descent, Card finds evidence of some increase in unemployment rates, but no effect on wages.

[3] This fact is not lost on the authors. The title of Card and Krueger (1994), for example, reads, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania."

awake with the hours at which the sun is up, thus reducing lighting needs. On the other hand, it may increase energy usage because people wake up when the sun rises (as opposed to after it has risen) and need to heat their homes during this time.

Kellogg and Wolff leverage an extension to DST in Australia that was put in place for the Summer 2000 Olympics. Some Australian states, including New South Wales (where the Olympics were held) and Victoria, extended DST beyond the date at which normally terminates. Other states, including South Australia, did not. They compare the change in electricity usage for Victoria (the "treated" state) to the change in electricity usage for South Australia (the "control" state). They are concerned, however, that electricity usage might be trending differently in these two states for reasons unrelated to the DST extension.

To address this concern, they observe that DST should not affect electricity usage during the middle of the day, when the sun is always in the sky regardless of whether you are on DST or standard time. The midday hours thus provide an extra "control" group that should be unaffected by DST. This allows them to implement a triple differences estimator. Specifically, they define the treated portion of the day as the hours from 0:00 to 12:00 and 14:30 to 24:00. They define the control portion of the day as the hours from 12:00 to 14:30.

Using a simple differences-in-differences estimator with electricity usage during the treated portion of the day as the outcome, they find that electricity usage fell by 0.4% in Victoria as compared to South Australia. However, they also find that electricity usage during the control portion of the day fell by 0.2% in Victoria as compared to South Australia. The triple differences estimator takes the difference between these two double differences estimators; thus their final estimate is that DST reduced electricity usage by 0.2% (with a standard error of 1.5%). The identifying assumption here is that, if Victoria and South Australia are trending differently from each other, these differential trends still have the same proportional effect on electricity usage from 12:00-14:30 and electricity usage from 0:00-12:00/14:30-24:00.

To increase the precision of their estimates, they also implement the triple differences estimator in a regression framework. The regression framework allows them to control for other determinants of electricity usage (e.g., day of week, weather, etc.). This reduces the unexplained variation in the outcome and thus reduces their standard errors. An observation in this regression is the half-hour-by-day-by-state. They regress electricity usage on the treatment variable (one if DST is in effect and it is before 12:00 or after 14:30, zero otherwise) and day-by-state indicators (which basically correspond to the state-by-time interactions in Section 2), hour-by-state indicators (which basically correspond to the state-by-age interactions in Section 2), hour-by-year indicators (which basically correspond to the time-by-age interactions in Section 2), and other control variables.

In the regression framework, they find that DST *increases* energy usage by 0.02% (if they impose a homogeneous effect across all treated hours) or 0.09% (if they allow for heterogeneous effects of DST across different times of day). The standard error drops to 0.4%, so they are able to rule out substantial electricity savings from DST – savings of 0.5% or higher, for example, are unlikely.

4 Event Study Designs

In practice, the data sets in most diffs-in-diffs analyses include multiple periods before and after treatment. With multiple observations prior to and following treatment, we can trace out "dynamic" treatment effects, examine whether the treatment effect materializes immediately, and look for evidence of differential pre-trends between treated and control units. We refer to regression models that estimate effects period by period as "event study" designs.

Consider the simple case of a single treated state ($s = 1$) and a single control state ($s = 0$), but assume that T is now greater than 2 (often much greater). Let T_0 be the period just prior to treatment; T_0 thus corresponds to the number of pre-treatment periods in the data set. Let D_{jst} be an indicator function for period t falling j periods after T_0 in the treated state (i.e., $\mathbf{1}(t - T_0 = j) \cdot \mathbf{1}(s = 1)$). We refer to the index j as "event time", and D_{jst} thus indicates whether the observation you are examining corresponds to event time j

in the treated state. The full event study design corresponds to estimating the regression:

$$Y_{ist} = \alpha + \sum_{j=-T_0}^{T-T_0} \tau_j D_{jst} + \gamma_s + \delta_t + \varepsilon_{st} + u_{ist}$$

$T_0 = 1 \quad T = 2$

$\sum_{j=-T_0}^{T-T_0} \quad j = -1 \quad 0 \quad 1$

$\sum_{j=-1}^{2-1=1}$

Note that the sum contains T dummy variables, which fully saturates event time. But in practice you cannot estimate this regression due to the dummy variable trap — the sum of the event time indicators (the D_{jst}) is colinear with the treated state's fixed effect. So we need to omit one of the event time dummies — typically the one with coefficient τ_{-1} (i.e. the one immediately preceding the treatment date) — and interpret the effects relative to that date.

In the regression above, τ_j represents the policy's effect j periods after the policy is implemented. If $j < 0$, then we're looking at pre-treatment periods, so τ_j should be zero. Examining the τ_j estimates for negative event time is thus a test for differential pre-trends. For $j > 0$, the event study coefficients τ_j allow you to trace out "dynamic" effects that may evolve over time. More generally, the patterns allow you to assess the credibility of the results and may yield evidence on the underlying data generating process. For example, an abrupt change in the outcome, immediately following treatment, is often viewed as more credible than a gradual change over time, because it is less likely to be generated by differential trends between the treated and control states. An abrupt change followed by a further upward trend could be evidence that some impacts of the policy only accumulate over sufficiently long time periods (particularly if there was no evidence of differential trends in the pre-treatment period).

While the framework is straightforward in the simple case with one treated state and one control state, it becomes more complicated when there are multiple treated states, with event dates that vary across units. For example, one state may have been treated in 2014, and another may have been treated in 2016. In this scenario, "event time" is no longer colinear with actual time, and the distribution of event time becomes unbalanced across units. Early event time dummies are missing for early-treated units, and late event time dummies are missing for late-treated units. Furthermore, if there are no “pure control” states (i.e. states that never get treated during your sample), then a second colinearity can arise even after omitting a single event time dummy. This second colinearity arises because event time equals calendar time minus treatment date. Without pure controls, time and treatment date fixed effects (the latter of which are proxied for by the state fixed effects), plus a full set of event time dummies (with one omitted), will be colinear.

To avoid this colinearity, and to address the unbalanced distribution of event time across units, it is standard practice to create “min” and “max” bins for early and late event time periods. For example, if $T = 40$ in your data, you might have event time dummies for $\tau_{-8}, \tau_{-7}, ..., \tau_{-2}, \tau_0, ..., \tau_{12}$, and then two binned dummies for $j < -9$ and $j > 12$. Note that the diffs-in-diffs estimator is in fact a special case of the event study design, in which all pre-treatment periods are binned into a single dummy (which is omitted from the regression), and all post-treatment periods are binned into a single dummy (i.e., the treatment indicator).