Class 11: Difference-in-Differences (Advanced)

POLS 6388: Causal Inference

Michael Kistner
Department of Political Science, University of Houston

April 01, 2025

Goals of Today's Class

1. Learn how to incorporate covariates into a DiD design
2. Understand the problems that arise in the staggered treatment rollout context
3. Explore the Callaway and Sant'Anna group-time ATT approach (implemented with the `did` R package), which accomodates most advanced DiD use cases

Incorporating Covariates in DiD

Re-introducing the DiD Design

Last week, we covered the canonical DiD design

- **The idea:** Address all confounders that are constant *within unit* or *within time periods* by comparing over-time change within a treated unit to over-time change in a control unit.
- Via the parallel trends assumption, can estimate a valid ATT
- Showed how to estimate using an interaction term (in the 2x2 setup) or with two-way FEs

Why two sessions on DiD? 1) It's used a *lot*, and 2) Many recent developments

A Trending Literature

Table 2: Statistical Packages for Recent DiD Methods

Package	Software	Description
Heterogeneity Robust Estimators for Staggered Treatment Timing		
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'Haultfœuille (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'Haultfœuille (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)

Note: This table lists R and Stata packages for recent DiD methods, and is based on Asjad Naqvi's repository at https://asjadnaqvi.github.io/DiD/. Several of the packages listed under "Heterogeneity Robust Estimators" also accommodate covariates.

A Trending Literature

Table 2: Statistical Packages for Recent DiD Methods

Package	Software	Description
Heterogeneity Robust Estimators for Staggered Treatment Timing		
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'Haultfœuille (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'Haultfœuille (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)

Note: This table lists R and Stata packages for recent DiD methods, and is based on Asjad Naqvi's repository at https://asjadnaqvi.github.io/DiD/. Several of the packages listed under "Heterogeneity Robust Estimators" also accommodate covariates.

Remembering the Basic DiD Setup

Recall the general set-up for the DiD design:

- Two groups of units i, treated ($D_i = 1$) and untreated ($D_i = 0$)
- A pre-treatment time period ($t = 0$) and a post-treatment time period ($t = 1$)

We want to estimate the ATT, which we can write as:

$$\tau_{ATT} = \mathbf{E}[Y_{i1}(1) - Y_{i1}(0)|D_i = 1]$$

This is identifiable given the Parallel Trends Assumptions (PTA):

$$\mathbf{E}[Y_{i1}(0) - Y_{i0}(0)|D_i = 1] = \mathbf{E}[Y_{i1}(0) - Y_{i0}(0)|D_i = 0]$$

Time-Varying Covariates in a DiD Model

One estimator we used for the DiD design was the 2WFE model. In practice, 2WFE models often include time-varying covariates, i.e.,

$$Y_{it} = \hat{\gamma}_i + \hat{\lambda}_t + \hat{\tau}_{2WFE} Post_{it} + \underbrace{\mathbf{X}_{it}}_{Covariates} \hat{\beta} + \epsilon_{it}$$

Two rationales:

1. Treatment is non-random, and $\mathbf{X}_{it}$ may include an omitted variable associated with both treatment assignment ($Post_{it}$) and outcome (Y_{it})
2. We might be better able to obtain more precise estimates by accounting for additional variables that affect outcome (but not treatment)

Assumptions When Using Time-Varying Covariates

When modeled in this manner, researchers are (implicitly) making the following assumptions:

1. All covariates in $\mathbf{X}_{it}$ are pre-treatment (not affected by D_i)

2. Treatment effects are homogeneous in $\mathbf{X}_{it}$, i.e.,

$$\mathbf{E}[Y_{i1}(1) - Y_{i1}(0)|X, D_i = 1] = \hat{\tau}_{2WFE}$$

3. There are no X-specific trends for either treated or untreated units, i.e.,

$$\mathbf{E}[Y_1 - Y_0|X, D = d] = \mathbf{E}[Y_1 - Y_0|D = d] \quad \text{for} \quad d \in \{0, 1\}$$

If these assumptions don't hold, $\hat{\tau}_{2WFE} \neq \tau_{ATT}$

Allowing for Conditional TEs + Parallel Trends (1/2)

Assumption 1 (no post-treatment covariates) cannot be relaxed

Assumptions 2 and 3 *can be relaxed*, however. Treatment effects and trends might be differet in different subgroups.

Example: What is the effect of passing a distracted driving law ($Post_{it}$) on traffic fatalities (Y_{it}) in specific states i?

- Large states (X_i) probably have larger effects
- If there are trends (e.g., roads becoming more dangerous), larger states might have larger (steeper) trends

Allowing for Conditional Parallel Trends (2/2)

The Conditional Parallel Trends Assumption (CPTA) - Outcome trends between treated and untreated groups are similar after adjusting for covariates $\mathbf{X}_i$:

$$\mathbf{E}[Y_{i1}(0) - Y_{i0}(0)|\mathbf{X}_i, D_i = 1] = \mathbf{E}[Y_{i1}(0) - Y_{i0}(0)|\mathbf{X}_i, D_i = 0]$$

There are two types of estimators for implementing this adjustment:

1. **Regression-augmented estimators**
2. **Matching-/Weighting-augmented estimators**

Let's discuss versions of each, as presented by Sant'Anna and Zhao (2020)…

Outcome Regression Adjustment

Outcome Regression (OR) Adjustment – A procedure for estimating a DiD design with pre-treatment covariates using a regression model to account for conditionally differential trends:

$$\hat{\tau}_{OR} = \mathbf{E}[Y_{i1} - Y_{i0} - \hat{Y}_{\Delta}(\mathbf{X_i})|D_i = 1]$$

where $\hat{Y}_{\Delta}$ is predicted difference in outcomes as a function of pre-treatment covariates, estimated using untreated observations ($D_i = 0$).

- Most often estimated using a linear first difference regression, i.e.,

$$\mathbf{E}[Y_{i1} - Y_{i0}|\mathbf{X_i}, D_i = 0] = \underbrace{\hat{Y}_{\Delta} = \mathbf{X}\hat{\beta}}_{\text{First Difference Regression}}$$

Estimating an Outcome Regression Adjusted DiD

Steps to estimate the OR-DiD ATT:

1. Among all untreated units, estimate $\hat{Y}_{\Delta} = \hat{f}(\mathbf{X})$
   - If using linear regression, $\hat{f}(\mathbf{X}) = \mathbf{X}\hat{\beta}$
2. For each treated unit i, calculate $\hat{\tau}_i = Y_{i1} - Y_{i0} - \hat{Y}_{\Delta}(\mathbf{X_i})$
3. To obtain $\hat{\tau}_{ATT}$, simply average the individual effects: $\frac{1}{n_{treat}} \sum_{i=1}^{n_{treat}} \hat{\tau}_i$
   - *Note:* This provides a sample ATT. Generalizing to a population requires assuming that your sample is representative of the broader population in terms of $Y_{i1}, Y_{i0}, D_i, \mathbf{X}_i$

Generating Data

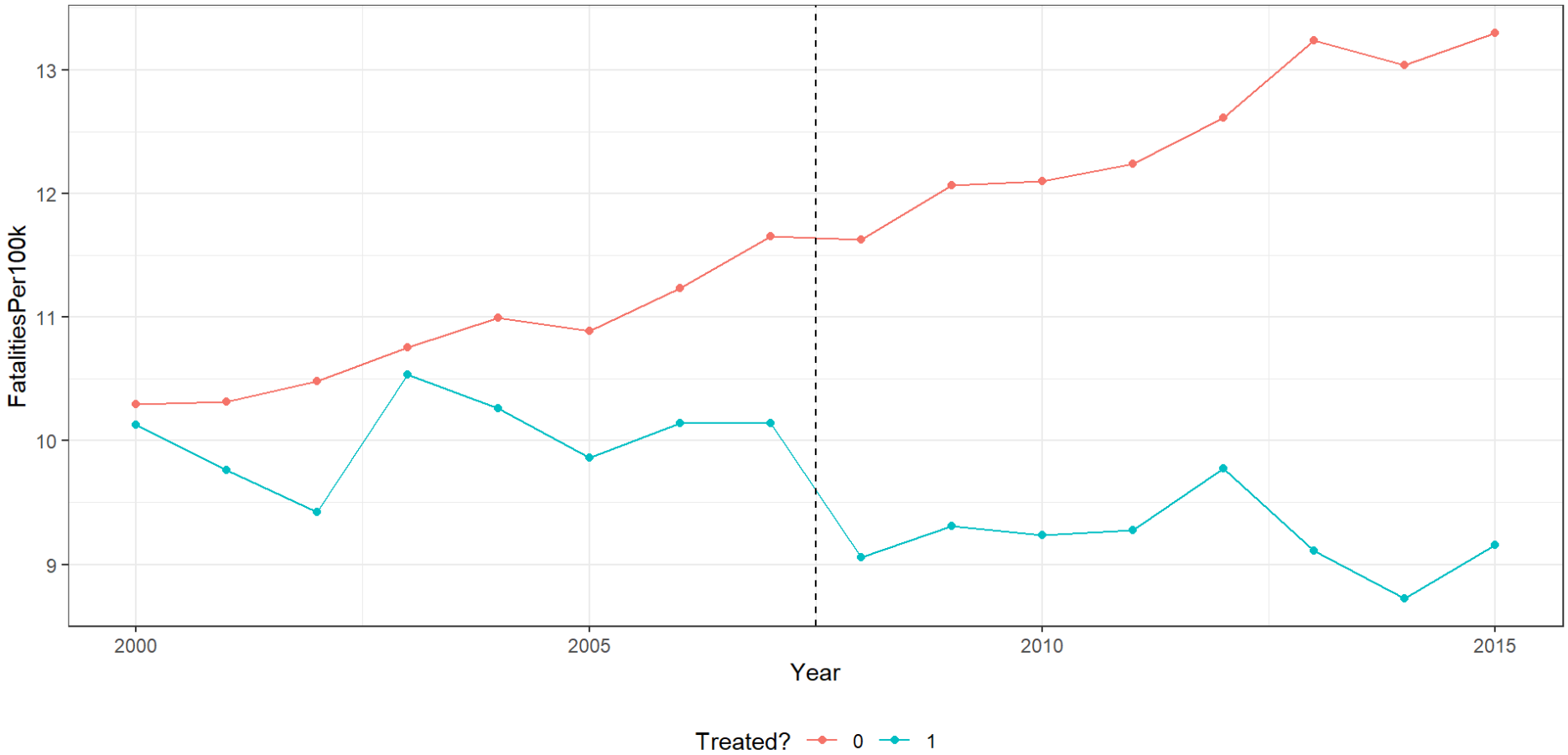
```
set.seed(2021)

distracted_df <- read_csv("distracted.csv") %>%
  mutate(SunBelt = ifelse(State %in% c("FL", "SC", "NC", "GA", "AL",
                                       "MS", "LA", "TX", "NM", "AZ"), 1, 0),
    Treated = ifelse(State %in% c("RI", "VT", "DE", "HI", "NH"), 1, 0),
    Post = ifelse(Year>= 2008, 1, 0),
    DD_Law = ifelse(Treated == 1 & Post == 1, 1, 0),
    FatalitiesPer100k = 10  + SunBelt * (Year - 2000) -
      1 * DD_Law + rnorm(800, mean = 0, sd = 1))
```

Implies $\hat{\tau}_{ATT} = 1$ fatality per 100k population per year.

```
# A tibble: 5 × 7
  State  Year SunBelt Treated  Post DD_Law FatalitiesPer100k
  <chr> <dbl>   <dbl>   <dbl> <dbl>  <dbl>             <dbl>
1 AL     2000       1       0     0      0              9.88
2 AL     2001       1       0     0      0             11.6
3 AL     2002       1       0     0      0             12.3
4 AL     2003       1       0     0      0             13.4
5 AL     2004       1       0     0      0             14.9
```

Parallel Pre-Trends?



The (un-adjusted) 2WFE Estimate

```
reg_1 <- feols(FatalitiesPer100k ~ DD_Law |
                 State + Year,
               data = distracted_df)

summary(reg_1)
```

```
OLS estimation, Dep. Var.: FatalitiesPer100k
Observations: 800
Fixed-effects: State: 50,  Year: 16
Standard-errors: Clustered (State)
       Estimate Std. Error  t value    Pr(>|t|)
DD_Law -2.52441   0.553479 -4.56098 0.000034218 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 2.11276     Adj. R2: 0.668208
                Within R2: 0.031122
```

Estimating the Outcome Regression

```r
# Create dataset with pre-/post- outcome differences among UNTREATED states
fd_untreated <- distracted_df %>%
  filter(Treated == 0, Year %in% c(2007, 2009)) %>%
  group_by(State, SunBelt) %>%
  summarize(DeltaY = FatalitiesPer100k[Year == 2009] - FatalitiesPer100k[Year == 2007])

# Estimate first-difference regression
reg_adjust <- feols(DeltaY ~ SunBelt,
                    data = fd_untreated)

# What's our estimated difference by state type?
summary(reg_adjust)
```

```
OLS estimation, Dep. Var.: DeltaY
Observations: 45
Standard-errors: IID
            Estimate Std. Error  t value  Pr(>|t|)
(Intercept) 0.004609   0.274884 0.016767 0.9866998
SunBelt     1.857287   0.583118 3.185098 0.0026926 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 1.58969   Adj. R2: 0.172074
```

Estimating the OR-Adjusted DiD

```r
# Create dataset with ACTUAL pre-/post- outcome differences among TREATED states
fd_treated <- distracted_df %>%
  filter(Treated == 1, Year %in% c(2007, 2009)) %>%
  group_by(State, SunBelt) %>%
  summarize(DeltaY = FatalitiesPer100k[Year == 2009] - FatalitiesPer100k[Year == 2007])

# Estimate PREDICTED pre-/post- outcome differences for treated states
y_hat <- predict(reg_adjust, fd_treated)

# Subtract predicted differences Y_hat from actual differences
tau_i <- fd_treated$DeltaY - y_hat
```

```
[1] "DE:0.49, HI:-0.9, NH:0.25, RI:-2.35, VT:-1.65"
```

```r
# Average over observations
mean(tau_i)
```

```
[1] -0.832
```

Matching / Weighting Adjustment

Matching / Weighting Adjustment for DiD – A procedure for estimating a DiD with pre-treatment covariates using a matching or weighting model to account for conditionally differential trends.

For example, an **inverse propensity weighting** estimator of the ATT:

$$\hat{\tau}_{IPW} = \frac{1}{\mathbf{E}[D_{it}]}\mathbf{E}\left[\frac{D - \hat{\pi}(\mathbf{X_i})}{1 - \hat{\pi}(\mathbf{X_i})}\left(Y_{i1} - Y_{i0}\right)\right],$$

where…

- $\hat{\pi}(\mathbf{X_i})$ is an estimate of the propensity score (i.e., probability of treatment)

Estimating an IPW-Adjusted DiD

Steps to estimate the IPW DiD ATT:

1. Estimate a binary regression model of treatment assignment probability using pre-treatment covariates $\mathbf{X_i}$
    - Sant'Anna and Zhao (2020) recommend a logistic regression:
2. Calculate a weighted average of the pre-/post- differences in Y_{it}.

- For treated units, the weight w_i equals 1
- For untreated units, the weight $w_i = \frac{\hat{\pi}}{1-\hat{\pi}}$

3. Divide by the proportion of treated units

Implementing the IPW-Adjusted DiD (1/2)

```r
# Estimate the propensity score model
reg_propensity <- feglm(Treated ~ SunBelt,
      family = binomial(link = "logit"),
      data = distracted_df)

# Create dataset with pre-/post- outcome differences among ALL states
fd_all <- filter(distracted_df, Year %in% c(2007, 2009)) %>%
  group_by(State, SunBelt, Treated) %>%
  summarize(DeltaY = FatalitiesPer100k[Year == 2009] - FatalitiesPer100k[Year == 2007])

# Create propensity score
fd_all$Propensity <- round(predict(reg_propensity, fd_all), 3)
```

```
# A tibble: 5 × 5
# Groups:   State, SunBelt [5]
  State SunBelt Treated DeltaY Propensity
  <chr>   <dbl>   <dbl>  <dbl>      <dbl>
1 AK          0       0 -1.10       0.125
2 AL          1       0  2.81       0
3 AR          0       0 -0.274      0.125
4 AZ          1       0  3.26       0
5 CA          0       0  1.53       0.125
```

Implementing the IPW-Adjusted DiD (2/2)

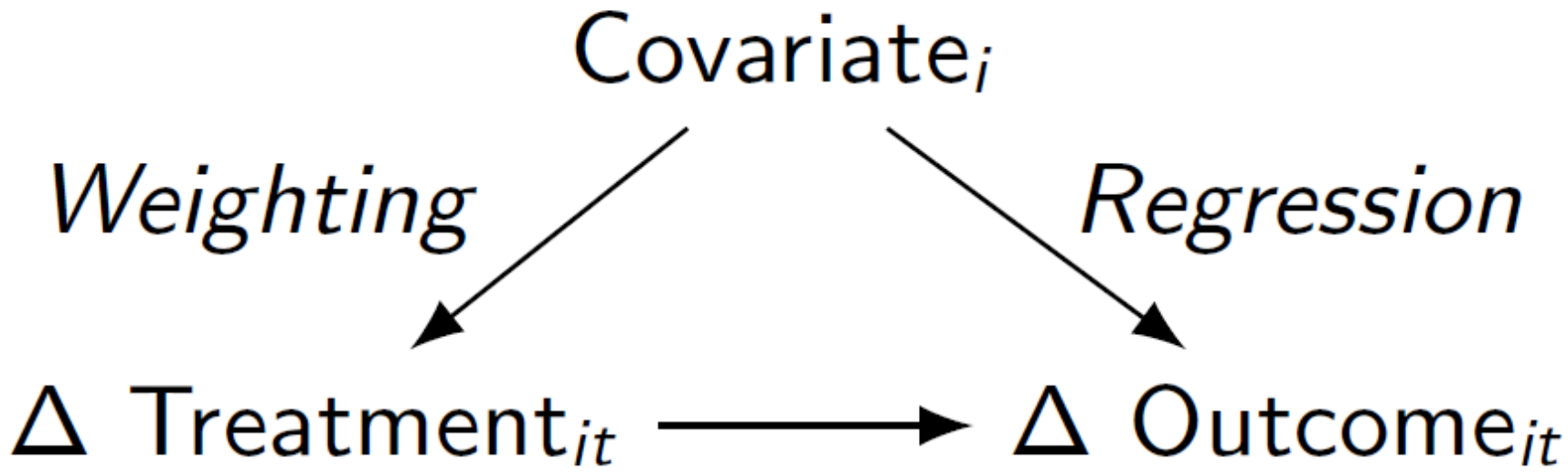
```r
# Create weights
fd_all <- mutate(fd_all,
                 weight = ifelse(Treated == 1, 1,
                                 -Propensity / (1 - Propensity)))
```

```
# A tibble: 6 × 6
# Groups:   State, SunBelt [6]
  State SunBelt Treated DeltaY Propensity weight
  <chr>   <dbl>   <dbl>  <dbl>      <dbl>  <dbl>
1 AK          0       0 -1.10       0.125 -0.143
2 AL          1       0  2.81       0      0
3 AR          0       0 -0.274      0.125 -0.143
4 AZ          1       0  3.26       0      0
5 CA          0       0  1.53       0.125 -0.143
6 CO          0       0  2.07       0.125 -0.143
```

```r
# Calculate weighted mean and divide by proportion treated
mean(fd_all$DeltaY * fd_all$weight) / mean(fd_all$Treated)
```

```
[1] -0.8319596
```

Intuition of The Two Approaches



A DAG of DiD with Covariates. Each of the two approaches closes one backdoor path.

A Doubly-Robust DiD Estimator (1/2)

Both approaches only work if modeled correctly.

- Regression adjustment assumes we can model the outcome
- Matching adjustment assumes we can model the propensity score

If done incorrectly, our estimate will suffer from misspecifiation bias

Fortunately, we don’t have to limit ourselves to one or the other…

A Doubly-Robust DiD Estimator (2/2)

Doubly-Robust Adjustment for DiD – a technique accounting for covariates using both matching/weighting *and* regression adjustment simultaneously

- Doubly-robust means if *either* matching or regression is done correctly, the estimate will be unbiased

The Sant'Anna and Zhao (2020) doubly-robust estimator:

$$\hat{\tau}_{DR} = \mathbf{E}\left[\frac{D_i}{\mathbf{E}[D_i]}\frac{(1-D_i)\hat{P}(\mathbf{X}_i)}{1-\mathbf{X}_i}\left(Y_{i1}-Y_{i0}-\hat{Y}_{\Delta}(\mathbf{X_i})\right)\right]$$

Good news: All of these estimators (including the doubly-robust DiD) work in the case of repeated cross-section data (not just panel data)

Understanding Issues with Staggered Treatment Timing

DiD in Multiple Time Periods

Staggered Rollout DiD - A DiD analysis where treatment occurs at different time periods for different units

- In practice, most DiD analyses have staggered rollouts

With staggered treatment timing, a 2WFE estimator will generally not return the ATT, even under random assignment of treatment

- Problem is caused by weighting + dynamics in treatment effects
- Event study analyses will be similarly biased

Let's walk through a step-by-step decomposition of these issues (from Goodman-Bacon 2021)…

Whose Comparison Is It, Anyway?

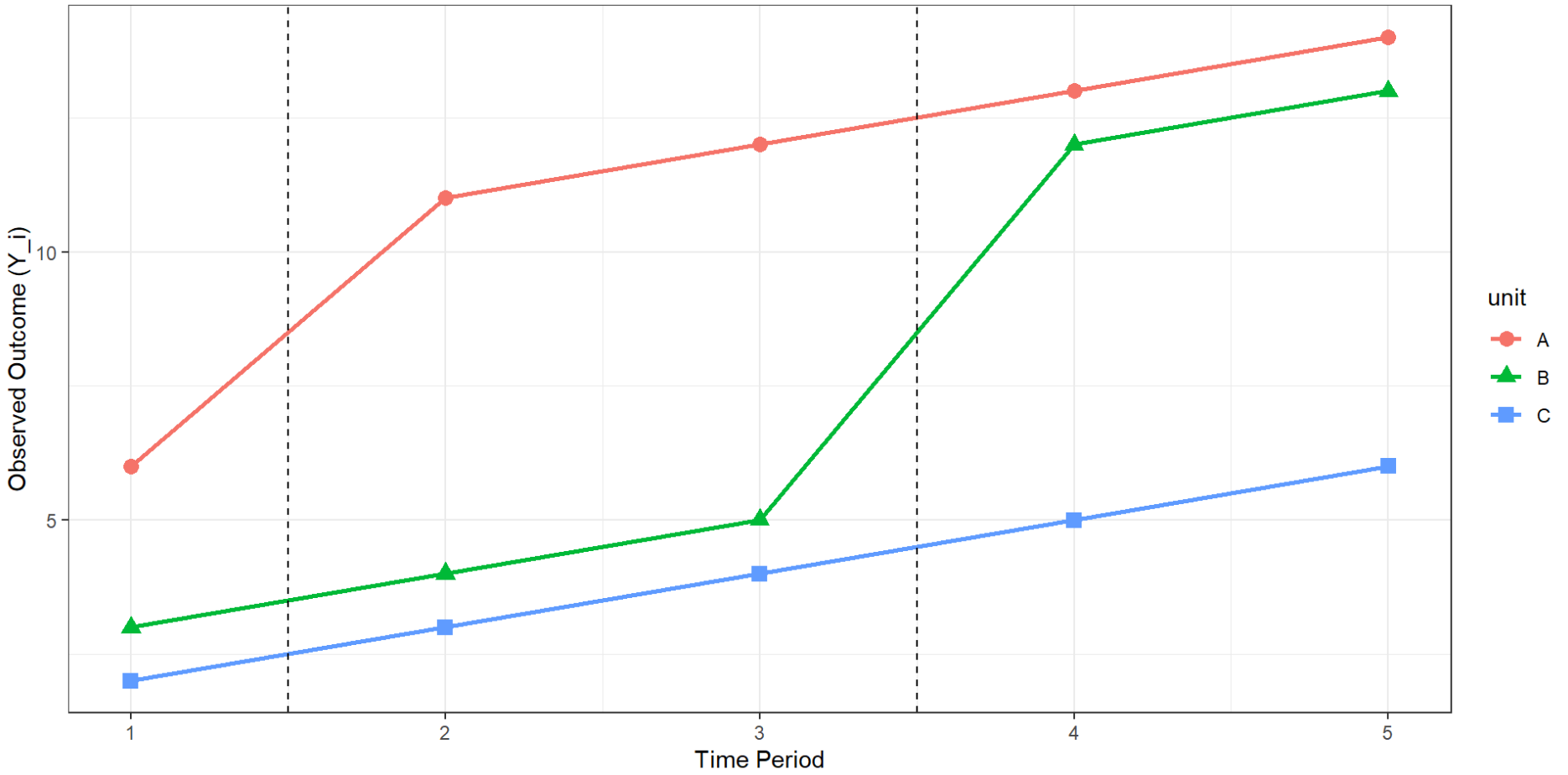
The problems emerge from the comparisons being made.

- When treatment is adopted by different groups at different times, who is the control group for each treated group? Not obvious…

Consider a simple example with three units $i \in \{A, B, C\}$ and five time periods $t \in \{1, \cdots, 5\}$.

- Unit A gets treated at time $t = 2$, unit B at time $t = 4$, and C never gets treated

Visualizing the Setup



2WFEs and Staggered Rollout

Suppose we use the basic 2WFE estimator to implement the DiD design.

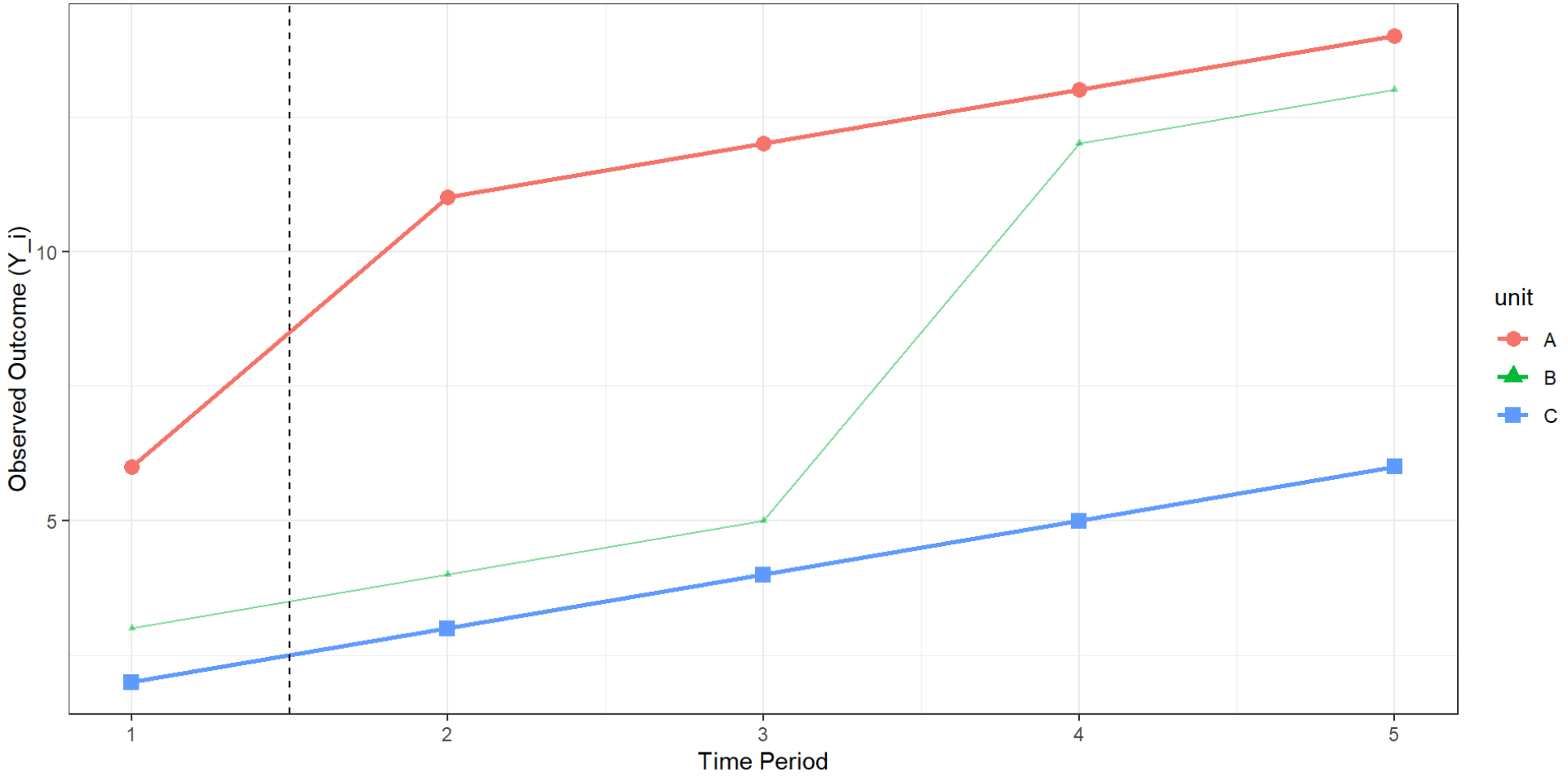
$$Y_{it} = \hat{\gamma}_i + \hat{\lambda}_t + \hat{\tau}_{2WFE} Post_{it} + \epsilon_{it}$$

What is $\hat{\tau}_{2WFE} Post_{it}$? What treatment effect, if any, are we estimating?

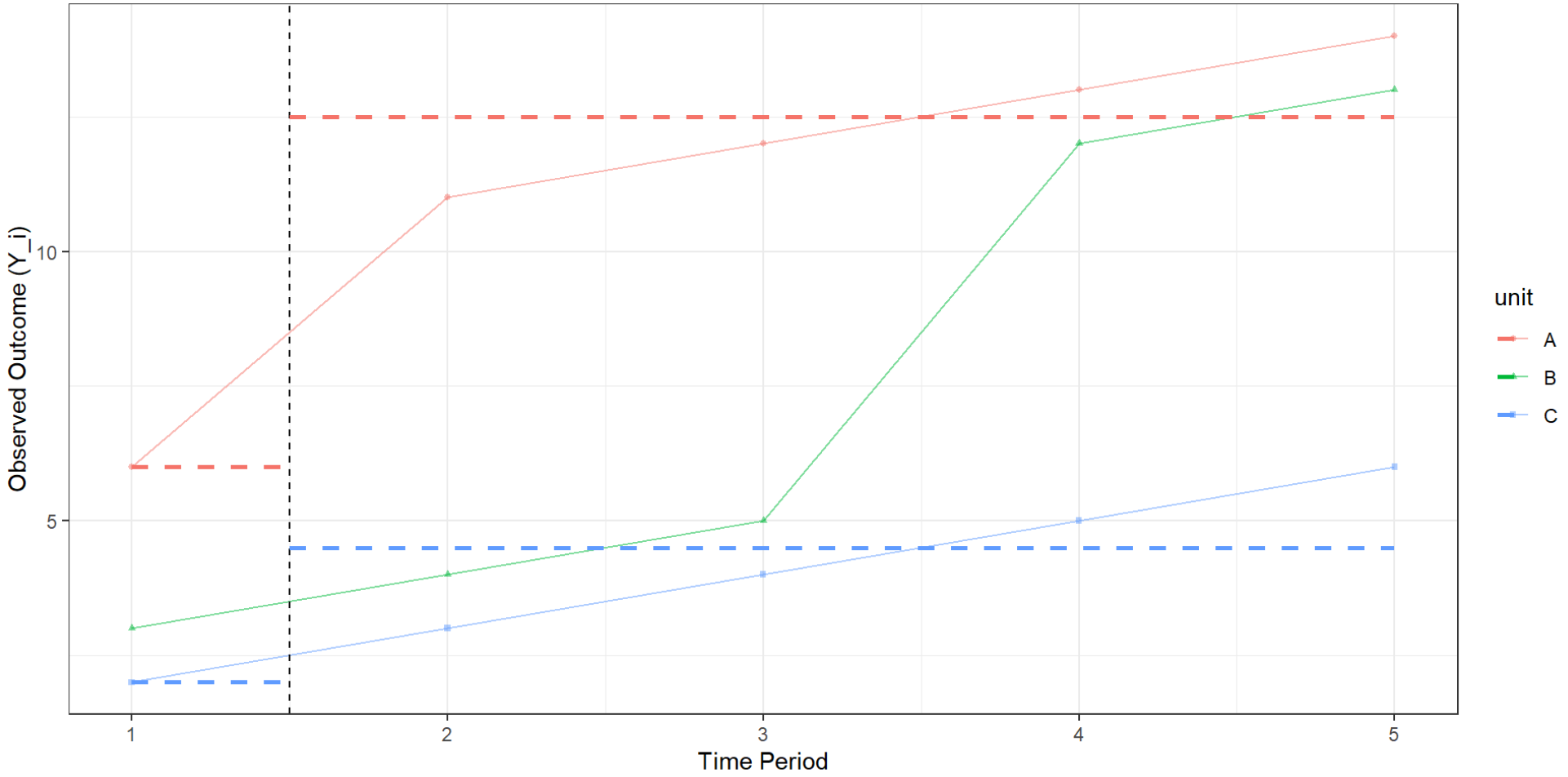
Answer: A weighted combination of (in this case, four) 2x2 comparisons

- Early Treated versus Never Treated
- Late Treated versus Never Treated
- Late Treated versus Early Treated
- Early Treated versus Late Treated

The First Comparison



Differences in Post-Pre Means



The First 2x2 Coefficient

This is a comparison that makes sense

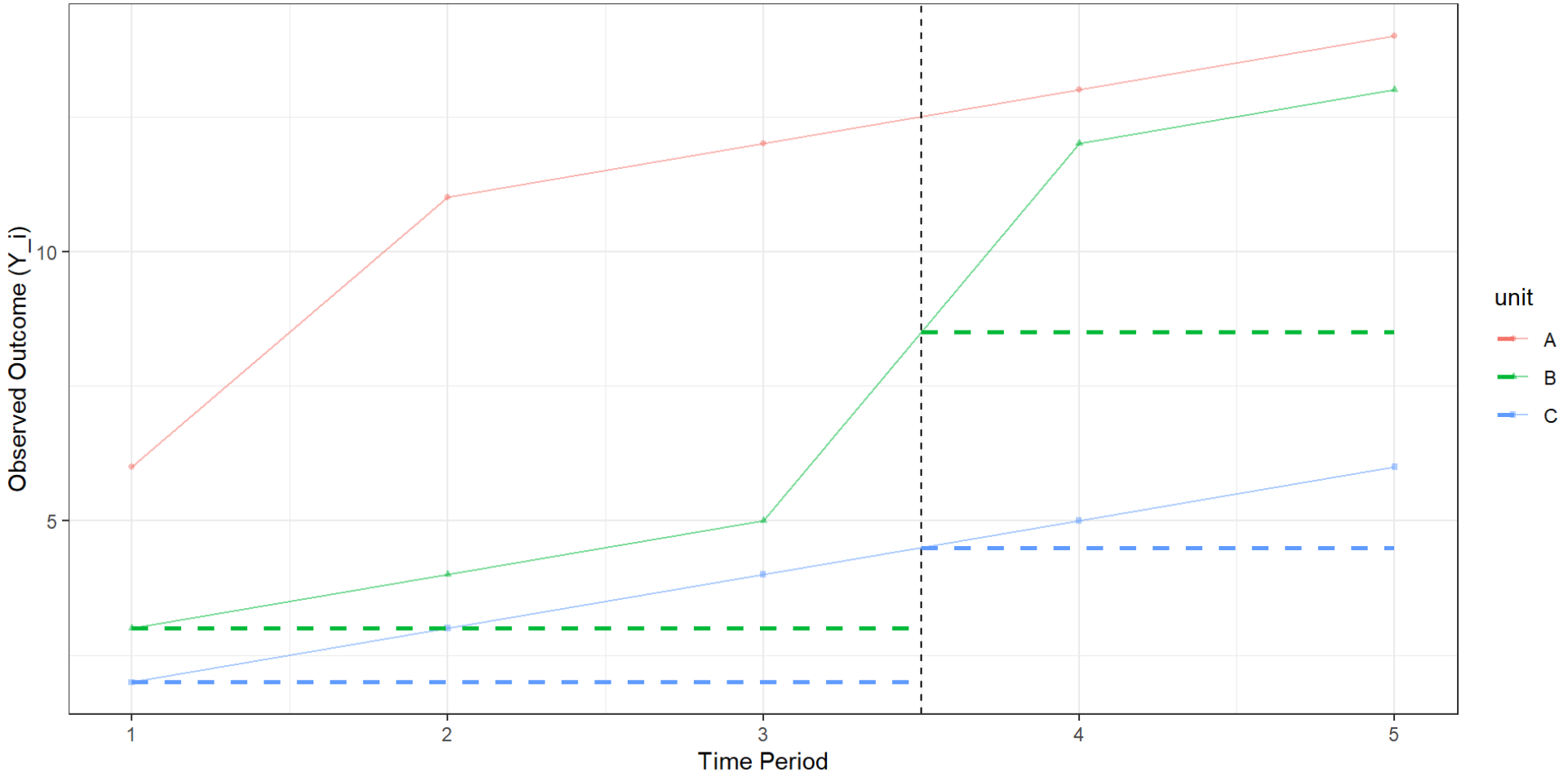
- A treated unit compared to a never-treated unit

Treatment occurs, calculate the difference in outcomes pre-/post- for the treated unit and subtract the pre-post difference for the never-treated unit

- Refer to this comparison as:

$$\hat{\tau}^{AC}_{2x2} = \left(\bar{Y}^{Post}_{A} - \bar{Y}^{Pre}_{A}\right) - \left(\bar{Y}^{Post}_{C} - \bar{Y}^{Pre}_{C}\right)$$

The Second Comparison



Another 2x2 Coefficient

This is another comparison that makes sense

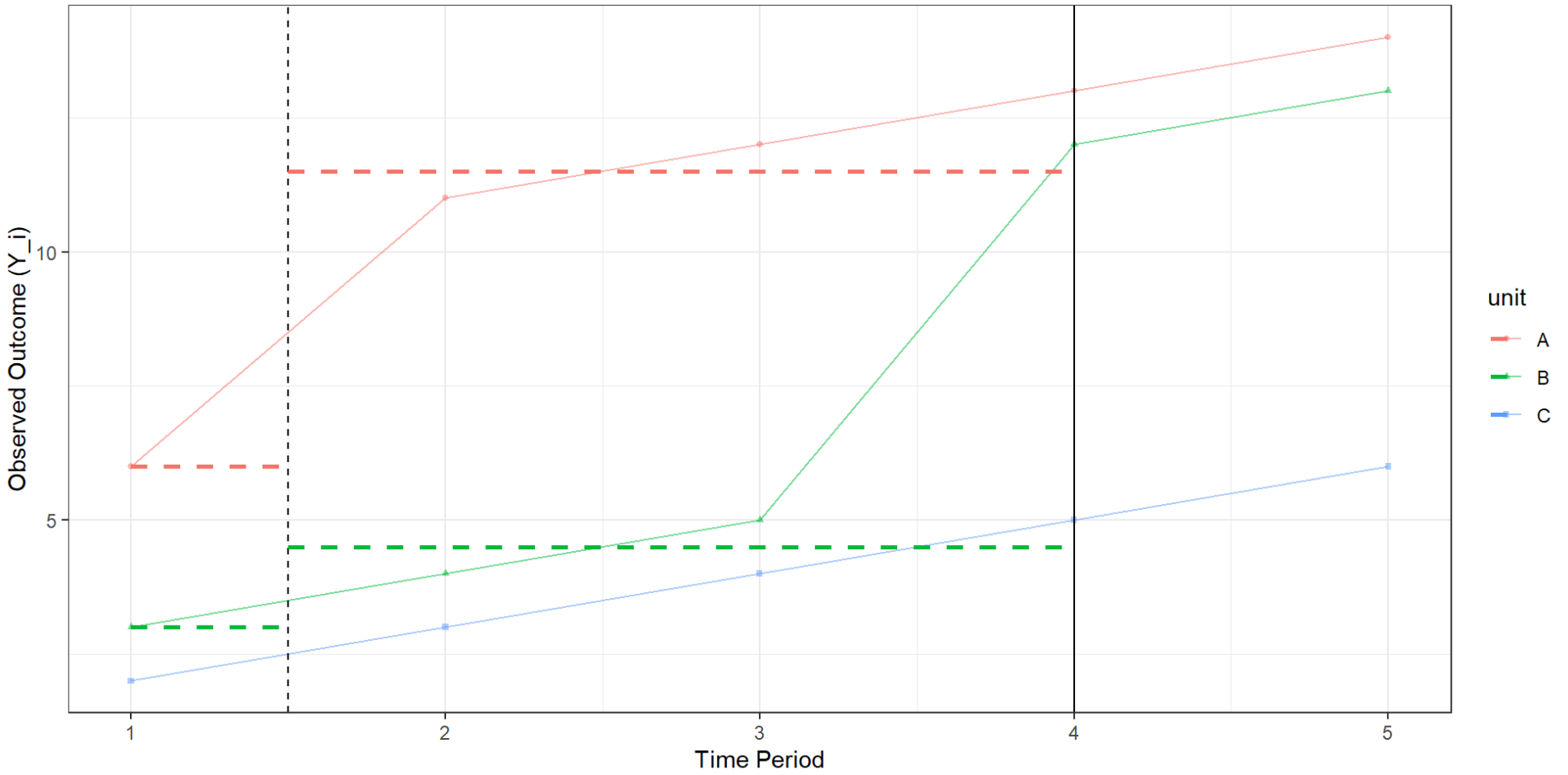
- A treated unit compared to a never-treated unit

Another treatment occurs, calculate the difference for this treated unit and subtract the difference for the never-treated unit

- Refer to this comparison as:

$$\hat{\tau}^{BC}_{2x2} = \left(\bar{Y}^{Post}_{B} - \bar{Y}^{Pre}_{B}\right) - \left(\bar{Y}^{Post}_{C} - \bar{Y}^{Pre}_{C}\right)$$

The Third Comparison



Yet Another 2x2 Coefficient

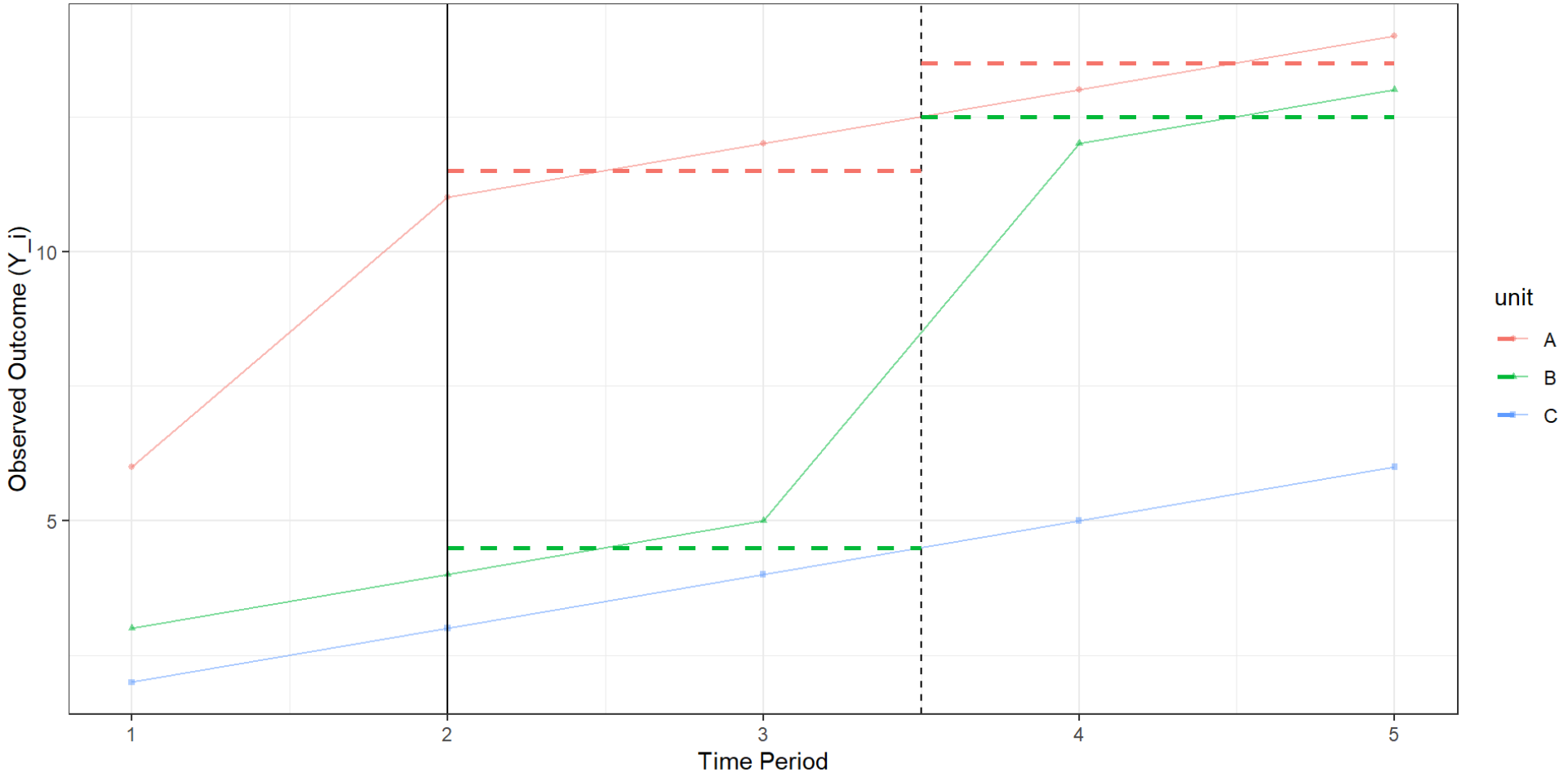
This is one more comparison that (probably) makes sense

- An early-treated unit compared to a late-treated unit *before its treatment*
- Refer to this comparison as:

$$\hat{\tau}^{AB}_{2x2} = \left(\bar{Y}^{Post}_{A} - \bar{Y}^{Pre}_{A}\right) - \left(\bar{Y}^{Post}_{B} - \bar{Y}^{Pre}_{B}\right)$$

where $Post$ refers to periods after A is treated but before B is.

The Fourth Comparison



The Final 2x2 Coefficient

This is probably **not** a comparison that makes sense

- A late-treated unit compared to an early-treated unit *after its treatment*
- Refer to this comparison as:

$$\hat{\tau}^{BA}_{2x2} = \left(\bar{Y}^{Post}_{B} - \bar{Y}^{Pre}_{B}\right) - \left(\bar{Y}^{Post}_{A} - \bar{Y}^{Pre}_{A}\right)$$

where Pre refers to before B is treated but after A is.

Decomposing the 2WFE Coefficient

The 2WFE-estimated coefficient gives us the following:

$$\hat{\tau}_{2WFE} = w^{AC}\hat{\tau}^{AC}_{2x2} + w^{BC}\hat{\tau}^{BC}_{2x2} + w^{AB}\hat{\tau}^{AB}_{2x2} + w^{BA}\hat{\tau}^{BA}_{2x2}$$

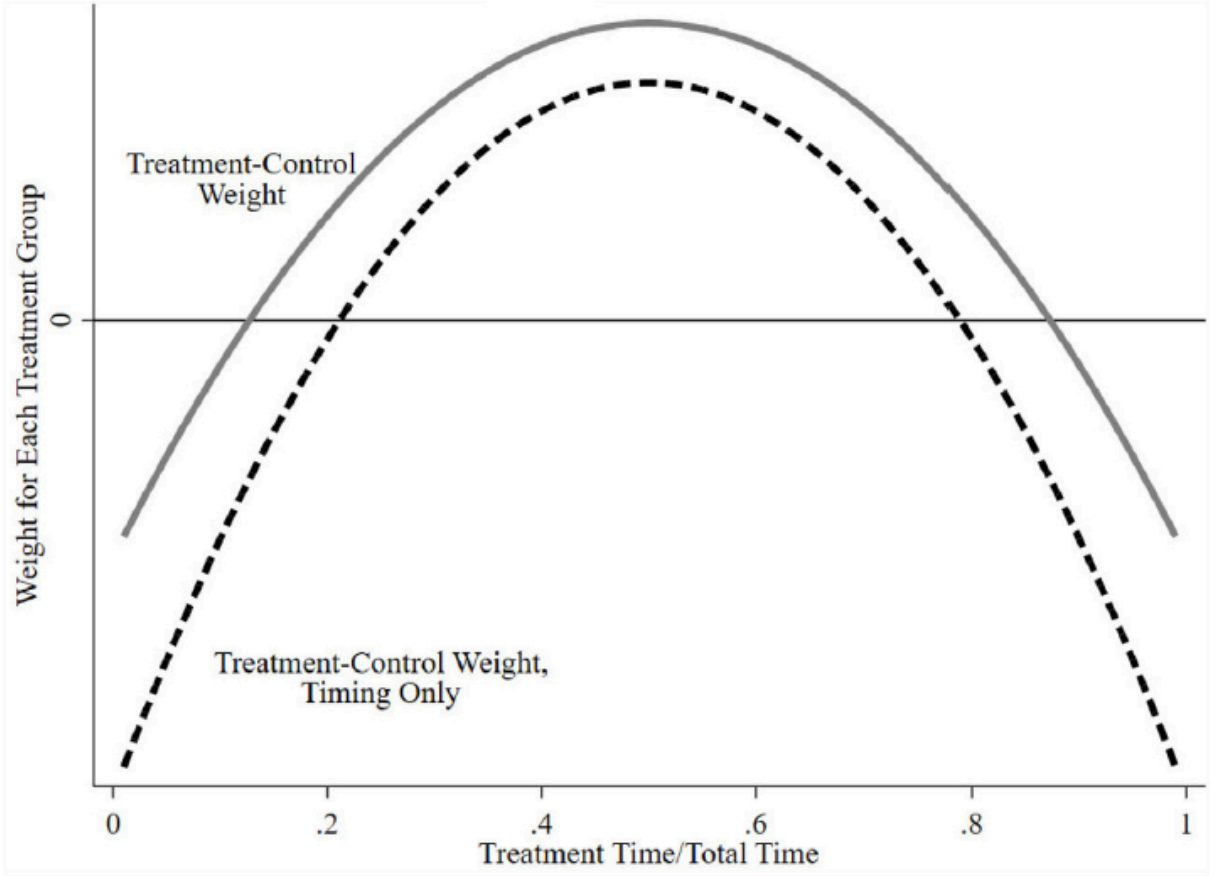
where w^{ij} is a weight that depends on two factors:

- $\uparrow$ in cohort size (# of units treated at time t)
- $\uparrow$ in treatment variance (treated in middle of time period)

In general, with K treatment cohorts, have $\frac{K!}{2!(K-2)!} + K$ comparisons

- Comparisons are always Treated vs. Never Treated, Earlier vs. Later Treated, and Later vs. Earlier Treated, however.

Visualizing Weights and Treatment Timing



Which Cohorts Are Weighted Most Heavily in a Staggered 2WFE Model

Estimands and Identifying Assumptions

If treatment effects are constant, $\hat{\tau}_{2WFE} = \tau_{ATT}$

- Identified under the typical PTA

If treatment effects vary across units, $\hat{\tau}_{2WFE} = \tau_{VWATT}$

- Identification now depends on a *Variance-Weighted PT* assumption

$$
\begin{aligned}
VWCT \equiv & \sum_{k \neq U} \sigma_{kU} \left[\Delta Y_k^0 \left(POST\left(k\right), PRE\left(k\right) \right) - \Delta Y_U^0 \left(POST\left(k\right), PRE\left(k\right) \right) \right] \\
& + \sum_{k \neq U} \sum_{\ell > k} \left[\sigma_{k\ell}^{k} \left\{ \Delta Y_k^0 \left(MID\left(k, \ell\right), PRE\left(k\right) \right) - \Delta Y_\ell^0 \left(MID\left(k, \ell\right), PRE\left(k\right) \right) \right\} \right. \\
& \left. + \sigma_{k\ell}^{\ell} \left\{ \Delta Y_\ell^0 \left(POST\left(\ell\right), MID\left(k, \ell\right) \right) - \Delta Y_k^0 \left(POST\left(\ell\right), MID\left(k, \ell\right) \right) \right\} \right].
\end{aligned}
$$

Time-Heterogeneous TEs

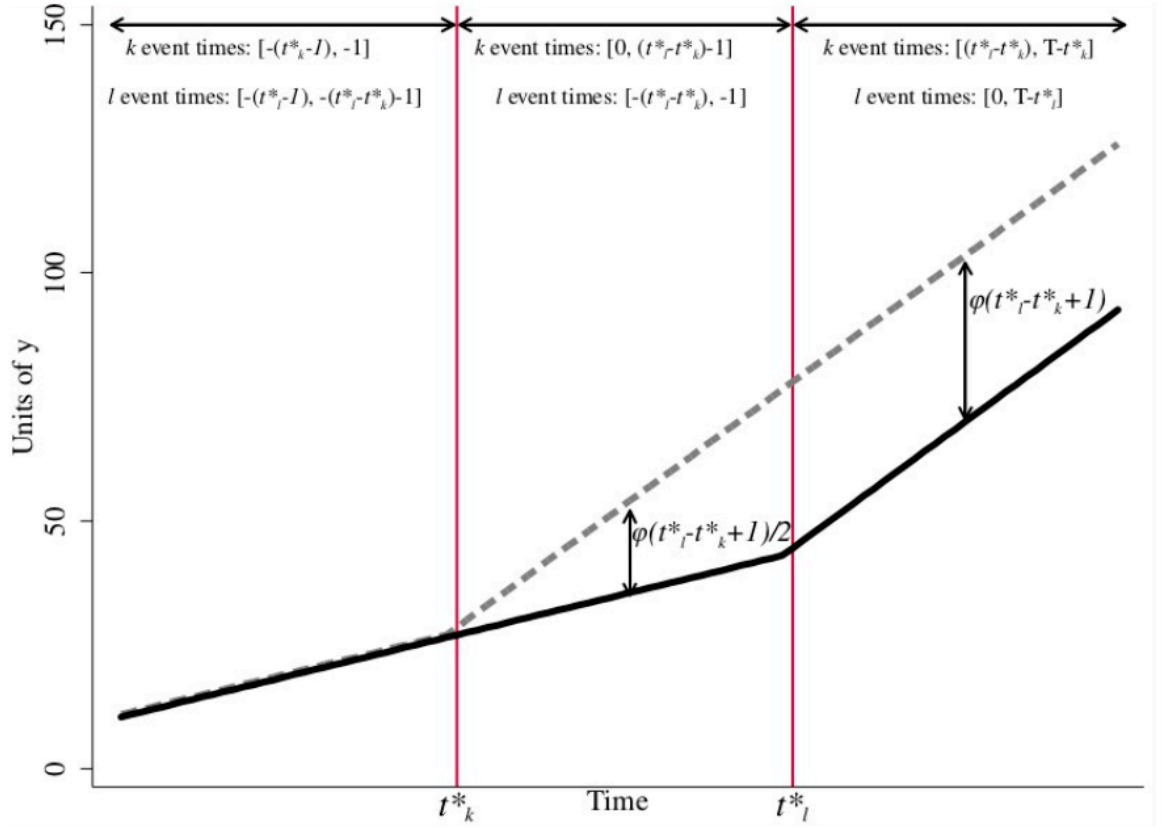
Worst case scenario – treatment effects that vary *within unit, across time*

- **Example:** effects that increase gradually as distance from treatment increases
- Does enforcement of/compliance with DD laws strengthen over time?

Biases arise because of *comparing Late-Treated to Early-Treated* groups.

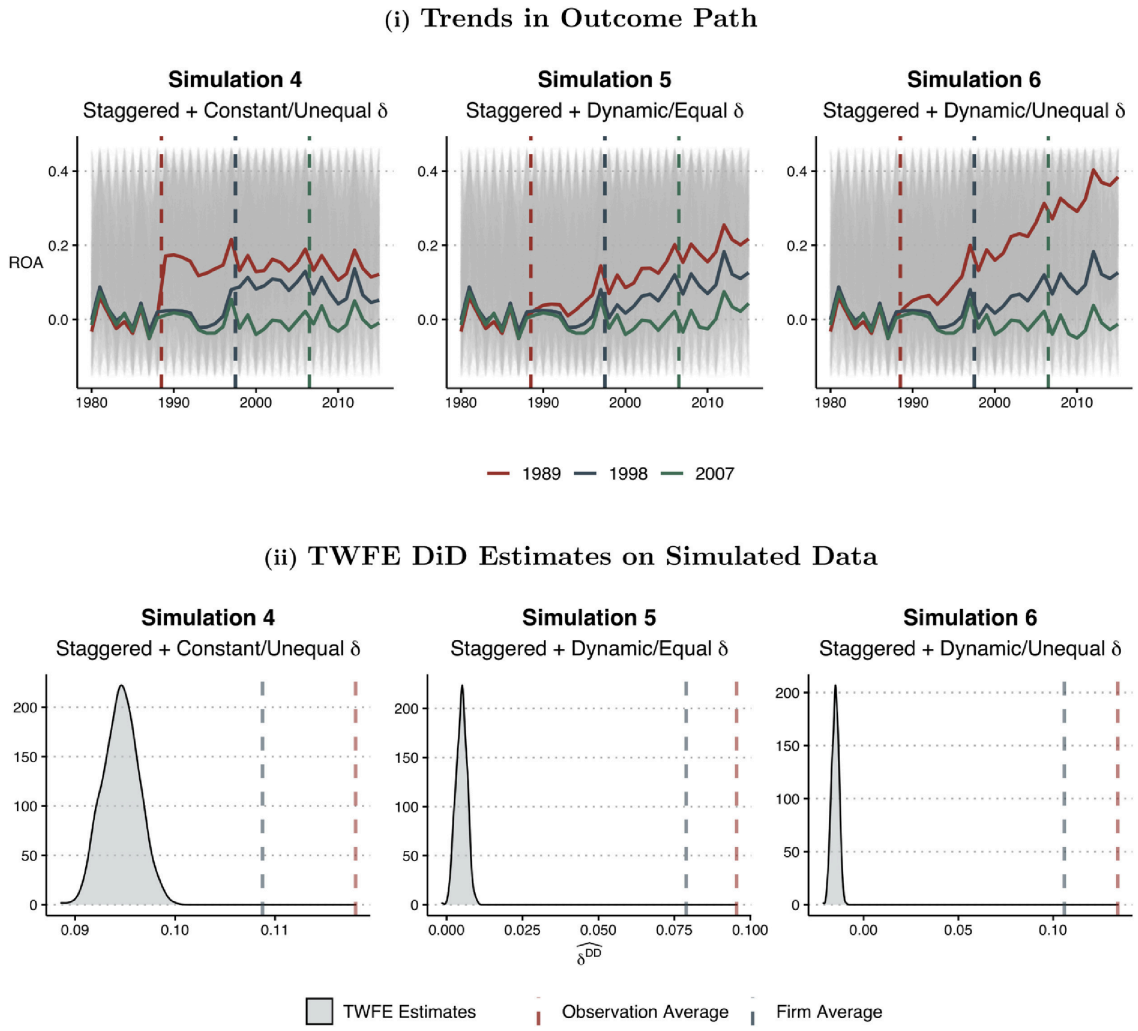
- Early-Treated make a poor comparison for Late-Treated under a counterfactual of no treatment, because former are being affected by treatment.
- The estimator will "subtract out" evolving treatment effects among the Early-Treated and attribute the difference to the effect for the Late-Treated

Visualizing the Problem



Effect Sizes that Vary Over Time Bias Our Estimates

Heterogeneity Bias



In Extreme Cases, Bias Can Reverse Sign of Estimated Effect

The Bacon-Goodman Decomposition

If using 2WFE models, should determine how much particular comparisons (especially Late-Treated to Early-Treated) shape our inferences

Bacon-Goodman Decomposition – the disaggregation of $\hat{\tau}_{2WFE}$ into the component 2x2 comparisons

- How do effect estimates differ between Treated vs. Never Treated, Earlier vs. Later Treated, and Later vs. Earlier Treated comparison?
- How much do each of these comparisons contribute to the overal estimated TE $\hat{\tau}_{2WFE}$?

Applying the Bacon-Goodman Decomposition

```r
library(bacondecomp)

bacon(outcome ~ treated,
      id_var = "unit",
      time = "time",
      data = staggered_df)
```

```
                    type weight avg_est
1 Earlier vs Later Treated  0.125     4.0
2 Later vs Earlier Treated  0.250     6.0
3     Treated vs Untreated  0.625     5.2
```

```
  treated untreated estimate weight                     type
2       4         2        6  0.250 Later vs Earlier Treated
4       2         4        4  0.125 Earlier vs Later Treated
7       2     99999        4  0.250     Treated vs Untreated
8       4     99999        6  0.375     Treated vs Untreated
```

Addressing Issues with Staggered Treatment Timing

Overcoming Issues with Staggered Treatments

Several estimators have been developed that work better than 2WFE in the case of staggered treatment timing

A solution from a familiar name: Callaway and Sant'Anna (2021)

- Implemented in the `did` package
- Can be used with panel or repeated cross-section data
- Can be used with balanced or imbalanced panels
- Does require "once-treated, always treated"
  $Post_{i,t-1} = 1 \implies Post_{i,t} = 1$
  - For an estimator for other cases, see de Chaisemartin and D'Haultfoeuille (2020) and the `DIDmultiplegt` package

Outline of the CS Approach

Callaway and Sant'Anna propose a **three step process:**

1. Identify disaggregated effect parameters by treatment cohort, using only "clean" comparisons
2. Aggregate these parameters to summarize effect direction and size
3. Perform statistical inference via bootstrapping

Let's consider each of these in turn.

Recovering Disaggregated Effects

Assume no units are treated in time period $t = 1$. Then let $G_i = g \in \{2, 3, \cdots, T\}$ indicate the time period that unit i was first-treated

- Thus G can be thought of as a *treatment cohort*
- Can denote Never-Treated groups using $G_i = \infty$

The treatment effect estimands the CS DiD procedure estimates:

- Group-Time ATT(g, t) – The average treatment effect on the outcome at time t for group G where all cohort members are first-treated at time g

$$ATT_{g,t} = \mathbf{E}[Y_t(g) - Y_t(0)|G_g = 1]$$

Taking Care with Comparison Groups

How do we construct the counterfactual outcome? By comparing to…

- Either A) Never-Treated groups, or B) Not-Yet-Treated groups
- **NOT** already treated groups

Example: TX and RI both pass a DD law in 2005. They are both group $g = 2005$. Can estimate the ATT for this group in t = 2005, 2006, etc.

- MA (during our timespan) never passes a distracted driving law, can be used as a comparison for TX and RI (or any other groups G)
- MS doesn't pass a DD law until 2018. MS can be used as a control comparison for TX and RI up until 2018 (but not vice-versa)

Identifying Assumptions (1/2)

Let's consider what the ATT(g, t) gives us, using PO notation:

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

The ATT estimated in the 2x2 case: ATT$(g = 2, t = 2)$

The group-time ATT can be identified using either of two assumptions:

Parallel Trends of Never-Treated Units: For all $g, t = 2, \cdots, T$, with $t \geq g$,

$$\mathbf{E}[Y_{i,t}(0) - Y_{i,t-1}(0)|G_i = g] = \mathbf{E}[Y_{it}(0) - Y_{i,t-1}(0)|G_i = \infty]$$

where $C_i = 1$ means unit i is in the never-treated group.

Identifying Assumptions (2/2)

Parallel Trends of Not-Yet-Treated Units: For all $g, s, t = 2, \cdots, T$, with $t \geq g$ and $s \geq 2$,

$$\mathbf{E}[Y_{i,t}(0) - Y_{i,t-1}(0)|G_i = g] = \mathbf{E}[Y_{it}(0) - Y_{i,t-1}(0)|D_{is} = 0, G \neq g]$$

where $D_{is} = 1$ is an indicator whether unit i has been treated by time s.

If both assumptions hold, can use both groups to estimate the ATT(g, t)'s

- Is one group more comparable to the treated group? (anticipation versus selection effects)

Estimation of the Group-Time ATTs

Using either of these "clean" comparisons, we can estimate:

$$ATT(g,t) = \mathbf{E}\Big[Y_{it}(1) - Y_{it}(0) | G_i = g\Big]$$

For instance, by comparison to the Never-Treated group, this is given by:

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

Just like in the 2x2 case, we're subtracting a set of four means.

The estimator using comparisons for the Not-Yet Treated group is analogous.

Aggregating ATT(g, t)'s

This procedure produces many $ATT(g, t)$'s

- Perhaps a feature, rather than a bug – we can assess the *distribution* of effects

Oftentimes we'll want a summary, however. Three options:

- **Simple:** a weighted average proportional to group size
- **Dynamic:** an average of each separate length of exposure (e) since treatment occurred
- **Group:** an average for each treatment cohort

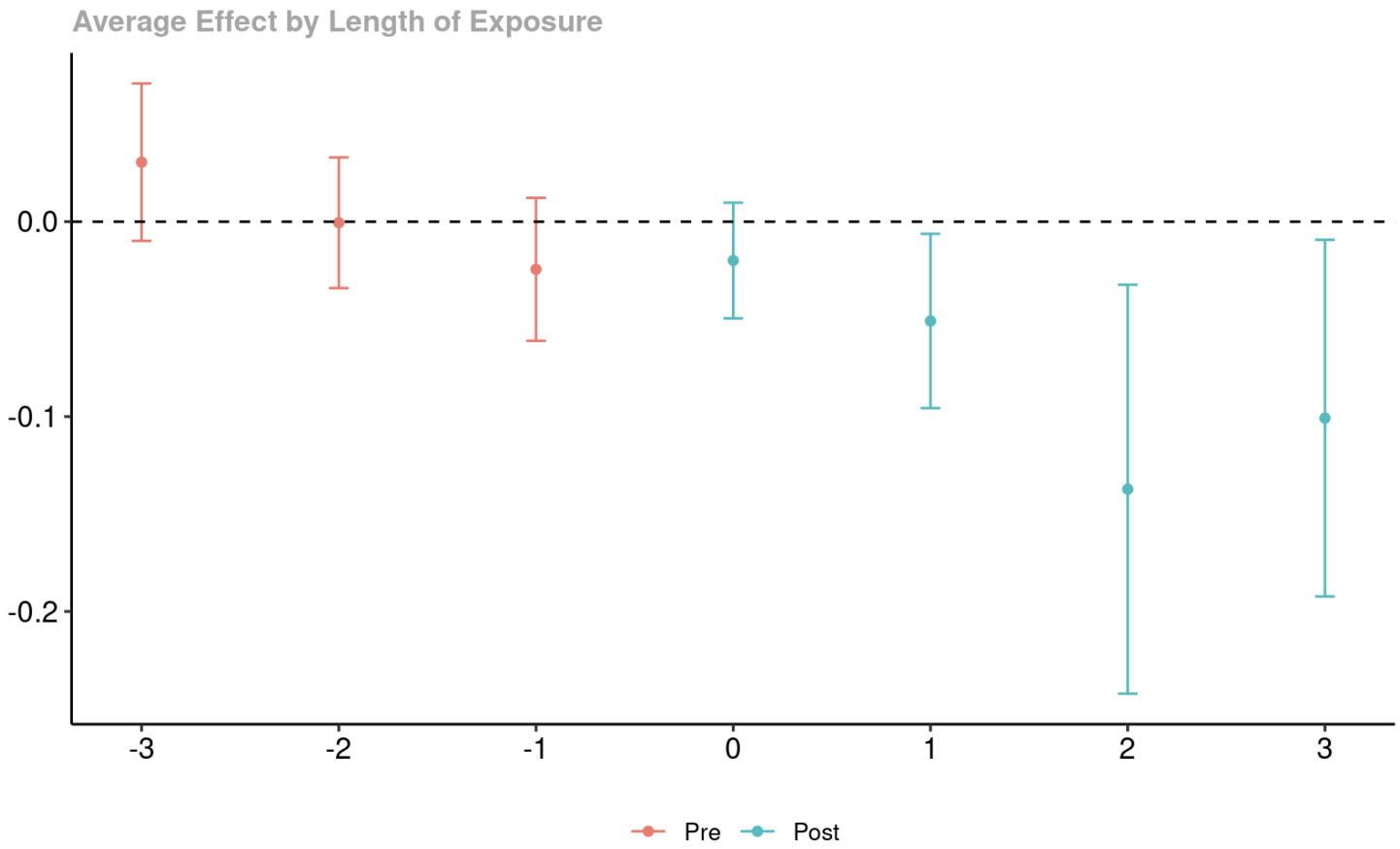
Example of Reporting ATT(g, t)'s

Table 1: Aggregated group-time ATTs

Group	ATT	SE	95% Confidence bands	
All groups	0.1075**	0.0358	0.0373	0.1778
2005	0.0952**	0.0321	0.0221	0.1682
2006	0.1074	0.0540	-0.0155	0.2304
2007	0.1332**	0.0559	0.0058	0.2606
2008	0.1181	0.0571	-0.0120	0.2483
2009	-0.0028	0.0390	-0.0916	0.0860

Both Simple and Group ATT Summaries

Using ATT(g, t)'s to Construct Event Studies



We Can Aggregate ATT Summaries by Group Time to Create an Event Study Plot

The CS Approach to Statistical Inference

Callaway and Sant'Anna recommend a *modified bootstrap procedure* to conduct statistical inference

Rather than re-sample observations or groups (could make it impossible to estimate a particular ATT(g, t)), randomly sample small perturbations to the *influence function* (how much each observation contributes to the estimate)

Some additional features:

- Can account for clustering of SEs
- Can create *confidence bands* for sets of parameters (rather than pointwise confidence intervals) to adjust for multiple hypothesis testing

Implementing Advanced DiD in R

Good news! Estimating advanced DiD models with covariates and/or with staggered treatment timing can be accomplished using a single R package: the `did` package

- Developed by Callaway and Sant'Anna

The `did` package incorporates regression-adjustment; weighting-adjustment; doubly-robust adjustment; $ATT(g,t)$ estimation; simple, dynamic, and group aggregation; event studies; inference; and more!

Let's see how to do this in `R`...