Part C: Panel Data Methods

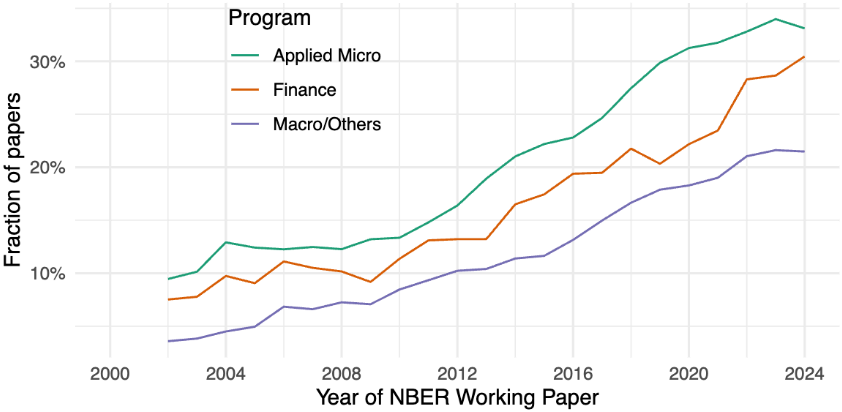
C2: Canonical Difference-in-Differences

Kirill Borusyak

ARE 213 Applied Econometrics

UC Berkeley, Fall 2024

It's everywhere



(a) Difference-in-differences

(Goldsmith-Pinkham (2024), Fig.5a)

C2 outline

1. $2x2$ DiD
2. DiD with multiple units and periods
3. Extensions

Difference-in-differences: Idea

- We are interested in causal effects of some binary treatment
- But treatment is not randomly assigned across units
- "Quasi-experimental" contrasts are easier to find in panel data
- Card and Krueger (1994): effect of minimum wages on employment
  - On April 1, 1992, NJ raised the min.wage from \$4.25 to \$5.05
  - Min.wage in PA stayed at \$4.25
  - Measure average employment at fast food restaurants before (Feb 1992) and after (Nov 1992)

Card and Krueger (1994)

Variable		PA (i)	NJ (ii)	Difference, NJ-PA (iii)
1.	FTE employment before,	23.33	20.44	-2.89
	all available observations	(1.35)	(0.51)	(1.44)
2.	FTE employment after,	21.17	21.03	-0.14
	all available observations	(0.94)	(0.52)	(1.07)
3.	Change in mean FTE	-2.16	0.59	2.76
	employment	(1.25)	(0.54)	(1.36)

(Card-Krueger Table 3, reproduced from MHE Table 5.2.1)

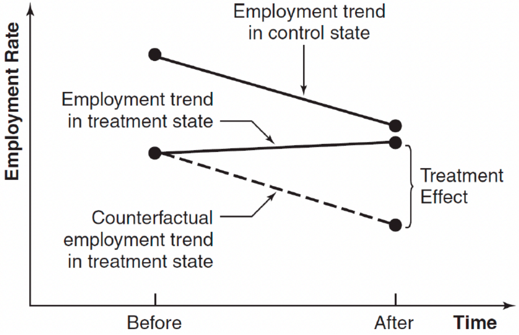
Is this causal?

- Define potential outcomes $Y_{it}(d)$ for $d = 0, 1$ (low/high min.wage), with $\tau_{it} = Y_{it}(1) - Y_{it}(0)$ and $Y_{it} = Y_{it}(D_{it})$

D_{it}	$i = PA$	$i = NJ$
$t = 1$	0	0
$t = 2$	0	1

- Estimand: $ATT = \tau_{NJ,2}$
- Assumptions for this potential outcomes formulation:
  - No spillovers
  - No anticipation effects: Y_{i1} does not depend on D_{i2}
  - No lagged effects: Y_{i2} does not depend on D_{i1}
  - Parallel trends...

Parallel trends (PTA)



(MHE Figure 5.2.1, corrected by Peter Hull)

Parallel trends: $\mathbb{E}\left[Y_{NJ,2}(0) - Y_{NJ,1}(0)\right] = \mathbb{E}\left[Y_{PA,2}(0) - Y_{PA,1}(0)\right]$

- Equivalently, $\mathbb{E}\left[Y_{it}(0)\right] = \alpha_i + \beta_t$ for $i = NJ, PA$ and $t = 1, 2$
- *Note*: notation for a fixed sample, $\mathbb{E}\left[Y_{it}(0)\right]$ depends on i, t. Expectations taken with respect to other determinants of the outcome

2x2 DiD estimator

- $\hat{\tau} = \left(\boxed{Y_{NJ,2}} - Y_{NJ,1} \right) - \left(Y_{PA,2} - Y_{PA,1} \right)$ is unbiased for ATT

- $\hat{\tau}$ can be obtained as OLS from the TWFE specification

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \tau D_{it} + \varepsilon_{it}$$

  - Proof: with 2 periods, the first-differenced equation $Y_{i2} - Y_{i1} = \beta + \tau \left(D_{i2} - D_{i1} \right) + \left(\varepsilon_{i2} - \varepsilon_{i1} \right)$ gives the same $\hat{\tau}$

Problems with 2x2 designs (1)

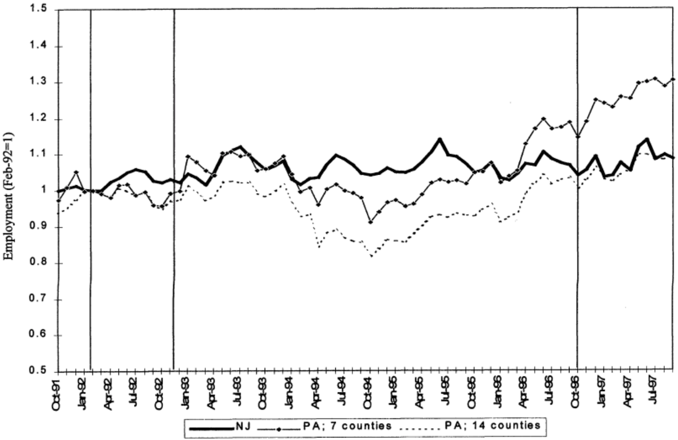


FIGURE 2. EMPLOYMENT IN NEW JERSEY AND PENNSYLVANIA FAST-FOOD RESTAURANTS, OCTOBER 1991 TO SEPTEMBER 1997

(Card-Krueger 2000 for an extended time period)

Are these fluctuations captured by SE in firm-level analysis of Card-Krueger (1994)?

Problems with 2x2 designs (2)

1. Effectively 4 observations: can't separate τ from other shocks in NJ relative to PA

Variable	PA (i)	NJ (ii)	Difference, NJ-PA (iii)
1. FTE employment before,	23.33	20.44	-2.89
all available observations	(1.35)	(0.51)	(1.44)
2. FTE employment after,	21.17	21.03	-0.14
all available observations	(0.94)	(0.52)	(1.07)
3. Change in mean FTE	-2.16	0.59	2.76
employment	(1.25)	(0.54)	(1.36)

These SE are not clustered, and cannot be!

Problems with 2x2 designs (2)

1. Effectively 4 observations: can't separate τ from other shocks in NJ relative to PA

2. Can't falsify parallel trends

Solution: have more (pre-)periods and more units

	$i = A$	$i = B$	$i = C$	$i = D$
$t = 1$				
$t = 2$				
$t = 3$				
$t = 4$				
$t = E = 5$				
$t = 6$				

Outline

Causal structure with multiple periods

This setting accommodates two scenarios:

1. Event happens at $t = 5$ only but can have persistent effects
   - D_{it} = having ever been exposed to the treatment/event
2. Policy switches on at $t = 5$ and stays on
   - Two types of effects at $t = 6$: contemporaneous effects of D_{i6}, delayed effects of D_{i5}
   - Potential outcomes $Y_{i6}(d_{i5}, d_{i6})$ would acknowledge this
   - But can't distinguish them here $\implies$ simplify notation to $Y_{i6}(d_{i6})$

PTA with multiple units and periods

- PTA at the unit level (in a fixed sample):

$$\mathbb{E}\left[Y_{it}(0)\right] = \alpha_i + \beta_t \qquad \forall i, t$$
$$\iff \mathbb{E}\left[Y_{it}(0) - Y_{i,t-1}(0)\right] = \mathbb{E}\left[Y_{jt}(0) - Y_{j,t-1}(0)\right] \qquad \forall i, j, t$$

- Or impose PTA at the treatment/control group level (in a random sample):

$$\mathbb{E}\left[Y_{it}(0) - Y_{i,t-1}(0) \mid Tr_i = 1\right] = \mathbb{E}\left[Y_{it}(0) - Y_{i,t-1}(0) \mid Tr_i = 0\right], \qquad \forall t$$

where Tr_i is a dummy of being in the treated group

Why might PTA hold?

- **Design-based approach** (approximating an experiment): assignment into treatment vs control group is random
  - Balance on levels in addition to trends $\Longrightarrow$ have better tools than DiD! (Roth and Sant'Anna, JPE:Micro 2023)
  - Justifies group-level PTA
- **Model-based approach** (using an outcome model): from contextual knowledge that unobserved factors are stable
  - Better with large shocks and in the short-run
  - Individual-level PTA is more appropriate
  - Levels vs logs matters (Roth and Sant'Anna, Ecma 2023)
- Common bad practice: don't justify PTA, just check for pre-trends

Justifying PTA assumptions is important

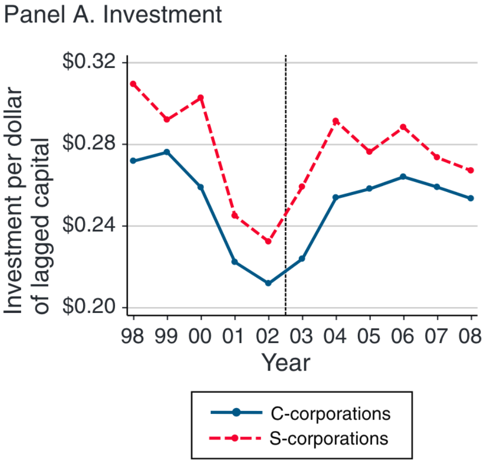
From David McKenzie's 2022 blog post:

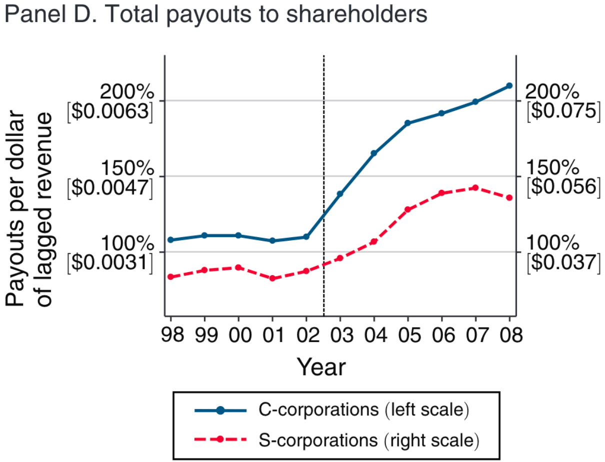
- [Three recent QJE] papers all use a variety of statistical methods to assess and assert the validity of their DiD estimates [...] But I found all three lacking in providing a discussion of why parallel trends should hold
- It is commonplace in papers that use IV to spend a lot of time arguing for the plausibility of the ultimately untestable exclusion restriction, but I think DiD papers have not done the same for the untestable assumption about future trends

Example: Yagan (2015, AER)

- Studies the effects of the 2003 dividend tax cut from 38.6% to 15%
- Treatment group: C-corporations; control group: S-corporations
- S-corporations don't pay dividend tax; otherwise similar taxation
- *"The identifying assumption is not random assignment of C- versus S-status; it is that C- and S-corporation outcomes would have trended similarly in the absence of the tax cut"*
- Justifications:
    1. *"C- and S-corporations of the same ages operate in the same narrow industries and at the same scale"*
    2. *"Contemporaneous stimulative tax provisions applied almost identically"*
    3. *"Key outcomes empirically trended similarly for C- and S-corporations before 2003"*

Yagan (2015): Plotting raw data





What would you ask the author?

“Static” and “dynamic” ATT estimation

Assume a balanced panel. Static TWFE specification: $Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \tau D_{it} + \varepsilon_{it}$

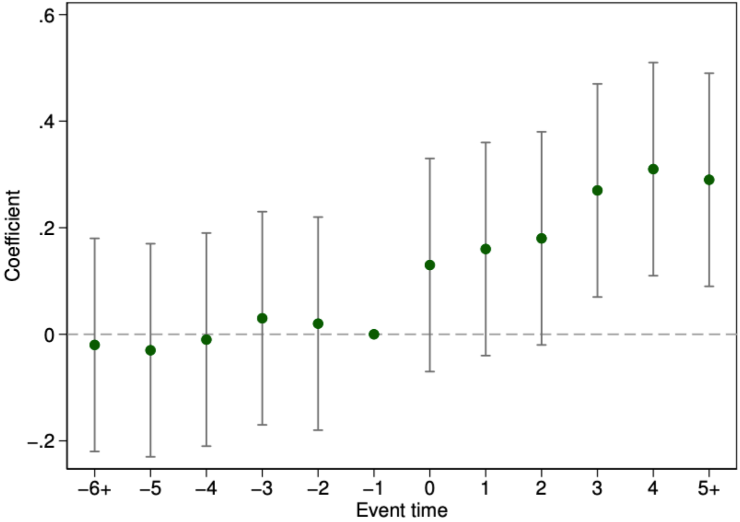
- $\hat{\tau} = (\bar{Y}_{\text{treated,post}} - \bar{Y}_{\text{treated,pre}}) - (\bar{Y}_{\text{control,post}} - \bar{Y}_{\text{control,pre}})$
- Under PTA, $\mathbb{E}[\hat{\tau}] = ATT \equiv$ avg τ_{it} across all treated units and “post” periods
- Doesn't *assume* static effects but not very useful if effects exhibit strong dynamics

“**Event study**” dynamic specification:

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \sum_{h=-(E-1)}^{T-E} \tau_h \mathbf{1}\left[t - E = h\right] Tr_i + \varepsilon_{it}, \qquad \tau_{-1} = 0$$

- $\hat{\tau}_h = (\bar{Y}_{\text{treated},E+h} - \bar{Y}_{\text{treated},E-1}) - (\bar{Y}_{\text{control},E+h} - \bar{Y}_{\text{control},E-1})$
- For $h \geq 0$, unbiased for ATT h periods after the event
- For $h < -1$, unbiased for zero under PTA: yields a “pre-trends” test

Example event study plot



(Example from Freyaldenhoven et al. 2021, Fig. 1)

Discussion of conventional event studies

Event study regression conflates estimation and testing

- If PTA holds, why not use all pre-periods for more efficient estimation?

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \sum_{\mathbf{h \geq 0}} \tau_h \mathbf{1}\left[t - E = h\right] Tr_i + \varepsilon_{it} \quad \text{in the full sample}$$

  - Using all periods before E as reference periods

- Then test for pre-trends on the sample before E only:

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \sum_{\mathbf{h < -1}} \tau_h \mathbf{1}\left[t - E = h\right] Tr_i + \varepsilon_{it}, \quad \text{for } t < E$$

Discussion of conventional event studies (2)

Pre-trend tests are at best suggestive:

- We imposed parallel trends in $Y_{it}(0)$ both pre- and post-treatment
  - Parallel trends after $t = E - 1$ are key and untestable
- Never say:
  - "My identification assumption is parallel pre-trends"
  - "The pre-trend test passes, which implies that my diff-in-diff strategy is solid"

Discussion of conventional event studies (3)

Pre-trend tests are not always informative:

- Mann and Pozzoli (2024) study the effects of low-skill immigration into Danish municipalities on industrial robot adoption by firms in 1995-2019
- How informative are their regressions of pre-period (1993-95) changes in robot adoption and other variables on the 1995-2019 change in low-skill immigration?

Table 3
Pre-sample trends and long-run changes in robotization.

	Δ robot users (broad) (1993–1995) (1)	Δ imports (1993–1995) (2)	Δ exports (1993–1995) (3)	Δ capital stock (1993–1995) (4)	ΔLow skill share (1993–1995) (5)
Δ Non-West Img IV (1995–2019)	0.476 (0.583)	0.300 (0.262)	−0.272 (0.482)	−0.390* (0.228)	−0.013 (0.026)
N	94	94	94	94	94
R-sq	0.355	0.166	0.067	0.277	0.332

Discussion of conventional event studies (4)

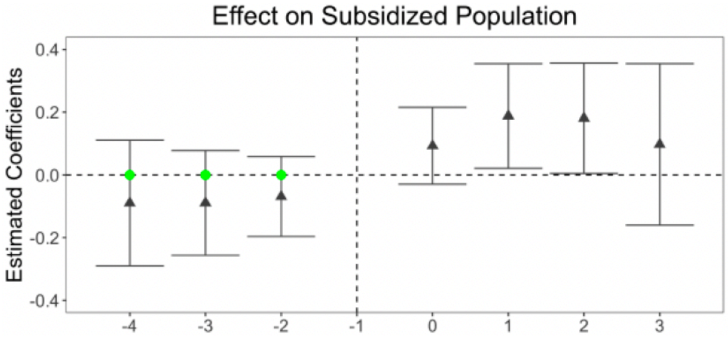
- Pre-trend tests can have low power against substantially different alternatives (Roth AER:I 2022)

    - He and Wang (2017 AEJ:Applied) study the effects of college-graduated bureaucrats placed to Chinese villages

    - Outcome = % of subsidized population (a measure of poverty)

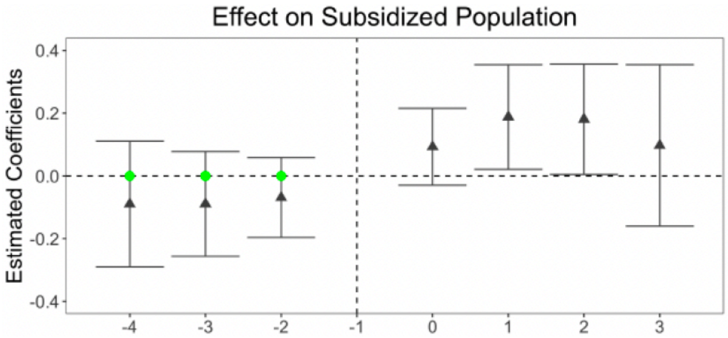
Discussion of conventional event studies (4)

- Pre-trend tests can have low power against substantially different alternatives (Roth AER:I 2022)
  - He and Wang (2017 AEJ:Applied) study the effects of college-graduated bureaucrats placed to Chinese villages
  - Outcome = % of subsidized population (a measure of poverty)



(From Jonathan Roth's Mixtape Session slides)

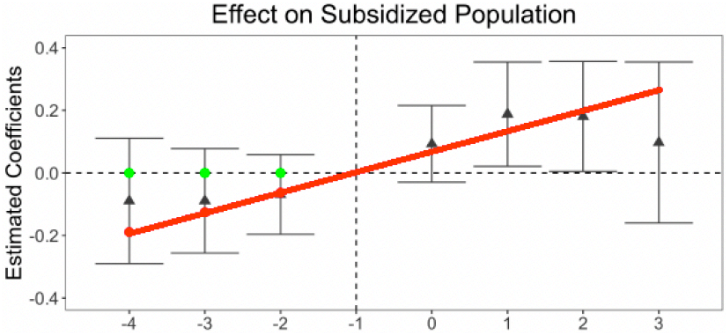
Discussion of conventional event studies (4)



(From Jonathan Roth's Mixtape Session slides)

- *"The estimated coefficients on the leads of treatment ... are statistically indifferent from 0. ... We conclude that the pretreatment trends in the outcomes in both groups of villages are similar, and villages without CGVOs [treatment] can serve as a suitable control group for villages with CGVOs in the treatment period."*

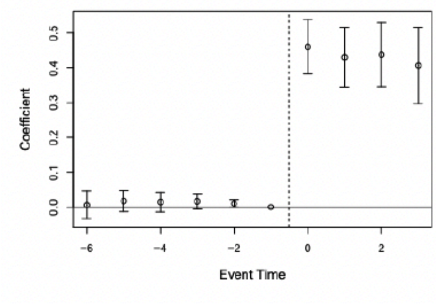
Discussion of conventional event studies (4)



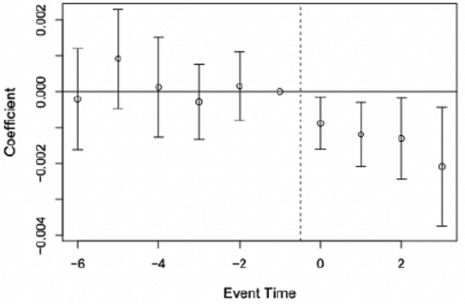
(From Jonathan Roth's Mixtape Session slides)

The “Straight line test”

- The more nonlinearity is needed to kill the effects, the more robust the results are



vs.



- See Stata package *pretrends* accompanying Roth (2022)

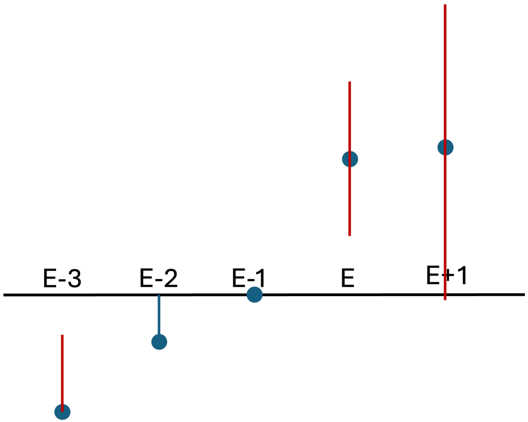
Checking robustness to PTA violations

Rambachan and Roth (2023): partial identification approach

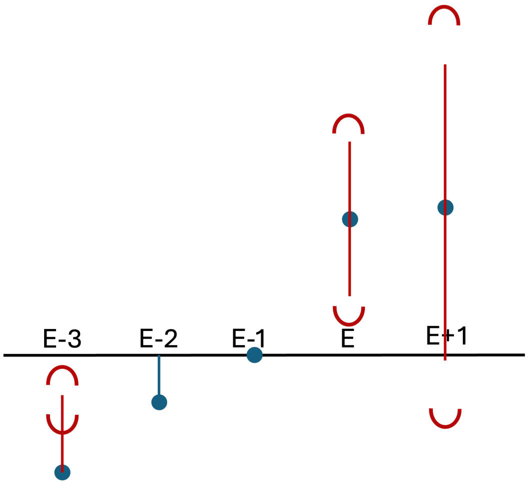
- Let $\delta_t = \mathbb{E}\left[Y_{it}(0) - Y_{i,t-1}(0) \mid Tr_i = 1\right] - \mathbb{E}\left[Y_{it}(0) - Y_{i,t-1}(0) \mid Tr_i = 0\right]$
- PTA requires $\delta_t = 0$; Estimable for $t < E$ but not $t \geq E$
- Weaker assumptions for $t \geq E$ as robustness checks:
    1. Differential trends are not too large: $|\delta_t| \leq M \cdot \max_{s<E} |\delta_s|$

No noise

Imagine a stylized event study plot [*not* a plot of δ_t]:



With noise



Checking robustness to PTA violations

Rambachan and Roth (2023): partial identification approach

- Let $\delta_t = \mathbb{E}\left[Y_{it}(0) - Y_{i,t-1}(0) \mid Tr_i = 1\right] - \mathbb{E}\left[Y_{it}(0) - Y_{i,t-1}(0) \mid Tr_i = 0\right]$
- PTA requires $\delta_t = 0$; Estimable for $t < E$ but not $t \geq E$
- Weaker assumptions for $t \geq E$ as robustness checks:
  1. Differential trends are not too large: $|\delta_t| \leq M \cdot \max_{s<E} |\delta_s|$
  2. Differential trends are smooth: $|\delta_t - \delta_{t-1}| \leq M$
     - *Note:* for $M = 0$ this is not PTA but a linear differential trend
- When pre-trends are noisily estimated, conf.interval will be wider
- Need to pick M or compute the largest M that doesn't kill your findings

Outline

Extensions

1. Continuous treatment intensity
2. Including covariates
3. Measuring effect heterogeneity
4. Triple-diffs
5. Spillovers

Continuous treatment intensity: Setting

- Suppose $D_{it} = \mathbf{1}[t \geq E] \times Tr_i$ where Tr_i is treatment dosage which is not binary
    - Enikolopov et al. (2011): effect of region i's exposure to independent TV channel (NTV) on election outcomes in Russia
    - In 1995, $D_{it} \equiv 0$; in 1999, $D_{it} = Tr_i =$ % of regional population with NTV access
- Estimands of interest:

$$\mathbb{E}\left[Y_{it}(d) - Y_{it}(d') \mid Tr_i = d\right] \quad \text{or} \quad \mathbb{E}\left[\frac{\partial Y_{it}(d)}{\partial d} \mid Tr_i = d\right]$$

- Can still impose PTA on $Y_{it}(0)$

Continuous treatment intensity: Event study

- Can still run the event study regression:

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \sum_{h \neq -1} \tau_h \mathbf{1}\left[t = E + h\right] \times Tr_i + \text{error}_{it}$$

- Or, equivalently, get $\hat{\tau}_h$ from h-specific differenced regression

$$Y_{i,E+h} - Y_{i,E-1} = \tilde{\beta}_h + \tau_h Tr_i + \text{error}_i$$

- But with heterogeneous effects and without randomization of Tr_i, don't get a causal estimand
  - PTA on $Y_{it}(0)$ does not allow comparing units with different dosages
  - OLS subtracts outcomes of units with low $Tr_i > 0 \implies$ puts "**negative weights**" on their effects

Continuous treatment intensity: What to do?

- If Tr_i is not 0 for any i, PTA on $Y_{it}(0)$ is entirely unhelpful
- If $Tr_i = 0$ for some i, PTA allows comparisons relative to the never-treated group
  - Yields $ATT(d) = \mathbb{E}\left[Y_{it}(d) - Y_{it}(0) \mid Tr_i = d\right]$; linked to **imputation** strategies
  - Still cannot identify $\mathbb{E}\left[Y_{it}(d) - Y_{it}(d') \mid Tr_i = d\right]$ for $d' \neq 0$
- Can get further with randomization, restrictions on heterogeneous effects, or stronger PTAs
  - See Callaway, Goodman-Bacon, Sant'Anna (2021), expect more work

DiD with covariates (1)

Two ways of thinking about covariates:

- *Borusyak, Jaravel, Spiess (2024)*: make the model of $Y_{it}(0)$ richer:

$$\mathbb{E}\left[Y_{it}(0)\right] = \alpha_i + \beta_t + \gamma' X_{it}$$

  - Can add unit-specific linear trends $\gamma_i \cdot t$ (to allow units to be on arbitrary non-parallel linear trends)
  - Or time-interacted baseline characteristics $\gamma_t' X_i$ (to allow units with different X_i to be on different nonlinear trends)
  - Or just regular X_{it}. (Of course, avoid bad controls!)

$\Rightarrow$ Estimate γ from untreated observations only, then "imputation"

DiD with covariates (2)

Two ways of thinking about covariates:

- *Abadie (2005), Sant'Anna and Zhao (2020)*: impose PTA conditional on time-invariant baseline characteristics

$$\mathbb{E}\left[\Delta Y_{it}(0) \mid Tr_i = 1, X_i\right] = \mathbb{E}\left[\Delta Y_{it}(0) \mid Tr_i = 0, X_i\right]$$

$\Rightarrow$ Do covariate adjustment for ΔY_{it} (e.g. IPW or AIPW)

DiD with covariates: What NOT to do

- With heterogeneous effects, do not just add covariates to TWFE and event study regressions, e.g. avoid

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \tau D_{it} + \gamma' X_{it} + \text{error}$$

  - $\hat{\gamma}$ will use treated observations too, misattributing some treatment effects

- *Example:* Wolfers (AER 2006) studies the effect of unilateral divorce laws on divorce rates

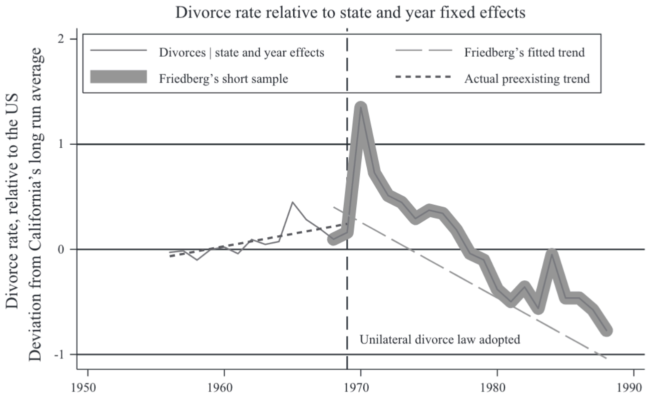
  - Reanalyzes Friedberg (1998) who found a positive effect

Results without and with state-specific trends

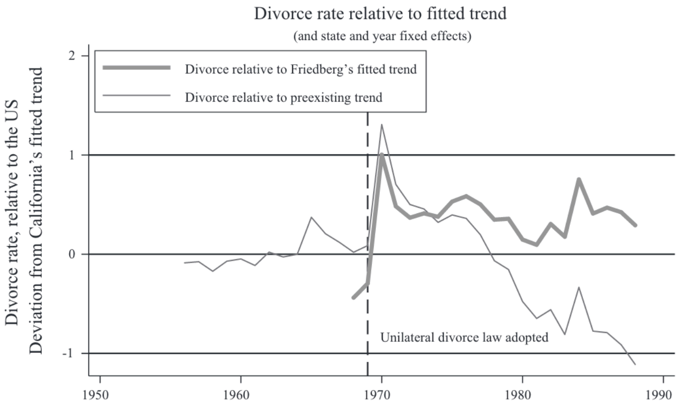
	(1) Basic specification	(2) State-specific trends linear
	Panel A. Friedberg (1998)	
Unilateral	0.004	0.447
	(0.056)	(0.050)
Year effects	$F = 89.0$	$F = 95.3$
State effects	$F = 217.3$	$F = 196.2$
State trend, linear	No	$F = 24.7$
State trend, quadratic	No	No
Adjusted R^2	0.946	0.976

Which results should we trust: with extra controls or without?

Example of California



Example of California (2)



- With appropriate estimators, state-specific trends don't change estimates (de Chaisemartin and D'Haultfœuille, Econometrics J. 2023)

Measuring effect heterogeneity

- If we are interested in the difference between ATT by group R_i (e.g. female dummy), can estimate

$$Y_{it} = \alpha_i + \beta_t + \gamma_t R_i + \tau_1 D_{it} + \tau_2 D_{it} R_i + \text{error}$$

or the event study version

Triple-differences (1)

- Suppose for each state $i = NJ, PA$ and period $t = 1, 2$ we observe groups $g = L, H$: low- and high-wage occupations
- Min.wage should affect group L only: $D_{igt} = \mathbf{1}[i = NJ] \times \mathbf{1}[g = L] \times \mathbf{1}[t = 2]$
- Strategy 1:
  - Use DiD on data from group L to learn the ATT for them
  - Also run **placebo** DiD for group H: test $\tau_H = 0$ in

$$Y_{iHt} = \alpha_{iH} + \beta_{Ht} + \tau_H \mathbf{1}[i = NJ] \times \mathbf{1}[t = 2] + \varepsilon_{iHt}$$

Triple-differences (2)

- Strategy 2 (**triple-differences**):
  - Allow non-parallel trends between NJ and PA: placebo would fail
  - But assume violations of PTA are the same for $g = L, H$:

$$\mathbb{E}\left[\Delta Y_{NJ,L}(0) - \Delta Y_{PA,L}(0)\right] = \mathbb{E}\left[\Delta Y_{NJ,H}(0) - \Delta Y_{PA,H}(0)\right] \equiv \alpha$$

  - Equivalent to $\mathbb{E}\left[\Delta Y_{ig}(0)\right] = \alpha_i + \beta_g$ and $\mathbb{E}\left[Y_{igt}(0)\right] = \alpha_{it} + \beta_{gt} + \gamma_{ig}$
  - Can estimate $ATT = \tau_{NJ,L,2}$ from 2

$$Y_{igt} = \tilde{\alpha}_{it} + \tilde{\beta}_{gt} + \tilde{\gamma}_{ig} + \tau D_{igt} + \text{error}_{igt}$$

Triple-diff $\neq$ effect heterogeneity

Do not confuse triple-differences with estimating effect heterogeneity: in triple-diff,

- Group H is untreated in NJ at $t = 2$
- Differential trends for group H between NJ and PA are not a consequence of the treatment, but unexplained
- The estimand is the causal effect on group L in NJ at $t = 2$, not some difference in effects
- The regression may look the same, but the underlying assumptions and the estimand are different

Spillovers (1)

- Consider a 2-period DiD with spillovers: $Y_{i2} = Y_{i2}(\boldsymbol{D}_2) \equiv Y_{i2}(\mathbf{0}) + \tau_{i2}(\boldsymbol{D}_2)$
  - $\tau_{i2}(\boldsymbol{d}_2)$ is the effect of a vector of treatments $\boldsymbol{d}_2$ on person *i*'s outcome relative to the counterfactual where nobody is treated
- Impose PTA on untreated outcomes:
  $\mathbb{E}\left[Y_{i2}(\mathbf{0}) - Y_{i1}(\mathbf{0}) \mid D_{i2} = 1\right] = \mathbb{E}\left[Y_{i2}(\mathbf{0}) - Y_{i1}(\mathbf{0}) \mid D_{i2} = 0\right]$
- Then DiD yields *relative* effect $\mathbb{E}\left[\tau_{i2}(\boldsymbol{D}_2) \mid D_{i2} = 1\right] - \mathbb{E}\left[\tau_{i2}(\boldsymbol{D}_2) \mid D_{i2} = 0\right]$
- Identifies the *aggregate* ATT if and only if $\mathbb{E}\left[\tau_{i2}(\boldsymbol{D}_2) \mid D_{i2} = 0\right] = 0$
  - Still cannot separate direct and indirect effects

Spillovers (2)

What can we do if the control group *can* be affected?

- Local spillovers are easier:
  - View untreated units that can be affected by spillovers as another treatment arm
  - E.g., "donut approach" in spatial settings
- Two-tier designs: e.g. use variation in % of treated across regions, on top of treated-control comparisons within regions
  - Help separate direct effects from spillovers
- Theoretical models can link spillovers to fundamental (identifiable) parameters
  - E.g. how import competition affect equilibrium wages