Difference-in-Differences: Theory

Huan Deng

April 29, 2025

Outline

Introduction

Potential Outcome Framework

The Canonical DID Model

Interpreting TWFE Under Staggered DID

The Bottom-Up/Forward-Engineering Approach

- Introduction
- Callaway and Sant'Anna (2021)
- Imputation-based Methods

Introduction

Overview

Today, we will consider causal inference for panel data and discuss the difference-in-differences (DID) research design.

1. We need to extend the potential outcome framework to this data structure
2. The recent progress on this topic extends the canonical DID to multiple periods and varying treatment timing
3. Two way fixed effect (TWFE) is an algorithm that aggregates causal estimands into a higher level "summary statistic"
   - The aggregation might involve negative weights
   - The aggregated estimate might not be policy-relevant or interesting
4. Various identification assumptions that differ in nature: design-wise? Functional form? Behavioral restrictions?

Overview

1. To facilitate discussion, we will first extend the potential outcome framework following Athey and Imbens (2022)
2. Then we will discuss the canonical DID model because it sercves as a building block or benchmark for later researches.
3. Then we will discuss the very popular model: Two Way Fixed Effect model (TWFE)
   - When does TWFE work?
   - When TWFE can be problematic?
   - What can we do if TWFE is not delivering a satisfactory causal estimate?
4. This literature is going really fast and seems daunting...
5. Don't panic!

Potential Outcome Framework

Setup: Potential Outcomes and Adoption Date

- Consider a panel data where there are N units, and T time periods.
- Now the potential outcome is defined over the **entire path of treatment status**: $Y_{it}(D_{i1}, D_{i2}, ..., D_{iT})$
- Suppose D_{it} is binary, then for a given individual at a specific time t, there are 2^T possible potential outcomes
- $Y_{it}(D_{i1}, D_{i2}, ..., D_{iT})$: the potential outcome of individual i at time t for the treatment path of $\{D_{i1}, D_{i2}, ..., D_{iT}\}$
- Remember that we define causal effect just as the comparison of potential outcomes, but now we have much more comparisons to make: for all t, and all pairs of treatment paths.

Setup: Potential Outcomes and Adoption Date

- In most of the applications we consider, the treatment is an absorbing state: once treated, stay treated forever. We make this assumption throughout the whole lecture.
- This greatly simplifies the space of the treatment paths: from 2^T to $T+1$
- Now the potential outcome is defined by the **adoption date** $A_i \in \{1, \ldots, T, \infty\}$ for unit i, where ∞ indicates “never treated”.
- $Y_{it}(a)$: the potential outcome for individual i at time t if the adoption time is a

Treatment Adoption Matrix

Definition: Define $W : A \times T \to \{0,1\}$ as the binary indicator for the adoption date a preceding time t:

$$W_{it} = W(A_i, t) = \mathbf{1}_{a_i \leq t}. \tag{1}$$

The $N \times T$ matrix $\mathbf{W}$ represents adoption as:

$$\mathbf{W}_{N \times T} = \begin{bmatrix} 0 & 0 & 0 & 0 & \dots & 0 & \text{(never adopter)} \\ 0 & 0 & 0 & 0 & \dots & 1 & \text{(late adopter)} \\ . & . & . & . & \dots & . & \\ 0 & 0 & 1 & 1 & \dots & 1 & \text{(medium adopter)} \\ 0 & 1 & 1 & 1 & \dots & 1 & \text{(early adopter)} \end{bmatrix}.$$

Adoption Distribution

- Define $N_a = \sum_{i=1}^{N} \mathbf{1}_{A_i = a}$ as the number of units adopting at a.
- The fraction of units adopting at a:

$$\pi_a = \frac{N_a}{N}, \quad a \in A. \tag{2}$$

- The cumulative fraction adopting by period t:

$$\Pi_t = \sum_{s=1}^{t} \pi_s, \quad t \in T. \tag{3}$$

Potential Outcomes and Causal Effects

Population potential outcome for period t given adoption date a:

$$\bar{Y}_t(a) = \frac{1}{N}\sum_{i=1}^{N} Y_{it}(a), \quad t \in T, a \in A. \tag{4}$$

Unit-level causal effect of adoption at a' versus a:

$$\tau_{it,aa'} = Y_{it}(a') - Y_{it}(a). \tag{5}$$

Average causal effect:

$$\tau_{t,aa'} = \frac{1}{N}\sum_{i=1}^{N} \big(Y_{it}(a') - Y_{it}(a)\big) = \bar{Y}_t(a') - \bar{Y}_t(a). \tag{6}$$

Alternative Building Blocks

- By varying a or t, we can define many building-block causal estimands.
- Sun and Abraham (2021):

$$CATT_{a,s} = E[\tau_{ia+s,\infty a} | A_i = a]. \tag{7}$$

- Callaway and Sant'Anna (2021):

$$E[\tau_{it,\infty a} | A_i = a] \tag{8}$$

Two Treatment Patterns

Block Adoption:

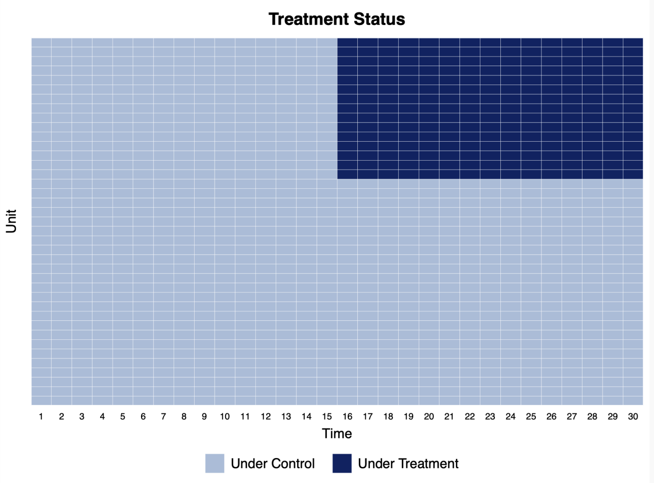
- $A_i = A$ for all i
- Each unit i has an adoption indicator $D_i \in \{0, 1\}$, so $W_{it} = D_i \, \mathbf{1}\{t > A\}$.

Staggered Adoption:

- A_i can vary
- Treatment for unit i in period t is $W_{it} = \mathbf{1}\{t > A_i\}$.

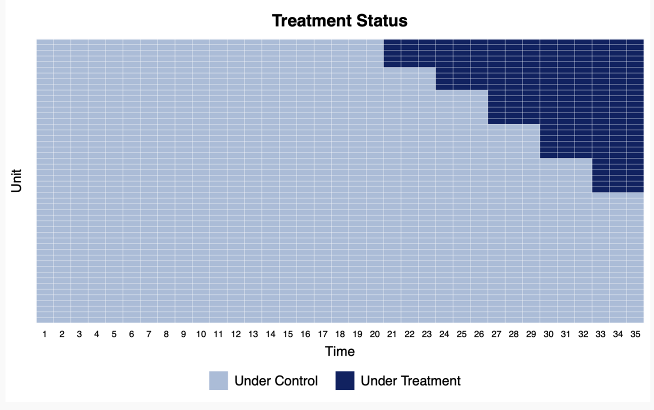
Graphic Illustrations of Treatment Patterns (from Xu's Method Lecture)

Block Adoption



Graphic Illustrations of Treatment Patterns (from Xu's Method Lecture)

Staggered Adoption



Three Sets of Assumptions

- As before, identification of causal effects require assumptions.
- Athey and Imbens (2022) discuss three sets of assumptions:
  - About design: how the adoption date is assigned
  - About potential outcomes: ruling out the presence of certain treatment effects
  - About heterogeneity in treatment effects

Assumption 1: Random Adoption Date

Definition: For some set of positive integers N_a, for $a \in A$, the probability of adoption follows:

$$\mathrm{pr}(\mathbf{A} = \mathbf{a}) = \left(\frac{N!}{\prod_{a \in A} N_a!} \right)^{-1}, \tag{9}$$

for all N-vectors $\mathbf{a}$ such that for all $a \in A$,

$$\sum_{i=1}^{N} \mathbf{1}_{a_i = a} = N_a. \tag{10}$$

This is a design-based perspective, analogous to a completely randomized experiment but for adoption dates.

Assumption 2: No Anticipation

No Anticipation: For $a > t$, $Y_{it}(a) = Y_{it}(\infty)$.

- If the unit has not yet adopted by time t, future adoption date does not affect current outcomes.
- Rules out anticipation effects.
- This assumption is key to many identification results of the DID design
- Which date should be encoded as the adoption date? The date the policy was announced or the date the policy was actually in effect?
- Suppose a new natural resource has been found, so the government announces a permanent tax reduction effective in two years, by permanent income hypothesis, today's consumption should immediately increase.
- We need to argue that the policy comes as a surprise.

Assumption 3: Invariance to History

Invariance to History: For $a \leq t$, $Y_{it}(a) = Y_{it}(1)$.

- If a unit has adopted by time t, the outcome at time t depends only on being treated, not on *when* it was adopted.
- The time exposure to the treatment doesn't matter.
- I bet many labor economists or health economists don't like this assumption.

Additional Assumptions (4 – 7)

- **Assumption 4: Constant Treatment Effects over Units**
  For all i, j, t, a, a', $Y_{it}(a') - Y_{it}(a) = Y_{jt}(a') - Y_{jt}(a)$.
- **Assumption 5: Constant Treatment Effects over Time**
  $Y_{it}(1) - Y_{it}(\infty)$ is the same for all t (often goes together with Assumptions 2 and 3).
- **Assumption 6: Random Sampling**
  We view $(A_i, Y_{it}(a))$ as drawn i.i.d. from a large population.
- **Assumption 7: Additivity**
  $E[Y_{it}(\infty)] = \alpha_i + \beta_t$ (two-way additive structure for untreated potential outcomes, which is closely related to Parallel Trend Assumption that we will discuss later).

The Canonical DID Model

The simplest case

- We first discuss the DID design in the simplest setup.
- This will help us better understand the recent research that relaxes one or more aspects of the canonical model.
- Panel data on Y_{it} for $t = 1, 2$ and $i = 1, ..., N$
- Only two groups: some units $(D_i = 1)$ are treated in period 2; others are never treated $(D_i = 0)$
- ATT is our causal estimand of interest: $E[Y_{i2}(1) - Y_{i2}(0) \mid D_i = 1]$
- We will show that with the No Anticipation and Parallel Trend assumptions, we can identify the ATT.

Key Assumption: Parallel Trends

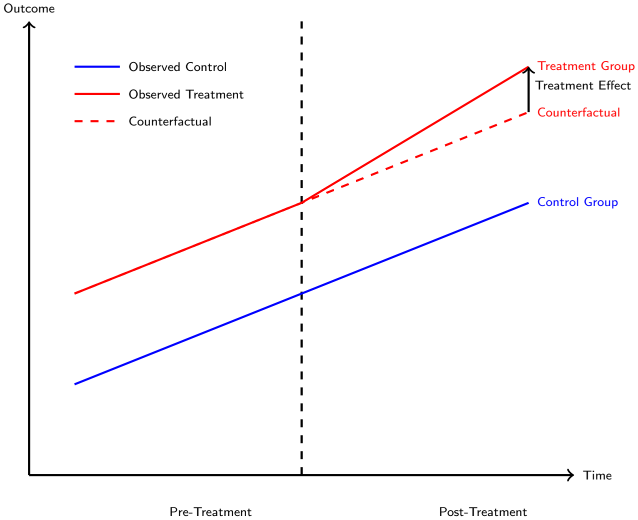
- The **parallel trends** assumption states that if the treatment hadn't occurred, average outcomes for the treatment and control groups would have evolved in parallel

$$\underbrace{E[Y_{i2}(0) - Y_{i1}(0) \mid D_i = 1]}_{\text{Counterfactual change for treated group}} = \underbrace{E[Y_{i2}(0) - Y_{i1}(0) \mid D_i = 0]}_{\text{Change for the control group}} \tag{11}$$

- Parallel Trend allows selection bias, as long as the selection bias is invariant over time:

$$\underbrace{E[Y_{i2}(0) \mid D_i = 1] - E[Y_{i2}(0) \mid D_i = 0]}_{\text{Selection bias in period 2}} = \underbrace{E[Y_{i1}(0) \mid D_i = 1] - E[Y_{i1}(0) \mid D_i = 0]}_{\text{Selection bias in period 1}} \tag{12}$$

A Visualization of Parallel Trend Assumption



Key Idea: The vertical difference between the observed treatment group and the counterfactual represents the treatment effect.

Identification

- **Identification:** ATT can be identified under the No Anticipation and Parallel Trend assumptions:

$$\tau_{ATT} = \underbrace{(E[Y_{i2}|D_i = 1] - E[Y_{i1}|D_i = 1])}_{\text{Change for treated}} - \underbrace{(E[Y_{i2}|D_i = 0] - E[Y_{i1}|D_i = 0])}_{\text{Change for control}} \tag{13}$$

- Thus ATT is just the difference-in-differences of CEFs.

Proof

- Below we prove the identification:

$$\begin{aligned}
\tau_{ATT} &= E[Y_{i2}(1) - Y_{i2}(0)|D_i = 1] \\
&= E[Y_{i2}(1) - Y_{i1}(0) + Y_{i1}(0) - Y_{i2}(0)|D_i = 1] \\
&= E[Y_{i2}(1) - Y_{i1}(0)|D_i = 1] - E[Y_{i2}(0) - Y_{i1}(0)|D_i = 1] \\
&= E[Y_{i2}(1) - Y_{i1}(0)|D_i = 1] - E[Y_{i2}(0) - Y_{i1}(0)|D_i = 0] \\
&= E[Y_{i2}(1) - Y_{i1}(1)|D_i = 1] - E[Y_{i2}(0) - Y_{i1}(0)|D_i = 0] \\
&= (E[Y_{i2}|D_i = 1] - E[Y_{i1}|D_i = 1]) - (E[Y_{i2}|D_i = 0] - E[Y_{i1}|D_i = 0])
\end{aligned} \tag{14}$$

- The fourth equation exploits the Parallel Trend assumption while the fifth equation uses the No Anticipation assumption.

Estimation

- The most straightforward 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

- $\hat{\tau}_{DID}$ is algebraically equal to OLS coefficient $\hat{\beta}$ from the below TWFE model:

$$Y_{it} = \alpha_i + \phi_t + W_{it}\beta + \epsilon_{it}, \tag{15}$$

where $W_{it} = D_i * 1[t = 2]$.

- It is easy to see the TWFE can also "run" for multiple periods, this is perhaps one of the reasons why TWFE has been so popular and only until recently we start to seriously doubt its validity.

Interpreting TWFE Under Staggered DID

Overview

- We have just showed that TWFE can identify the ATT under certain assumptions in the canonical DID setting.
- So it is natural to ask whether TWFE still has a causal interpretation in multiple periods with staggered adoption, which is the focus of of the recent literature.
- Many real-world policies roll out at different times
- A main focus of recent literature: difference-in-differences under staggered adoption.
- When there is treatment effect heterogeneity under the staggered setting, then we may run into the "negative weights" problem.

Extending the Identifying Assumptions

- Now we extend the No Anticipation and Parallel Trend assumptions to the staggered setting.
- **Parallel trends:** if treatment hadn't happened, all "adoption cohorts" would have parallel average outcomes for all periods:

$$E[Y_{it}(\infty) - Y_{i,t-1}(\infty)|A_i = a] = E[Y_{it}(\infty) - Y_{i,t-1}(\infty)|A_i = a'] \text{ for all } a, a', t \tag{16}$$

- **No anticipation:**

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

Negative Weights

- Consider the static TWFE model:

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

  where $W_{it} = 1[t \geq A_i]$ is a treatment indicator.
- If each unit has a constant treatment effect over time, $\tau_{it}(a) = Y_{it}(a) - Y_{it}(\infty) = \tau$, then $\hat{\beta}$ is a consistent estimate for τ
- if treatment effects are heterogeneous (within unit over time), then β may put negative weights on treatment effects for some units and time periods
- For example, if treatment effect depends only on time since treatment, $\tau_r = Y_{it}(t-r) - Y_{it}(\infty)$, then some τ_r may get negative weight

The Source of the Negative Weights

The intuition for these negative results is that the OLS is too greedy an aggregation scheme by aggregating all types of comparisons.

- **Good comparisons:** between treated and not-yet-treated units
- **Forbidden comparisons:** between newly-treated and already-treated units

These forbidden comparisons are the source of negative weights: the early adopters serve as controls for later adopters. This is problematic if their treatment effects change over time

Goodman-Bacon (2021)

- Goodman-Bacon (2021) provides helpful intuition for understanding the source of negative weights.
- $\hat{\beta}$ can be written as a weighted average of different two-groups-and-two-periods comparisons.
- The weights on the 2×2 DID are proportional to timing group sizes and the variance of the treatment dummy in each cell, which is highest for units treated in the middle of the panel.
- Some comparisons involve early-treated units as controls for later-treated units.
- This can introduce negative weights, making $\hat{\beta}$ problematic when treatment effects differ across units or time.

Graphic Illustrations of Comparisons (from Goodman-Bacon (2021))

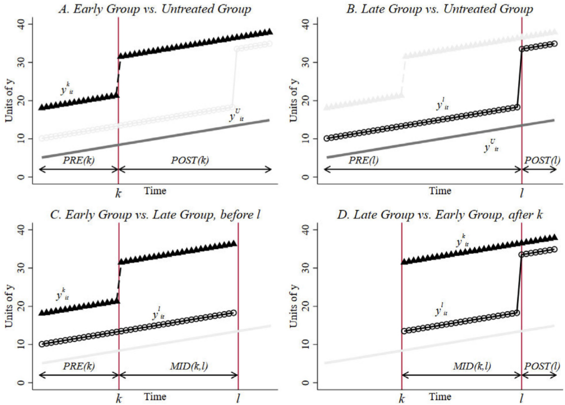


Fig. 2. The four simple (2x2) difference-in-differences estimates in the three group case. Notes: The figure plots outcomes for the subsamples that generate the four simple 2x2 difference-in-difference estimates in the three timing group case from Fig. 1. Each panel plots the data structure for one 2x2 DD. Panel A compares early treated units to untreated units ($\hat{\beta}^{DD}_{kU}$); panel B compares late treated units to untreated units ($\hat{\beta}^{DD}_{\ell U}$); panel C compares early treated units to late treated units during the late timing group's pre-period ($\hat{\beta}^{DD,k}_{k\ell}$); panel D compares late treated units to early treated units during the early timing group's post-period ($\hat{\beta}^{DD,\ell}_{k\ell}$). The treatment times mean that $\overline{D}_k = 0.67$ and $\overline{D}_\ell = 0.16$, so with equal group sizes, the decomposition weights on the 2x2 estimate from each panel are 0.365 for panel A, 0.222 for panel B, 0.278 for panel C, and 0.135 for panel D.

Another Intuitive Explanation for Negative Weights

- By FWL, we have:

$$\beta = \frac{Cov(Y_{it}, W_{it} - \hat{W}_{it})}{Var(W_{it} - \hat{W}_{it})} = \frac{E(Y_{it}(W_{it} - \hat{W}_{it}))}{Var(W_{it} - \hat{W}_{it})} \tag{17}$$

  where $W_{it} - \hat{W}_{it}$ is the residual from regressing W_{it} on the individual and time fixed effects.

- If $W_{it} = 1$ but $W_{it} - \hat{W}_{it} < 0$, then β will be decreasing in Y_{it} even though i is treated at time t.
- These negative weights will tend to arise for early-treated units in later time, when they serve as controls for later adopters as discussed by Goodman-Bacon (2021).

Dynamic TWFE

- Now consider the dynamic TWFE specifications:

$$Y_{it} = \alpha_i + \lambda_t + \sum_{k \neq 0} \gamma_k W_{it}^k + \varepsilon_{it},$$

  where $W_{it}^k = 1\{t - A_i = k\}$ are "event-time" indicators.
- If all units have treatment effect τ_k in the k-th period after adoption, then $\beta_k = \tau_k$ under the parallel trends and no anticipation assumptions
- However, as shown by Sun and Abraham (2021), when there are heterogeneous dynamic treatment effects across adoption cohorts, γ_k may be a non-convex weighted average of the dynamic treatment effect k periods after treatment
- Sun and Abraham (2021) also show that γ_k may be "contaminated" by treatment effects at lags $k' \neq k$

Summary

- We have learned that the TWFE might not be a good aggregation scheme when there is heterogeneity in treatment effect.
- The dynamic TWFE allows for heterogeneity in treatment effects by the time spent on treatment, but not by adoption cohort. This might not be desirable
- Even if the weights are all positive, the weighted average treatment effects might not be the most interesting or not even policy relevant.
- This “reverse engineering” approach has limited potentials.

The Bottom-Up/Forward-Engineering Approach

The Bottom-Up/Forward-Engineering Approach

Introduction

Overview

- Several recent papers have proposed alternative estimators that more sensibly aggregate heterogeneous treatment effects in settings with staggered treatment timing.
- These estimators avoid the bias from TWFE by taking a bottom-up approach.
- The TWFE approach starts from an estimate and asks whether it has a causal interpretation (i.e. a positively weighted average of building-block causal estimands), and if yes, whether this estimate is interesting.

Overview

- The recent estimators start from the causal estimand we care about and propose a proper weighting scheme to form a higher-level estimand.
- Those new methods differ in the choice of the building-block estimands, the groups that serve as controls, and weights.
- But they all avoid forbidden comparisons and are transparent about the estimand and weighting schemes.
- Here we discuss two representative estimators: Callaway and Sant'Anna (2021), and Borusyak et al. (2024)

The Bottom-Up/Forward-Engineering Approach

Callaway and Sant'Anna (2021)

Callaway and Sant'Anna (2021)

- The ATT for group a at time t is given by:

$$ATT(a,t) = \mathbb{E}[Y_{i,t} - Y_{i,a-1} \mid A_i = a] - \mathbb{E}[Y_{i,t} - Y_{i,a-1} \mid A_i = a'], \quad \text{for any } a' > t \tag{18}$$

This is a multi-period generalization of the identification result for the canonical DID model.

If this holds for any comparison group $a' > t$, then it also holds if we average over some set of comparisons A_{comp} such that $a' > t$ for all $a' \in G_{\text{comp}}$:

$$ATT(a,t) = \mathbb{E}[Y_{i,t} - Y_{i,a-1} \mid A_i = a] - \mathbb{E}[Y_{i,t} - Y_{i,a-1} \mid A_i \in A_{\text{comp}}] \tag{19}$$

Estimating ATT Using Sample Analogs

$ATT(a,t)$ can be estimated by replacing the expectations with their sample analogs:

$$\widehat{ATT}(a,t) = \frac{1}{N_a} \sum_{i:A_i=a} [Y_{i,t} - Y_{i,a-1}] - \frac{1}{N_{A_{\text{comp}}}} \sum_{i:A_i \in A_{\text{comp}}} [Y_{i,t} - Y_{i,a-1}] \quad (20)$$

where:

- N_a is the number of treated units in group a.
- $N_{A_{\text{comp}}}$ is the number of comparison units.

Callaway and Sant'Anna (2021)

- Callaway and Sant'Anna (2021) propose two possible choices for the comparison group A_{comp}:
  - Option 1: Use only never-treated units:

$$A_{\text{comp}} = \{\infty\} \tag{21}$$

  - Option 2: Use all not-yet-treated units:

$$A_{\text{comp}} = \{a' : a' > t\} \tag{22}$$

- When there are relatively few time periods and treatment cohorts, reporting $\widehat{ATT}(a, t)$ for all relevant (a, t) is feasible.
- With many treated periods and/or cohorts, reporting all $ATT(a, t)$ can be inconvinient and imprecise.

Aggregation schemes

- With many treated periods and/or cohorts, it is desirable to report higher-level estimand by averaging the $\widehat{ATT}(a, t)$ using some sensible weights.
- For example, the event-study parameter: $\hat{\theta}_k = \sum_a \omega_a \widehat{ATT}(a, a + k)$ that aggregates effects for cohorts in the kth period after adoption
- The weights ω_a can be chosen to weight different cohorts equally, or by their relative frequencies in the treated population
- We can also construct for $k < 0$ to estimate “pre-trends”

The Bottom-Up/Forward-Engineering Approach

Imputation-based Methods

Borusyak et al. (2024)

- Estimate the individual fixed effect and time fixed effect by regressing the below model only on pre-treatment data:

$$Y_{it}(\infty) == \lambda_t + \gamma_i + \epsilon_{it} \tag{23}$$

- For each treated unit, compute $\hat{\tau}_{it} = Y_{it} - \widehat{Y_{it}}(\infty)$
- Then we get:

$$\widehat{\tau}_{BJS} = \sum_{W_{it}=1} \widehat{\tau}_{it} \Big/ |\{W_{it} = 1\}| \tag{24}$$

- Individual $\widehat{\tau}_{it}$ estimates may not noisy and inconsistent.
- However, their aggregate $\widehat{\tau}_{BJS}$ is designed to average out errors, recovering consistency.

Summary

- Sun and Abraham (2021) propose a similar estimator but with different comparisons groups (using last-to-be treated rather than not-yet-treated)
- Borusyak et al. (2024) uses more pre-treatment periods than Callaway and Sant'Anna (2021), which just uses period $a - 1$
- This can sometimes be more efficient, but also relies on a stronger Parallel Trend (for all groups and time periods) than SA and CS, which may be more susceptible to bias