DID with Staggered Adoption

Lee Kennedy-Shaffer, PhD

2025-06-10

Motivating Example: COVID-19 Vaccine Mandates

Question

What happens when multiple units adopt the intervention at different times?

Motivating Example: COVID-19 Vaccine Mandates

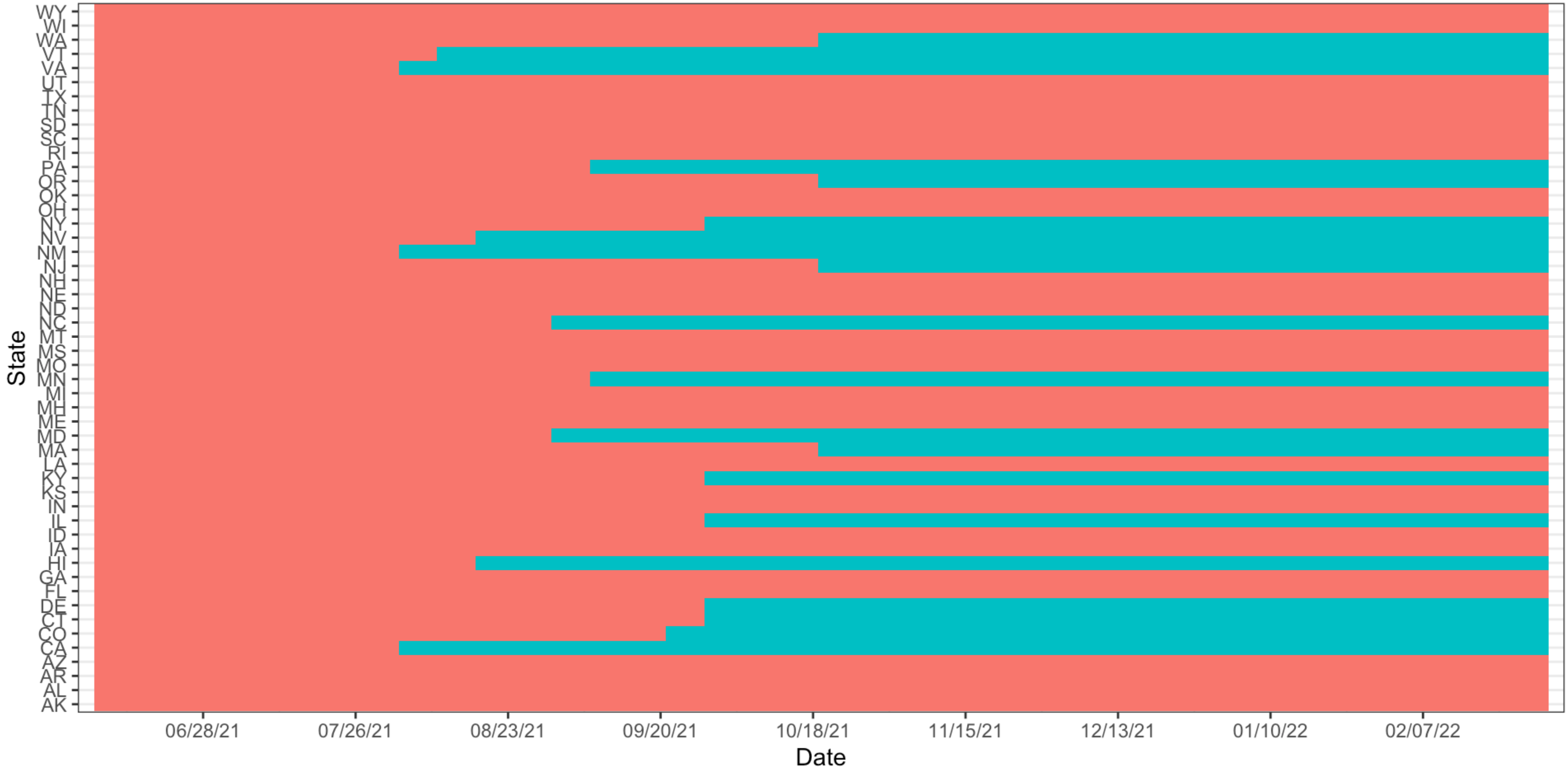
Question

What happens when multiple units adopt the intervention at different times?

Example

Twenty states adopted COVID-19 vaccine mandates for state employees at different times.

COVID-19 Vaccine Mandate Timing

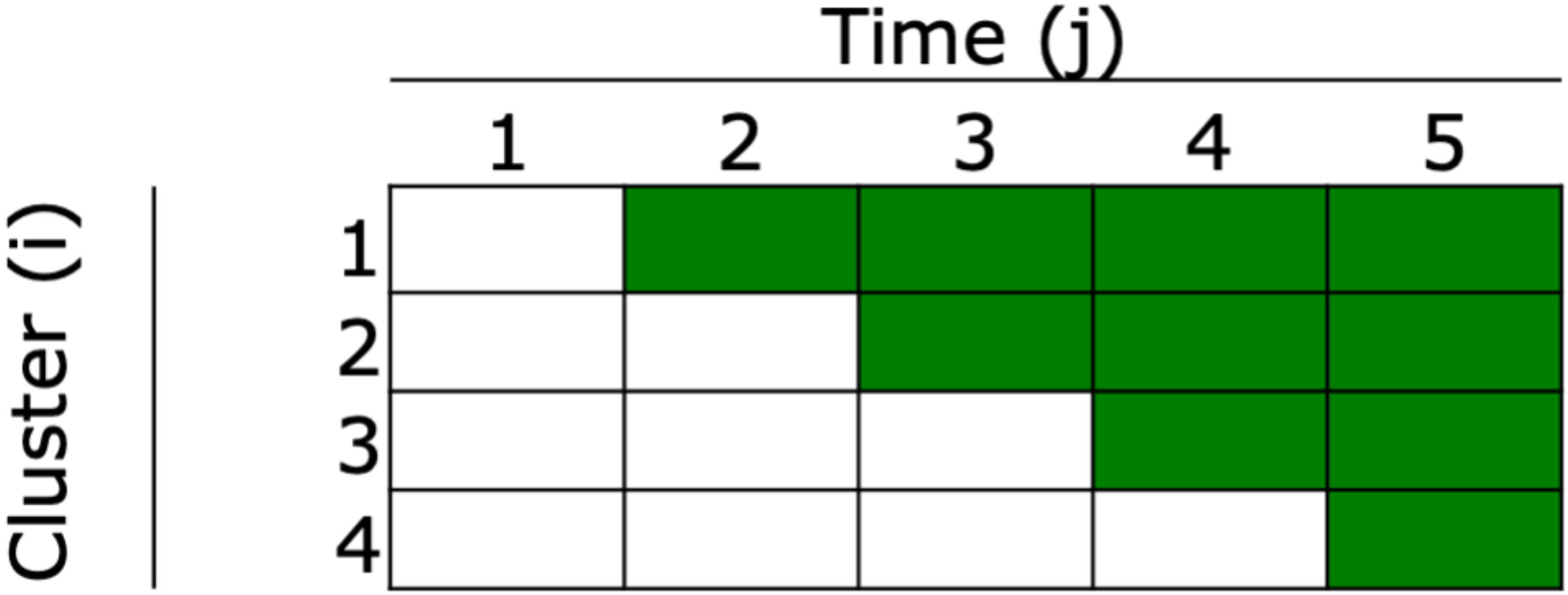


Plot of state employee vaccination mandate timings, U.S. states, June 2021–February 2022

Staggered Adoption

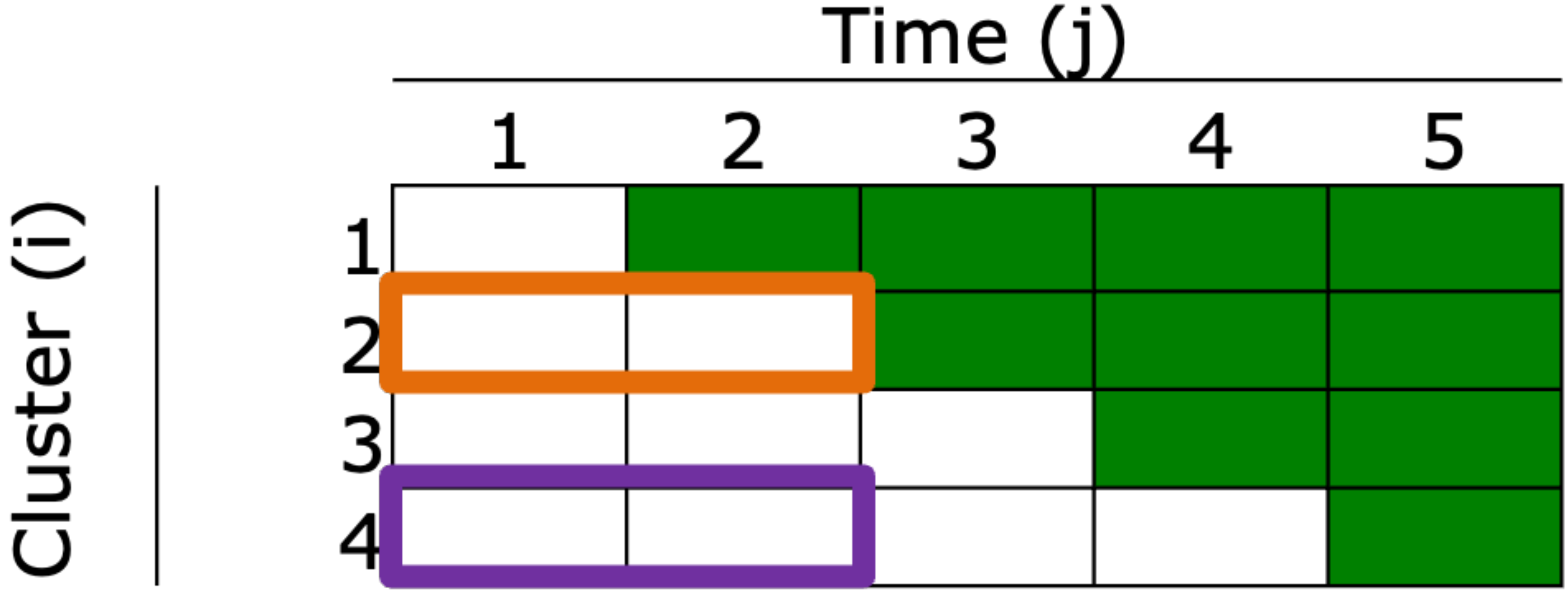
Panel Data Setting

- Multiple units, treated at different time points
- Multiple time points of observations

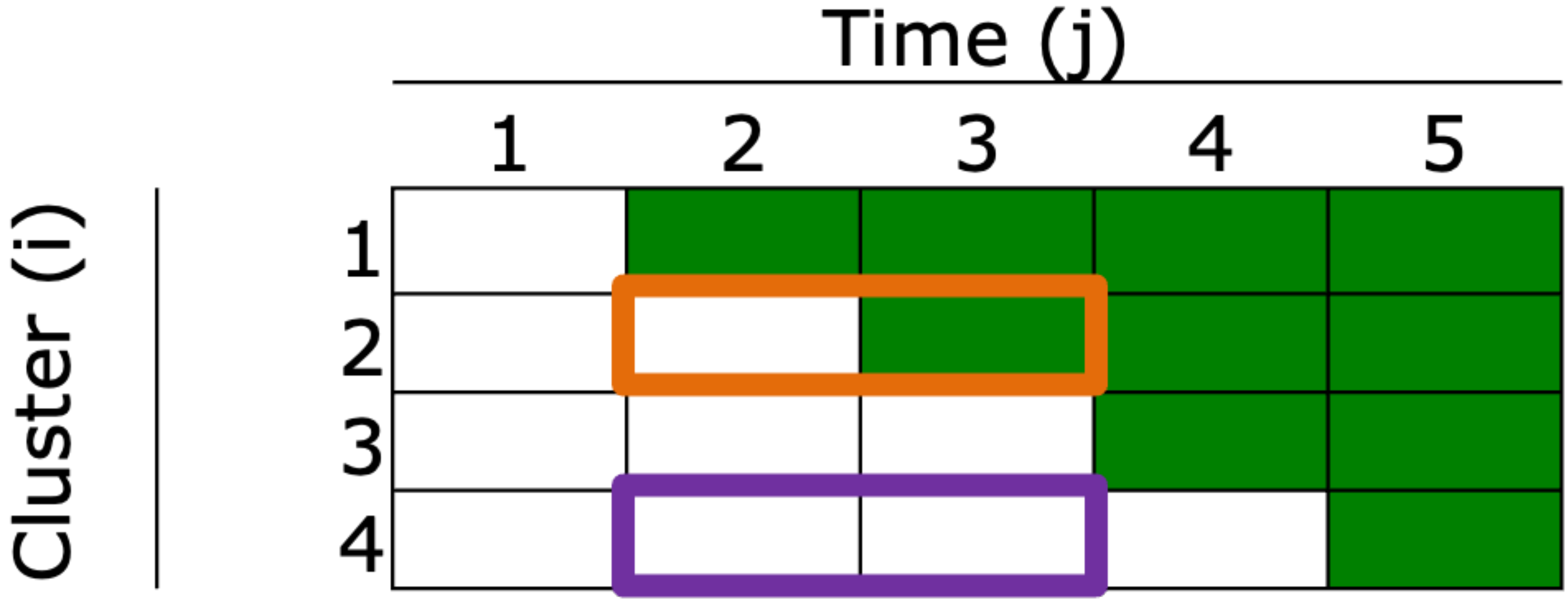


Cluster (i) \ Time (j)	1	2	3	4	5
1		🟩	🟩	🟩	🟩
2			🟩	🟩	🟩
3				🟩	🟩
4					🟩

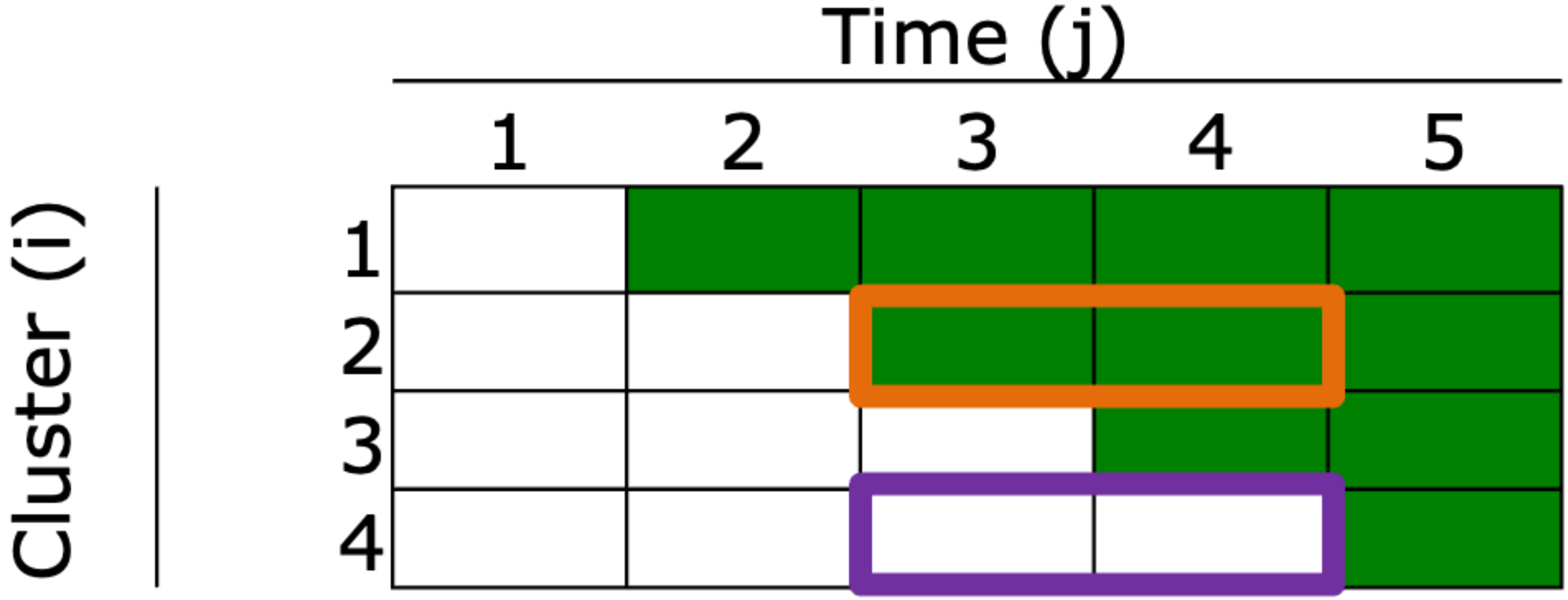
Types of 2x2s: Always Control vs. Always Control



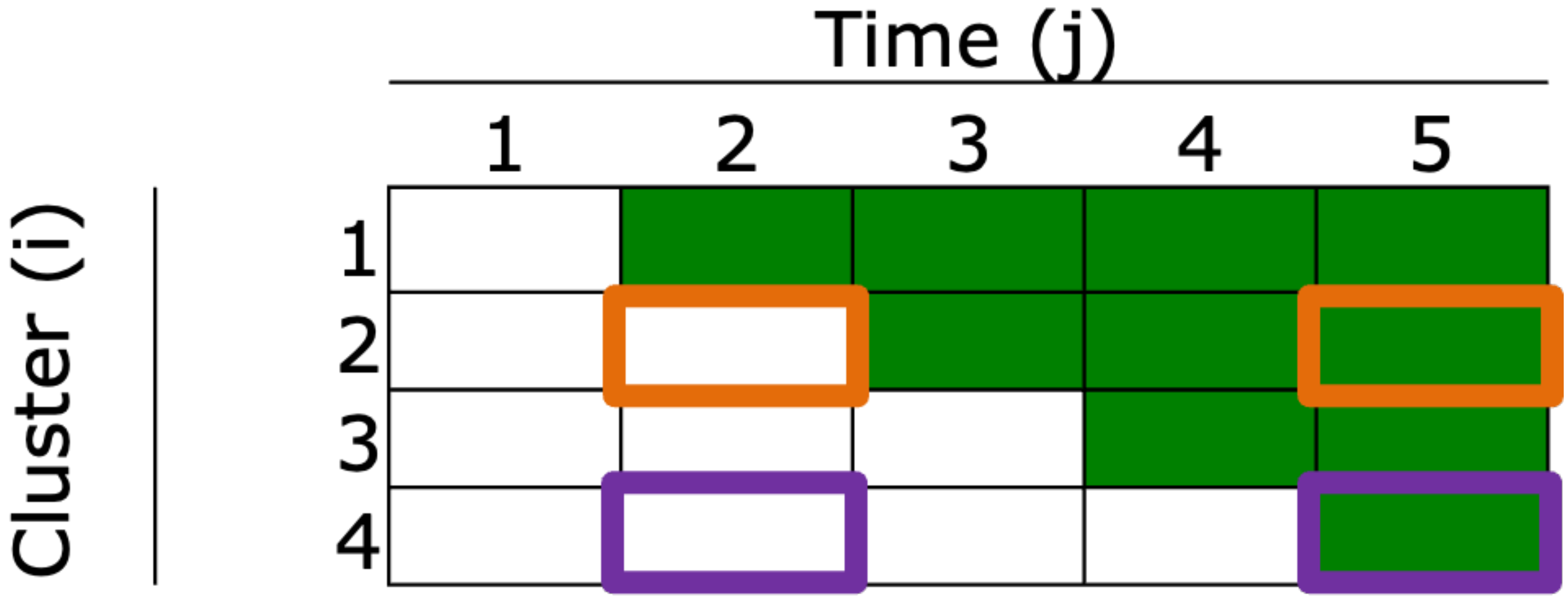
Types of 2x2s: Switch vs. Always Control



Types of 2x2s: Always Treated vs. Always Control

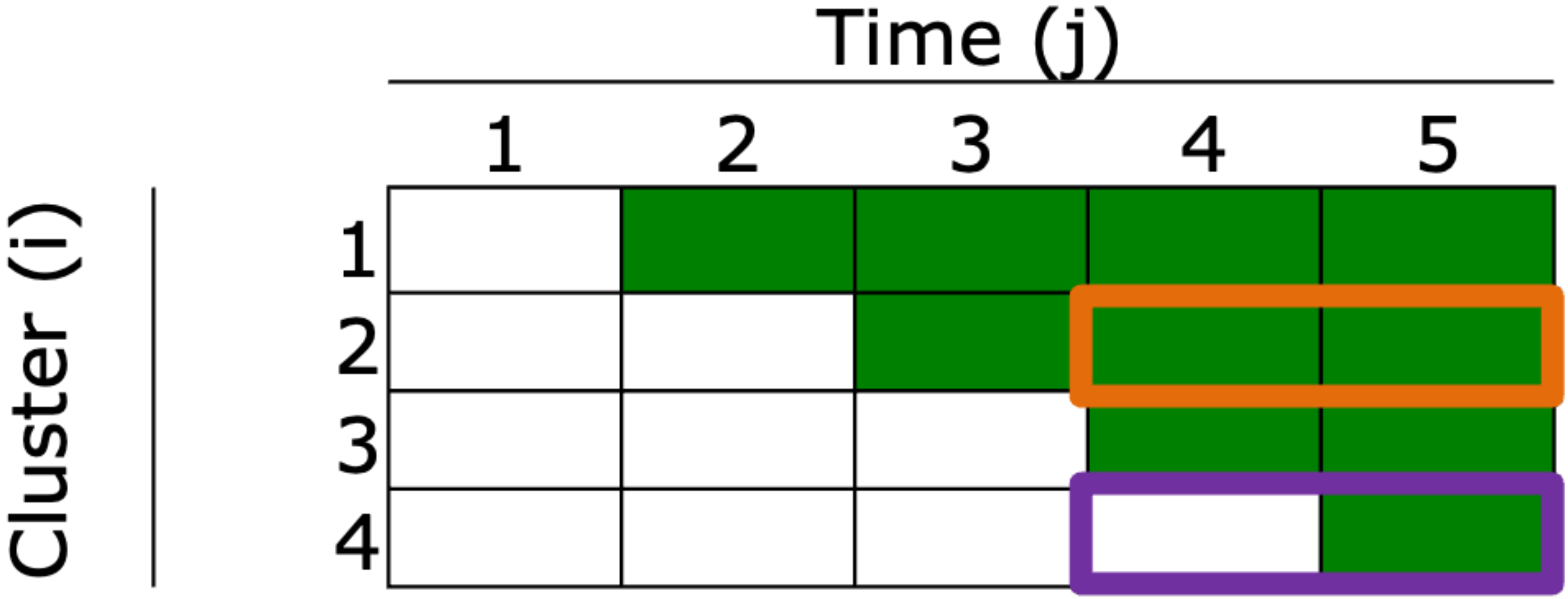


Types of 2x2s: Switch vs. Switch



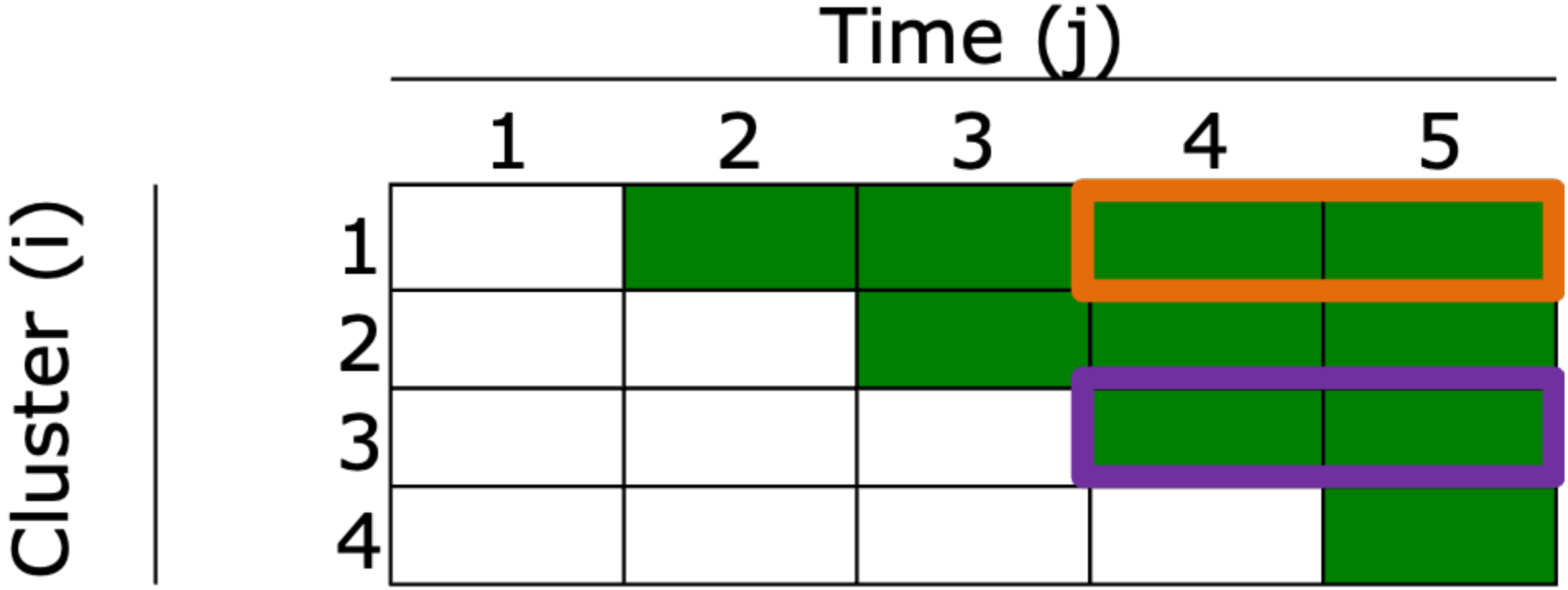
	1	2	3	4	5
1		🟩	🟩	🟩	🟩
2			🟩	🟩	🟩
3				🟩	🟩
4					🟩

Types of 2x2s: Always Treated vs. Switch



Cluster (i)	1	2	3	4	5
1		🟩	🟩	🟩	🟩
2			🟩	🟩	🟩
3				🟩	🟩
4					🟩

Types of 2x2s: Always Treated vs. Always Treated



Cluster (i)	1	2	3	4	5
1		🟩	🟩	🟩	🟩
2			🟩	🟩	🟩
3				🟩	🟩
4					🟩

Two-Way Fixed Effects Model

Caution

The TWFE model can accommodate this *statistically*:

$$Y_{it} = \alpha_i + \gamma_t + \theta I(X_{it} = 1) + \epsilon_{it}$$

But as written it assumes **treatment effect homogeneity** across time periods, time-on-treatment, and units: there is a single treatment effect θ.

Challenge: Heterogeneity

Treatment Effect Heterogeneity

In many settings, especially in epidemiology, heterogeneity is common, especially with non-randomized adoption.

> **Question**
>
> What might cause heterogeneity in the effect of a state employee COVID-19 vaccine mandate?

Treatment Effect Heterogeneity

In many settings, especially in epidemiology, heterogeneity is common, especially with non-randomized adoption.

> **ⓘ Question**
>
> What might cause heterogeneity in the effect of a state employee COVID-19 vaccine mandate?

Journal of Econometrics 225 (2021) 254–277

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

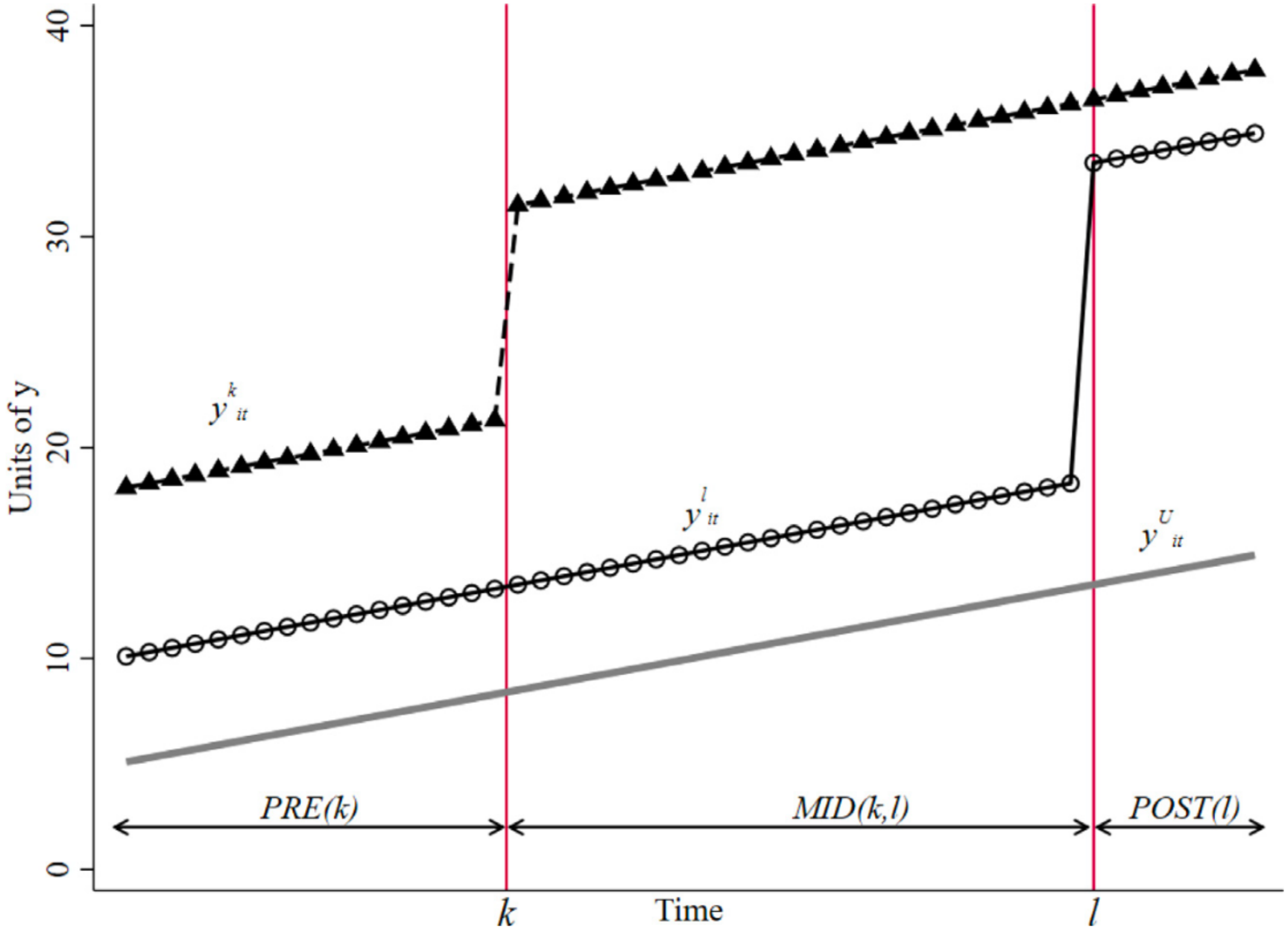
ELSEVIER
JOURNAL OF Econometrics
Check for updates

Difference-in-differences with variation in treatment timing☆

Andrew Goodman-Bacon *

Opportunity and Inclusive Growth Institute, Federal Reserve Bank of Minneapolis, 90 Hennepin Ave, Minneapolis, MN 55401, USA
National Bureau of Economic Research, USA

Weighted Average of Effects

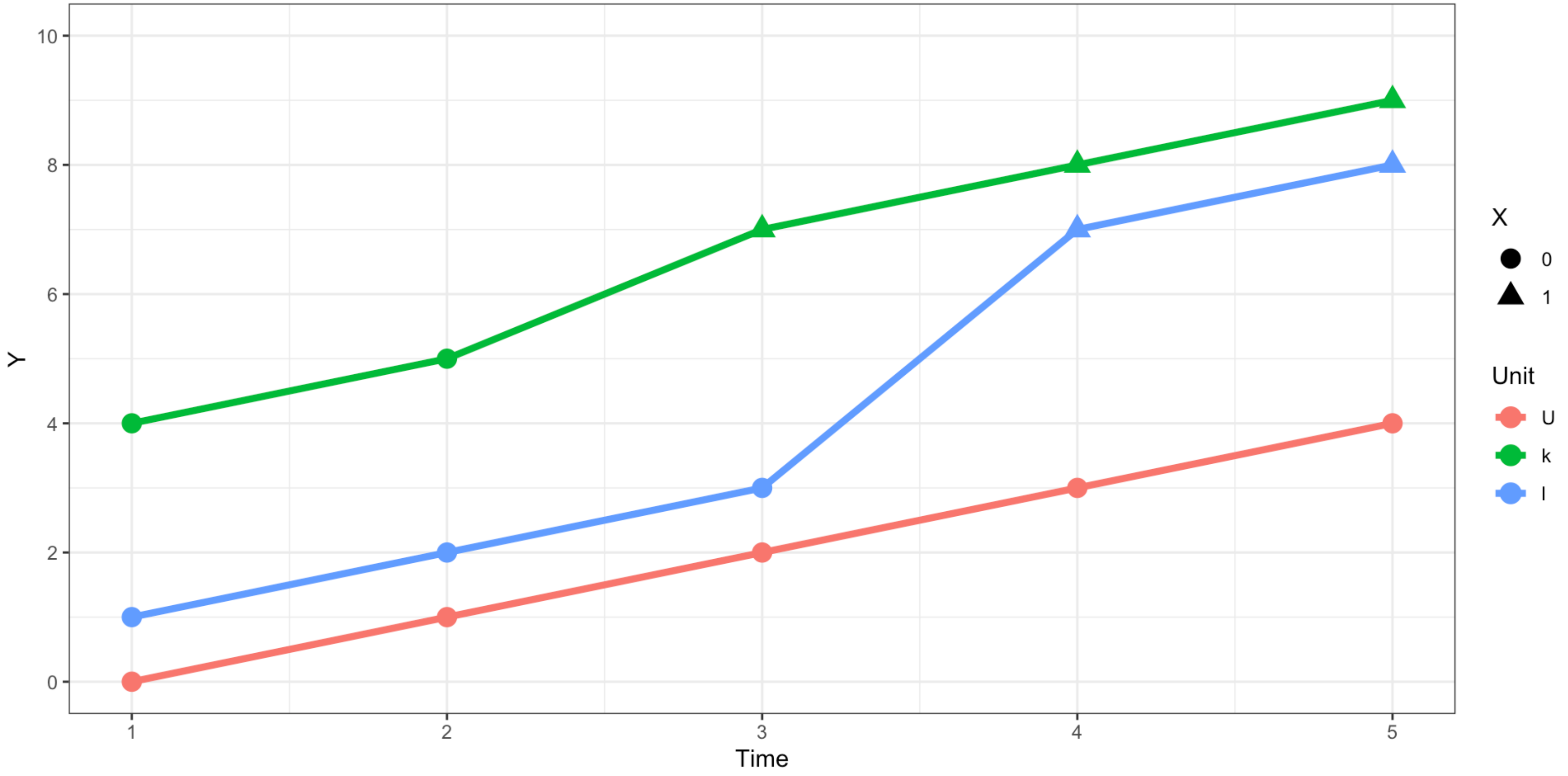


Goodman-Bacon (2021), Figure 1.

Weighted Average of Effects

This complicates the interpretation of the estimand: it is still an **ATT**, but now that average is a weighted average that depends on the number of units at each switch point and the number of periods observed.

Weighed Average of Effects: Simplified Example



Plot of three-unit staggered adoption example with outcomes

Weighed Average of Effects: Simplified Example

Using all periods:

$$\theta = \frac{1+1+1+3+3}{5} = 1.8$$

Using only the first four periods:

$$\theta = \frac{1+1+3}{3} \approx 1.67$$

Weighted Average of Effects: Regression Estimates

Using all periods:

```
(Intercept)        Time       Unitk       Unitl          X1
       -1.0         1.0         3.4         1.4         2.0
```

Using only the first four periods:

```
(Intercept)        Time       Unitk       Unitl          X1
  -1.000000    1.000000    3.571429    1.285714    1.857143
```

Weightings

If all units are of the same size/variance, per Goodman-Bacon (2021) Thm. 1:

$$
\begin{aligned}
s_{kU} &\propto \bar{D}_k(1-\bar{D}_k) \\
s_{k\ell}^{k} &\propto (\bar{D}_k-\bar{D}_\ell)(1-\bar{D}_k) \\
s_{k\ell}^{\ell} &\propto (\bar{D}_k-\bar{D}_\ell)(1-\bar{D}_\ell)
\end{aligned}
$$

where $\bar{D}$ is the proportion of periods treated.

Weighted Average of Effects: In Regression Estimator

Using all periods:

$$\bar{D}_k = 3/5, \quad \bar{D}_\ell = 2/5$$

$$s_{kU} \propto 6/25, \quad s_{\ell U} \propto 6/25$$

$$s_{k\ell}^{k} \propto 2/25, \quad s_{k\ell}^{\ell} \propto 2/25$$

$$\hat{\theta} = \frac{6 \cdot 1 + 6 \cdot 3 + 2 \cdot 1 + 2 \cdot 3}{6 + 6 + 2 + 2} = 2$$

Weighted Average of Effects: In Regression Estimator

Using only the first four periods:

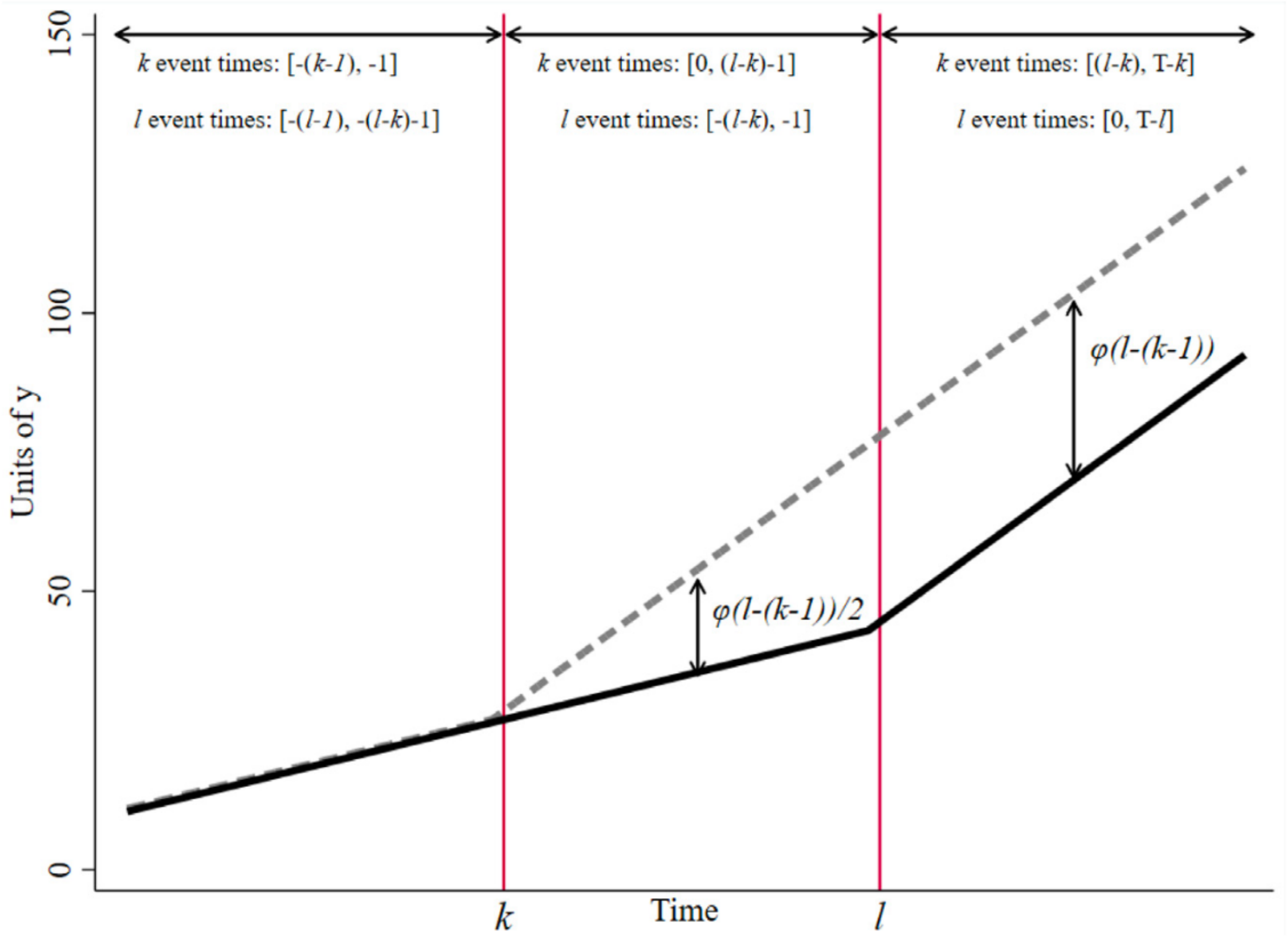
$$\bar{D}_k = 2/4, \quad \bar{D}_\ell = 1/4$$

$$s_{kU} \propto 2/16, \quad s_{\ell U} \propto 2/16$$

$$s_{k\ell}^{k} \propto 2/16, \quad s_{k\ell}^{\ell} \propto 1/16$$

$$\hat{\theta} = \frac{2 \cdot 1 + 2 \cdot 3 + 2 \cdot 1 + 1 \cdot 3}{2 + 2 + 2 + 1} \approx 1.857$$

Time-Varying Effects



Goodman-Bacon (2021), Figure 3.

Forbidden Comparisons

The weights on treatment effects can be non-convex (i.e., negative) if there are **time-varying treatment effects**.

This gives an uninterpretable TWFE estimand, and can even switch the sign of the estimate.

Forbidden Comparisons

Time-varying effects can be broadly categorized into:

- **Dynamic treatment effects** that depend on how long a unit has been treated, and
- **Calendar time heterogeneous effects** that differ for all units based on the period of observation.

Both can be problematic, although calendar time effects can be handled similarly to unit-varying effects.

Goodman-Bacon Decomposition: General

ⓘ **Goodman-Bacon (2021), Eqn 15**

$$plim_{N \to \infty} \hat{\theta} = \beta^{DD} = VWATT + VWCT - \Delta ATT,$$

where:

$VWATT$ is the variance-weighted ATT (as in computation above),

$VWCT$ is the variance-weighted deviation from parallel trends, and

ΔATT is the weighted sum of changes in the treatment effect within timing groups.

Goodman-Bacon Decomposition: 2x2s

Goodman-Bacon (2021), Theorem 1

$$\hat{\theta} = \sum_{k\neq U} \hat{\beta}_{kU}^{2\times 2} + \sum_{k\neq U}\sum_{\ell>k}\left[\boldsymbol{s}_{k\ell}^{k}\hat{\beta}_{k\ell}^{2\times 2,k} + \boldsymbol{s}_{k\ell}^{\ell}\hat{\beta}_{k\ell}^{2\times 2,\ell}\right]$$

where:

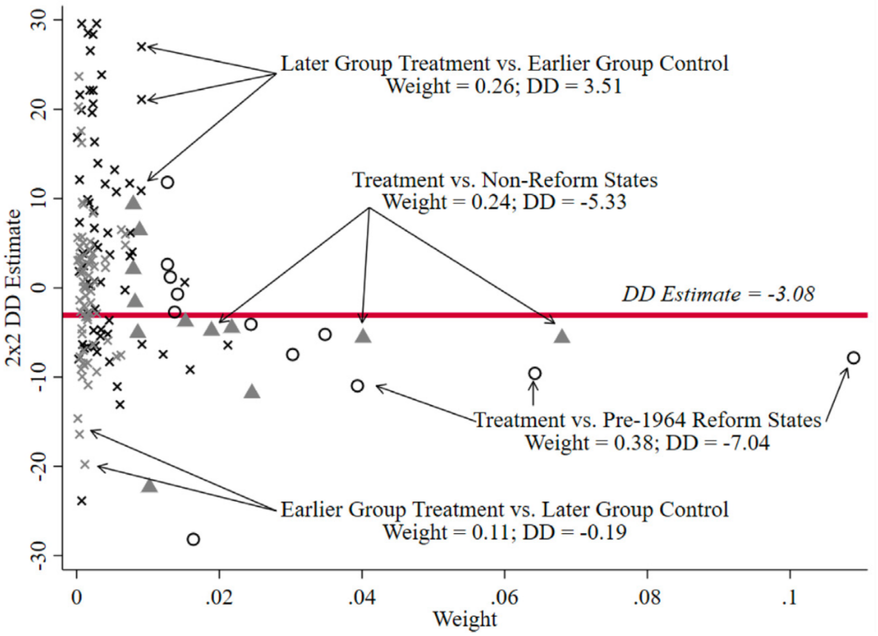
$\hat{\beta}_{kU}^{2\times 2}$ is a comparison of timing group k to untreated U,

$\hat{\beta}_{k\ell}^{2\times 2,k}$ is a comparison of an early-treated group k to a late-treated group ℓ in the time between the two switches, and

$\hat{\beta}_{k\ell}^{2\times 2,\ell}$ is a comparison of a late-treated group ℓ to an early-treated group k in the time after ℓ switched.

Goodman-Bacon Decomposition of DID Comparisons

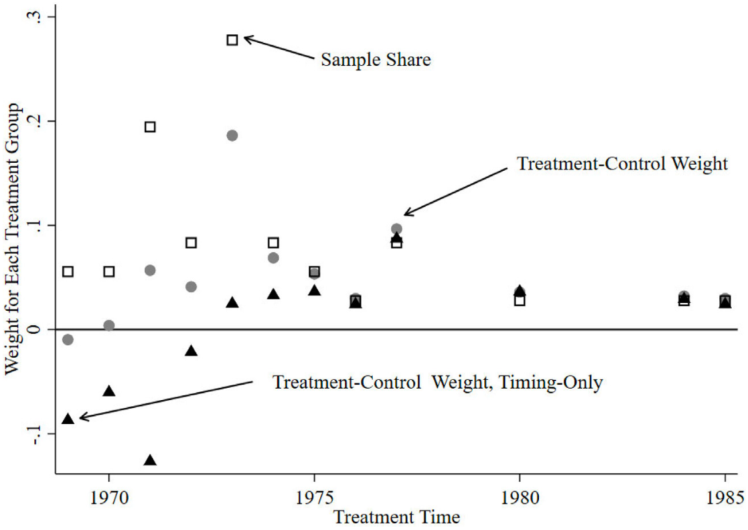
The TWFE model estimates a weighted average of all 2x2 DID comparisons.



Goodman-Bacon (2021), Figure 6.

Goodman-Bacon Decomposition of Treatment Effects

We can observe the overall weight given to each treatment timing group, which may be negative if it is more often used as a control than a treated group.



Goodman-Bacon (2021), Figure 7.

Issues with TWFE

Issues with TWFE

1. Under unit-varying treatment effects, $VWATT$ may not be a readily interpretable average of effects (and some may get negative weights).

Issues with TWFE

1. Under unit-varying treatment effects, $VWATT$ may not be a readily interpretable average of effects (and some may get negative weights).
2. Under any individual deviations from parallel trends (even if average does not deviate), $VWCT$ may not be 0.

Issues with TWFE

1. Under unit-varying treatment effects, $VWATT$ may not be a readily interpretable average of effects (and some may get negative weights).
2. Under any individual deviations from parallel trends (even if average does not deviate), $VWCT$ may not be 0.
3. Under time-varying treatment effects, $\Delta ATT \neq 0$: can result in negative weights on some effects.

Issues with TWFE

1. Under unit-varying treatment effects, $VWATT$ may not be a readily interpretable average of effects (and some may get negative weights).
2. Under any individual deviations from parallel trends (even if average does not deviate), $VWCT$ may not be 0.
3. Under time-varying treatment effects, $\Delta ATT \neq 0$: can result in negative weights on some effects.

> ⚠ **Warning**
>
> Negative weights on some treatment effects *can* lead to averages outside of the range, or even a changed treatment effect sign.

Proposed Solutions

Proposed Solutions (Broadly)

1. Dynamic Specification
2. Dynamic Specification with Restricted Observations
3. Interacted Specification
4. Restricting Periods: First Differences
5. Explicit Weightings

1. Dynamic Regression Specification

$$Y_{it} = \alpha_i + \gamma_t + \sum_{k \neq -1} \delta_k I(K_{it} = k) + \epsilon_{it},$$

where K_{it} is the lead/lag for unit i in period t (e.g., $K_{it} = 0$ in the first exposed period).

See Borusyak and Jaravel (2018) and Borusyak et al. (2024). Captures time-on-treatment heterogeneity.

Also useful to test for “pre-trends” in single intervention time setting: δ_k should be 0 for $k < 0$.

Issues

If there are any heterogeneities beyond time-on-treatment, then this dynamic specification can still give bias or uninterpretable estimands.

- Negative weights on some units
- “Contaminated” period effects

See Sun and Abraham (2021).

2. Restricting Observations

One approach to avoiding forbidden comparisons is restricting the observations used to fit the model.

Borusyak et al. (2024) fit this model using only the not-yet-treated observations. They then use that to derive counterfactual outcomes for comparison.

2. Restricting Observations

One approach to avoiding forbidden comparisons is restricting the observations used to fit the model.

Borusyak et al. (2024) fit this model using only the not-yet-treated observations. They then use that to derive counterfactual outcomes for comparison.

Sun and Abraham (2021) use the approach only comparing to a clean "control" C that is either never-treated or last-treated. Their regression approach then implicitly weights by population share in each timing group.

2. Restricting Observations

Advantages:

- Simple to implement
- Straightforward interpretations for each treated unit

Disadvantages/Limitations:

- Still opaque weighting of unit- or time-varying effects
- Throws out potentially valuable information: inefficient

3. Interacted Dynamic Effects

We can account for timing cohort heterogeneity as well by further allowing the effect to vary by adoption timing group (G_i):

$$Y_{it} = \alpha_i + \gamma_t + \sum_{g \notin C} \sum_{k \neq -1} \delta_{g,k} I(G_i = g) I(K_{it} = k) + \epsilon_{it}$$

3. Interacted Dynamic Effects

We can account for timing cohort heterogeneity as well by further allowing the effect to vary by adoption timing group (G_i):

$$Y_{it} = \alpha_i + \gamma_t + \sum_{g \notin C} \sum_{k \neq -1} \delta_{g,k} I(G_i = g) I(K_{it} = k) + \epsilon_{it}$$

Various methods use this approach (e.g., Sun and Abraham (2021) can be put in this form), and differ in which comparisons/observations they allow and how they combine results (SA focuses on time-on-treatment averaging). This implies different assumptions and bias-variance tradeoffs.

3. Stacked DID Regression

Cengiz et al. (2019) and others weight through “stacked” regression. Get data for a clean comparison for each treated group and put these data sets together, with an indicator for each one. Then conduct TWFE, accounting for these data indicators.

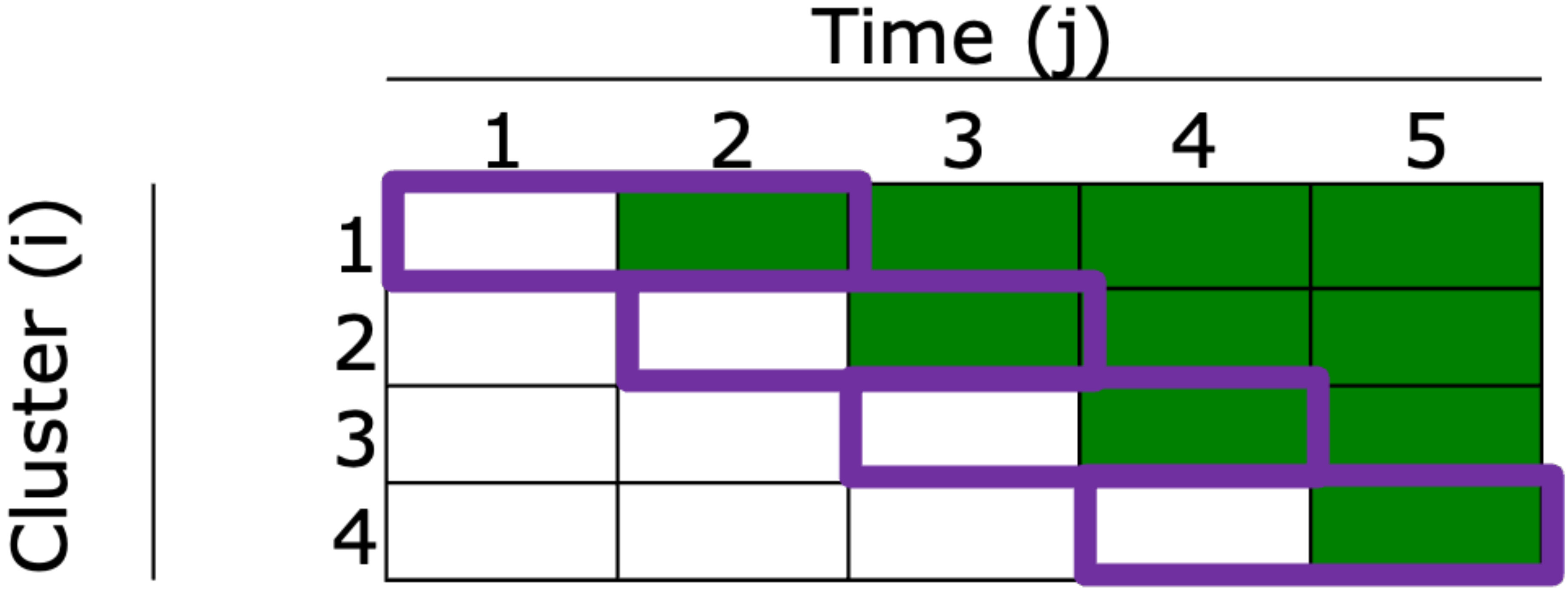
This essentially estimates each $\delta_{g,t}$ for each unit-time effect and then inverse-variance weights them through OLS. May have good efficiency, but limits interpretability of estimand if there are dynamic heterogeneities.

4. Restricting Periods: First-Difference

Use only the immediate switching effect. For each time period t with at least one unit untreated at $t-1$ and treated at t and at least one unit untreated at both $t-1$ and t, compute:

$$\widehat{DID}_{+,t} = \frac{1}{N_{1,0,t}} \sum_{i:D_{i,t}=1,D_{i,t-1}=0} (Y_{i,t} - Y_{i,t-1}) \\ - \frac{1}{N_{0,0,t}} \sum_{i:D_{i,t}=D_{i,t-1}=0} (Y_{i,t} - Y_{i,t-1})$$

4. Restricting Periods: First-Difference



	Time (j)				
Cluster (i)	1	2	3	4	5
1		🟩	🟩	🟩	🟩
2			🟩	🟩	🟩
3				🟩	🟩
4					🟩

4. Restricting Periods: First-Difference

Average these switcher estimates across all time periods t, weighted by number of units or individuals.

See de Chaisemartin and d'Haultfoeuille (2020) and de Chaisemartin and d'Haultfoeuille (2023), or the crossover (CO) estimator in Kennedy-Shaffer et al. (2020).

4. Restricting Periods: First-Difference

Advantages:

- All switching units (except possibly last) are included equally
- Can restrict to only clean comparisons
- Avoids dynamic treatment effects

Disadvantages/Limitations:

- Throws out a lot of information: inefficient
- Does not capture the full scope of treatment effects
- Need to be very careful about wash-out periods and interpretations

5. Weighting: Group-Time ATT

Let

$$ATT(g,t) = E[Y_{it}(g) - Y_{it}(0)],$$

the group-time ATT in period t for a unit first treated in period g, compared to if it had never been treated (or not yet treated by period t).

5. Weighting: Group-Time ATT

Let

$$ATT(g, t) = E[Y_{it}(g) - Y_{it}(0)],$$

the group-time ATT in period t for a unit first treated in period g, compared to if it had never been treated (or not yet treated by period t).

Many solutions boil down to considering which group-time ATTs should be included in the estimand, how they differ, and how to weight them.

TWFE assumes $ATT(g, t) = \theta$ for all $g \leq t$.

5. Weighting: Callaway and Sant’Anna

Callaway and Sant’Anna (2021) propose to estimate $\widehat{ATT}_{g,t}$ for each timing group g and period t using a non-parametric scheme compared to the **last** pre-treatment period, comparing group g to a never-, last-, or not yet-treated control. Simple DID can be used, or regression-based estimators (IPW, OR, DR) to accomodate covariates. CS favor aggregation by cohort. Also has highly developed inference.

Then summarize to an overall average effect weighted by $w_{g,t}$:

$$\theta = \sum_{g} \sum_{t=2}^{T} w_{g,t} ATT_{g,t}.$$

5. Weighting: 2x2 Building Blocks

Another framework allows weights across all 2x2 DID comparisons, with weights chosen to target a specific estimand or minimize variance.

$$\hat{\theta} = \sum_{i,i',t,t'} w_{i,i',t,t'} \left[\left(Y_{i,t'} - Y_{i,t} \right) - \left(Y_{i',t'} - Y_{i',t} \right) \right]$$

See Kennedy-Shaffer (2024) and Baker et al. (2025).

Dynamic and Weighting Approaches

Advantages:

- Highly flexible across heterogeneities
- Can ensure clean comparisons for each treated unit-period
- Provides event study results: time-varying effect estimates

Disadvantages/Limitations:

- Not necessarily efficient weighting (especially if only use last period)
- Complex to decide, fit, and interpret: many researcher degrees of freedom
- Complicates covariate adjustment

Summary of Options

Review/survey papers:

- Baker et al. (2022)
- Butts and Gardner (2022)
- Roth et al. (2023)
- de Chaisemartin and d’Haultfoeuille (2023)
- Roth (2024)
- Wing et al. (2024)

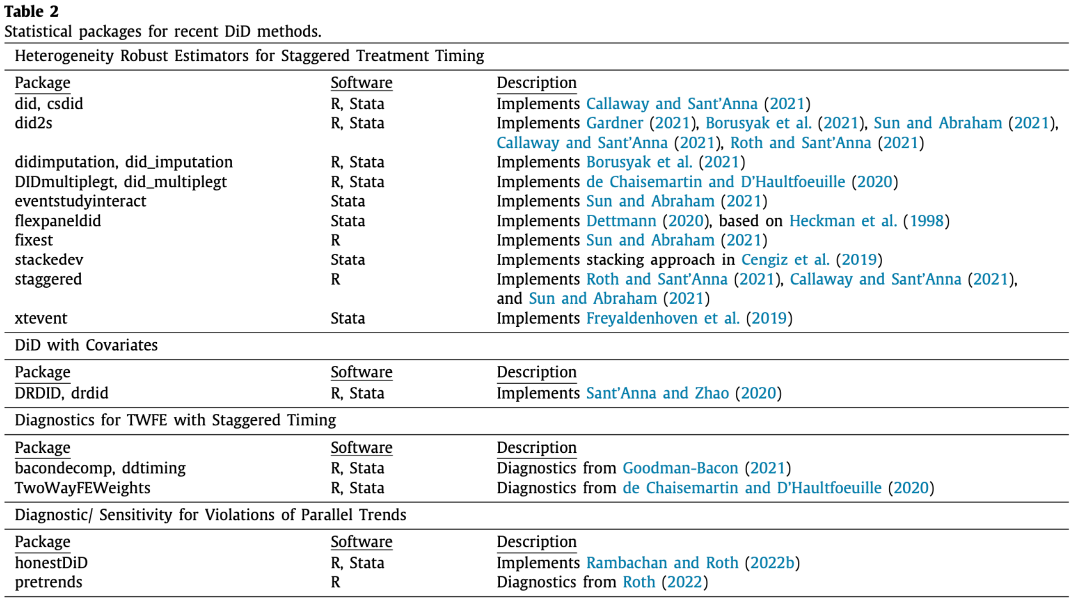


Table 2
Statistical packages for recent DiD methods.

Heterogeneity Robust Estimators for Staggered Treatment Timing		
Package	Software	Description
did, csdid	R, Stata	Implements Callaway and Sant’Anna (2021)
did2s	R, Stata	Implements Gardner (2021), Borusyak et al. (2021), Sun and Abraham (2021), Callaway and Sant’Anna (2021), Roth and Sant’Anna (2021)
didimputation, did_imputation	R, Stata	Implements Borusyak et al. (2021)
DIDmultiplegt, did_multiplegt	R, Stata	Implements de Chaisemartin and D’Haultfoeuille (2020)
eventstudyinteract	Stata	Implements Sun and Abraham (2021)
flexpaneldid	Stata	Implements Dettmann (2020), based on Heckman et al. (1998)
fixest	R	Implements Sun and Abraham (2021)
stackedev	Stata	Implements stacking approach in Cengiz et al. (2019)
staggered	R	Implements Roth and Sant’Anna (2021), Callaway and Sant’Anna (2021), and Sun and Abraham (2021)
xtevent	Stata	Implements Freyaldenhoven et al. (2019)
DiD with Covariates		
Package	Software	Description
DRDID, drdid	R, Stata	Implements Sant’Anna and Zhao (2020)
Diagnostics for TWFE with Staggered Timing		
Package	Software	Description
bacondecomp, ddtiming	R, Stata	Diagnostics from Goodman-Bacon (2021)
TwoWayFEWeights	R, Stata	Diagnostics from de Chaisemartin and D’Haultfoeuille (2020)
Diagnostic/ Sensitivity for Violations of Parallel Trends		
Package	Software	Description
honestDiD	R, Stata	Implements Rambachan and Roth (2022b)
pretrends	R	Diagnostics from Roth (2022)

Roth et al. (2023), Table 2.

Recommendations

Recommendations

- Consider whether there is staggered adoption and display extent.

Recommendations

- Consider whether there is staggered adoption and display extent.
- State assumed homogeneities clearly and justify. Also state what heterogeneities are considered.

Recommendations

- Consider whether there is staggered adoption and display extent.
- State assumed homogeneities clearly and justify. Also state what heterogeneities are considered.
- Report TWFE estimate and its decomposition.

Recommendations

- Consider whether there is staggered adoption and display extent.
- State assumed homogeneities clearly and justify. Also state what heterogeneities are considered.
- Report TWFE estimate and its decomposition.
- Report event study (dynamic regression) specification results.

Recommendations

Recommendations

- Consider alternative weightings and restrictions:

Recommendations

- Consider alternative weightings and restrictions:
    - Is there a never-treated group? Or last-treated group that can be excluded?

Recommendations

- Consider alternative weightings and restrictions:
    - Is there a never-treated group? Or last-treated group that can be excluded?
    - What is the desired estimand? Value of bias vs. variance vs. interpretability

Recommendations

- Consider alternative weightings and restrictions:
  - Is there a never-treated group? Or last-treated group that can be excluded?
  - What is the desired estimand? Value of bias vs. variance vs. interpretability
  - Is parallel trends reasonable across all units? With or without covariate adjustment or restrictions?

Recommendations

- Consider alternative weightings and restrictions:
    - Is there a never-treated group? Or last-treated group that can be excluded?
    - What is the desired estimand? Value of bias vs. variance vs. interpretability
    - Is parallel trends reasonable across all units? With or without covariate adjustment or restrictions?
- Select main approach (ideally pre-specified), conduct, and report clearly (with necessary assumptions).

Other Challenges

Parallel Trends and Covariates

These are complicated by staggered adoption and the longer time frames implied by panel data. Recent work has focused on how to interpret and test for these assumptions and how to incorporate time-varying covariates.

Other Assumptions and Limitations

The no-anticipation (or known/limited anticipation) assumption still must hold, as must the no-spillover assumption.

Other Assumptions and Limitations

The no-anticipation (or known/limited anticipation) assumption still must hold, as must the no-spillover assumption.

All of these approaches change the precise specification of the estimand as well: the **ATT** must be interpreted in terms of the included time periods, lags, and units, and how they are weighted.

When is DID useful? When is it lacking?

Important

It’s easy to ignore the fundamentals when using the more advanced methods. Consider the validity of the data, the question being asked, and the feasibility of the effect.

Questions?