# Lecture 10
# Difference-in-Differences

(Chapter 13.1-13.2 in the textbook)

Econ 322 Econometrics

Spring 2025

1. Pooled Cross Sections

2. Difference-in-Differences

# Table of Contents

1 Pooled Cross Sections

2 Difference-in-Differences

# What is a Pooled Cross Section?

- Data obtained by pooling cross sections are very useful for establishing trends and conducting policy analysis.
- Pooled cross sections are available whenever surveys are repeated over time, and new samples are obtained each time.
- Examples include the General Social Survey and the Current Population Survey.

- Statistically, analyzing pooled cross sections is similar to a single cross section, provided we assume random sampling was used to collect each cross section.
- With a PCS, often a goal is to see how the mean value of a variable has changed over time in ways that cannot be explained by observables. (For example, how has fertility changed in ways that cannot be explained by education or labor force participation?)

- From a policy perspective, PCSs are at the foundation of *difference-in-differences* estimation.
- The typical DID setup is that data can be collected both before and after an intervention (or "treatment" ), and there is (at least) one "control group" and (at least) one "treatment" group.
- Often the intervention is of a yes/no form. But other nonbinary treatments (such as class size) can be handled, too.

- In the simplest case, outcomes are observed for two groups over two time periods. One of the groups is exposed to a "treatment" (or intervention) in the second period but not in the first period. The second group is not exposed to the treatment during either period.
- The setup is used often to study the effects of policy interventions.

**EXAMPLE:** Suppose we can observe test scores for a sample of 4th graders in two large school districts at the end of the 2008 and 2009 school years.

- One district was given a grant to make fourth-grade class sizes smaller for the 2009 year.
- In 2008, we would have a random sample of 4th graders from each district. We would sample the same districts in 2009, and get a different set of 4th graders. (Any overlap would be small and due to repeating a grade.)
- We effectively have random samples from four different populations (two districts and two time periods each).

# Table of Contents

1 Pooled Cross Sections

2 Difference-in-Differences

# The Differences-in-Differences Estimator

- Let $A$ be the control group and $B$ the treatment group. Let $d2$ be a time period dummy equal to one for a unit in the second time period. Write

$$y = \beta_0 + \beta_1 dB + \delta_0 d2 + \delta_1 d2 \cdot dB + u, \tag{1}$$

  where $y$ is the outcome of interest.
- Treatment $X = d2 \cdot dB$
- The mean value of $u$ is zero (essentially by definition), so we can read off the means of the response for different combinations.

| | Before (1) | After (2) | After − Before |
|---|---|---|---|
| Control (A) | $\beta_0$ | $\beta_0 + \delta_0$ | $\delta_0$ |
| Treatment (B) | $\beta_0 + \beta_1$ | $\beta_0 + \delta_0 + \beta_1 + \delta_1$ | $\delta_0 + \delta_1$ |
| Treatment − Control | $\beta_1$ | $\beta_1 + \delta_1$ | $\delta_1$ |

$$y = \beta_0 + \beta_1 dB + \delta_0 d2 + \delta_1 d2 \cdot dB + u$$

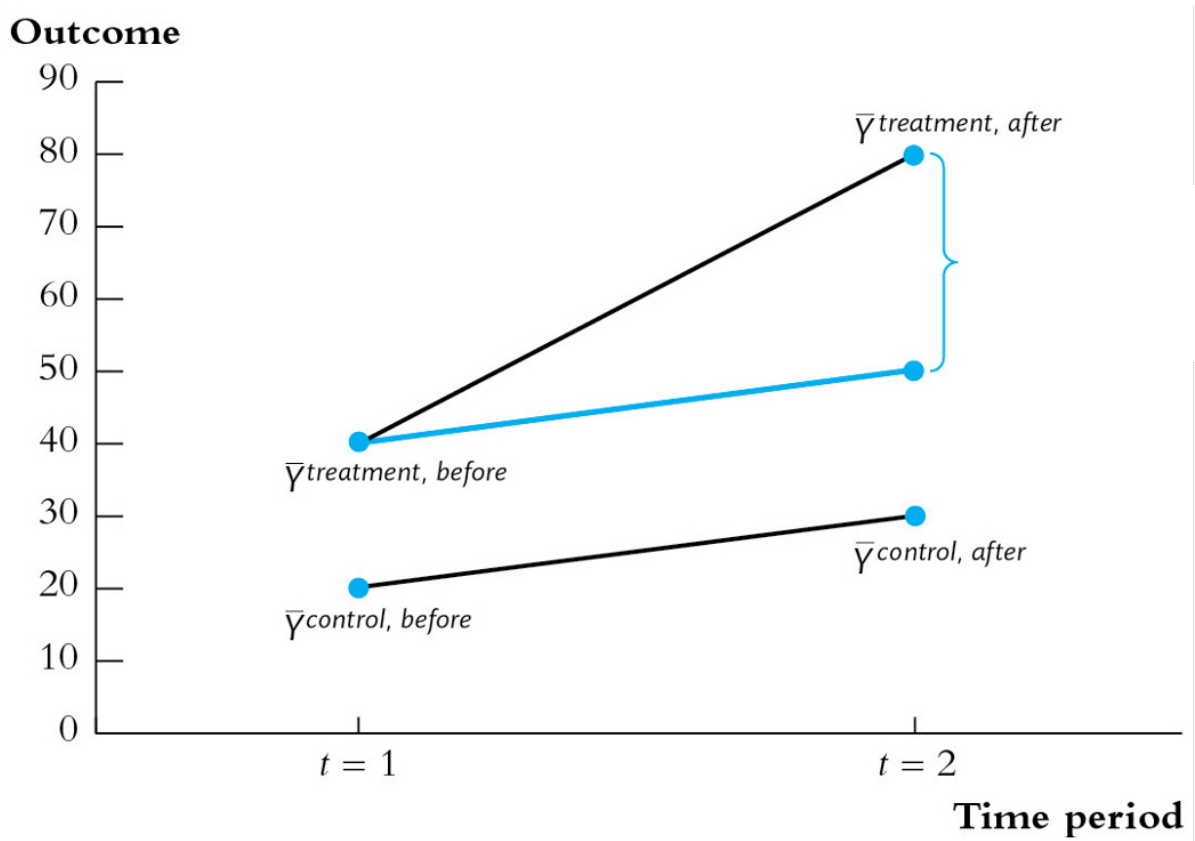
- *dB* captures possible differences between the treatment and control groups prior to the policy change. Its coefficient, $\beta_1$, is the difference between treatment and control *before* the intervention.
- *d*2 captures aggregate factors that would cause changes in *y* over time even in the absense of an intervention. Notice its coefficient, $\delta_0$, is the change in the mean of the control group across the two periods.

- The change in the mean over time for the treatment group is $\delta_0 + \delta_1$.
- The coefficient of interest is $\delta_1 = (\delta_1 + \delta_0) - \delta_0$, the difference in the average changes over time for the treatment and control groups.
- Conveniently, $\delta_1$ is the coefficient on the interaction $d2 \cdot dB$, which is one if and only if the unit is in the treatment group in period 2.
- If $y$ is a logarithm then, as usual, $\delta_1$ is a proportionate effect (multiply by 100 to approximate the percentage effect).

- The difference-in-differences (DD) estimate can be obtained by applying OLS to equation (1). Or, we can just use the averages directly:

$$\begin{aligned} \hat{\delta}_{DD} &= \hat{\delta}_1 = (\bar{y}_{B,2} - \bar{y}_{B,1}) - (\bar{y}_{A,2} - \bar{y}_{A,1}) \\ &= (\bar{y}_{B,2} - \bar{y}_{A,2}) - (\bar{y}_{B,1} - \bar{y}_{A,1}) \end{aligned}$$

- OLS makes inference straightforward. Heteroskedasticity-robust inference allows different group/time period variances in regression framework.


Outcome
90
80
70
60
50
40
30
20
10
0
$\bar{Y}$ treatment, after
$\bar{Y}$ treatment, before
$\bar{Y}$ control, after
$\bar{Y}$ control, before
$t = 1$
$t = 2$
Time period

- Just using $\bar{y}_{B,2} - \bar{y}_{B,1}$ – the difference over time in the means of the treatment group – attributes all change to the intervention.
- Just using $\bar{y}_{B,2} - \bar{y}_{A,2}$ – that is, the difference in treatment and control means post-treatment – attributes any differences in the groups to the treatment.
- Writing $\hat{\delta}_1 = (\bar{y}_{B,2} - \bar{y}_{B,1}) - (\bar{y}_{A,2} - \bar{y}_{A,1})$ shows that we are comparing the change in means over time for the treatment to the change in means for the control.

- While powerful, the basic approach can suffer from several problems. First, there may be compositional effects. For example, when looking at math scores and an intervention to reduce class size, maybe (say) the 4th graders in the two years are not comparable. Perhaps the smaller class sizes attracts new students to that district. This is the problem of *compositional changes*.
- Can control for changes in composition to some extent by including regressors as controls. For example, family background variables. The regression framework makes this easy:

$$y = \beta_0 + \beta_1 dB + \delta_0 d2 + \delta_1 d2 \cdot dB + W\gamma + u$$

# Parallel Trends

- A potential problem with using only two periods is that the control and treatment groups may be trending at different rates having nothing to do with the intervention.
- In the class-size example, what if performance in district $B$ was at an initial lower level (on average) than district $A$ but was increasing faster? $dB$ accounts for initial level differences, but the effect of the policy is measured by the growth in the averages. Only way to solve this problem is to get another control group or more years of data (preferably at least two before the intervention).
- The key assumption is that, absent the treatment, the differences between treated and control observations would be constant over time – what's known as the parallel trends assumption

**EXAMPLE:** Effects of Attendance Monitoring on Attendance

- Two years of data (2005 and 2006) on two instructors (lectures 1 and 2). In lecture 2 in 2006, electronic attendance taking was used (card swiping).
- Year dummy is $d06$, lecture dummy is $dlec2$, and the interaction is $d06_dlec2$.

```
. use attendance

. des

Contains data from attendance.dta
  obs:         1,338
 vars:             8                          10 Oct 2010 23:11
 size:        26,760 (99.9% of memory free)    (_dta has notes)
-------------------------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------------------------
attend          byte   %8.0g                  classes attended out of 32
priGPA          float  %9.0g                  cumulative GPA prior to term
ACT             byte   %8.0g                  ACT score
year            int    %9.0g                  2005 or 2006
lecture         byte   %9.0g                  1 or 2
d06             byte   %9.0g                  =1 if year == 2006
dlec2           byte   %9.0g                  =1 if lecture == 2
d06_dlec2       byte   %9.0g                  d06*dlec2
-------------------------------------------------------------------------------------------------
Sorted by:
```

```
. tab year

    2005 or |
       2006 |      Freq.     Percent        Cum.
------------+-----------------------------------
       2005 |        680       50.82       50.82
       2006 |        658       49.18      100.00
------------+-----------------------------------
      Total |      1,338      100.00

. tab lecture

     1 or 2 |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        683       51.05       51.05
          2 |        655       48.95      100.00
------------+-----------------------------------
      Total |      1,338      100.00
```

```
. tab attend
```

| classes attended out of 32 | Freq. | Percent | Cum. |
|---|---|---|---|
| 0 | 4 | 0.30 | 0.30 |
| 1 | 1 | 0.07 | 0.37 |
| 2 | 1 | 0.07 | 0.45 |
| 3 | 2 | 0.15 | 0.60 |
| 4 | 2 | 0.15 | 0.75 |
| 5 | 4 | 0.30 | 1.05 |
| 6 | 5 | 0.37 | 1.42 |
| 7 | 4 | 0.30 | 1.72 |
| 8 | 8 | 0.60 | 2.32 |
| 9 | 5 | 0.37 | 2.69 |
| 10 | 9 | 0.67 | 3.36 |
| 11 | 5 | 0.37 | 3.74 |
| 12 | 9 | 0.67 | 4.41 |
| 13 | 17 | 1.27 | 5.68 |
| 14 | 13 | 0.97 | 6.65 |
| 15 | 23 | 1.72 | 8.37 |
| 16 | 32 | 2.39 | 10.76 |
| 17 | 22 | 1.64 | 12.41 |
| 18 | 32 | 2.39 | 14.80 |
| 19 | 26 | 1.94 | 16.74 |
| 20 | 36 | 2.69 | 19.43 |

| | | | |
|---|---|---|---|
| 21 | 62 | 4.63 | 24.07 |
| 22 | 68 | 5.08 | 29.15 |
| 23 | 81 | 6.05 | 35.20 |
| 24 | 96 | 7.17 | 42.38 |
| 25 | 132 | 9.87 | 52.24 |
| 26 | 126 | 9.42 | 61.66 |
| 27 | 137 | 10.24 | 71.90 |
| 28 | 138 | 10.31 | 82.21 |
| 29 | 109 | 8.15 | 90.36 |
| 30 | 74 | 5.53 | 95.89 |
| 31 | 30 | 2.24 | 98.13 |
| 32 | 25 | 1.87 | 100.00 |
| Total | 1,338 | 100.00 | |

```
. reg attend d06

      Source |       SS       df       MS              Number of obs =    1338
-------------+------------------------------           F(  1,  1336) =   30.76
       Model |    910.7359     1    910.7359            Prob > F      =  0.0000
    Residual |  39557.0728  1336  29.6085874            R-squared     =  0.0225
-------------+------------------------------           Adj R-squared =  0.0218
       Total |  40467.8087  1337  30.2676205            Root MSE      =  5.4414

------------------------------------------------------------------------------
      attend |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         d06 |    1.650277   .2975565     5.55   0.000     1.066548    2.234006
       _cons |    23.17647   .2086673   111.07   0.000     22.76712    23.58582
------------------------------------------------------------------------------

. * Shows that attendance increased by 1.65 classes, on average, from 2005
. * to 2006 (across both lectures).
```

```
. reg attend dlec2

      Source |       SS       df       MS              Number of obs =    1338
-------------+------------------------------           F(  1,  1336) =    1.57
       Model |  47.3565824     1  47.3565824           Prob > F      =  0.2111
    Residual |  40420.4521  1336  30.2548294           R-squared     =  0.0012
-------------+------------------------------           Adj R-squared =  0.0004
       Total |  40467.8087  1337  30.2676205           Root MSE      =  5.5004

------------------------------------------------------------------------------
      attend |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dlec2 |   .3763459   .3008115     1.25   0.211    -.2137683    .9664602
       _cons |   23.80381   .2104684   113.10   0.000     23.39092    24.21669
------------------------------------------------------------------------------

. * Average attendance is slightly higher in lecture 2, but not statistically
. * significant.
```

```
. reg attend d06 dlec2 d06_dlec2

      Source |       SS       df       MS              Number of obs =    1338
-------------+------------------------------           F(  3,  1334) =   12.60
       Model |  1114.82893     3  371.609644            Prob > F      =  0.0000
    Residual |  39352.9797  1334  29.4999848            R-squared     =  0.0275
-------------+------------------------------           Adj R-squared =  0.0254
       Total |  40467.8087  1337  30.2676205            Root MSE      =  5.4314

------------------------------------------------------------------------------
      attend |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         d06 |    1.009064   .4156566     2.43   0.015     .1936517    1.824476
       dlec2 |   -.2411765   .4165685    -0.58   0.563    -1.058377    .5760243
   d06_dlec2 |    1.328705   .5942944     2.24   0.026     .1628514    2.494558
       _cons |    23.29706   .2945584    79.09   0.000     22.71921    23.87491
------------------------------------------------------------------------------

. * The difference-in-differences estimate is 1.33, and it is statistically
. * significant. Note that attendance increased by 1 for the control group.
. * A spillover effect, or something else?
```

```
. reg attend d06 dlec2 d06_dlec2, robust

Linear regression                               Number of obs =    1338
                                                F(  3,  1334) =   11.11
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0275
                                                Root MSE      =  5.4314

------------------------------------------------------------------------------
             |               Robust
      attend |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         d06 |   1.009064   .3603792     2.80   0.005      .302092    1.716035
       dlec2 |  -.2411765   .4165156    -0.58   0.563    -1.058273    .5759204
   d06_dlec2 |   1.328705   .5976782     2.22   0.026     .1562132    2.501196
       _cons |   23.29706   .2560285    90.99   0.000      22.7948    23.79932
------------------------------------------------------------------------------
```

```
. * Now add prior GPA and ACT score:

. reg attend d06 dlec2 d06_dlec2 priGPA ACT, robust

Linear regression                               Number of obs =    1338
                                                F(  5,  1332) =   91.21
                                                Prob > F      =  0.0000
                                                R-squared     =  0.3034
                                                Root MSE      =  4.6004

------------------------------------------------------------------------------
             |               Robust
      attend |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         d06 |   .9429047   .3209404     2.94   0.003     .3133011    1.572508
       dlec2 |   .1362344   .3508858     0.39   0.698    -.5521145    .8245834
   d06_dlec2 |   1.471297   .5055328     2.91   0.004     .4795696    2.463024
      priGPA |   5.443933   .2791131    19.50   0.000     4.896383    5.991482
         ACT |  -.5445861   .0404013   -13.48   0.000    -.6238432   -.4653289
       _cons |   21.28492   .8457649    25.17   0.000     19.62574    22.94409
------------------------------------------------------------------------------

. * Does improve precision of estimate.
```