The Role of Parallel Trends in Event Study Settings: An Application to Environmental Economics

Michelle Marcus
Vanderbilt University

Pedro H. C. Sant'Anna
Vanderbilt University

Journal of the Association of Environmental and Resource Economists, 2021

Difference-in-Differences

- Difference-in-Differences (DiD) is one of the most popular designs for causal inference.
- Canonical format:
  - 2 groups: $G = 0$ and $G = 1$;
  - 2 time periods: $t = 1$ and $t = 2$.
- Parameter of interest:

$$ATT \equiv \mathbb{E}\left[Y_{i,2}\left(1\right)|G_i = 1\right] - \mathbb{E}\left[Y_{i,2}\left(0\right)|G_i = 1\right].$$

DiD and event-study in practice

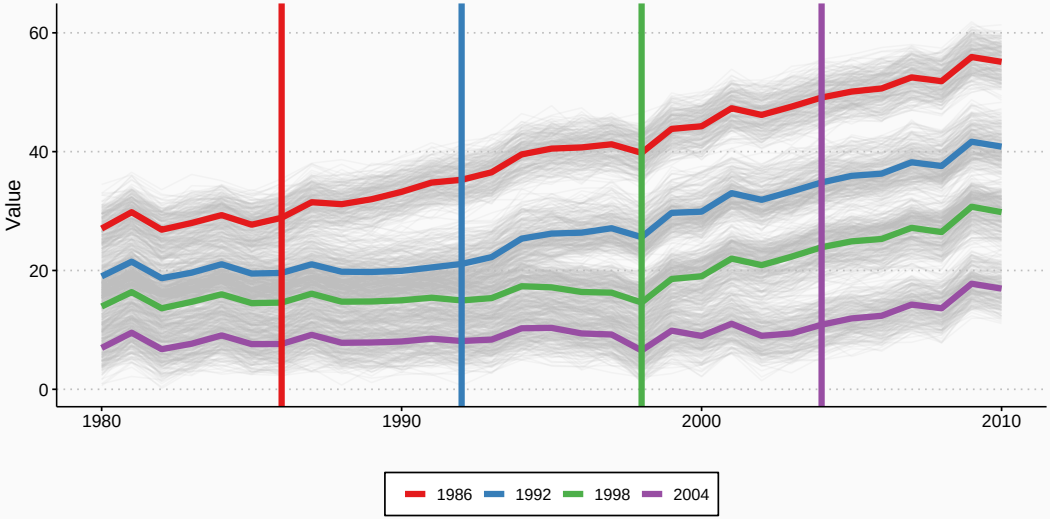
- Many DiD applications, however, deviates from the classical 2x2 DiD setup:
    1. Multiple time periods,
    2. Variation in treatment timing.
- Do these “generalizations” pose a practical problem?

Stylized example using simulated data

- 1000 units ($i = 1, 2, \ldots, 1000$) from 40 states ($state = 1, 2, \ldots, 40$).
- Data from 1980 to 2010 (31 years).
- 4 different groups based on treatment starting year: $g = 1986, 1992, 1998, 2004$.
- Randomly assign each state to a group.
- Outcome:

$$Y_{i,t} = \underbrace{(2010 - g)}_{\text{cohort-specific intercept}} + \underbrace{\alpha_i}_{N\left(\frac{state}{5}1,\right)} + \underbrace{\alpha_t}_{\frac{(t-g)}{10}+N(0,1)} + \underbrace{\tau_{i,t}}_{(t-g+1)\cdot 1\{t \geq g\}} + \underbrace{\varepsilon_{i,t}}_{N\left(0,\left(\frac{1}{2}\right)^2\right)}.$$

- ATT at the first treatment period is 1, at the second period since treatment is 2, etc.
- Evolution of the treatment effects is homogeneous across treatment groups.


Value
60
40
20
0
1980
1990
2000
2010
1986
1992
1998
2004

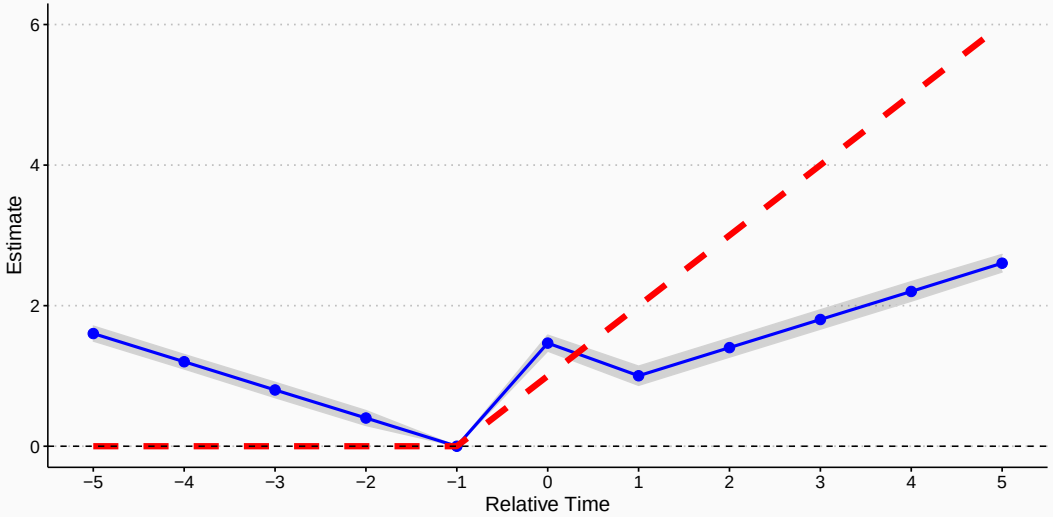
Traditional methods: TWFE event-study regression

- What if we tried to estimate the treatment effects using traditional TWFE event-study regressions?

$$Y_{i,t} = \alpha_i + \alpha_t + \gamma_k^{-K} D_{i,t}^{<-K} + \sum_{k=-K}^{-2} \gamma_k^{lead} D_{i,t}^{k} + \sum_{k=0}^{L} \gamma_k^{lags} D_{i,t}^{k} + \gamma_k^{L+} D_{i,t}^{>L} + \varepsilon_{i,t}.$$

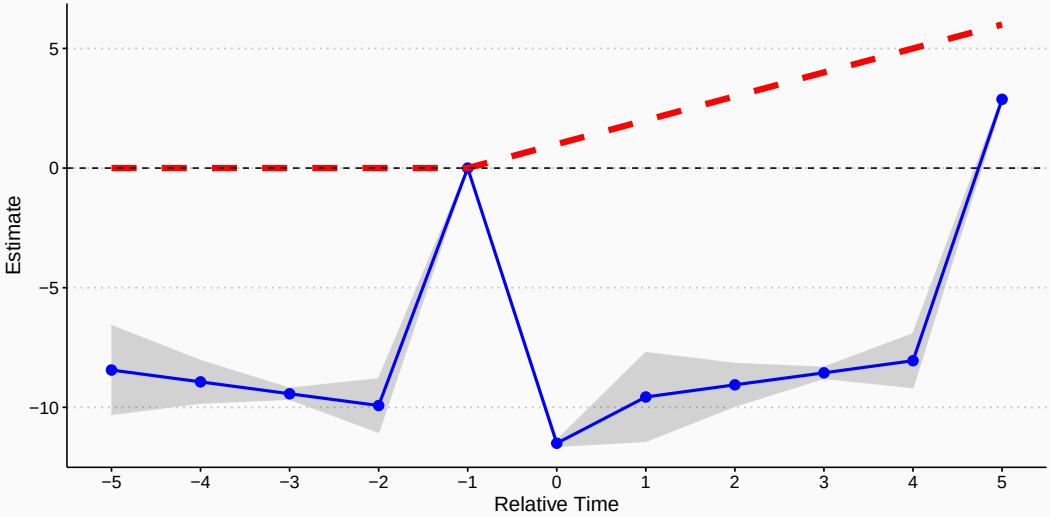
- Set K and L to be equal to 5.

- Simulate data and repeat 1,000 times to compute bias and simulation standard deviations.


6
4
2
0
Estimate
−5
−4
−3
−2
−1
0
1
2
3
4
5
Relative Time

Traditional methods: TWFE event-study regression

- What if we include all possible leads and lags in the TWFE event study specification, i.e., to set K and L to the maximum allowable in the data making inclusion of $D_{i,t}^{<-K}$ and of $D_{i,t}^{>L}$ unnecessary ?

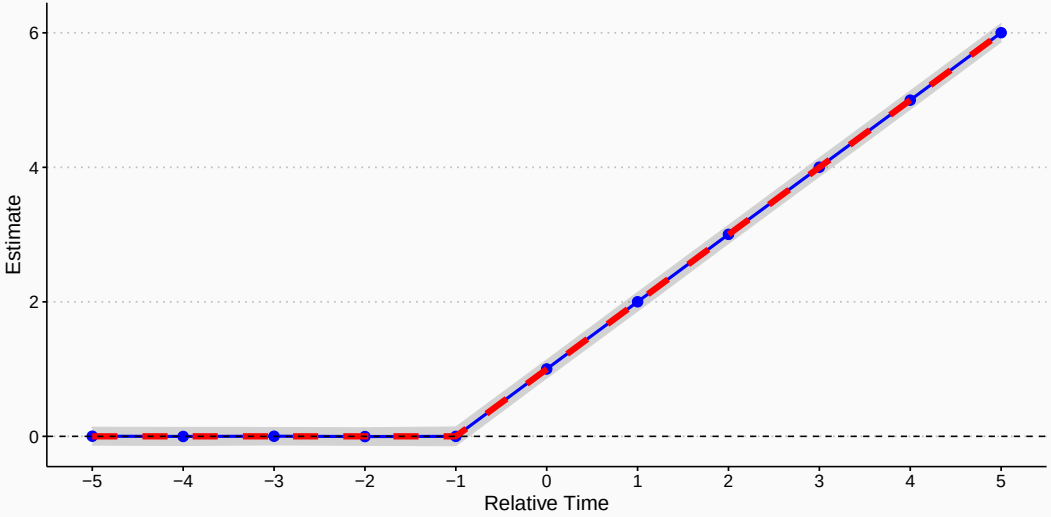

5
0
−5
−10
Estimate
−5
−4
−3
−2
−1
0
1
2
3
4
5
Relative Time


WARNING
PROCEED WITH CAUTION!
MORE ECONOMETRICS MAY BE NEEDED!

Recent literature

- In light of these problems, a recent econometrics literature is rising.
  - Athey and Imbens (2018)
  - Borusyak and Jaravel (2017)
  - Callaway and Sant'Anna (2020) (henceforth CS)
  - de Chaisemartin and D'Haultfouille (2020) (henceforth dCD)
  - Goodman-Bacon (2019)
  - Sun and Abraham (2020) (henceforth SA)
- In contrast to the 2x2 case, these papers rely on different parallel trends assumptions and/or recover different causal parameters.

Event-study plot using CS proposed estimator



Main goals of Marcus and Sant'Anna (2020)

- Highlight the different target causal parameters proposed by CS, dCD and SA.
- Highlight the different versions of parallel trends assumptions invoked by CS, dCD and SA.
- Not all parallel trends assumptions (PTAs) are made equal:
  - Some PTAs do restrict pre-treatment trends across groups whereas others do not;
  - Some PTAs are directly testable whereas others are not.
- Different PTAs suggest different estimators via the choice of the possible comparison groups.
  - Explicitly stating the invoked PTA adds transparency and objectivity, too!
- Propose new DiD and event study estimators that "better exploit" the restrictions imposed by the invoked PTA.

Application to Environmental Economics

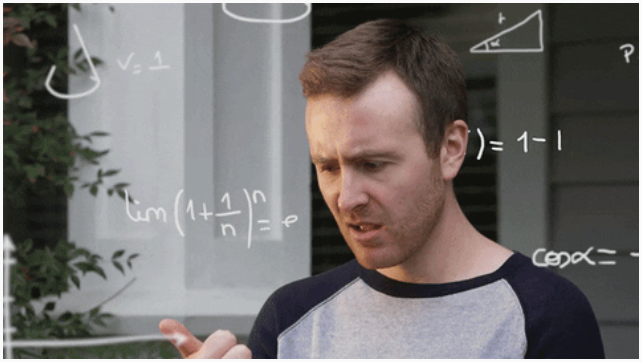
- Revisit Grooms (2015) and examine the effect of the transition from federal to state management of the Clean Water Act (CWA) on violation rates.
    - Transition from federal to state control has little to no effect on violation rates (robust across PTAs and parameters of interest).
    - Analyze whether states with a long prevalence of corruption see a large decrease in the violation rate after authorization relative to states without corruption and find that conclusion depends on the PTA.
    - If we allow for “corruption-specific” trends, we find no effect in violation rates, which is in sharp contrast to the TWFE specification adopted by Grooms (2015).

Main take-away message from this paper

Whenever possible, separate the analysis into two steps:

1. identification analysis, paying particular attention to the empirical content of the invoked PTA;

2. data analysis and estimation procedure.



Time for some notation

Framework

- Consider a random sample

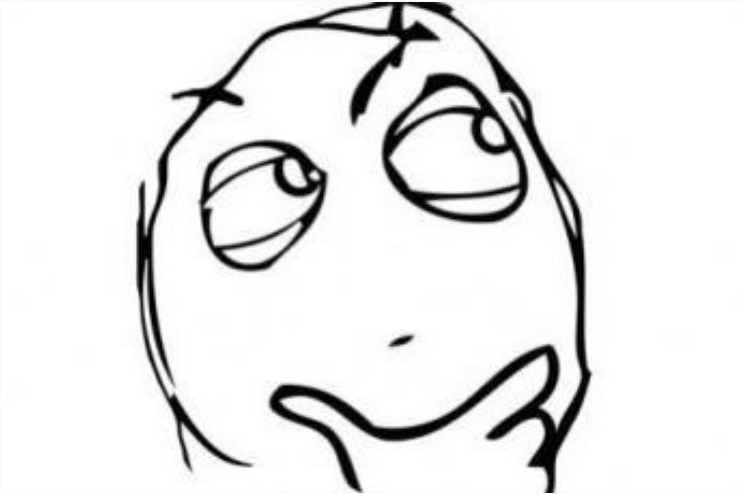
$$\{(Y_{i,1}, Y_{i,2}, \ldots, Y_{i,\mathcal{T}}, D_{i,1}, D_{i,2}, \ldots, D_{i,\mathcal{T}}, X_i)\}_{i=1}^{n}.$$

  where $D_{i,t}$ takes value 1 for individuals treated in period t and 0 otherwise.
- D_t are treatment dummies at time $t = 1, \ldots, \mathcal{T}$.
- $G_{i,g} = 1$ if individual i is first treated at time g, and zero otherwise ("Treatment start-time dummies").
- $C = 1$ is a "never-treated" comparison group.
- Staggered treatment design: $D_{i,t} = 1 \implies D_{i,t+1} = 1$, *for* $t = 1, 2, \ldots, \mathcal{T}$.
- No Anticipation: For all $t = 1, \ldots, \mathcal{T}$, $g = 2, \ldots, \mathcal{T}$ such that $t < g$, $\mathbb{E}[Y_{i,t} | G_g = 1] = \mathbb{E}[Y_{i,t}(0) | G_g = 1]$.
- Overlap: $P(G_1 = 1) = 0$ and, for some $\epsilon > 0$, and all $g = 2, \ldots, \mathcal{T}$, $P(G_g = 1) > \epsilon$.

Important building block

- We aim to express different causal parameters of interest as functionals of the "Group-time average treatment effects", i.e., the average treatment effect at time t, for those units first treated at time g:

$$\begin{aligned} ATT\left(g,t\right) &\equiv \mathbb{E}\left[\left.Y_{i,t}\left(1\right)\right|G_g=1\right]-\mathbb{E}\left[\left.Y_{i,t}\left(0\right)\right|G_g=1\right] \\ &= \alpha_{g,t}\left(1\right)-\alpha_{g,t}\left(0\right). \end{aligned} \tag{1}$$



What are the different PTAs?

Parallel trend assumptions

Assumption (dCD & SA "stronger PTA")
For all $t = 2, \ldots, \mathcal{T}$, and all $g = 2, \ldots, \mathcal{T}$,

$$\begin{aligned} \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | G_g = 1 \right] &= \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | C = 1 \right] \\ &= \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \right] \end{aligned}$$

Assumption (CS PTA based on "never treated")
For all $g, t = 2, \ldots, \mathcal{T}$, such that $t \geq g$,

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | G_g = 1 \right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | C = 1 \right].$$

Assumption (CS PTA based on "not-yet-treated")
For all $g, s, t = 2, \ldots, \mathcal{T}$, such that $t \geq g$, $s \geq t$,

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | G_g = 1 \right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | D_s = 0 \right].$$

Let's illustrate how these PTAs differ from each other

- Let's consider a simple, stylized example.
- Assume that we observe Y_{it} for a sample of units $i = 1, \ldots, n$ in four time periods, $t = 1, 2, 3, 4$.
- Some units are first treated at time 3 ($G_{i3} = 1$), others at time 4, ($G_{i4} = 1$), and the remaining units are not treated in the entire observation window $(C_i = 1)$.
- Once a unit i is treated at time g, it remains treated for all time periods $t \geq g$.
- Let $W = (Y_1, Y_2, Y_3, Y_4, G_3, G_4, C)'$, and assume that we observe a random sample $\{W_i\}_{i=1}^n$ of W.

Stronger PTA: dCD and SA

- Withing the framework described before, the "stronger" PTA invoked by dCD and SA is equivalent to

$$\begin{aligned}
\alpha_{3,3}(0) &= \mathbb{E}[Y_3 - Y_2 | C = 1] + \mathbb{E}[Y_2 | G_3 = 1], && (2)\\
\alpha_{3,3}(0) &= \mathbb{E}[Y_3 - Y_2 | G_4 = 1] + \mathbb{E}[Y_2 | G_3 = 1], && (3)\\
\alpha_{3,4}(0) &= \mathbb{E}[Y_4 - Y_3 | C = 1] + \alpha_{3,3}(0), && (4)\\
\alpha_{4,4}(0) &= \mathbb{E}[Y_4 - Y_3 | C = 1] + \mathbb{E}[Y_3 | G_4 = 1], && (5)\\
\mathbb{E}[Y_2 - Y_1 | G_3 = 1] &= \mathbb{E}[Y_2 - Y_1 | C = 1], && (6)\\
\mathbb{E}[Y_2 - Y_1 | G_3 = 1] &= \mathbb{E}[Y_2 - Y_1 | G_4 = 1] && (7)\\
\mathbb{E}[Y_3 - Y_2 | G_4 = 1] &= \mathbb{E}[Y_3 - Y_2 | C = 1]. && (8)
\end{aligned}$$

- 3 unknowns and 7 moment restrictions;
- However (8) is a linear combination of the moment conditions (2) and (3), so (8) is redundant.
- **Over-identified system of equations.**

PTA based on “never-treated”: CS

- Within the framework described before, the PTA based on “never-treated” units invoked by CS is equivalent to

$$\alpha_{3,3}(0) = \mathbb{E}[Y_3 - Y_2 | C = 1] + \mathbb{E}[Y_2 | G_3 = 1], \tag{9}$$

~~$\alpha_{3,3}(0) = \mathbb{E}[Y_3 - Y_2 | G_4 = 1] + \mathbb{E}[Y_2 | G_3 = 1]$~~

$$\alpha_{3,4}(0) = \mathbb{E}[Y_4 - Y_3 | C = 1] + \alpha_{3,3}(0), \tag{10}$$

$$\alpha_{4,4}(0) = \mathbb{E}[Y_4 - Y_3 | C = 1] + \mathbb{E}[Y_3 | G_4 = 1] \tag{11}$$

~~$\mathbb{E}[Y_2 - Y_1 | G_3 = 1] = \mathbb{E}[Y_2 - Y_1 | C = 1]$,~~

~~$\mathbb{E}[Y_2 - Y_1 | G_3 = 1] = \mathbb{E}[Y_2 - Y_1 | G_4 = 1]$,~~

~~$\mathbb{E}[Y_3 - Y_2 | G_4 = 1] = \mathbb{E}[Y_3 - Y_2 | C = 1]$~~

- 3 unknowns and 3 moment restrictions
- **Just-identified system of equations.**

PTA based on “not-yet-treated”: CS

- Withing the framework described before, the PTA based on “not-yet-treated” units invoked by CS is equivalent to

$$\alpha_{3,3}(0) = \mathbb{E}[Y_3 - Y_2 | C = 1] + \mathbb{E}[Y_2 | G_3 = 1], \tag{12}$$

$$\alpha_{3,3}(0) = \mathbb{E}[Y_3 - Y_2 | G_4 = 1] + \mathbb{E}[Y_2 | G_3 = 1] \tag{13}$$

$$\alpha_{3,4}(0) = \mathbb{E}[Y_4 - Y_3 | C = 1] + \alpha_{3,3}(0), \tag{14}$$

$$\alpha_{4,4}(0) = \mathbb{E}[Y_4 - Y_3 | C = 1] + \mathbb{E}[Y_3 | G_4 = 1] \tag{15}$$

~~$\mathbb{E}[Y_2 - Y_1 | G_3 = 1] = \mathbb{E}[Y_2 - Y_1 | C = 1]$,~~

~~$\mathbb{E}[Y_2 - Y_1 | G_3 = 1] = \mathbb{E}[Y_2 - Y_1 | G_4 = 1]$,~~

~~$\mathbb{E}[Y_3 - Y_2 | G_4 = 1] = \mathbb{E}[Y_3 - Y_2 | C = 1]$~~

- 3 unknowns and 4 moment restrictions
- **Over-identified system of equations.**
- The PTA based on “not-yet-treated” units does not restrict pre-trends involving time periods before the first unit is treated, and does not restrict pre-trends for the earliest treatment group.

Summary of PTAs

- PTA based on “never-treated” does not restrict pre-trends and is weaker than the “stronger” PTA and the one based on “not-yet-treated” units, though it requires the existence of a “never treated” group.

- PTA based on “not-yet-treated” is arguably weaker then the “stronger” PTA , as the latter restricts all pre-trends in all pre-treatment periods, while the former does not restrict pre-trends involving time periods before the first unit is treated.

- This can be practically relevant in applications where data are available on many time periods before the first group of units is treated.

Parameters of interest

Parameter of interest: dCD

- dCD focuses on an instantaneous treatment effect measure across all “ever treated” groups.

- More precisely, dCD is mainly interested in estimating

$$\delta^{S} \equiv \mathbb{E}\left[\frac{\sum_{i=1}^{n}\sum_{t=2}^{\mathcal{T}} G_{ig}\cdot\left(Y_{it}\left(1\right)-Y_{it}\left(0\right)\right)}{\sum_{i=1}^{n}\sum_{t=2}^{\mathcal{T}} G_{it}}\right], \tag{16}$$

  the average of the treatment effect at the time when a group starts receiving the treatment, across all groups that become treated at some point (see Section 4 of dCD).

Parameter of interest: dCD

- dCD also proposes an easy-to-implement estimator for δ^S. To better understand their estimator, let

$$\widehat{ATT}_{ny}(g,t) = \frac{n^{-1}\sum_{i=1}^{n} G_{ig}\left(Y_{it} - Y_{ig-1}\right)}{n^{-1}\sum_{i=1}^{n} G_{it}} - \frac{n^{-1}\sum_{i=1}^{n}\left(1-D_{it}\right)\left(1-G_{ig}\right)\left(Y_{it} - Y_{ig-1}\right)}{n^{-1}\sum_{i=1}^{n}\left(1-D_{it}\right)\left(1-G_{ig}\right)} \tag{17}$$

  be a DiD estimator for $ATT(g,t)$ that uses not-yet treated units by time t as a comparison group for treatment group g, at time t.

- Let $N_{g \cap \geq e}$ denotes the number of observations in group g among those units that have been treated for at least e periods;

- $N_{\geq e}$ is the number of units who have been treated for at least e periods.

Parameter of interest: dCD

- Consider the estimator for the probability of a unit being in group g given that it is among the units that are treated for at least $e = t - g + 1$ periods given by

$$\widehat{w}(g; e) \equiv \widehat{P}\left(G_g = 1 | \text{Treated for} \geq \text{e periods}\right) = \frac{N_{g \cap \geq e}}{N_{\geq e}}, \tag{18}$$

- dC&D then show that, under the "stronger PTA" and some additional regularity conditions,

$$\widehat{\delta^S} = \sum_{g=2}^{\mathcal{T}} \widehat{P}\left(G_g = 1 | \text{Treated for} \geq 1 \text{ period}\right) \cdot \widehat{ATT}_{ny}\left(g, g\right), \tag{19}$$

  is an unbiased estimator of δ^S, and, as (effective) sample size grows, $\widehat{\delta^S}$ is also consistent and asymptotically normal.

Parameter of interest: CS

- CS propose to analyze the $ATT(g,t)$'s and their functionals so we have a better understanding of treatment effect heterogeneity.
- A "simple" average of all $ATT\,(g,t)$'s,

$$ATT^{simple} = \frac{\sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} \mathbf{1}\{g \leq t\} P(G=g) \cdot ATT(g,t)}{\sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} \mathbf{1}\{g \leq t\} P(G=g)}, \tag{20}$$

- Treatment effect dynamics (event-study-type causal parameters)

$$\delta^{es}(e) = \sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} 1\{t-g+1=e\}\, P(G_g = 1 | \text{Treated for} \geq \text{e periods})\, ATT(g,t), \tag{21}$$

  which provides the average treatment effect for units that have been treated for e periods.

- Average of $\delta^{es}(e)$ over all possible (positive) values of e,

$$\delta^{e,avg} = \frac{1}{\mathcal{T}-1} \sum_{e=1}^{\mathcal{T}-1} \delta^{es}(e). \tag{22}$$

Estimators proposed by CS

- CS proposed estimators depend on the underlying PTA one is willing to invoke.
- For instance, when one imposes the PTA that the "never-treated" or the "not-yet-treated" units serve as comparison groups, one can estimate the "event-study" type estimands $\delta^{es}(e)$ by

$$\widehat{\delta}^{es}_{never}(e) = \sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} 1\{t-g+1=e\}\, \widehat{w}(g;e) \widehat{ATT}_{never}(g,t),$$

$$\widehat{\delta}^{es}_{ny}(e) = \sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} 1\{t-g+1=e\}\, \widehat{w}(g;e) \widehat{ATT}_{ny}(g,t),$$

  respectively, where $e \geq 1$, $\widehat{w}(g;e)$ is defined as in (18), and

$$\widehat{ATT}_{never}(g,t) = \frac{n^{-1}\sum_{i=1}^{n} G_{ig}\left(Y_{it}-Y_{ig-1}\right)}{n^{-1}\sum_{i=1}^{n} G_{ig}} - \frac{n^{-1}\sum_{i=1}^{n} C_{i}\left(Y_{it}-Y_{ig-1}\right)}{n^{-1}\sum_{i=1}^{n} C_{i}} \quad (23)$$

- The aggregated estimators for ATT^{simple} and for $\delta^{e,avg}$ are formed analogously.

SA: Parameter of interest and estimators

- SA is mainly interested in recovering event-study-type parameters $\delta^{es}(e)$.
- Their estimation procedure differs from CS, though.
- Interaction-weighted estimator
  1. use the linear TWFE specification that interacts relative time indicators with treatment group indicator:

$$Y_{it} = \lambda_i + \lambda_t + \sum_{g=2}^{\mathcal{T}-1} \sum_{e \neq 0} \delta_{ge} \cdot G_{ig} 1\{t - G_i + 1 = e\} + v_{it} \tag{24}$$

     on observations from $t = 1, \ldots, \mathcal{T} - 1$, where the last time period $\mathcal{T}$ is dropped in order to accommodate the case where there is no "never treated" group; if there is a never-treated group available, dropping data from time period $\mathcal{T}$ is unnecessary.

  2. estimate $\delta^{es}(e)$ using

$$\widehat{\delta}^{es}_{S\&A}(e) = \sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} 1\{t - g + 1 = e\} \widehat{w}(g; e) \widehat{\delta}_{gt}$$

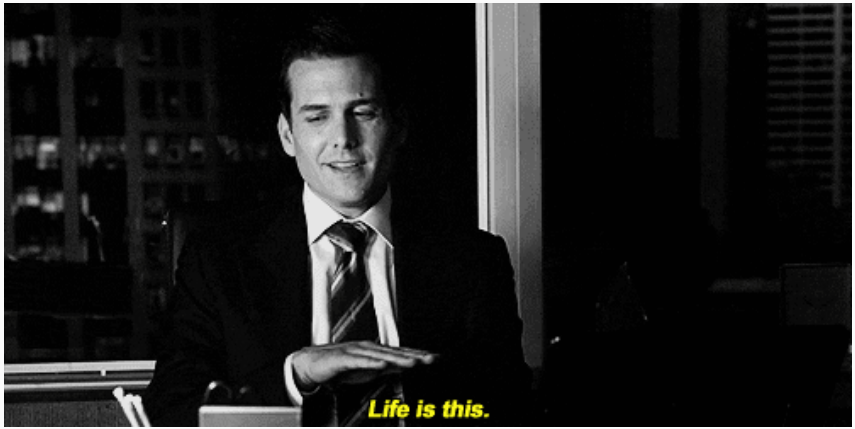
PTAs and the choice of DiD estimators

What if I want to impose the PTA based on "never-treated" comparison groups?

- This PTA is the weakest PTA among the three we have considered so far as it does not impose any restriction on pre-treatment trends across group.
- Only available estimator is $\widehat{ATT}_{never}(g,t)$, though you can obtain it using different algorithms.
- Of course, to use this PTA and estimator, we must have a set of units that do not experience treatment in the time-window we want to analyze.
- When such a group of units is available but its relative size is small, inference procedures based on (23) may not be as precise as one wishes.
- However, it is important to stress that this potential "loss of efficiency" is a direct consequence of not exploiting restrictions on pre-treatment trends across groups.
- In practice, we foresee researchers taking into account this "robustness" versus "efficiency" trade-off when deciding if the PTA based on "never-treated" units is the most suitable for the given application.

What if I want to impose one of the other two PTAs that we have already discussed?

- "Never-treated" group may not be available or may be too small.
- Alternatively, researchers may be comfortable a priori ruling out (some) non-parallel pre-treatment trends
- In these cases, one can rely on the "stronger" PTA or the PTA that uses the not-yet-treated as comparison groups.
- In both cases, we can use the DiD estimator $\widehat{ATT}_{ny}(g,t)$.
- However, this estimator does not fully exploit the information available in the data given either PTA.
- You are leaving money on the table!



Using GMM to get efficient DiD estimators

GMM

- To avoid repetition, we focus on the case where researchers impose the PTA based on "not-yet-treated" units.
- The implementation based on the "stronger" PTA is completely analogous.
- The key to implement the GMM is to list all moment restrictions we are imposing to recover the $ATT(g,t)$'s
  - Involves the moment restrictions implied by the PTA
  - Involves the observational restrictions that, for all $t \geq g$,

$$\begin{aligned} \alpha_{g,t}(1) &\equiv \mathbb{E}[Y_t(1)|G_g = 1] = \mathbb{E}[Y_t|G_g = 1], \\ \alpha_g^{prop} &\equiv \mathbb{E}[G_g], \\ \alpha_C^{prop} &\equiv \mathbb{E}[C]. \end{aligned}$$

- We can then use these "augmented" moment restrictions to efficiently estimate all the unknown parameters involved in our problem by following Hansen(1982).

GMM

- Let's illustrate this in the context of stylized example.
- We want to estimate

$$\alpha \equiv \left(\alpha_{3,3}(1), \alpha_{3,3}(0), \alpha_{3,4}(1), \alpha_{3,4}(0), \alpha_{4,4}(1), \alpha_{4,4}(0), \alpha_C^{prop}, \alpha_3^{prop}, \alpha_4^{prop}\right)'.$$

- Efficient GMM objective function is given by

$$\widehat{\alpha}^{gmm} = \arg\min_{\alpha \in \Theta} \bar{g}_\alpha(W)' \widehat{\Sigma}_{\hat{\alpha},gmm}^{-1} \bar{g}_\alpha(W), \tag{25}$$

with

$$\bar{g}_\alpha(W) = \frac{1}{n}\sum_{i=1}^{n} g_\alpha(W_i),$$

$g_a(W_i)$ combining all (linearly independent) moment conditions, and $\widehat{\Sigma}_{\hat{\alpha},gmm}$ an estimator of the efficient weighting matrix.

$$g_a(W_i) = \begin{pmatrix} \frac{(1-D_{i3})(Y_{i3}-Y_{i2})}{a_C^{prop}+a_4^{prop}} + \frac{G_{i3}Y_{i2}}{a_3^{prop}} - a_{3,3}(0) \\ \frac{(1-D_{i4})(Y_{i3}-Y_{i2})}{a_C^{prop}} + \frac{G_{i3}Y_{i2}}{a_3^{prop}} - a_{3,3}(0) \\ \frac{(1-D_{i4})(Y_{i4}-Y_{i3})}{a_C^{prop}} + a_{3,3}(0) - a_{3,4}(0) \\ \frac{(1-D_{i4})(Y_{i4}-Y_{i3})}{a_C^{prop}} + \frac{G_{i4}Y_{i3}}{a_4^{prop}} - a_{4,4}(0) \\ \frac{G_{i3}Y_{i3}}{a_3^{prop}} - a_{3,3}(1) \\ \frac{G_{i3}Y_{i4}}{a_3^{prop}} - a_{3,4}(1) \\ \frac{G_{i4}Y_{i4}}{a_4^{prop}} - a_{4,4}(1) \\ C_i - a_C^{prop} \\ G_{i3} - a_3^{prop} \\ G_{i4} - a_4^{prop} \end{pmatrix}. \tag{26}$$

GMM

$$\widehat{\Sigma}_{\check{\alpha},gmm} = \frac{1}{n}\sum_{i=1}^{n} g_{\check{\alpha}}\left(W_i\right) g_{\check{\alpha}}\left(W_i\right)',$$

$\check{\alpha}$ being a preliminary consistent estimator for α, say the minimizer of (25) with $\widehat{\Sigma}_{\check{\alpha},gmm}$ replaced by the identity matrix.

GMM

- With $\widehat{\alpha}^{gmm}$, one can then efficiently estimate the parameters of interest: $ATT\,(3,3)$, $ATT\,(3,4)$ and $ATT\,(4,4)$ by

$$\widehat{ATT}_{gmm}\begin{pmatrix} 3,3 \\ 3,4 \\ 4,4 \end{pmatrix} = \begin{pmatrix} \widehat{\alpha}^{gmm}_{3,3}(1) - \widehat{\alpha}^{gmm}_{3,3}(0) \\ \widehat{\alpha}^{gmm}_{3,4}(1) - \widehat{\alpha}^{gmm}_{3,4}(0) \\ \widehat{\alpha}^{gmm}_{4,4}(1) - \widehat{\alpha}^{gmm}_{4,4}(0) \end{pmatrix}. \tag{27}$$

- Properties follows from delta method.
- We can also assess the validity of the PTA using the Sargan-Hansen J-test.
- Challenge: computationally more complicated than $\widehat{ATT}_{ny}\,(g,t)$ when there are many g's and t's.
- In the application, we have 16 treatment groups and 33 time periods, the GMM involves 780 moments with 195 overidentification restrictions, whereas sample size (state-year pairs) is equal to 759.

What if there is not “never-treated” and I do not want to restrict pre-trends?

An alternative DiD estimator

- We want some estimator that is alternative to the GMM one that
  1. easy to compute
  2. does not require the existence of a “never-treated” group
  3. exploits more data than $\widehat{ATT}_{ny}(g, t)$
  4. does not explicitly restrict pre-trends.
- Key: Pay attention to the PTA

Another PTA

Assumption ("Weaker" Parallel trends assumption based on "not-yet treated" units)

For all $g, t = 2, \ldots, \mathcal{T}$, such that $t \geq g$,

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \middle| G_g = 1 \right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \middle| D_t = 0 \right].$$

- The PTA 4 imposes that the evolution of the outcome at time t among those units that have not yet experienced treatment by time t can help us identify the $ATT(g,t)$'s.
- Unlike CS PTA, it does not impose that every individual not-yet-treated group can be used as a comparison group, which, in turn, suggests that the $ATT(g,t)$, $t \geq g$ are nonparametrically just-identified by

$$ATT_{ny+}(g,t) \equiv \mathbb{E}\left[Y_t - Y_{g-1} \middle| G_g = 1 \right] - \left(\sum_{s=g}^{t} \mathbb{E}\left[\Delta Y_s \middle| D_s = 0, G_g = 0 \right] \right). \quad (28)$$

How do we get that formula?

- Let's focus on our stylized example
- There, the above PTA is equivalent to

$$\alpha_{3,3}(0) = \mathbb{E}[Y_3 - Y_2 | C = 1] + \mathbb{E}[Y_2 | C_3 = 1],$$
$$\alpha_{3,3}(0) = \mathbb{E}[Y_3 - Y_2 | C_4 = 1] + \mathbb{E}[Y_2 | C_3 = 1]$$
$$\alpha_{3,3}(0) = \mathbb{E}[Y_3 - Y_2 | D_3 = 0] + \mathbb{E}[Y_2 | G_3 = 1], \quad (29)$$
$$\alpha_{3,4}(0) = \mathbb{E}[Y_4 - Y_3 | D_4 = 0] + \alpha_{3,3}(0), \quad (30)$$
$$\alpha_{4,4}(0) = \mathbb{E}[Y_4 - Y_3 | D_4 = 0] + \mathbb{E}[Y_3 | G_4 = 1] \quad (31)$$

$$\mathbb{E}[Y_2 - Y_1 | C_3 = 1] = \mathbb{E}[Y_2 - Y_1 | C = 1],$$
$$\mathbb{E}[Y_2 - Y_1 | C_3 = 1] = \mathbb{E}[Y_2 - Y_1 | C_4 = 1],$$
$$\mathbb{E}[Y_3 - Y_2 | C_4 = 1] = \mathbb{E}[Y_3 - Y_2 | C = 1]$$

How do we get that formula?

- Let's focus on our stylized example
- There, the above PTA is equivalent to

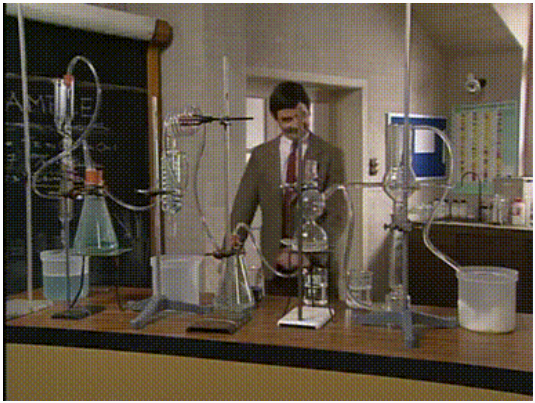
$$
\begin{aligned}
\alpha_{3,3}(0) &= \mathbb{E}[Y_3 - Y_2 | C = 1] + \mathbb{E}[Y_2 | G_3 = 1], \\
\alpha_{3,3}(0) &= \mathbb{E}[Y_3 - Y_2 | G_4 = 1] + \mathbb{E}[Y_2 | G_3 = 1] \\
\alpha_{3,3}(0) &= \mathbb{E}[Y_3 - Y_2 | D_3 = 0] + \mathbb{E}[Y_2 | G_3 = 1], \\
\alpha_{3,4}(0) &= \mathbb{E}[Y_4 - Y_3 | C = 1] + \alpha_{3,3}(0), \\
\alpha_{4,4}(0) &= \mathbb{E}[Y_4 - Y_3 | C = 1] + \mathbb{E}[Y_3 | G_4 = 1] \\
\mathbb{E}[Y_2 - Y_1 | G_3 = 1] &= \mathbb{E}[Y_2 - Y_1 | C = 1], \\
\mathbb{E}[Y_2 - Y_1 | G_3 = 1] &= \mathbb{E}[Y_2 - Y_1 | G_4 = 1], \\
\mathbb{E}[Y_3 - Y_2 | G_4 = 1] &= \mathbb{E}[Y_3 - Y_2 | C = 1]
\end{aligned}
$$

New DiD estimator

- The above results suggest the following simple-to-compute DiD estimator:

$$\widehat{ATT}_{ny+}(g,t) = \frac{n^{-1}\sum_{i=1}^{n} G_{ig}\left(Y_{it} - Y_{ig-1}\right)}{n^{-1}\sum_{i=1}^{n} G_{it}} - \sum_{s=g}^{t}\left(\frac{n^{-1}\sum_{i=1}^{n}\left(1-D_{is}\right)\left(1-G_{ig}\right)\Delta Y_{is}}{n^{-1}\sum_{i=1}^{n}\left(1-D_{is}\right)\left(1-G_{ig}\right)}\right). \quad (32)$$

- We establish its large sample properties in the paper.

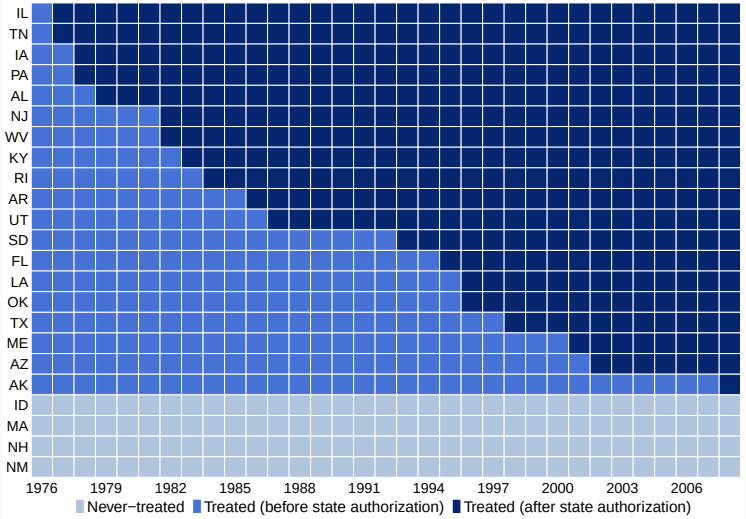


Empirical Application

The effect enforcing the Clean Water Act

- We replicate Katherine Grooms' (2015) analysis of the transition from federal to state management of the Clean Water Act (CWA).
- Environmental policy mandated at the federal level is often implemented at the state level. Yet, there exists variation in the level of enforcement across states.
- Grooms(2015) exploits the staggered timing of the transfer from federal to state monitoring and enforcement of the CWA.
- Let's apply our proposed tools to revisit this debate.

Timing of State Authorization



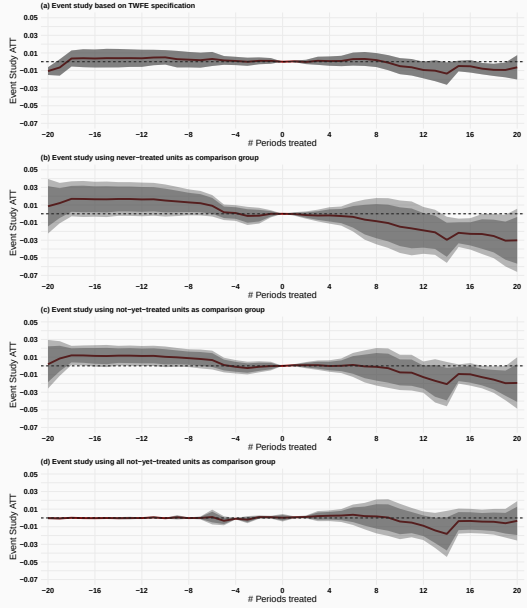
TWFE specifications vs. carefully-crafted event-study estimators

- Start with the TWFE specification

$$Y_{it} = \lambda_i + \lambda_t + \sum_{e=-30, e\neq 0}^{32} \beta_e 1\{t - G_i + 1 = e\} + v_{it}, \tag{33}$$

  which includes 30 treatment lead indicators (all the indicators associated with β_e with $e < 0$) and 32 treatment lag indicators (all the indicators associated with β_e with $e > 0$).

- Compare it with the different event-study-type estimators that explicitly rely on a given PTA.


(a) Event study based on TWFE specification
(b) Event study using never−treated units as comparison group
(c) Event study using not−yet−treated units as comparison group
(d) Event study using all not−yet−treated units as comparison group
Event Study ATT
0.05
0.03
0.01
−0.01
−0.03
−0.05
−0.07
−20
−16
−12
−8
−4
0
4
8
12
16
20
Periods treated

Summary measures	Never-treated (1)	Not-yet-treated (2)	All Not-yet-treated (3)	TWFE (4)
ATT^{simple}	-0.017	-0.010	-0.003	—
	(0.009)	(0.009)	(0.006)	—
	[-0.032, 0.001]	[-0.024, 0.004]	[-0.014, 0.008]	—
$\delta^{e,avg}$	-0.015	-0.008	-0.003	—
	(0.007)	(0.006)	(0.004)	—
	[-0.027, -0.002]	[-0.017, 0.002]	[-0.010, 0.004]	—
TWFE	—	—	—	-0.003
	—	—	—	(0.010)
	—	—	—	[-0.019, 0.013]

Notes: The point estimates, cluster-robust standard errors (in parenthesis), and 90% confidence interval (in brackets) for the effect of state authorization on violation rates. ATT^{simple} is as defined in (20) and denotes the weighted average of all post-treatment $ATT(g,t)'s$. $\delta^{e,avg}$ is as defined in (22) and denotes the time-average of all event-study parameters $\delta^{es}(e)$, $e > 0$. TWFE refers to the ordinary least square estimates of β_{fe} in the TWFE linear regression specification (**??**), which is invariant to the comparison group being used. Column (1) display the results that uses (23) as an estimator for $ATT(g,t)$, column (2) displays the results that uses (17) as an estimator for $ATT(g,t)$, and column (3) displays the results that uses (32) as an estimator for $ATT(g,t)$ Column (4) displays the result using the TWFE regression specification. Standard errors are clustered at the state level, and, with the exception of the TWFE summary measure, are computed using the multiplicative bootstrap procedure described in Algorithm B.1, which is akin to the one proposed by C&S. We use 1,000 bootstrap draws.

Does the effect vary depending on whether a state is "corrupt"?

- Next, we analyze whether the effect of state authorization on violation rates vary depending on whether a state has a long prevalence of corruption.
- We follow Grooms (2015) and consider the following TWFE specification

$$Y_{it} = \alpha_i + \alpha_t + \sum_{e=-30, e\neq 0}^{32} \beta_e 1\{t - G_i + 1 = e\} + \sum_{e=-30, e\neq 0}^{32} \beta_e^c (1\{t - G_i + 1 = e\} \times Corrupt_i) + v_{it}, \quad (34)$$

  where the β_e^c's are considered to be a measure of how treatment effects vary depending on whether a state is "corrupt" or not: positive (negative) point estimates suggest that the violation rates increased (decreased) more in corrupt states than in non-corrupt states.

- **But what type of PTA is being made in the TWFE specification?**
- **Is the TWFE regression model above susceptible to the potential pitfalls discussed in the begining of the slides?**

Being explicit about the PTA

Assumption (PTA based on "never treated" w/ corruption-specific trends)

For $c = 0, 1$, and all g, $t = 2, \ldots, \mathcal{T}$, such that $t \geq g$,

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | G_g = 1, Corr = c \right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | C = 1, Corr = c \right].$$

Assumption (PTA based on "not-yet treated" w/ corruption-specific trends)

For $c = 0, 1$, and all g, s, $t = 2, \ldots, \mathcal{T}$, such that $t \geq g$, $s \geq t$,

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | G_g = 1, Corr = c \right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | D_s = 0, Corr = c \right].$$

Assumption ("Weaker" PTA based on "not-yet treated" w/ corruption-specific trends)

For $c = 0, 1$, and all g, $t = 2, \ldots, \mathcal{T}$, such that $t \geq g$,

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | G_g = 1, Corr = c \right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | D_t = 0, Corr = c \right].$$

Being explicit about the PTA (cont.)

Assumption (PTA based on "never treated" w/o corruption-specific trends)

For $c = 0, 1$, and all g, $t = 2, \ldots, \mathcal{T}$, such that $t \geq g$,

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | G_g = 1, Corr = c \right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | C = 1 \right].$$

Assumption (PTA based on "not-yet treated" w/o corruption-specific trends)

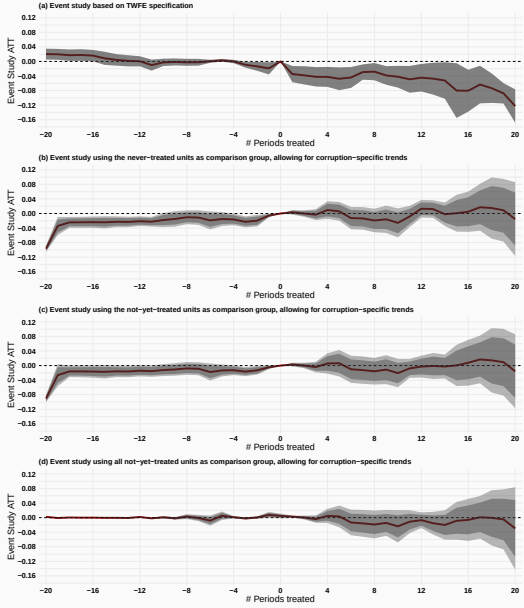
For $c = 0, 1$, and all g, s, $t = 2, \ldots, \mathcal{T}$, such that $t \geq g$, $s \geq t$,

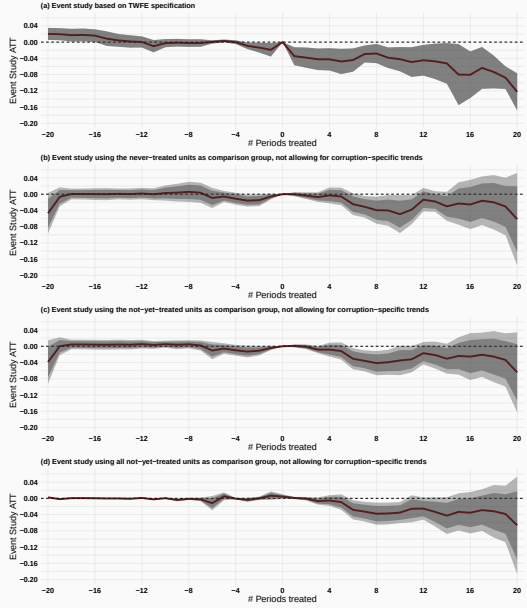
$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | G_g = 1, Corr = c \right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | D_s = 0 \right].$$

Assumption ("Weaker" PTA based on "not-yet treated" w/o corruption-specific trends)

For $c = 0, 1$, and all g, $t = 2, \ldots, \mathcal{T}$, such that $t \geq g$,

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | G_g = 1, Corr = c \right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | D_t = 0 \right].$$


(a) Event study based on TWFE specification
(b) Event study using the never−treated units as comparison group, allowing for corruption−specific trends
(c) Event study using the not−yet−treated units as comparison group, allowing for corruption−specific trends
(d) Event study using all not−yet−treated units as comparison group, allowing for corruption−specific trends
Event Study ATT
0.12
0.08
0.04
0.00
−0.04
−0.08
−0.12
−0.16
−20
−16
−12
−8
−4
0
4
8
12
16
20
Periods treated


(a) Event study based on TWFE specification
(b) Event study using the never−treated units as comparison group, not allowing for corruption−specific trends
(c) Event study using the not−yet−treated units as comparison group, not allowing for corruption−specific trends
(d) Event study using all not−yet−treated units as comparison group, not allowing for corruption−specific trends
Event Study ATT
0.04
0.00
−0.04
−0.08
−0.12
−0.16
−0.20
−20
−16
−12
−8
−4
0
4
8
12
16
20
Periods treated

Summary measures	Allow for corrupt-specific trends			Not allow corrupt-specific trend			TWFE
	Never-treated (1)	Not-yet-treated (2)	All Not-yet-treated (3)	Never-treated (4)	Not-yet-treated (5)	All Not-yet-treated (6)	(7)
$ATT^{simple,1-0}$	-0.007	-0.008	-0.014	-0.035	-0.035	-0.033	—
	(0.014)	(0.014)	(0.013)	(0.012)	(0.012)	(0.010)	—
	[-0.030, 0.017]	[-0.031, 0.016]	[-0.036, 0.008]	[-0.054, -0.016]	[-0.054, -0.015]	[-0.049, -0.017]	—
$\delta^{e,avg,1-0}$	-0.001	-0.002	-0.009	-0.024	-0.025	-0.028	—
	(0.014)	(0.016)	(0.014)	(0.013)	(0.012)	(0.011)	—
	[-0.024, 0.022]	[-0.028, 0.024]	[-0.031, 0.013]	[-0.045, -0.003]	[-0.045, -0.005]	[-0.047, -0.009]	—
TWFE	—	—	—	—	—	—	-0.037
	—	—	—	—	—	—	(0.010)
	—	—	—	—	—	—	[-0.054, -0.020]

Notes: The point estimates, cluster-robust standard errors (in parenthesis), and 90% confidence interval (in brackets) for the effect of state authorization on violation rates. $ATT^{simple,2-0}$ is as defined in (**??**) and denotes the difference of the weighted average of all post-treatment $ATT(g,t;c)'s$ between corrupt and non-corrupt states. $\delta^{e,avg,1-0}$ is as defined in (**??**) and denotes difference of the time-average of all event-study parameters $\delta^{es}(e)$, $e>0$, between corrupt and non-corrupt states. TWFE refers to the ordinary least square estimates of β^{c}_{fe} in the TWFE linear regression specification (**??**), which is invariant to the comparison group being used. Columns (1)-(6) display the results that relies on the PTA 5-10, respectively. Standard errors are clustered at the state level, and, with the exception of the TWFE summary measure, are computed using the multiplicative bootstrap procedure presented in Algorithm B.1, which is akin to C&S proposal. We use 1,000 bootstrap draws.