Applied Economics

Evaluating Policies Using Pooled Cross-Sections

Department of Economics
Universidad Carlos III de Madrid

See also Wooldridge, chapter 13

Policy Analysis in Pooled Cross-Sections: Differences-in-Differences

- In many cases the variable of interest changes over time for a group of individuals, a whole state, a cohort. For example, state policies regarding unemployment benefits for long term unemployed may change over time but are fixed across workers within states.
- The source of omitted variable bias in those cases might be the presence of unobserved variables at the state and year level.
- For these cases we use the Differences-in-Differences (DD or diff-in-diff) identification strategy.
- For this strategy we can use pooled cross-sections.

Differences-in-Differences: intuition

- Suppose we randomly assign treatment to some units (or treatment can be considered “as if” randomly assigned).
- To estimate the treatment effect, we could just compare the treated units before and after treatment.
- The problem is that we might pick up the effects of other factors that changed during the treatment.
- Therefore, we use a control group to discount these confounding factors and isolate the treatment effect.

Differences-in-Differences: the first example 1/2

- The first example of DD seems to be Snow (1855). This physician wanted to test the hypothesis that cholera was transmitted by contaminated drinking water. In 1854, he found a natural experiment useful to address this question.
- In 1849 two water companies, SVC and LC, got their water from the dirty Thames in central London. In 1852 only one of these companies, LC, started to obtain its water supply from a different and cleaner part of the Thames.
- A naive analyst could have compared death rates in the districts serviced by the company that changed the source of the water, in 1852 vs 1849.

Differences-in-Differences: the first example 2/2

- But, as we said, we might pick up the effects of other factors that changed in London between 1849 and 1852 (like new methods to fight cholera).
- Snow used as a control group the districts served by the company that did not change its water supply.
- He compared changes in death rates from cholera between 1854 and 1849 in districts supplied by these two water companies: SVC and LC.
- He found that death rates in districts supplied by LC fell in comparison to the change in death rates in districts supplied by SVC.

New Notation

- $Y(t)$: observed outcome in period t
- $Y_1(t)$: outcome in period t (if treated)
- $Y_0(t)$: outcome in period t (if not treated)
- $t = 0$: before treatment
- $t = 1$: after treatment

$$Y(t) = D * Y_1(t) + (1 - D) * Y_0(t)$$

Differences-in-Differences: a simple computation

- One approach is to take the difference of the averages between treated and controls before (D_1) and after treatment (D_2), and compute the "difference of the differences".

Average Outcome Variables by Group and Period

	Treated	Control	Difference
Before	$E[Y_0(0)\|D=1]$	$E[Y_0(0)\|D=0]$	D_1
After	$E[Y_1(1)\|D=1]$	$E[Y_0(1)\|D=0]$	D_2

- The "diff-in-diff" is then: $DID = D_2 - D_1$

$$\begin{aligned} DID = \quad & \{E[Y_1(1)|D=1] - E[Y_0(1)|D=0]\} \\ & -\{E[Y_0(0)|D=1] - E[Y_0(0)|D=0]\} \end{aligned}$$

Differences-in-Differences: a simple computation

- It is also possible to compute the observed change for the treated (D_T), for the controls (D_C) and then calculate the "difference of the differences".

Average Outcome Variables by Group and Period

	Treated	Control	Difference
Before	$E[Y_0(0)\|D=1]$	$E[Y_0(0)\|D=0]$	D_1
After	$E[Y_1(1)\|D=1]$	$E[Y_0(1)\|D=0]$	D_2
Difference	D_T	D_C	DID

- In this case: $DID = D_T - D_C$

$$
\begin{aligned}
DID = \quad & \{E[Y_1(1)|D=1] - E[Y_0(0)|D=1]\} \\
& - \{E[Y_0(1)|D=0] - E[Y_0(0)|D=0]\}
\end{aligned}
$$

Snow example

Snow example: death rates per 100,000 indiv.

	Districts served by LC	Districts served by SVC
1849	131	165
1853	60	114

Two ways:

$D_2 = (60 - 114)$ y $D_1 = (131 - 165)$: $DID = -54 - (-34) = -20$

$D_T = (60 - 131)$ y $D_C = (114 - 165)$: $DID = -71 - (-51) = -20$

Effect on the treated: α_{ATT}

How does the treatment affect those who are treated?

- $\alpha_{ATT} = E\left[Y_1(1) - Y_0(1)|D = 1\right]$

- Problem: We do not observe the outcome without the treatment among those who were treated, in the period after the treatment: $E[Y_0(1)|D = 1]$.
- Idea: use the control group to get this value.

What information do we have?

Both in pooled cross-sections and panel data we can consistently estimate:

- the average output for those treated under non-treatment before they were treated: $E[Y_0(0)|D=1]$
- the average change in output for those non-treated under no treatment: $E[Y_0(1)-Y_0(0)|D=0]$

The Parallel Paths Assumption

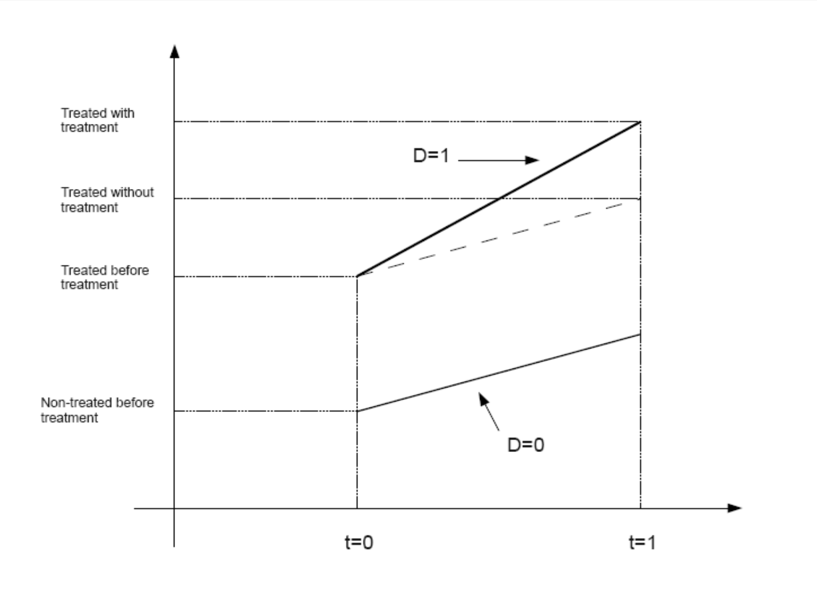
Parallel paths Assumption

- The output change for those treated would have been the same as the output change for those non-treated if there had been no policy change.

$$E[Y_0(1) - Y_0(0)|D = 1] = E[Y_0(1) - Y_0(0)|D = 0]$$

- Example: Death rates in districts served by LC company would have decreased just as in the districts served by SVC if LC had not changed its water. If there had been no change in water supply we would have expected a decrease in death rates in LC districts of 51, instead of the observed 71.

A Graphical Interpretation



Estimation of α_{ATT}

$$\alpha_{ATT} = E[Y_1(1) - Y_0(1)|D=1] = E[Y_1(1)|D=1] - E[Y_0(1)|D=1]$$

- Under the parallel paths assumption:

$$E[Y_0(1)|D=1] = E[Y_0(0)|D=1] + E[Y_0(1) - Y_0(0)|D=0]$$

- Then:

$$\begin{aligned}\alpha_{ATT} &= E[Y_1(1)|D=1] - E[Y_0(0)|D=1] - E[Y_0(1) - Y_0(0)|D=0] \\ &= E[Y_1(1) - Y_0(0)|D=1] - E[Y_0(1) - Y_0(0)|D=0]\end{aligned}$$

- And we observe all the averages we need!

$$\hat{\alpha}_{ATT} = \{\overline{Y}_1(D=1) - \overline{Y}_0(D=1)\} - \{\overline{Y}_1(D=0) - \overline{Y}_0(D=0)\}$$

- under general conditions, this estimator is consistent and asymptotically normal

OLS and the diff-in-diff estimator (1/2)

- We can get this result in a regression framework:

$$Y_i = \beta_0 + \beta_1 D_i^{treated} + \beta_2 D_i^{after} + \beta_3 D_i^{treated} D_i^{after} + u_i$$

  where $D^{treated} = 1$ if observation in treatment group and $D^{after} = 1$ if period after treatment.

- Therefore:
  - $E[Y_1(1)|D = 1] = E[Y|treated, after] = \beta_0 + \beta_1 + \beta_2 + \beta_3$
  - $E[Y_0(0)|D = 1] = E[Y|treated, before] = \beta_0 + \beta_1$
  - $E[Y_0(1)|D = 0] = E[Y|nontreated, after] = \beta_0 + \beta_2$
  - $E[Y_0(0)|D = 0] = E[Y|nontreated, before] = \beta_0$

OLS and the diff-in-diff estimator (2/2)

- Going back to our table:

Expected Outcome by Group and Period

	Treated	Control
Before	$\beta_0 + \beta_1$	β_0
After	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_0 + \beta_2$
Difference	$\beta_2 + \beta_3$	β_2

- First “diff” (for treated): $E[Y_1(1)|D=1] - E[Y_0(0)|D=1] = \beta_2 + \beta_3$
- Second “diff” (for controls): $E[Y_0(1)|D=0] - E[Y_0(0)|D=0] = \beta_2$
- The “diff-in-diff”: $(\beta_2 + \beta_3) - (\beta_2) = \beta_3$

Conditional diff-in-diff

$$Y_{it} = \beta_0 + \gamma X_i + \beta_1 D_i^{treated} + \beta_2 D_i^{after} + \beta_3 D_i^{treated} D_i^{after} + u_i$$

- The model can be extended to include additional regressors: X_i represents a set of variables that may affect Y_i, for instance individual characteristics before treatment.
- The OLS estimator for β_3 is the conditional diff-in-diff estimator.
- The parallel paths assumption is assumed to be satisfied after controlling for this vector of covariates.
- Implementable both in panel data and in pooled cross-sections.

A Simple Example: The Effect of Worker Compensation Laws on Injury Duration

- Meyer, Viscusi, and Durbin,[1] studied the length of time that an injured worker receives workers' compensation.
- On July 15, 1980, Kentucky raised the cap on weekly earnings that were covered by workers' compensation.
- An increase in the cap has no effect on the benefit for low-income workers, but it makes it less costly for a high-income worker to stay on workers' compensation.
- Therefore, the treatment group is high-income workers, and the control group low-income workers.

[1] Workers' Compensation and Injury Duration: Evidence from a Natural Experiment: AER (1995)

Diff-in-diffs

Using data before and after the cap is raised: 1979 and 1981

$$log(durat)_i = \beta_0 + \beta_1 highearn_i + \beta_2 D_i^{1981} + \beta_3 highearn_i D_i^{1981} + u_i$$

where $log(durat)$ is the time spent on workers' compensation (in logs), *highearn* a dummy for high income workers and D^{1981} a dummy for 1981.

- β_1 captures the duration differential for high and low income workers before the cap is raised
- β_2 captures average duration changes for low income workers
- β_3 is the diff-in-diffs estimator: consistently estimates the policy effect if the high income workers and low-income workers had no other different changes, only the cap raising.