# Difference-in-Differences

## Mixtape Session

# Roadmap

- Weighted Group-Time ATT
  - Bacon decomposition
  - Castle doctrine reform
  - CS
  - SA
  - dCH

# Differential timing

- We covered mostly the simple two group case
- In the two group case, we can estimate the ATT under parallel trends using OLS with unit and time fixed effects
- If we have covariates, then we can use TWFE under restrictive assumptions, or we have other options (OR, IPW, DR)
- Now let's move to a more common scenario where we have more than two groups who get treated at various times

# 2x2 versus differential timing

- For this next part, similar to how we did with Sant'Anna and Zhao (2020), we will decompose TWFE to understand what it needs for unbiasedness under differential timing
- All of this is from Goodman-Bacon (2021, forthcoming) though the expression of the weights is from 2018 for personal preference
- Goodman-Bacon (2021, forthcoming) shows that parallel trends is **not enough** for TWFE to be unbiased when treatment adoption is described by differential timing
- TWFE with differential timing uses treated groups as controls – not all estimators do – and this can introduce bias

# Decomposition Preview

- TWFE estimates a parameter that is a weighted average over all 2x2 in your sample
- TWFE assigns weights that are a function of sample sizes of each "group" and the variance of the treatment dummies for those groups

# Decomposition (cont.)

- TWFE needs two assumptions: that the variance weighted parallel trends are zero (far more parallel trends iow) and no dynamic treatment effects (not the case with 2x2)
- Under those assumptions, TWFE estimator estimates the variance weighted ATT as a weighted average of all possible ATTs

# $K^2$ distinct DDs

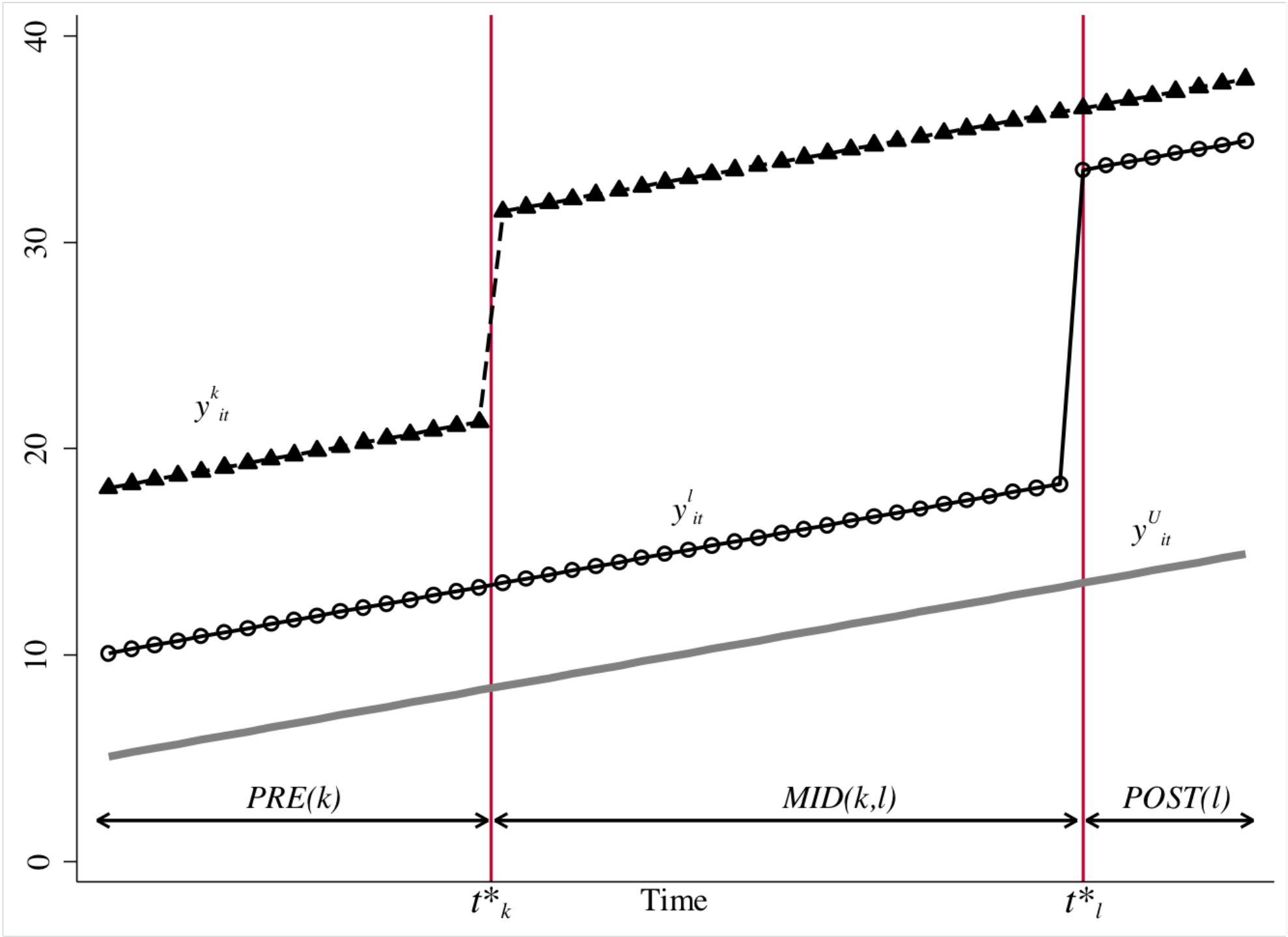
Let's look at 3 timing groups (a, b and c) and one untreated group (U). With 3 timing groups, there are 9 2x2 DDs. Here they are:

| | | |
|---|---|---|
| a to b | b to a | c to a |
| a to c | b to c | c to b |
| a to U | b to U | c to U |

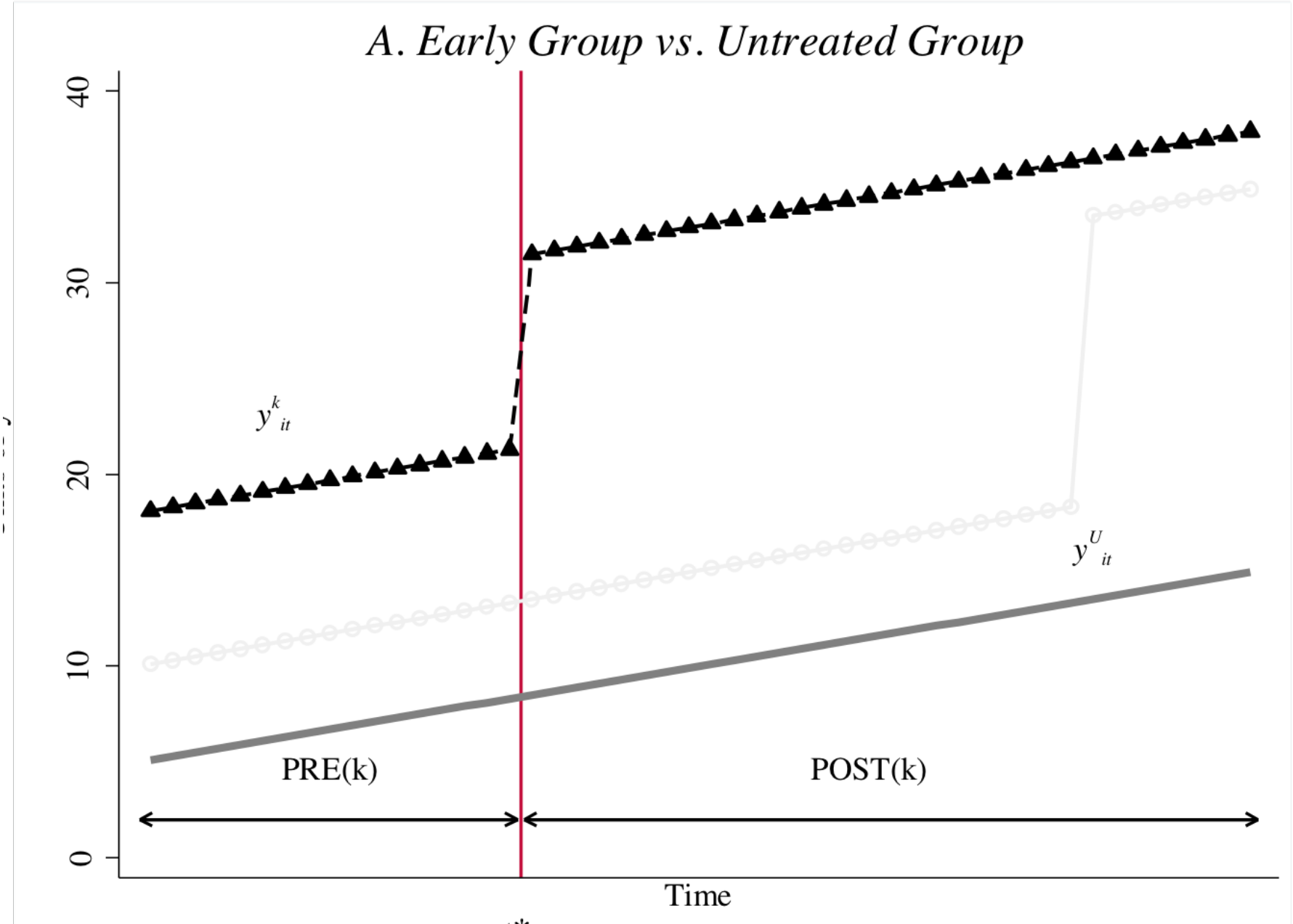
Let's return to a simpler example with only two groups – a $k$ group treated at $t_k^*$ and an $l$ treated at $t_l^*$ plus an never-treated group called the $U$ untreated group

# Terms and notation

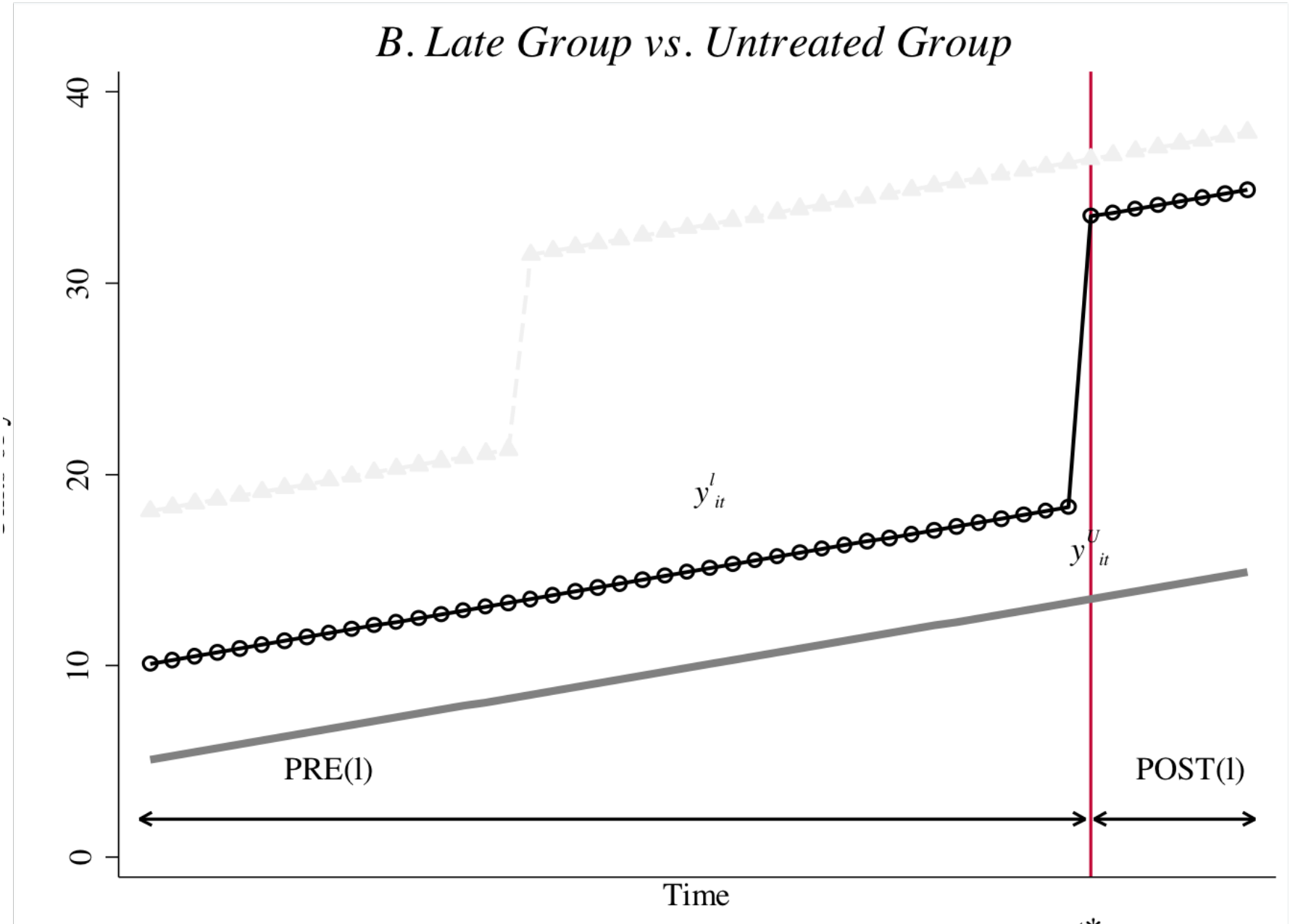
- Let there be two treatment groups (k,l) and one untreated group (U)
- k,l define the groups based on when they receive treatment (differently in time) with k receiving it earlier than l
- Denote $\overline{D}_k$ as the share of time each group spends in treatment status
- Denote $\widehat{\delta}_{jb}^{2x2}$ as the canonical $2 \times 2$ DD estimator for groups $j$ and b where $j$ is the treatment group and $b$ is the comparison group


40
30
20
10
0
$y^{k}_{it}$
$y^{l}_{it}$
$y^{U}_{it}$
PRE(k)
MID(k,l)
POST(l)
$t^{*}_{k}$
Time
$t^{*}_{l}$

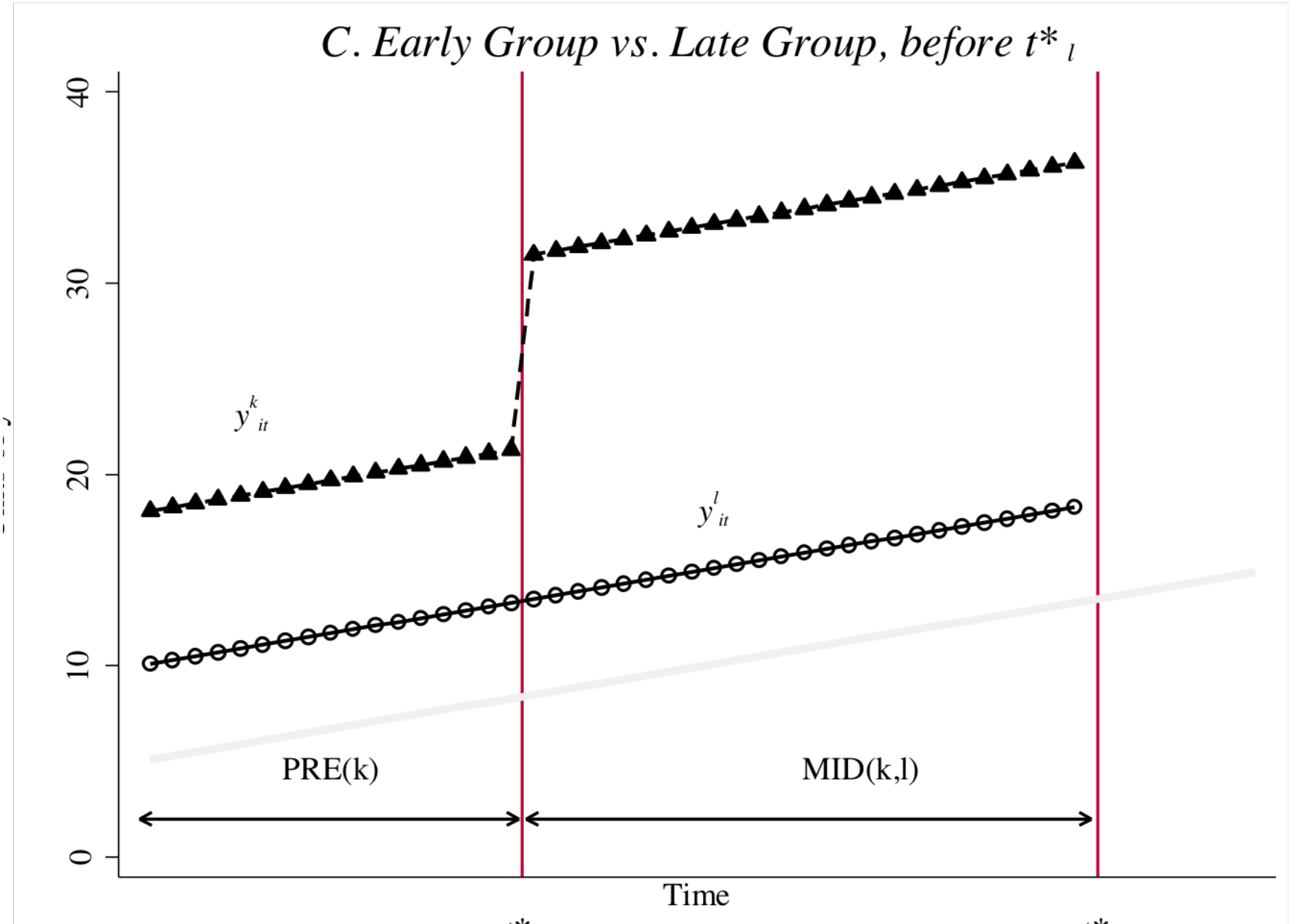
$$\widehat{\delta}_{kU}^{2x2} = \left(\overline{y}_k^{post(k)} - \overline{y}_k^{pre(k)}\right) - \left(\overline{y}_U^{post(k)} - \overline{y}_U^{pre(k)}\right)$$



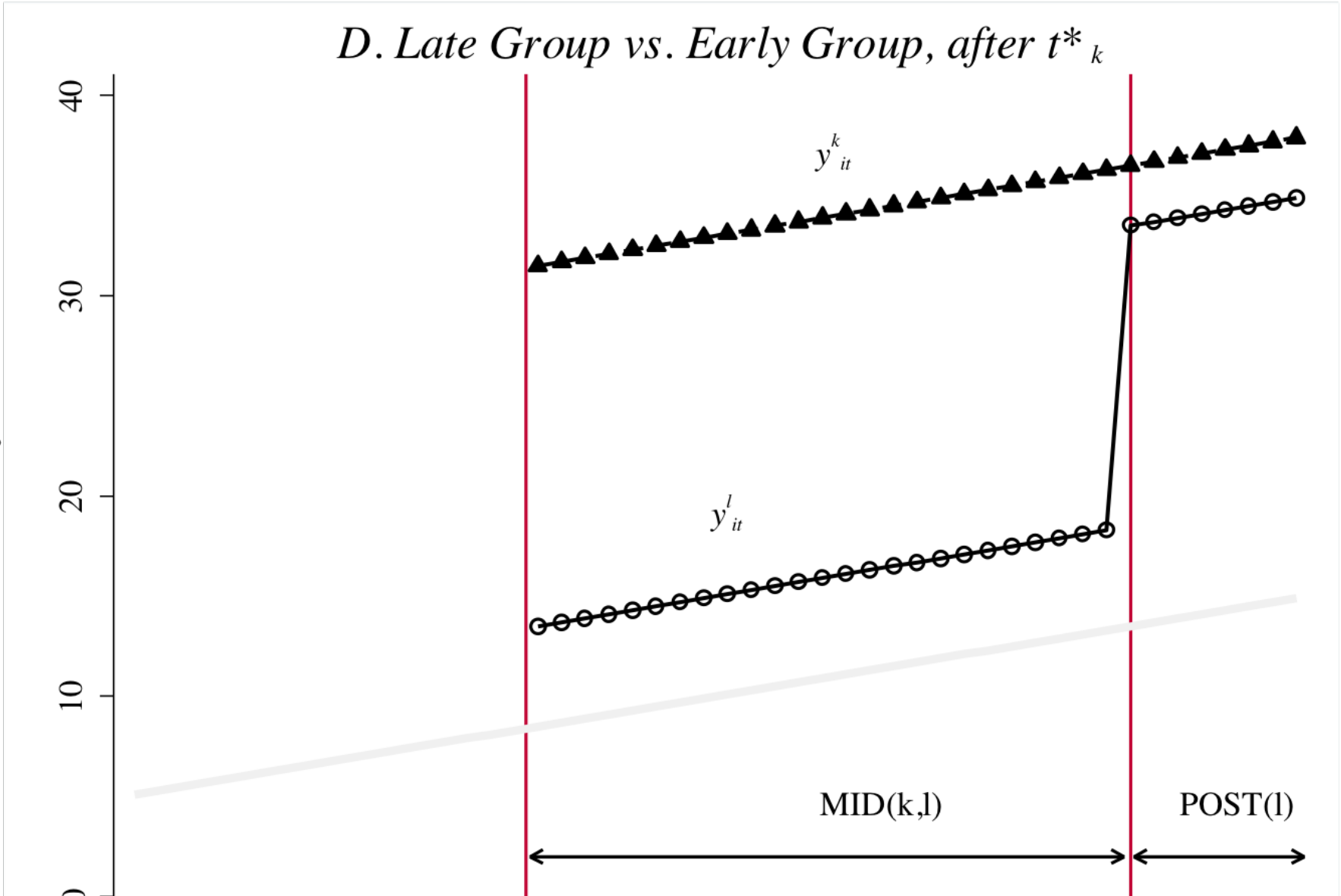
$$\widehat{\delta}_{lU}^{2x2} = \left(\overline{y}_{l}^{post(l)} - \overline{y}_{l}^{pre(l)}\right) - \left(\overline{y}_{U}^{post(l)} - \overline{y}_{U}^{pre(l)}\right)$$



$$\delta_{kl}^{2x2,k} = \left( \overline{y}_k^{MID(k,l)} - \overline{y}_k^{Pre(k,l)} \right) - \left( \overline{y}_l^{MID(k,l)} - \overline{y}_l^{PRE(k,l)} \right)$$



$$\delta_{lk}^{2x2,l} = \left( \overline{y}_{l}^{POST(k,l)} - \overline{y}_{l}^{MID(k,l)} \right) - \left( \overline{y}_{k}^{POST(k,l)} - \overline{y}_{k}^{MID(k,l)} \right)$$



# Bacon decomposition

TWFE estimate yields a weighted combination of each groups' respective 2x2 (of which there are 4 in this example)

$$\widehat{\delta}^{DD} = \sum_{k \neq U} s_{kU} \widehat{\delta}^{2x2}_{kU} + \sum_{k \neq U} \sum_{l > k} s_{kl} \left[ \mu_{kl} \widehat{\delta}^{2x2,k}_{kl} + (1 - \mu_{kl}) \widehat{\delta}^{2x2,l}_{lk} \right]$$

where that first 2x2 combines the k compared to U and the l to U (combined to make the equation shorter)

# Third, the Weights

$$
\begin{aligned}
s_{ku} &= \frac{n_k n_u \overline{D}_k (1 - \overline{D}_k)}{\widehat{Var}(\tilde{D}_{it})} \\
s_{kl} &= \frac{n_k n_l (\overline{D}_k - \overline{D}_l)(1 - (\overline{D}_k - \overline{D}_l))}{\widehat{Var}(\tilde{D}_{it})} \\
\mu_{kl} &= \frac{1 - \overline{D}_k}{1 - (\overline{D}_k - \overline{D}_l)}
\end{aligned}
$$

where $n$ refer to sample sizes, $\overline{D}_k(1 - \overline{D}_k)$ $(\overline{D}_k - \overline{D}_l)(1 - (\overline{D}_k - \overline{D}_l))$ expressions refer to variance of treatment, and the final equation is the same for two timing groups.

# Weights discussion

- Two things to note:
  - → More units in a group, the bigger its 2x2 weight is
  - → Group treatment variance weights up or down a group's 2x2
- Think about what causes the treatment variance to be as big as possible. Let's think about the $s_{ku}$ weights.
  - → $\overline{D} = 0.1$. Then $0.1 \times 0.9 = 0.09$
  - → $\overline{D} = 0.4$. Then $0.4 \times 0.6 = 0.24$
  - → $\overline{D} = 0.5$. Then $0.5 \times 0.5 = 0.25$
  - → $\overline{D} = 0.6$. Then $0.6 \times 0.4 = 0.24$
- This means the weight on treatment variance is maximized for *groups treated in middle of the panel*

# More weights discussion

- But what about the “treated on treated” weights (i.e., $\overline{D}_k - \overline{D}_l$)
- Same principle as before - when the difference between treatment variance is close to 0.5, those 2x2s are given the greatest weight
- For instance, say $t_k^* = 0.15$ and $t_l^* = 0.67$. Then $\overline{D}_k - \overline{D}_l = 0.52$. And thus $0.52 \times 0.48 = 0.2496$.

# Summarizing TWFE centralities

- Groups in the middle of the panel weight up their respective 2x2s via the variance weighting
- Decomposition highlights the strange role of panel length when using TWFE
- Different choices about panel length change both the 2x2 and the weights based on variance of treatment

# Moving from 2x2s to causal effects and bias terms

Let's start breaking down these estimators into their corresponding estimation objects expressed in causal effects and biases

$$
\begin{aligned}
\widehat{\delta}_{kU}^{2x2} &= ATT_k Post + \Delta Y_k^0(Post(k), Pre(k)) - \Delta Y_U^0(Post(k), Pre) \\
\widehat{\delta}_{kl}^{2x2} &= ATT_k(MID) + \Delta Y_k^0(MID, Pre) - \Delta Y_l^0(MID, Pre)
\end{aligned}
$$

These look the same because you're always comparing the treated unit with an untreated unit (though in the second case it's just that they haven't been treated *yet*).

# The dangerous 2x2

But what about the 2x2 that compared the late groups to the already-treated earlier groups? With a lot of substitutions we get:

$$\widehat{\delta}_{lk}^{2x2} = ATT_{l,Post(l)} + \underbrace{\Delta Y_l^0(Post(l), MID) - \Delta Y_k^0(Post(l), MID)}_{\text{Parallel trends bias}} - \underbrace{(ATT_k(Post) - ATT_k(Mid))}_{\text{Heterogeneity bias!}}$$

# Substitute all this stuff into the decomposition formula

$$\widehat{\delta}^{DD} = \sum_{k \neq U} s_{kU} \widehat{\delta}^{2x2}_{kU} + \sum_{k \neq U} \sum_{l > k} s_{kl} \left[ \mu_{kl} \widehat{\delta}^{2x2,k}_{kl} + (1 - \mu_{kl}) \widehat{\delta}^{2x2,l}_{kl} \right]$$

where we will make these substitutions

$$\begin{aligned}
\widehat{\delta}^{2x2}_{kU} &= ATT_k(Post) + \Delta Y^0_l(Post, Pre) - \Delta Y^0_U(Post, Pre) \\
\widehat{\delta}^{2x2,k}_{kl} &= ATT_k(Mid) + \Delta Y^0_l(Mid, Pre) - \Delta Y^0_l(Mid, Pre) \\
\widehat{\delta}^{2x2,l}_{lk} &= ATT_l Post(l) + \Delta Y^0_l(Post(l), MID) - \Delta Y^0_k(Post(l), MID) \\
&\quad -(ATT_k(Post) - ATT_k(Mid))
\end{aligned}$$

Notice all those potential sources of biases!

# Potential Outcome Notation

$$p\,lim\,\widehat{\delta}^{TWFE}_{n\to\infty} \quad = \quad VWATT + VWPT - \Delta ATT$$

- Notice the number of assumptions needed *even* to estimate this very strange weighted ATT (which is a function of how you drew the panel in the first place).
- With dynamics, it attenuates the estimate (bias) and can even reverse sign depending on the magnitudes of what is otherwise effects in the sign in a reinforcing direction!
- Let's look at each of these three parts more closely

# Variance weighted ATT

$$
\begin{aligned}
VWATT &= \sum_{k \neq U} \sigma_{kU} ATT_k(Post(k)) \\
&+ \sum_{k \neq U} \sum_{l > k} \sigma_{kl} \left[ \mu_{kl} ATT_k(MID) + (1 - \mu_{kl}) ATT_l(POST(l)) \right]
\end{aligned}
$$

where $\sigma$ is like $s$ only population terms not samples.

- Weights sum to one.
- Note, if all the ATT are identical, then the weighting is irrelevant.
- But otherwise, it's basically weighting each of the individual sets of ATT we have been discussing, where weights depend on group size and variance

# Variance weighted parallel trends

$$
\begin{aligned}
VWPT &= \sum_{k \neq U} \sigma_{kU} \Big[ \Delta Y_k^0(Post(k), Pre) - \Delta Y_U^0(Post(k), Pre) \Big] \\
&+ \sum_{k \neq U} \sum_{l > k} \sigma_{kl} \Big[ \mu_{kl} \{ \Delta Y_k^0(Mid, Pre(k)) - \Delta Y_l^0(Mid, Pre(k)) \} \\
&+ (1 - \mu_{kl}) \{ \Delta Y_l^0(Post(l), Mid) - \Delta Y_k^0(Post(l), Mid) \} \Big]
\end{aligned}
$$

There are $K^2$ parallel trends inside the weights. Their weighted average must equal zero.

# Heterogeneity bias

$$\Delta ATT = \sum_{k \neq U} \sum_{l > k} (1 - \mu_{kl}) \Big[ ATT_k(Post(l) - ATT_k(Mid)) \Big]$$

Now, if the ATT is constant over time, then this difference is zero, but what if the ATT is not constant? Then TWFE is biased, and depending on the dynamics and the VWATT, may even flip signs

JHR
The Journal of Human Resources





# Does Strengthening Self-Defense Law Deter Crime or Escalate Violence?
## Evidence from Expansions to Castle Doctrine

**Cheng Cheng**

**Mark Hoekstra**

## Abstract

From 2000 to 2010, more than 20 states passed so-called "Castle Doctrine" or "stand your ground" laws. These laws expand the legal justification for the use of lethal force in self-defense, thereby lowering the expected cost of using lethal force and increasing the expected cost of committing violent crime. This paper exploits the within-state variation in self-defense law to examine their effect on homicides and violent crime. Results indicate the laws do not deter burglary, robbery, or aggravated assault. In contrast, they lead to a statistically significant 8 percent net increase in the number of reported murders and nonnegligent manslaughters.

# Case study: Castle doctrine reforms

- Cheng and Hoekstra (2013) is a good, clean example of a differential timing for us to practice on
- In 2005, Florida passed a law called Stand Your Ground that expanded self-defense protections beyond the house
- More “castle doctrine” reforms followed from 2006 to 2009

# Description

Details of castle doctrine reforms

- “Duty to retreat” is removed versus castle doctrine reforms; expanded where you can use lethal force
- Presumption of reasonable fear is added
- Civil liability for those acting under the law is removed

# Ambiguous predictions

Castle reforms $\rightarrow$ homicides: Increase by removing homicide penalties and increasing opportunities

- Castle doctrine expansions lowered the (expected) cost of killing someone in self-defense
- Lowering the price of lethal self-defense should increase lethal homicides

Castle reforms $\rightarrow$ homicides: decrease through deterrence

# Cheng and Hoekstra's estimation model

- TWFE model

$$Y_{it} = \beta_1 D_i + \beta_2 T_t + \beta_3 (CDL_{it}) + \alpha_1 X_{it} + c_i + u_t + \varepsilon_{it}$$

- $CDL$ is a fraction between 0 and 1 depending on the percent of the year the state has a castle doctrine law
- Preferred specifications includes "region-by-year fixed effects" (see next slide)
- Estimation with TWFE and Poisson with and without population weights
- Models will include covariates (e.g., police, imprisonment, race shares, state spending on public assistance)

# Publicly available crime data

Main data: FBI Uniform Crime Reports Part 1 Offenses (2000-2010)

- Main outcomes: log homicides
- Falsification outcomes: motor vehicle theft and larceny
- Deterrence outcomes: burglary, robbery, assault

# Region-by-year fixed effects

- **Parallel trends assumption**: imposed structurally with region-by-year dummies
- **Argument**: unobserved changes in crime are running “parallel” to the treatment states within region over time
- **SUTVA** and **No Anticipation**: No spillovers, no hidden variation in treatment, no behavioral change today in response to tomorrow’s law

# Results – Deterrence

| | OLS - Weighted by State Population | | | | | | OLS - Unweighted | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Panel A: Burglary | Log (Burglary Rate) | | | | | | Log (Burglary Rate) | | | | | |
| Castle Doctrine Law | 0.0780*** | 0.0290 | 0.0223 | 0.0164 | 0.0327* | 0.0237 | 0.0572** | 0.00961 | 0.00663 | 0.00277 | 0.00683 | 0.0207 |
| | (0.0255) | (0.0236) | (0.0223) | (0.0247) | (0.0165) | (0.0207) | (0.0272) | (0.0291) | (0.0268) | (0.0304) | (0.0222) | (0.0259) |
| One Year Before Adoption of Castle Doctrine Law | | | | -0.0201 | | | | | | -0.0154 | | |
| | | | | (0.0139) | | | | | | (0.0214) | | |
| Panel B: Robbery | Log (Robbery Rate) | | | | | | Log (Robbery Rate) | | | | | |
| Castle Doctrine Law | 0.0408 | 0.0344 | 0.0262 | 0.0216 | 0.0376** | 0.0515* | 0.0448 | 0.0320 | 0.00839 | 0.00552 | 0.00874 | 0.0267 |
| | (0.0254) | (0.0224) | (0.0229) | (0.0246) | (0.0181) | (0.0274) | (0.0331) | (0.0421) | (0.0387) | (0.0437) | (0.0339) | (0.0299) |
| One Year Before Adoption of Castle Doctrine Law | | | | -0.0156 | | | | | | -0.0115 | | |
| | | | | (0.0167) | | | | | | (0.0283) | | |
| Panel C: Aggravated Assault | Log (Aggravated Assault Rate) | | | | | | Log (Aggravated Assault Rate) | | | | | |
| Castle Doctrine Law | 0.0434 | 0.0397 | 0.0372 | 0.0362 | 0.0424 | 0.0414 | 0.0555 | 0.0698 | 0.0343 | 0.0305 | 0.0341 | 0.0317 |
| | (0.0387) | (0.0407) | (0.0319) | (0.0349) | (0.0291) | (0.0285) | (0.0604) | (0.0630) | (0.0433) | (0.0478) | (0.0405) | (0.0380) |
| One Year Before Adoption of Castle Doctrine Law | | | | -0.00343 | | | | | | -0.0150 | | |
| | | | | (0.0161) | | | | | | (0.0251) | | |
| Observations | 550 | 550 | 550 | 550 | 550 | 550 | 550 | 550 | 550 | 550 | 550 | 550 |
| State and Year Fixed Effects | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Region-by-Year Fixed Effects | | Yes | Yes | Yes | Yes | Yes | | Yes | Yes | Yes | Yes | Yes |
| Time-Varying Controls | | | Yes | Yes | Yes | Yes | | | Yes | Yes | Yes | Yes |
| Contemporaneous Crime Rates | | | | | Yes | | | | | | Yes | |
| State-Specific Linear Time Trends | | | | | | Yes | | | | | | Yes |

# Results – Homicides

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Panel C: Homicide (Negative Binomial - Unweighted) | | | | | | |
| Castle Doctrine Law | 0.0565* | 0.0734** | 0.0879*** | 0.0783** | 0.0937*** | 0.108*** |
| | (0.0331) | (0.0305) | (0.0313) | (0.0355) | (0.0302) | (0.0346) |
| One Year Before Adoption of Castle Doctrine Law | | | | -0.0352 | | |
| | | | | (0.0260) | | |
| Observations | 550 | 550 | 550 | 550 | 550 | 550 |
| Panel D: Log Murder Rate (OLS - Weighted) | | | | | | |
| Castle Doctrine Law | 0.0906** | 0.0955** | 0.0916** | 0.0884** | 0.0981** | 0.0813 |
| | (0.0424) | (0.0389) | (0.0382) | (0.0404) | (0.0391) | (0.0520) |
| One Year Before Adoption of Castle Doctrine Law | | | | -0.0110 | | |
| | | | | (0.0230) | | |
| Observations | 550 | 550 | 550 | 550 | 550 | 550 |
| State and Year Fixed Effects | Yes | Yes | Yes | Yes | Yes | Yes |
| Region-by-Year Fixed Effects | | Yes | Yes | Yes | Yes | Yes |
| Time-Varying Controls | | | Yes | Yes | Yes | Yes |
| Contemporaneous Crime Rates | | | | | Yes | |
| State-Specific Linear Time Trends | | | | | | Yes |

# Interpretation

- Series of robustness checks (falsifications on larceny and motor vehicle theft; deterrence; many different specifications)
- Castle doctrine reforms are associated with an 8% net increase in homicide rates per year across the 21 adopting states
- Interpretation is these would not have occurred without castle doctrine reforms
- But is this robust to alternative models? Today we will check

# Callaway and Sant'Anna 2020

- New papers are coming out focused on the issues that we are seeing with TWFE
- I'll discuss one though by Callaway and Sant'anna (2020) due to time constraints (call it CS)
- If we have time, I'll run through a simulation illustrating both the bias of TWFE and the unbiased estimation of this CS estimator
- Interesting ancestry – CS is a descendent of Abadie (2005) from earlier

# Preliminary

CS considers identification, aggregation, estimation and inference procedures for ATT in DD designs with

1. multiple time periods
2. variation in treatment timing (i.e., differential timing)
3. parallel trends only holds after conditioning on observables

# When might you use this estimator

Probably in the very situations describing your own study

1. When treatment effects heterogenous by time of adoption
2. When treatment effects change over time
3. When shortrun effects more pronounced than longrun effects
4. When treatment effect dynamics differ if people are first treated in a recession relative to expansion years

# Group-time ATT is the parameter of interest in CS

$$ATT(g,t) = E[Y_t^1 - Y_t^0 | G_g = 1]$$

# Group-time ATT

Group-time ATT is the ATT for a specific group and time

- Groups are basically cohorts of units treated at the same time
- CS will calculate an ATT per group/time which will be the sum of all $T - t_k$ for all groups (i.e., a lot)
- Group-time ATT estimates are not determined by the estimation method one adopts (first difference or FE) bc they are simple differences in means
- Does not directly restrict heterogeneity with respect to observed covariates, timing or the evolution of treatment effects over time
- Provides a way to aggregate over these to get a single ATT
- Inference is the bootstrap

# Notation

- $T$ periods going from $t = 1, \ldots, T$
- Units are either treated ($D_t = 1$) or untreated ($D_t = 0$) but once treated cannot revert to untreated state
- $G_g$ signifies a group and is binary. Equals one if individual units are treated at time period $t$.
- $C$ is also binary and indicates a control group unit equalling one if "never treated" (can be relaxed though to "not yet treated")
  - $\rightarrow$ Recall the problem with TWFE on using treatment units as controls
- Generalized propensity score enters into the estimator as a weight:

$$\widehat{p(X)} = Pr(G_g = 1 | X, G_c + C = 1)$$

# Assumptions

Assumption 1: Sampling is iid (panel data)

Assumption 2: Conditional parallel trends (for either never treated or not yet treated)

$$E[Y_t^0 - Y_{t-1}^0 | X, G_g = 1] = [Y_t^0 - Y_{t-1}^0 | X, C = 1]$$

Assumption 3: Irreversible treatment

Assumption 4: Common support (propensity score)

Assumption 5: Limited treatment anticipation (i.e., treatment effects are zero pre-treatment)

# CS Estimator (the IPW version)

$$ATT(g,t) = E\left[\left(\frac{G_g}{E[G_g]} - \frac{\frac{\hat{p}(X)C}{1-\hat{p}(X)}}{E\left[\frac{\hat{p}(X)C}{1-\hat{p}(X)}\right]}\right)(Y_t - Y_{g-1})\right]$$

This is the inverse probability weighting estimator. Alternatively, there is an outcome regression approach and a doubly robust. Sant'Anna recommends DR. Notice hw CS doesn't use already-treated as controls.

# Staggered adoption (i.e., universal coverage)

**Proof.**

**Remark 1:** In some applications, eventually all units are treated, implying that $C$ is never equal to one. In such cases one can consider the "not yet treated" ($D_t = 0$) as a control group instead of the "never treated?" ($C = 1$). □

# Aggregated vs single year/group ATT

- The method they propose is really just identifying very narrow ATT per group time.
- But we are often interested in more aggregate parameters, like the ATT across all groups and all times
- They present two alternative methods for building “interesting parameters”
- Inference from a bootstrap

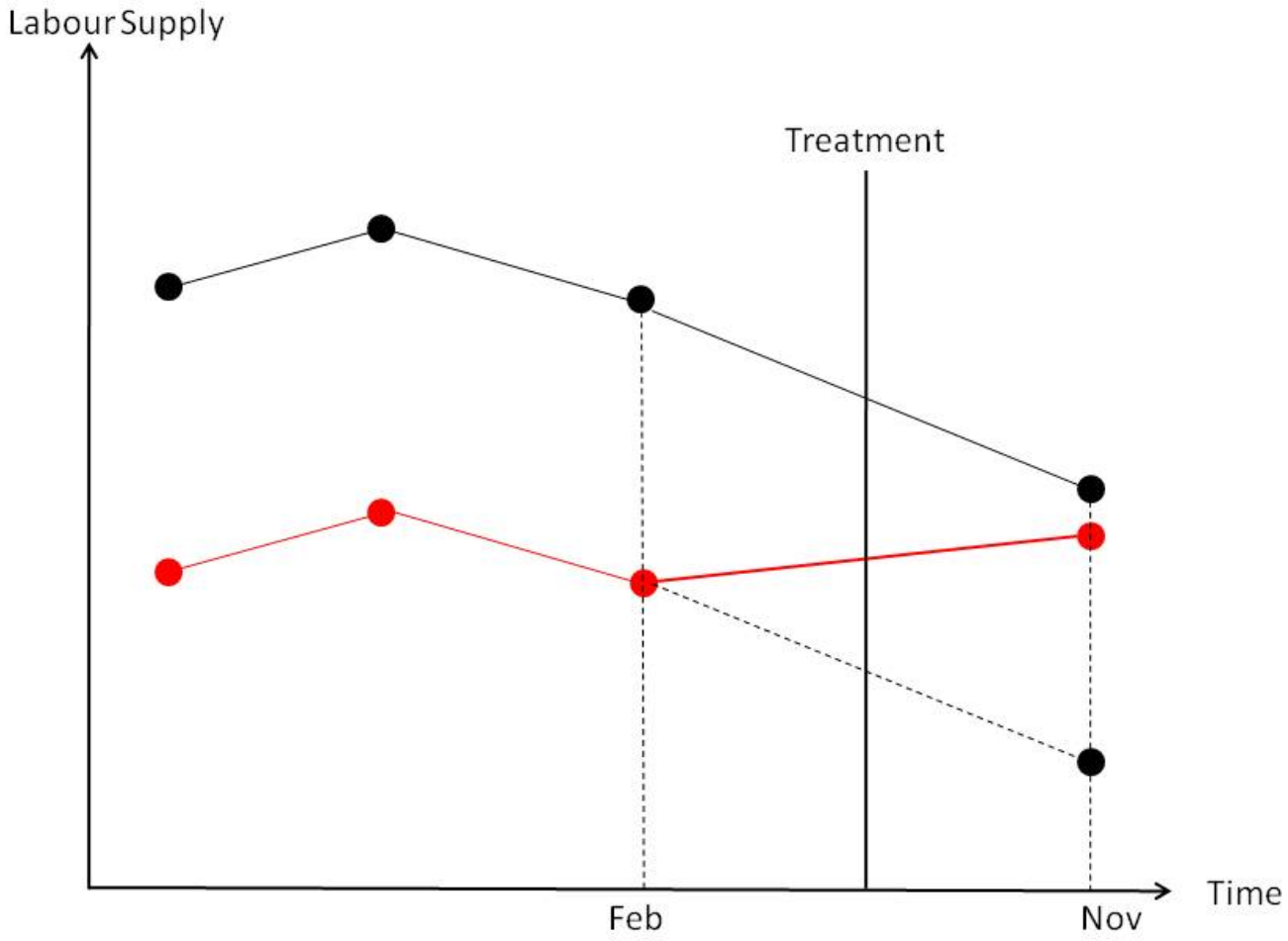
# Stata simulation

Let's now review a simulation in Stata which can be downloaded from my github repo called **`baker.do`**.

# Pre-trends

- The identifying assumption for all DD designs is parallel trends
- Parallel trends cannot be directly verified because technically one of the parallel trends is an unobserved counterfactual
- But one often will check a hunch for parallel trends using pre-trends
- But, even if pre-trends are the same one still has to worry about other policies changing at the same time (omitted variable bias)

# Plot the raw data when there's only two groups



# Event study regression

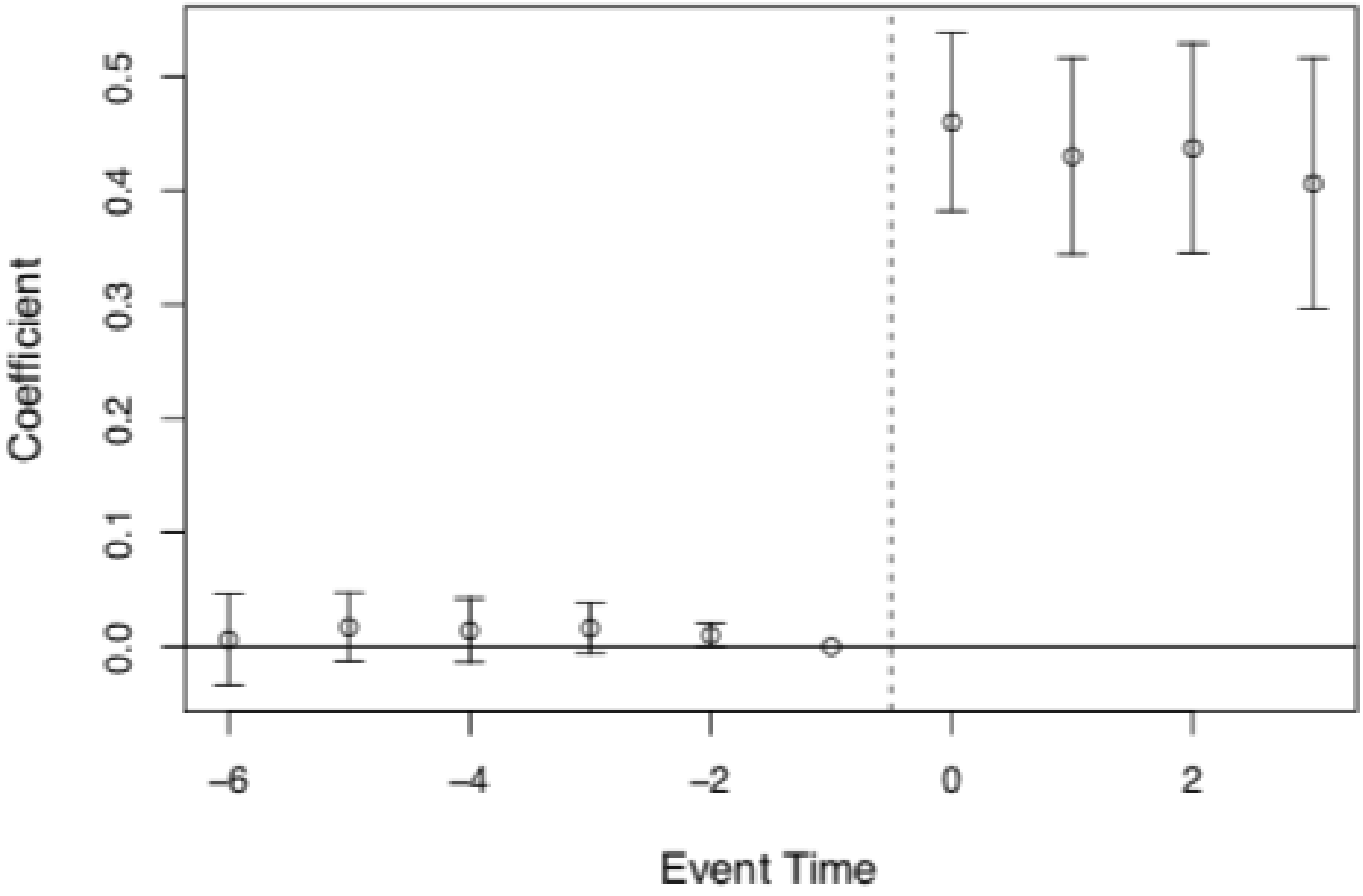
- Including leads into the DD model is an easy way to analyze pre-treatment trends
- Lags can be included to analyze whether the treatment effect changes over time after assignment
- The estimated regression would be:

$$Y_{its} = \gamma_s + \lambda_t + \sum_{\tau=-2}^{-q} \gamma_\tau D_{s\tau} + \sum_{\tau=0}^{m} \delta_\tau D_{s\tau} + \varepsilon_{ist}$$

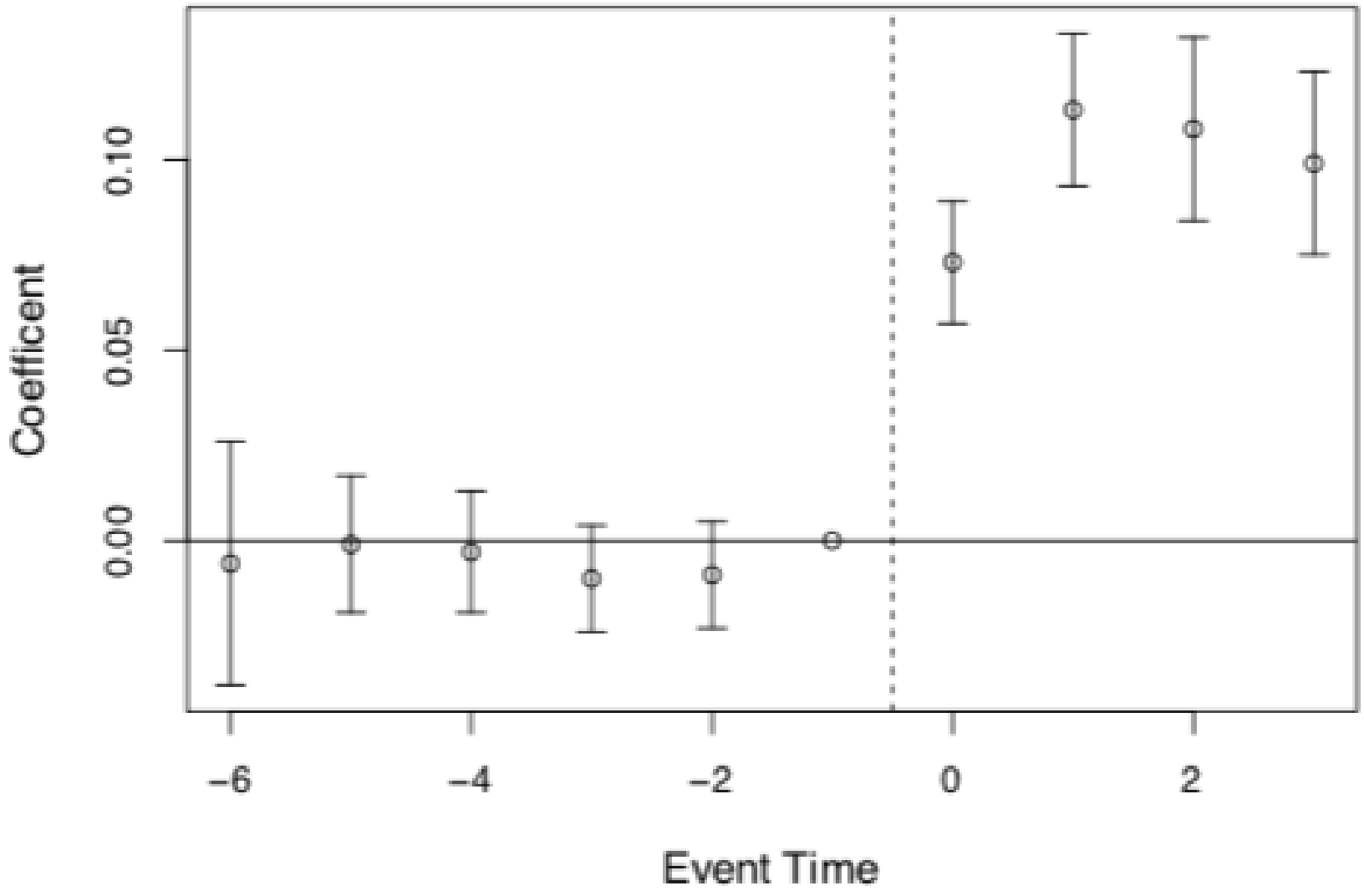
→ Treatment occurs in year 0
→ Includes $q$ leads or anticipatory effects
→ Includes $m$ leads or post treatment effects

# Medicaid and Affordable Care Act example

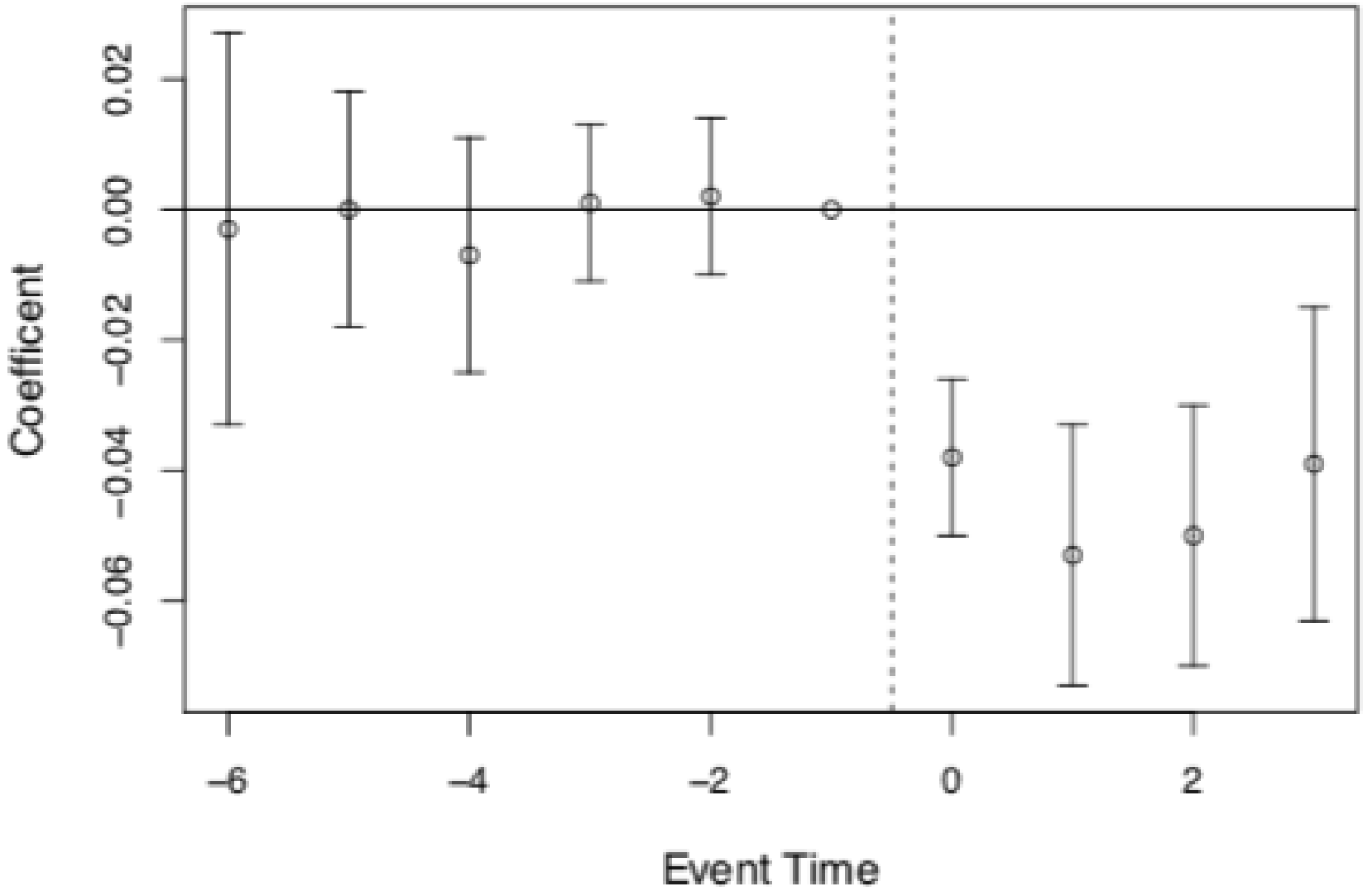
- Miller, et al. (2019) examine a rollout of Medicaid under the Affordable Care Act
- They link large-scale survey data with administrative death records
- 9.3 reduction in annual mortality caused by Medicaid expansion
- Driven by a reduction in disease-related deaths which grows over time




(a) Medicaid Eligibility




(b) Medicaid Coverage




(c) Uninsured

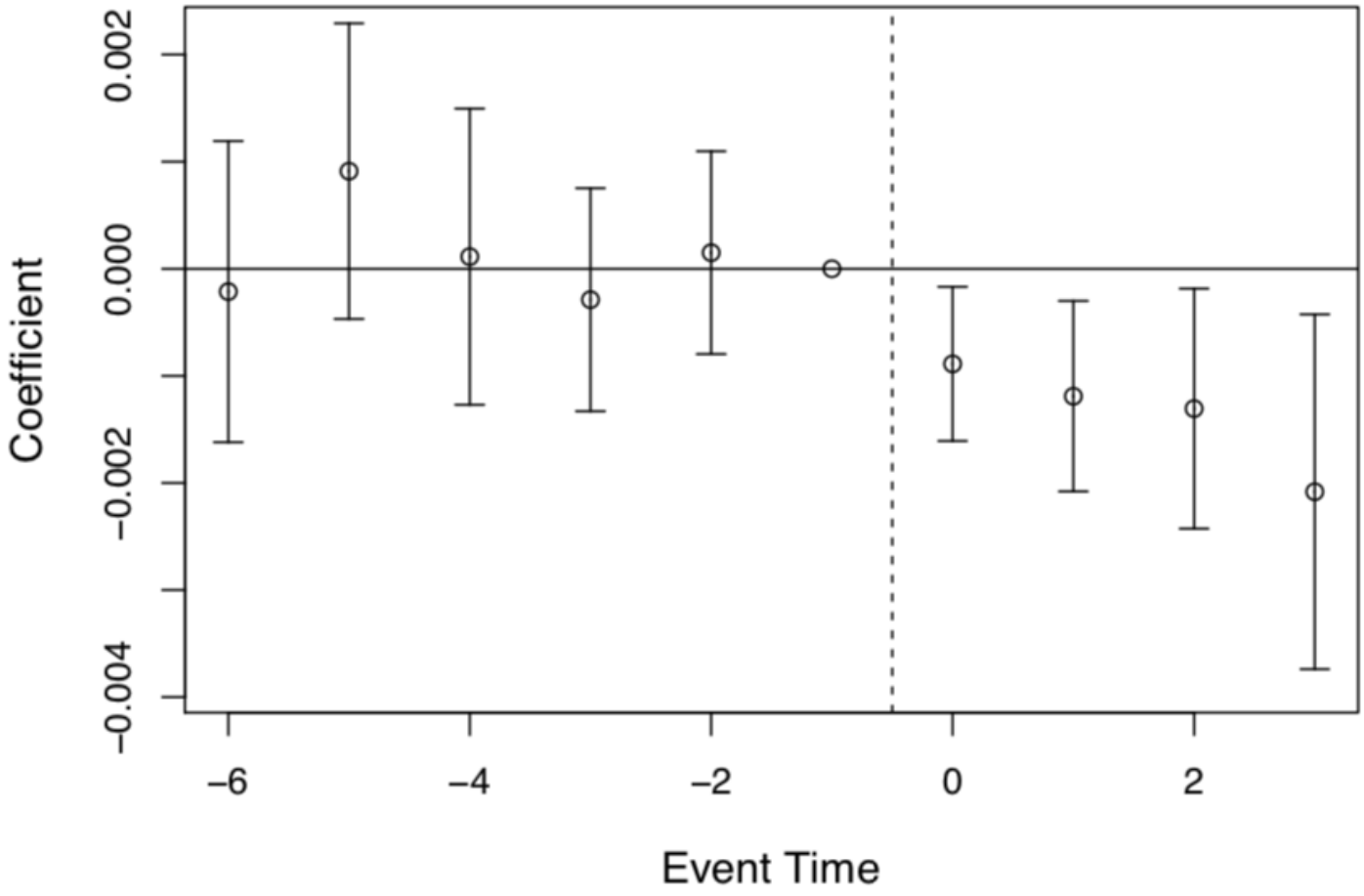



*Figure:* Miller, et al. (2019) estimates of Medicaid expansion's effects on on annual mortality

# Differential timing complicates plotting sample averages

- New Jersey treated in late 1992, New York in late 1993, Pennsylvania never treated
- Pre-treatment:
  - → New Jersey: $<1992$
  - → New York: $<1993$
  - → Pennsylvania: undefined
- So how do we check parallel leads?

# Early efforts at event studies

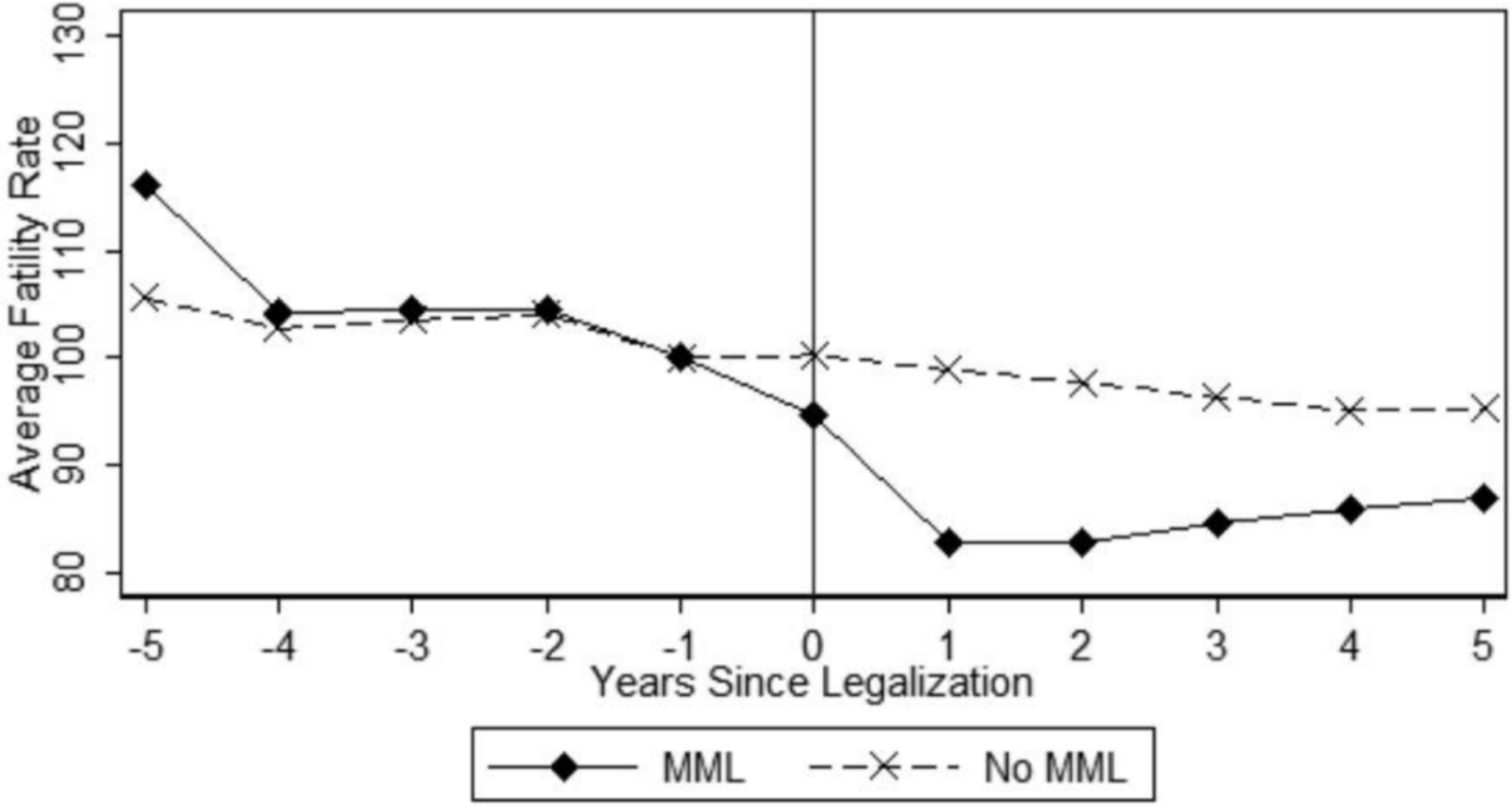



*Figure:* Anderson, et al. (2013) display of raw traffic fatality rates for re-centered treatment states and control states with randomized treatment dates

# Randomized control counties to receive arbitrary dates as treatment can be misleading

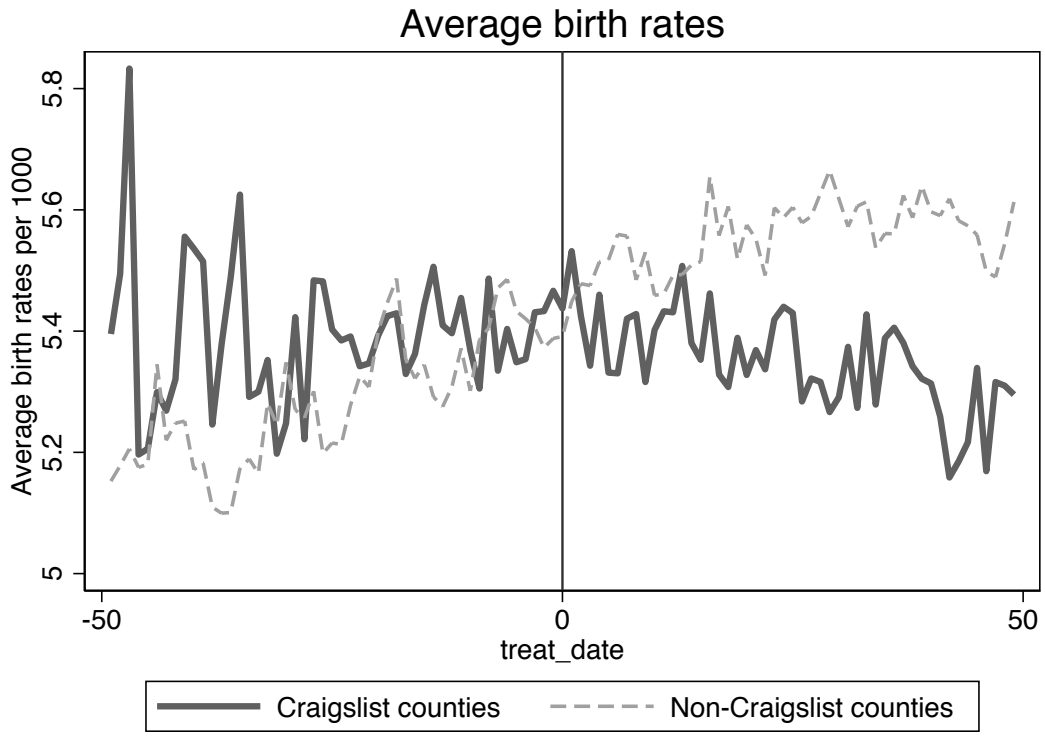
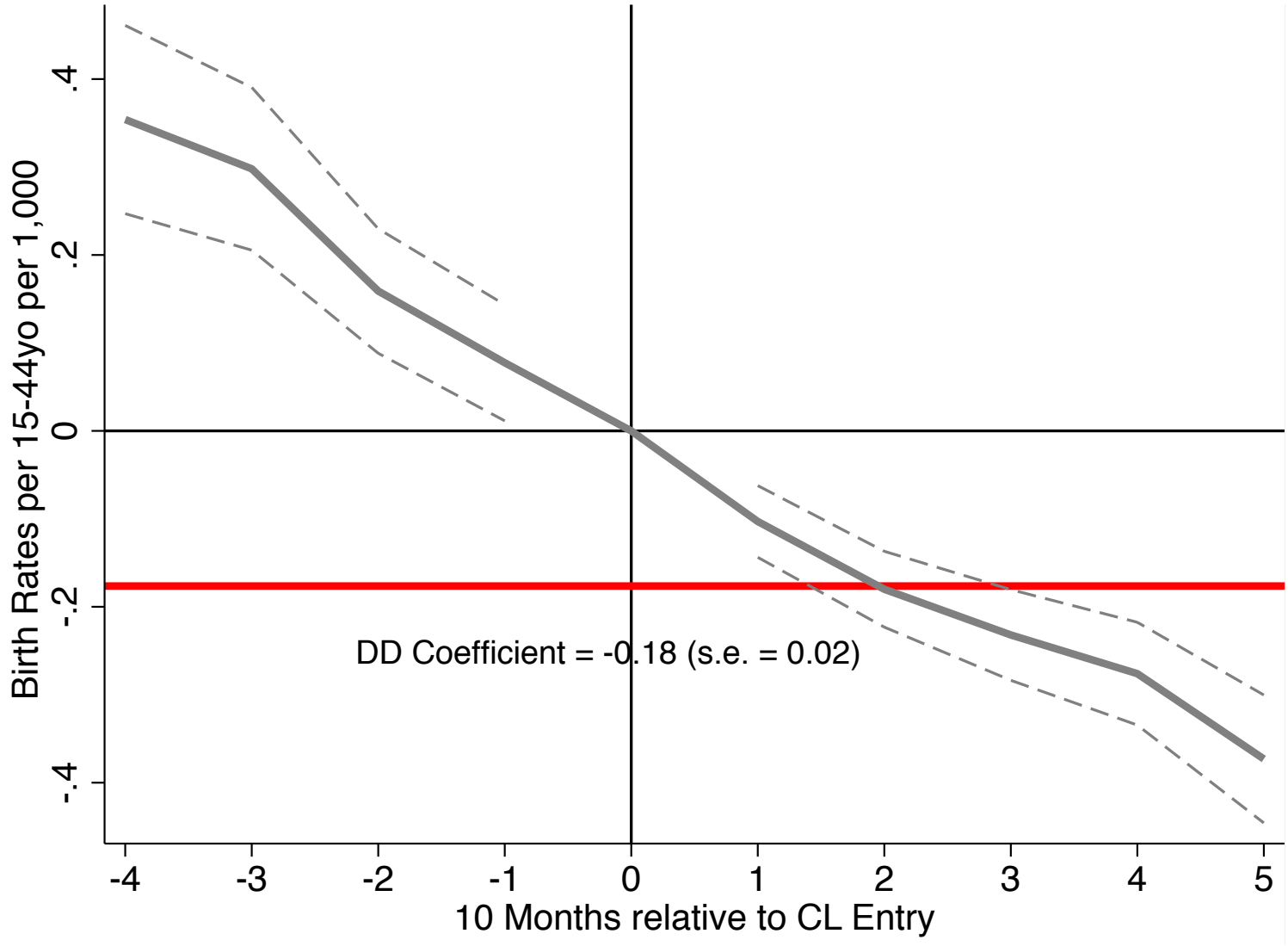



*Figure:* From one of my studies. Looks decent right?




Same data as a couple slides ago, leads don't look good

# Sun and Abraham 2020

- Recall our discussion of event studies estimated with TWFE under differential timing
- Now that we know about the biases of TWFE when estimating aggregate DD parameters, let's revisit event studies under differential timing
- Callaway and Sant'Anna (2020) propose alternative estimators for event studies that estimate group-time ATT in relative event time
- But now we will discuss Sun and Abraham (2020) [SA] which is like a blend of Goodman-Bacon's decomposition and Callaway and Sant'anna alternative estimator to TWFE

# Summarizing

- Goodman-Bacon (2021, forthcoming) focused on decomposition of TWFE to show bias under differential timing
- Callaway and Sant'anna (2020) presents alternative estimator that yields unbiased estimates of group-time ATTs which can be aggregated or put into event study plots
- Sun and Abraham (SA) is like a combination of the two papers

# Summarizing (cont.)

1. SA is a decomposition of the population regression coefficient on event study leads and lags with differential timing estimated with TWFE
2. They show that the population regression coefficient is "contaminated" by information from other leads and lags
3. SA presents an alternative estimator that is not so dissimilar to CS

# Summarizing (cont.)

- Problems seem to occur with DD when we introduce treatment effect heterogeneity
- Under treatment effect heterogeneity, spurious non-zero positive lead coefficients even when there is no pretrend
- This problem is exacerbated by the TWFE related weights as under some scenarios, the weights sum to zero and“cancel out” the treatment effects from other periods
- They present a 3-step TWFE based alternative estimator which addresses the problems that they find

# Summarizing (cont.)

- Only decomposition of TWFE estimating dynamic leads and lags (Goodman-Bacon focused on a “static” specification)
- Contamination of coefficients on leads and lags by treatment effects depends on the magnitude of the weights on the true group-time ATT, or “cohort-specific ATT”
- Weights are a function of cohort composition
- Examining weights lets you gauge how treatment effect heterogeneity would interact with potential non-zero and non-convex weighting in population regression coefficients on the leads and lags

# Difficult notation sadly

- When treatment occurs at the same time, we say they are part of the same cohort, $e$
- If we bin the data, then a lead or lag $l$ will appear in the bin $g$ so sometimes they use $g$ instead of $l$ or $l \in g$
- Building block is the "cohort-specific ATT" or $CATT_{e,l}$ – same thing as CS group-time ATT
- Estimate $CATT_{e,l}$ with population regression coefficient $\mu_l$

# Difficult notation (cont.)

- At each time $t$ there are two possible treatment status $D_{i,t} \in \{0, 1\}$ over $T + 1$ time periods
- Path of treatment status scales exponentially with $T$ and an take on $2^{T+1}$ possible values
- They focus on irreversible treatment where treatment status is non-decreasing sequence of zeroes and ones

# Difficult notation (cont.)

- If a group is never treated, the $\infty$ symbol is used to either describe the group ($E_i = \infty$) or the potential outcome ($Y^{\infty}$)
- $Y_{i,t}^{\infty}$ is is the potential outcome for unit $i$ if it had never received treatment (versus received it later), also called the baseline outcome
- Other counterfactuals are possible – maybe unit $i$ isn't "never treated" but treated later in counterfactual

# More difficult notation (cont.)

- Treatment effects are the difference between the observed outcome relative to the never-treated counterfactual outcome: $Y_{i,t} - Y_{i,t}^{\infty}$
- We can take the average of treatment effects at a given relative time period across units first treated at time $E_i = e$ (same cohort) which is what we mean by $CATT_{e,l}$
- Doesn't use $t$ index time ("calendar time"), rather uses $l$ which is time until or time after treatment date $e$ ("relative time")
- Think of it as $l =$year - treatment date

# Definition 1

**Definition 1:** The cohort-specific ATT $l$ periods from initial treatment date $e$ is:

$$CATT_{e,l} = E[Y_{i,e+l} - Y^{\infty}_{i,e+l} | E_i = e]$$

# Identifying assumption 1

**Assumption 1: Parallel trends in baseline outcomes**:
$E[Y_{i,t}^{\infty} - Y_{i,s}^{\infty} | E_i = e]$ is the same for all $e \in supp(E_i)$ and for all $s$, $t$ and is equal to $E[Y_{i,t}^{\infty} - Y_{i,s}^{\infty}]$

Interesting SA comment: Never-treated units are likely to differ from ever-treated units in many ways; think of a Roy model. What does it imply that they chose not to get treated? It may imply net negative treatment effects and that could mean they may not share the same evolution of baseline outcomes as the treatment groups. If you think they are unlikely to satisfy this assumption, then drop them. Almost like a synthetic control approach.

# Assumption 2

**Assumption 2: No anticipator behavior in pre-treatment periods**:
There is a set of pre-treatment periods such that
$E[Y^{e}_{i,e+l} - Y^{\infty}_{i,e+l} | E_i = e] = 0$ for all possible leads.

Basically means that potential outcomes prior to treatment at baseline by on average the same. This means there is no pre-trends, essentially. This is most plausible if the full treatment paths are not known to the units (e.g., Craigslist opening erotic services without announcement)

# Assumption 3

**Assumption 3: Treatment effect homogeneity**: For each relative time period $l$, the $CATT_{e,l}$ doesn't depend on the cohort and is equal to $CATT_l$.

Assumption 3 requires each cohort experience the same path of treatment effects. Treatment effects need to be the same across cohorts in every relative period for homogeneity to hold, whereas for heterogeneity to occur, treatment effects just need to differ across cohorts in one relative time period. Doesn't preclude dynamic treatment effects, though. It just imposes that cohorts share the same treatment path.

# Treatment effect heterogeneity

- Assumption 3 is violated when different cohorts experience different paths of treatment effects
- Cohorts may differ in their covariates which affect how they respond to treatment (e.g., if treatment effects vary with age, and there is variation in age across units first treated at different times, then there will be heterogeneous treatment effects)
- Doesn’t rule out parallel trends

# TWFE Regression

$$Y_{i,t} = \alpha_i + \delta_t + \sum_{g \in G} \mu_g 1\{t - E_i \in g\} + \varepsilon_{i,t}$$

They say $E_i$ is the initial time of a binary variable absorbing treatment for unit $i$. Fixed effects should be obvious. $\mu_g$ is the population regression coefficient on the leads and lags that we want to estimate. We estimate this using OLS and get $\widehat{\mu_g}$.

We are interested in the properties of $\mu_g$ under differential timing as well as whether there are any never-treated units

# Specifying the leads and lags

How will we specify the $1\{t - E_i \in g\}$ term? SA considers a couple:

1. Static specification:

$$Y_{i,t} = \alpha_i + \delta_t + \mu_g \sum_{l \geq 0} D_{i,t}^l + \varepsilon_{i,t}$$

2. Dynamic specification:

$$Y_{i,t} = \alpha_i + \delta_t + \sum_{l=-K}^{-2} \mu_l D_{i,t}^l + \sum_{l=0}^{L} \mu_l D_{i,t}^l + \varepsilon_{i,t}$$

# Multicollinearity

Dynamic specification requires deciding which leads to drop. They recommend dropping two: $l = -1$ and some other one (they seem to favor $l = -4$). The reason is twofold. You drop one of them to avoid multicollinearity in the relative time indicators. You drop a second one because of the multicollinearity coming from the linear relationship between TWFE and the relative period indicators.

# Trimming and binning

- First some terms: trimming and binning, I do both in the Mixtape when analyzing Cheng and Hoekstra (2013)
- Binning means placing all “distant” relative time indicators into a single one. Done because of the sparseness of units in such distant bins. So if there’s 3 distant leads and lags that aren’t balanced, combine them all into the last lead and lag
- Trimming means excluding any relative period for which you don’t have balance in relative time. This creates a balanced panel “in relative time”, but imbalanced panel length overall.
- They’ll analyze both and how they affect $\widehat{\mu_g}$ estimation using TWFE

## Interpreting $\widehat{\mu_g}$ under no to all assumptions

**Proposition 1 (no assumptions):** The population regression coefficient on relative period bin $g$ is a linear combination of differences in trends from its own relative period $l \in g$, from relative periods $l \in g'$ of other bins $g' \neq g$, and from relative periods excluded from the specification (e.g., trimming).

$$
\begin{aligned}
\mu_g \quad = \quad & \underbrace{\sum_{l \in g} \sum_{e} w^g_{e,l} (E[Y_{i,e+l} - Y^{\infty}_{i,0} | E_i = e] - E[Y^{\infty}_{i,e+l} - Y^{\infty}_{i,0}])}_{\text{Good stuff}} \\
+ \quad & \underbrace{\sum_{g' \neq g} \sum_{l \in g'} \sum_{e} w^g_{e,l} (E[Y_{i,e+l} - Y^{\infty}_{i,0} | E_i = e] - E[Y^{\infty}_{i,e+l} - Y^{\infty}_{i,0}])}_{\text{Bleh – Other included relative time}} \\
+ \quad & \underbrace{\sum_{l \in g^{excl}} \sum_{e} w^g_{e,l} (E[Y_{i,e+l} - Y^{\infty}_{i,0} | E_i = e] - E[Y^{\infty}_{i,e+l} - Y^{\infty}_{i,0}])}_{\text{More bleh – Excluded}}
\end{aligned}
$$

Superscript $g$ associates the weight with coefficient $\mu_g$. The weight associated with cohort $e$ in relative period $l$ is equal to the population regression coefficient on the $1\{t - E_i \in g\}$ from regression $D^l_{i,t} \times 1\{E_i = e\}$ on all bin indicators included in the regression and TWFE. Just the mechanics of double demeaning from TWFE

# Weight ($w^g_{e,l}$) summation cheat sheet

1. For relative periods of $\mu_g$ own $l \in g$, $\sum_{l \in g} \sum_e w^g_{e,l} = 1$
2. For relative periods belonging to some other bin $l \in g'$ and $g' \neq g$, t $\sum_{l \in g'} \sum_e w^g_{e,l} = 0$
3. For relative periods not included in $G$, $\sum_{l \in g^{excl}} \sum_e w^g_{e,l} = -1$

# Estimating the weights

Regress $D_{i,t}^{l} \times 1\{E_i = e\}$ on:

1. all bin indicators included in the main TWFE regression,
2. $\{1\{t - E_i \in g\}\}_{g \in G}$(i.e., leads and lags) and
3. the unit and time fixed effects

# Interpretation of coefficients under parallel trends only

**Proposition 2**: Under the parallel trends only, the population regression coefficient on the indicator for relative period bing $g$ is a linear combination of $CATT_{e,l\in g}$ as well as $CATT_{d,l'}$ from other relative periods $l' \notin g$ with the same weights stated in Proposition 1:

$$\begin{aligned}\mu_g \quad = \quad & \underbrace{\sum_{l\in g}\sum_{e} w^g_{e,l} CATT_{e,l}}_{\text{Desirable}} \\ & + \underbrace{\sum_{g'\neq g, g'\in G}\sum_{l'\in g'}\sum_{e} w^g_{e,l'} CATT_{e,l'}}_{\text{Undesirable – other specified bins}} \\ & + \underbrace{\sum_{l'\in g^{excl}}\sum_{e} w^g_{e,l'} CATT_{e,l'}}_{\text{Undesirable – excluded relative time indicators}}\end{aligned}$$

# Comment on Proposition 2

The coefficient $\mu_g$ can be written as an average of $CATT_{e,l}$ from own periods but also $CATT_{e,l'}$ from other periods.

The weights are still functions of cohort comparisons, like in Proposition 1, which means $\mu_g$ can be written as non-convex averages of not only $CATT_{e,l}$ from own periods $l \in g$, but also $CATT_{e,l'}$ from other periods.

Means $\mu_g$ could in fact be the wrong sign to all $CATT_{e,l \in g}$.

Weights can help us gauge the severity of this problem.

When the weights have larger magnitude, treatment effect heterogeneity matters more as a particular $CATT_{e,l}$ can drive the overall estimates. But when weights are uniform, treatment effect heterogeneity matters less.

# Interpretation under parallel trends and no anticipation

**Proposition 3**: If parallel trends holds and no anticipation holds for all $l < 0$ (i.e., no anticipatory behavior pre-treatment), then the population regression coefficient $\mu_g$ for $g$ is a linear combination of post-treatment $CATT_{e,l'}$ for all $l' \geq 0$.

$$\begin{aligned}
\mu_g \quad = \quad & \sum_{l' \in g, l' \geq 0} \sum_{e} w^g_{e,l'} CATT_{e,l'} \\
& + \sum_{g' \neq g, g' \in G} \sum_{l' \in g', l' \geq 0} \sum_{e} w^g_{e,l'} CATT_{e,l'} \\
& + \sum_{l' \in g^{excl}, l' \geq 0} \sum_{e} w^g_{w,l'} CATT_{e,l'}
\end{aligned}$$

# Proposition 3 comment

Notice how once we impose zero pre-treatment treatment effects, those terms are gone (i.e., no $l \in g, l < 0$). But the second term remains unless we impose treatment effect homogeneity (homogeneity causes terms due to weights summing to zero to cancel out). Thus $\mu_g$ may be non-zero for pre-treatment periods *even though parallel trends hold in the pre period.*

# Proposition 4

**Proposition 4**: If parallel trends and treatment effect homogeneity, then $CATT_{e,l} = ATT_l$ is constant across $e$ for a given $l$, and the population regression coefficient $\mu_g$ is equal to a linear combination of $ATT_{l \in g}$, as well as $ATT_{l' \notin g}$ from other relative periods

$$\begin{aligned} \mu_g &= \sum_{l \in g} w_l^g ATT_l \\ &\quad + \sum_{g' \neq g} \sum_{l' \in g'} w_{l'}^g ATT_{l'} \\ &\quad + \sum_{l' \in g^{excl}} w_{l'}^g ATT_{l'} \end{aligned}$$

# Proposition 4 comment

The weight $w_l^g = \sum_e w_{e,l}^g$ sums over the weights $w_{e,l}^g$ from Proposition 1 and is equal to the population regression coefficient from the following auxiliary regression:

$$D_{i,t}^l = \alpha_i + \lambda_t + \sum_{g \in G} w_l^g \cdot 1\{t - E_i \in g\} + u_{i,t}$$

which regresses $D_{i,t}^l$ on all bin indicators and TWFE

# On binning

- Many propose either binning or trimming to create “balanced” panels (in relative event time)
- But SA notes that binning in simulations creates uninterpretable weights (due to the binned $CATT_{e,l'}$ inclusion in $\mu_g$), whereas trimming creates weights that are more reasonable
- This may be because trimming subtracts the corresponding $CATT_{e,l'}$ from $\mu$ regression coefficient

# Intuition for contamination

- Stupid notation make Hulk smash!
- Let's do a simple toy example instead

Balanced panel $T = 2$ with cohorts $E_i \in \{1, 2\}$. We drop two relative time periods to avoid multicollinearity, so we will include bins $\{-2, 0\}$ and drop $\{-1, 1\}$.

# Toy example

$$\mu_{-2} = \underbrace{CATT_{2,-2}}_{\text{own period}} + \underbrace{\frac{1}{2}CATT_{1,0} - \frac{1}{2}CATT_{2,0}}_{\text{other included bins}}$$
$$+ \underbrace{\frac{1}{2}CATT_{1,1} - CATT_{1,-1} - \frac{1}{2}CATT_{2,-1}}_{\text{Excluded bins}}$$

- Parallel trends gets us to all of the $CATT$
- No anticipation makes $CATT = 0$ for all $l < 0$ (all $l < 0$ cancel out)
- Homogeneity cancels second and third terms
- Still leaves $\frac{1}{2}CATT_{1,1}$ – you chose to exclude a group with a treatment effect

Lesson: drop the relative time indicators on the left, not things on the right, bc lagged effects will contaminate through the excluded bins

# Interaction-weighted estimator

- They propose an interacted weighted estimator (IW) as a consistent estimator for $\mu_g$
- Estimator uses either never-treated as controls or “last cohort treated” if no never-treated (contra CS which uses “not yet treated”)
- No covariates bc this is a regression with fixed effects and time-varying covariates create own biases, although they note you can plug in CS for the DD calculation and recover $CATT$ that way
- The interaction is a TWFE regression specification that interacts relative period indicators with cohort/group indicators, excluding indicators for never-treated cohorts

# Interaction-weighted estimator

- **Step one**: Do this DD regression and hold on to $\widehat{\delta}_{e,l}$

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{e \notin C} \sum_{l \neq -1} \delta_{e,l} (1\{E_i = e\} \cdot D_{i,t}^l) + \varepsilon_{i,t}$$

Can use never-treated or last-treated cohort. Drop always treated. The $\delta_{e,l}$ is a DD estimator for $CATT_{e,l}$ with particular choices for pre-period and cohort controls

# Interaction-weighted estimator

- **Step two**: Estimate weights using sample shares of each cohort in the relevant periods:

$$Pr(E_i = e | E_i \in [-l, T - l])$$

# IW estimator

- **Step three**: Take a weighted average of estimates for $CATT_{e,l}$ from Step 1 with weight estimates from step 2

$$\widehat{v}_g = \frac{1}{|g|} \sum_{l \in g} \sum_{e} \widehat{\delta}_{e,l} \widehat{Pr}\{E_i = e | E_i \in [-l, T-l]\}$$

# Consistency and Inference

- Under parallel trends and no anticipation, $\widehat{\delta}_{e,l}$ is consistent, and sample shares are also consistent estimators for population shares.
- Thus IV estimator is consistent for a weighted average of $CATT_{e,l}$ with weights equal to the share of each cohort in the relevant period(s).
- They show that each IW estimator is asymptotically normal and derive its asymptotic variance. Doesn't rely on bootstrap like CS.

# DD Estimator of CATT

**Definition 2**: DD estimator with pre-period $s$ and control cohorts $C$ estimates $CATT_{e,l}$ as:

$$\widehat{\delta_{e,l}} = \frac{E_N[(Y_{i,e+l} - Y_{i,s}) \times 1\{E_i = e\}]}{E_N[1\{E_i = e\}]} - \frac{E_N[(Y_{i,e+l} \times 1\{E_i \in C\}]}{E_N[1\{E_i \in C\}]}$$

**Proposition 5**: If parallel trends and no anticipation both hold for all pre-periods, then the DD estimator using any pre-period and non-empty control cohorts (never-treated or not-yet-treated) is an unbiased estimate for $CATT_{e,l}$

# Software

- **Stata**: eventstudyinteract (can be installed from ssc)
- **R**: fixest with subab() option (see `https://lrberge.github.io/fixest/reference/sunab.html/`)

# Conclusion of SA

- Bacon shows the TWFE coefficient on the static parameter is “contaminated” by other periods leads and lags
- Three strong assumptions needed for TWFE to be unbiased: parallel trends, no anticipation, and treatment homogeneity
- SA doesn’t restrict to treatment profile homogeneity, though; very similar to CS
- SA is a three step interaction-weighted estimator
- Callaway and Sant’Anna (2020) uses not-yet-treated or never treated, but Sun and Abraham (2020) uses last-treated or never treated

# de Chaisemartin and D'Haultfoeulle 2020

de Chaisemartin and D'Haultfouelle 2020 (dCdH) is different from the other papers in several ways

- Like SA, it's a diagnosis and a cure
- TWFE decomposition shows coefficient a weighted average of underlying treatment effects, but weights can be negative negating causal interpretation
- Propose a solution for both static and dynamic specification which does not use already treated as controls
- Treatment can turn on and off

# Comment on Bacon

- Recall the Bacon decomposition – TWFE coefficients are decomposed into weighted average of all underlying 2x2s. Weights were non-negative and summed to one.
- But this decomposition was more a numerical decomposition – what exactly adds up to equal the TWFE coefficient using the data we observe?
- Bacon's decomposition is not "theoretical" – not in the way that other decompositions are. He is just explaining what OLS "does" when it calculates $\widehat{\delta}$
- Just explains what comparisons OLS is using to calculate the TWFE coefficient – just peels back the curtain.

# Causal effects

- dCdH impose causal assumptions and try a different decomposition strategy
- Uses as its building block the unit-specific treatment effects
- This is hopefully going to help us better understand where these negative weights are coming from
- Note that their model is very general in that the treatment is reversible (meaning you can turn it on and off)

# Terms

- Target parameter:

$$\Delta_{i,t}^{g} = Y_{i,t}^{1} - Y_{i,t}^{\infty}$$

  but where the treatment is in time period $g$. Notice –it's not the ATT (it's $i$ individual treatment effect)

- TWFE terms. Define the error term as $\varepsilon_{i,t}$:

$$D_{i,t} = \alpha_i + \alpha_t + \varepsilon_{i,t}$$

- Weights:

$$w_{i,t} = \frac{\varepsilon_{i,t}}{\frac{1}{N^T}\sum_{i,t:D_{i,t}=1}\varepsilon_{i,t}}$$

  Basically divide the error by the average of the error for all treated units.

# Assumptions

## Strong unconditional PT

Assume that for every time period $t$ and every group $g, g'$,

$$E[Y_t^\infty - Y_{t-1}^\infty | G = g] = E[Y_t^\infty - Y_{t-1}^\infty | G = g']$$

Assume parallel trends for every unit in every cohort in every time period.

# dCdH Theorem

## Theorem – dCdH decomposition

Assuming SUTVA, no anticipation and the strong PT, then let $\beta$ be the TWFE estimand associated with

$$Y_{i,t} = \alpha_i + \alpha_t + \beta D_{i,t} + \varepsilon_{i,t}$$

Then it follows that

$$\beta = E\left[ \sum_{i,t:D_{i,t}=1} \frac{1}{N^T} w_{i,t} \cdot \Delta^g_{i,t} \right]$$

where $\sum_{i,t:D_{i,t}=1} \frac{w_{i,t}}{N^T} = 1$ but $w_{i,t}$ can be negative

So once you run that specification, $\beta$ is going to recover a non-convex average over all unit level treatment effects (weights can be negative). dCdH was the first I think.

# Negative weights

- Very common now to hear about negative weights, and furthermore, that negative weights wipe out any causal interpretation, but why?
- What if every unit treatment effect was positive, but some of the weights were negative?
- It's possible it could flip the sign, but it would definitely at least pull the estimate away from the true effect
- This is dangerous – and it's caused by the forbidden contrasts (treated to already treated)

# Negative weights

- Doesn’t always pose a problem, but no proofs for this intuition known yet
- A large number of never-treated seems to make this less an issue
- Shrinking the spacing between treatment dates also can drive it down
- But does that mean that TWFE works, and what does it mean to work?
- TWFE still even when all the weights are positive the weighted average may not aggregate to what we think it does

# Weighting

- The weights in OLS all come out of the model itself, *not the economic question*
- The economic question is “what parameter do you want? What does it look like? Who is in it?”
- And when you define the parameter up front, you’ve more or less defined the economic question you’re asking
- But OLS sort of ignores your question and just gives you what it wants

# Weighting

- What makes something a good vs a bad weight?
- Not being negative is the absolute minimal requirement
- But it’s also not a good sign if you can’t really explain the weights

# dCdH Solution

- dCdH propose an alternative that doesn't have the problems of TWFE – both avoiding negative weights and improving interpretability
- Recall, their model can handle reversible treatments