Difference in Differences

Lukáš Laffers

Matej Bel University, Dept. of Mathematics

One of the current leading research designs for estimating causal effects.

It is based on the assumption that differences across units in time should be the same (similar) absent the treatment.

Any time-constant unobservables are taken care of.

It is very popular (26% of the most cited paper published in 2015-2019 used DiD)

This lecture

- Examples
- 2x2 setup
- Identification
- Regression formulation + covariates
- Different complications
- DiD with covariates (without linearity)
- Two-way fixed effects model (TWFE)
- (*) Recent developments (problems with TWFE)

John Snow - Cholera (1854)

The first careful analysis of this type was done by epidemiologist John Snow in the 19th century in Soho, London.

John Snow - Cholera (1854)

The first careful analysis of this type was done by epidemiologist John Snow in the 19th century in Soho, London.

At the time of the cholera outbreak, it was believed it was spread via *miasma* (via ”air”)

John Snow - Cholera (1854)

The first careful analysis of this type was done by epidemiologist John Snow in the 19th century in Soho, London.

At the time of the cholera outbreak, it was believed it was spread via *miasma* (via ”air”)

Snow challenged this view via his careful analysis.

John Snow - Cholera (1854)

The first careful analysis of this type was done by epidemiologist John Snow in the 19th century in Soho, London.

At the time of the cholera outbreak, it was believed it was spread via *miasma* (via ”air”)

Snow challenged this view via his careful analysis.

Snow compared the evolution of cholera related deaths with 2 groups of (otherwise similar) houses where one group had their water supply changed for a cleaner one.

Source: `https://www.rcseng.ac.uk/library-and-publications/library/blog/mapping-disease-john-snow-and-cholera/`

TREATED CONTROL

Cholera deaths

Water company	year 1849	year 1854	Difference
Lambeth	85	19	-66
Soutwark and Vauxhall	135	147	12
Difference in differences			(-66) - 12 = -78

$$(Y^{L}_{1854} - Y^{L}_{1849}) - (Y^{SV}_{1854} - Y^{SV}_{1849}) = (-66) - 12 = -78$$

TREATED CONTROL

Cholera deaths

Water company	year 1849	year 1854
Lambeth	85	19
Soutwark and Vauxhall	135	147
Difference	-50	-138
Difference in differences	-138 - (-50) = -78	

$$(Y^{L}_{1854} - Y^{SV}_{1854}) - (Y^{L}_{1849} - Y^{SV}_{1849}) = -138 - (-50) = -78$$

Example: Minimum wage and employment

What is the impact of minimum wages on employment?

Example: Minimum wage and employment

What is the impact of minimum wages on employment? From February '92 to November '92:

Pennysylvania (control): \$4.25 $\rightarrow$ \$4.25
New Yersey (treated): \$4.25 $\rightarrow$ \$5.05

Example: Minimum wage and employment

What is the impact of minimum wages on employment? From February '92

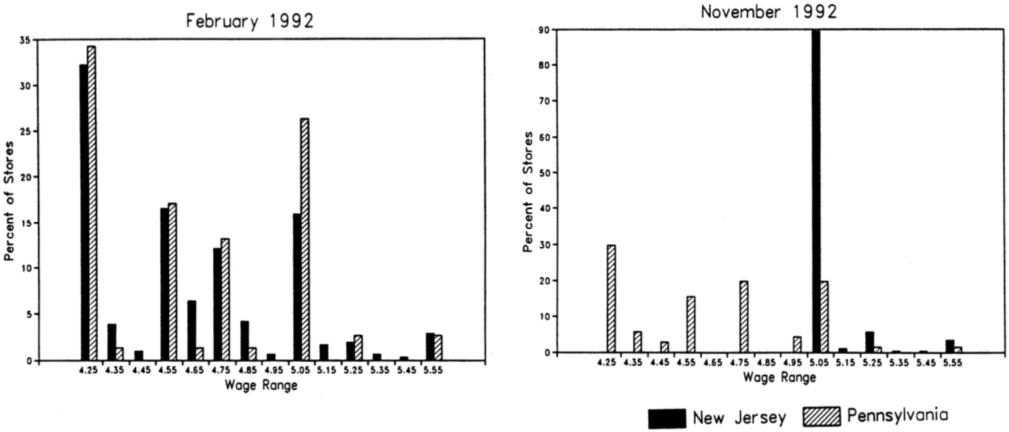
to November '92:
Pennysylvania (control): $\$4.25 \rightarrow \$4.25$
New Yersey (treated): $\$4.25 \rightarrow \$5.05$

They look at the subpopulation where minimum wage mattered: surveyed 400 fast-food restaurants.

Outcome variable was the average number of employees per store.

Card and Krueger (1994)

Was minimum wage binding?



Source: Figure 2 in Card and Krueger (1994).

Card and Krueger (1994)

Average employment per store

State	February	November	Difference
Pennsylvania (control)	23.3	21.14	-2.16
New Yersey (treated)	20.44	21.0	0.56
Difference	-2.86	-0.14	
Difference in differences	-0.14 - (-2.86) = 2.72		0.56 - (-2.16) = 2.72

Card and Krueger (1994)

Average employment per store

State	February	November	Difference
Pennsylvania (control)	23.3	21.14	-2.16
New Yersey (treated)	20.44	21.0	0.56
Difference	-2.86	-0.14	
Difference in differences	-0.14 - (-2.86) = 2.72		0.56 - (-2.16) = 2.72

$$(E[Y_{Nov}|NY] - E[Y_{Nov}|PA]) - (E[Y_{Feb}|NY] - E[Y_{Feb}|PA]) = -0.14 - (-2.86) = 2.72$$

Card and Krueger (1994)

Average employment per store

State	February	November	Difference
Pennsylvania (control)	23.3	21.14	-2.16
New Yersey (treated)	20.44	21.0	0.56
Difference	-2.86	-0.14	
Difference in differences	-0.14 - (-2.86) = 2.72		0.56 - (-2.16) = 2.72

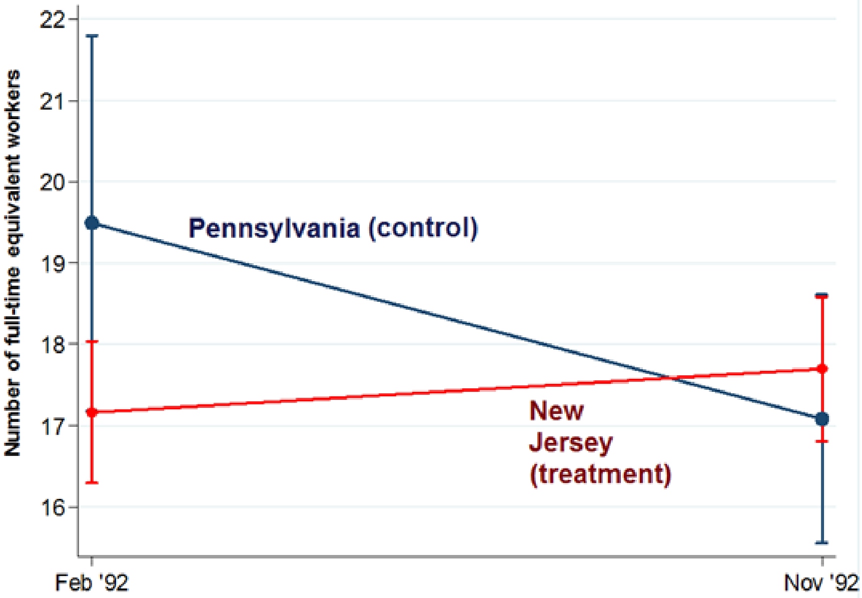
$$(E[Y_{Nov}|NY] - E[Y_{Nov}|PA]) - (E[Y_{Feb}|NY] - E[Y_{Feb}|PA]) = -0.14 - (-2.86) = 2.72$$

$$(E[Y_1|D=1] - E[Y_1|D=0]) - (E[Y_0|D=1] - E[Y_0|D=0]) = -0.14 - (-2.86) = 2.72$$

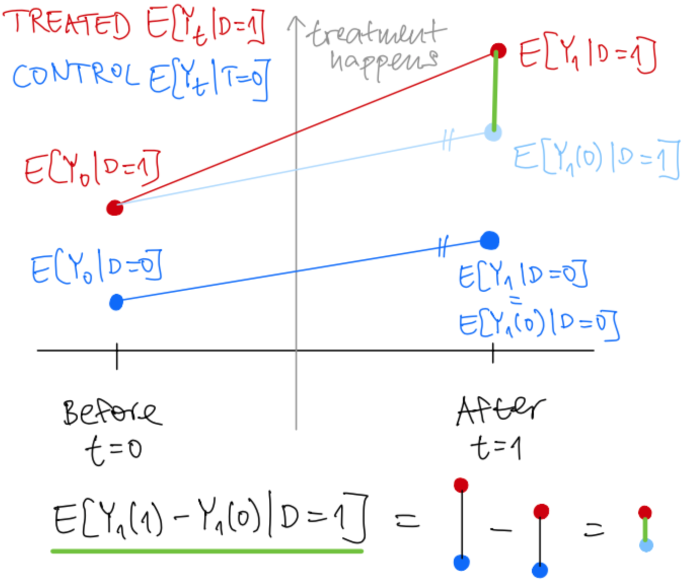
Again. How comparable are the units?

Again. How comparable are the units?

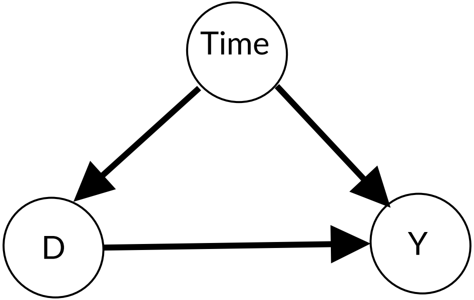
Work hard to convince your reader it is the treatment that matters. Apples to Apples.


Number of full-time equivalent workers
22
21
20
19
18
17
16
Pennsylvania (control)
New Jersey (treatment)
Feb '92
Nov '92

Basic 2x2 case



Causal Graphical Model



Outcomes are changing in time and this is unrelated to the treatment.

Identification

What we have seen before:

- Under $(Y(0), Y(1)) \perp\!\!\!\perp D$, we have
  $ATE = E[Y(1) - Y(0)] = E[Y|D = 1] - E[Y|D = 0]$

Identification

What we have seen before:

- Under $(Y(0), Y(1)) \perp\!\!\!\perp D$, we have
  $ATE = E[Y(1) - Y(0)] = E[Y|D = 1] - E[Y|D = 0]$
- Under $Y(0) \perp\!\!\!\perp D$, we have
  $ATT = E[Y(1) - Y(0)|D = 1] = E[Y|D = 1] - E[Y|D = 0]$

Identification

What we have seen before:

- Under $(Y(0), Y(1)) \perp\!\!\!\perp D$, we have
  $ATE = E[Y(1) - Y(0)] = E[Y|D = 1] - E[Y|D = 0]$
- Under $Y(0) \perp\!\!\!\perp D$, we have
  $ATT = E[Y(1) - Y(0)|D = 1] = E[Y|D = 1] - E[Y|D = 0]$

Here, we have introduced time, thus we have countrafactuals $Y_t(1), Y_t(0)$ and observed Y_t.

$Y_0(d) = Y_{\text{before}}(d)$ and $Y_1(d) = Y_{\text{after}}(d)$

Identification

What we have seen before:

- Under $(Y(0), Y(1)) \perp\!\!\!\perp D$, we have
  $ATE = E[Y(1) - Y(0)] = E[Y|D = 1] - E[Y|D = 0]$
- Under $Y(0) \perp\!\!\!\perp D$, we have
  $ATT = E[Y(1) - Y(0)|D = 1] = E[Y|D = 1] - E[Y|D = 0]$

Here, we have introduced time, thus we have countrafactuals $Y_t(1), Y_t(0)$ and observed Y_t.

$Y_0(d) = Y_{\text{before}}(d)$ and $Y_1(d) = Y_{\text{after}}(d)$

This is the object of interest:

$$ATT = E[Y_1(1) - Y_1(0)|D = 1] = E[Y_1(1)|D = 1] - \underbrace{E[Y_1(0)|D = 1]}_{\text{unobserved}}$$

Identification

How do we identify *ATT* ?

Identification

How do we identify *ATT* ?

Assumption 1: Consistency assumption

$$\forall t : D = d \implies Y_t = Y_t(d)$$

Identification

How do we identify *ATT* ?

Assumption 1: Consistency assumption

$\forall t : D = d \implies Y_t = Y_t(d)$

Assumption 2: Parallel trends

$E[Y_1(0) - Y_0(0)|D = 1] = E[Y_1(0) - Y_0(0)|D = 0]$

(weaker than $(Y_1(0) - Y_0(0)) \perp\!\!\!\perp D$

Identification

How do we identify ATT ?

Assumption 1: Consistency assumption

$\forall t : D = d \implies Y_t = Y_t(d)$

Assumption 2: Parallel trends

$E[Y_1(0) - Y_0(0)|D = 1] = E[Y_1(0) - Y_0(0)|D = 0]$

(weaker than $(Y_1(0) - Y_0(0)) \perp\!\!\!\perp D$

Assumption 3: No pre-treatment effect

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

Identification

How do we identify *ATT* ?

Assumption 1: Consistency assumption
$\forall t : D = d \implies Y_t = Y_t(d)$

Assumption 2: Parallel trends
$E[Y_1(0) - Y_0(0)|D = 1] = E[Y_1(0) - Y_0(0)|D = 0]$
(weaker than $(Y_1(0) - Y_0(0)) \perp\!\!\!\perp D$

Assumption 3: No pre-treatment effect
$E[Y_0(1)|D = 1] - E[Y_0(0)|D = 1] = 0$

Assumption 4: SUTVA (often not stated explicitly)
No interactions between individuals and no hidden versions of the treatment (no hidden variability, everyone receives the same treatment)

Identification

How do we identify ATT ?

$$ATT = E[Y_1(1) - Y_1(0) | D = 1] \quad \text{(definition)}$$

Identification

How do we identify *ATT* ?

$$\begin{aligned} ATT &= E[Y_1(1) - Y_1(0)|D=1] \quad \text{(definition)} \\ &= E[Y_1(1)|D=1] - E[Y_1(0)|D=1] \quad \text{(linearity of } E(\cdot)\text{)} \end{aligned}$$

Identification

How do we identify *ATT* ?

$$
\begin{aligned}
ATT &= E[Y_1(1) - Y_1(0)|D=1] \quad \text{(definition)} \\
&= E[Y_1(1)|D=1] - E[Y_1(0)|D=1] \quad \text{(linearity of } E(\cdot)\text{)} \\
&= E[Y_1|D=1] - E[Y_1(0)|D=1]
\end{aligned}
$$

Identification

How do we identify ATT ?

$$
\begin{aligned}
ATT &= E[Y_1(1) - Y_1(0)|D=1] \quad \text{(definition)} \\
&= E[Y_1(1)|D=1] - E[Y_1(0)|D=1] \quad \text{(linearity of } E(\cdot)\text{)} \\
&= E[Y_1|D=1] - E[Y_1(0)|D=1] \\
&= E[Y_1|D=1] - \big(E[Y_0(0)|D=1] + E[Y_1(0)|D=0] - E[Y_0(0)|D=0]\big)
\end{aligned}
$$

Identification

How do we identify ATT ?

$$
\begin{aligned}
ATT &= E[Y_1(1) - Y_1(0)|D=1] \quad \text{(definition)} \\
&= E[Y_1(1)|D=1] - E[Y_1(0)|D=1] \quad \text{(linearity of } E(\cdot)\text{)} \\
&= E[Y_1|D=1] - E[Y_1(0)|D=1] \\
&= E[Y_1|D=1] - \big(E[Y_0(0)|D=1] + E[Y_1(0)|D=0] - E[Y_0(0)|D=0]\big) \\
&= E[Y_1|D=1] - \big(E[Y_0(0)|D=1] + E[Y_1|D=0] - E[Y_0|D=0]\big)
\end{aligned}
$$

Identification

How do we identify *ATT* ?

$$
\begin{aligned}
ATT &= E[Y_1(1) - Y_1(0)|D=1] \quad \text{(definition)} \\
&= E[Y_1(1)|D=1] - E[Y_1(0)|D=1] \quad \text{(linearity of } E(\cdot)\text{)} \\
&= E[Y_1|D=1] - E[Y_1(0)|D=1] \\
&= E[Y_1|D=1] - \big(E[Y_0(0)|D=1] + E[Y_1(0)|D=0] - E[Y_0(0)|D=0]\big) \\
&= E[Y_1|D=1] - \big(E[Y_0(0)|D=1] + E[Y_1|D=0] - E[Y_0|D=0]\big) \\
&= E[Y_1|D=1] - \big(E[Y_0(1)|D=1] + E[Y_1|D=0] - E[Y_0|D=0]\big)
\end{aligned}
$$

Identification

How do we identify ATT ?

$$
\begin{aligned}
ATT &= E[Y_1(1) - Y_1(0)|D=1] \quad \text{(definition)} \\
&= E[Y_1(1)|D=1] - E[Y_1(0)|D=1] \quad \text{(linearity of } E(\cdot)) \\
&= E[Y_1|D=1] - E[Y_1(0)|D=1] \\
&= E[Y_1|D=1] - \big(E[Y_0(0)|D=1] + E[Y_1(0)|D=0] - E[Y_0(0)|D=0]\big) \\
&= E[Y_1|D=1] - \big(E[Y_0(0)|D=1] + E[Y_1|D=0] - E[Y_0|D=0]\big) \\
&= E[Y_1|D=1] - \big(E[Y_0(1)|D=1] + E[Y_1|D=0] - E[Y_0|D=0]\big) \\
&= E[Y_1|D=1] - \big(E[Y_0|D=1] + E[Y_1|D=0] - E[Y_0|D=0]\big)
\end{aligned}
$$

Identification

How do we identify ATT ?

$$
\begin{aligned}
ATT &= E[Y_1(1) - Y_1(0)|D=1] \quad \text{(definition)} \\
&= E[Y_1(1)|D=1] - E[Y_1(0)|D=1] \quad \text{(linearity of } E(\cdot)\text{)} \\
&= E[Y_1|D=1] - E[Y_1(0)|D=1] \\
&= E[Y_1|D=1] - \big(E[Y_0(0)|D=1] + E[Y_1(0)|D=0] - E[Y_0(0)|D=0]\big) \\
&= E[Y_1|D=1] - \big(E[Y_0(0)|D=1] + E[Y_1|D=0] - E[Y_0|D=0]\big) \\
&= E[Y_1|D=1] - \big(E[Y_0(1)|D=1] + E[Y_1|D=0] - E[Y_0|D=0]\big) \\
&= E[Y_1|D=1] - \big(E[Y_0|D=1] + E[Y_1|D=0] - E[Y_0|D=0]\big) \\
&= \underbrace{\big(E[Y_1|D=1] - E[Y_0|D=1]\big) + \big(E[Y_1|D=0] - E[Y_0|D=0]\big)\big)}_{\text{observed quantities only}}
\end{aligned}
$$

Regression formulation

- Treatment assignment: $D \in \{0, 1\}$
- Time pre/post, before/after: $T \in \{0, 1\}$

$$Y = \beta_0 + \beta_1 D + \beta_2 T + \beta_3 D \cdot T + \varepsilon$$

Regression formulation

- Treatment assignment: $D \in \{0, 1\}$
- Time pre/post, before/after: $T \in \{0, 1\}$

$$Y = \beta_0 + \beta_1 D + \beta_2 T + \beta_3 D \cdot T + \varepsilon$$

This is a saturated model.

- $\beta_0 = E[Y_0 | D = 0]$
- $\beta_1 = E[Y_1 | D = 0] - E[Y_0 | D = 0]$
- $\beta_2 = E[Y_0 | D = 1] - E[Y_0 | D = 0]$
- $\beta_3 = (E[Y_1 | D = 1] - E[Y_1 | D = 0]) - (E[Y_0 | D = 1] - E[Y_0 | D = 0])$

Complications

- Parallel trends may only hold conditional on X
  $E[Y_1(0) - Y_0(0)|X, D = 1] = E[Y_1(0) - Y_0(0)|X, D = 0]$

Complications

- Parallel trends may only hold conditional on X
  $E[Y_1(0) - Y_0(0)|X, D = 1] = E[Y_1(0) - Y_0(0)|X, D = 0]$
- Parallel trends assumption is NOT scale invariant
  $E[Y_1(0) - Y_0(0)|D = 1] = E[Y_1(0) - Y_0(0)|D = 0] \not\Rightarrow$
  $E[\log Y_1(0) - \log Y_0(0)|D = 1] = E[\log Y_1(0) - \log Y_0(0)|D = 0]$
  (unless D is randomly assigned: Roth and Sant'Anna (2020))

Complications

- Parallel trends may only hold conditional on X
  $E[Y_1(0) - Y_0(0)|X, D = 1] = E[Y_1(0) - Y_0(0)|X, D = 0]$
- Parallel trends assumption is NOT scale invariant
  $E[Y_1(0) - Y_0(0)|D = 1] = E[Y_1(0) - Y_0(0)|D = 0] \;\not\Rightarrow$
  $E[\log Y_1(0) - \log Y_0(0)|D = 1] = E[\log Y_1(0) - \log Y_0(0)|D = 0]$
  (unless D is randomly assigned: Roth and Sant'Anna (2020))
- Effects may be heterogenous
- Units may be treated in different times

Differential timing

$$Y_{it} = \delta D_{it} + \gamma X_{it} + \alpha_{i\cdot} + \alpha_{\cdot t} + \varepsilon_{it}$$

Differential timing

$$Y_{it} = \delta D_{it} + \gamma X_{it} + \alpha_{i\cdot} + \alpha_{\cdot t} + \varepsilon_{it}$$

Differential timing with state level (or any group) treatments:

$$Y_{ist} = \delta D_{st} + \gamma X_{ist} + \alpha_{s\cdot} + \alpha_{\cdot t} + \varepsilon_{ist}$$

Differential timing

$$Y_{it} = \delta D_{it} + \gamma X_{it} + \alpha_{i\cdot} + \alpha_{\cdot t} + \varepsilon_{it}$$

Differential timing with state level (or any group) treatments:

$$Y_{ist} = \delta D_{st} + \gamma X_{ist} + \alpha_{s\cdot} + \alpha_{\cdot t} + \varepsilon_{ist}$$

Aggregated version: this will lead to the same estimate δ but with higher standard errors:

$$Y_{st} = \delta D_{st} + \gamma X_{st} + \alpha_{s\cdot} + \alpha_{\cdot t} + \varepsilon_{ist}$$

- $D_{it} = 1$ if the unit i is treated at time t
- $D_{st} = 1$ if the state s is treated at time t
- $\alpha_{i\cdot}$ - constant for unit i
- $\alpha_{s\cdot}$ - constant for state s
- $\alpha_{\cdot t}$ - constant for time t
- X_{it}, X_{ist} - covariates - (beware of colliders!!)

Statistical inference?

Estimate $\hat{\delta}$ via OLS.

- BUT: Observations are likely serially correlated across states (groups) and thus standard errors may be too optimistic (small).
- Panels are long.
- Often very little variation in D_{st}

Statistical inference?

Estimate $\hat{\delta}$ via OLS.

- BUT: Observations are likely serially correlated across states (groups) and thus standard errors may be too optimistic (small).
- Panels are long.
- Often very little variation in D_{st}
- Simulations in Bertrand et al. (2004) show you can reject correct null in 45% cases! (instead of 5%)

Statistical inference?

Estimate $\hat{\delta}$ via OLS.

- BUT: Observations are likely serially correlated across states (groups) and thus standard errors may be too optimistic (small).
- Panels are long.
- Often very little variation in D_{st}
- Simulations in Bertrand et al. (2004) show you can reject correct null in 45% cases! (instead of 5%)

How to fix this?

Statistical inference?

Estimate $\hat{\delta}$ via OLS.

- BUT: Observations are likely serially correlated across states (groups) and thus standard errors may be too optimistic (small).
- Panels are long.
- Often very little variation in D_{st}
- Simulations in Bertrand et al. (2004) show you can reject correct null in 45% cases! (instead of 5%)

How to fix this?

- Block bootstrap. (Sample states with replacement)

Statistical inference?

Estimate $\hat{\delta}$ via OLS.

- BUT: Observations are likely serially correlated across states (groups) and thus standard errors may be too optimistic (small).
- Panels are long.
- Often very little variation in D_{st}
- Simulations in Bertrand et al. (2004) show you can reject correct null in 45% cases! (instead of 5%)

How to fix this?

- Block bootstrap. (Sample states with replacement)
- Ignore the time dimension altogether. (We're in 2x2 table)

Statistical inference?

Estimate $\hat{\delta}$ via OLS.

- BUT: Observations are likely serially correlated across states (groups) and thus standard errors may be too optimistic (small).
- Panels are long.
- Often very little variation in D_{st}
- Simulations in Bertrand et al. (2004) show you can reject correct null in 45% cases! (instead of 5%)

How to fix this?

- Block bootstrap. (Sample states with replacement)
- Ignore the time dimension altogether. (We're in 2x2 table)
- Cluster standard errors (at the level of groups or individuals) - we may allow arbitrary correlation between outcomes within a certain state (or individual) over time.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. "How much should we trust differences-in-differences estimates?." The Quarterly journal of economics 119.1 (2004): 249-275.

Pre-treatment trends? Event study

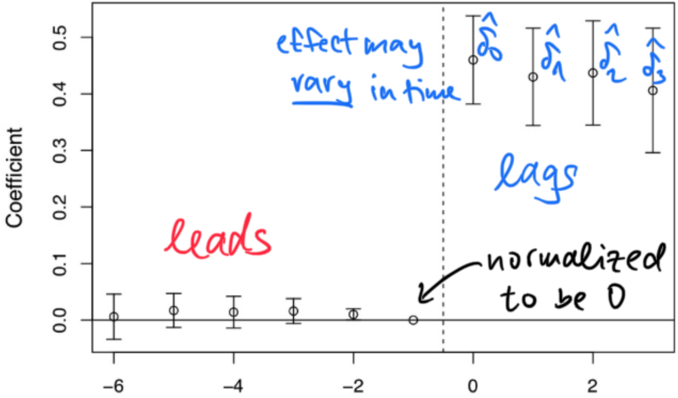
$$Y_{it} = \sum_{\tau=-q}^{-2} \underbrace{\delta_\tau D_{it}^\tau}_{\text{leads}} + \sum_{\tau=0}^{m} \underbrace{\delta_\tau D_{it}^\tau}_{\text{lags}} + \gamma X_{it} + \alpha_{i.} + \alpha_{.t} + \varepsilon_{it}$$

D_{it}^τ is an indicator for unit i being τ periods away from the initial treatment at time t

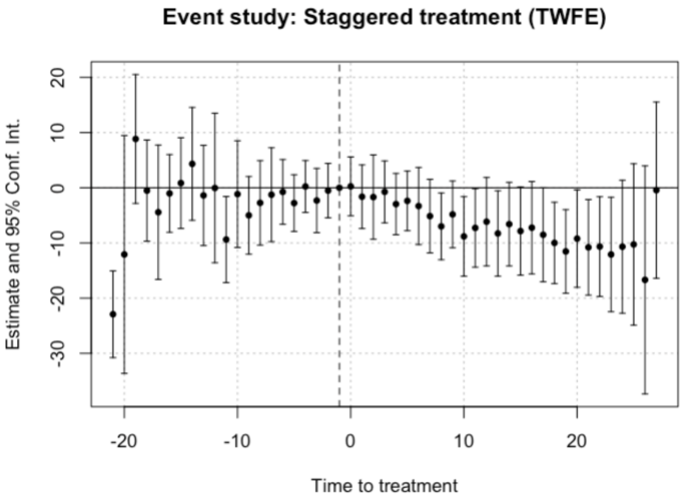
If state i adopted a new policy in $t = 2000$, then
$D_{i,1999}^{-1} = D_{i,2000}^{0} = D_{i,2001}^{1} = \ldots = 1$ and e.g.
$D_{i,1999}^{-2} = D_{i,1999}^{0} = D_{i,1999}^{1} = D_{i,1999}^{2} = 0$.

Pre-treatment trends? Event study

$$Y_{it} = \underbrace{\sum_{\tau=-q}^{-2} \delta_\tau D_{it}^\tau}_{\text{leads}} + \underbrace{\sum_{\tau=0}^{m} \delta_\tau D_{it}^\tau}_{\text{lags}} + \gamma X_{it} + \alpha_{i.} + \alpha_{.t} + \varepsilon_{it}$$



Pre-treatment trends? Event study



(The previous figure was too beautiful, normally it looks more like this one.)

Placebo tests

There is a lot of room for creativity

- choose workers unaffected by the minimum wage

Placebo tests

There is a lot of room for creativity

- choose workers unaffected by the minimum wage
- change treatment date to a fake one

Placebo tests

There is a lot of room for creativity

- choose workers unaffected by the minimum wage
- change treatment date to a fake one
- choose a fake treatment group

Placebo tests

There is a lot of room for creativity

- choose workers unaffected by the minimum wage
- change treatment date to a fake one
- choose a fake treatment group
- change the outcome to the one that should plausibly be unaffected

Placebo tests

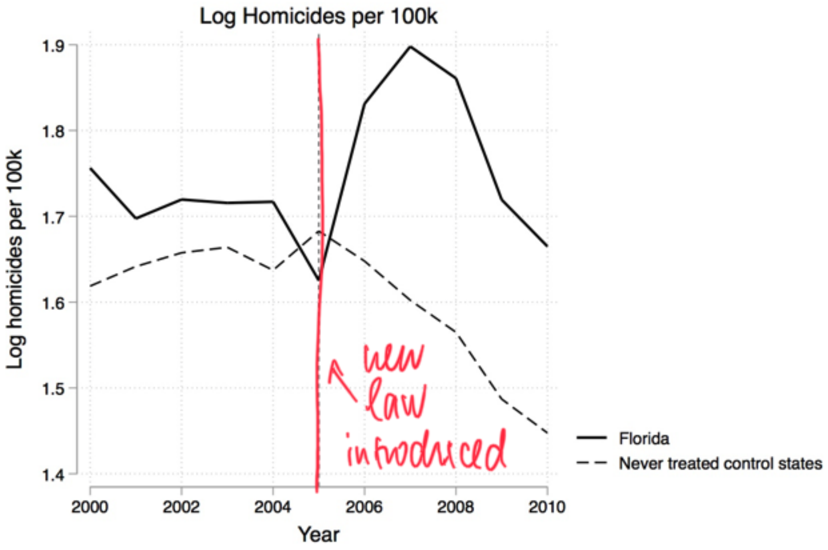
There is a lot of room for creativity

- choose workers unaffected by the minimum wage
- change treatment date to a fake one
- choose a fake treatment group
- change the outcome to the one that should plausibly be unaffected
- look at different subgroups - use your domain knowledge

Empirical Application - Cheng and Hoekstra (2013)

- had gun reform had impact on violance?
- different states adopted the law in different times
- ChH provide evidence that it is not associated with other types of crimes (e.g. cars theft)
- The new law was associated with an increase 8-10% in homicides

Source: Chapter 9.6.6 in `https://mixtape.scunning.com/difference-in-differences.html`



Source: Chapter 9.6.6 in `https://mixtape.scunning.com/difference-in-differences.html` very robust

Panel A. Homicide	Log(Homicide Rates)					
OLS–Weights	1	2	3	4	5	6
Castle Doctrine Law	0.0801**	0.0946***	0.0937**	0.0955**	0.0985***	0.100**
	(0.0342)	(0.0279)	(0.0290)	(0.0367)	(0.0299)	(0.0388)
0 to 2 years before adoption of castle doctrine law					0.00398	
					(0.0222)	
Observation	550	550	550	550	550	550
State and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Region-by-year Fixed		Yes	Yes	Yes	Yes	Yes
Effects						
Time-Varying Controls			Yes	Yes	Yes	Yes
Controls for Larceny or Motor Theft						Yes
State-specific Linear Time Trends						Yes

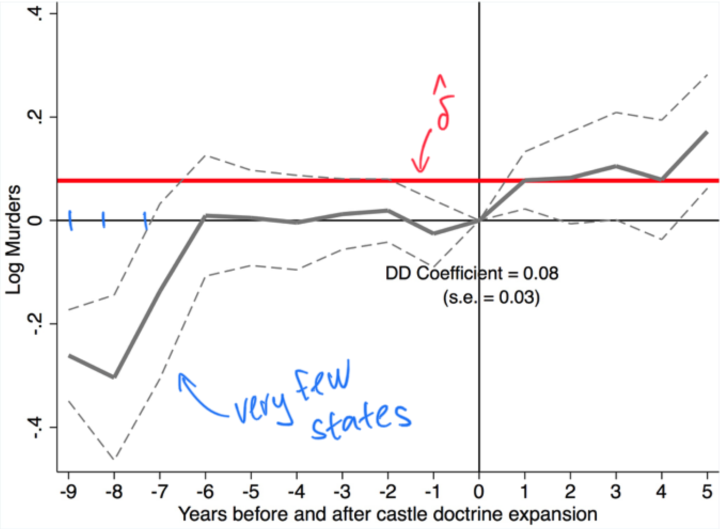
Method	Average estimate	Estimates larger than actual estimate
Weighted OLS	−0.003	0/40
Unweighted OLS	0.001	1/40
Negative binomial	0.001	0/40

no effect

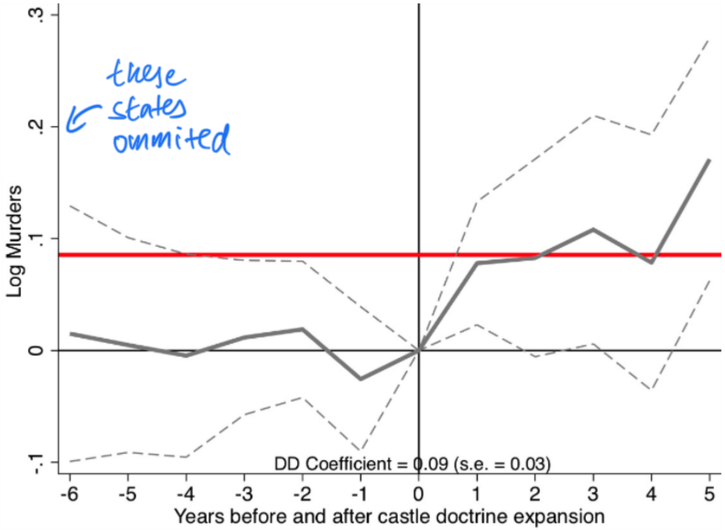
- using 20 different placebo dates
- the average estimates essentially zero

Source: Chapter 9.6.6 in

`https://mixtape.scunning.com/difference-in-differences.html`



Chapter 9.6.6 in `https://mixtape.scunning.com/difference-in-differences.html`



Chapter 9.6.6 in `https://mixtape.scunning.com/difference-in-differences.html`

DiD with covariates based on IPW

- Parallel trends cond. on X:
  $E[Y_1(0) - Y_0(0)|X, D = 1] = E[Y_1(0) - Y_0(0)|X, D = 0]$

DiD with covariates based on IPW

- Parallel trends cond. on X:
  $E[Y_1(0) - Y_0(0)|X, D = 1] = E[Y_1(0) - Y_0(0)|X, D = 0]$
- No effect of D on X: $X(1) = X(0) = X$

DiD with covariates based on IPW

- Parallel trends cond. on X:
  $E[Y_1(0) - Y_0(0)|X, D = 1] = E[Y_1(0) - Y_0(0)|X, D = 0]$
- No effect of D on X: $X(1) = X(0) = X$
- No pretreatment effect: $E[Y_0(1)|D = 1] - E[Y_0(0)|D = 1] = 0$

DiD with covariates based on IPW

- Parallel trends cond. on X:
  $E[Y_1(0) - Y_0(0)|X, D = 1] = E[Y_1(0) - Y_0(0)|X, D = 0]$
- No effect of D on X: $X(1) = X(0) = X$
- No pretreatment effect: $E[Y_0(1)|D = 1] - E[Y_0(0)|D = 1] = 0$
- Common support: $P(D = 1, T = 1|X, (D, T) \in \{(d, t), (1, 1)\}) < 1$ for all $(d, t) \in \{(1, 0), (0, 1), (0, 0)\}$

DiD with covariates based on IPW

- Parallel trends cond. on X:
  $E[Y_1(0) - Y_0(0)|X, D = 1] = E[Y_1(0) - Y_0(0)|X, D = 0]$
- No effect of D on X: $X(1) = X(0) = X$
- No pretreatment effect: $E[Y_0(1)|D = 1] - E[Y_0(0)|D = 1] = 0$
- Common support: $P(D = 1, T = 1|X, (D, T) \in \{(d, t), (1, 1)\}) < 1$ for all $(d, t) \in \{(1, 0), (0, 1), (0, 0)\}$

$$ATT = E\left[Y \cdot \left\{\frac{D \cdot T}{\Pi} - \frac{D \cdot (1 - T) \cdot \rho_{1,1}(X)}{\rho_{1,0}(X) \cdot \Pi} - \left(\frac{(1 - D) \cdot T \cdot \rho_{1,1}(X)}{\rho_{0,1}(X) \cdot \Pi} - \frac{(1 - D) \cdot T \cdot \rho_{1,1}(X)}{\rho_{0,0}(X) \cdot \Pi}\right)\right\}\right]$$

where $\Pi = P(D = 1, T = 1)$ and $\rho_{d,t}(X) = p(D = d, T = t|X)$

Lechner, Michael. "The Estimation of Causal Effects by Difference-in-Difference Methods." Foundations and Trends (R) in Econometrics 4.3 (2011): 165-224.

Two-way fixed effects model (TWFE)

$$Y_{it} = \delta D_{it} + \gamma X_{it} + \alpha_{i.} + \alpha_{.t} + \varepsilon_{it}$$

Two-way fixed effects model (TWFE)

$$Y_{it} = \delta D_{it} + \gamma X_{it} + \alpha_{i\cdot} + \alpha_{\cdot t} + \varepsilon_{it}$$

it looks reasonable: we extend the basic 2x2 setup into multiple time-periods, covariates and differential timing. Units can be treated at

different time-periods. We even plugged in dummies for greater flexibility (but hey, more is better, right?).

But, after all, what is this δ ?

Goodman-Bacon (2021) decomposition

We estimate $Y_{it} = \delta D_{it} + \alpha_{i\cdot} + \alpha_{\cdot t} + \varepsilon_{it}$ to get $\hat{\delta}$

Staggered rollout setup. Once treated, then treated forever.
$D_{it} = 1 \implies D_{it+1} = 1$

Goodman-Bacon (2021) decomposition

We estimate $Y_{it} = \delta D_{it} + \alpha_{i.} + \alpha_{.t} + \varepsilon_{it}$ to get $\hat{\delta}$

Staggered rollout setup. Once treated, then treated forever.
$D_{it} = 1 \implies D_{it+1} = 1$

Goodman-Bacon (2021) shows this $\hat{\delta}$ is a weighted average of different $\hat{\delta}^{2x2}$. These are based on different 2x2 comparisons! Just like the Card and Krueger (1994).

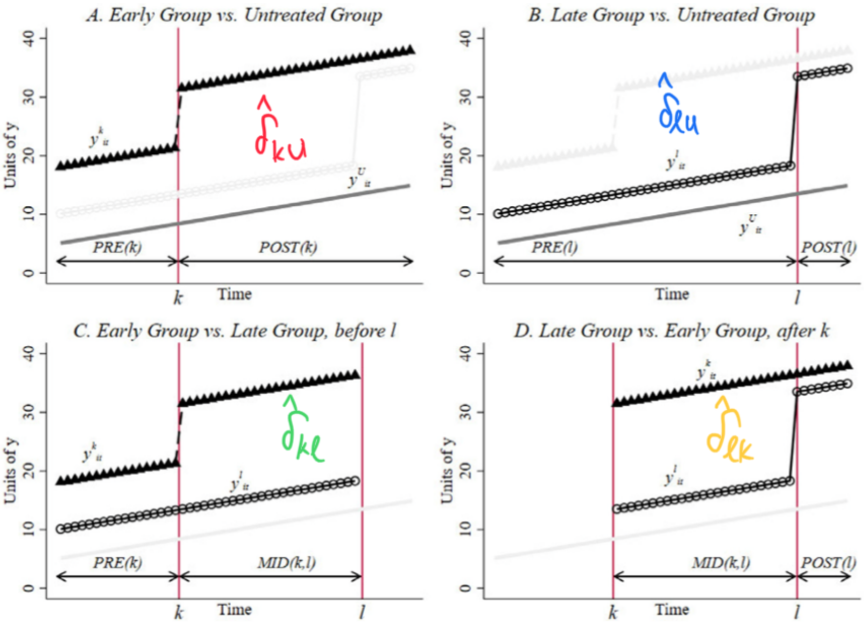
Goodman-Bacon (2021) decomposition

We estimate $Y_{it} = \delta D_{it} + \alpha_{i.} + \alpha_{.t} + \varepsilon_{it}$ to get $\hat{\delta}$

Staggered rollout setup. Once treated, then treated forever.
$D_{it} = 1 \implies D_{it+1} = 1$

Goodman-Bacon (2021) shows this $\hat{\delta}$ is a weighted average of different $\hat{\delta}^{2x2}$. These are based on different 2x2 comparisons! Just like the Card and Krueger (1994).

This is great, because we understand what $\hat{\delta}^{2x2}$ from canonical 2x2 setup means!


A. Early Group vs. Untreated Group
B. Late Group vs. Untreated Group
C. Early Group vs. Late Group, before l
D. Late Group vs. Early Group, after k
Units of y
Time
PRE(k)
POST(k)
PRE(l)
POST(l)
MID(k,l)
k
l
0
10
20
30
40

There are 3 groups: k - early adopters, l - late adopters, U - untreated

There are 3 groups: k - early adopters, l - late adopters, U - untreated

$$\hat{\delta} = w_{kU}\hat{\delta}^{2x2}_{kU} + w_{lU}\hat{\delta}^{2x2}_{lU} + w_{kl}\hat{\delta}^{2x2}_{kl} + w_{lk}\hat{\delta}^{2x2}_{lk}$$

- Weights depend on: (i) how large the groups are, (ii) how much variation there is in the treatments.

There are 3 groups: k - early adopters, l - late adopters, U - untreated

$$\hat{\delta} = w_{kU}\hat{\delta}^{2x2}_{kU} + w_{lU}\hat{\delta}^{2x2}_{lU} + w_{kl}\hat{\delta}^{2x2}_{kl} + w_{lk}\hat{\delta}^{2x2}_{lk}$$

- Weights depend on: (i) how large the groups are, (ii) how much variation there is in the treatments.
- Just like in OLS, large weights are given to groups with higher variation.

There are 3 groups: k - early adopters, l - late adopters, U - untreated

$$\hat{\delta} = w_{kU}\hat{\delta}_{kU}^{2x2} + w_{lU}\hat{\delta}_{lU}^{2x2} + w_{kl}\hat{\delta}_{kl}^{2x2} + w_{lk}\hat{\delta}_{lk}^{2x2}$$

- Weights depend on: (i) how large the groups are, (ii) how much variation there is in the treatments.
- Just like in OLS, large weights are given to groups with higher variation.
- This result is about **estimators** not estimands.

There are 3 groups: k - early adopters, l - late adopters, U - untreated

$$\hat{\delta} = w_{kU}\hat{\delta}_{kU}^{2x2} + w_{lU}\hat{\delta}_{lU}^{2x2} + w_{kl}\hat{\delta}_{kl}^{2x2} + w_{lk}\hat{\delta}_{lk}^{2x2}$$

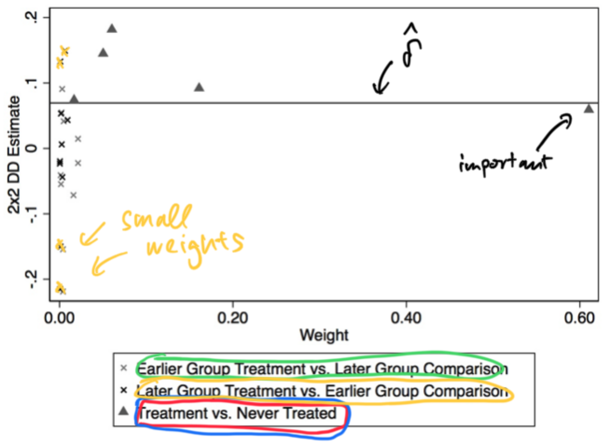
- Weights depend on: (i) how large the groups are, (ii) how much variation there is in the treatments.
- Just like in OLS, large weights are given to groups with higher variation.
- This result is about **estimators** not estimands.
- Adding/removing time periods changes the weights.

Diagnostics

Similar decomposition could be done if you have many different groups.

Diagnostics

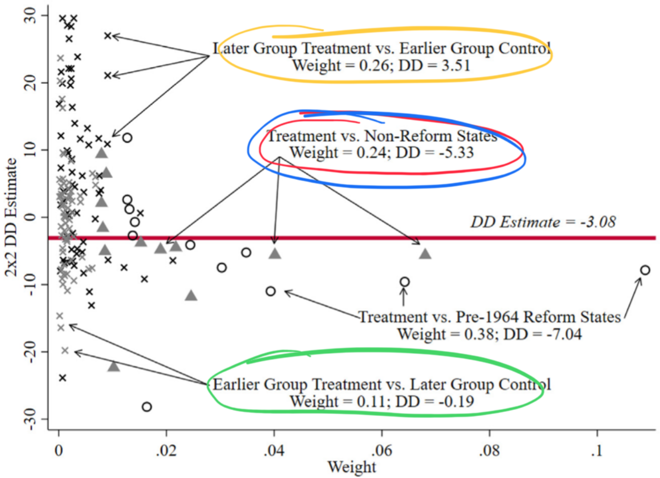
Similar decomposition could be done if you have many different groups.



Source: Cunningham's reconstruction of Cheng and Hoekstra (2013)

Diagnostics (different paper)

Additional control group here (circles).



Source: Goodman-Bacon (2021)

What is TWFE really?

Static specification (a single δ)

$$Y_{it} = \delta \sum_{\tau=0}^{m} D_{it}^{\tau} + \gamma X_{it} + \alpha_{i.} + \alpha_{.t} + \varepsilon_{it}$$

What is TWFE really?

Static specification (a single δ)

$$Y_{it} = \delta \sum_{\tau=0}^{m} D_{it}^{\tau} + \gamma X_{it} + \alpha_{i\cdot} + \alpha_{\cdot t} + \varepsilon_{it}$$

D_{it}^{τ} is an indicator for unit i being τ periods away from the initial treatment at time t

If state i adopted a new policy in $t = 2000$, then
$D_{i,1999}^{-1} = D_{i,2000}^{0} = D_{i,2001}^{1} = \ldots = 1$

What is TWFE really?

Dynamic specification (multiple δ_τ-s)

$$Y_{it} = \sum_{\tau=-q}^{-2} \delta_\tau D_{it}^{\tau} + \sum_{\tau=0}^{m} \delta_\tau D_{it}^{\tau} + \gamma X_{it} + \alpha_{i.} + \alpha_{.t} + \varepsilon_{it}$$

What is TWFE really?

Dynamic specification (multiple δ_τ-s)

$$Y_{it} = \sum_{\tau=-q}^{-2} \delta_\tau D_{it}^{\tau} + \sum_{\tau=0}^{m} \delta_\tau D_{it}^{\tau} + \gamma X_{it} + \alpha_{i.} + \alpha_{.t} + \varepsilon_{it}$$

Yes, we run some regressions. But what do we actually get? How do we interpret these $\hat{\delta}$ or $\hat{\delta}_\tau$?

Sun and Abraham (2021)

Consider e.g.

$$Y_{it} = \sum_{\tau=-q}^{-2} \delta_\tau D_{it}^{\tau} + \sum_{\tau=0}^{m} \delta_\tau D_{it}^{\tau} + \alpha_{i\cdot} + \alpha_{\cdot t} + \varepsilon_{it}$$

Sun and Abraham (2021)

Consider e.g.

$$Y_{it} = \sum_{\tau=-q}^{-2} \delta_\tau D_{it}^{\tau} + \sum_{\tau=0}^{m} \delta_\tau D_{it}^{\tau} + \alpha_{i\cdot} + \alpha_{\cdot t} + \varepsilon_{it}$$

- Common practice is to use leads to test for a pre-trend differences.

Sun and Abraham (2021)

Consider e.g.

$$Y_{it} = \sum_{\tau=-q}^{-2} \delta_\tau D_{it}^{\tau} + \sum_{\tau=0}^{m} \delta_\tau D_{it}^{\tau} + \alpha_{i\cdot} + \alpha_{\cdot t} + \varepsilon_{it}$$

- Common practice is to use leads to test for a pre-trend differences.
- But these coefficients are contaminated by both the pre-trends and heterogeneity

Sun and Abraham (2021)

Consider e.g.

$$Y_{it} = \sum_{\tau=-q}^{-2} \delta_\tau D_{it}^{\tau} + \sum_{\tau=0}^{m} \delta_\tau D_{it}^{\tau} + \alpha_{i\cdot} + \alpha_{\cdot t} + \varepsilon_{it}$$

- Common practice is to use leads to test for a pre-trend differences.
- But these coefficients are contaminated by both the pre-trends and heterogeneity
- They propose a way how to examine how much of a problem this is

Sun and Abraham (2021)

Consider e.g.

$$Y_{it} = \sum_{\tau=-q}^{-2} \delta_\tau D_{it}^{\tau} + \sum_{\tau=0}^{m} \delta_\tau D_{it}^{\tau} + \alpha_{i\cdot} + \alpha_{\cdot t} + \varepsilon_{it}$$

- Common practice is to use leads to test for a pre-trend differences.
- But these coefficients are contaminated by both the pre-trends and heterogeneity
- They propose a way how to examine how much of a problem this is
- They also propose an estimator that uses **never-treated** as a comparison group

Callaway and Sant'Anna (2021)

Staggered treatment adoption setup. $D_{it} = 1 \implies D_{it+1} = 1$

Callaway and Sant'Anna (2021)

Staggered treatment adoption setup. $D_{it} = 1 \implies D_{it+1} = 1$

Decompose everything into "lego" pieces:

$$ATT(g,t) = E[Y_t(g) - Y_t(0) | G_g = 1]$$

ATT in time t for group treated in time g. ($G_g = 1$)

Callaway and Sant'Anna (2021)

Staggered treatment adoption setup. $D_{it} = 1 \implies D_{it+1} = 1$

Decompose everything into "lego" pieces:

$$ATT(g,t) = E[Y_t(g) - Y_t(0) | G_g = 1]$$

ATT in time t for group treated in time g. ($G_g = 1$)

They make

- Limited treatment anticipation assumption

Callaway and Sant'Anna (2021)

Staggered treatment adoption setup. $D_{it} = 1 \implies D_{it+1} = 1$

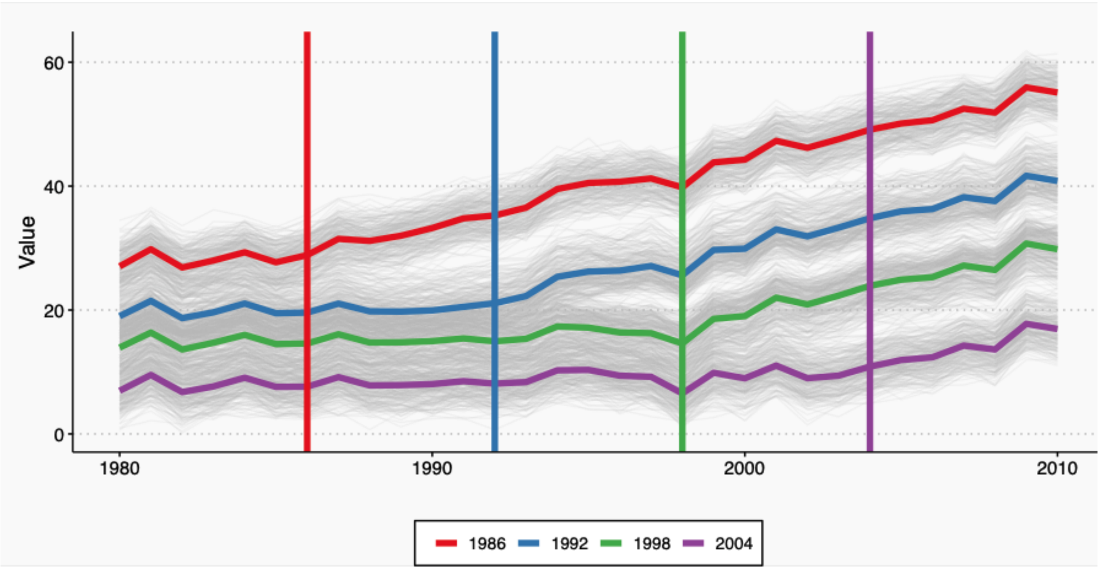
Decompose everything into "lego" pieces:

$$ATT(g,t) = E[Y_t(g) - Y_t(0)|G_g = 1]$$

ATT in time t for group treated in time g. ($G_g = 1$)

They make

- Limited treatment anticipation assumption
- Different Conditional parallel trend assumptions
  - Comparing to never-treated individuals
  - Comparing to not-yet-treated individuals



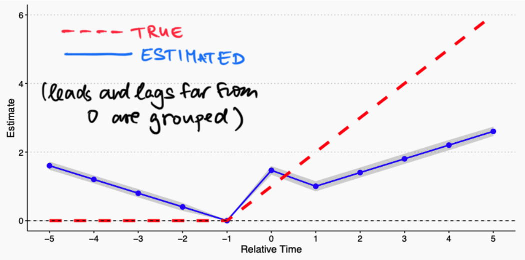
Source: `https://pedrohcgs.github.io/files/Callaway_SantAnna_2020_slides.pdf`

Estimate via OLS?

$$Y_{it} = \sum_{\tau=-q}^{-2} \delta_\tau D_{it}^{\tau} + \sum_{\tau=0}^{m} \delta_\tau D_{it}^{\tau} + \gamma X_{ist} + \alpha_{i\cdot} + \alpha_{\cdot t} + \varepsilon_{it}$$

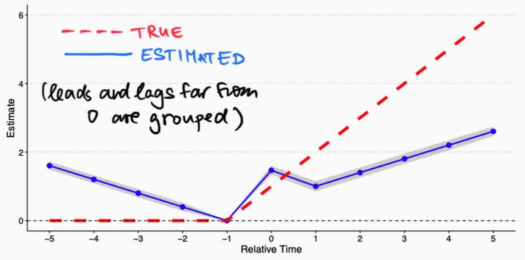
Estimate via OLS?

$$Y_{it} = \sum_{\tau=-q}^{-2} \delta_\tau D_{it}^\tau + \sum_{\tau=0}^{m} \delta_\tau D_{it}^\tau + \gamma X_{ist} + \alpha_{i.} + \alpha_{.t} + \varepsilon_{it}$$



Estimate via OLS?

$$Y_{it} = \sum_{\tau=-q}^{-2} \delta_\tau D_{it}^\tau + \sum_{\tau=0}^{m} \delta_\tau D_{it}^\tau + \gamma X_{ist} + \alpha_{i\cdot} + \alpha_{\cdot t} + \varepsilon_{it}$$



Source: `https://pedrohcgs.github.io/files/Callaway_SantAnna_2020_slides.pdf`

E.g. based on comparing to never-treated individuals (denoted as $C = 1$), they get:

$$ATT(g,t) = E\left[\left(\frac{G_g}{E[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{E\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)(Y_t - Y_{g-1})\right]$$

E.g. based on comparing to never-treated individuals (denoted as $C = 1$), they get:

$$ATT(g,t) = E\left[\left(\frac{G_g}{E[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{E\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)(Y_t - Y_{g-1})\right]$$

- $p_g(X) =$ is a propensity score

E.g. based on comparing to never-treated individuals (denoted as $C = 1$), they get:

$$ATT(g,t) = E\left[\left(\frac{G_g}{E[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{E\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)(Y_t - Y_{g-1})\right]$$

- $p_g(X) =$ is a propensity score
- Comparing to never-treated individuals
- Never-treated are re-weighted to match those treated in time g (IPW style)

E.g. based on comparing to never-treated individuals (denoted as $C = 1$), they get:

$$ATT(g,t) = E\left[\left(\frac{G_g}{E[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{E\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)(Y_t - Y_{g-1})\right]$$

- $p_g(X) =$ is a propensity score
- Comparing to never-treated individuals
- Never-treated are re-weighted to match those treated in time g (IPW style)
- They have a doubly robust version of this expression.

E.g. based on comparing to never-treated individuals (denoted as $C = 1$), they get:

$$ATT(g,t) = E\left[\left(\frac{G_g}{E[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{E\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)(Y_t - Y_{g-1})\right]$$

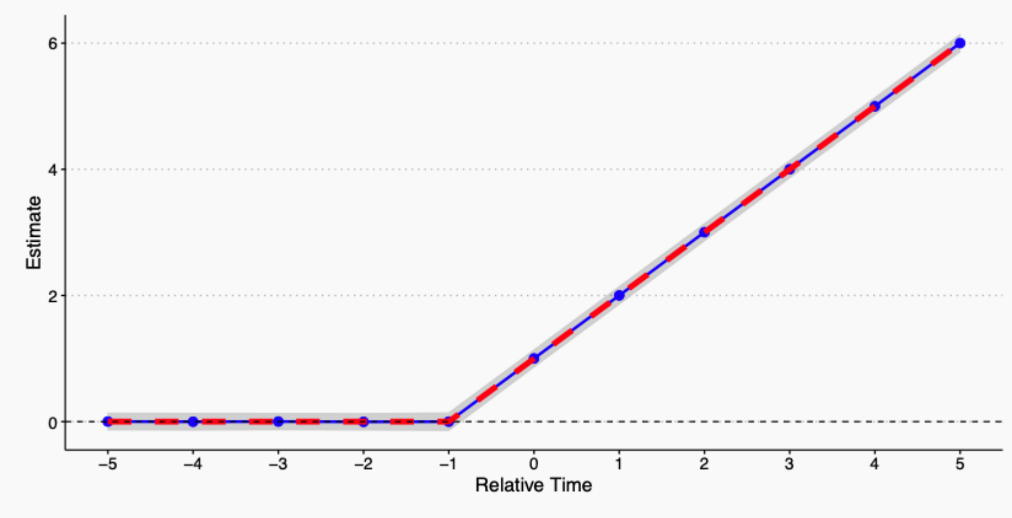
- $p_g(X) =$ is a propensity score
- Comparing to never-treated individuals
- Never-treated are re-weighted to match those treated in time g (IPW style)
- They have a doubly robust version of this expression.
- Different $ATT(g,t)$ are weighted into forming different parameters of interest

E.g. based on comparing to never-treated individuals (denoted as $C = 1$), they get:

$$ATT(g,t) = E\left[\left(\frac{G_g}{E[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{E\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)(Y_t - Y_{g-1})\right]$$

- $p_g(X) =$ is a propensity score
- Comparing to never-treated individuals
- Never-treated are re-weighted to match those treated in time g (IPW style)
- They have a doubly robust version of this expression.
- Different $ATT(g,t)$ are weighted into forming different parameters of interest
- `did` and `DRDID` packages

Much nicer with their method



Source: `https://pedrohcgs.github.io/files/Callaway_SantAnna_2020_slides.pdf`

de Chaisemartin and d'Haultfoeuille (2020)

Consider the following object of interest

$$ATT(g,t) = E[Y_t(g) - Y_t(0) | G_g = 1]$$

de Chaisemartin and d'Haultfoeuille (2020)

Consider the following object of interest

$$ATT(g,t) = E[Y_t(g) - Y_t(0) | G_g = 1]$$

Let δ be TWFE estimand from this regression

$$Y_{it} = \delta D_{it} + \alpha_{i.} + \alpha_{.t} + \varepsilon_{it}$$

de Chaisemartin and d'Haultfoeuille (2020)

Consider the following object of interest

$$ATT(g,t) = E[Y_t(g) - Y_t(0)|G_g = 1]$$

Let δ be TWFE estimand from this regression

$$Y_{it} = \delta D_{it} + \alpha_{i.} + \alpha_{.t} + \varepsilon_{it}$$

Then

$$\delta = E\left[\sum_{i,t:D_{it}=1} \frac{1}{N_1} w_{it} \cdot ATT(g,t)\right]$$

de Chaisemartin and d'Haultfoeuille (2020)

Consider the following object of interest

$$ATT(g,t) = E[Y_t(g) - Y_t(0) | G_g = 1]$$

Let δ be TWFE estimand from this regression

$$Y_{it} = \delta D_{it} + \alpha_{i.} + \alpha_{.t} + \varepsilon_{it}$$

Then

$$\delta = E\left[\sum_{i,t:D_{it}=1} \frac{1}{N_1} w_{it} \cdot ATT(g,t)\right]$$

- But the weights w_{it} can be negative(!)

de Chaisemartin and d'Haultfoeuille (2020)

Consider the following object of interest

$$ATT(g,t) = E[Y_t(g) - Y_t(0) | G_g = 1]$$

Let δ be TWFE estimand from this regression

$$Y_{it} = \delta D_{it} + \alpha_{i.} + \alpha_{.t} + \varepsilon_{it}$$

Then

$$\delta = E\left[\sum_{i,t:D_{it}=1} \frac{1}{N_1} w_{it} \cdot ATT(g,t)\right]$$

- But the weights w_{it} can be negative(!)
- So $\delta \neq ATT$. What is the δ then?

de Chaisemartin and d'Haultfoeuille (2020)

Consider the following object of interest

$$ATT(g,t) = E[Y_t(g) - Y_t(0) | G_g = 1]$$

Let δ be TWFE estimand from this regression

$$Y_{it} = \delta D_{it} + \alpha_{i.} + \alpha_{.t} + \varepsilon_{it}$$

Then

$$\delta = E\left[\sum_{i,t:D_{it}=1} \frac{1}{N_1} w_{it} \cdot ATT(g,t)\right]$$

- But the weights w_{it} can be negative(!)
- So $\delta \neq ATT$. What is the δ then?
- It depends on the assumptions you impose (have a look at dCh & d'H (2020)

Two very recent reviews!

The status quo has been changed.

New papers emerging very rapidly.

- de Chaisemartin and D'Haultfœuille - Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey (**Dec 15 2021**)
- Roth, Sant'Anna, Bilinski and Poe - What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature (**Jan 3 2022**)

Concluding remarks

- The stream of new papers show rather depressing set of results.

Concluding remarks

- The stream of new papers show rather depressing set of results.
- Note that this is relevant only if there is differential treatment timing

Concluding remarks

- The stream of new papers show rather depressing set of results.
- Note that this is relevant only if there is differential treatment timing
- TWFE is not what we would like it to be and all these papers show various degrees of hopelessness.

Concluding remarks

- The stream of new papers show rather depressing set of results.
- Note that this is relevant only if there is differential treatment timing
- TWFE is not what we would like it to be and all these papers show various degrees of hopelessness.
- But

Concluding remarks

- The stream of new papers show rather depressing set of results.
- Note that this is relevant only if there is differential treatment timing
- TWFE is not what we would like it to be and all these papers show various degrees of hopelessness.
- But
- They also provide alternative estimators and implementations in R/STATA

Concluding remarks

What are the important questions we should ask?

- Who to compare with whom?
- What is the the object of interest?
- What kind of parallel trends assumptions will we impose?

Thank you for your attention!

References

- Chapter on Dif-in-dif in Cunningham's book is long, but fun nevertheless. I found the notation somewhat inconsistent. Cunningham, Scott. Causal Inference. Yale University Press, 2021. Free here: https://mixtape.scunning.com/difference-in-differences.html
- Introductory video on 2x2 DiD identification etc: Brady Neal, Causal Inference course https://www.youtube.com/watch?v=2nDgrNP7XSE
- Chapter 18 in Bruce Hansen's Econometrics book is a good start.
- Inference problems with DiD: Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. "How much should we trust differences-in-differences estimates?." The Quarterly journal of economics 119.1 (2004): 249-275.
- Parallel trends and functional forms: Roth, Jonathan, and Pedro HC Sant'Anna. "When Is Parallel Trends Sensitive to Functional Form?." arXiv preprint arXiv:2010.04814 (2020)
- DiD with covariates based on IPW: Lechner, Michael. "The Estimation of Causal Effects by Difference-in-Difference Methods." Foundations and Trends (R) in Econometrics 4.3 (2011): 165-224.
- Cheng, Cheng, and Mark Hoekstra. 2013. "Does Strengthening Self-Defense Law Deter Crime or Escalate Violence? Evidence from Expansions to Castle Doctrine." Journal of Human Resources 48 (3): 821–54.
- Recent advances: Taylor Wright's DiD reading group: https://taylorjwright.github.io/did-reading-group/ This is the best source. Videos of presentations by the authors of some of the most important recent contributions in the DiD literature.
- Goodman-Bacon, Andrew. "Difference-in-differences with variation in treatment timing." Journal of Econometrics (2021).
- Sun, Liyang, and Sarah Abraham. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." Journal of Econometrics 225.2 (2021): 175-199.
- Callaway, Brantly, and Pedro HC Sant'Anna. "Difference-in-differences with multiple time periods." Journal of Econometrics 225.2 (2021): 200-230.
- De Chaisemartin, Clément, and Xavier d'Haultfoeuille. "Two-way fixed effects estimators with heterogeneous treatment effects." American Economic Review 110.9 (2020): 2964-96.
- de Chaisemartin, Clément, and Xavier D'Haultfœuille. "Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey." Available at SSRN (2021).
- Jonathan Roth, Pedro H. C. Sant'Anna, Alyssa Bilinski and John Poe - What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature