Econometrics II

Lecture 10: Dynamic Difference-in-Differences

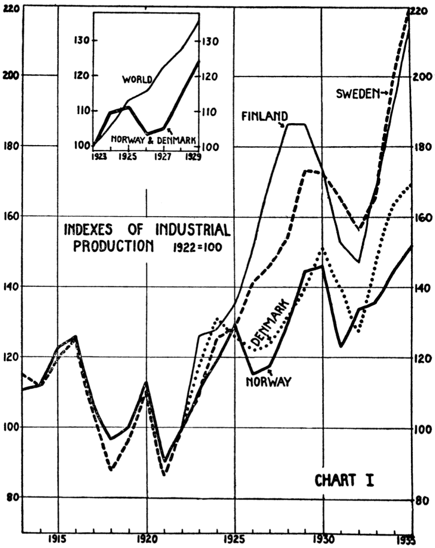
David Schönholzer

Stockholm University

May 2, 2024

Plan for Today

Example: Lester (1937)



- Great identification in macro exists!
  - DK and NO reintroduced gold standard
  - FI and SE did not
  - Large difference in industrial production

- Issues we study today:
  1. Parallel trends:
     - Are FI and SE good counterfactuals?
     - Did they evolve similarly before 1925?
  2. Time heterogeneity / dynamics:
     - How did effect change over time?
  3. Unit heterogeneity:
     - Did DK and NO respond differently?
     - Different in short-run or long-run?

Table of Contents

Parallel Trends Assumption and Pre-Trends

- PTA is untestable! We can never observe $\mathbb{E}\left[Y_{it}(0) \mid D_{it}=1\right]$
- However, we can test for parallel *pre-trends*
  - Absence of pre-trends provides evidence for PT
  - But common shocks remain a threat
- Need at least one $G_i \in [3, \infty)$ to test for pre-trends, e.g.

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- *Placebo test*: $\mathbb{E}\left[Y_{i2} - Y_{i1} \mid D_i = 1\right] = \mathbb{E}\left[Y_{i2} - Y_{i1} \mid D_i = 0\right]$
- Note that with $g \geq 3$, this corresponds to

$$\mathbb{E}\left[Y_{i2}(0) - Y_{i1}(0) \mid D_i = 1\right] = \mathbb{E}\left[Y_{i2}(0) - Y_{i1}(0) \mid D_i = 0\right]$$

- And if it holds, we might believe

$$\mathbb{E}\left[Y_{i3}(0) - Y_{i2}(0) \mid D_i = 1\right] = \mathbb{E}\left[Y_{i3}(0) - Y_{i2}(0) \mid D_i = 0\right]$$

Testing for Parallel Trends in Staggered Rollout

- Consider staggered rollout with at least one $G_i \in [3, \infty)$
- We run (see `tsvarlist`)

$$Y_{it} = \alpha_i + \gamma_t + \sum_{\ell \in \mathcal{L}} 1 \left[t = G_i + \ell \right] \beta_\ell + \tau D_{it} + \varepsilon_{it}$$

  where $\mathcal{L} = \{-g+1, -g+2, ..., -2\}$ is set of lags
- Test $\beta_\ell = 0$ jointly or individually
- Note we set $\beta_{-1} = 0$ at the outset
  - Avoids multicollinearity (due to $\gamma_t, \beta_\ell, \tau$)
  - Expresses β_ℓ (and τ) relative to the period before treatment
- E.g. in 2×3 block structure with $g = 3$ such that $\mathcal{L} = \{-2\}$:

$$Y_{it} = \alpha_i + \gamma_t + D_i 1 \left[t = 1 \right] \beta_{-2} + \tau D_i P_t + \varepsilon_{it}$$

  and test $\beta_{-2} = 0$; if reject, then there are pre-trends

Pre-Trend Test Example: Duflo (2001)

TABLE 3—MEANS OF EDUCATION AND LOG(WAGE) BY COHORT AND LEVEL OF PROGRAM CELLS

	Years of education			Log(wages)		
	Level of program in region of birth			Level of program in region of birth		
	High (1)	Low (2)	Difference (3)	High (4)	Low (5)	Difference (6)
Panel A: Experiment of Interest						
Aged 2 to 6 in 1974	8.49	9.76	−1.27	6.61	6.73	−0.12
	(0.043)	(0.037)	(0.057)	(0.0078)	(0.0064)	(0.010)
Aged 12 to 17 in 1974	8.02	9.40	−1.39	6.87	7.02	−0.15
	(0.053)	(0.042)	(0.067)	(0.0085)	(0.0069)	(0.011)
Difference	0.47	0.36	0.12	−0.26	−0.29	0.026
	(0.070)	(0.038)	(0.089)	(0.011)	(0.0096)	(0.015)
Panel B: Control Experiment						
Aged 12 to 17 in 1974	8.02	9.40	−1.39	6.87	7.02	−0.15
	(0.053)	(0.042)	(0.067)	(0.0085)	(0.0069)	(0.011)
Aged 18 to 24 in 1974	7.70	9.12	−1.42	6.92	7.08	−0.16
	(0.059)	(0.044)	(0.072)	(0.0097)	(0.0076)	(0.012)
Difference	0.32	0.28	0.034	0.056	0.063	0.0070
	(0.080)	(0.061)	(0.098)	(0.013)	(0.010)	(0.016)

Notes: The sample is made of the individuals who earn a wage. Standard errors are in parentheses.

Limitations of Pre-Trend Testing

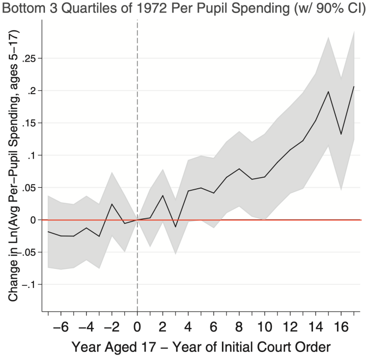


FIGURE I

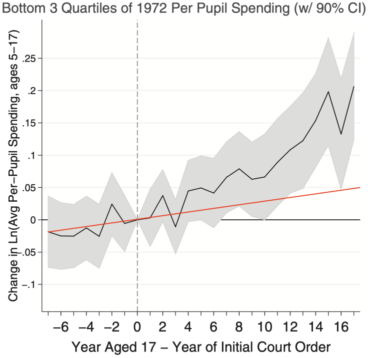


FIGURE I

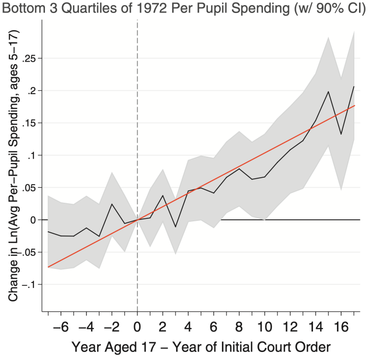
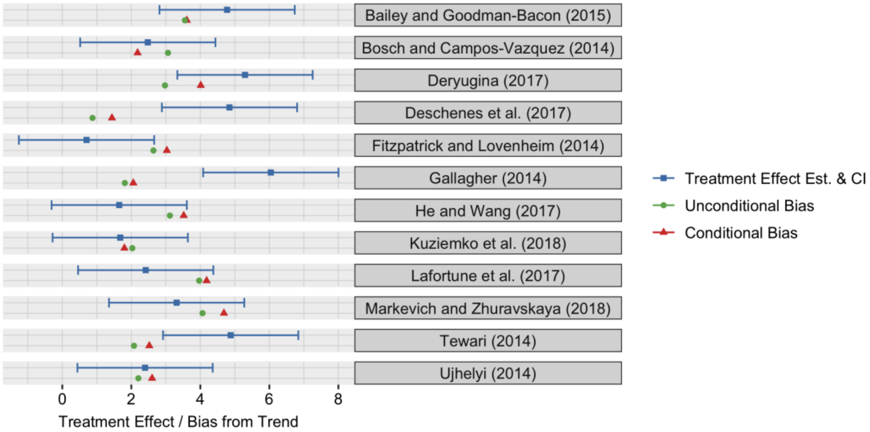


FIGURE I

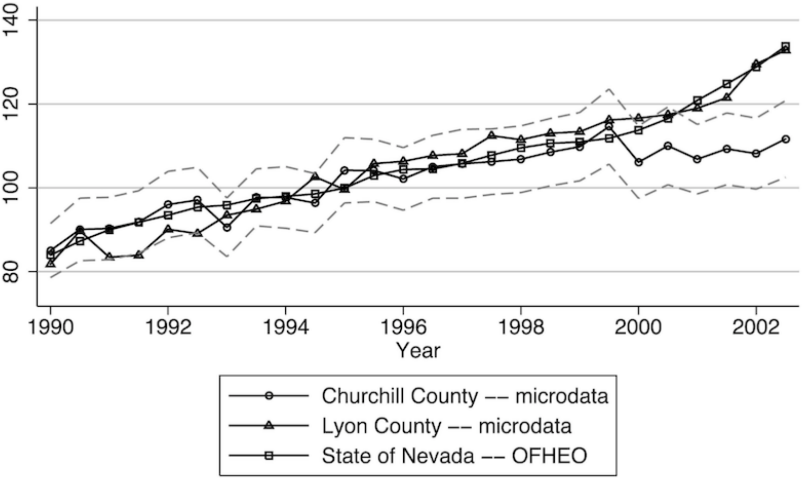
- Example: Jackson et al (2016)
  - Cannot reject zero pre-trend
  - But also cannot reject moderate or strong pre-trend
- How good does pre-trend testing work to detect potential biases?

Roth (2022): Pre-Trend Testing Issues

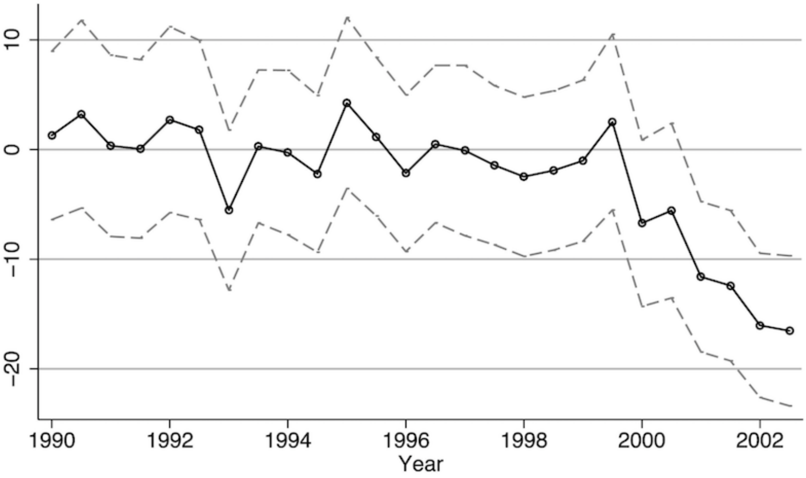
1. Low power: pre-tests struggle to detect meaningful biases from pre-trends
2. Pre-test bias: passing a pre-test can increase bias



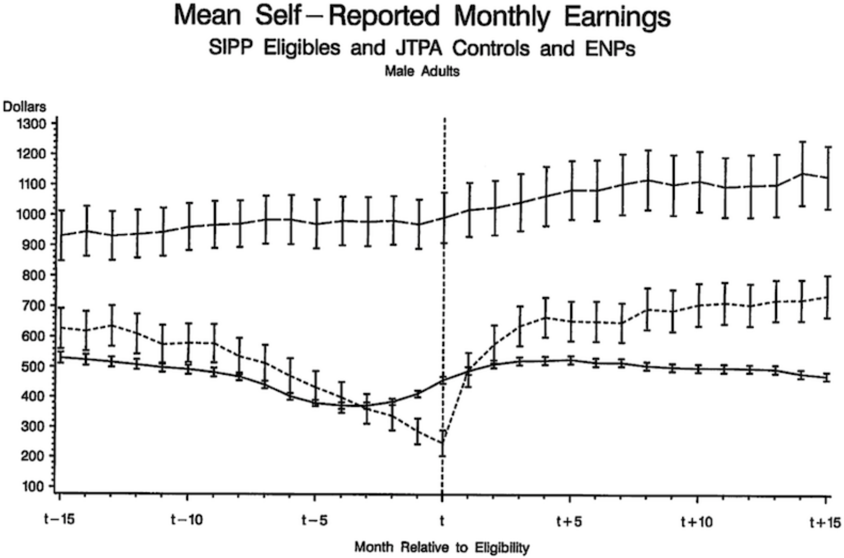
Davis (2004) Example: Raw Data in Block Structure



Davis (2004) Example: Dynamic DID Estimates



Example of Failed Pre-Trend: Ashenfelter's Dip



What if There Are Pre-Trends?

- Parallel trends unlikely to hold in the presence of pre-trends
- Several ways to proceed:
  1. Linear time trends $\kappa_g \times t$, but may absorb part of τ and generally unreliable
  2. Richer fixed effects: e.g. state-by-year FE
  3. IV to correct for pre-trend (Freyaldenhoven et al. 2019)
  4. Set-identify treatment effects (Roth and Rambachan 2022)
  5. Find different design
- Next lecture: new methods for when parallel trends are unlikely to hold
  1. Matched DID
  2. Synthetic control methods
- Making causal claims with pre-trends will meet resistance
- But DID with pre-trend can be descriptively interesting

Famous Despite Pre-Trend: Mass Layoffs

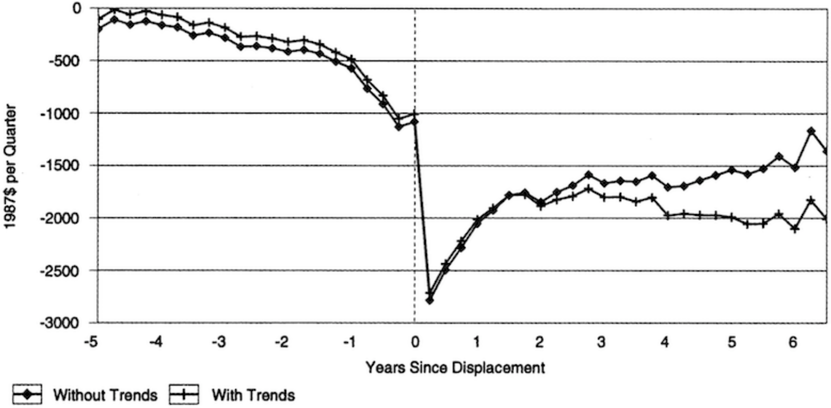


FIGURE 2. EARNINGS LOSSES FOR SEPARATORS IN MASS-LAYOFF SAMPLE

Table of Contents

Group-Dependent Potential Outcomes

- Now interested in *heterogeneous*, *dynamic* ATT
- Need to generalize our PO for treatment timing
  - Define $Y_{it}(g)$ as PO if treated starting in $g \geq 2$ (i.e. ignore always-treated case)
  - E.g. for $T = 3$, there are three counterfactuals for each (i, t):

$$Y_{it}(2), Y_{it})(3), Y_{it}(\infty)$$

    where $Y_{it}\left(\infty\right)$ is the PO for the never-treated
  - Corresponding treatment paths:

$$g = 2 : \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}, \quad g = 3 : \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}, \quad \text{and } g = \infty : \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$$

  - How many POs for unit *i* are there in this example? $T \times |\mathcal{G}| = 3 \times 3 = 9$
- So even for small structures, there are *many* POs

Observed Outcome as Function of Potential Outcomes

- Compare to $Y_{it}(d)$ with $d \in \{0,1\}$
  - Were only able to consider static counterfactuals:

    What if (i,t) had not been treated?

  - But now, can consider dynamic counterfactuals:

    What if (i,t) had been treated earlier/later?

- How do we relate POs to observed outcome?
  - Let $D_i^g = 1\left[G_i = g\right]$
  - Then the observed outcome is given by

$$Y_{it} = \sum_{g \in \mathcal{G}} D_i^g Y_{it}\left(g\right)$$

- Assume *no anticipation*: $Y_{it}\left(g\right) = Y_{it}\left(g'\right)$ for all i; $g, g' > t$

Dynamic Group-Time ATT

- We now define the *dynamic group-time ATT* as

$$\tau_t\left(g\right) \equiv \mathbb{E}\left[Y_{it}\left(g\right) - Y_{it}\left(\infty\right) \middle| G_i = g\right]$$

- $\tau_t\left(g\right)$ formalize heterogeneity along two dimensions:
  1. *Period-specific* heterogeneity: $\tau_t\left(g\right) \neq \tau_{t'}\left(g\right)$
  2. *Group-specific* heterogeneity: $\tau_t\left(g\right) \neq \tau_t\left(g'\right)$
- Consider simple example: $T = 3$ and $G_i \in \{2, 3\}$
  - Period-specific effects might differ: $\tau_2\left(2\right) \neq \tau_3\left(2\right)$

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & {\color{red}1} & {\color{red}1} \end{bmatrix}$$

  - Group-specific effects might differ: $\tau_3\left(2\right) \neq \tau_3\left(3\right)$

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & {\color{red}1} \\ 0 & 1 & {\color{red}1} \end{bmatrix}$$

  - Effects could differ along both dimensions

Table of Contents

Event Study Parameters

- We focus first on *event-specific* heterogeneity
- To this end, define *event study* parameters:

$$\tau_{g+\ell}(g) = \mathbb{E}\left[Y_{ig+\ell}(g) - Y_{ig+\ell}(\infty) \,|\, G_i = g\right]$$

  where ℓ is the number of periods ("lags") since treatment
  - $\ell < 0$: periods before first treatment period
  - $\ell \geq 0$: periods after treatment "switched on"
- In block structures (i.e. single $G_i < \infty$), $\tau_{g+\ell}(g)$ collapses to

$$\tau_\ell \equiv \mathbb{E}\left[Y_{ig+\ell}(g) - Y_{ig+\ell}(\infty) \,|\, D_i = 1\right]$$

- For $T = 5$ and $g = 3$, we would have $\tau_{-2}, \tau_{-1}, \tau_0, \tau_1, \tau_2$
- Due to *no anticipation*: $\tau_\ell = 0$ for $\ell < 0$
- So we are after τ_0, τ_1 and τ_2: *dynamic effects*
- E.g. policy effects in year of enactment; one, two years later

Identification of τ_ℓ in Staggered Rollouts

- No surprise: requires *parallel trends assumption* — assume

$$\mathbb{E}\left[Y_{it}(\infty) - Y_{it-1}(\infty) \,|\, G_i = g\right]$$

  is the same for all $g \in \mathcal{G}$

- Logic:
  - First identify τ_0 using PTA (same as static case)
  - Then identify τ_1 using PTA and τ_0
  - And so forth, up to τ_{T-g+1}

- Relies on *no anticipation*: $Y_{it}(g) = Y_{it}(g')$ for $g, g' > t$
  - Specifically, $Y_{it}(\infty) = Y_{it}(g)$ for $g > t$
  - Which is why this works even if no $G_i = \infty$

Estimation of τ_ℓ in Block Structures

- Construct event-time indicators:
  - Define $P_t^\ell = 1\left[t = g + \ell\right]$ for $\ell \in \{-g+1, ..., T-g\}$, e.g. $\{-2, -1, 0, 1, 2\}$
  - Set $P_t^{-1} = 0$ as reference time (and avoid multicollinearity)
  - So relevant set of event-time indicators is e.g. $\{-2, 0, 1, 2\}$

- Then run *dynamic DID regression*:

$$Y_{it} = \alpha_i + \gamma_t + \sum_{\ell \in \mathcal{L}} D_i P_t^\ell \tau_\ell + \varepsilon_{it}$$

- By including $\ell < 0$ in $\mathcal{L}$, estimate pre-trend
  - Normalization $P_t^{-1} = 0$ equivalent to $\tau_{-1} = 0$
  - $\tau_\ell = 0$ for $\ell < 0$ if *no anticipation* holds
  - Violated by e.g. Ashenfelter's Dip

Matrix Visualization of Dynamic DID Estimation

- What are post indicators picking up? Assume:

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

  and note that (demonstrating on three out of four P_t^ℓ)

$$P_t^{-2}: \begin{bmatrix} \color{blue}{0} & 0 & 0 & 0 & 0 \\ \color{blue}{0} & 0 & 1 & 1 & 1 \end{bmatrix} \quad P_t^{0}: \begin{bmatrix} 0 & 0 & \color{blue}{0} & 0 & 0 \\ 0 & 0 & \color{blue}{1} & 1 & 1 \end{bmatrix} \quad P_t^{1}: \begin{bmatrix} 0 & 0 & 0 & \color{blue}{0} & 0 \\ 0 & 0 & 1 & \color{blue}{1} & 1 \end{bmatrix}$$

- Thus, the interactions $D_i P_t^\ell$ capture (for each τ_ℓ):

$$D_i P_t^{-2}: \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ \color{blue}{0} & \color{green}{0} & 1 & 1 & 1 \end{bmatrix} \quad D_i P_t^{0}: \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \color{green}{0} & \color{blue}{1} & 1 & 1 \end{bmatrix} \quad D_i P_t^{1}: \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \color{green}{0} & 1 & \color{blue}{1} & 1 \end{bmatrix}$$

  i.e. measure treatment mean in ℓ relative to -1

Estimation of τ_ℓ in Staggered Rollout

- Treatment timing differs across groups G_i
  - Varying number of pre- and post-periods per group
- Data determines bounds (ℓ_L, ℓ_U): largest number of pre/post periods
  - Display only subset of pre/post estimates with sufficient precision: judgement call
- Then define event-time treatments:
  - Let $D_{it}^{\ell} = 1\left[t = G_i + \ell\right]$ for $\ell_L < \ell < \ell_U$
  - Again set $D_{it}^{-1} = 0$ as reference and avoid multicollinearity
- Then run *event study regression*:

$$Y_{it} = \alpha_i + \gamma_t + \sum_{\ell \in \mathcal{L}} D_{it}^{\ell} \tau_\ell + \varepsilon_{it}$$

- Note: imposing $\tau_\ell(g) = \tau_\ell$ (homogeneity) for all g
  - Leads to similar issues as τ_{TWFE} if violated
  - In fact, it is also TWFE!

Matrix Visualization of Event Study Estimation

- What are event-study indicators picking up?

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

- Here $(\ell_L, \ell_U) = (-4, 4)$, hence we want $\tau_{-4}, \tau_{-3}, \tau_{-2}, \tau_0, \tau_1, \tau_2, \tau_3, \tau_4$
  - But might want to display only e.g. $\tau_{-2}, \ldots, \tau_2$ for clarity
  - $\tau_{-4}, \tau_{-3}, \tau_{-2}$ are placebo tests; $\tau_0, \ldots \tau_4$ are treatment effect estimates
  - $\tau_{-1} = 0$ is reference category
  - So for example, looking at 3 (of 8) D_{it}^{ℓ}:

D_{it}^{-2} :

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

D_{it}^{0} :

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

D_{it}^{1} :

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

again measure treatment mean in ℓ relative to -1

Event Time Visualization

- Event study is equivalent to the following procedure:
  1. Residualize Y_{it} with respect to α_i and γ_t
  2. Move all treatment paths into *event time* indexed by ℓ:

$$\begin{bmatrix} 0 & 0 & 0 & \color{green}{0} & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & 0 & \color{green}{0} & 1 & 1 & 1 & \cdot \\ \cdot & \cdot & \cdot & \color{green}{0} & 1 & 1 & 1 & 1 \end{bmatrix}$$

  3. Compute means of $\tilde{Y}_{it}$ per event time, e.g. for τ_1 (i.e. D_{it}^1):

$$\begin{bmatrix} 0 & 0 & 0 & \color{green}{0} & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & 0 & \color{green}{0} & 1 & \color{blue}{1} & 1 & \cdot \\ \cdot & \cdot & \cdot & \color{green}{0} & 1 & \color{blue}{1} & 1 & 1 \end{bmatrix}$$

relative to $\color{green}{-1}$

Jensen (2007) Example: Raw Data in Staggered Rollout

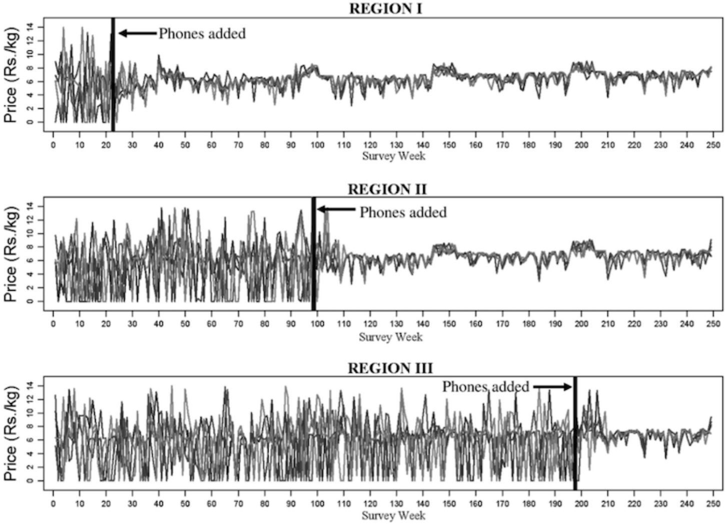


FIGURE IV
Prices and Mobile Phone Service in Kerala

Table of Contents

Elementary Target Parameter

- Recall dynamic group-time ATT as

$$\tau_t(g) = \mathbb{E}\left[Y_{it}(g) - Y_{it}(\infty) \,|\, G_i = g\right]$$

- We have already seen special case τ_ℓ
  - Here we assumed $\tau_\ell(g) = \tau_\ell$ (i.e. homogeneity across groups)
  - But TWFE nature of estimator may lead to bias

- It turns out $\tau_t(g)$ itself is non-parametrically identified
  - Shown in Callaway and Sant'Anna (published in 2021 and 4,538 citations)

- Once we have an estimand $\tau_t(g)$, can aggregate:
  1. By *period*: $\tau_t = \frac{1}{\sum_{g:g\leq t} N_g} \sum_{g:g\leq t} \tau_t(g)$
  2. By *group*: $\tau(g) = \frac{1}{T-g+1} \sum_{t:g\leq t} \tau_t(g)$
  3. By *event time*: $\tau_{\ell,\text{CS}} = \frac{1}{\sum_{g:g+\ell\leq T} N_g} \sum_{g:g+\ell\leq T} \tau_{g+\ell}(g)$

Aggregate Target Parameter

- Imagine want to study minimum wage changes that occurred over 2004-2007
- Want to know effect of MW increases on employment
- These estimators allow us to make comparisons like:
  1. τ_t: 2006 vs 2007 effect of MW increase occurring in 2004
  2. $\tau(g)$: average post-effect of 2004 vs 2006 MW increase
  3. $\tau_\ell(g)$: effect after 3 years of 2004 vs 2006 MW increase

- Can further aggregate $\tau_t(g)$:
  - Mean period effect: $\tau_{\text{period}} = \frac{1}{T}\sum_t \tau_t$
  - Mean group effect: $\tau_{\text{group}} = \frac{1}{|\mathcal{G}\backslash\infty|}\sum_{g:g\neq\infty}\tau(g)$
  - "Simple" effect: $\tau_{\text{simple}} = \frac{1}{\sum_t \sum_{g:g\leq t} N_g}\sum_t \sum_{g:g\leq t}\tau_t(g)$

- Good news! τ_{simple} is unbiased even under heterogeneity

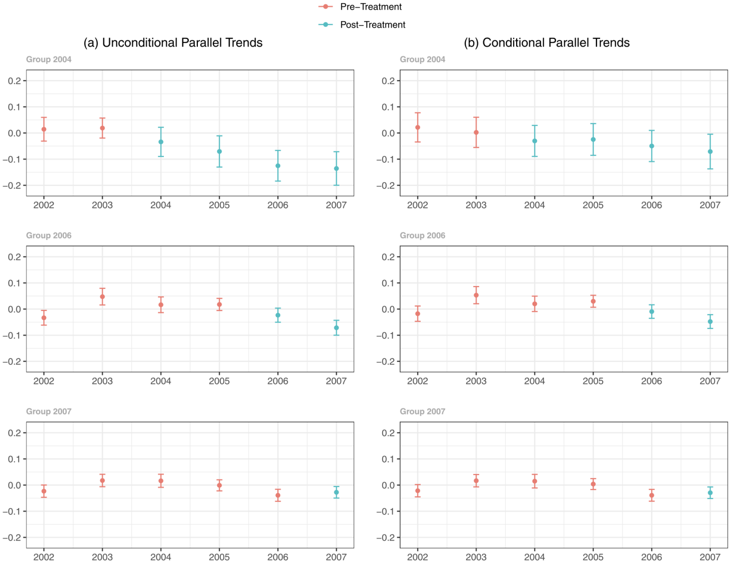
Callaway & Sant'Anna Estimator

- Basic idea of (one of many) estimators:
  - Consider staggered rollout with $\infty \in \mathcal{G}$
  - Then let
  $$\widehat{\tau}_{t,\text{CS}}(g) \equiv (\bar{Y}_{g,t} - \bar{Y}_{\infty,t}) - (\bar{Y}_{g,g-1} - \bar{Y}_{\infty,g-1})$$
  where $\bar{Y}_{gt} = \frac{1}{N_{gt}} \sum_{i:G_i=g} Y_{it}$
- This is the DID that compares outcomes ...
  - ... between t and period right before treatment $g-1$ ...
  - ... for the group first treated in period g ...
  - ... relative to the never-treated group ∞
- For example, for $T = 5, \mathcal{G} = \{3, 4, \infty\}$:

$$\widehat{\tau}_4(4) : \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}; \widehat{\tau}_4(3) : \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}; \widehat{\tau}_3(3) : \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

- Implementation: Run dynamic DID group-by-group, then aggregate

Minimum Wage Example: Estimates of $\tau_t(g)$



Alternative: Imputation Estimator (Borusyak et al 2024)

- Recall parallel trends $\leftrightarrow$ additive separability:

$$Y_{it}(\infty) = \alpha_i + \gamma_t + \varepsilon_{it}$$

  using **only not-yet-treated** observations, i.e. those with $g > t$
- Can use this to infer $\widehat{Y}_{it}(\infty)$ and construct estimate

$$\widehat{\tau}_{it} = Y_{it} - \widehat{Y}_{it}(\infty)$$

  where $Y_{it} = Y_{it}(g)$ for $G_i = g$
- In the example with $T = 5$ and $\mathcal{G} = \{3, 4, \infty\}$:

$$\text{Use } \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot \end{bmatrix} \text{ to infer } \begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \color{red}{0} & \color{red}{0} \\ \cdot & \cdot & \color{red}{0} & \color{red}{0} & \color{red}{0} \end{bmatrix} \text{ compared to } \begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \color{blue}{1} & \color{blue}{1} \\ \cdot & \cdot & \color{blue}{1} & \color{blue}{1} & \color{blue}{1} \end{bmatrix}$$

Imputation Estimator for Block Structures

- Can aggregate $\widehat{\tau}_{it}$ similarly as with the CS estimator
- In general hard to characterize estimator in closed form
- For block structures with $G_i = g$ for all i, it turns out we can write

$$\widehat{\tau}_{t,\text{IE}} = \left(\bar{Y}_{g,t} - \bar{Y}_{g,1:g-1}\right) - \left(\bar{Y}_{\infty,t} - \bar{Y}_{\infty,1:g-1}\right)$$

  where $\bar{Y}_{g,1:g'-1} = \frac{1}{N_{g,1:g'-1}} \sum_{t=1}^{g'-1} Y_{it}$ is mean in g for $t = 1, \ldots, g' - 1$
- For example if $T = 5$ and $\mathcal{G} = \{3, \infty\}$:

$$\widehat{\tau}_3 : \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} ; \widehat{\tau}_4 : \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

- We can see the following tradeoff:
  1. $\widehat{\tau}_{t,\text{IE}}(g)$ uses more data than $\widehat{\tau}_{t,\text{CS}}(g)$: smaller SEs
  2. However, requires stronger parallel trend assumption: for periods $t = 1, \ldots, g - 1$

How to Choose an Estimator: Two Decisions

1. Can I still use a classical dynamic DID or event study estimator?
   - Dynamic DID in block structures are fine: no group het, allowing for time het
   - With staggered rollout, TWFE leads to forbidden comparisons
   - These are generally problematic when lots of large negative weights

2. Which heterogeneous-robust estimator should I use?
   - Besides CS and EI, there are others (Sun and Abraham 2021, Wooldridge 2021)
   - Tradeoff: efficiency versus strength of parallel trends assumption (see also here)
   - Packages exist to implement all these methods (see here)
   - Interpretation in these new estimators can be misleading (see here and here)
   - Typically, these estimators produce similar results (often similar to event studies)
   - Can be useful to build them yourself to have full control

Table of Contents

What is the DID Functional Form (FF) Issue?

- For 2×2, parallel trends is *invariant to transformations* if:

$$\mathbb{E}\left[h\left(Y_{i1}\left(0\right)\right) - h\left(Y_{i0}\left(0\right)\right) \middle| D_i = d\right]$$

  is the same for $d \in \{0, 1\}$ for all strictly monotonic $h\left(\cdot\right)$
- Says: validity of DID assumption does not depend on units

- For example, does PT hold in logs or levels?
  - Consider $i \in \{0, 1\}$ (only two units)

$$\begin{bmatrix} Y_{00}(0) & Y_{01}(0) \\ Y_{10}(0) & Y_{11}(0) \end{bmatrix} = \begin{bmatrix} 1 & e \\ e & e^2 \end{bmatrix} \text{ and log transform: } \begin{bmatrix} 0 & 1 \\ 1 & 2 \end{bmatrix}$$

    where $e \approx 2.72$ is Euler's number
  - Note that $2 - 1 = 1 - 0$ (PT holds in logs)
  - However, $e^2 - e \neq e - 1$ (PT does *not* hold in levels)

When is Parallel Trends Invariant to FF?

Requires parallel trends in CDFs of untreated POs: Define

- $F_{D=d,t=s}^{Y(0)}(y)$: the CDF of $Y_{i,t=s}(0)\,|D_i = d$ (group & time)
- $F_{t=s}^{Y(0)}(y)$: CDF of $Y_{is}(0)$ depending only on time
- $F_{D=d}^{Y(0)}(y)$: CDF of $Y_{it}(0)\,|D_i = d$ depending only on group

Proposition (Roth and Sant'Anna 2021)

PT is invariant to transformation iff

$$F_{D=1,t=1}^{Y(0)}(y) - F_{D=1,t=0}^{Y(0)}(y) = F_{D=0,t=1}^{Y(0)}(y) - F_{D=0,t=0}^{Y(0)}(y)$$

This holds iff

$$F_{D=d,t=s}^{Y(0)}(y) = \theta F_{t=s}^{Y(0)}(y) + (1-\theta)\, F_{D=d}^{Y(0)}(y)$$

for all $y \in \mathbb{R}$ *and* $d, s \in \{0,1\}$, *where* $\theta \in [0,1]$

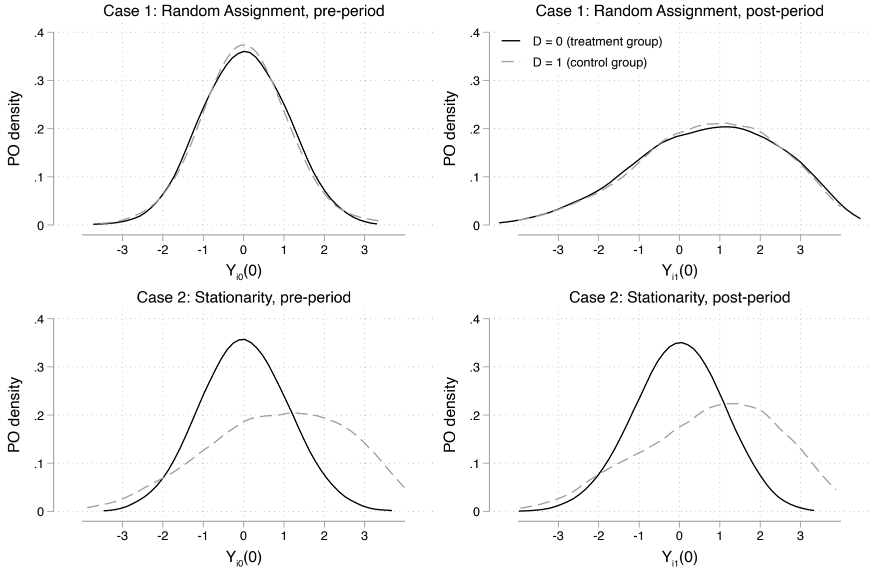
When is Parallel Trends in CDFs Satisfied?

- What does this proposition say?
  - Need strong distributional assumptions for FF *not* to matter
  - These assumptions fall into three specific cases

- The three cases for which PT in CDFs can hold:
  1. *Random assignment* ($\theta = 1$): $F^{Y(0)}_{D=1,t}(y) = F^{Y(0)}_{D=0,t}(y)$
  2. *Stationary* $\mathbf{Y}_i(0)$ ($\theta = 0$): $F^{Y(0)}_{D=d,t=1}(y) = F^{Y(0)}_{D=d,t=0}(y)$
  3. *Partially random and partially stationary* ($\theta \in (0,1)$)

- Remarks:
  - Binary outcomes $Y_{it} \in \{0,1\}$: always invariant to FF
  - PT can hold in logs and levels even if $F^{Y(0)}_{D=0,t=0} \neq F^{Y(0)}_{D=1,t=0}$
  - Can hold even if $\mathbb{E}\left[Y_{i0}(0) \,|D_i = 0\right] \neq \mathbb{E}\left[Y_{i0}(0) \,|D_i = 0\right]$

Visualization of Cases: Monte Carlo Sample of POs



What Are Practical Implications of this Result?

Three important take-aways:

1. If treatment is (as-if) random, PT is invariant of FF

2. If not, then DID is sensitive to FF, unless CDFs are parallel
   - Three cases: as-if random; stationarity; or combination

3. If PT likely to be sensitive to FF, then justify a specific FF
   - For example, context may say $\mathbb{E}\left[\log Y_{it}\left(0\right)|\alpha_{i,}\gamma_{t}\right] = \alpha_i + \gamma_t$
   - Then not necessary to be invariant to FF!