Econometric Analysis of Cross Section and Panel Data

Lecture 9: Difference-in-Differences

Zhian Hu

Fall 2024

This lecture

- DiD has become the single most popular research design in the quantitative social sciences
  - → Simplest form: a group of units is treated at the same time
  - → Staggered form: groups of units are treated at different points in time

This lecture

- DiD has become the single most popular research design in the quantitative social sciences
  - → Simplest form: a group of units is treated at the same time
  - → Staggered form: groups of units are treated at different points in time
- This lecture focuses on
  - → Identifying assumption
  - → Estimation and inference
  - → Recent development

John Snow's cholera hypothesis

- The majority medical opinion about cholera transmission in the nineteenth century was *miasma*.
  → But methods like quarantining the sick were strangely ineffective at slowing down this plague.

John Snow's cholera hypothesis

- The majority medical opinion about cholera transmission in the nineteenth century was *miasma*.
  - → But methods like quarantining the sick were strangely ineffective at slowing down this plague.
- John Snow asked: Is it possible that Cholera was transmitted by water?

John Snow's cholera hypothesis

- The majority medical opinion about cholera transmission in the nineteenth century was *miasma*.
  - → But methods like quarantining the sick were strangely ineffective at slowing down this plague.
- John Snow asked: Is it possible that Cholera was transmitted by water?
- If Snow was a dictator with unlimited wealth and power, how could he test his theory that cholera is waterborne?

John Snow’s cholera hypothesis

- In the 1800s, neighborhoods were served by water companies, which took water from the Thames (polluted by victims’ evacuations via runoff).
- In 1849, the Lambeth water company had moved its intake pipes upstream higher up the Thames, above the main sewage discharge point, thus giving its customers uncontaminated water.

John Snow’s cholera hypothesis

- In the 1800s, neighborhoods were served by water companies, which took water from the Thames (polluted by victims’ evacuations via runoff).
- In 1849, the Lambeth water company had moved its intake pipes upstream higher up the Thames, above the main sewage discharge point, thus giving its customers uncontaminated water.

Company name	1849	1954
Lambeth	85	19

John Snow's cholera hypothesis

- In the 1800s, neighborhoods were served by water companies, which took water from the Thames (polluted by victims' evacuations via runoff).
- In 1849, the Lambeth water company had moved its intake pipes upstream higher up the Thames, above the main sewage discharge point, thus giving its customers uncontaminated water.

Company name	1849	1954
Lambeth	85	19
Southwark and Vauxhall	135	147

John Snow's cholera hypothesis

- In the 1800s, neighborhoods were served by water companies, which took water from the Thames (polluted by victims' evacuations via runoff).
- In 1849, the Lambeth water company had moved its intake pipes upstream higher up the Thames, above the main sewage discharge point, thus giving its customers uncontaminated water.

Company name	1849	1954
Lambeth	85	19
Southwark and Vauxhall	135	147

- A DiD estimate: [19-85]-[147-135]=-78

The simplest case

- We will start a description of DiD in the simplest "canonical" case
- Why? Because recent DiD literature can be viewed as relaxing various components of the canonical model while preserving others

The simplest case

In the canonical DiD model, we have:

- 2 periods: treatment occurs (for some units) in period 2
- Identification of the ATT from parallel trends and no anticipation
- Estimation using sample analogs, equivalent to OLS with TWFE
- A large number of independent observations (or clusters)

Canonical DiD – with math

- Panel data on Y_{it} for $t = 1, 2$ and $i = 1, ..., N$
- **Treatment timing:** Some units $(D_i = 1)$ are treated in period 2; everyone else is untreated $(D_i = 0)$

Canonical DiD – with math

- Panel data on Y_{it} for $t = 1, 2$ and $i = 1, ..., N$
- **Treatment timing:** Some units $(D_i = 1)$ are treated in period 2; everyone else is untreated $(D_i = 0)$
- **Potential outcomes:** Observe $Y_{it}(1) \equiv Y_{it}(0, 1)$ for treated units; and $Y_{it}(0) \equiv Y_{it}(0, 0)$ for comparison

Key identifying assumptions

- **Parallel trends:**

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

Key identifying assumptions

- **Parallel trends:**

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

- **No anticipation:** $Y_{i1}(1) = Y_{i1}(0)$

Key identifying assumptions

- **Parallel trends:**

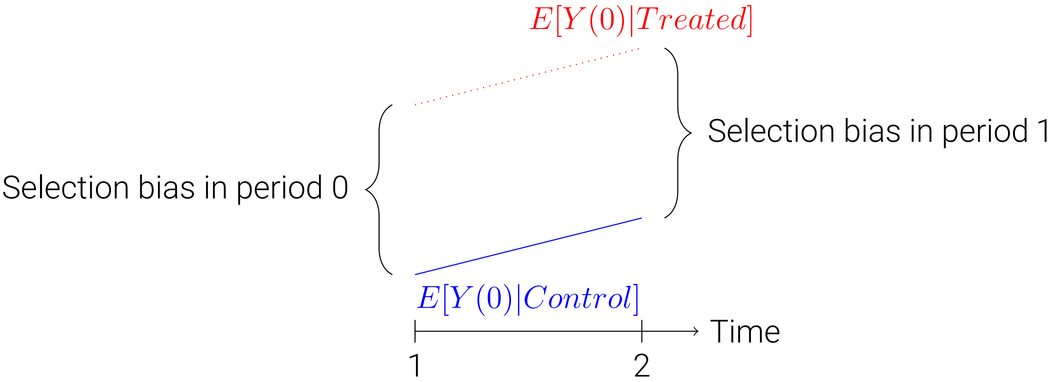
$$\mathbb{E}\left[Y_{i2}(0) - Y_{i1}(0) \mid D_i = 1\right] = \mathbb{E}\left[Y_{i2}(0) - Y_{i1}(0) \mid D_i = 0\right]. \tag{1}$$

- **No anticipation:** $Y_{i1}(1) = Y_{i1}(0)$

  $\rightarrow$ Intuitively, outcome in period 1 isn't affected by treatment status in period 2

  $\rightarrow$ Often left implicit in notation, but important for interpreting DiD estimand as a causal effect in period 2

Visualizing PT



Identification

- **Target parameter:** Average treatment effect on the treated (ATT) in period 2

$$\tau_{ATT} = E[Y_{i2}(1) - Y_{i2}(0)|D_i = 1]$$

Identification

- **Target parameter:** Average treatment effect on the treated (ATT) in period 2

$$\tau_{ATT} = E[Y_{i2}(1) - Y_{i2}(0)|D_i = 1]$$

- Under parallel trends and no anticipation, can show that

$$\tau_{ATT} = \underbrace{(E[Y_{i2}|D_i = 1] - E[Y_{i1}|D_i = 1])}_{\text{Change for treated}} - \underbrace{(E[Y_{i2}|D_i = 0] - E[Y_{i1}|D_i = 0])}_{\text{Change for control}},$$

  a "difference-in-differences" of population means

Proof of identification argument

- Start with

$$E[Y_{i2} - Y_{i1}|D_i = 1] - E[Y_{i2} - Y_{i1}|D_i = 0]$$

Proof of identification argument

- Start with
$$E[Y_{i2} - Y_{i1}|D_i = 1] - E[Y_{i2} - Y_{i1}|D_i = 0]$$
- Apply definition of POs to obtain:

$$E[Y_{i2}(1) - Y_{i1}(1)|D_i = 1] - E[Y_{i2}(0) - Y_{i1}(0)|D_i = 0]$$

Proof of identification argument

- Start with
$$E[Y_{i2} - Y_{i1} | D_i = 1] - E[Y_{i2} - Y_{i1} | D_i = 0]$$
- Apply definition of POs to obtain:
$$E[Y_{i2}(1) - Y_{i1}(1) | D_i = 1] - E[Y_{i2}(0) - Y_{i1}(0) | D_i = 0]$$
- Use No Anticipation to substitute $Y_{i1}(0)$ for $Y_{i1}(1)$:
$$E[Y_{i2}(1) - Y_{i1}(0) | D_i = 1] - E[Y_{i2}(0) - Y_{i1}(0) | D_i = 0]$$

Proof of identification argument

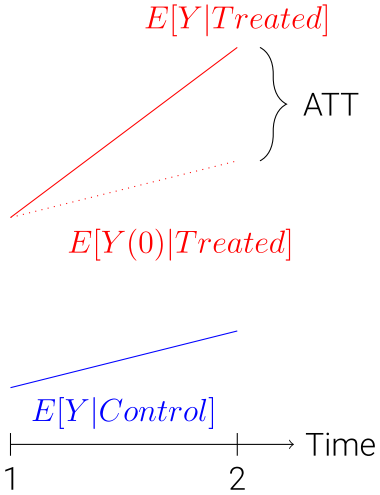
- Start with
$$E[Y_{i2} - Y_{i1}|D_i = 1] - E[Y_{i2} - Y_{i1}|D_i = 0]$$
- Apply definition of POs to obtain:
$$E[Y_{i2}(1) - Y_{i1}(1)|D_i = 1] - E[Y_{i2}(0) - Y_{i1}(0)|D_i = 0]$$
- Use No Anticipation to substitute $Y_{i1}(0)$ for $Y_{i1}(1)$:
$$E[Y_{i2}(1) - Y_{i1}(0)|D_i = 1] - E[Y_{i2}(0) - Y_{i1}(0)|D_i = 0]$$
- Add and subtract $E[Y_{i2}(0)|D_i = 1]$ to obtain:

$$E[Y_{i2}(1) - Y_{i2}(0)|D_i = 1] +$$
$$[(E[Y_{i2}(0)|D_i = 1] - E[Y_{i1}(0)|D_i = 1]) - (E[Y_{i2}(0)|D_i = 0] - E[Y_{i1}(0)|D_i = 0])$$

Proof of identification argument

- Start with
$$E[Y_{i2} - Y_{i1} | D_i = 1] - E[Y_{i2} - Y_{i1} | D_i = 0]$$
- Apply definition of POs to obtain:
$$E[Y_{i2}(1) - Y_{i1}(1) | D_i = 1] - E[Y_{i2}(0) - Y_{i1}(0) | D_i = 0]$$
- Use No Anticipation to substitute $Y_{i1}(0)$ for $Y_{i1}(1)$:
$$E[Y_{i2}(1) - Y_{i1}(0) | D_i = 1] - E[Y_{i2}(0) - Y_{i1}(0) | D_i = 0]$$
- Add and subtract $E[Y_{i2}(0) | D_i = 1]$ to obtain:
$$E[Y_{i2}(1) - Y_{i2}(0) | D_i = 1] +$$
$$[(E[Y_{i2}(0) | D_i = 1] - E[Y_{i1}(0) | D_i = 1]) - (E[Y_{i2}(0) | D_i = 0] - E[Y_{i1}(0) | D_i = 0])]$$
- Cancel the last terms using PT to get $E[Y_{i2}(1) - Y_{i2}(0) | D_i = 1] = \tau_{ATT}$

Visualizing identification



Estimation and inference

- The most conceptually simple estimator replaces population means with sample analogs:

$$\hat{\tau}_{DiD} = (\bar{Y}_{12} - \bar{Y}_{11}) - (\bar{Y}_{02} - \bar{Y}_{01})$$

where $\bar{Y}_{dt}$ is sample mean for group d in period t

Estimation and inference

- The most conceptually simple estimator replaces population means with sample analogs:

$$\hat{\tau}_{DiD} = (\bar{Y}_{12} - \bar{Y}_{11}) - (\bar{Y}_{02} - \bar{Y}_{01})$$

  where $\bar{Y}_{dt}$ is sample mean for group d in period t

- Conveniently, $\hat{\tau}_{DID}$ is algebraically equal to OLS coefficient $\hat{\beta}$ from

$$Y_{it} = \alpha_i + \phi_t + D_{it}\beta + \epsilon_{it}, \tag{2}$$

  where $D_{it} = D_i * 1[t = 2]$. Also equivalent to β from $\Delta Y_i = \alpha + \Delta D_i \beta + u_{it}$.

Estimation and inference

- The most conceptually simple estimator replaces population means with sample analogs:
$$\hat{\tau}_{DiD} = (\bar{Y}_{12} - \bar{Y}_{11}) - (\bar{Y}_{02} - \bar{Y}_{01})$$
where $\bar{Y}_{dt}$ is sample mean for group d in period t
- Conveniently, $\hat{\tau}_{DID}$ is algebraically equal to OLS coefficient $\hat{\beta}$ from
$$Y_{it} = \alpha_i + \phi_t + D_{it}\beta + \epsilon_{it}, \tag{2}$$
where $D_{it} = D_i * 1[t = 2]$. Also equivalent to β from $\Delta Y_i = \alpha + \Delta D_i \beta + u_{it}$.
- **Inference:** And clustered standard errors are valid as number of clusters grows large

Support PT

- Parallel trend assumption, by definition, cannot be fully tested
- Given multiple periods, you can support parallel trend by showing parallel pre-trend
  - → If they had been similar before, then why wouldn't they continue to be in the absence of treatment?

Support PT

- Parallel trend assumption, by definition, cannot be fully tested
- Given multiple periods, you can support parallel trend by showing parallel pre-trend
  - → If they had been similar before, then why wouldn't they continue to be in the absence of treatment?
- One way is to simply show the raw data and just visually inspect whether the pre-treatment dynamics of the treatment group differed from that of the control group units

Support PT

- Galiani et al. (2005)

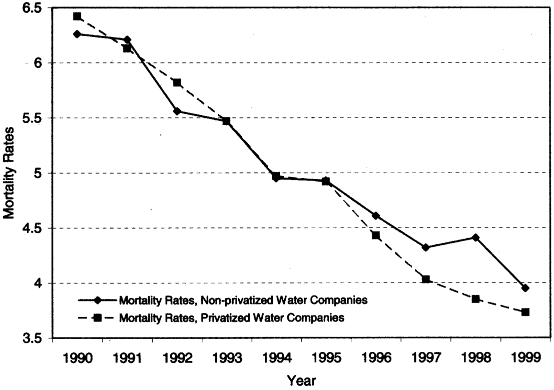


Fig. 1.—Evolution of mortality rates for municipalities with privatized vs. nonprivatized water services

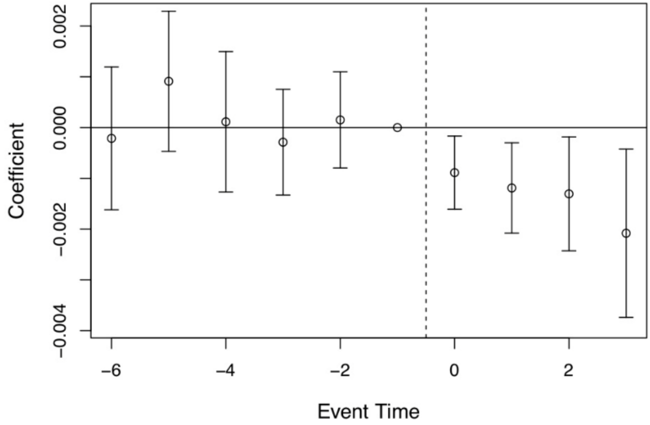
Support PT

- Or you can estimate the differences in each period using an event study

$$Y_{it} = \alpha_i + \phi_t + \sum_{\tau \neq -1} \beta_\tau D_i * T_\tau + \epsilon_{it}, \tag{3}$$

where $T_\tau = 1[t - t^* = \tau]$ and t^* indicates the treatment period

Support PT



Characterizing the recent literature

We can group the recent innovations in DiD lit by which elements of the canonical model they relax:

- **Multiple periods and staggered treatment timing**
- **Relaxing or allowing PT to be violated**
- **Inference with a small number of clusters**

Staggered timing

- Remember that in the canonical DiD model we had:
  - $\rightarrow$ Two periods and a common treatment date
  - $\rightarrow$ Identification from parallel trends and no anticipation
  - $\rightarrow$ A large number of clusters for inference
- A very active recent literature has focused on relaxing the first assumption: **what if there are multiple periods and units adopt treatment at different times?**
- This literature typically maintains the remaining ingredients: parallel trends and many clusters

Overview of staggered timing literature

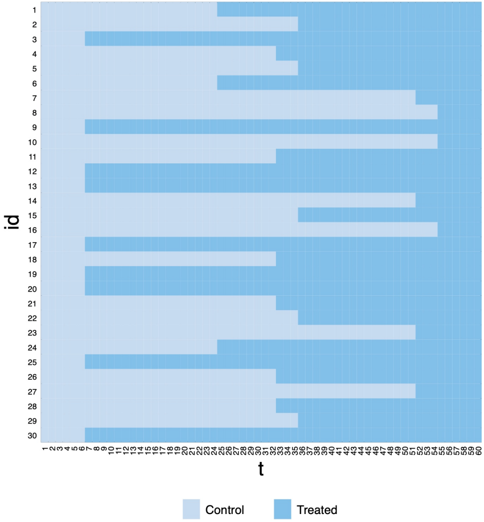
1. Negative results: TWFE OLS doesn't give us what we want with treatment effect heterogeneity
2. New estimators: perform better under treatment effect heterogeneity

Staggered timing set-up

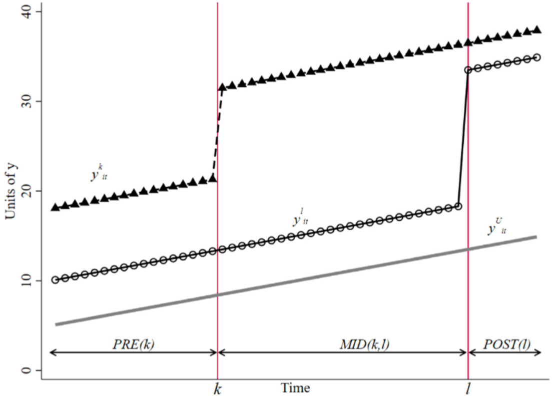
- Panel of observations for periods $t = 1, ..., T$
- Suppose units adopt a binary treatment at different dates $G_i \in \{1, ..., T\} \cup \infty$ (where $G_i = \infty$ means "never-treated")
  - $\rightarrow$ Literature is now starting to consider cases with continuous treatment & treatments that turn on/off – that lit is still developing
- Potential outcomes $Y_{it}(g)$ – depend on time and time you were first-treated

Staggered timing set-up

panelview



Staggered timing set-up



Extending the Identifying Assumptions

- The key identifying assumptions from the canonical model are extended in the natural way

Extending the Identifying Assumptions

- The key identifying assumptions from the canonical model are extended in the natural way

- **Parallel trends:** Intuitively, says that if treatment hadn't happened, all "adoption cohorts" would have parallel average outcomes in all periods

$$E[Y_{it}(\infty) - Y_{i,t-1}(\infty) | G_i = g] = E[Y_{it}(\infty) - Y_{i,t-1}(\infty) | G_i = g'] \text{ for all } g, g', t$$

  Note: can impose slightly weaker versions (e.g. only require PT post-treatment)

Extending the Identifying Assumptions

- The key identifying assumptions from the canonical model are extended in the natural way

- **Parallel trends:** Intuitively, says that if treatment hadn't happened, all "adoption cohorts" would have parallel average outcomes in all periods

$$E[Y_{it}(\infty) - Y_{i,t-1}(\infty) | G_i = g] = E[Y_{it}(\infty) - Y_{i,t-1}(\infty) | G_i = g'] \text{ for all } g, g', t$$

  Note: can impose slightly weaker versions (e.g. only require PT post-treatment)

- **No anticipation:** Intuitively, says that treatment has no impact before it is implemented

$$Y_{it}(g) = Y_{it}(\infty) \text{ for all } t < g$$

Negative results

- Suppose we again run the regression

$$Y_{it} = \alpha_i + \phi_t + D_{it}\beta + \epsilon_{it},$$

  where $D_{it} = 1[t \geq G_i]$ is a treatment indicator.
- Suppose we're willing to assume no anticipation and parallel trends across all adoption cohorts as described above

Negative results

- Suppose we again run the regression

$$Y_{it} = \alpha_i + \phi_t + D_{it}\beta + \epsilon_{it},$$

  where $D_{it} = 1[t \geq G_i]$ is a treatment indicator.
- Suppose we're willing to assume no anticipation and parallel trends across all adoption cohorts as described above
- Good news: if treatment effects are constant across time and units, $Y_{it}(g) - Y_{it}(\infty) \equiv \tau$, then $\beta = \tau$

Negative results

- Suppose we again run the regression

$$Y_{it} = \alpha_i + \phi_t + D_{it}\beta + \epsilon_{it},$$

  where $D_{it} = 1[t \geq G_i]$ is a treatment indicator.
- Suppose we're willing to assume no anticipation and parallel trends across all adoption cohorts as described above
- Good news: if treatment effects are constant across time and units, $Y_{it}(g) - Y_{it}(\infty) \equiv \tau$, then $\beta = \tau$

异质性处理效应双向固定效应模型存在偏误

- Bad news: if treatment effects are heterogeneous, then β may put negative weights on treatment effects for some units and time periods
  - $\rightarrow$ E.g., if treatment effect depends on time since treatment, $Y_{it}(t-r) - Y_{it}(\infty) = \tau_r$, then some τ_rs may get negative weight

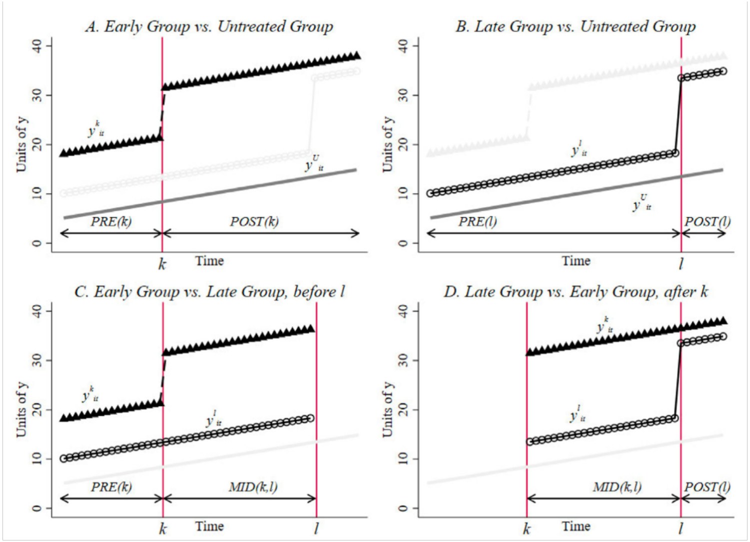
Where do these negative results come from?

- The intuition for these negative results is that the TWFE OLS specification combines two sources of comparisons:

    1. **Clean comparisons:** DiD's between treated and not-yet-treated units
    2. **Forbidden comparisons:** DiD's between two sets of already-treated units (who began treatment at different times)

- These forbidden comparisons can lead to negative weights: the "control group" is already treated, so we run into problems if their treatment effects change over time

Bacon decomposition



forbidden

Some intuition for forbidden comparisons

- Consider the two period model, except suppose now that our two groups are **always-treated** units (treated in both periods) and **switchers** (treated only in period 2)
- With two periods, the coefficient β from $Y_{it} = \alpha_i + \phi_t + D_{it}\beta + \epsilon_{it}$ is the same as from the first-differenced regression $\Delta Y_i = \alpha + \Delta D_i \beta + u_i$

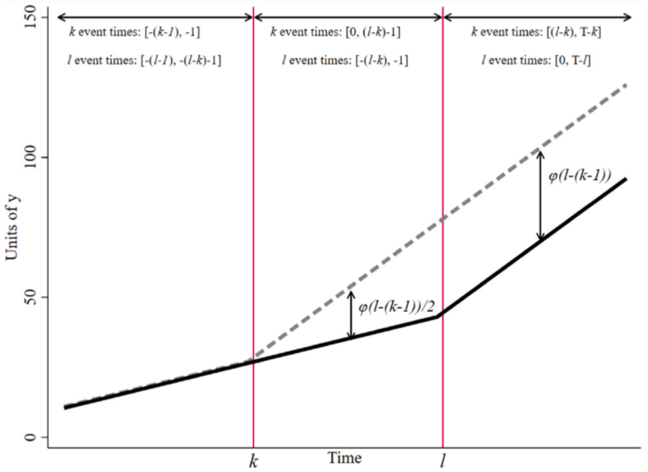
Some intuition for forbidden comparisons

- Consider the two period model, except suppose now that our two groups are **always-treated** units (treated in both periods) and **switchers** (treated only in period 2)
- With two periods, the coefficient β from $Y_{it} = \alpha_i + \phi_t + D_{it}\beta + \epsilon_{it}$ is the same as from the first-differenced regression $\Delta Y_i = \alpha + \Delta D_i \beta + u_i$
- Observe that ΔD_i is one for switchers and zero for stayers.
  That is, the stayers are the control group! Thus,

$$\hat{\beta} = \underbrace{\left(\bar{Y}_{Switchers,2} - \bar{Y}_{Switchers,1}\right)}_{\text{Change for switchers}} - \underbrace{\left(\bar{Y}_{AT,2} - \bar{Y}_{AT,1}\right)}_{\text{Change for always treated}}$$

- Problem: if the treatment effect for the always-treated grows over time, that will enter $\hat{\beta}$ negatively!

Some intuition for forbidden comparisons



Second intuition for negative weights

- The Frisch-Waugh-Lovell theorem says that we can obtain the coefficient β in $Y_{it} = \alpha_i + \phi_t + D_{it}\beta + \epsilon_{it}$ by the following two-step procedure.

Second intuition for negative weights

- The Frisch-Waugh-Lovell theorem says that we can obtain the coefficient β in $Y_{it} = \alpha_i + \phi_t + D_{it}\beta + \epsilon_{it}$ by the following two-step procedure.

- First, regress the treatment indicator D_{it} on the FEs (a linear probability model): $D_{it} = \tilde{\alpha}_i + \tilde{\phi}_t + \tilde{\epsilon_{it}}$

- Then run a univariate regression of Y_{it} on $D_{it} - \hat{D}_{it}$ to obtain β.
  $\rightarrow$ Thus, $\beta = \frac{Cov(Y_{it}, D_{it} - \hat{D}_{it})}{Var(D_{it} - \hat{D}_{it})} = \frac{E(Y_{it}(D_{it} - \hat{D}_{it}))}{Var(D_{it} - \hat{D}_{it})}$

Second intuition for negative weights

- The Frisch-Waugh-Lovell theorem says that we can obtain the coefficient β in $Y_{it} = \alpha_i + \phi_t + D_{it}\beta + \epsilon_{it}$ by the following two-step procedure.

- First, regress the treatment indicator D_{it} on the FEs (a linear probability model): $D_{it} = \tilde{\alpha}_i + \tilde{\phi}_t + \tilde{\epsilon_{it}}$

- Then run a univariate regression of Y_{it} on $D_{it} - \hat{D}_{it}$ to obtain β.
  $\rightarrow$ Thus, $\beta = \frac{Cov(Y_{it}, D_{it} - \hat{D}_{it})}{Var(D_{it} - \hat{D}_{it})} = \frac{E(Y_{it}(D_{it} - \hat{D}_{it}))}{Var(D_{it} - \hat{D}_{it})}$

- However, it's well known that the linear probability model for D_{it} may have predictions outside the unit interval. If $\hat{D}_{it} > 1$ even though unit i is treated in period t, then $D_{it} - \hat{D}_{it} < 0$, and thus Y_{it} gets negative weight.

Not just negative but weird...

The literature has placed a lot of emphasis on the fact that some treatment effects may get negative weights

- But even if the weights are non-negative, they might not give us the most intuitive parameter

- For example, suppose each unit i has treatment effect τ_i in every period if they are treated (no dynamics). Then β gives a weighted average of the τ_i where the weights are largest for units treated closest to the middle of the panel

- It is not obvious that these weights are relevant for policy, even if they are all non-negative!

Issues with dynamic TWFE

- Sun and Abraham (2021) show that similar issues arise with dynamic TWFE specifications:

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{k \neq -1} \gamma_k D_{i,t}^k + \varepsilon_{i,t},$$

where $D_{i,t}^k = 1\{t - G_i = k\}$ are "event-time" dummies.

- Like for the static spec, γ_k may put negative weight on treatment effects after k periods for some units

- SA also show that γ_k may be "contaminated" by treatment effects at lags $k' \neq k$

Dynamic TWFE - Continued

- The results in SA suggest that interpreting the $\hat{\gamma}_k$ for $k = 1, 2, ...$ as estimates of the dynamic effects of treatment may be misleading
- These results also imply that pre-trends tests of the γ_k for $k < 0$ may be misleading – could be non-zero even if parallel trends holds, since they may be "contaminated" by post-treatment effects!

Dynamic TWFE - Continued

- The results in SA suggest that interpreting the $\hat{\gamma}_k$ for $k = 1, 2, \ldots$ as estimates of the dynamic effects of treatment may be misleading
- These results also imply that pre-trends tests of the γ_k for $k < 0$ may be misleading – could be non-zero even if parallel trends holds, since they may be "contaminated" by post-treatment effects!
- The issues discussed in SA arise if dynamic path of treatment effects is heterogeneous across adoption cohorts
  - $\rightarrow$ Biases may be less severe than for "static" specs if dynamic patterns are similar across cohorts

New estimators (and estimands!)

- Several new (closely-related) estimators have been proposed to try to address these negative weighting issues
- The key components of all of these are:
  1. Be precise about the target parameter (estimand) – i.e., how do we want to aggregate treatment effects across time/units
  2. Estimate the target parameter using only "clean-comparisons"

Example – Callaway and Sant'Anna (2020)

- Define $ATT(g, t)$ to be ATT in period t for units first treated at period g,

$$ATT(g, t) = E[Y_{it}(g) - Y_{it}(\infty) | G_i = g]$$

Example – Callaway and Sant'Anna (2020)

- Define $ATT(g,t)$ to be ATT in period t for units first treated at period g,

$$ATT(g,t) = E[Y_{it}(g) - Y_{it}(\infty) | G_i = g]$$

- Under PT and No Anticipation, $ATT(g,t)$ is identified as

$$ATT(g,t) = \underbrace{E[Y_{it} - Y_{i,g-1} | G_i = g]}_{\text{Change for cohort g}} - \underbrace{E[Y_{it} - Y_{i,g-1} | G_i = \infty]}_{\text{Change for never-treated units}}$$

- Why?

Example – Callaway and Sant'Anna (2020)

- Define $ATT(g,t)$ to be ATT in period t for units first treated at period g,

$$ATT(g,t) = E[Y_{it}(g) - Y_{it}(\infty)|G_i = g]$$

- Under PT and No Anticipation, $ATT(g,t)$ is identified as

$$ATT(g,t) = \underbrace{E[Y_{it} - Y_{i,g-1}|G_i = g]}_{\text{Change for cohort g}} - \underbrace{E[Y_{it} - Y_{i,g-1}|G_i = \infty]}_{\text{Change for never-treated units}}$$

- Why? This is a two-group two-period comparison, so the argument is the same as in the canonical case!

Proof of identification argument

- Start with
$$E[Y_{it} - Y_{i,g-1} | G_i = g] - E[Y_{it} - Y_{i,g-1} | G_i = \infty]$$

Proof of identification argument

- Start with
$$E[Y_{it} - Y_{i,g-1} | G_i = g] - E[Y_{it} - Y_{i,g-1} | G_i = \infty]$$
- Apply definition of POs to obtain:

$$E[Y_{it}(g) - Y_{i,g-1}(g) | G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty) | G_i = \infty]$$

Proof of identification argument

- Start with
$$E[Y_{it} - Y_{i,g-1}|G_i = g] - E[Y_{it} - Y_{i,g-1}|G_i = \infty]$$
- Apply definition of POs to obtain:
$$E[Y_{it}(g) - Y_{i,g-1}(g)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]$$
- Use No Anticipation to substitute $Y_{i,g-1}(\infty)$ for $Y_{i,g-1}(g)$:
$$E[Y_{it}(g) - Y_{i,g-1}(\infty)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]$$

Proof of identification argument

- Start with
$$E[Y_{it} - Y_{i,g-1}|G_i = g] - E[Y_{it} - Y_{i,g-1}|G_i = \infty]$$
- Apply definition of POs to obtain:
$$E[Y_{it}(g) - Y_{i,g-1}(g)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]$$
- Use No Anticipation to substitute $Y_{i,g-1}(\infty)$ for $Y_{i,g-1}(g)$:
$$E[Y_{it}(g) - Y_{i,g-1}(\infty)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]$$
- Add and subtract $E[Y_{it}(\infty)|G_i = g]$ to obtain:

$$E[Y_{it}(g) - Y_{it}(\infty)|G_i = g] +$$
$$[E[Y_{it}(\infty) - Y_{i,g-1}(\infty)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]]$$

Proof of identification argument

- Start with
$$E[Y_{it} - Y_{i,g-1}|G_i = g] - E[Y_{it} - Y_{i,g-1}|G_i = \infty]$$
- Apply definition of POs to obtain:
$$E[Y_{it}(g) - Y_{i,g-1}(g)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]$$
- Use No Anticipation to substitute $Y_{i,g-1}(\infty)$ for $Y_{i,g-1}(g)$:
$$E[Y_{it}(g) - Y_{i,g-1}(\infty)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]$$
- Add and subtract $E[Y_{it}(\infty)|G_i = g]$ to obtain:
$$E[Y_{it}(g) - Y_{it}(\infty)|G_i = g] +$$
$$[E[Y_{it}(\infty) - Y_{i,g-1}(\infty)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]]$$
- Cancel the last term using PT to get $E[Y_{it}(g) - Y_{it}(\infty)|G_i = g] = ATT(g,t)$

Example – Callaway and Sant'Anna (2020)

- Define $ATT(g,t)$ to be ATT in period t for units first treated at period g,

$$ATT(g,t) = E[Y_{it}(g) - Y_{it}(\infty)|G_i = g]$$

- Under PT and No Anticipation,

$$ATT(g,t) = \underbrace{E[Y_{it} - Y_{i,g-1}|G_i = g]}_{\text{Change for cohort g}} - \underbrace{E[Y_{it} - Y_{i,g-1}|G_i = \infty]}_{\text{Change for never-treated}}$$

Example – Callaway and Sant'Anna (2020)

- Define $ATT(g,t)$ to be ATT in period t for units first treated at period g,

$$ATT(g,t) = E[Y_{it}(g) - Y_{it}(\infty) | G_i = g]$$

- Under PT and No Anticipation,

$$ATT(g,t) = \underbrace{E[Y_{it} - Y_{i,g-1} | G_i = g]}_{\text{Change for cohort g}} - \underbrace{E[Y_{it} - Y_{i,g-1} | G_i = \infty]}_{\text{Change for never-treated}}$$

- We can then estimate this with sample analogs:

$$\widehat{ATT}(g,t) = \underbrace{\widehat{E}[Y_{it} - Y_{i,g-1} | G_i = g]}_{\text{Sample change for cohort g}} - \underbrace{\widehat{E}[Y_{it} - Y_{i,g-1} | G_i = \infty]}_{\text{Sample change for never-treated}}$$

Aggregation schemes

- If have a large number of observations and relatively few groups/periods, can report $\widehat{ATT}(g,t)$'s directly.
- If there are many groups/periods, the $\widehat{ATT}(g,t)$ may be very imprecisely estimated and/or too numerous to report concisely

Aggregation schemes

- In these cases, it is often desirable to report sensible averages of the $\widehat{ATT}(g,t)$'s.

Aggregation schemes

- In these cases, it is often desirable to report sensible averages of the $\widehat{ATT}(g,t)$'s.
- One of the most useful is to report event-study parameters which aggregate $\widehat{ATT}(g,t)$'s at a particular lag since treatment
  - $\rightarrow$ E.g. $\hat{\theta}_k = \sum_g \widehat{ATT}(g, g+k)$ aggregates effects for cohorts in the kth period after treatment
  - $\rightarrow$ Can also construct for $k < 0$ to estimate "pre-trends"

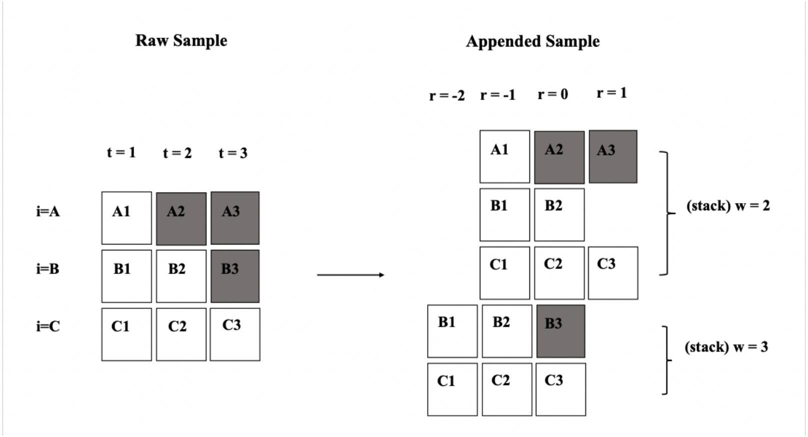
Aggregation schemes

- In these cases, it is often desirable to report sensible averages of the $\widehat{ATT}(g,t)$'s.
- One of the most useful is to report event-study parameters which aggregate $\widehat{ATT}(g,t)$'s at a particular lag since treatment
  - → E.g. $\hat{\theta}_k = \sum_g \widehat{ATT}(g, g+k)$ aggregates effects for cohorts in the kth period after treatment
  - → Can also construct for $k < 0$ to estimate "pre-trends"
- C&S discuss other sensible aggregations too – e.g., if interested in whether treatment effects differ across good/bad economies, may want to "calendar averages" that pool the $\widehat{ATT}(g,t)$ for the same year

Comparisons of new estimators

- Callaway and Sant'Anna also propose an analogous estimator using *not-yet-treated* rather than never-treated units.
- Sun and Abraham (2021) propose a similar estimator but with different comparisons groups (e.g. using last-to-be treated rather than not-yet-treated)
- Borusyak et al. (2024), Wooldridge (2021), Gardner (2021) propose "imputation" estimators that estimate the counterfactual $\hat{Y}_{it}(0)$ using a TWFE model that is fit using only pre-treatment data
  - → Main difference from C&S is that this uses more pre-treatment periods, not just period $g-1$
  - → This can sometimes be more efficient (if outcome not too serially correlated), but also relies on a stronger PT assumption that may be more susceptible to bias

Stacked DiD



Raw Sample
Appended Sample
r = -2
r = -1
r = 0
r = 1
t = 1
t = 2
t = 3
i=A
i=B
i=C
A1
A2
A3
B1
B2
B3
C1
C2
C3
(stack) w = 2
(stack) w = 3

Personal advice

- Don’t freak out about this new literature!

Personal advice

- Don't freak out about this new literature!
- In most cases, using the "new" DiD methods will not lead to a big change in your results (empirically, TE heterogeneity is not *that* large in most cases)
  - → The exceptions are cases where there are periods where almost all units are treated – this is when "forbidden comparisons" get the most weight

Personal advice

- Don't freak out about this new literature!
- In most cases, using the "new" DiD methods will not lead to a big change in your results (empirically, TE heterogeneity is not *that* large in most cases)
  - → The exceptions are cases where there are periods where almost all units are treated – this is when "forbidden comparisons" get the most weight
- The most important thing is to be precise about who you want the comparison group to be and to choose a method that only uses these "clean comparisons"

Personal advice

- Don't freak out about this new literature!
- In most cases, using the "new" DiD methods will not lead to a big change in your results (empirically, TE heterogeneity is not *that* large in most cases)
  - → The exceptions are cases where there are periods where almost all units are treated – this is when "forbidden comparisons" get the most weight
- The most important thing is to be precise about who you want the comparison group to be and to choose a method that only uses these "clean comparisons"
- In my experience, the difference between the new estimators is typically not that large – can report multiple new methods for robustness (to make your referees happy!)

References I

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting Event-Study Designs: Robust and Efficient Estimation," *The Review of Economic Studies*, 2024, *91* (6), 3253–3285.

Galiani, Sebastian, Paul Gertler, and Ernesto Schargrodsky, "Water for Life: The Impact of the Privatization of Water Services on Child Mortality," *Journal of Political Economy*, 2005.

Gardner, John, "Two-stage differences in differences," *Working Paper*, 2021.

Sun, Liyang and Sarah Abraham, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, 2021, *225* (2), 175–199.

References II

Wooldridge, Jeffrey M, “Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators,” *Working Paper*, 2021, pp. 1–89.