BUSFIN 8240
Empirical Corporate Finance

Lecture 4 – Natural Experiments [DiD]

Professor Michael Wittry

Background readings

- Roberts and Whited
  - *Sections 2.2, 4*
- Angrist and Pischke
  - *Section 5.2*
- *Baker, Larker, and Wang (2022)*
  - *Forthcoming JFE*

Natural Experiments – *Outline*

- Motivation and definition
- Understanding treatment effects
- Two types of simple differences
- Difference-in-differences

Recall… CMI assumption is key

- A violation of conditional mean independence (CMI), such that $E(u|x) \neq E(u)$ precludes our ability to make causal inferences

$$y = \beta_0 + \beta_1 x + u$$

  - $Cov(x,u) \neq 0$ implies CMI is violated

CMI violation implies non-randomness

- Another way to think about violation is that it indicates that our x is non-random
  - **I.e. the distribution of *x* (or the distribution of *x* after controlling for other observable covariates) isn't random**
    - E.g. firms with high x might have higher y (beyond just the effect of x on y) because high x is more likely for firms with some omitted variable contained in u…

Randomized experiments are great…

- In many of the "hard" sciences, the researcher can simply design experiment to achieve the necessary randomness
  - Ex. #1 – To determine effect of new drug, you randomly give it to certain patients
  - Ex. #2 – To determine effect of certain gene, you modify it in a random sample of mice

But, we simply can't do them ☹

- We can't do this in corporate finance!
  - E.g. we can't randomly assign a firm's leverage to determine it's effect on investment
  - And, we can't randomly assign CEOs' # of options to determine their effect on risk-taking
- Therefore, we need to rely on what we call "Natural experiments"

Defining a Natural Experiment

- Natural experiment is basically when some event causes a random assignment of (or change in) a variable of interest, x
  - Ex. #1 – Some weather event increases leverage for a random subset of firms
  - Ex. #2 – Some change in regulation reduces usage of options at a random subset of firms

Nat. Experiments Provide Randomness

- We can use such "natural" experiments to ensure that randomness (i.e. CMI) holds and make causal inferences!
  - E.g., we use the randomness introduced into x by the natural experiment to uncover the causal effect of x on y

NEs can be used in many ways

- Technically, natural experiments can be used in many different ways
  - Use them to construct IV
    - E.g., gender of first child being a boy used in Bennedsen, et al. (2007) is an example NE
  - Use them to construct regression discontinuity
    - E.g., cutoff for securitizing loans at credit score of 620 used in Keys, et al. (2010) is a NE

And, the Difference-in-Differences…

- But admittedly, when most people refer to natural experiment, they are talking about a difference-in-differences (D-i-D) estimator
  - Basically, compares outcome *y* for a "treated" group to outcome *y* for "untreated" group where treatment is randomly assigned by the natural experiment
  - **This is how I'll use NE in this class**

Natural Experiments – *Outline*

- Motivation and definition
- Understanding treatment effects
  - Notation and definitions
  - Selection bias and why randomization matters
  - Regression for treatment effects
- Two types of simple differences
- Difference-in-differences

Treatment Effects

- Before getting into natural experiments in context of difference-in-difference, it is first helpful to describe "treatment effects"

Notation and Framework

- Let *d* equal a treatment indicator from the experiment we will study
  - $d = 0$ → untreated by experiment *(i.e. control group)*
  - $d = 1$ → treated by experiment *(i.e. treated group)*
- Let *y* be the potential outcome of interest
  - $y = y(0)$ for untreated group
  - $y = y(1)$ for treated group

Example treatments in corp. fin…

- Ex. #1 – Treatment might be that your firm's state passed anti-takeover law
  - $d = 1$ for firms incorporated in those states
  - y could be a number of things, e.g. ROA
- Ex. #2 – Treatment is that your firm discovers workers exposed to carcinogen
  - $d = 1$ if have exposed workers
  - y could be a number of things, like M&A

Average Treatment Effect (ATE)

- Can now define some useful things
  - **Average Treatment Effect** (ATE) is given by

$$\mathbf{E}[y(1) - y(0)]$$

    - What does this mean in words?
    - **Answer:** The expected change in *y* from being treated by the experiment; **<u>this is the causal effect we are typically interested in uncovering!</u>**

But, ATE is unobservable

$$\mathbf{E}[y(1) - y(0)]$$

- Why can't we actually directly observe ATE?
  - **Answer** = We only observe one outcome…
    - If treated, we observe *y(1)*; if untreated, we observe *y(0)*. We never observe both.
    - I.e., we cannot observe the counterfactual of what your *y* would have been absent treatment

Defining ATT

- **Average Treatment Effect if Treated (ATT)** is given by $\mathbf{E[}y(1) - y(0) \mid d = 1\mathbf{]}$
  - This is the effect of treatment on those that are treated; i.e change in y we'd expect to find if treated random sample from population of observations that are treated
  - **What don't we observe here?**
  - **Answer** $= y(0) \mid d = 1$

Defining ATU

- **Average Treatment Effect if Untreated (ATU)** is given by $\mathbf{E[}y\mathbf{(1) - }y\mathbf{(0) \mid }d\mathbf{ = 0]}$
  - This is what the effect of treatment would have been on those that are not treated by the experiment
  - We don't observe $y(1) \mid d = 0$

Uncovering ATE *[Part 1]*

- So how do we estimate ATE, $E[y(1) - y(0)]$?
  - **Answer =** We instead rely on $E[y(1) \mid d = 1]$ – $E[(y(0) \mid d = 0]$ as our way to *infer* the ATE

In words, what are we doing & what are we assuming?

Uncovering ATE *[Part 2]*

- In words, we compare average *y* of treated to average *y* of untreated observations
  - If we interpret this as the ATE, we are assuming that absent the treatment, the treated group would, on average, have had same outcome *y* as the untreated group
  - We can show this formally by simply working out $E[y(1) \mid d = 1] - E[y(0) \mid d = 0]$…

Uncovering ATE *[Part 3]*

$$\{E[y(1)\,|\,d=1]-E[y(0)\,|\,d=1]\}+\{E[y(0)\,|\,d=1]-E[y(0)\,|\,d=0]\}$$

First bracket is <u>ATT</u>

Just added and subtracted the same term

Second bracket is what we call the "selection bias"

- Simple comparison doesn't give us the ATE! In fact, the comparison is rather meaningless!
- What is the **"selection bias"** in words?

Natural Experiments – *Outline*

- Motivation and definition
- Understanding treatment effects
  - Notation and definitions
  - Selection bias and why randomization matters
  - Regression for treatment effects
- Two types of simple differences
- Difference-in-differences

Selection bias defined

- Selection bias: $E[y(0) \mid d = 1] - E[y(0) \mid d = 0]$
  - **Definition =** What the difference in average *y* would have been for treated and untreated observations absent any treatment
  - *We do not observe this counterfactual!*

- Now let's see why randomness is key!

Introducing random treatment

- A random treatment, *d*, implies that *d* is independent of potential outcomes; i.e.

$$E[y(0) \mid d = 1] = E[y(0) \mid d = 0] = E[y(0)]$$

and

$$E[y(1) \mid d = 1] = E[y(1) \mid d = 0] = E[y(1)]$$

In words, the expected value of *y* is the same for treated and untreated absent treatment

  - With this, easy to see that selection bias = 0
  - **And**, remaining ATT is equal to ATE!

Random treatment makes life easy

- I.e. with random assignment of treatment, our simple comparison gives us the ATE!
  - This is why we like randomness!
  - But, absent randomness, we must worry that any observed difference is driven by selection bias

Natural Experiments – *Outline*

- Motivation and definition
- Understanding treatment effects
  - Notation and definitions
  - Selection bias and why randomization matters
  - Regression for treatment effects
- Two types of simple differences
- Difference-in-differences

ATE in Regression Format *[Part 1]*

- Can re-express everything in regression format

$$y = \beta_0 + \beta_1 d + u$$

$$\beta_0 = E[y(0)]$$

$$\text{where} \quad \beta_1 = y(1) - y(0)$$

$$u = y(0) - E[y(0)]$$

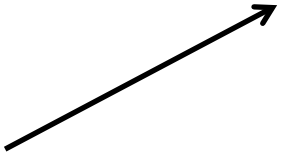
This regression will only give consistent estimate of β_1 if cov(d, u) = 0; i.e. treatment, d, is random, and hence, uncorrelated with $y(0)$!

  - If you plug-in, it will get you back to what the true model, $y = y(0) + d[y(1) - y(0)]$

ATE in Regression Format *[Part 2]*

- We are interested in $\mathrm{E}[y \mid d = 1] - \mathrm{E}[y \mid d = 0]$
  - But, can easily show that this expression is equal to

$$\beta_1 + E[y(0) \mid d = 1] - E[y(0) \mid d = 0]$$



Our estimate will equal true effect plus selection bias term

Note: Selection bias term occurs only if CMI isn't true!

Treatment effect – *Example*

- Suppose we compare the leverage of firms with and without a credit rating [or equivalently, regress leverage on indicator for rating]
  - Treatment is having a credit rating
  - Outcome of interest is leverage

Why might our estimate not equal ATE of rating?

Why might controls not help us much?

Treatment effect – *Example Answer*

- **Answer #1:** Having a rating isn't random
  - Firms with rating likely would have had higher leverage anyway because they are larger, more profitable, etc.; selection bias will be positive
  - Selection bias is basically an omitted var.!
- **Answer #2:** Even adding controls might not help if firms also differ in unobservable ways, like investment opportunities

Natural Experiments – *Outline*

- Motivation and definition
- Understanding treatment effects
- Two types of simple differences
  - Cross-sectional difference & assumptions
  - Time-series difference & assumptions
  - Miscellaneous issues & advice
- Difference-in-differences

We actually just did this one!

Cross-sectional Simple Difference

- Very intuitive idea
  - Compare post-treatment outcome, *y*, for treated group to the untreated group
  - I.e. just run following regression…

In regression format…

- Cross-section simple difference

$$y_{i,t} = \beta_0 + \beta_1 d_i + u_{i,t}$$

  - $d = 1$ if observation i is in treatment group and equals zero otherwise
  - Regression only contains post-treatment time periods

What is needed for β_1 to capture the true (i.e. causal) treatment effect?

Identification Assumption

- **Answer:** $E(u \mid d) = 0$; i.e. treatment, *d*, is uncorrelated with the error
  - In words… after accounting for effect of treatment, the expected level of *y* in post-treatment period isn't related to whether you're in the treated or untreated group
  - ***I.e.***, expected *y* of treated group would have been same as untreated group *absent* treatment

Natural Experiments – *Outline*

- Motivation and definition
- Understanding treatment effects
- Two types of simple differences
  - Cross-sectional difference & assumptions
  - Time-series difference & assumptions
  - Miscellaneous issues & advice
- Difference-in-differences

Time-series Simple Difference

- Again, very intuitive idea
  - Compare pre- and post-treatment outcomes, *y*, for just the treated group *[i.e. pre-treatment period acts as 'control' group]*
  - I.e. run following regression…

In Regression Format

- Time-series simple difference

$$y_{i,t} = \beta_0 + \beta_1 p_t + u_{i,t}$$

  - $p_t = 1$ if period t occurs after treatment and equals zero otherwise
  - Regression contains only observations that are treated by "experiment"

What is needed for β_1 to capture the true (i.e. causal) treatment effect?

Identification Assumption

- **Answer:** $E(u|p) = 0$; i.e. post-treatment indicator, p, is uncorrelated with the error
  - I.e., after accounting for effect of treatment, p, the expected level of y in post-treatment period wouldn't have been any different than expected y in pre-treatment period

Using a First-Difference (FD) Approach

- Could also run regression using first-differences specification

$$y_{i,t} - y_{i,t-1} = \beta_1 \left(p_t - p_{t-1} \right) + \left(u_{i,t} - u_{i,t-1} \right)$$

  - If just one pre- and one post-treatment period (i.e. *t-1* and *t*), then will get identical results
  - But, if more than one pre- and post-treatment period, the results will differ…

FD *versus* Standard Approach *[Part 1]*

- Why might these two models give different estimates of β_1 when there are more than one pre- and post-treatment periods?

$$y_{i,t} = \beta_0 + \beta_1 p_t + u_{i,t}$$

versus

$$y_{i,t} - y_{i,t-1} = \beta_1 \left(p_t - p_{t-1} \right) + \left(u_{i,t} - u_{i,t-1} \right)$$

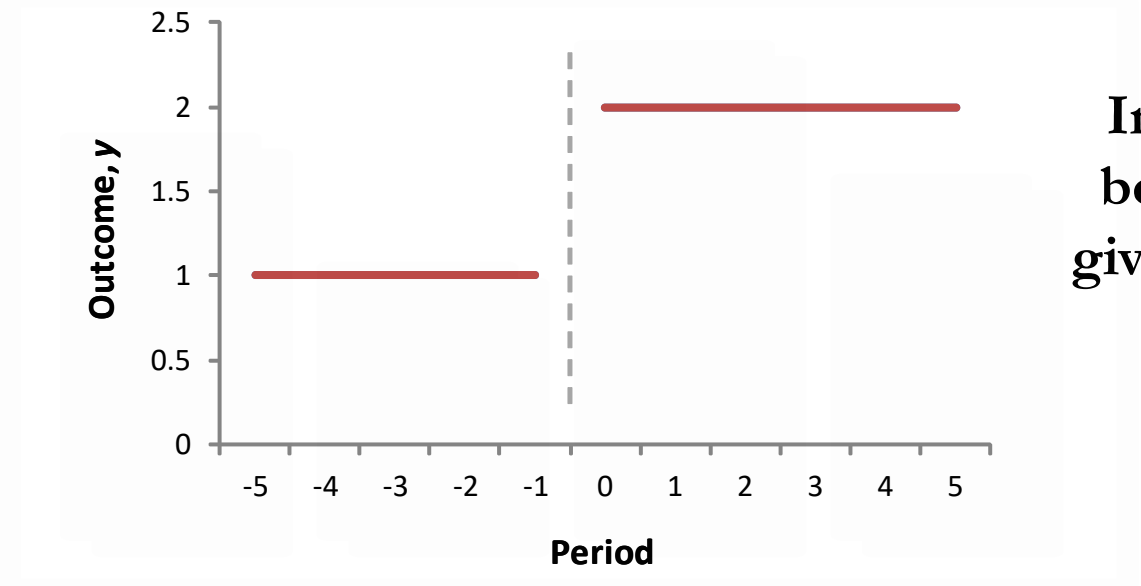
FD *versus* Standard Approach *[Part 2]*

How might this matter in practice?

- **Answer:**
  - In 1st regression, β_1 captures difference between avg. y pre-treatment *versus* avg. y post-treatment
  - In 2nd regression, β_1 captures difference in Δy immediately after treatment versus Δy in all other pre- and post-treatment periods
    - I.e. the Δp variable equals 1 only in immediate post-treatment period, and 0 for all other periods

FD *versus* Standard Approach *[Part 3]*

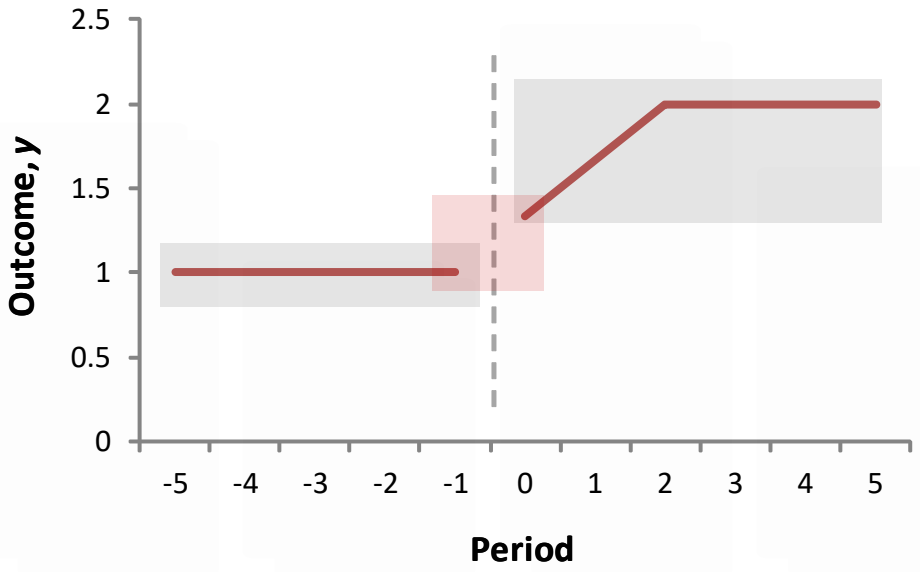
- Both approaches assume the effect of treatment is immediate and persistent, e.g.



In this scenario, both approaches give same estimate

FD *versus* Standard Approach *[Part 4]*

- But, suppose the following is true...



In this scenario, FD approach gives much smaller estimate

1st approach compares avg. pre- versus post

FD compares Δy from $t=0$ to $t=-1$ against Δy elsewhere (which isn't always zero!)

Correct way to do difference

- Correct way to get a 'differencing' approach to match up with the more standard simple diff specification in multi-period setting is to instead use

$$\bar{y}_{i,post} - \bar{y}_{i,pre} = \beta_1 + \left(\bar{u}_{i,post} - \bar{u}_{i,pre}\right)$$

  - This is exactly the same as simple difference

Natural Experiments – *Outline*

- Motivation and definition
- Understanding treatment effects
- Two types of simple differences
  - Cross-sectional difference & assumptions
  - Time-series difference & assumptions
  - Miscellaneous issues & advice
- Difference-in-differences

Treatment effect isn't always immediate

- In prior example, the specification is wrong because the treatment effect only slowly shows up over time
  - Why might such a scenario be plausible?
  - **Answer =** Many reasons. E.g. firms might only slowly respond to change in regulation, or CEO might only slowly change policy in response to compensation shock

Accounting for a delay…

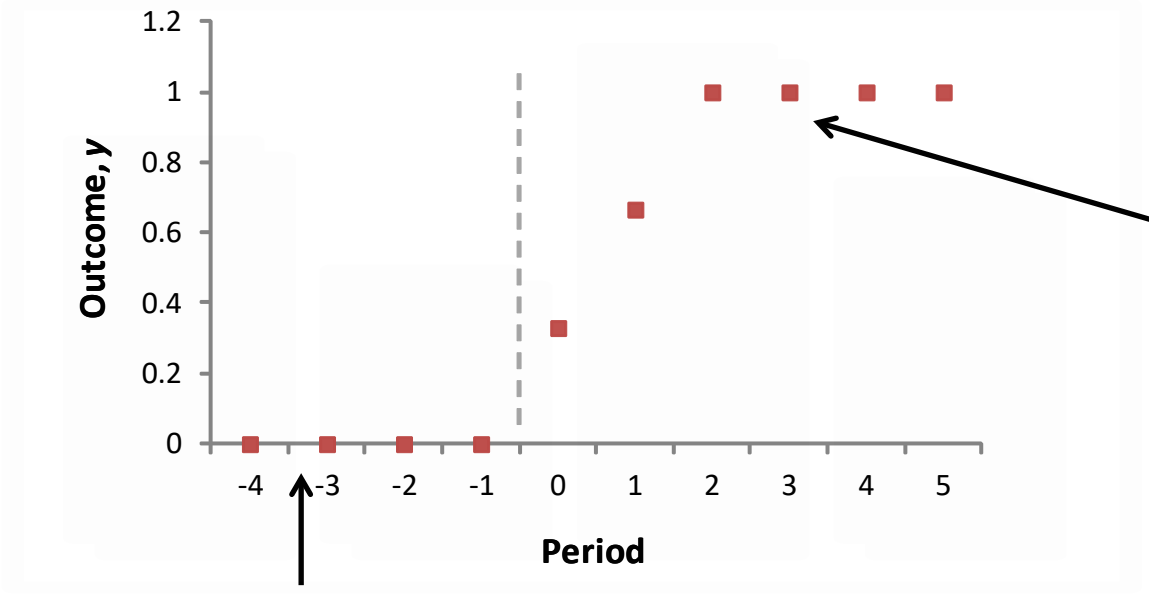
- Simple-difference misses this subtlety; it assumes effect was immediate
- For this reason, it is always helpful to run regression that allows effect to vary by period
  - **How can you do this?**
  - **Answer =** Insert indicators for each year relative to the treatment year *[see next slide]*

Non-parametric approach

- If have 5 pre- and 5 post-treatment obs.; could estimate : $y_{i,t} = \beta_0 + \sum_{t=-4}^{5} \beta_t p_t + u_{i,t}$
  - p_t is now an indicator that equals 1 if year = t and zero otherwise; e.g.
    - $t = 0$ is the period treatment occurs
    - $t = -1$ is period before treatment
  - β_t estimates change in y relative to excluded periods; you then plot these in graph

Non-parametric approach – *Graph*

- Plot estimates to trace out effect of treatment



Approach allows effect of treatment to vary by year!

Estimates capture change relative to excluded period (*t*-5)

Could easily plot confidence intervals as well

These equal zero because *y* was same as *y* in excluded period (*t*-5)

Simple Differences – *Advice*

- In general, simple differences are not that convincing in practice…
  - Cross-sectional difference requires us to assume the average *y* of treated and untreated would have been same absent treatment
  - Time-series difference requires us to assume the average *y* would have been same in post- and pre-treatment periods absent treatment
- **Is there a better way?**

Natural Experiments – *Outline*

- Motivation and definition
- Understanding treatment effects
- Two types of simple differences
- Difference-in-differences
  - Intuition & implementation
  - "Parallel trends" assumption
  - Staggered treatment and heterogeneous treatment effects

Difference-in-differences

- Yes, we can do better!
- We can do a difference-in-differences that combines the two simple differences
  - **Intuition** = compare change in *y* pre- versus post-treatment for treated group *[1st difference]* to change in *y* pre- versus post-treatment for untreated group *[2nd difference]*

Implementing diff-in-diff

- Difference-in-differences estimator

$$y_{i,t} = \beta_0 + \beta_1 p_t + \beta_2 d_i + \beta_3 \left(d_i \times p_t \right) + u_{i,t}$$

  - $p_t = 1$ if period t occurs after treatment and equals zero otherwise
  - $d_i = 1$ if unit is in treated group and equals zero otherwise

What do β_1, β_2, and β_3 capture?

Interpreting the estimates *[Part 1]*

- **Here is how to interpret everything…**
  - β_1 captures the average change in *y* from the pre- to post-treatment periods that is common to both treated and untreated groups
  - β_2 captures the average difference in level of *y* between treated and untreated groups that is common to both pre- and post-treatment periods

Interpreting the estimates *[Part 2]*

- β_3 captures the average <u>differential</u> change in *y* from the pre- to post-treatment period for the treatment group ***relative*** to the change in *y* for the untreated group

- β_3 is what we call the diff-in-diff estimate

Comparing group means approach *[P1]*

- Again, the regression is…

$$y_{i,t} = \beta_0 + \beta_1 p_t + \beta_2 d_i + \beta_3 \left(d_i \times p_t \right) + u_{i,t}$$

  - And, the four possible combinations are:

$$E(y \mid d = 1, p = 1) = \beta_0 + \beta_1 + \beta_2 + \beta_3$$
$$E(y \mid d = 1, p = 0) = \beta_0 + \beta_2$$
$$E(y \mid d = 0, p = 1) = \beta_0 + \beta_1$$
$$E(y \mid d = 0, p = 0) = \beta_0$$

What assumption did I make in doing this?

Answer: $E(u \mid d,p)=0$; i.e. the "experiment" is random

Comparing group means approach *[P2]*

$$E(y \mid d = 1, p = 1) = \beta_0 + \beta_1 + \beta_2 + \beta_3$$
$$E(y \mid d = 1, p = 0) = \beta_0 + \beta_2$$
$$E(y \mid d = 0, p = 1) = \beta_0 + \beta_1$$
$$E(y \mid d = 0, p = 0) = \beta_0$$

- These can be arranged in two-by-two table

Comparing group means approach *[P3]*

- First, take the simple differences

	Post-Treatment, (1)	Pre-Treatment, (2)	Difference, (1)-(2)
Treatment, (a)	$\beta_0+\beta_1+\beta_2+\beta_3$	$\beta_0+\beta_2$	$\beta_1+\beta_3$
Control, (b)	$\beta_0+\beta_1$	β_0	β_1
Difference, (a)-(b)	$\beta_2+\beta_3$	β_2	

Comparing group means approach *[P4]*

- Then, take difference-in-differences!

	Post-Treatment, (1)	Pre-Treatment, (2)	Difference, (1)-(2)
Treatment, (a)	$\beta_0+\beta_1+\beta_2+\beta_3$	$\beta_0+\beta_2$	$\beta_1+\beta_3$
Control, (b)	$\beta_0+\beta_1$	β_0	β_1
Difference, (a)-(b)	$\beta_2+\beta_3$	β_2	**β_3**

This is why they call it the difference-in-differences estimate; regression gives you same estimate as if you took differences in the group averages. Again, β_3 has a *causal* interpretation when $E(u|d,p)=0$.

Simple difference – Revisited *[Part 1]*

- Useful to look at simple differences

	Post-Treatment, (1)	Pre-Treatment, (2)	Difference, (1)-(2)
Treatment, (a)	$\beta_0+\beta_1+\beta_2+\beta_3$	$\beta_0+\beta_2$	$\beta_1+\beta_3$
Control, (b)	$\beta_0+\beta_1$	β_0	β_1
Difference, (a)-(b)	$\beta_2+\beta_3$	β_2	**β_3**

This was cross-sectional simple difference

When does that simple diff give effect of treatment, β_3?

Answer = when β_2 equals zero; i.e. no difference in level of *y* absent treatment

Simple difference – Revisited *[Part 2]*

- Now, look at time-series simple diff…

	Post-Treatment, (1)	Pre-Treatment, (2)	Difference, (1)-(2)
Treatment, (a)	$\beta_0+\beta_1+\beta_2+\beta_3$	$\beta_0+\beta_2$	$\beta_1+\beta_3$
Control, (b)	$\beta_0+\beta_1$	β_0	β_1
Difference, (a)-(b)	$\beta_2+\beta_3$	β_2	**β_3**

This was time-series simple difference

When does that simple diff give effect of treatment, β_3?

Answer = when β_1 equals zero; i.e. no change in *y* absent treatment

Natural Experiments – *Outline*

- Motivation and definition
- Understanding treatment effects
- Two types of simple differences
- Difference-in-differences
  - Intuition & implementation
  - "Parallel trends" assumption
  - Staggered treatment and heterogeneous treatment effects

"Parallel trends" assumption

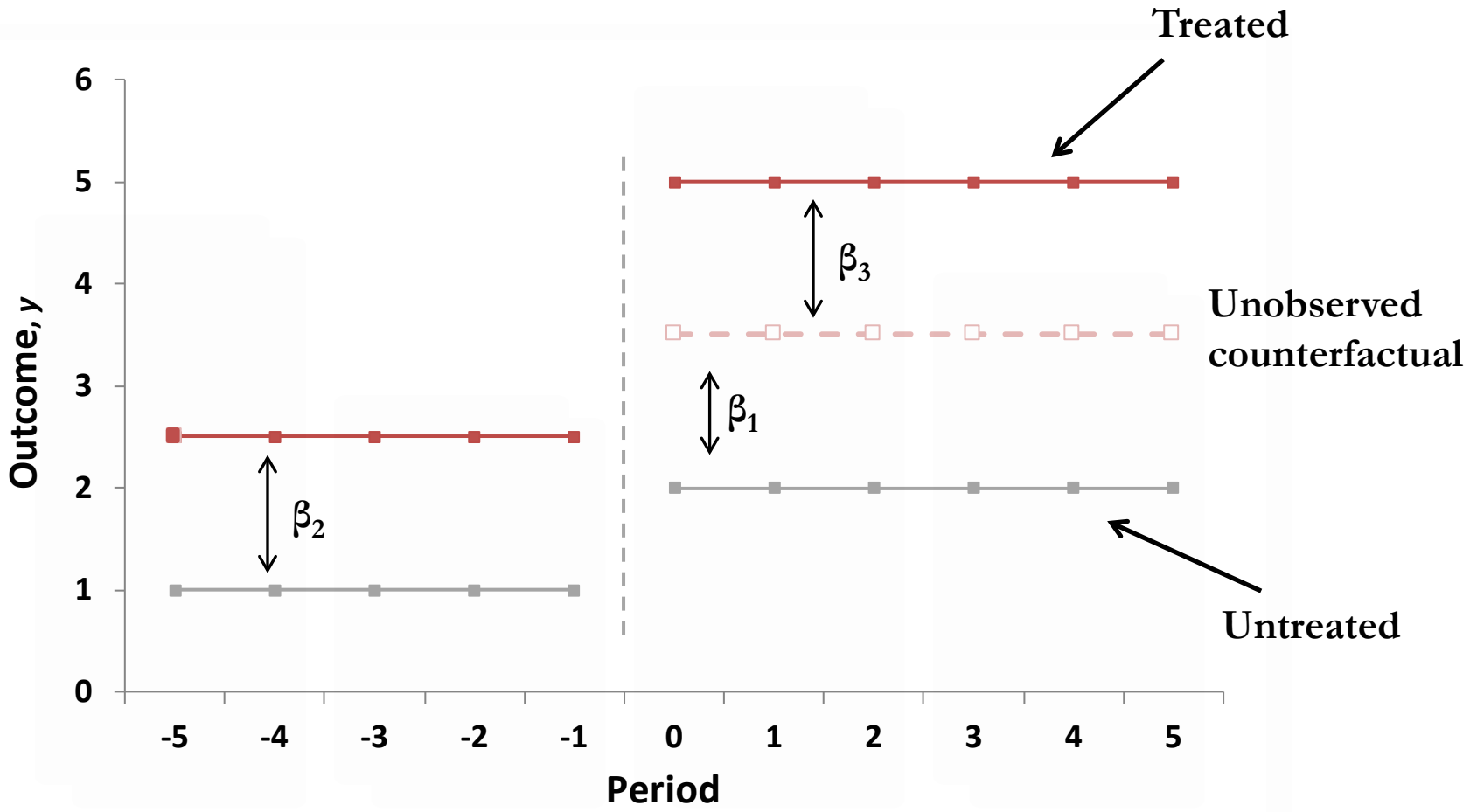
- Identification assumption is what we call the **parallel trends assumption**
  - Absent treatment, the change in *y* for treated would not have been different than the change in *y* for the untreated observations

Difference-in-differences – *Visually*

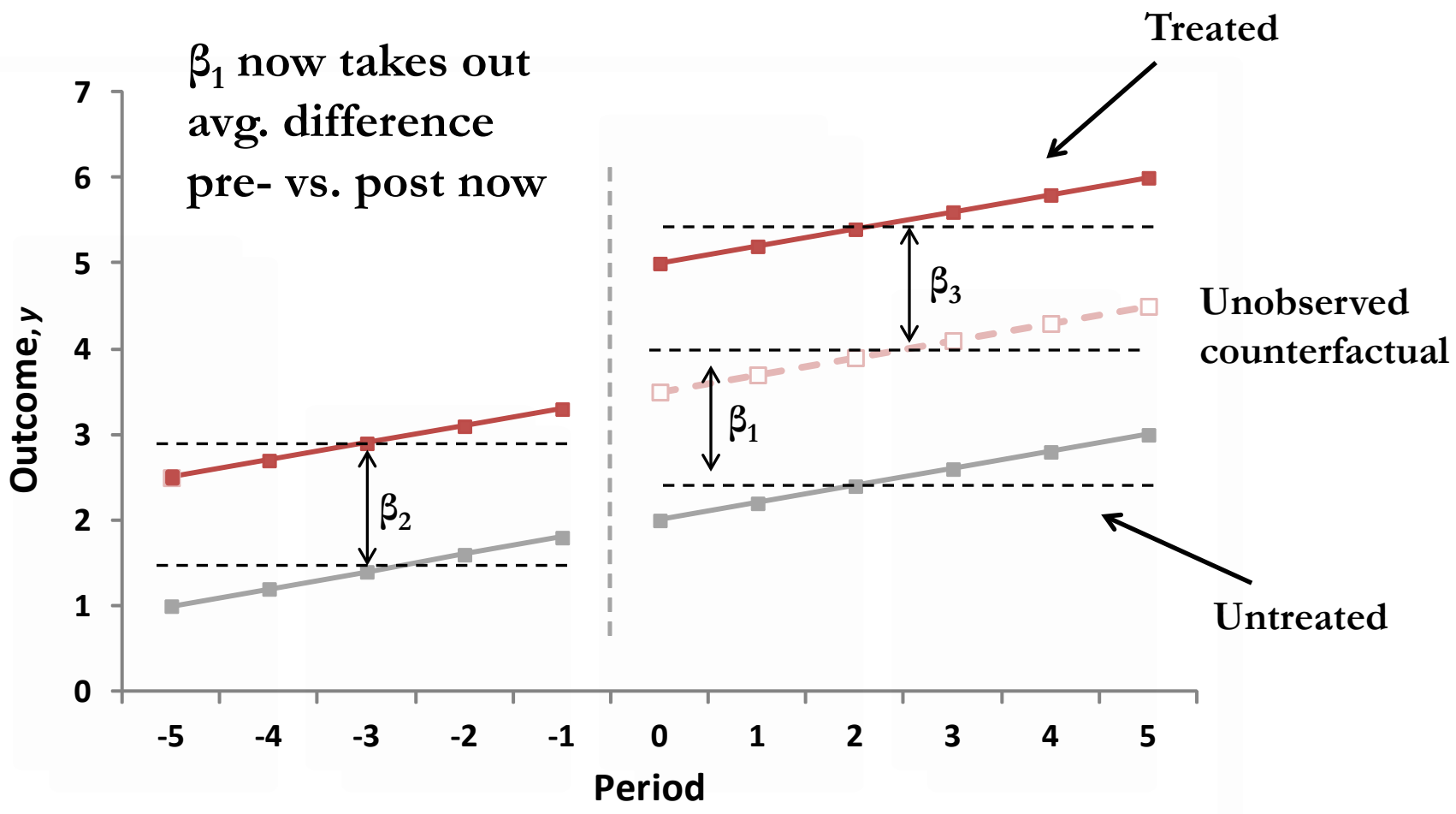
- Looking at what difference-in-differences estimate is doing in graphs will also help you see why the parallel trends assumption is key

$$y_{i,t} = \beta_0 + \beta_1 p_t + \beta_2 d_i + \beta_3 \left(d_i \times p_t \right) + u_{i,t}$$

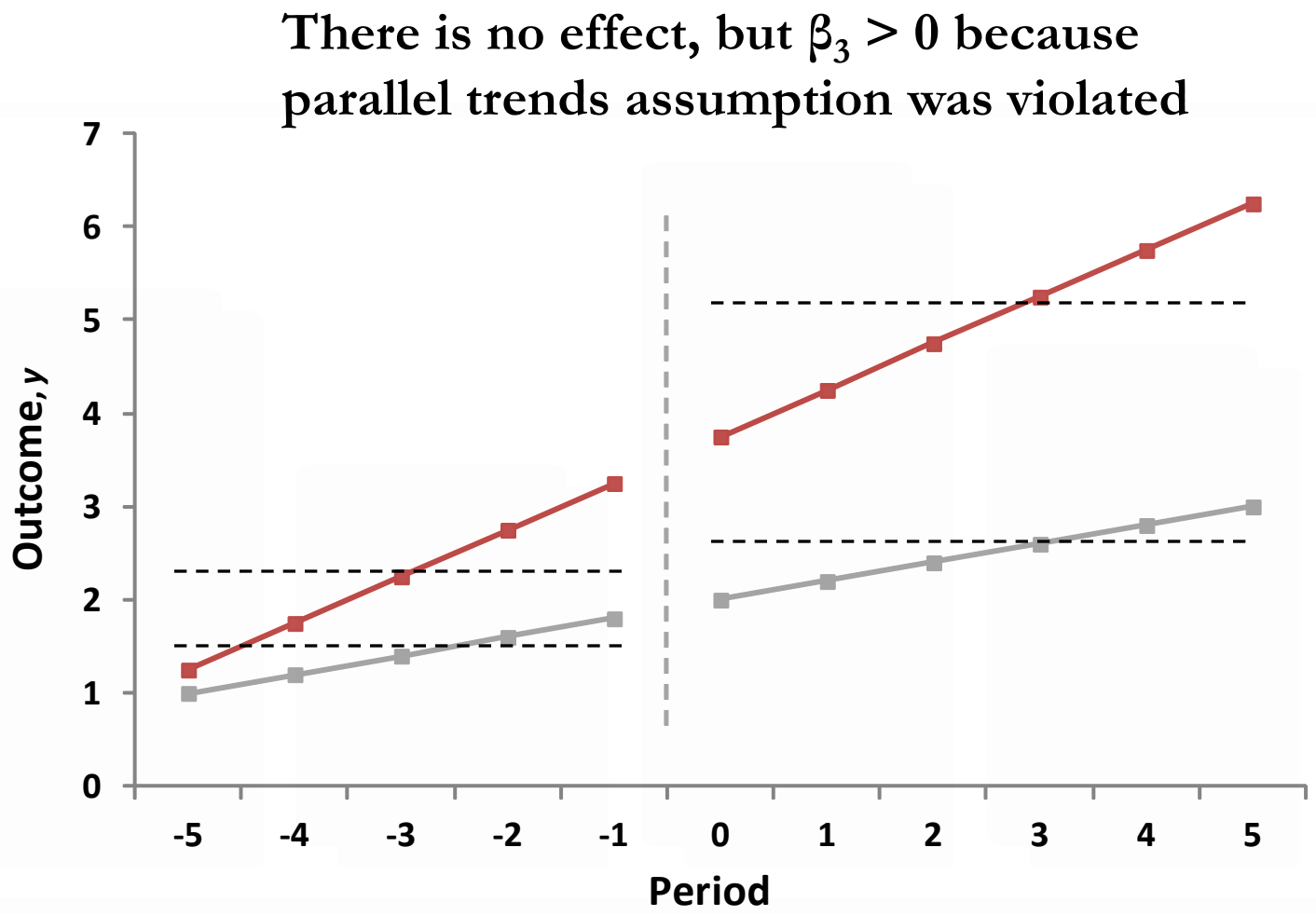
Diff-in-diffs – *Visual Example #1*



Diff-in-diff – *Visual Example #2*



Violation of parallel trends – *Visual*

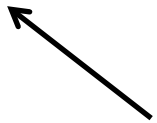


Why we like diff-in-diff *[Part 1]*

- With simple difference, any of the below arguments would prevent causal inference
  - Cross-sectional diff – "Treatment and untreated avg. *y* could be different for reasons a, b, and c, that just happen to be correlated with whether you are treated or not"
  - Time-series diff – "Treatment group's avg. *y* could change post- treatment for reasons a, b, and c, that just happen to be correlated with the timing of treatment"

Why we like diff-in-diff *[Part 2]*

- But, now the required argument to suggest the estimate isn't causal is…
  - "The change in *y* for treated observations after treatment would have been different than change in *y* for untreated observations for reasons a, b, and c, that just happen to be correlated with **both** whether you are treated and when the treatment occurs"



This is (usually) a much harder story to tell

Example…

- Bertrand & Mullainathan (JPE 2003) use state-by-state changes in regulations that made it harder for firms to do M&A
  - They compare wages at firms pre- versus post-regulation in treated versus untreated states
  - Are the below valid concerns about their difference-in-differences…

Are these concerns for internal validity?

- The regulations were passed during a time period of rapid growth of wages nationally…
  - **Answer = No.** Indicator for post-treatment accounts for common growth in wages
- States that implement regulation are more likely have unions, and hence, higher wages…
  - **Answer =** No. Indicator for treatment accounts for this average difference in wages

Example continued…

- However, ex-ante average differences is troublesome in some regard…
  - Suggests treatment wasn't random
  - And, ex-ante differences can be problematic if we think they their effect may vary with time…
    - Time-varying omitted variables **<u>are</u>** problematic because they can cause violation of "parallel trends"
    - E.g. states with more unions were trending differently at that time because of changes in union power

Natural Experiments – *Outline*

- Motivation and definition
- Understanding treatment effects
- Two types of simple differences
- Difference-in-differences
  - Intuition & implementation
  - "Parallel trends" assumption
  - Staggered (multiple) treatments and heterogeneous treatment effects

Natural Experiment – *Outline*

- Staggered (multiple) treatments and heterogeneous treatment effects
  - How to handle multiple events
    - Why they are useful
    - Simple estimation approach
  - Problems with this design
    - Solutions to this problem

Motivating example…

- Gormley and Matsa (2011) looked at firms' responses to increased left-tail risk
  - Used discovery that workers were exposed to harmful chemical as exogenous increase in risk
  - One discovery occurred in 2000; a chemical heavily used by firms producing semiconductors was found to be harmful
- Can you think of any concerns about parallel trends assumption of this setting?

Motivating Example – *Answer*

- **Answer:** Yes… This coincides with bursting of technology bubble; technology firms might arguably trend differently after 2000 for this reasons unrelated to chemical
  - How might multiple treatment events, occurring at different times (which is what Gormley and Matsa used), help?

Multiple treatment events

- Sometimes, the natural experiment is repeated a multiple points in times for multiple groups of observations
  - E.g. U.S. states make a particular regulatory change at different points in time
- **These settings are particularly useful in mitigating concerns about violation of parallel trends assumption…**

How multiple events are helpful

- Can show that effect of treatment is similar across different time periods
- Can show effect of treatment isn't driven by a particular set of treated firms
  - I.e. now the "identification police" would need to come up with story as to why parallel trends is violated for <u>each</u> unique event

Natural Experiment – *Outline*

- Staggered (multiple) treatments and heterogeneous treatment effects
  - How to handle multiple events
    - Why they are useful
    - Simple estimation approach
  - Problems with this design
    - Solutions to this problem

Estimation with Multiple Events

- Estimating model with multiple events is still relatively easy to do
  - Use approach of Bertrand and Mullainathan (JPE 2003)

Estimation with Multiple Events *[P1]*

- Just estimate the following estimation

$$y_{ict} = \beta d_{ict} + p_t + m_c + u_{ict}$$

  - y_{ict} is outcome for unit *i* (e.g. firm) in period *t* (e.g. year) and cohort *c*, where "cohort" indexes the different sets of firms <u>treated</u> by each event
    - E.g. different firms might be affected by a change in regulation at different points in time; firms **<u>affected</u>** at one point in time are a 'cohort'

Estimation with Multiple Events *[P2]*

$$y_{ict} = \beta d_{ict} + p_t + m_c + u_{ict}$$

d_{ict} = indicator on whether cohort *c* is affected by time *t*; this is the interaction between *treatment* & *post*

Time period fixed effects; they will control for *post* dummy in each event

Cohort fixed effects; they are the control for the *treatment* dummy in each event

Estimation with Multiple Events *[P3]*

- Intuition of this approach…
  - Every untreated observation at a particular point in time acts as control for treated observations in that time period
    - E.g. a firm treated in 1999 by some event will act as a control for a firm treated in 1994 until itself becomes treated in 1999
  - β will capture <u>average</u> treatment effect across the multiple events

Natural Experiment – *Outline*

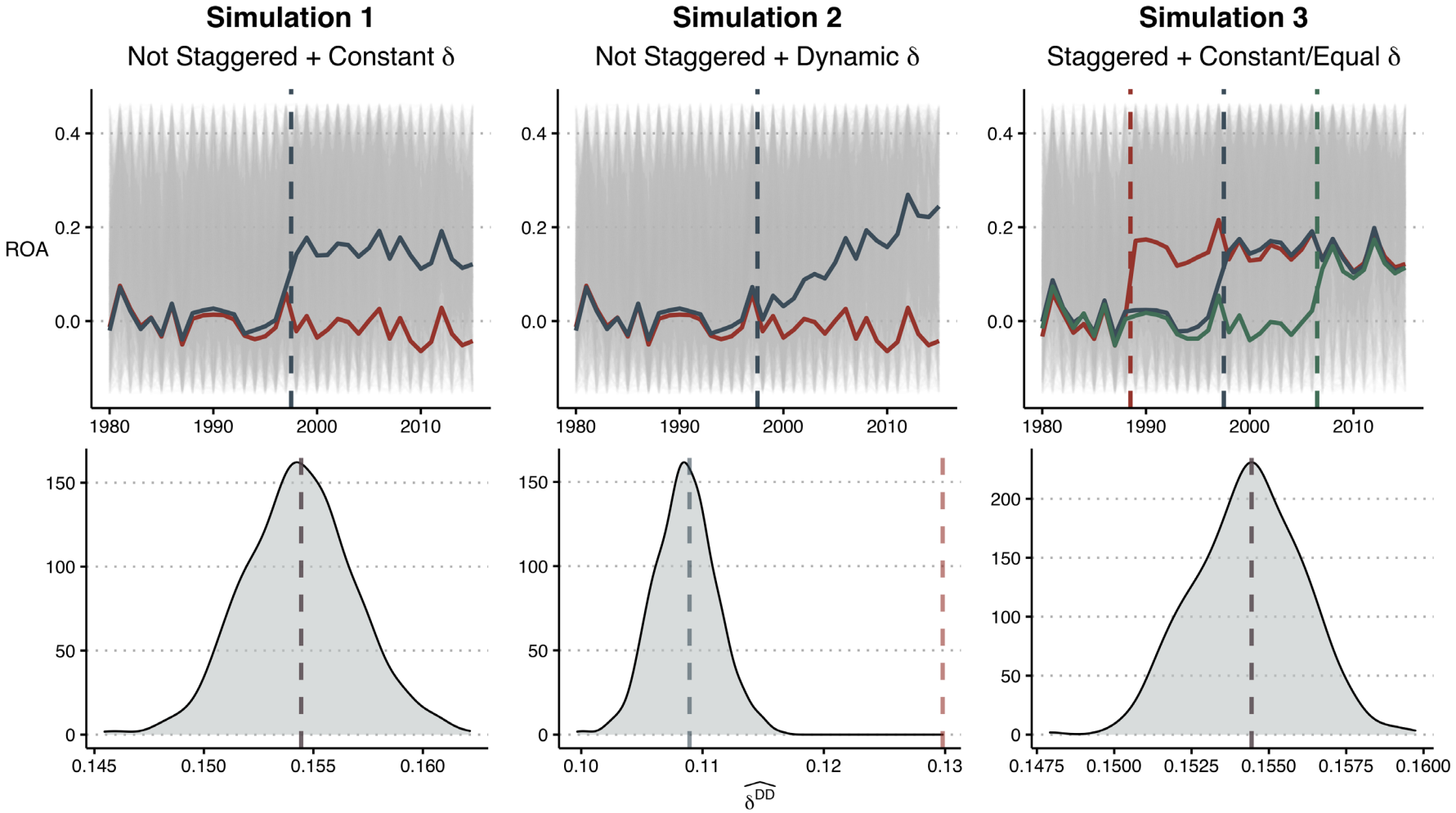
- Staggered (multiple) treatments and heterogeneous treatment effects
  - How to handle multiple events
    - Why they are useful
    - Simple estimation approach
  - Problems with this design
    - Solutions to this problem

A Potential Problem

- Early treated units are used as a control for later treated units
  - E.g., because they have already been treated, they don't change treatment status for the later units
  - The answer depends on your assumptions on treatment effect heterogeneity

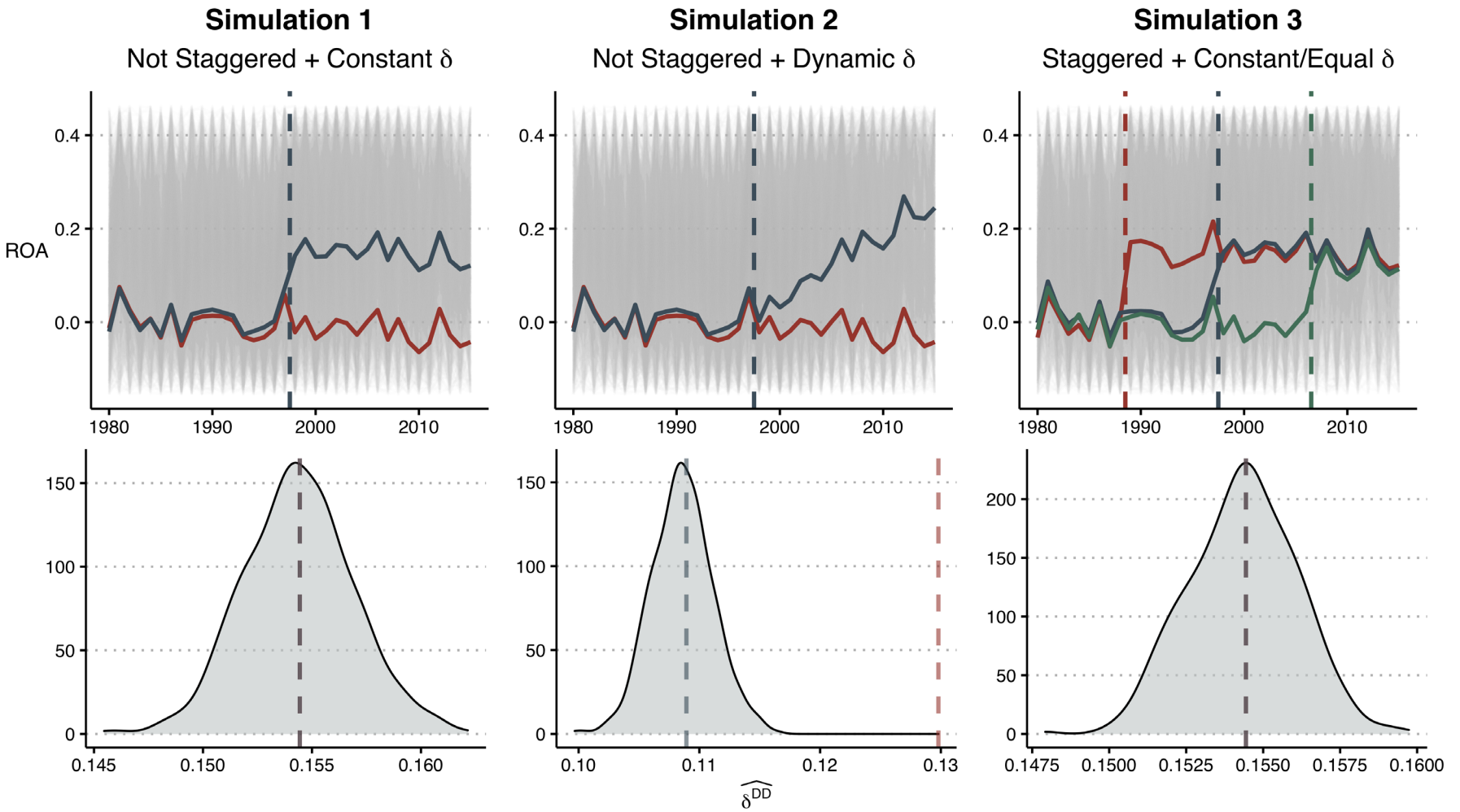
Do you see a problem with this?

Assumptions about timing and heterogeneity



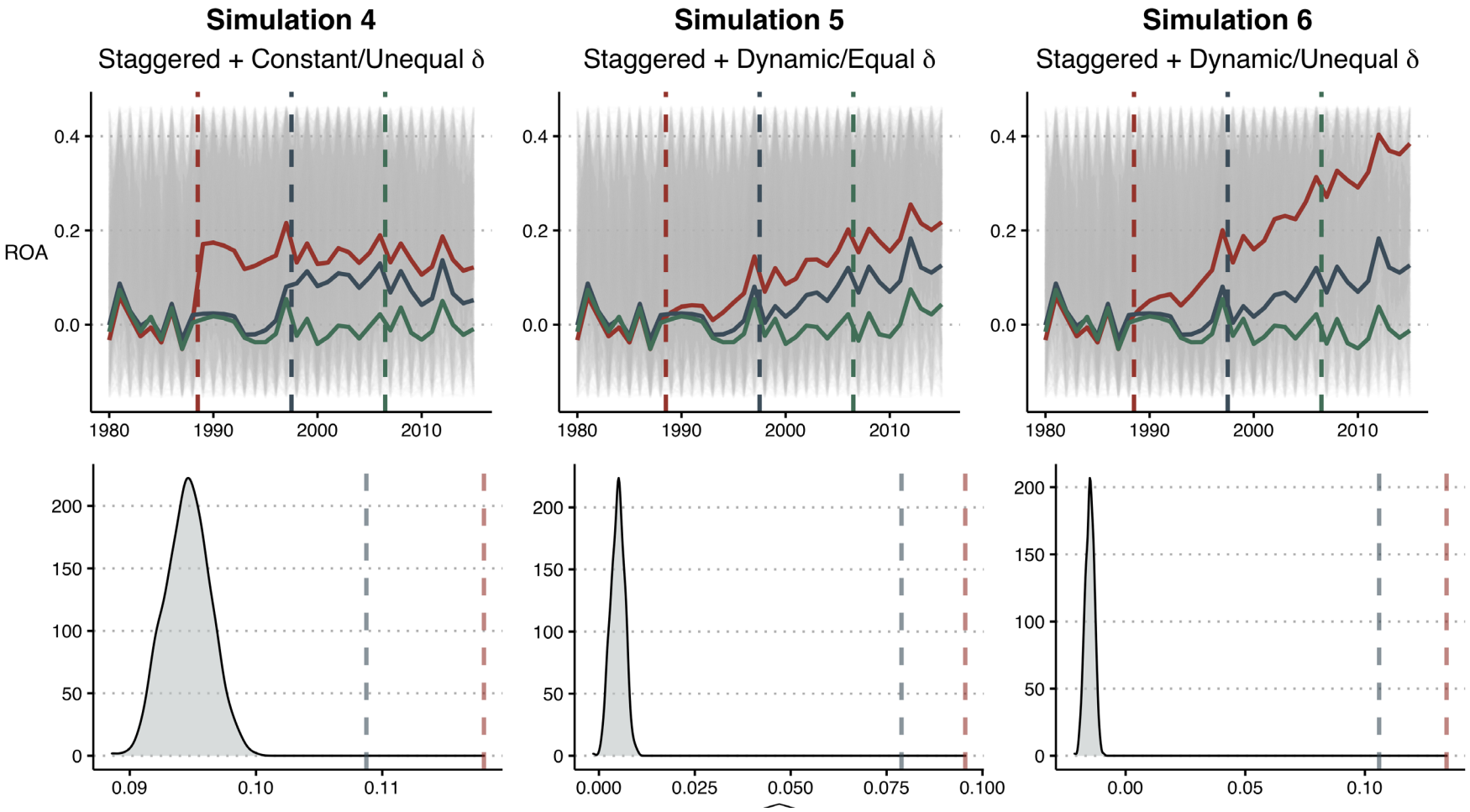
Source: Baker, Larcker, and Wang (JFE 2022)

Assumptions about timing and heterogeneity



Source: Baker, Larcker, and Wang (JFE 2022)

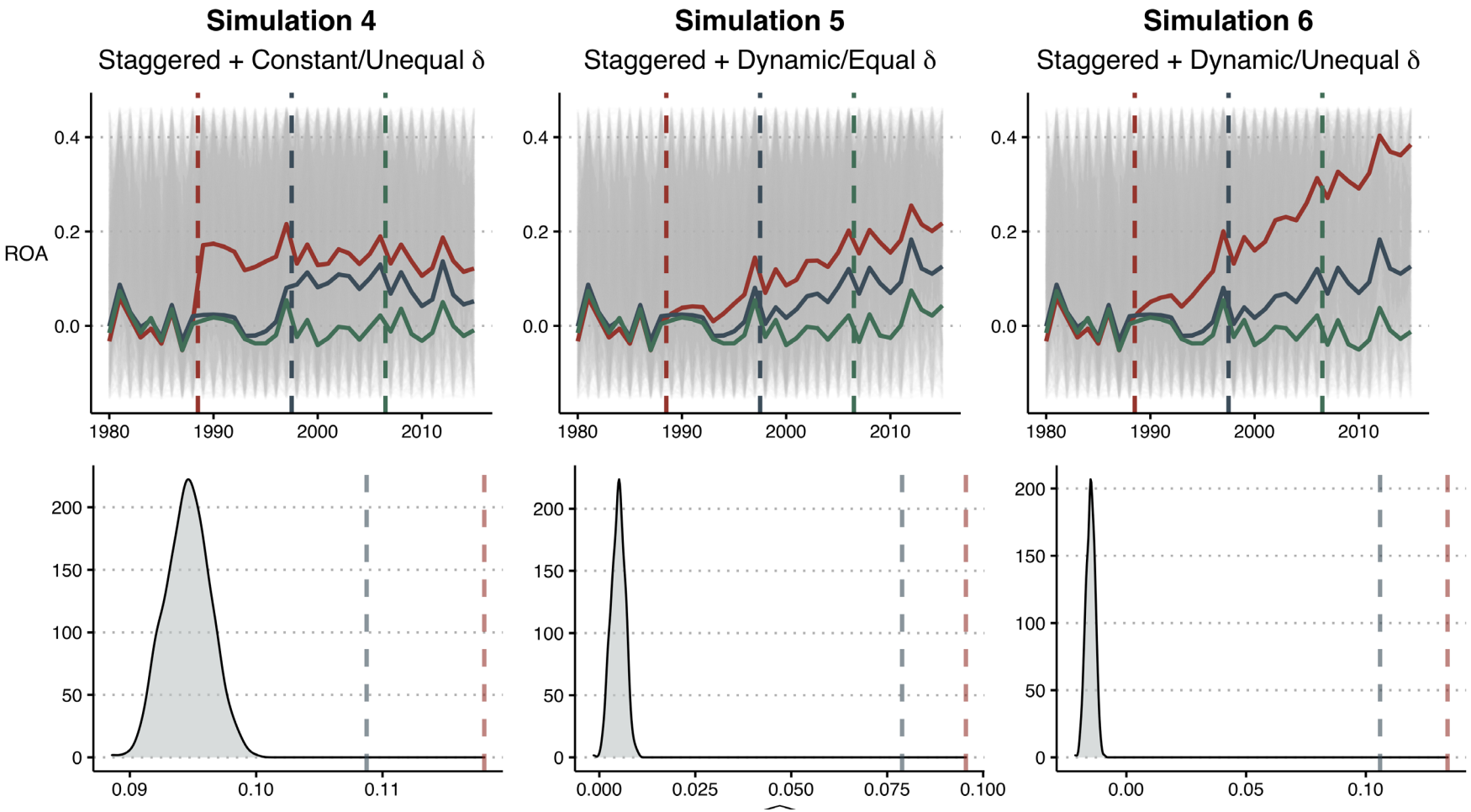
Assumptions about timing and heterogeneity



Source: Baker, Larcker, and Wang (JFE 2022)

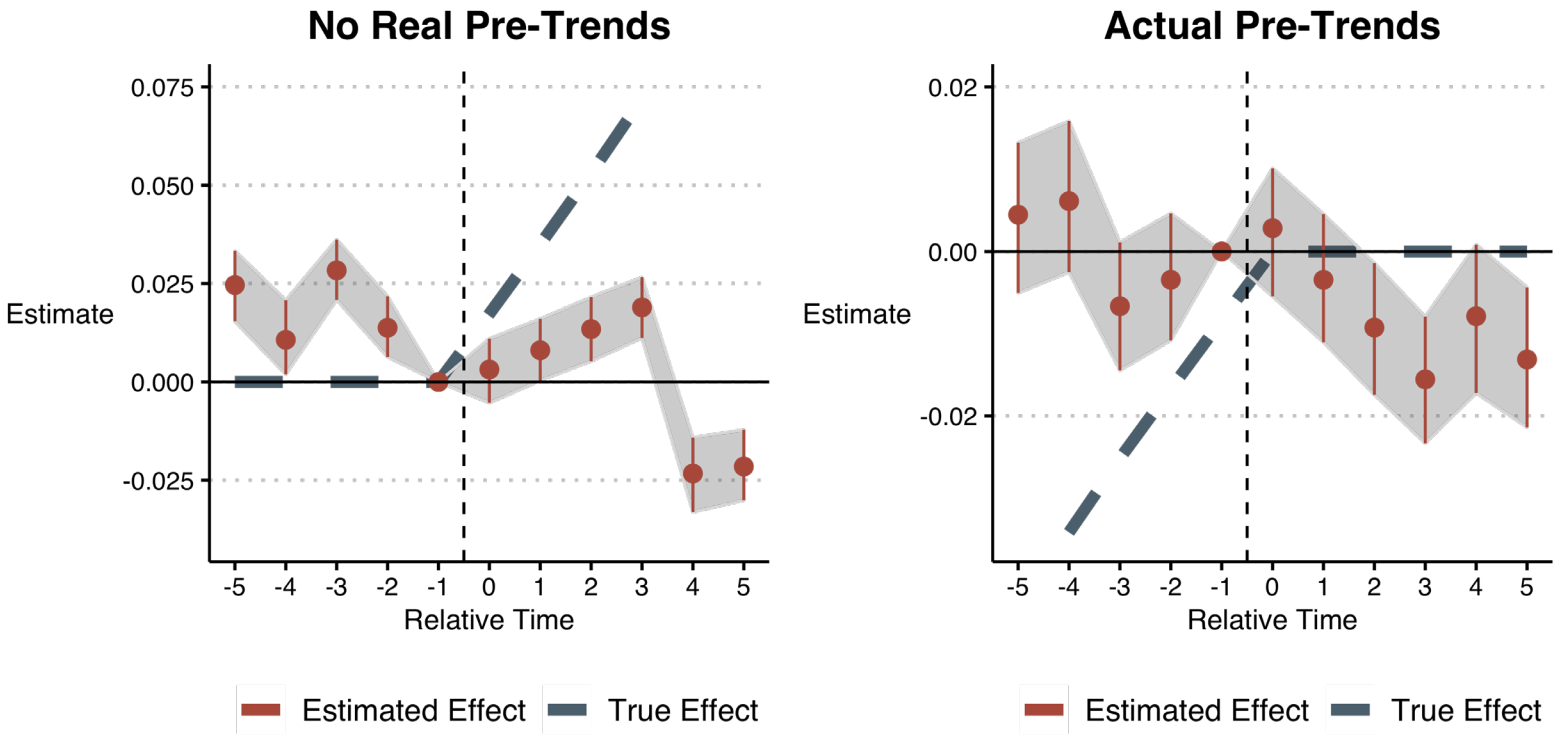
Assumptions about timing and heterogeneity

Problems with 5 and 6



Source: Baker, Larcker, and Wang (JFE 2022)

Can we solve the issue with dynamic treatment effect estimates?



Probably not…

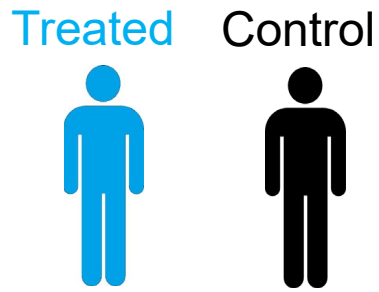
Source: Baker, Larcker, and Wang (JFE 2022)

Staggered + Dynamic Treatment Effect

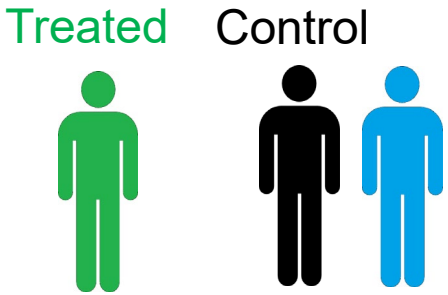
- Why is this so problematic?
  - E.g., in simulations 5 and 6

Staggered + Constant Treatment Effect

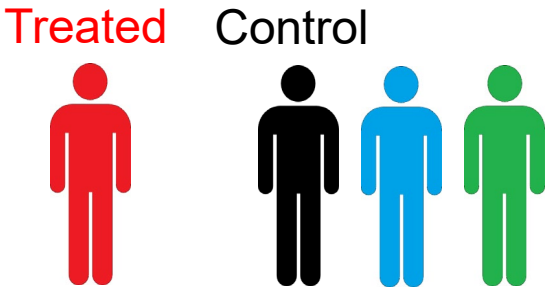
- Treatment 1



- Treatment 2

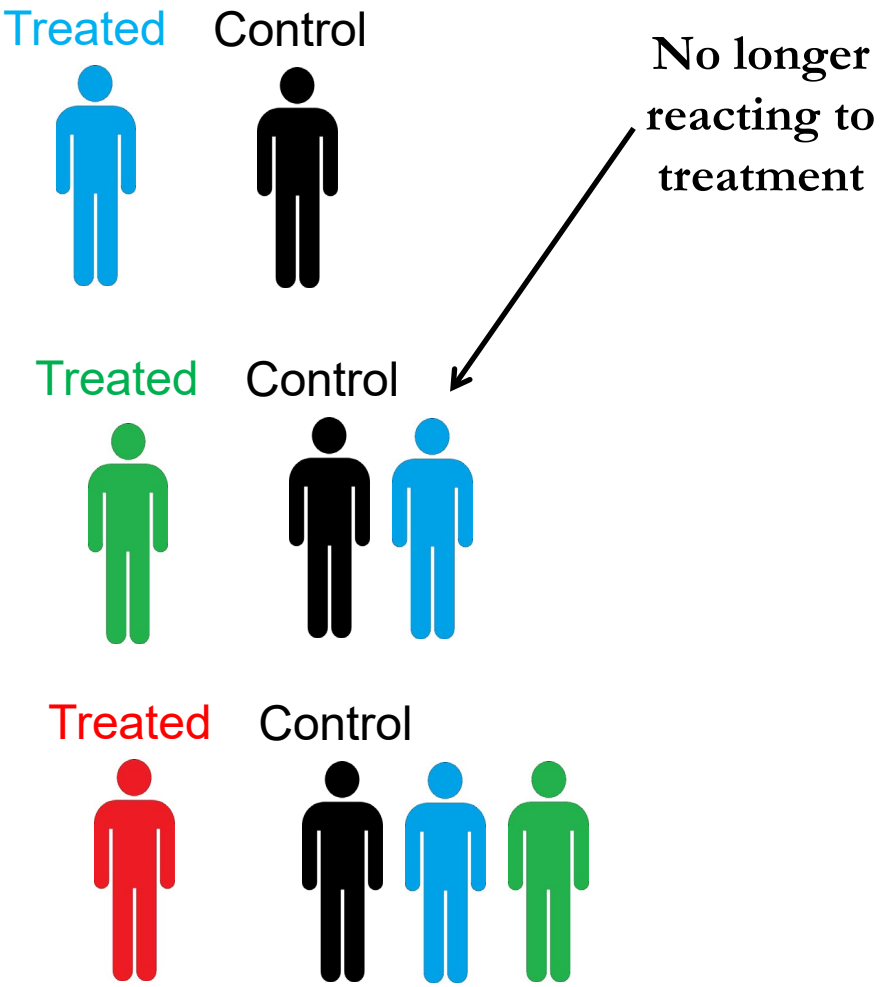


- Treatment 3



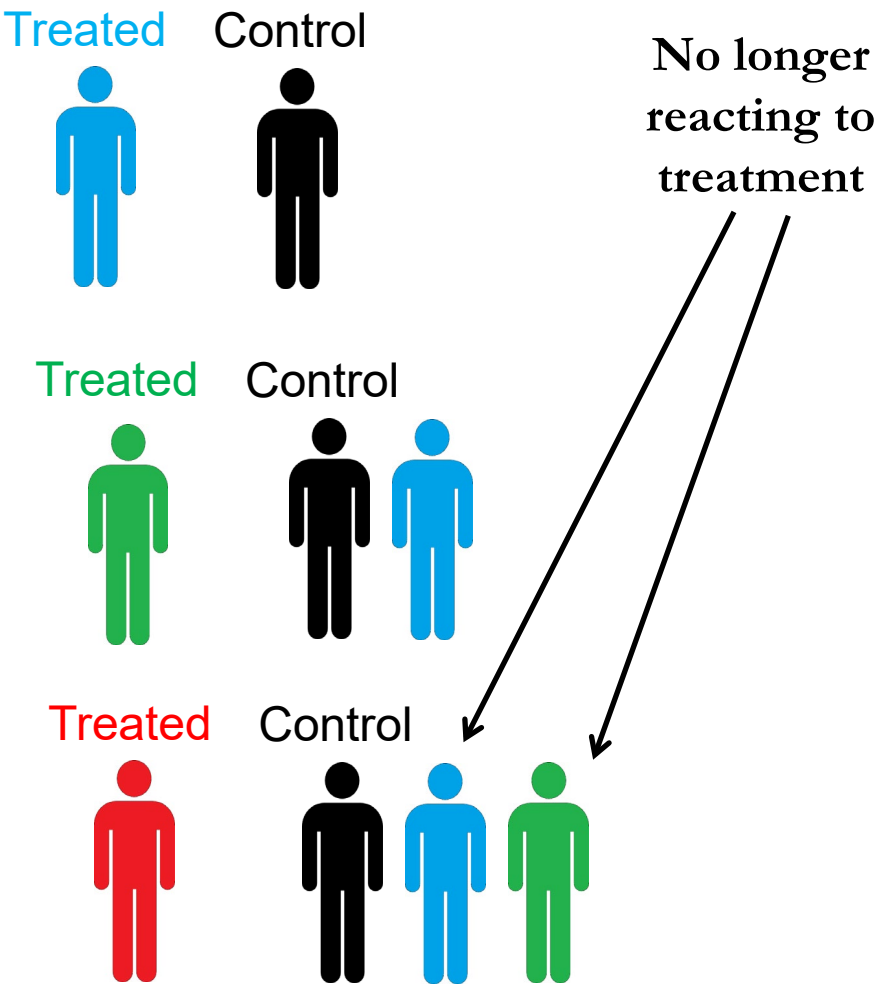
Staggered + Constant Treatment Effect

- Treatment 1
- Treatment 2
- Treatment 3



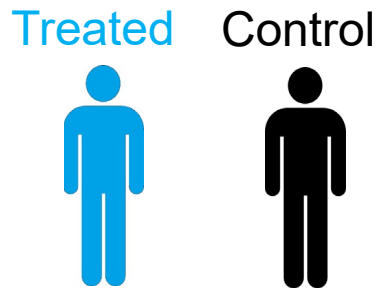
Staggered + Constant Treatment Effect

- Treatment 1
- Treatment 2
- Treatment 3

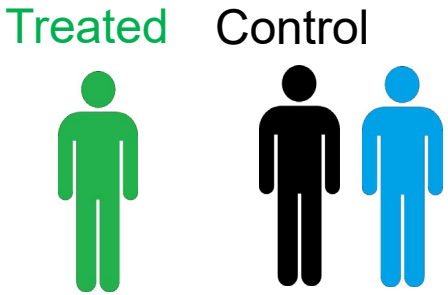


Staggered + Dynamic Treatment Effect

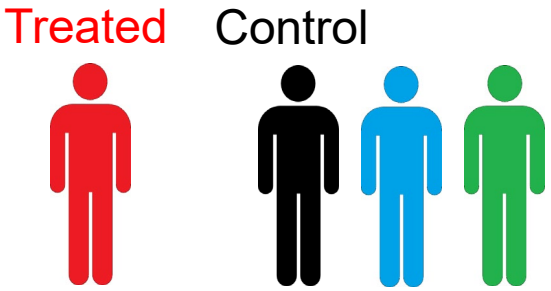
- Treatment 1



- Treatment 2

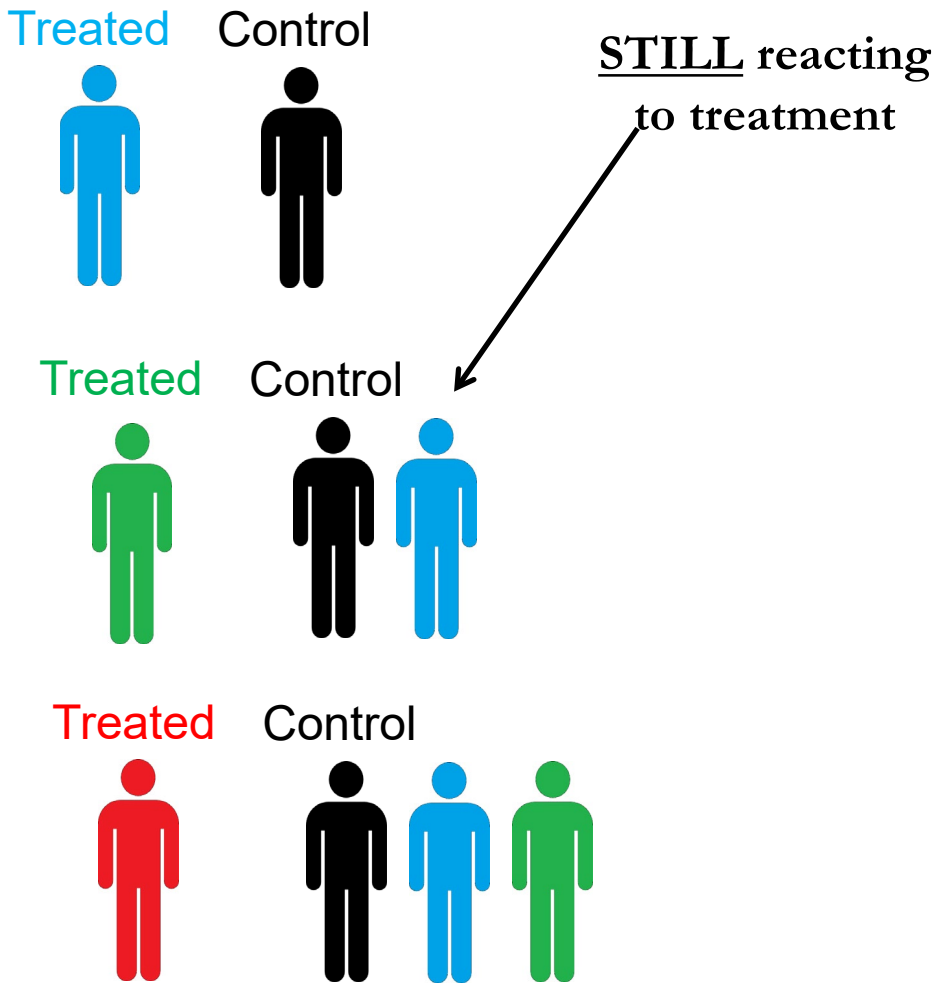


- Treatment 3



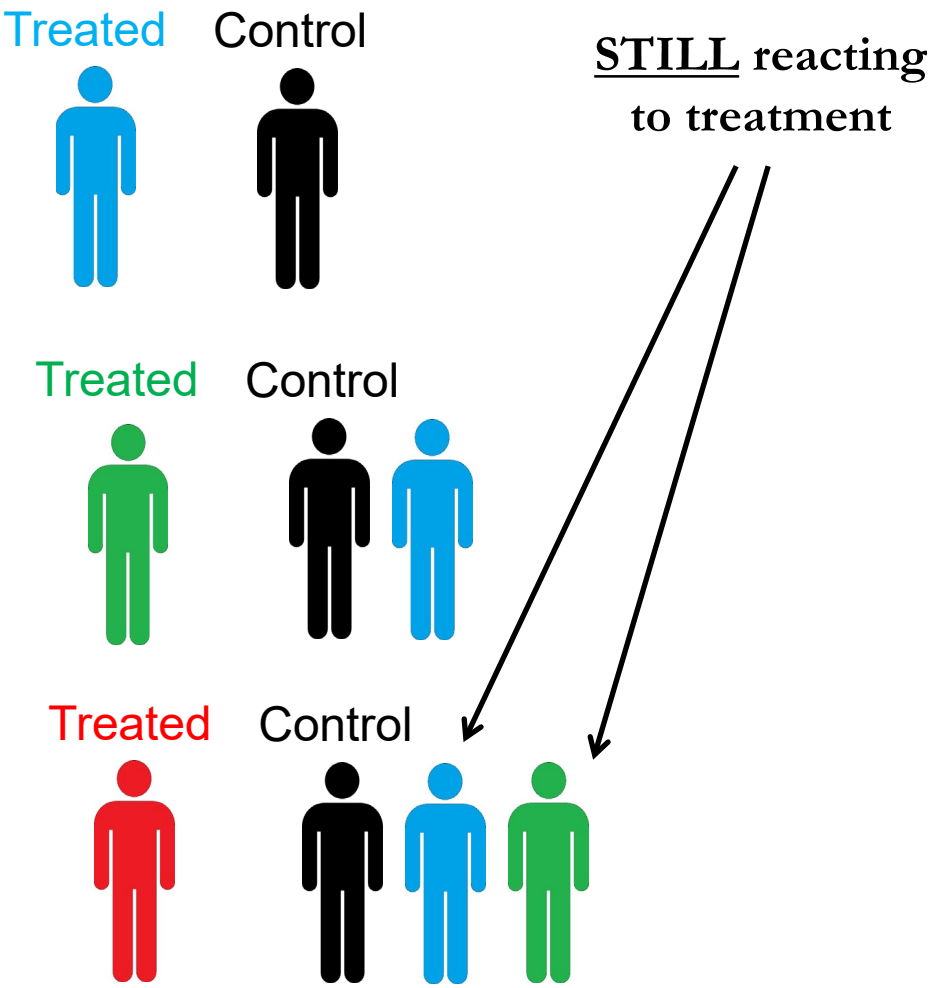
Staggered + Dynamic Treatment Effect

- Treatment 1
- Treatment 2
- Treatment 3

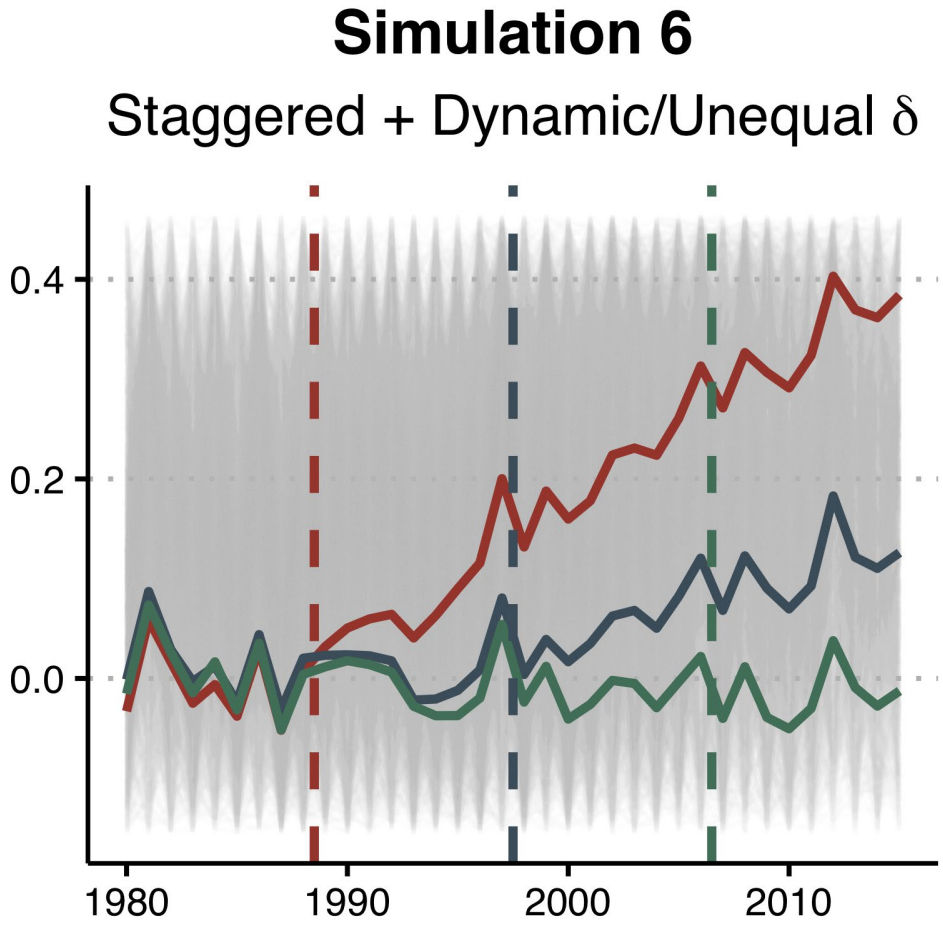


Staggered + Dynamic Treatment Effect

- Treatment 1
- Treatment 2
- Treatment 3

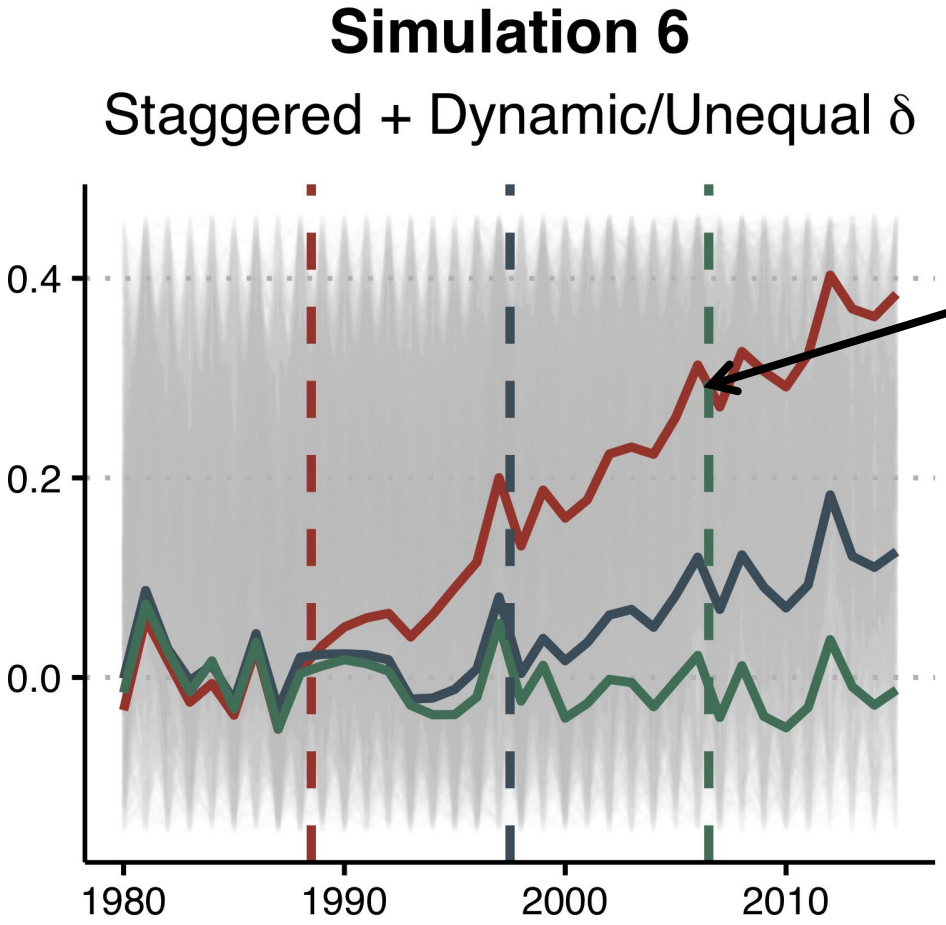


Staggered + Dynamic Treatment Effect



Source: Baker, Larcker, and Wang (JFE 2022)

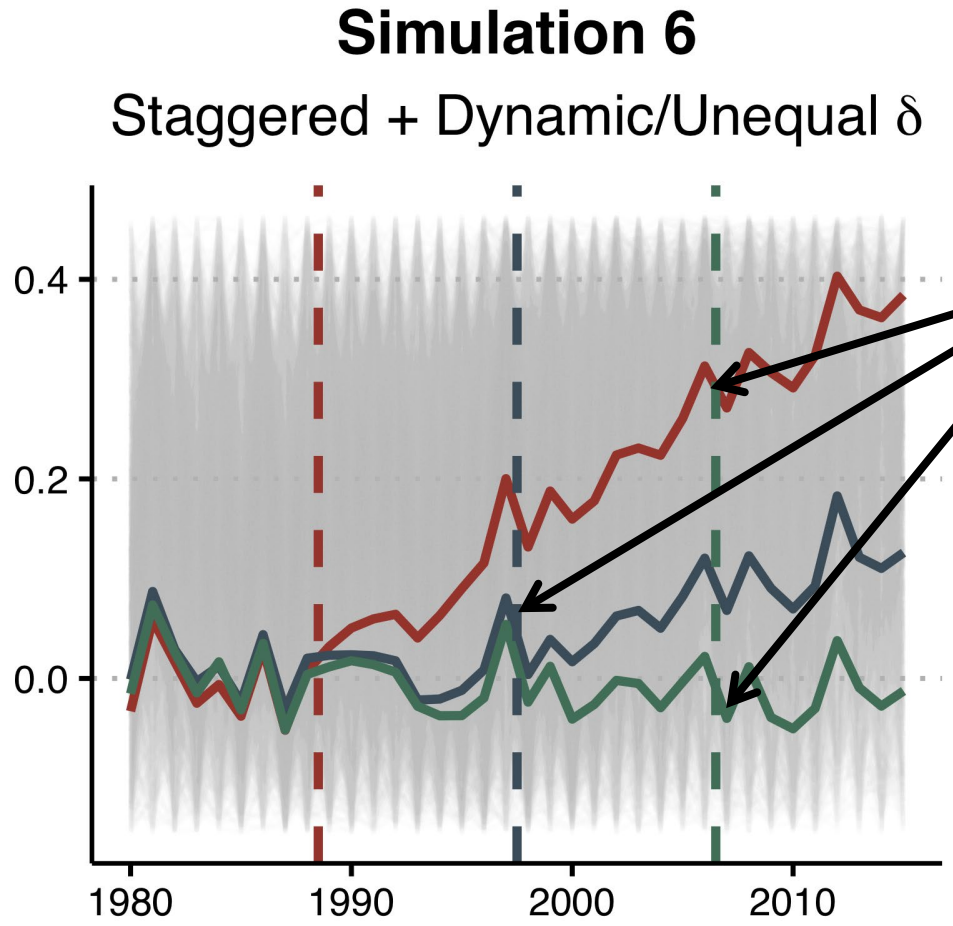
Staggered + Dynamic Treatment Effect



If the 1989 treatment group is part of the control for the 2007 treatment, easy to see why the ATT for that cohort could be negative!

Source: Baker, Larcker, and Wang (JFE 2022)

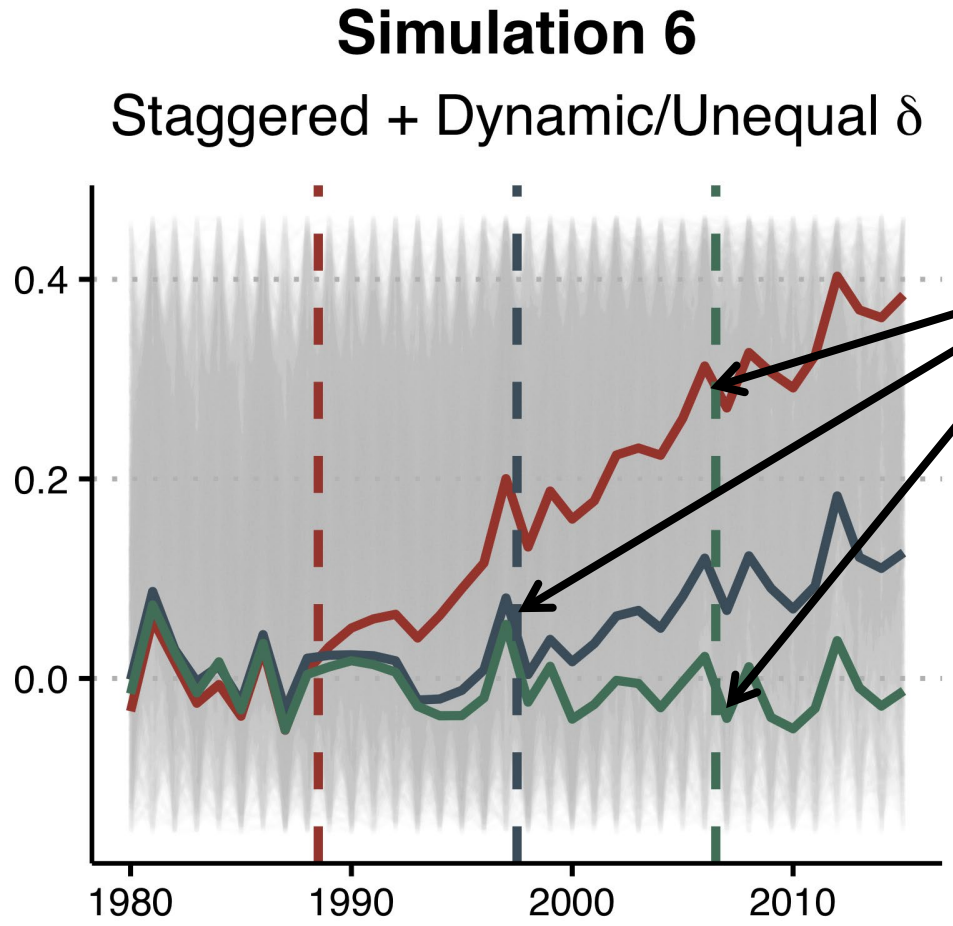
Staggered + Dynamic Treatment Effect



We need OLS to properly weight the different cohorts in aggregating the treatment effects.

Source: Baker, Larcker, and Wang (JFE 2022)

Staggered + Dynamic Treatment Effect



We need OLS to properly weight the different cohorts in aggregating the treatment effects.

Unfortunately, it does not.

Source: Baker, Larcker, and Wang (JFE 2022)

Natural Experiment – *Outline*

- Staggered (multiple) treatments and heterogeneous treatment effects
  - How to handle multiple events
    - Why they are useful
    - Simple estimation approache
  - Problems with this design
    - Solutions to this problem

Three Potential Solutions to the Problem

- Calloway and Sant'Anna (CS) (2021) Estimator
- Sun and Abram (SA) (2021) Estimator
- Stacked Regression (Gormley and Matsa, 2011)

CS and SA Estimators *[P1]*

- Both CS and SA
  - First estimate the individual cohort-time-specific treatment effects allowing for treatment effect heterogeneity,
  - Then, aggregate them to produce measures of overall treatment effects.
- However, significant differences in flexibility, covariates, control groups and inference

CS and SA Estimators *[P2]*

- CS - *csdid*
  - Allows for greater flexibility in aggregation technique (e.g., can get to SA's)
  - Allows for greater control in selecting control group
  - Allows for pre-treatment and static covariates
  - Uses simultaneous CI's estimated with a multiplier bootstrapping procedure

CS and SA Estimators *[P3]*

- SA – *eventstudyinteract*
  - Fully parametric regression-based estimator with full set of cohort-specific relative-time treatment effects
  - Weights cohort-specific ATTs with each cohort's share of sample
  - Limits control group to never-treated or last-treated, no covariates
  - Uses pointwise inference with asymptotic SE's (e.g., no bootstrapping)

Three Potential Solutions to the Problem

- Calloway and Sant'Anna (CS) (2021) Estimator
- Sun and Abram (SA) (2021) Estimator
- Stacked Regression (Gormley and Matsa, 2011)

Stacked Regression *[P1]*

- Think of running generalized diff-in-diff for just one of the multiple events…

$$y_{it} = \beta\left(d_i \times p_t\right) + \alpha_i + \delta_t + u_{it}$$

  - d_i = indicator for unit i (e.g. firm) being a treated firm in that particular event
  - p_t = indicator for treatment having occurred by period t (e.g. year)
  - Unit i and period t FE control for the independent effects of d_i and p_t

Stacked Regression *[P2]*

- But, contrary to standard difference-in-difference, your sample is…
  - Restricted to a small window around event; e.g. 5 years pre- and post- event
  - And, drops any observations that are treated by another event
    - I.e. your sample starts only with previously untreated observations, and if a 'control' observation later gets treated by a different event, those post-event observations are dropped

Stacked Regression *[P3]*

- Now, create a similar sample for each "event" being analyzed
- Then, "stack" the samples into one dataset and create a variable that identifies the event (i.e. 'cohort') each observation belongs to
  - **Note:** some observation units will appear multiple times in the data [e.g. firm 123 might be a control in event year 1999 but a treated firm in a later event in 2005]

Stacked Regression *[P4]*

- Then, estimate the following on the stacked dataset you've created

$$y_{ict} = \beta d_{ict} + \delta_{tc} + \alpha_{ic} + u_{ict}$$

d_{ict} = indicator on whether cohort *c* is affected by time *t*; this is the interaction between *treatment* & *post*

Time-cohort period fixed effects; they control for *post* dummy in each event (i.e. for each 'stack')

Unit-cohort FE; they control for the *treatment* dummy in each cohort (i.e. in each 'stack')

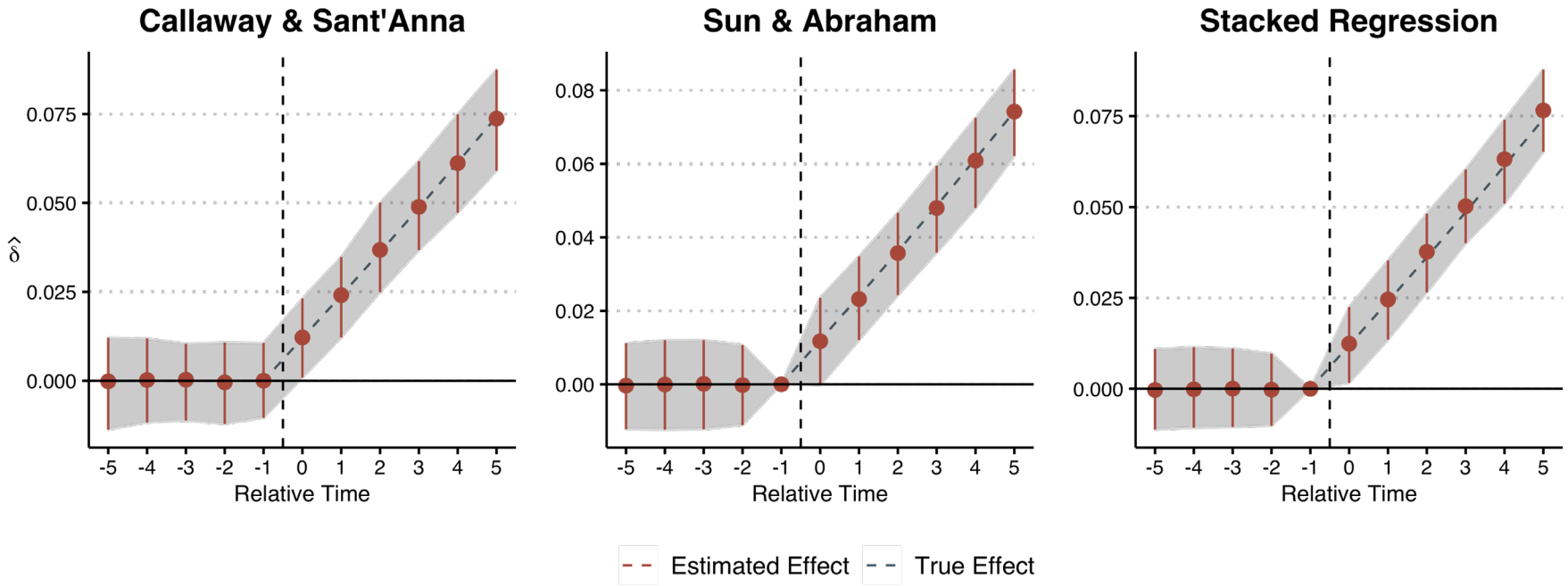
Stacked Regression *[P5]*

- Still uses OLS to weight cohorts
  - However, because each cohort ATT is estimated separately, the variance weighting is efficient
  - In other words, removing the "already treated" observations from the later control group, helps avoid creating biased estimates due to treatment effect heterogeneity (Calloway and Sant'Anna, 2021, Goodman-Bacon, 2021)

Stacked Regression *[P6]*

- This approach has same intuition of the BM(2003) approach, but has a couple further advantages
  - Can more easily isolate a particular window of interest around each event
    - Prior approach compared all pre- versus post-treatment observations against each other
  - Can more easily extend this into a triple-difference type specification *[more on that later]*

Back to Simulation 6



Source: Baker, Larcker, and Wang (JFE 2022)

Natural Experiment – *Outline*

- Difference-in-difference continued…
  - When additional controls are appropriate
- Falsification tests
- Triple differences

Adding controls to diff-in-diff

- Easy to add controls to regression

$$y_{i,t} = \beta_0 + \beta_1 p_t + \beta_2 d_i + \beta_3 \left(d_i \times p_t \right) + \Gamma X_{i,t} + u_{i,t}$$

  - X is some vector of controls
  - Γ is vector of coefficients

- $E[y|d,p]$ in prior proofs just becomes $E[y|d,p,X]$

From earlier lecture, what type of controls should you NEVER add?

When controls are inappropriate

- Remember! You should never add controls that might themselves be affected by treatment
  - Angrist-Pischke call this a "bad control"
  - You won't be able to get a consistent estimate of β_3 from estimating the equation
- If you have a treatment that is truly random, do not put in controls affected by the treatment!

When controls are appropriate

- Two main reasons to add controls
  - Improve precision (i.e. lower standard errors)
  - Restore 'random' assignment of treatment

#1 – To improve precision

- Adding controls can soak up some of residual variation (i.e. noise) allowing you to better isolate the treatment effect
  - Should the controls change the estimate?
    - NO! If treatment is truly random, adding controls shouldn't affect actual estimate; they should only help lower the standard errors!
  - If adding controls changes estimates, you might have 'bad controls' *or* worse, non-random treatment ☹

Example – Improving precision

- Suppose you have firm-level panel data
- Some natural experiment 'treats' some firms but not other firms
  - Could just estimate the standard diff-in-diff

$$y_{i,t} = \beta_0 + \beta_1 p_t + \beta_2 d_i + \beta_3 \left(d_i \times p_t \right) + u_{i,t}$$

  - *Or*, could add fixed effects (like firm and year FE) to get more precise estimate…

Example – Improving precision *[Part 2]*

- So, suppose you estimate…

$$y_{i,t} = \beta_0 + \beta_1 p_t + \beta_2 d_i + \beta_3 \left(d_i \times p_t\right) + \alpha_i + \delta_t + u_{i,t}$$

Firm fixed effects

Year fixed effects

- **What meaning does β_1 have now?**
- **What meaning does β_2 have now?**

Example – Improving precision *[Part 3]*

- Trick question! They have no meaning!
  - p_t is perfectly collinear with year FE
    [because it doesn't vary across firms]
  - d_i is perfectly collinear with firm FE
    [because it doesn't vary across time for each firm]
- Stata just randomly drops a couple of the FE
  - The estimates on p_t and d_i are just random intercepts with **no** meaning

Example – Improving precision *[Part 4]*

- Instead, you should estimate…

$$y_{i,t} = \beta_0 + \beta_3 \left(d_i \times p_t \right) + \alpha_i + \delta_t + u_{i,t}$$

Firm fixed effects control for treatment

Year fixed effects control for post-treatment

  - **This is what some call the generalized difference-in-differences estimator**

Generalized Difference-in-differences

- Advantage of generalized differences-in-differences is that it can improve precision and provide better fit of model
  - It doesn't assume all firms in treatment (or untreated) group have same average *y*; it allows intercept to vary for each firm
  - It doesn't assume that common change in *y* around event is a simple change in level; it allows common change in *y* to vary by year

When controls are appropriate

- Two main reasons to add controls
  - Improve precision (i.e. lower standard errors)
  - Restore 'random' assignment of treatment

#2 – Restore randomness of treatment

- Suppose the following is true…
  - Observations of certain characteristic, e.g. high x, are more likely to be treated
  - **And,** firms with this characteristic are likely to have differential trend in outcome y
- Adding control for x could restore 'randomness'; i.e. being treated is random after controlling for x!

I.e. treatment isn't random

And, non-randomness is problematic for identification

Restoring randomness – *Example*

- Natural experiment is change in regulation
  - Firms affected by regulation is random, except that it is more likely to hit firms that are larger
  - ***And***, we think larger firms might have different trend in outcome *y* afterwards for other reasons
  - ***And,*** firm size is not going to be affected by the change in regulation in any way
- If all true, adding size as control would be an appropriate and desirable thing to do

Controls continued…

- In prior example, suppose size is potentially effected by the change in regulation…
  - What would be another approach that won't run afoul of the 'bad control' problem?
    - **Answer:** Use firm size in year prior to treatment and it's interaction with post-treatment dummy
    - This will control for non-random assignment (based on size) and differential trend (based on size)

Restoring randomness – **Caution!**

- In practice, don't often see use of controls to restore randomness
  - Requires assumption that non-random assignment isn't also correlated with unobservable variables…
  - So, not that plausible unless there are very specific reasons for non-randomness
- **But, regression discontinuity is one example of this; we'll see it next week**

One last note… be careful about SEs

- When you have multiple pre- and post-treatment periods, need to be careful with standard errors
  - Either cluster SEs at level of each unit
  - Or, collapse data down to one pre- and one post-treatment observation for each cross-section
- We will discuss more about standard errors in lecture on "standard errors"

Natural Experiment [P2] – *Outline*

- Difference-in-difference continued…
- Falsification tests
- Triple differences

Falsification Tests for D-i-D

- Can never directly test underlying identification assumption, but can do some falsification tests to support its validity

  #1 – Compare pre-treatment observables

  #2 – Check that timing of observed change in *y* coincides with timing of event *[i.e. no pre-trend]*

  #3 – Check for treatment reversal

  #4 – Check variables that shouldn't be affected

  #5 – Add a triple-difference

#1 – Pre-treatment comparison *[Part 1]*

- Idea is that experiment 'randomly' treats some subset of observations
  - If true, then ex-ante characteristics of 'treated' observations should be similar to ex-ante characteristics of 'untreated' observations
  - Showing treated and untreated observations are comparable in dimensions thought to affect *y* can help ensure assignment was random

#1 – Pre-treatment comparison *[Part 2]*

- If find ex-ante difference in some variable z, is difference-in-difference is invalid?
  - **Answer** = Not necessarily.
    - We need some story as to why units are expected to have differential trend in *y* after treatment (for reasons unrelated to treatment) that is correlated with z for this to actually be a problem for identification
    - **And,** even with this story, we could just control for z and it's interaction with time
    - **But, what would be the lingering concern?**

#1 – Pre-treatment comparison *[Part 3]*

- **Answer =** unobservables!
  - If the treated and control differ ex-ante in observable ways, we worry they might differ in unobservable ways that related to some violation of the parallel trends assumption

#2 – Check for pre-trend *[Part 1]*

- Can just allow effect of treatment to vary by period to non-parametrically map out the timing
  - "Parallel trends" suggest we shouldn't observe any differential trend prior to treatment for the observations that are eventually treated

#2 – Check for pre-trend *[Part 2]*

- Estimate the following:

$$y_{i,t} = \beta_0 + \beta_1 d_i + \beta_2 p_t + \sum_t \gamma_t \left(d_i \times \lambda_t \right) + u_{i,t}$$

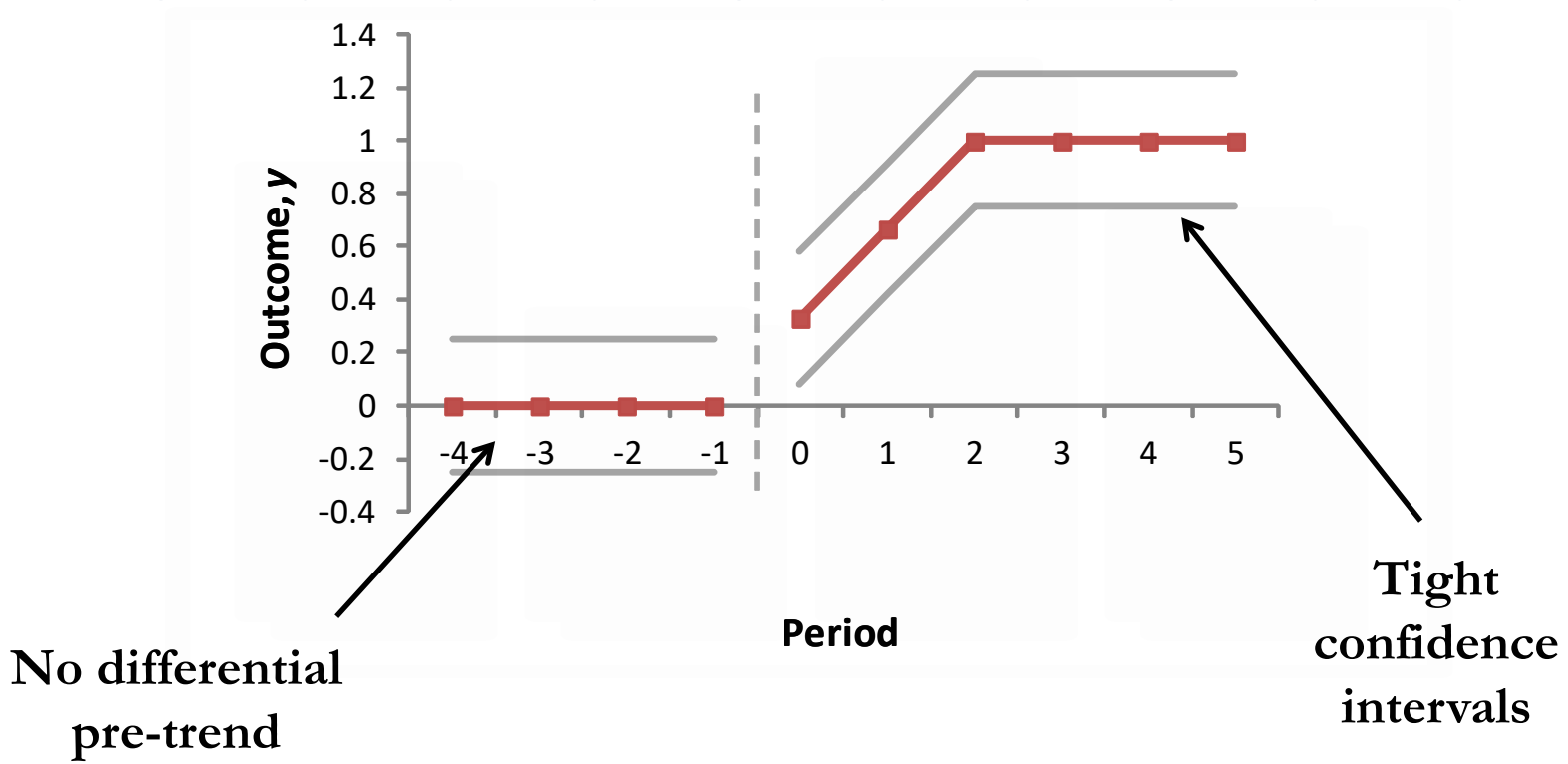
  - d_i and p_t are defined just as before
  - λ_t is indicator that equals 1 if event time = t and zero otherwise, where
    - $t = 0$ is the period treatment occurs
    - $t = -1$ is period before treatment

#2 – Check for pre-trend *[Part 3]*

- γ_t estimates change in *y* relative to excluded periods; you then plot these in graph
  - Easiest to **fully saturate** the model (i.e. include λ_t for every period but the very first one); then all estimates γ_t are relative to this period
  - Can also plot confidence interval for each γ_t

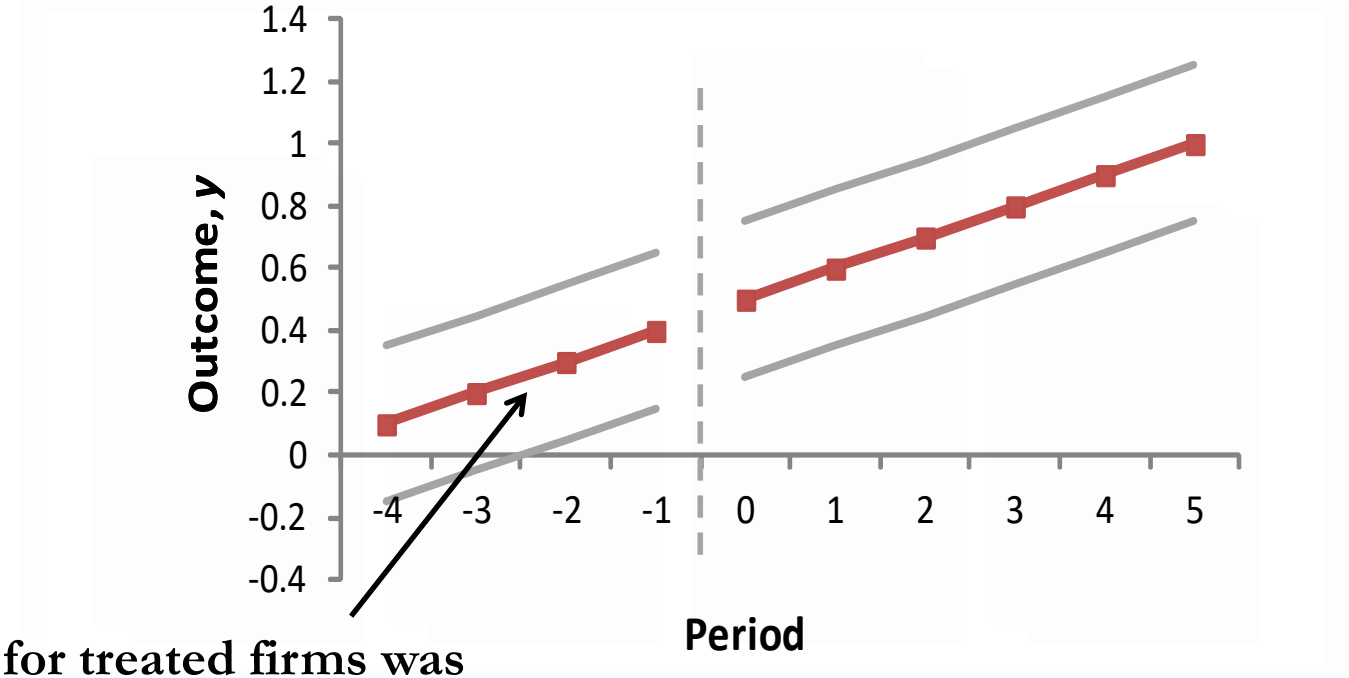
#2 – Check for pre-trend *[Part 4]*

- Something like this is ideal…



#2 – Check for pre-trend *[Part 5]*

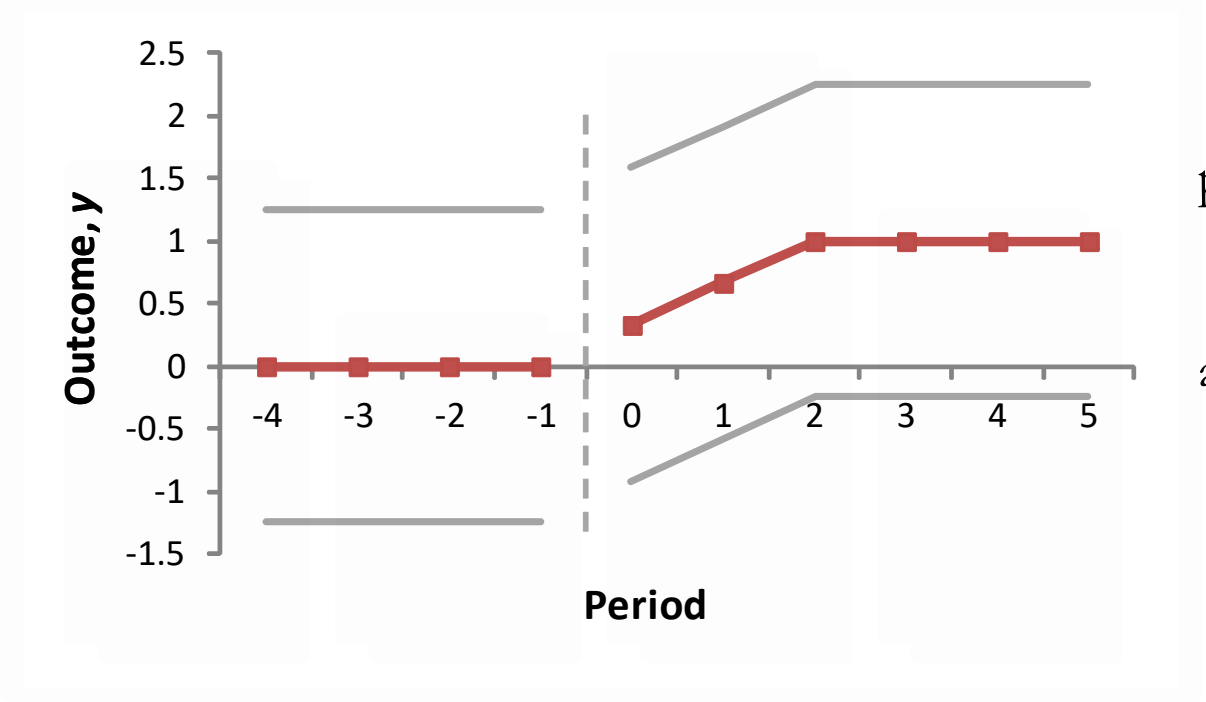
- Something like this is very bad



***y* for treated firms was already going up at faster rate prior to event!**

#2 – Check for pre-trend *[Part 6]*

- Should we make much of wide confidence intervals in these graphs? E.g.



Answer: Not too much… Each period point estimate might be noisy; diff-in-diff will tell us whether post-average *y* is significantly different then pre-average *y*

#3 – Treatment reversal

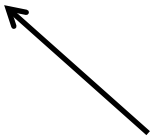
- In some cases, the "natural experiment" is subsequently reversed
  - E.g. regulation is subsequently undone
- If we expect the reversal should have the opposite effect, it is good to confirm this

#4 – Unaffected variables

- In some cases, theory provides guidance on what variables should be unaffected by the "natural experiment"
  - If natural experiment is what we think it is, we should see this in the data… so check

#5 – Add Triple difference

- If theory tells us treatment effect should be larger for one subset of observations, we can check this with triple difference
  - Pre- *versus* post-treatment
  - Untreated *versus* treated
  - Less sensitive *versus* More sensitive



This is the third difference

Natural Experiment Outline – *Part 2*

- Difference-in-difference continued…
- Falsification tests
- Triple differences
  - How to estimate & interpret it
  - Using the popular subsample approach

Diff-in-diff-in-diff – Regression

$$y_{i,t} = \beta_0 + \beta_1 p_t + \beta_2 d_i + \beta_3 h_i + \beta_4 \left(p_t \times h_i \right) + \beta_5 \left(d_i \times h_i \right) + \beta_6 \left(p_t \times d_i \right) + \beta_7 \left(p_t \times d_i \times h_i \right) + u_{i,t}$$

- $p_t = 1$ if period t occurs after treatment and equals zero otherwise
- $d_i = 1$ if unit is in treated group and equals zero otherwise
- $h_i = 1$ if unit is group that is expected to be more sensitive to treatment

Diff-in-diff-in-diff – Regression *[Part 2]*

- How to choose and set h_i
  - E.g. If theory says effect is bigger for larger firms; could set $h_i = 1$ if assets of firm in year prior to treatment is above the median size
  - **Note:** Remember to use ex-ante measures to construct indicator if you think underlying variable (that determines sensitivity) might be affected by treatment… Why?
    - **Answer =** To avoid bad controls!

Diff-in-diff-in-diff – Regression *[Part 3]*

$$y_{i,t} = \beta_0 + \beta_1 p_t + \beta_2 d_i + \beta_3 h_i + \beta_4 \left(p_t \times h_i \right) + \beta_5 \left(d_i \times h_i \right) + \beta_6 \left(p_t \times d_i \right) + \beta_7 \left(p_t \times d_i \times h_i \right) + u_{i,t}$$

- **What do β_6 and β_7 capture?**

Interpreting the estimates *[Part 1]*

- β_6 diff-in-diff estimate for the less-sensitive obs.
  - Captures average <u>differential</u> change in *y* from the pre- to post-treatment period for the less sensitive observations in the treatment group ***relative*** to the change in *y* for the less sensitive observations in the untreated group

Interpreting the estimates *[Part 2]*

- β_7 is the triple diff estimate; it tells us how much larger effect is for the more sensitive obs.
  - β_7 captures how different the difference-in-difference estimate is for observations considered more sensitive to the treatment
  - **What is total treatment effect for these firms?**
  - **Answer =** $\beta_6+\beta_7$

Tangent – Continuous vs. Indicator?

- Can also do the triple difference replacing h_i with a continuous measure instead of indicator
  - E.g. suppose we expect treatment effect is bigger for larger firms; rather than constructing indicator based on ex-ante size, could just use ex-ante size
  - **What are the advantages, disadvantages of this?**

Tangent – Continuous vs. Indicator?

- Advantages
  - Makes better use of variation available in data
  - Provides estimate on magnitude of sensitivity
- Disadvantages
  - Makes linear functional form assumption; indicator imposes less structure on the data
  - More easily influenced by outliers

Generalized Triple-Difference

- Similar to diff-in-diff, can add in FE to soak up the various terms and improve precision
- E.g. in firm-level panel regression with firm and year fixed effects, you'd estimate

$$y_{i,t} = \beta_1 \left(p_t \times h_i \right) + \beta_2 \left(p_t \times d_i \right) + \beta_3 \left(p_t \times d_i \times h_i \right) + \delta_t + \alpha_i + u_{i,t}$$

  - The other terms (including the constant) all drop out; they are collinear with the FE

Natural Experiment [P2] – *Outline*

- Difference-in-difference continued…
- Falsification tests
- Triple differences
  - How to estimate & interpret it
  - Using the popular subsample approach

Subsample Approach

- Instead of doing full-blown triple-difference, you can also just estimate the double-difference in the two separate subsamples
  - Double-difference for low sensitive obs. (i.e. $h_i = 0$)
  - Double-difference for more sensitive obs. (i.e. $h_i = 1$)
- **Note:** the estimates won't directly match the β_2, $\beta_2+\beta_3$ effects in prior estimation… **Why?**

Subsample Approach Differences…

- **Answer** = In subsample approach year FE are allowed to differ by sub-sample
  - Therefore, subsample approach is actually controlling for **more** things
  - However, one can easily recover the subsample estimates in one regression (and test the statistical difference) between subsamples by estimating…

Matching Subsample to Combined *[P1]*

$$y_{i,t} = \beta_2 \left(p_t \times d_i \right) + \beta_3 \left(p_t \times d_i \times h_i \right) + \delta_t + \left(\delta_t \times h_i \right) + \alpha_i + u_{i,t}$$

Year FE interacted with sensitivity indicator

- Just add interaction between year FE and indicator for being more sensitivity…
  - This allows for different year FE for each subsample, which is what happened when we estimated the subsamples in two separate regressions

Matching Subsample to Combined *[P2]*

- **In prior regression…**
  - β_2 will equal coefficient from diff-in-diff using just the subsample of less sensitive observations
  - $\beta_2+\beta_3$ will equal coefficient from diff-in-diff using just the subsample of more sensitive observations
  - t-test on β_3 tells you whether effect for more sensitive subsample is statistically different from that of the less sensitive subsample

Triple Diff – Stacked Regression *[Part 1]*

- Another advantage of stacked regression approach to multiple events is ability to more easily incorporate a triple diff
  - Can simply run stacked regression in separate subsamples to create triple-diff or run it in one regression as shown previously

Triple Diff – Stacked Regression *[Part 2]*

- Can't easily do either of these in approach of Bertrand and Mullainathan (2003)
  - Some observations act as both 'control' and 'treated' at different points in sample; not clear how create subsamples in such a setting

External Validity – **Final Note**

- While randomization ensures internal validity (i.e. causal inferences), external validity might still be an issue
  - Is the experimental setting representative of other settings of interest to researchers?
    - I.e. can we extrapolate the finding to other settings?
    - A careful argument that the setting isn't unique or that the underlying theory (for why you observe what you observe) is likely to apply elsewhere is necessary

Summary of Today *[Part 1]*

- Diff-in-diff & control variables
  - Don't add controls affected by treatment
  - Controls shouldn't affect estimates, but can help improve precision
- Multiple events are helpful in mitigating concerns about parallel trends assumption

Summary of Today *[Part 2]*

- Many falsification tests one should do to help assess internal validity
  - Ex. #1 – Compare ex-ante characteristics
  - Ex. #2 – Check timing of observed effect
- Triple difference is yet another way to check internal validity and mitigate concerns about identification

Summary of Today *[Part 3]*

- Simple OLS approach for staggered DiD can be problematic
  - Use CS, SA estimators, or
  - Stacked regression approach

Next time

- Regression discontinuity
  - What are they?
  - How are they useful?
  - How do we implement them?
- Related readings… see syllabus