Topic 7: Difference-in-Differences and Factor Models

ECON 5783 — University of Arkansas

Prof. Kyle Butts

December 2024

Difference-in-Differences

- Initial Difference-in-difference usage
- Classic Example: Card and Krueger (2000, AER)
- Econometric formulation to DID
- Event-study

Conditional Parallel Trends

Staggered Treatment Timing

- Estimating Group-Time ATTs (Callaway and Sant'Anna)
- Imputation based estimators

What is difference-in-differences (DiD)

Difference-in-differences compares a group assigned to treatment versus a group not assigned to treatment

- The estimator compares the treated groups change in outcomes before and after the treatment to the control groups change in outcomes before and after the treatment

One of the most widely used quasi-experimental methods in economics and increasingly in industry

Difference-in-Differences

- Initial Difference-in-difference usage
- Classic Example: Card and Krueger (2000, AER)
- Econometric formulation to DID
- Event-study

Conditional Parallel Trends

Staggered Treatment Timing

- Estimating Group-Time ATTs (Callaway and Sant'Anna)
- Imputation based estimators

Ignaz Semmelweis and washing hands

Early 1820s, Vienna passed legislation requiring that if a pregnant women giving birth went to a public hospital (free care)

- depending on the day of week and time of day, she would be routed to either the midwife wing or the physician wing

Pregnant women died after delivery in the (male) wing at a rate of 13-18%, but only 3% in the (female) midwife wing

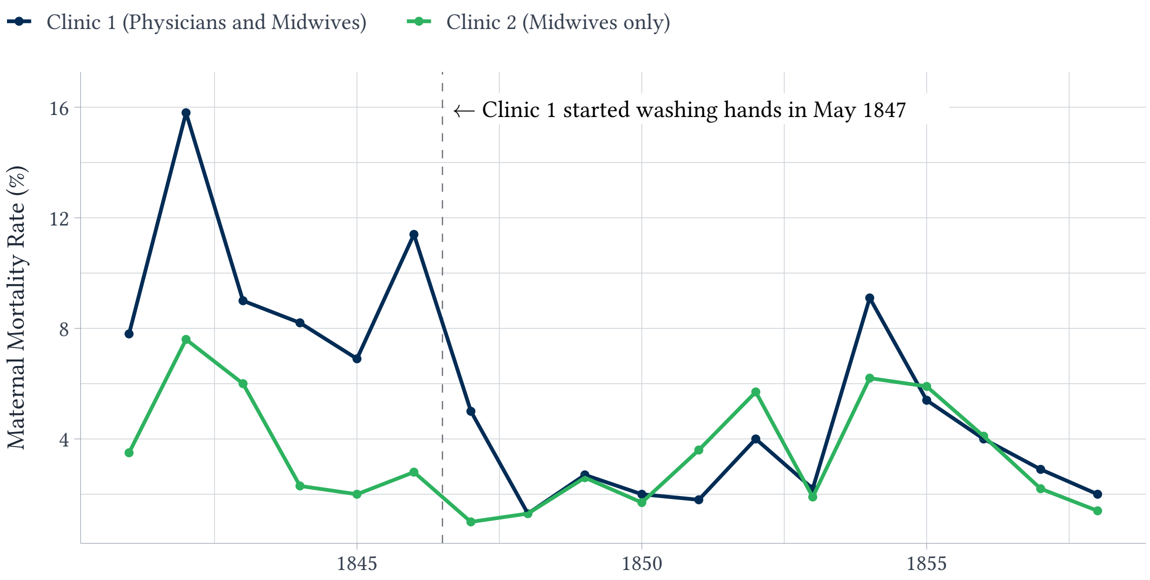
Ignaz Semmelweis and washing hands

Ignaz Semmelweis, after a lot of observation, conjectures that the cause is:

- the teaching faculty would teach anatomy using cadavers and then delivering babies without washing hands

Convinced the hospital to have physicians wash their hands in chlorine but not the midwives

- Compared mortality rates in treated Clinic 1 (Physicians and Midwives) vs. untreated Clinic 2 (Midwives only)



Clinic 1 (Physicians and Midwives)
Clinic 2 (Midwives only)
← Clinic 1 started washing hands in May 1847
Maternal Mortality Rate (%)
16
12
8
4
1845
1850
1855

Identifying assumptions

While this, today, seems like an obvious treatment effect, people at the time did not believe this result

- In fact, Semmelweis was fired about a year and a half later and his life was ruined by critics

It is worth asking for this topic, "What do we need to assume to believe this result?"

Identifying assumptions

Looking at the previous figure, we see that prior to treatment, mortality rates were way higher in the physicians clinic than midwives. Then, right when treatment starts we see a large drop in the mortality rate

The main issue is that we can not be sure what would happen had the physician clinic not been required to wash their hands

- Do not observe the post-treatment $y(0)$

Identifying assumptions

Looking at the previous figure, we see that prior to treatment, mortality rates were way higher in the physicians clinic than midwives. Then, right when treatment starts we see a large drop in the mortality rate

The main issue is that we can not be sure what would happen had the physician clinic not been required to wash their hands

- Do not observe the post-treatment $y(0)$

We, however, do not see a similar drop in the second clinic, so this rules out many shocks that would impact both clinics

Identifying assumptions

What we will come to formalize is the parallel counterfactual trends assumption:

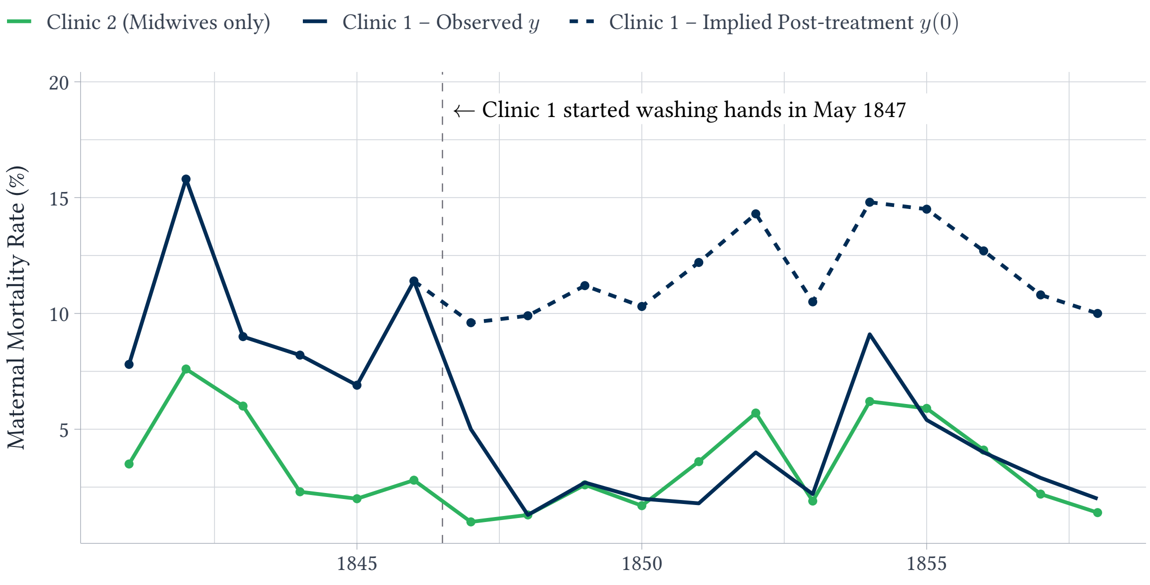
- In the absence of treatment, the treated units would be on the same counterfactual trend as we observe in the untreated units

You can imagine taking the trend from Clinic 2 and appending that onto the start of the post-period for Clinic 1

- The implied $Y(0)$ if indeed the two clinics would have the same counterfactual trends

People typically call it the *parallel trends* assumption

- But I prefer the full phrase because it emphasizes this is about trends for the treated units *had they not been treated*


Clinic 2 (Midwives only)
Clinic 1 – Observed y
Clinic 1 – Implied Post-treatment $y(0)$
← Clinic 1 started washing hands in May 1847
Maternal Mortality Rate (%)
20
15
10
5
1845
1850
1855

Difference-in-Differences

- Initial Difference-in-difference usage
- Classic Example: Card and Krueger (2000, AER)
- Econometric formulation to DID
- Event-study

Conditional Parallel Trends

Staggered Treatment Timing

- Estimating Group-Time ATTs (Callaway and Sant'Anna)
- Imputation based estimators

Card and Krueger (1994, AER)

The first “modern” economics paper to use difference-in-differences

Card and Krueger studied the 1992 minimum wage increase in New Jersey from $4.25 to $5.05

- The story goes that they heard about the minimum wage change and *ran to the field* to start collecting data on fast-food employment prior to the minimum wage

  $\rightarrow$

Card and Krueger (1994, AER)

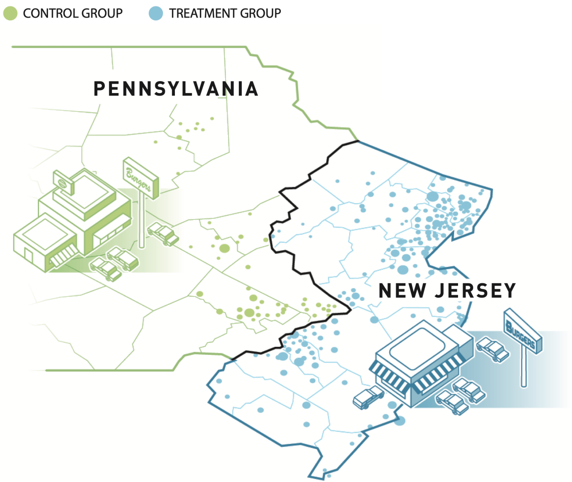
The first “modern” economics paper to use difference-in-differences

Card and Krueger studied the 1992 minimum wage increase in New Jersey from \$4.25 to \$5.05

- The story goes that they heard about the minimum wage change and *ran to the field* to start collecting data on fast-food employment prior to the minimum wage
    - $\rightarrow$

Their strategy was to compare changes to New Jersey fast-food employment to those in Eastern Pennsylvania

- 331 in New Jersey (treated)
- 79 in Eastern Pennsylvania (untreated)

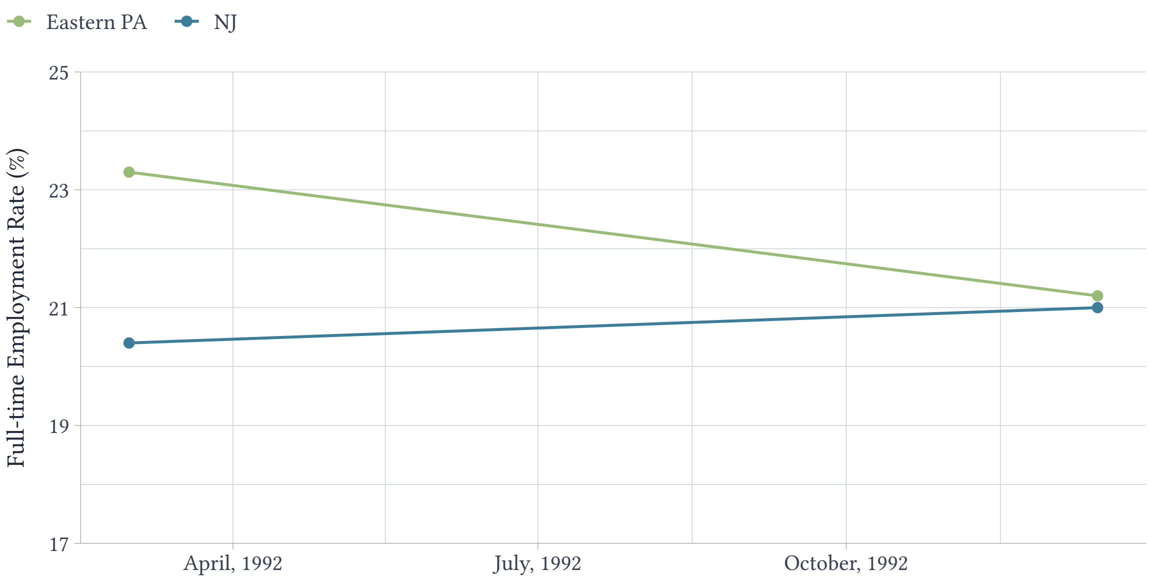


Source: Nobel Prize summary

Measurements

They measured employment before (in March 1992) and after (in December 1992) the minimum wage passed

- This is a relatively small survey, but it was novel because no one really tried to see what the actual impacts of minimum wage changes was


Eastern PA
NJ
Full-time Employment Rate (%)
25
23
21
19
17
April, 1992
July, 1992
October, 1992

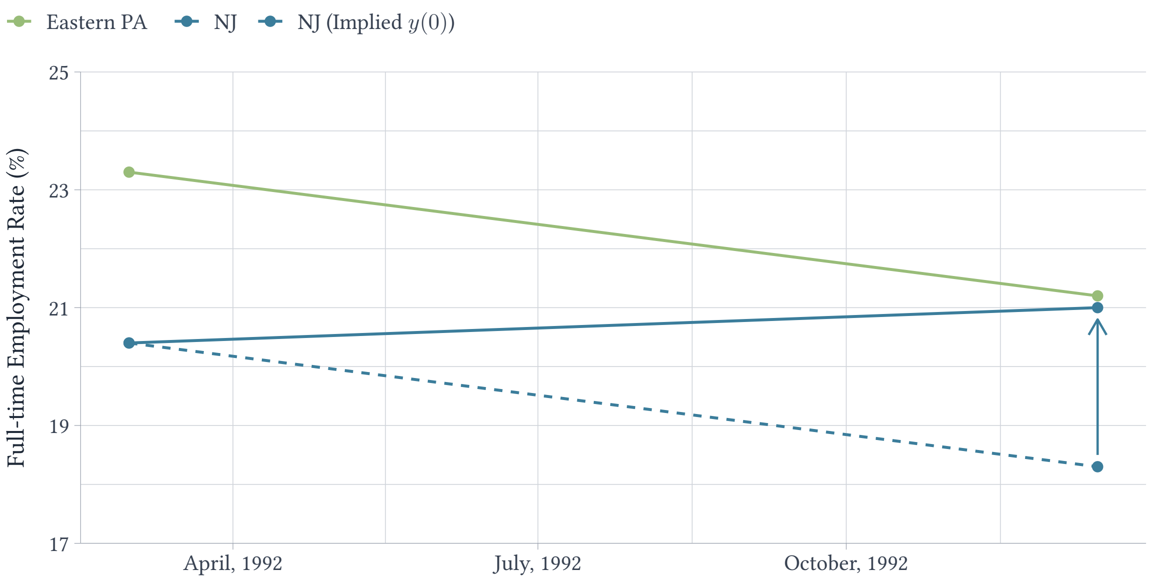
Identification

So, we see that NJ employment went up slightly and Eastern PA employment went down a bit more

- We infer that NJ would have went down by the same amount as Eastern PA had the minimum wage not passed

That is, we assume that there are "common shocks" to both areas and assume that there are no additional shocks that impact *only one* of the two regions

- They are on "parallel counterfactual trends"

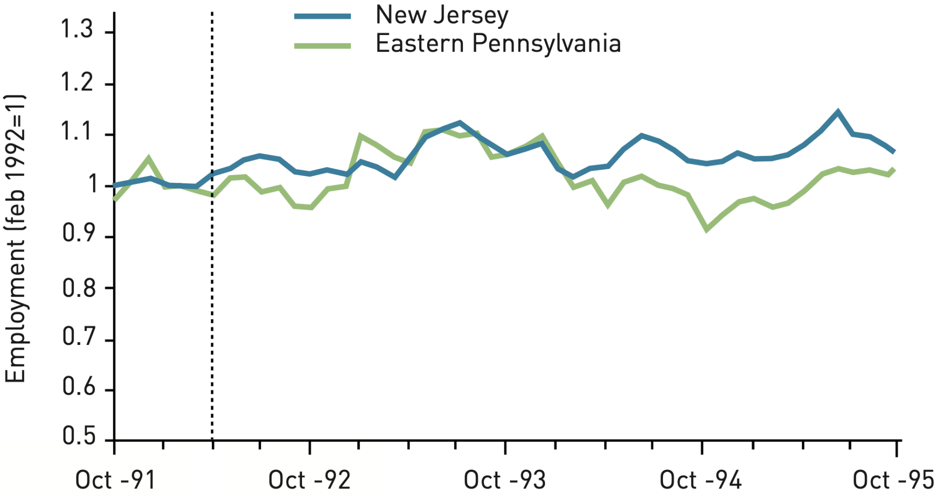

Eastern PA
NJ
NJ (Implied $y(0)$)
Full-time Employment Rate (%)
25
23
21
19
17
April, 1992
July, 1992
October, 1992

Is this believable?

From this graph, it’s not clear what to think about this assumption of parallel counterfactual trends

For this reason, it is (now) typical to compare the treated and control units *prior* to treatment uptake to see if they are on similar trends

- Using many observations before and after treatment are called ‘event-study’ estimates



Source: Nobel Prize summary

Pre-trends

The previous figure shows that for a few months prior to the minimum wage change, the employment trends of Eastern PA and New Jersey followed closely to one another

- This supports the idea that in the absence of treatment the NJ and Eastern PA trends would be similar in the post-period

Pre-trends

The previous figure shows that for a few months prior to the minimum wage change, the employment trends of Eastern PA and New Jersey followed closely to one another

- This supports the idea that in the absence of treatment the NJ and Eastern PA trends would be similar in the post-period

To be clear, parallel counterfactual trends involves the *post-treatment* y_{it}

- Having similar trends prior to treatment helps support this assumption, but does not *prove it*

Ashenfelter’s dip

Orley Ashenfelter’s 1978 paper entitled “Estimating the Effect of Training Programs on Earnings” is a great example to illustrate the different between common trends before treatment and the *parallel counterfactual trends* assumption

Ashenfelter’s dip

Orley Ashenfelter’s 1978 paper entitled “Estimating the Effect of Training Programs on Earnings” is a great example to illustrate the different between common trends before treatment and the *parallel counterfactual trends* assumption

He looks at individuals that sign-up for a work training program on their future earnings

- For many years prior to treatment, the workers that do and do not enter the training have common earnings trends
- Just prior to treatment, the workers that do enter the program face a sudden *dip* in earnings
- Then, after the program, the workers’ earnings go back up towards the original level

Ashenfelter's dip

What was happening was that workers just prior to treatment lost their job (hence trying to learn new labor force skills)

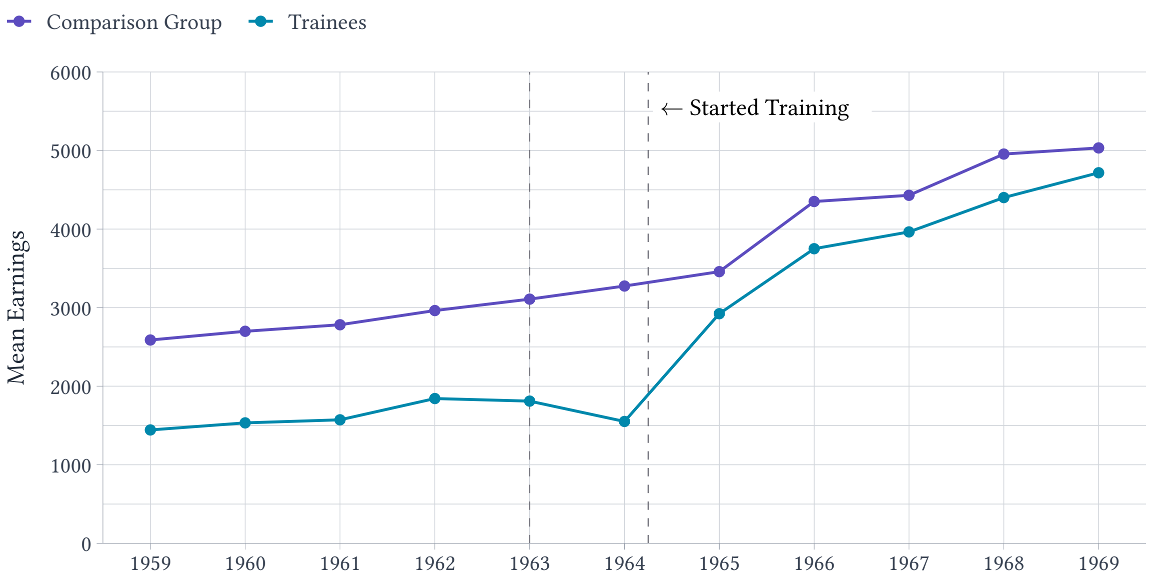
Ashenfelter’s dip

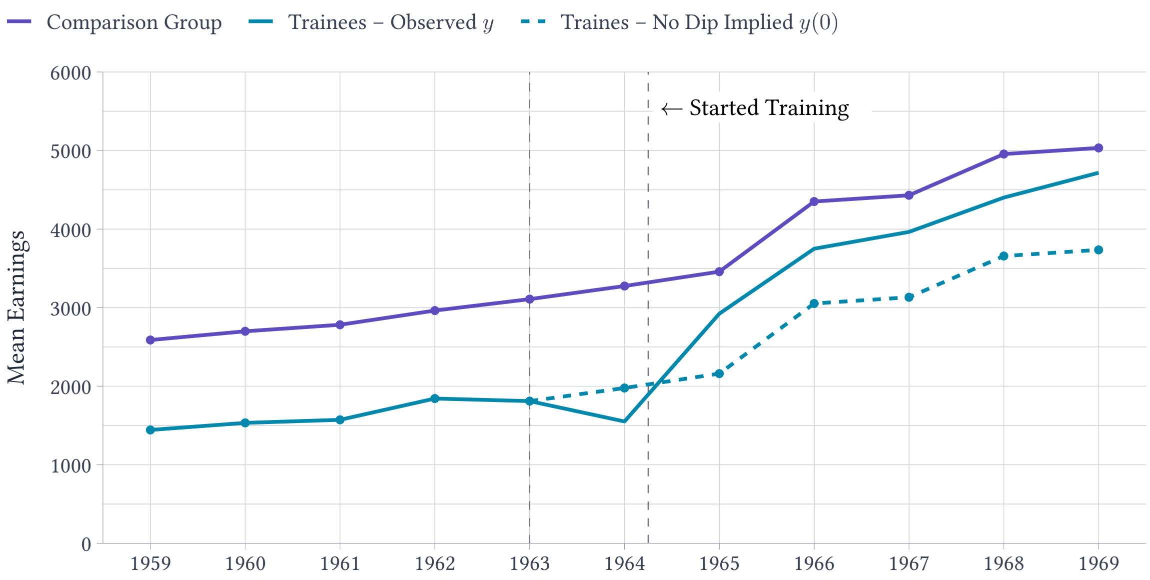
What was happening was that workers just prior to treatment lost their job (hence trying to learn new labor force skills)

In the absence of the training, we would expect those workers to have a raise in earnings anyways because they would likely be hired somewhere

- The treated workers and the untreated workers have different earning dynamics

So even though they have similar trends prior to treatment, the parallel counterfactual trends assumption does not hold in this setting


Comparison Group
Trainees
← Started Training
Mean Earnings
6000
5000
4000
3000
2000
1000
0
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969


Comparison Group
Trainees – Observed y
Traines – No Dip Implied $y(0)$
Mean Earnings
← Started Training
0
1000
2000
3000
4000
5000
6000
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969

DID Key Ideas

Difference-in-differences compares a group assigned to treatment versus a group not assigned to treatment

- The estimator compares the treated groups change in outcomes before and after the treatment to the control groups change in outcomes before and after the treatment

The key assumption we make is the parallel counterfactual trends assumption

- The change in outcomes over time for control units are an appropriate stand-in for the treated unit's change in outcomes *if they did not receive treatment*

Difference-in-Differences

- Initial Difference-in-difference usage
- Classic Example: Card and Krueger (2000, AER)
- Econometric formulation to DID
- Event-study

Conditional Parallel Trends

Staggered Treatment Timing

- Estimating Group-Time ATTs (Callaway and Sant'Anna)
- Imputation based estimators

2×2 Difference-in-Differences

The Card and Krueger minimum wage paper is an example of the canonical 2×2 DID, so we will begin there

We observe units $i \in \{1, \ldots, N\}$ for two periods (before and after), $t = 0$ and $t = 1$

- Let D_i be an indicator for which units receive treatment
- Let $\text{Post}_t = \mathbb{1}[t = 1]$ be an indicator for being in the post-period

2×2 Difference-in-Differences

The Card and Krueger minimum wage paper is an example of the canonical 2×2 DID, so we will begin there

We observe units $i \in \{1, \ldots, N\}$ for two periods (before and after), $t = 0$ and $t = 1$

- Let D_i be an indicator for which units receive treatment
- Let Post$_t = \mathbb{1}[t = 1]$ be an indicator for being in the post-period

Then, we have potential outcomes for each unit in the post-period:

- $y_{i0}(D_i)$ and $y_{i1}(D_i)$
  - $\rightarrow$ We typically assume that treatment does not impact y_{i0}, i.e. $y_{i0} = y_{i0}(1) = y_{i0}(0)$. This is called the "no anticipation" assumption

Treatment effect of interest

The treatment effect of interest is the average effect of treatment in period 1 for the treated units:

$$\text{ATT}_1 = \mathbb{E}[y_{i1}(1) - y_{i1}(0) \mid D_i = 1]$$

The counterfactual compares the period 1 outcome under treatment to the period 1 outcome in the absence of treatment

- This is ***not*** the post-y minus pre-y!

Parallel Counterfactual Trends assumption

Our Parallel Counterfactual Trends imposes restrictions on the change in untreated potential outcomes:

$$\mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = 1] = \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = 0]$$

This says, in the absence of treatment, the change in y is on average the same for the treated and the control group

Observed difference in y

For the treated unit, we can do an econometrician's favorite math trick (add and subtract something) to analyze the observed change in y for the treated units:

$$\mathbb{E}[y_{i1} - y_{i0} \mid D_i = 1] = \mathbb{E}[y_{i1}(1) - y_{i0}(0) \mid D_i = 1]$$

Observed difference in y

For the treated unit, we can do an econometrician's favorite math trick (add and subtract something) to analyze the observed change in y for the treated units:

$$\begin{aligned}\mathbb{E}[y_{i1} - y_{i0} \mid D_i = 1] &= \mathbb{E}[y_{i1}(1) - y_{i0}(0) \mid D_i = 1] \\ &= \mathbb{E}\Big[y_{i1}(1) \boxed{-y_{i1}(0) + y_{i1}(0)} - y_{i0}(0) \mid D_i = 1\Big]\end{aligned}$$

Observed difference in y

For the treated unit, we can do an econometrician's favorite math trick (add and subtract something) to analyze the observed change in y for the treated units:

$$
\begin{aligned}
\mathbb{E}[y_{i1} - y_{i0} \mid D_i = 1] &= \mathbb{E}[y_{i1}(1) - y_{i0}(0) \mid D_i = 1] \\
&= \mathbb{E}\Big[y_{i1}(1) \boxed{- y_{i1}(0) + y_{i1}(0)} - y_{i0}(0) \mid D_i = 1\Big] \\
&= \mathbb{E}[y_{i1}(1) - y_{i1}(0) \mid D_i = 1] + \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = 1]
\end{aligned}
$$

Observed difference in y

For the treated unit, we can do an econometrician's favorite math trick (add and subtract something) to analyze the observed change in y for the treated units:

$$
\begin{aligned}
\mathbb{E}[y_{i1} - y_{i0} \mid D_i = 1] &= \mathbb{E}[y_{i1}(1) - y_{i0}(0) \mid D_i = 1] \\
&= \mathbb{E}\Big[y_{i1}(1) \boxed{-y_{i1}(0) + y_{i1}(0)} - y_{i0}(0) \mid D_i = 1\Big] \\
&= \mathbb{E}[y_{i1}(1) - y_{i1}(0) \mid D_i = 1] + \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = 1] \\
&= \text{ATT}_1 + \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = 1]
\end{aligned}
$$

$\implies$ The change in outcome for the treated units is the effect of treatment plus the treated groups' counterfactual trend

Observed difference in y

$$\mathbb{E}[y_{i1} - y_{i0} \mid D_i = 1] = \textsf{ATT}_1 + \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = 1]$$

For control units, the math is simpler

$$\mathbb{E}[y_{i1} - y_{i0} \mid D_i = 0] = \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = 0]$$

$\implies$ The change in outcome for the control units is the effect of treatment plus the control groups' counterfactual trend

Difference-in-differences

Now, the difference-in-differences estimand is formed by subtracting the two change in outcomes:

$$\tau_{\texttt{DID}} = \mathbb{E}[y_{i1} - y_{i0} \mid D_i = 1] - \mathbb{E}[y_{i1} - y_{i0} \mid D_i = 0]$$

Difference-in-differences

Now, the difference-in-differences estimand is formed by subtracting the two change in outcomes:

$$
\begin{aligned}
\tau_{\texttt{DID}} &= \mathbb{E}[y_{i1} - y_{i0} \mid D_i = 1] - \mathbb{E}[y_{i1} - y_{i0} \mid D_i = 0] \\
&= \textsf{ATT}_1 + \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = 1] - \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = 0]
\end{aligned}
$$

The difference-in-differences estimand compares treated unit's change in y to control unit's change in y

- This estimates the effect of treatment plus the difference in trends between the two groups

Difference-in-differences

$$\begin{aligned}\tau_{\texttt{DID}} &= \mathbb{E}[y_{i1} - y_{i0} \mid D_i = 1] - \mathbb{E}[y_{i1} - y_{i0} \mid D_i = 0] \\ &= \text{ATT}_1 + \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = 1] - \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = 0]\end{aligned}$$

For example, if the treated group had a larger counterfactual growth in y (like in Ashenfelter's dip example), then the treatment effect will be biased upwards

Difference-in-differences

However, assuming parallel counterfactual trends implies that these two counterfactual trend terms are the same and therefore cancel out

$$\begin{aligned}
\tau_{\texttt{DID}} &= \mathbb{E}[y_{i1} - y_{i0} \mid D_i = 1] - \mathbb{E}[y_{i1} - y_{i0} \mid D_i = 0] \\
&= \textsf{ATT}_1 + \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = 1] - \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = 0]
\end{aligned}$$

Difference-in-differences

However, assuming parallel counterfactual trends implies that these two counterfactual trend terms are the same and therefore cancel out

$$
\begin{aligned}
\tau_{\texttt{DID}} &= \mathbb{E}[y_{i1} - y_{i0} \mid D_i = 1] - \mathbb{E}[y_{i1} - y_{i0} \mid D_i = 0] \\
&= \textsf{ATT}_1 + \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = 1] - \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = 0] \\
&= \textsf{ATT}_1 + \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = 1] - \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = 1] \\
&= \textsf{ATT}_1
\end{aligned}
$$

Difference-in-differences as an imputation estimator

Remember in the selection on observables topic, we used a regression imputation estimator to explicitly estimate the treated units' $y_i(0)$.

It turns out, we can write the difference-in-differences estimator as an imputation estimator

Difference-in-differences as an imputation estimator

Our imputation for $y_{i1}(0)$ is given as:

$$\hat{y}_{i1}(0) = y_{i0} + \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = 0]$$

In words, take the unit's period $t = 0$ outcome and add to it the average change in y for the comparison group.

- This is what I was drawing in the figures at the start of the slides

2×2 DID Estimation

Our estimation strategy replaces these terms with their sample averages:

$$\hat{\tau}_{\texttt{DID}} = \hat{\mathbb{E}}[y_{i1} - y_{i0} \mid D_i = 1] - \hat{\mathbb{E}}[y_{i1} - y_{i0} \mid D_i = 0]$$

We could do this as four averages

$$\left(\hat{\mathbb{E}}[y_{i1} \mid D_i = 1] - \hat{\mathbb{E}}[y_{i0} \mid D_i = 1] \right) - \left(\hat{\mathbb{E}}[y_{i1} \mid D_i = 0] - \hat{\mathbb{E}}[y_{i0} \mid D_i = 0] \right)$$

2×2 DID Estimation

Our estimation strategy replaces these terms with their sample averages:

$$\hat{\tau}_{\texttt{DID}} = \hat{\mathbb{E}}[y_{i1} - y_{i0} \mid D_i = 1] - \hat{\mathbb{E}}[y_{i1} - y_{i0} \mid D_i = 0]$$

We could do this as four averages

$$\left(\hat{\mathbb{E}}[y_{i1} \mid D_i = 1] - \hat{\mathbb{E}}[y_{i0} \mid D_i = 1]\right) - \left(\hat{\mathbb{E}}[y_{i1} \mid D_i = 0] - \hat{\mathbb{E}}[y_{i0} \mid D_i = 0]\right)$$

Or just do a difference-in-means using $y_{i1} - y_{i0}$ as the outcome variable

- Be careful to only have one row per unit when running this regression

A note on the name 'Difference-in-Differences'

The correct name is Difference in Difference*s*

- You are taking the difference between *two* averages of first-differences

Personal pet-peeve, but this is the one and only name for this estimator

2×2 in regression form

Just like difference-in-means, it turns out you can use OLS regression to estimate $\hat{\tau}_{\text{DID}}$

$$y_{it} = \alpha + \gamma D_i + \lambda \text{Post}_t + \tau d_{it} + u_{it}$$

- $d_{it} = D_i \text{Post}_t$ is an indicator for when a unit is actively under treatment

2×2 in regression form

$$y_{it} = \alpha + \gamma D_i + \lambda \text{Post}_t + \tau d_{it} + u_{it}$$

Since these are just a bunch of indicator variables, we can derive what they estimate:

$$\mathbb{E}[Y_{it} \mid D_i = 0, \text{Post}_t = 0] = \mathbb{E}[Y_{i0} \mid D_i = 0] \qquad = \alpha \qquad (1)$$

$$\mathbb{E}[Y_{it} \mid D_i = 0, \text{Post}_t = 1] = \mathbb{E}[Y_{i1} \mid D_i = 0] \qquad = \alpha + \lambda \qquad (2)$$

$$\mathbb{E}[Y_{it} \mid D_i = 1, \text{Post}_t = 0] = \mathbb{E}[Y_{i0} \mid D_i = 1] \qquad = \alpha + \gamma \qquad (3)$$

$$\mathbb{E}[Y_{it} \mid D_i = 1, \text{Post}_t = 1] = \mathbb{E}[Y_{i1} \mid D_i = 1] \qquad = \alpha + \gamma + \lambda + \tau \qquad (4)$$

Solving these equations for τ give the DID estimate: $\tau = [(4) - (3)] - [(2) - (1)]$

2×2 in regression form

$$y_{it} = \alpha + \gamma D_i + \lambda \text{Post}_t + \tau d_{it} + u_{it}$$

First, we have $\hat{\alpha} = \hat{\mathbb{E}}[y_{i0} \mid D_i = 0]$ and $\hat{\gamma} = \hat{\mathbb{E}}[y_{i0} \mid D_i = 1]$

Second, we have $\hat{\lambda} = \hat{\mathbb{E}}[y_{i1} - y_{i0} \mid D_i = 0]$

Last, we have

$$\hat{\tau}_{\text{OLS}} = \hat{\mathbb{E}}[y_{i1} - y_{i0} \mid D_i = 1] - \hat{\mathbb{E}}[y_{i1} - y_{i0} \mid D_i = 0]$$

2×2 in regression form

You can also use unit and time fixed-effects to estimate this

$$y_{it} = \mu_i + \lambda_t + \tau d_{it} + u_{it}$$

It is also true that, $\hat{\tau}_{\text{OLS}} = \hat{\tau}_{\text{DID}}$!

- Note that one of the time fixed-effects will need to be omitted for collinearity

This form is usually more common, so we will focus on that

- "Absorb the time-fixed effects of individuals and common time-shocks"

Users Beware !!

The equivalence between OLS and 2×2 DID only holds in this case. Using OLS in other cases will turn out to bite us in the butt later on

- People have been using OLS for DID for a long time and it turns out to create problems when treatment starts at different points in time for different units

Difference-in-Differences

- Initial Difference-in-difference usage
- Classic Example: Card and Krueger (2000, AER)
- Econometric formulation to DID
- Event-study

Conditional Parallel Trends

Staggered Treatment Timing

- Estimating Group-Time ATTs (Callaway and Sant'Anna)
- Imputation based estimators

Multiple pre- and post- periods

Now, consider an extension of the 2×2 DID where we observe N units over $t = 1, \ldots, T_0, T_0 + 1, \ldots, T$ time periods

Treatment turns on for some individuals at period $T_0 + 1$. Here, we still have D_i be the treatment indicator

Dynamic treatment effects

However, we now have many post-periods, so we will look at ATT^ℓ coefficients:

$$\text{ATT}^\ell = \mathbb{E}[y_{it}(1) - y_{it}(0) \mid D_i = 1]$$

- The impact of being treated in period $t > T_0$
- No anticipation implies $\text{ATT}^\ell = 0$ for $l \leq T_0$

Dynamic treatment effects

However, we now have many post-periods, so we will look at ATT$^\ell$ coefficients:

$$\text{ATT}^\ell = \mathbb{E}[y_{it}(1) - y_{it}(0) \mid D_i = 1]$$

- The impact of being treated in period $t > T_0$
- No anticipation implies ATT$^\ell = 0$ for $l \leq T_0$

These are called dynamic treatment effects

- E.g. policies might take a while to be in full-effect (ATT$^\ell$ growing)
- Or, the policy has a large initial shock that fades away as people adapt

Dynamic Treatment Effects

These can be identified using a similar 2×2 strategy:

$$\text{ATT}^{\ell} = \mathbb{E}[y_{it} - y_{iT_0} \mid D_i = 1] - \mathbb{E}[y_{it} - y_{iT_0} \mid D_i = 0]$$

Comparing "long-differences" from period T_0 to period t. This requires parallel trends holds for all post-periods:

$$\mathbb{E}[y_{it}(0) - y_{iT_0}(0) \mid D_i = 1] = \mathbb{E}[y_{it}(0) - y_{iT_0}(0) \mid D_i = 0]$$

Estimating dynamic effects using OLS

Similarly, we can estimate these using OLS:

$$y_{it} = \mu_i + \lambda_t + \sum_{\ell=T_0+1}^{T} d_{it}^{\ell} \tau^{\ell} + v_{it},$$

where $d_{it}^{\ell} = \mathbb{1}[D_i = 1] * \mathbb{1}[t = \ell]$ are called the event-study indicators

Just like the 2×2, the OLS estimates are the sample analogue

$$\hat{\tau}^{\ell} = \hat{\mathbb{E}}[y_{it} - y_{iT_0} \mid D_i = 1] - \hat{\mathbb{E}}[y_{it} - y_{iT_0} \mid D_i = 0]$$

Pre-trends estimates

If you recall, one thing that made us confident in our time-series plots was that the treated and control units had similar trends prior to treatment (bolstering confidence in our parallel trends assumption)

Pre-trends estimates

If you recall, one thing that made us confident in our time-series plots was that the treated and control units had similar trends prior to treatment (bolstering confidence in our parallel trends assumption)

It is common, to include similar d_{it}^{ℓ} event-study indicators for $t < T_0$

$$y_{it} = \mu_i + \lambda_t + \sum_{\ell=1,\ldots,T_0-1,T_0+1,\ldots,T} d_{it}^{\ell} \tau^{\ell} + v_{it},$$

The pre-treatment ℓ estimate 'placebo estimates' and if parallel trends holds in the pre-periods, they should be zero.

“Event-time”

An equivalent way of writing this is to instead use event-time: $t - (T_0 + 1)$

- event-time $= 0$ is the first period of treatment
- event-time $= 1$ is the second period of treatment
- event-time $= -1$ is the period before treatment

Then, you estimate

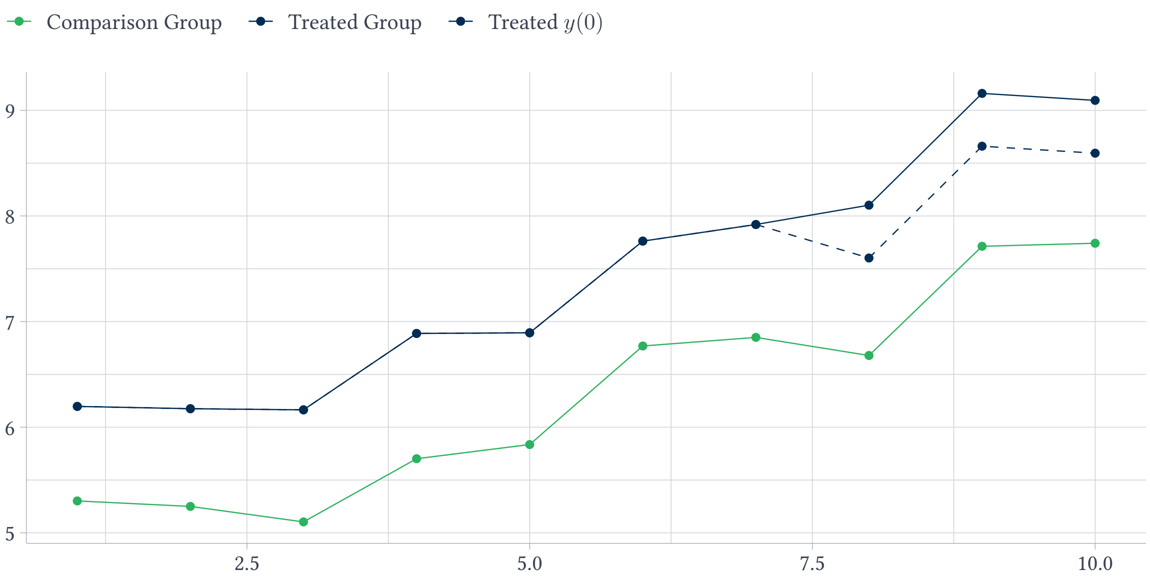
$$y_{it} = \mu_i + \lambda_t + \sum_{\ell = -T_0, \dots, -2, 0, \dots, T - T_0} d_{it}^{\ell} \tau^{\ell} + v_{it},$$

- In this case, the points are the same; just relabeling

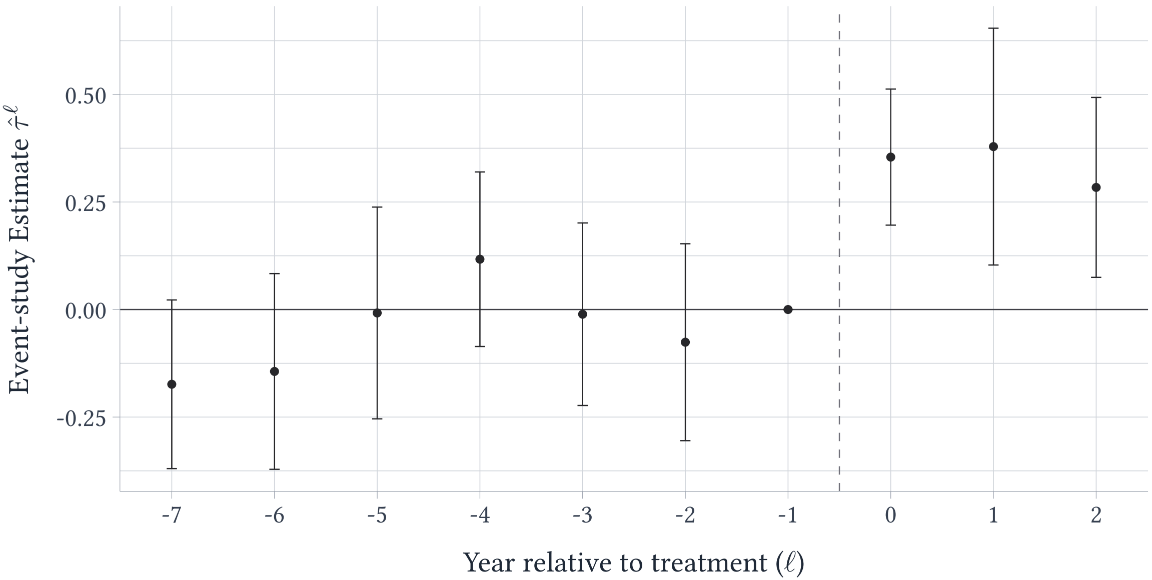
Event-study plots

It is common to plot the coefficients $\hat{\tau}^{\ell}$ to show two things:

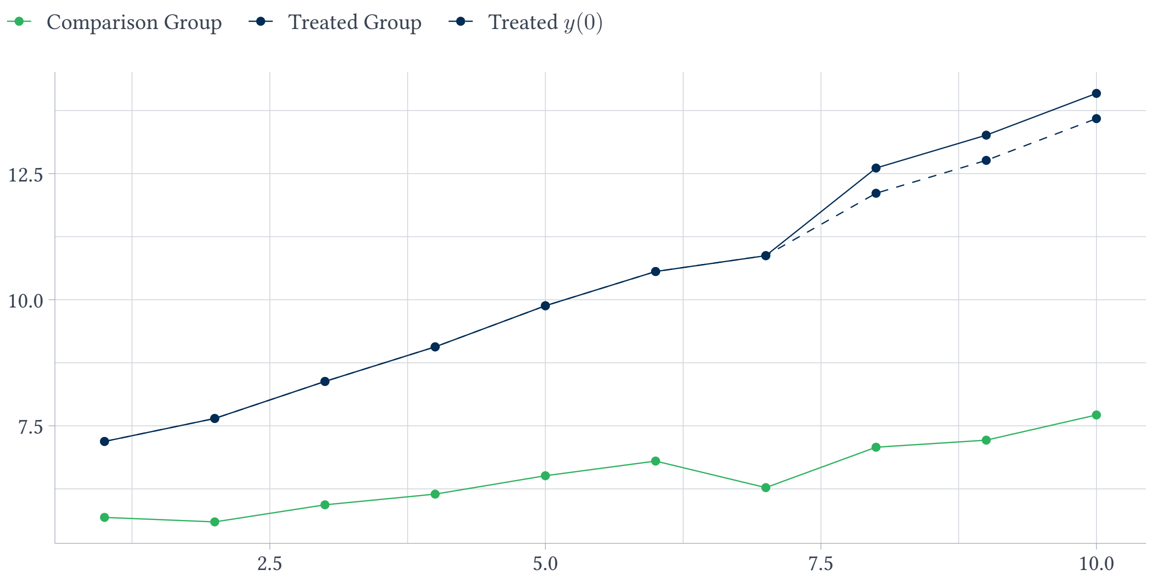
1. The pre-treatment estimates should be near zero and not show any trends
2. The post-treatment effect dynamics

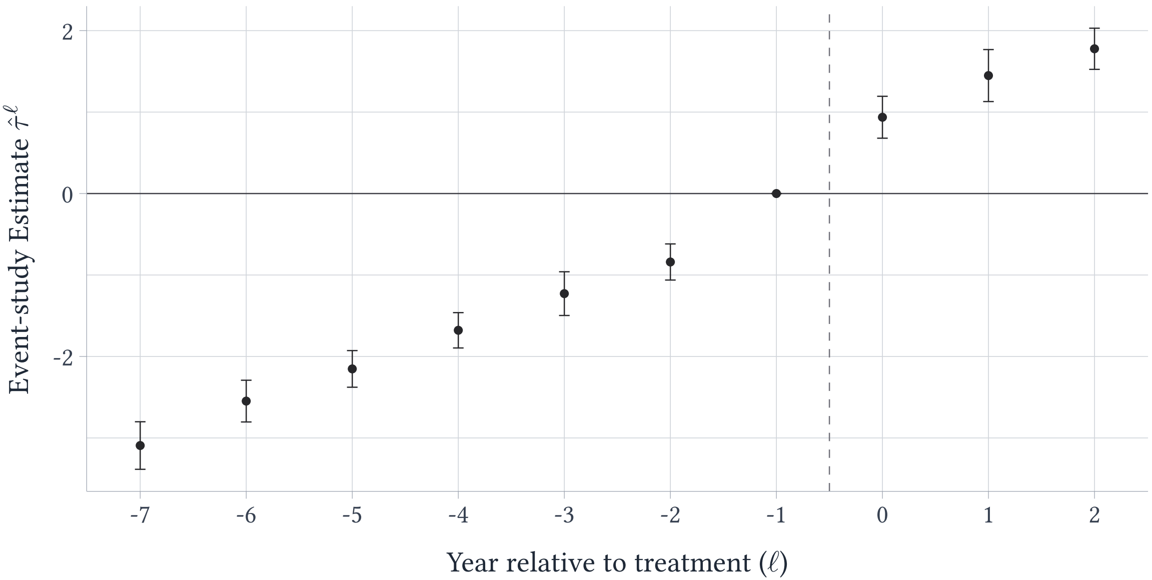


Comparison Group
Treated Group
Treated $y(0)$
9
8
7
6
5
2.5
5.0
7.5
10.0

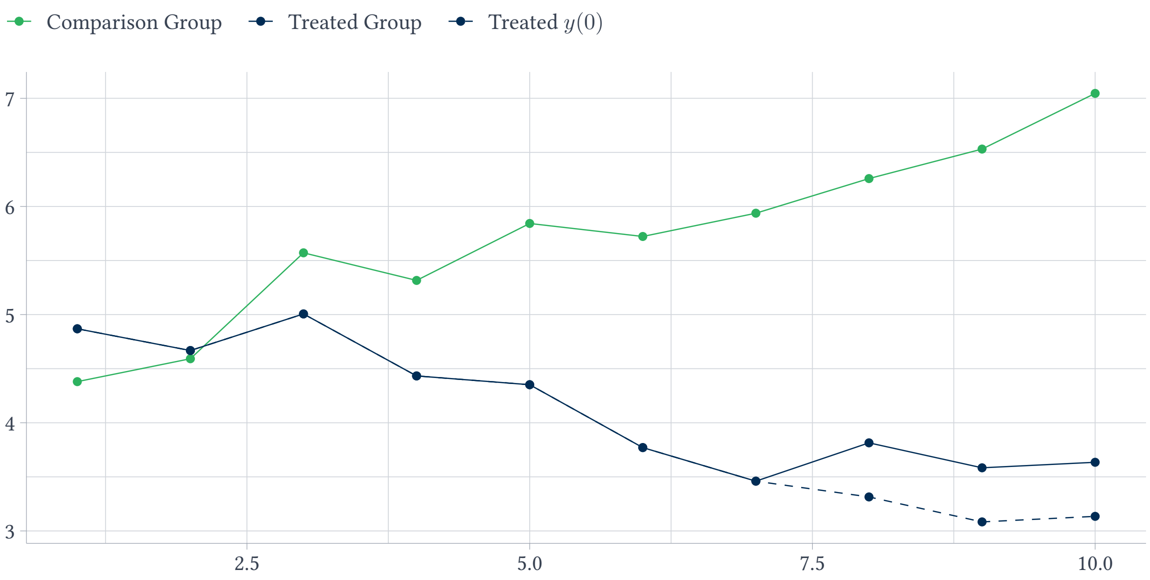


Event-study Estimate $\hat{\tau}^{\ell}$
0.50
0.25
0.00
-0.25
-7
-6
-5
-4
-3
-2
-1
0
1
2
Year relative to treatment (ℓ)

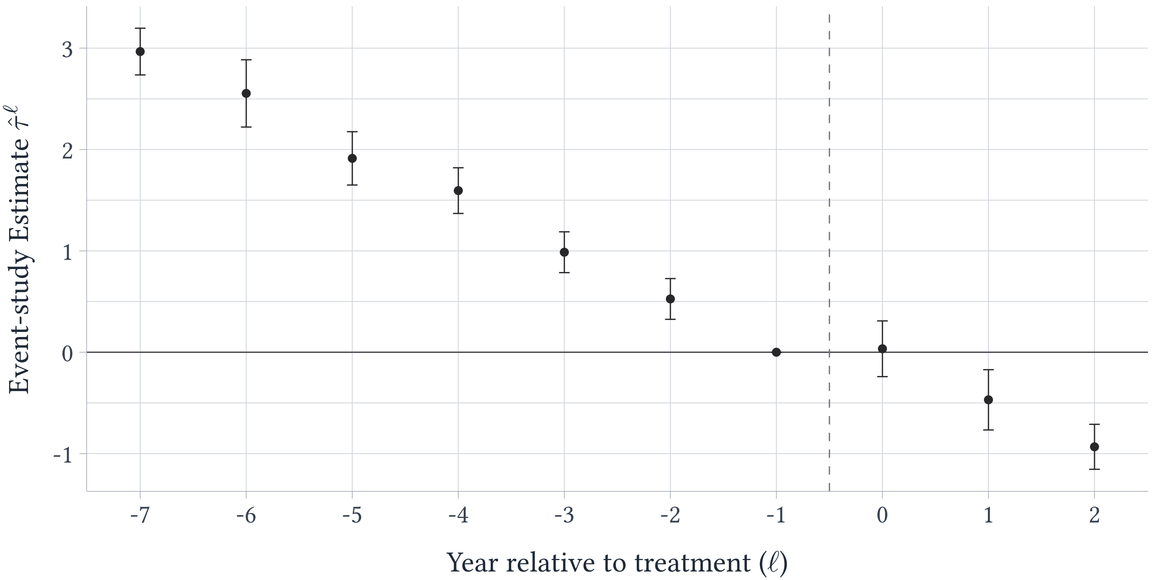

Comparison Group
Treated Group
Treated $y(0)$
12.5
10.0
7.5
2.5
5.0
7.5
10.0



Event-study Estimate $\hat{\tau}^{\ell}$
2
0
-2
-7
-6
-5
-4
-3
-2
-1
0
1
2
Year relative to treatment (ℓ)



Comparison Group
Treated Group
Treated $y(0)$
7
6
5
4
3
2.5
5.0
7.5
10.0



Event-study Estimate $\hat{\tau}^{\ell}$
3
2
1
0
-1
-7
-6
-5
-4
-3
-2
-1
0
1
2
Year relative to treatment (ℓ)

Difference-in-Differences

- Initial Difference-in-difference usage
- Classic Example: Card and Krueger (2000, AER)
- Econometric formulation to DID
- Event-study

Conditional Parallel Trends

Staggered Treatment Timing

- Estimating Group-Time ATTs (Callaway and Sant'Anna)
- Imputation based estimators

Conditional Parallel trends

Just like in the first two topics of the course, economists often times find the parallel trends assumption to be too strong and want to use covariates to relax this assumption

- Similar to completely randomly assigned compared to assigned randomly conditional on observables

Conditional Parallel trends

Just like in the first two topics of the course, economists often times find the parallel trends assumption to be too strong and want to use covariates to relax this assumption

- Similar to completely randomly assigned compared to assigned randomly conditional on observables

Say we have X_i be a unit's gender. We might think men and women have different wage trends in the sample

- I.e. we think looking among people of the same gender, trends are parallel

Conditional Parallel Trends

Just like selection on observables, there is an obvious estimation strategy

- Look within workers with the same gender and estimate a male DID and a female DID

For the overall ATT, we can take a weighted average of gender-specific DID estimates

Conditional Parallel Trends

Let $\boldsymbol{X}_i$ be a set of time-invariant characteristics of a unit i

- E.g. gender, years of education, etc.

Our Conditional Parallel Trends Assumption says

$$\mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = 1, \boldsymbol{X}_i = \boldsymbol{x}] = \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = 0, \boldsymbol{X}_i = \boldsymbol{x}]$$

for all values of $\boldsymbol{x}$

Conditional Parallel Trends does not imply Parallel Trends

Say we have X_i be a unit's gender and conditional parallel trends holds. The unconditional trends for the treated and untreated groups are:

$$
\begin{aligned}
&\mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = d] = \\
&\qquad \mathbb{P}(X_i = F \mid D_i = d)\, \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = d, X_i = F] \\
&\qquad + \mathbb{P}(X_i = M \mid D_i = d)\, \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = d, X_i = M]
\end{aligned}
$$

Conditional Parallel Trends does not imply Parallel Trends

Say we have X_i be a unit's gender and conditional parallel trends holds. The unconditional trends for the treated and untreated groups are:

$$
\begin{aligned}
&\mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = d] = \\
&\qquad \mathbb{P}(X_i = F \mid D_i = d)\, \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = d, X_i = F] \\
&\qquad + \mathbb{P}(X_i = M \mid D_i = d)\, \mathbb{E}[y_{i1}(0) - y_{i0}(0) \mid D_i = d, X_i = M]
\end{aligned}
$$

If the gender composition of the treated group differs from the untreated group, PTs does not hold (even though conditional PTs does)

Estimation Strategies with Conditional Parallel Trends

One way of rewriting the conditional parallel trends assumption is as follows:

$$\mathbb{E}[\Delta y_i(0) \mid D_i = d, \boldsymbol{X}_i = \boldsymbol{x}] = \mathbb{E}[\Delta y_i(0) \mid \boldsymbol{X}_i = \boldsymbol{x}]$$

That is, treatment is mean-independent of the *change in* $y_i(0)$ conditional on $\boldsymbol{X}_i$

Estimation Strategies with Conditional Parallel Trends

One way of rewriting the conditional parallel trends assumption is as follows:

$$\mathbb{E}[\Delta y_i(0) \mid D_i = d, \boldsymbol{X}_i = \boldsymbol{x}] = \mathbb{E}[\Delta y_i(0) \mid \boldsymbol{X}_i = \boldsymbol{x}]$$

That is, treatment is mean-independent of the *change in* $y_i(0)$ conditional on $\boldsymbol{X}_i$

$\implies$ We can treat this like a selection-on-observables problem where $\Delta y_i = y_{i1} - y_{i0}$ is our outcome variable

- Two potential outcomes: $\Delta y_i(1) = y_{i1}(1) - y_{i0}(0)$ and $\Delta y_i(0) = y_{i1}(0) - y_{i0}(0)$
- This unlocks all of our selection-on-observables strategies

Example 1: Matching

$$\mathbb{E}[\Delta y_i(0) \mid D_i = d, \boldsymbol{X}_i = \boldsymbol{x}] = \mathbb{E}[\Delta y_i(0) \mid \boldsymbol{X}_i = \boldsymbol{x}]$$

For instance, we could perform a nearest-neighbor matching exercise where we match up treated and control units with similar $\boldsymbol{X}_i$ and perform a difference-in-means of Δy_i with the matched control group

Example 1: Matching

For event-study type setups, we can do a version of this where we use $y_{it} - y_{iT_0}$ as the outcome variable

- For each $t \in 1, \dots, T_0 - 1, T_0 + 1, \dots, T$, we can compute an estimate using the matched-control group

Collect the $\hat{\tau}^\ell$ and we can make our event-study plot

Regression Adjustment

It is worth spending a moment thinking about the regression adjustment estimator.

Using just the control units, we estimate this model:

$$y_{i1}(0) - y_{i0}(0) = \alpha + \boldsymbol{X}_i \beta + u_i$$

Regression Adjustment

It is worth spending a moment thinking about the regression adjustment estimator.

Using just the control units, we estimate this model:

$$y_{i1}(0) - y_{i0}(0) = \alpha + \boldsymbol{X}_i \beta + u_i$$

$\hat{\beta}$ will estimate how the change in $y_i(0)$ varies by $\boldsymbol{X}_i$

- For e.g. if $\boldsymbol{X}_i$ was an indicator for being a female
  - $\rightarrow$ $\hat{\alpha}$ would estimate the average male trend in $y(0)$
  - $\rightarrow$ $\hat{\beta}$ would estimate the difference in average female trend in $y(0)$ relative to the male trend

Regression Adjustment

$$y_{i1}(0) - y_{i0}(0) = \alpha + \boldsymbol{X}_i \beta + u_i$$

What regression adjustment is doing, in effect, is estimating how trends across units based on their $\boldsymbol{X}_i$ and adjusting for it

Regression Adjustment

$$\Delta y_i(0) = y_{i1}(0) - y_{i0}(0) = \alpha + \boldsymbol{X}_i \beta + u_i$$

Once we have $\hat{\alpha}$ and $\hat{\beta}$ estimated using our $D_i = 0$ units, we can form our DID estimate as

$$\hat{\tau}_{\text{RA}} = \frac{1}{N_1} \sum_{i \,:\, D_i = 1} \underbrace{(\Delta y_i)}_{\Delta y_i(0) + \tau_i} - \underbrace{\left(\hat{\alpha} + \boldsymbol{X}_i \hat{\beta}\right)}_{\widehat{\Delta y_i(0)}}$$

Difference-in-Differences

- Initial Difference-in-difference usage
- Classic Example: Card and Krueger (2000, AER)
- Econometric formulation to DID
- Event-study

Conditional Parallel Trends

Staggered Treatment Timing

- Estimating Group-Time ATTs (Callaway and Sant'Anna)
- Imputation based estimators

Staggered Treatment Timing

Now we turn to settings where treatment begins at different time for different units

- E.g. states roll out a policy over time

Let G_i denote the time-period where a unit i is first starts treatment. The literature refers to G_i as 'treatment-timing group', or 'group' for short

- By convention, units that are never treated in the sample have $G_i = \infty$
- Can not estimate effects for units treated in period 1 ("always-treated"), so we drop them

High-level Overview

There are two prevalent strategies in the literature for estimating effects:

1. The 'building blocks' approach where you estimate small parameters using 2×2 DID and aggregate them up
   - $\rightarrow$ E.g. Callaway and Sant'Anna (2021); deChasiemartin and D'Haultfoeuille (many); Callaway Goodman-Bacon and Sant'Anna (2024)

2. The 'imputation' approach where you try and estimate $y_{it}(0)$ explicitly
   - $\rightarrow$ Borusyak, Jaravel, and Spiess (2024); Gardner (WP)
   - $\rightarrow$ Extends into more complex models (e.g. my own work)

Difference-in-Differences

- Initial Difference-in-difference usage
- Classic Example: Card and Krueger (2000, AER)
- Econometric formulation to DID
- Event-study

Conditional Parallel Trends

Staggered Treatment Timing

- Estimating Group-Time ATTs (Callaway and Sant'Anna)
- Imputation based estimators

Heterogeneity

There are going to be two sources of heterogeneity in this setting:

1. Treatment effects might vary over time for a unit, sometimes called “dynamics”
   - E.g. treatment effects start big but fade out as units adapt
   - Or, effects start small but grow over time as the policy is fully implemented

2. Treatment effects might vary by when you receive treatment
   - E.g. units with larger treatment effects start treatment earliest
   - E.g. entering treatment during a recession has larger impacts on the outcome

Defining potential outcomes

Because of the sources of heterogeneity before, it matters when a unit starts treatment when keeping track of potential outcomes

Unit i at period t has potential outcomes $y_{it}(g)$

- The outcome for unit i in period t in the counterfactual world where they started treatment in period g
- There are now more than 2 potential outcomes!

Defining potential outcomes

For units with $G_i = g$, we observe $y_{it} = y_{it}(g)$ when $t \geq g$ or $y_{it} = y_{it}(g) = y_{it}(\infty)$ when $t < g$

- This is our "switching equation" and a no-anticipation assumption

Defining potential outcomes

For units with $G_i = g$, we observe $y_{it} = y_{it}(g)$ when $t \geq g$ or $y_{it} = y_{it}(g) = y_{it}(\infty)$ when $t < g$

- This is our "switching equation" and a no-anticipation assumption

The treatment effect can be defined as $y_{it}(g) - y_{it}(\infty)$

- For unit i, the treatment effect at time t depends on when you started treatment
  - $\rightarrow$ if I've been treated for 5 periods or 1 period matters (dynamic effects)

Group-time Average Treatment Effect

Since we can only observe $y_{it}(g)$ for units that start treatment in period $G_i = g$, we can only estimate a "group-time ATT":

$$\text{ATT}(g, t) \equiv \mathbb{E}[y_{it}(g) - y_{it}(\infty) \mid G_i = g]$$

- Just like in 2×2, we can only estimate the treatment effect on the treated units

This represents the effect in period t of being treated since period g, averaged over the units that receive treatment in period g

Group-time Average Treatment Effect

$$\text{ATT}(g,t) \equiv \mathbb{E}[y_{it}(g) - y_{it}(\infty) \mid G_i = g]$$

Our previous discussion of heterogeneity can be described as follows:

1. Treatment effects might vary over time for a unit:

   $\rightarrow$ For a given g, $\text{ATT}(g,t)$ varies over t

2. Treatment effects might vary by when you receive treatment

   $\rightarrow$ For a given t, $\text{ATT}(g,t)$ varies over g

Aggregating Effects

The ATT(g, t) parameters are sometimes referred to as the 'building blocks' because you can take weighted averages of these parameters to get more aggregate parameters

- Pro: ATT(g, t) can be quite noisily estimated and be too numerous to report, so summarizing effects can be helpful
- Con: may be meaningless if effects are too heterogeneous

Overall ATT

To get an overall treatment effect, you could do:

$$\text{ATT}_{\text{overall}} = \sum_g \sum_t \mathbb{1}[t \geq g] \frac{1}{N_{g,t}} \text{ATT}(g,t),$$

where $N_{g,t}$ is the number of units in group g in period t

Overall ATT

To get an overall treatment effect, you could do:

$$\text{ATT}_{\text{overall}} = \sum_g \sum_t \mathbb{1}[t \geq g] \frac{1}{N_{g,t}} \text{ATT}(g,t),$$

where $N_{g,t}$ is the number of units in group g in period t

- A single summary measure
- If effects do not vary too much by g and t, then this will typically have the smallest standard errors

Dynamic ATT / Event-Study

We can estimate 'dynamic effects' / 'event-study effects' by averaging over an *event-time* ℓ:

$$\text{ATT}^{\ell} = \sum_{g} \sum_{t} \mathbb{1}[t - g = \ell] \frac{1}{N_{g,t}} \text{ATT}(g, t)$$

- Summarize how effects change over time for a unit, i.e. *dynamic effects*

Who to compare to?

There are now multiple different comparison groups:

- The units that never start treatment ($G_i = \infty$)
- The units that start treatment later than a given unit
- Or both

Our parallel trends assumption will be modified to basically say 'the comparison group we use has the same counterfactual trends as group g'

Parallel Trends Assumption

If we want to use treated units compared to never-treated units, we need to assume for all t and all g

$$\mathbb{E}[y_{it}(\infty) - y_{i,t-1}(\infty) \mid G_i = g] = \mathbb{E}\Big[y_{it}(\infty) - y_{i,t-1}(\infty) \mid \underbrace{\boxed{G_i = \infty}}_{\textit{never-treated units}}\Big]$$

Parallel Trends Assumption

If we want to use treated units compared to never-treated units, we need to assume for all t and all g

$$\mathbb{E}[y_{it}(\infty) - y_{i,t-1}(\infty) \mid G_i = g] = \mathbb{E}\left[y_{it}(\infty) - y_{i,t-1}(\infty) \mid \underbrace{\boxed{G_i = \infty}}_{\textit{never-treated units}}\right]$$

If we want to use treated units compared to not-yet-treated units, we need to assume for all t and all g, g' with $g' > g$:

$$\mathbb{E}[y_{it}(\infty) - y_{i,t-1}(\infty) \mid G_i = g] = \mathbb{E}\left[y_{it}(\infty) - y_{i,t-1}(\infty) \mid \underbrace{\boxed{G_i = g', t < g'}}_{\textit{not-yet-treated units}}\right]$$

Using only not-yet-treated units

Note, you do not want to use units treated before period t. Why?

Using only not-yet-treated units

Note, you do not want to use units treated before period t. Why?

We do not observe $y_{it}(\infty)$ for these units when these units are already treated by time t

- We will come back to this

Estimation of ATT(g, t)

Estimation of ATT(g, t) can be done in the same way as the 2×2, but requires a bit more care:

1. Subset to units that are never-treated and/or not treated by time t
   - $\rightarrow$ The choice of this depends on which parallel trends assumption you believe
2. Then, estimate a 2×2 estimate using periods t and $g - 1$ and $D_i = \mathbb{1}[G_i = g]$
   - $\rightarrow$ Compare change in y for group $G_i = g$ to the not-yet and/or never treated units

Repeating this for all g and t pairs estimates $\hat{\text{ATT}}(g, t)$

Pre-trends Estimates

Similar to before, we can estimate pre-trend estimates by using $t < g - 1$ and doing the same way

- Treat t as the "post"-period and $g - 1$ as the pre-period and estimate 2×2 estimate
  - $\rightarrow$ I.e. regress $y_{i,t} - y_{i,g-1}$ on $D_i = \mathbb{1}[G_i = g]$

In this way, we can estimate an event-study plot for each group g and plot them:

- E.g. make an event-study plot for units treated in 2012 and a separate plot for those treated in 2013

R code

The package `did` implements this looping over g and t for you

- Note of caution later on that's important regarding `base_period` argument

When using covariates, will (internally) use the `DRDID` package to estimate each `ATT`(g,t)

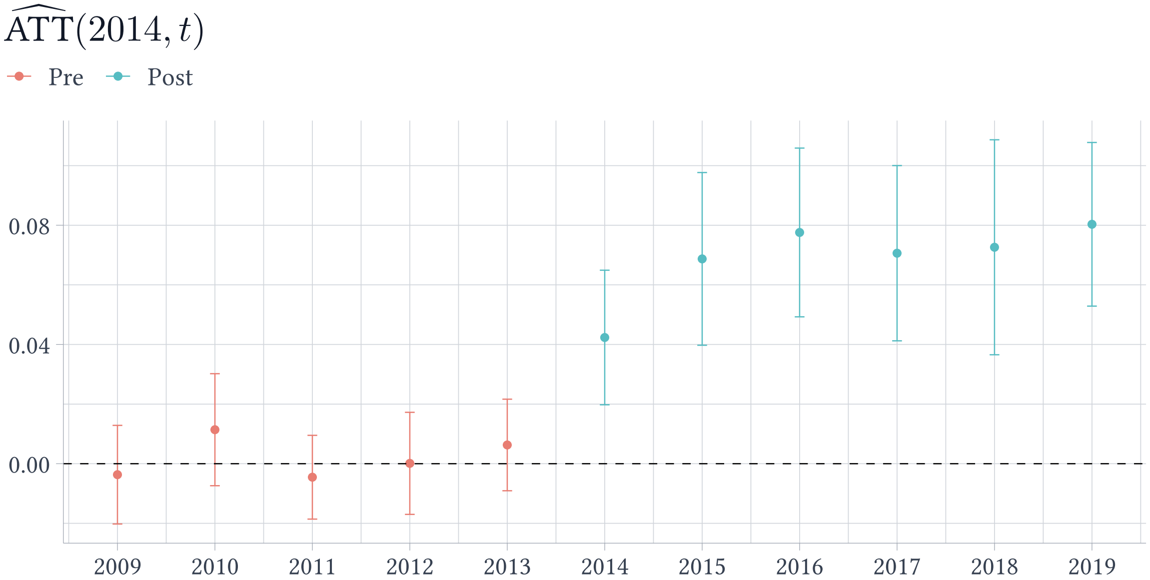
- Gives you immediate access to regression adjustment, IPTW, and doubly-robust estimators for ATT(g,t) estimation

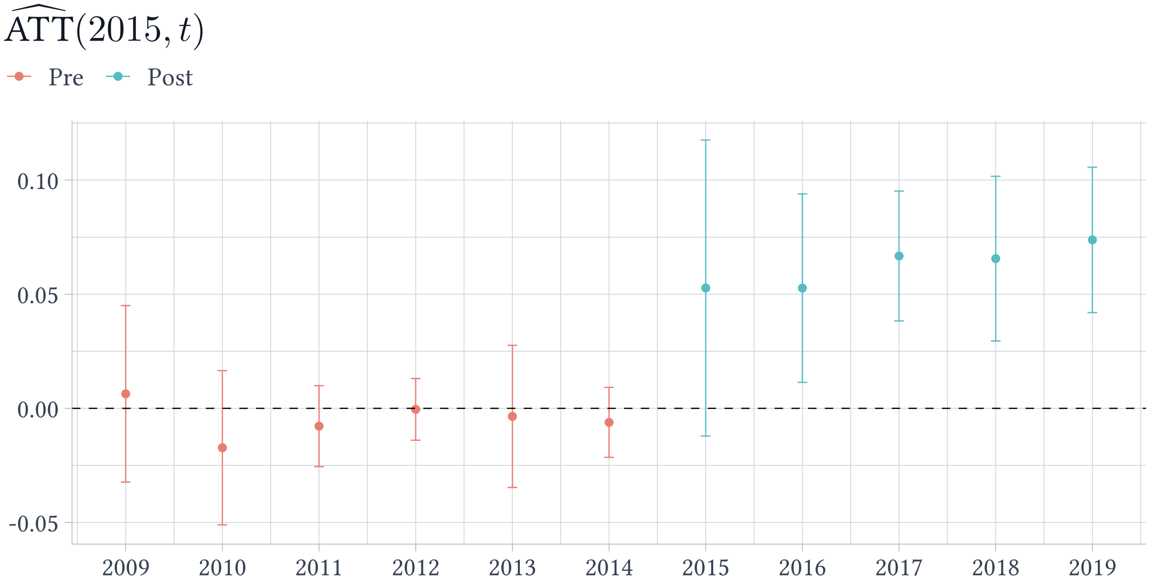
Empirical Example

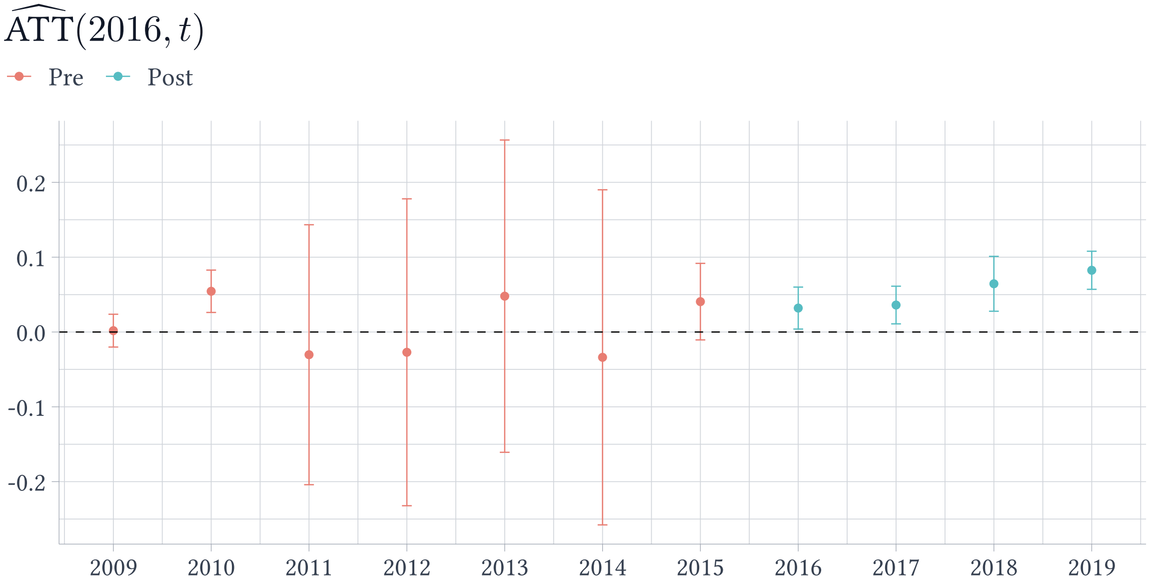
We will look at a very simple example looking at state roll-out of medicaid expansion on the rate of being insured.

- Use public ACS survey data to have state-by-year panel of insured rate
- Similar to Carey, Miller, and Wherry (2020, AEJ applied), although they use confidential data.

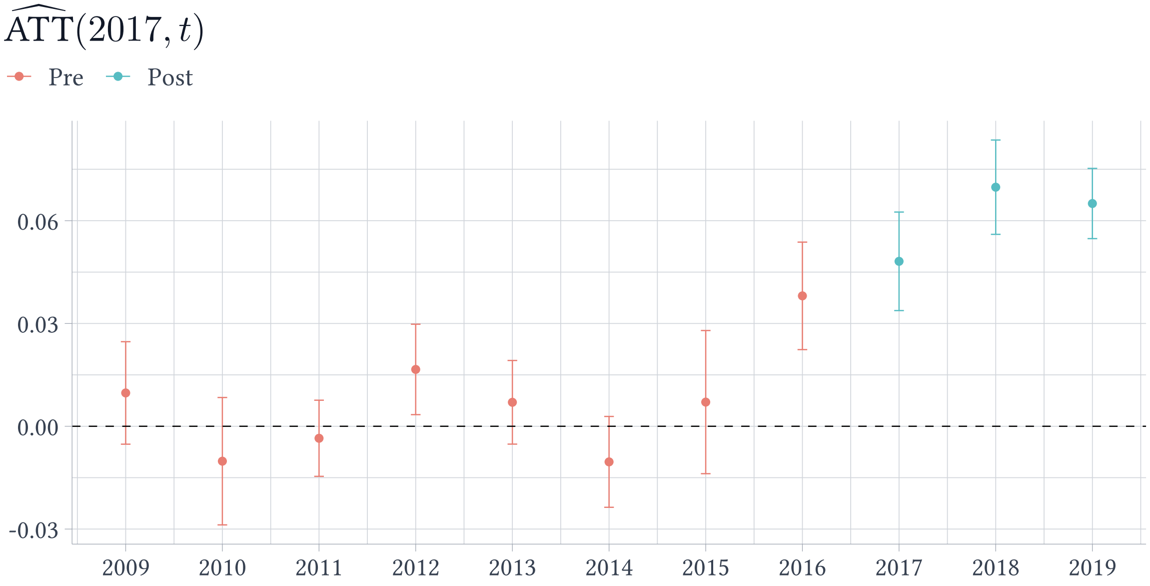
First, use `did::att_gt` function to estimate ATT(g, t) parameters. Then, plot with `did::ggdid`.


$\widehat{\text{ATT}}(2014, t)$
Pre
Post
0.08
0.04
0.00
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019

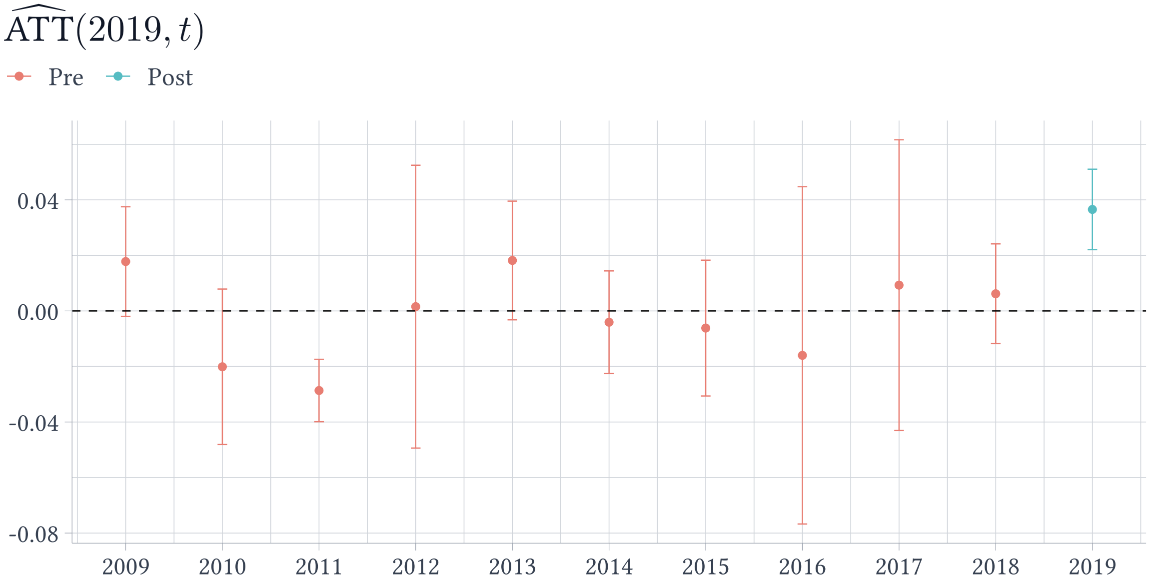

$\widehat{\text{ATT}}(2015, t)$
Pre
Post
0.10
0.05
0.00
-0.05
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019



$\widehat{\text{ATT}}(2016, t)$
Pre
Post
0.2
0.1
0.0
-0.1
-0.2
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019



$\widehat{\text{ATT}}(2017, t)$
Pre
Post
0.06
0.03
0.00
-0.03
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019


$\widehat{\text{ATT}}(2019, t)$
Pre
Post
0.04
0.00
-0.04
-0.08
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019

Warnings with Small Groups

In this example I get a warning

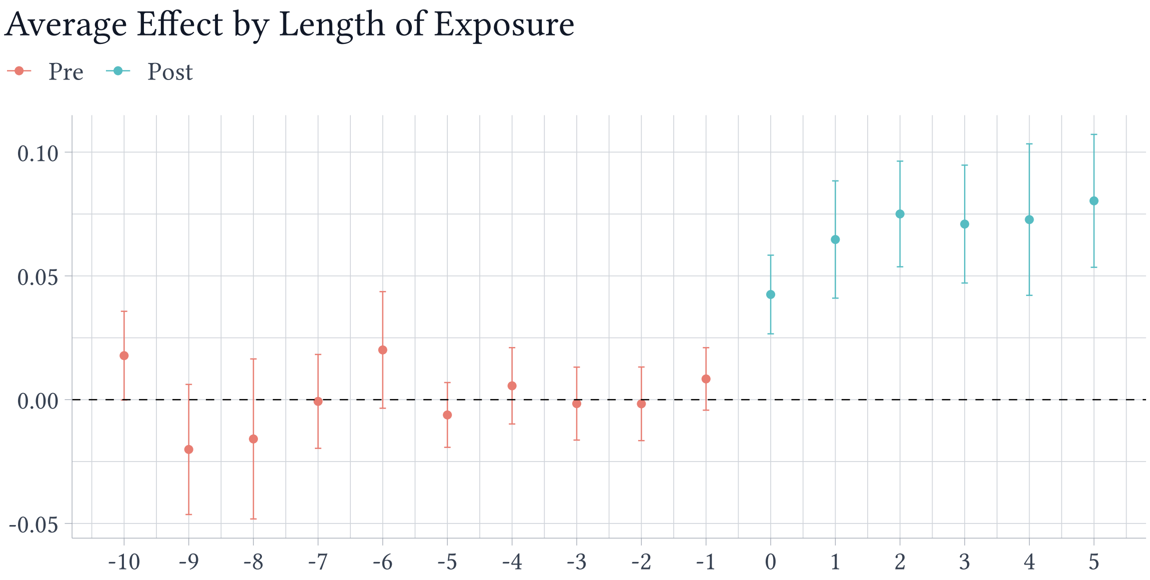
> `Be aware that there are some small groups in your dataset.`
> `Check groups: 2015, 2016, 2017, 2019.`

This is because there are only a few states in some of the groups

- With N_g small, inference based on the normality assumption is problematic

Aggregating to dynamic effects

Last, I can use the `did::aggte(MP, type = "dynamic")` with the results from `att_gt()` to aggregate to $\hat{\text{ATT}}^{\ell}$.


Average Effect by Length of Exposure
Pre
Post
0.10
0.05
0.00
-0.05
-10
-9
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5

Pre-trend estimates

By default the `did::att_gt` function does something weird with pre-trends

- I will not go into the details, but the default argument makes the event-study plots not what people expect them to be
- My recommendation is to add the argument `base_period = "universal"` to all of your `att_gt` calls

See Jon Roth's working paper "Interpreting Event-Studies from Recent Difference-in-Differences Methods" for a few page super applied-friendly summary of this

Difference-in-Differences

- Initial Difference-in-difference usage
- Classic Example: Card and Krueger (2000, AER)
- Econometric formulation to DID
- Event-study

Conditional Parallel Trends

Staggered Treatment Timing

- Estimating Group-Time ATTs (Callaway and Sant’Anna)
- Imputation based estimators

Imputation-based Estimation Strategies

Now, let's turn to the alternate estimation strategy of imputation. At a high-level, imputation estimators:

1. impose a model for the missing untreated counterfactual, $y_{it}(\infty)$
2. fit that model using untreated / not-yet-treated observations
3. then predict out-of-sample for the post-treatment observations, $\hat{y}_{it}(\infty)$

Treatment effects are then calculated as averages of the difference between observed y_{it} and the imputed counterfactual $\hat{y}_{it}(\infty)$

- This is a lot like the regression adjustment estimate we have discussed

Two-way fixed effect model

The standard model to impose is the two-way fixed effect model:

$$y_{it}(\infty) = \mu_i + \lambda_t + u_{it},$$

where we assume a version of parallel-trends that $\mathbb{E}[u_{it}] = 0$ for all (i, t)

The λ_t are the common trends between units

Estimate the model

$$y_{it}(\infty) = \mu_i + \lambda_t + u_{it},$$

We will estimate this model using all observations with $d_{it} = 0$

- Both never-treated units and treated units prior to treatment
- Do not include $d_{it} = 1$ because their $y_{it} = y_{it}(g)$, not $y_{it}(\infty)$

Estimate the model

$$y_{it}(\infty) = \mu_i + \lambda_t + u_{it},$$

We will estimate this model using all observations with $d_{it} = 0$

- Both never-treated units and treated units prior to treatment
- Do not include $d_{it} = 1$ because their $y_{it} = y_{it}(g)$, not $y_{it}(\infty)$

Collect $\hat{\mu}_i$ and $\hat{\lambda}_t$ and predict out of sample $\hat{y}_{it}(\infty)$ for the full sample.

Imputation-based estimation of ATT(g, t)

With an estimated $\hat{y}_{it}(\infty)$, we can plug directly into estimands and take sample averages, e.g.:

$$\begin{aligned} \text{ATT}(g,t) &\equiv \mathbb{E}[y_{it}(g) - y_{it}(\infty) \mid G_i = g] \\ &= \mathbb{E}[y_{it} - y_{it}(\infty) \mid G_i = g] \end{aligned}$$

Our estimate then becomes

$$\hat{\mathbb{E}}[y_{it} - \hat{y}_{it}(\infty) \mid G_i = g]$$

Imputation-based estimation of ATT^ℓ

Or, we can do the same with event-study estimands, ATT^ℓ:

$$\text{ATT}^\ell = \mathbb{E}[y_{it} - \hat{y}_{it}(\infty) \mid t - G_i = \ell]$$

Can more easily calculate these averages with regression of $y_{it} - \hat{y}_{it}(\infty)$ on event-study indicators

- That is, $D_i \times$ event-time indicators

Two-stage DID

To summarize, our estimation procedure is

1. Estimate model for $y_{it}(\infty)$ using observations with $d_{it} = 0$ and get fitted values for full sample, $\hat{y}_{it}(\infty)$
2. Regress $y_{it} - \hat{y}_{it}(\infty)$ on event-study indicators

Two-stage DID

To summarize, our estimation procedure is

1. Estimate model for $y_{it}(\infty)$ using observations with $d_{it} = 0$ and get fitted values for full sample, $\hat{y}_{it}(\infty)$
2. Regress $y_{it} - \hat{y}_{it}(\infty)$ on event-study indicators

Note that the second-stage regression estimates will have incorrect standard errors

- The outcome variable $y_{it} - \hat{y}_{it}(\infty)$ is generated from a first-stage regression
- Inference is corrected in my package `did2s`

Adding Covariates

To relax the parallel trends assumption, we can add covariates to our model for $y_{it}(\infty)$:

$$y_{it}(\infty) = \mu_i + \lambda_t + f_t(\boldsymbol{X}_i) + u_{it},$$

where f_t is a time-varying function of $\boldsymbol{X}_i$ to allow differential trends based on a unit's characteristics

Adding Covariates

To relax the parallel trends assumption, we can add covariates to our model for $y_{it}(\infty)$:

$$y_{it}(\infty) = \mu_i + \lambda_t + f_t(\boldsymbol{X}_i) + u_{it},$$

where f_t is a time-varying function of $\boldsymbol{X}_i$ to allow differential trends based on a unit's characteristics

For example, you can assume $f_t(\boldsymbol{X}_i) = \boldsymbol{X}_i \beta_t$

- If $\boldsymbol{X}_i$ are indicators, then this allows completely general trends
- If $\boldsymbol{X}_i$ is continuous, then this allows trends to depend *linearly* on X_i with β_t being time "shocks"

Adding Covariates

$$y_{it}(\infty) = \mu_i + \lambda_t + \boldsymbol{X}_i \beta_t + u_{it},$$

With our modified model, we can then proceed as usual:

- Estimate $y_{it}(\infty)$ using never-treated and not-yet-treated observations
- Predict $\hat{y}_{it}(\infty)$ out of sample and regress $y_{it} - \hat{y}_{it}(\infty)$ on event-study dummies

Flexibility of imputation

One advantage of imputation estimators is the flexibility they offer

- We write an explicit model for the never-treated potential outcome and use that model to estimate treatment effects

For our last topic, we will discuss how to adapt this procedure for settings where parallel trends fails.