Recent Advances in DID methods
Overview

Manuel Bagues

University of Warwick

Credits

Links to relevant material

The material used in this presentation comes from:

- Silvia Vannutelli
- Andrew Baker
- Callaway and Sant'Anna
- Scott Cunningham Carolina Kansikas

The following video presentations might be useful:

- Goodman-Bacon
- Callaway and Sant'Anna

Recent review of the literature :

- Jonathan Roth, Pedro H. C. Sant'Anna, Alyssa Bilinski, John Poe (2022), 'What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature'.

Roadmap

- Canonical model
- Recent advances in three dimensions
  1. allowing for multiple periods and variation in treatment timing
  2. consider potential violations of parallel trends
  3. depart from the assumption of observing a sample of many independent clusters sampled from a super-population

Allowing for multiple periods and variation in treatment timing

- Differences-in-differences designs are widely used in empirical research.
- 2X2 comparisons (treatment, control, before, after) are well understood
- With staggered treatment implementation, two-way fixed effects estimators are commonly used.
- The interpretation of treatment effects for staggered DiD is not straightforward in these cases.
    - Weights + Bias.
- Recent papers propose solutions to diagnose and correct these issues.

Agenda for today

1. Problems with TWFE for staggered DiD designs: Weights + Dynamics
   - Illustration of the comparison and weighting problem; diagnostics (Goodman-Bacon 2019).
   - Callaway and Sant'Anna 2020, dynamics.
2. Solutions (in an event study setting)
   - Abraham and Sun 2020: "saturated´´ event study design.
   - Group average treatment effects with covariates (Callaway Sant'Anna 2020).
3. Example application
   - Shover et al. (PNAS): Does the adoption of medical cannabis have a causal effect on opioid overdose mortality? (Andrew Baker's discussion)

Recent advances in differences-in-differences with heterogeneous treatment effects

Some recent literature I

- Borusyak and Jaravel (2017) with TWFE (unit and time fixed effects) one cannot identify the linear component of pre-trends and dynamic treatment effects. They propose a solution for this.
- Athey and Imbens (2018) - treatment assignment time is random. Inference focus (randomization inference), suitable for settings with staggered treatment adoption.
- Goodman-Bacon (2019) TWFE linear regression with staggered treatment adoption. Diagnostic tool for DiD designs with multiple periods, decomposition theorem.

Recent advances in differences-in-differences with heterogeneous treatment effects

Some recent literature II

- Deshpande and Li (2019), Cengiz et al. (2019) for event studies. "Stacked dataset" structure ensures that you can only make only easily interpretable comparisons.
- Imai and Kim (2019) units can switch in and out of treatment at different periods in time - need not to stay treated as in CS(2020). "Matching estimator".
- Sun and Abraham (2020) treatment effect dynamics, in particular propose a way for aggregating treatment effects using event studies
- Callaway and Sant'Anna (2020), in addition to SA(2020) propose a way of aggregation that accounts for covariates (and allows for pre-trends to hold conditional on covariates)

Differences-in-differences

Canonical setting - 2 X 2 Comparison

Standard two-period setting, with one treated and one control group. Denote t as treatment group and c as control. In the standard 2X2 case, we will compare:

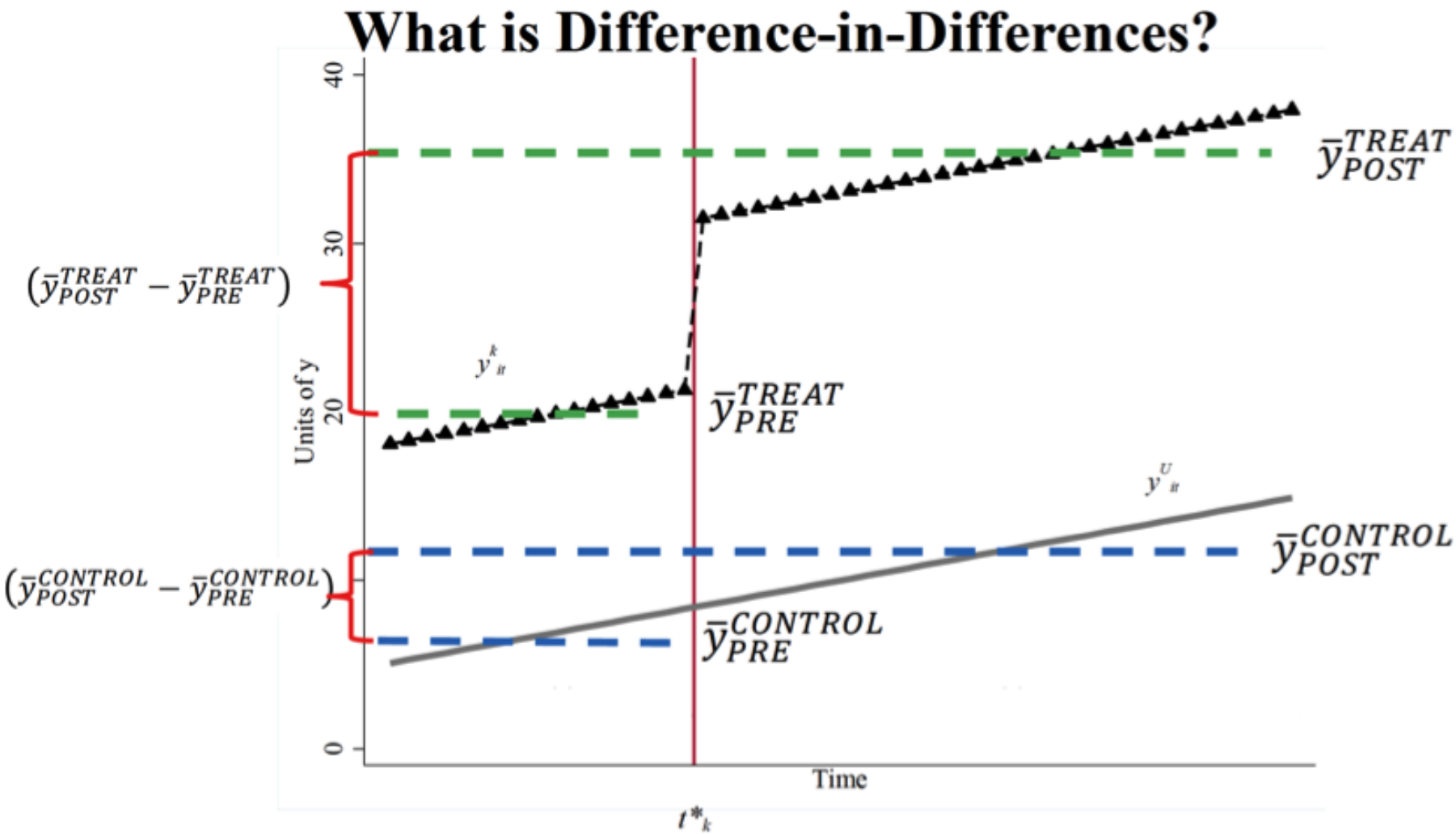
$$\widehat{\delta}_{TC}^{2\times 2} = \left(\bar{y}_T^{\text{post}} - \bar{y}_T^{pre}\right) - \left(\bar{y}_C^{post} - \bar{y}_C^{pre}\right) \tag{1}$$

Key assumption is additivity of potential outcomes, in absence of the treatment:

$$E\left[Y_i \mid D_i = 0, t\right] = \gamma_s + \lambda_t \tag{2}$$

Commonly described as the "parallel trends assumption" - entails (1) constant selection bias (2) same time trend for treatment and control groups λ_t.

Graphically



Differences-in-differences

With two groups and two time periods, the typical regression model is:

$$y_{it} = \beta_1 + \beta_2 Treat_i + \beta_3 Post_t + \beta_4 (Treat_{it} \text{ x } Post_{it}) + \epsilon_{it} \qquad (3)$$

- Treat is a dummy for treatment group
- Post is a post-treatment dummy
- β_1=
- β_2=
- β_3=
- β_4=

$$y_{it} = \beta_1 + \beta_2 Treat_i + \beta_3 Post_t + \beta_4 (Treat_{it} \text{ x } Post_{it}) + \epsilon_{it} \quad (4)$$

- Treat is a dummy for treatment group
- Post is a post-treatment dummy
- β_1=pre-program mean in control group
- β_2=Treatment vs. Control comparison - captures selection bias (assumed to be time-invariant)
- β_3=Pre vs. Post comparison, capturing time trend
- β_4=is the DD effect, identifying ATT

Differences-in-differences

Multiple time periods - TWFE

When more than 2 periods and 2 units are available, and units potentially experience treatment at different times, the typical model becomes:

$$y_{it} = \beta_i + \beta_t + \beta^{DD} D_{it} + \epsilon_{it} \tag{5}$$

- α_i are unit-level fixed-effects
- α_t time fixed-effects
- D_{it} unit-time indicator for treatment
- cluster S.E. at the group level to allow for serial correlation (Bertrand, Duflo and Mullainathan 2004)

β^{DD} is the coefficient of interest. But the interpretation of β^{DD} is not straightforward.

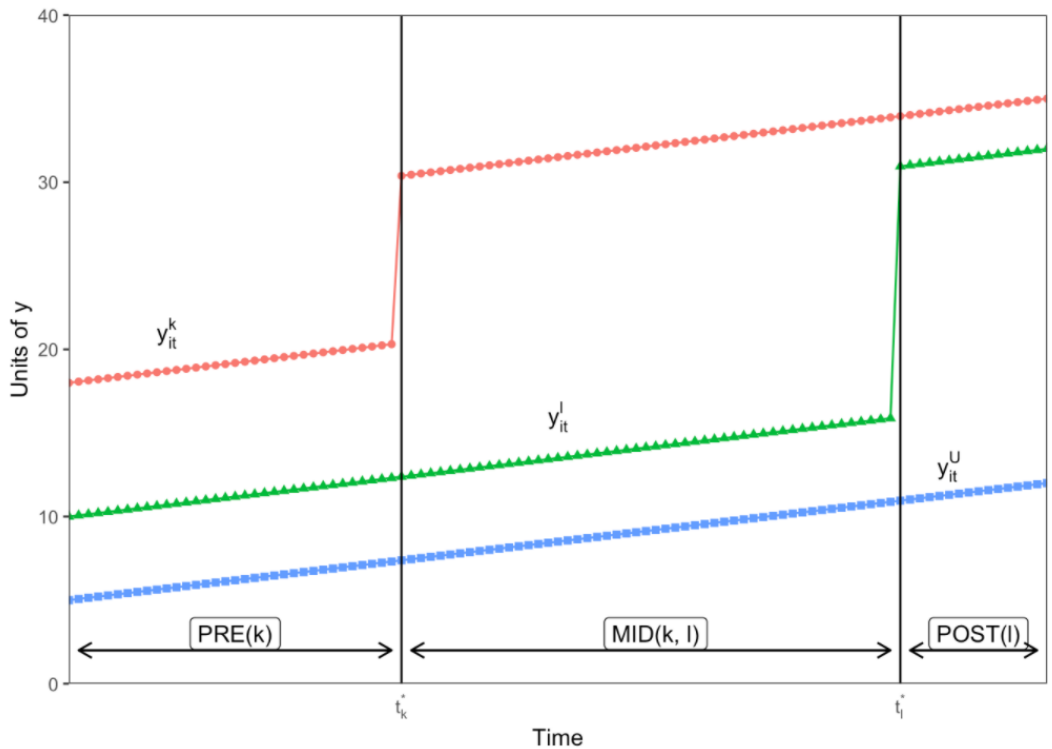
DD With Variation in Treatment Timing (Goodman-Bacon 2019)

- A common deviation from standard 2x2 set up is DD with staggered adoption: different units receive treatment at different times
- With multiple groups and staggered adoption, β_{DD} is a weighted average of all the possible 2x2 comparisons across groups over time
- Weights depend on a) group size b) variance of treatment dummies for the groups
- Groups with more units and/or treated in the 'middle' of sample get more weight
- Treated units act as both controls and treatment depending on the comparison: problems in presence of heterogeneous/dynamic TE

Example

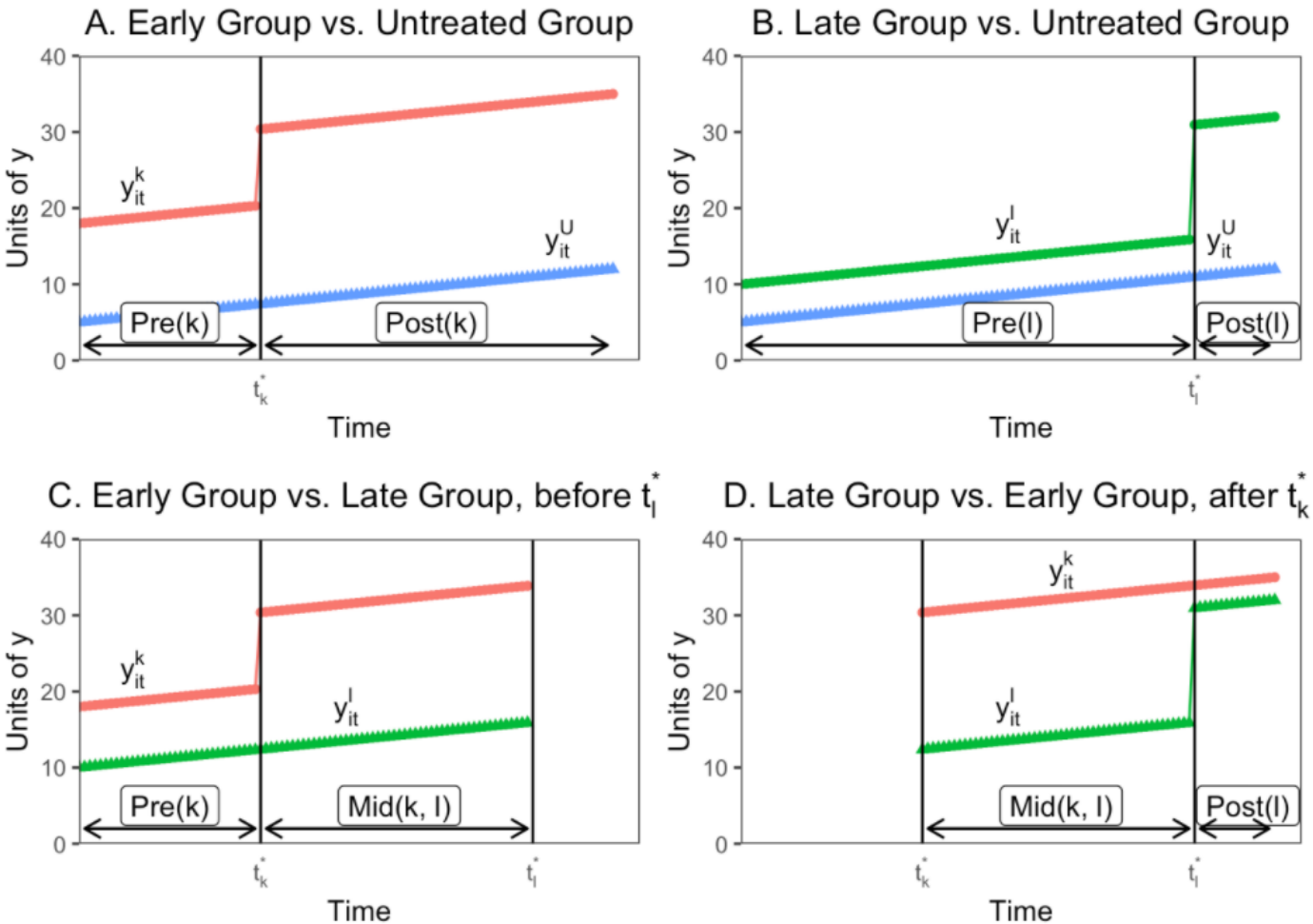
Goodman-Bacon 2019 - Multiple time periods

- 3 groups: Never treated U, early treatment k, late treated l.
- 3 windows: "PRE", "MID", "POST".



Decomposition of β^{DD}

Goodman-Bacon 2019



Decomposition of β^{DD}

Goodman-Bacon 2019

- Treated vs. Untreated comparisons:

$$\hat{\beta}_{jU}^{2x2} = \left(\bar{y}_j^{POST(j)} - \bar{y}_j^{PRE(j)}\right) - \left(\bar{y}_U^{POST(j)} - \bar{y}_U^{PRE(J)}\right), j = k, l \quad (6)$$

- Treated and not-yet treated comparison:

$$\hat{\beta}_{kl}^{2x2,k} = \left(\bar{y}_k^{MID(k,l)} - \bar{y}_k^{PRE(k)}\right) - \left(\bar{y}_l^{MID(k,l)} - \bar{y}_l^{PRE(k)}\right) \quad (7)$$

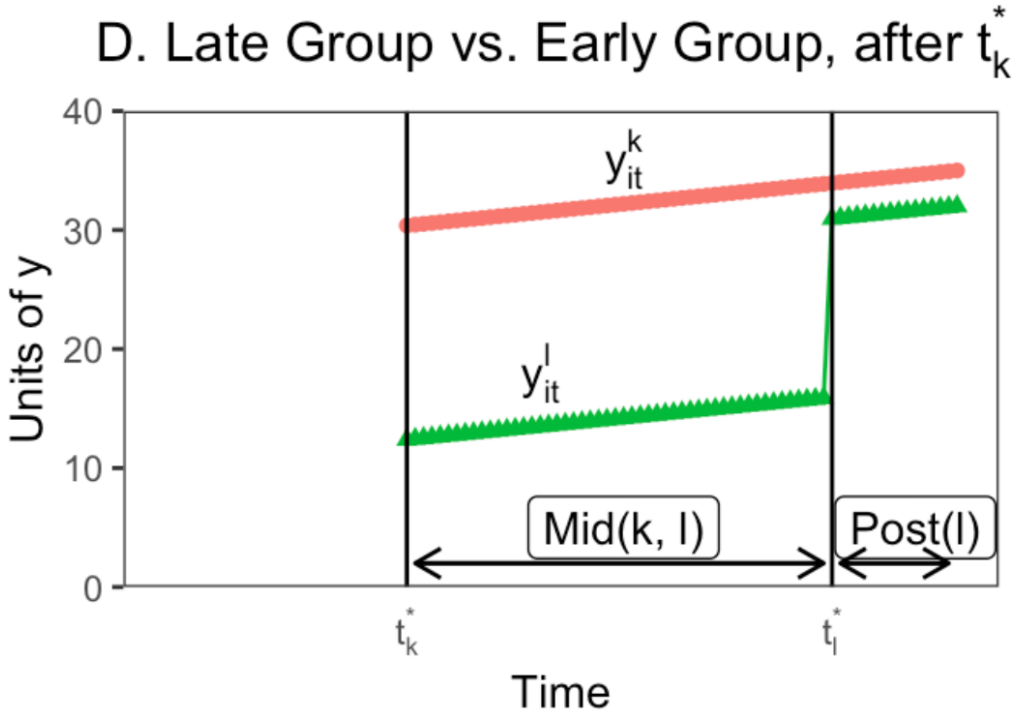
- Treated later and treated before comparison:

$$\hat{\beta}_{k,l}^{2x2,l} = \left(\bar{y}_l^{POST(l)} - \bar{y}_l^{MID(k,l)}\right) - \left(\bar{y}_k^{POST(l)} - \bar{y}_k^{MID(k,l)}\right) \quad (8)$$

Potentially problematic case

Using already treated as control, if there are dynamics

Already treated unit is treated as a control in the regression, their indicator does not switch from a period to the other.



Comparisons

- With K timing groups, you can form $K^2 - K$ timing only estimates comparing earlier and later treated groups.
- With an untreated group U you could form K treated/untreated 2x2 DiDs for a total of K^2 DiD estimates (4 in this case)
- Bias may arise when using already treated as control, if treatment effects are dynamic (there is a trend in treatment effects). This bias feeds through to the later comparisons, according to the size of the weights.

Decomposition Theorem

Comparisons and weights

The DiD decomposition theorem, 2 treatment groups and 2 time periods:

$$\hat{\beta}^{DD} = s_{kU}\hat{\beta}^{2x2}_{kU} + s_{\ell U}\hat{\beta}^{2x2}_{\ell U} + s^{k}_{k\ell}\hat{\beta}^{2x2,k}_{k\ell} + s^{\ell}_{k\ell}\hat{\beta}^{2x2,\ell}_{k\ell} \tag{9}$$

Note that this is a weighted average of all possible pairwise DD comparisons (slides before).

Weights

Estimated weights depend on sample shares and variance of treatment effects

Example: weight of the comparison between k and U will be:

$$s_{ku} = (n_k + n_u)^2 \frac{n_{ku}\left(1 - n_{ku}\right)\bar{D}_k\left(1 - \bar{D}_k\right)}{\widehat{\text{Var}}\left(\tilde{D}_{it}\right)} \tag{10}$$

- n_{ku} refer to relative sample sizes (Concentration) ($n_{ku} = n_k/\left(n_k + n_u\right)$)
- $\bar{D}_k$ = share of time group k spends treated.

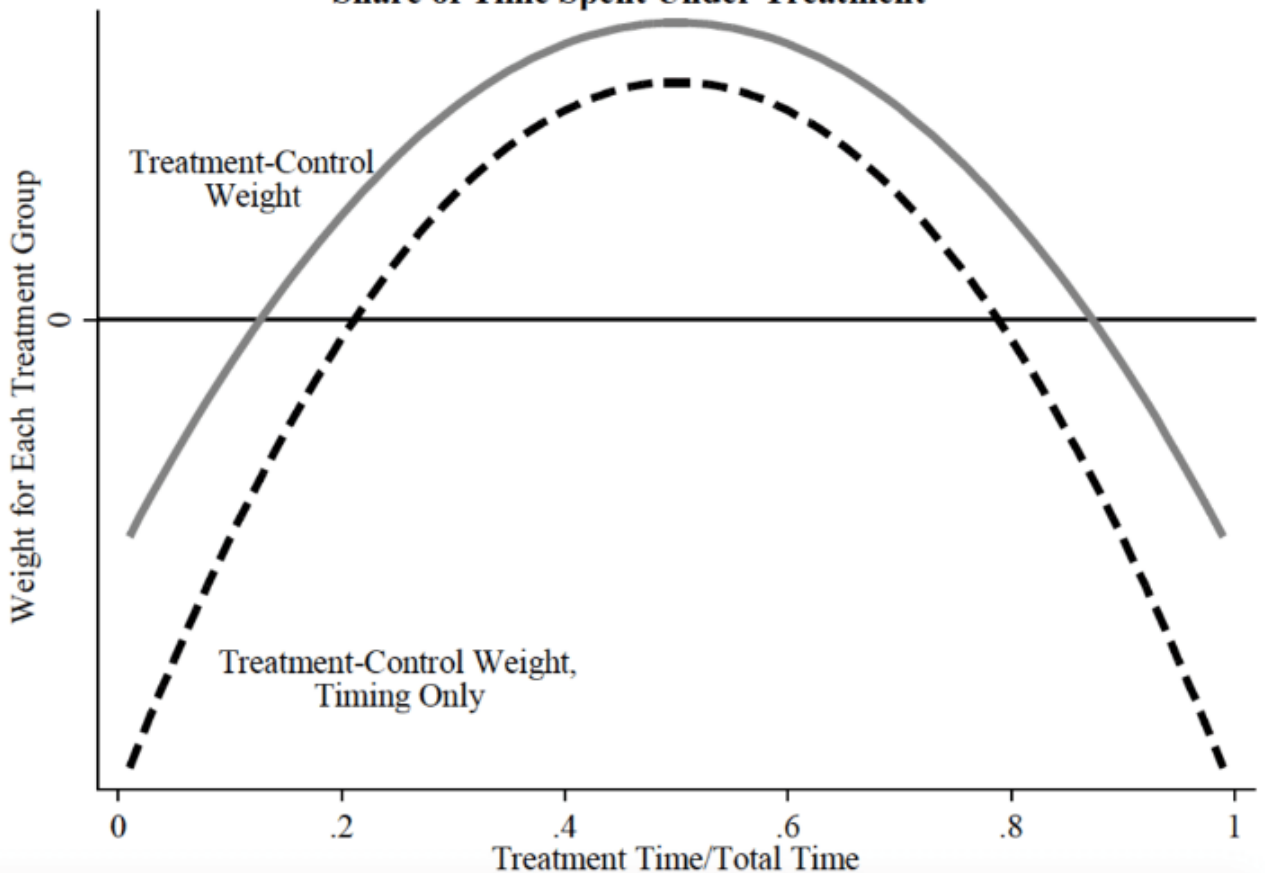
Notes:

- Each of the 2x2 DiDs identified by the treatment indicator variation in the sub-sample over which it is estimated.
- The share of the sample these observations represent also enter the weighting.
- At the middle of the panel (n_{ku} and $\bar{D}_k$=1/2) the weight is maximized.

Weights

Weights are maximized at 1/2 Treatment Time (Middle of panel)

Figure 4. Weighted Common Trends: The Treatment/Control Weights as a Function of the Share of Time Spent Under Treatment



Weights

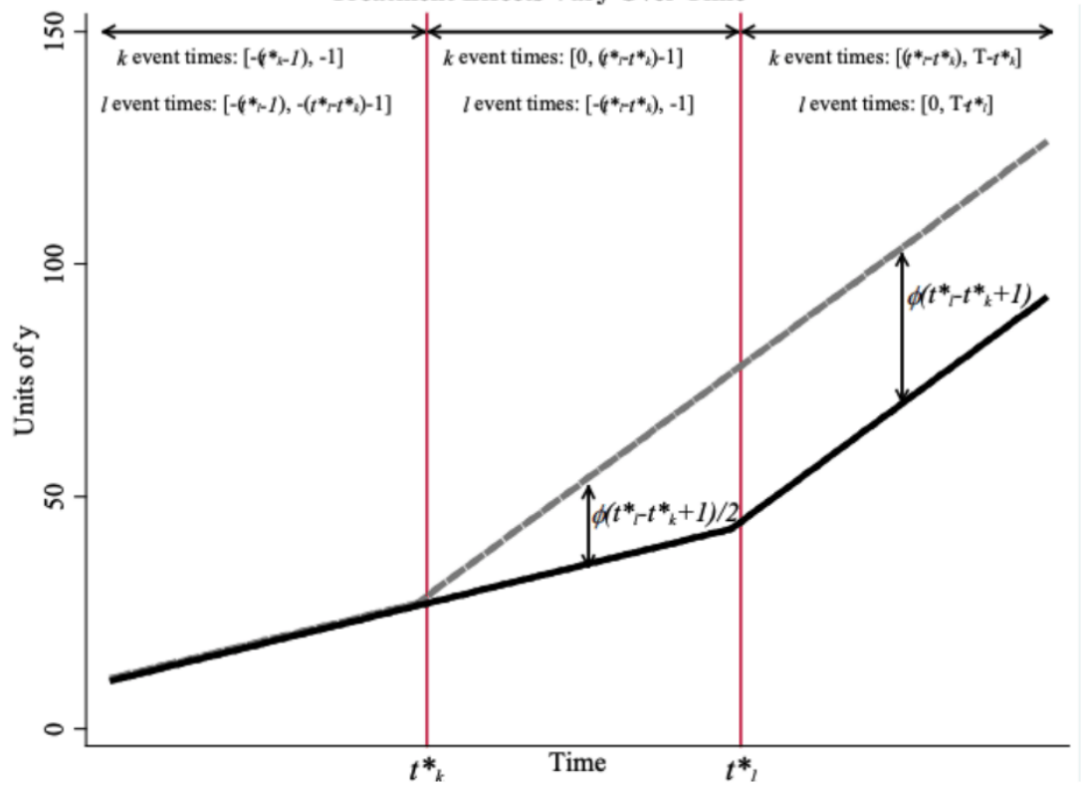
Take-aways

- **Weights are not equal to sample shares, in general.**
- Even if the treatment effects are constant, panel length by itself may affect the estimates.
- Estimates closer to the "middle" of the panel get more weight.
- Diagnostics proposed by Goodman-Bacon enable to assess which groups contribute the most to the observed treatment effect.
- So far we concentrated on weights, but we might also have dynamics (trends) in each unit's treatment effect.

Bias from dynamics

Goodman-Bacon 2019

Figure 3. Difference-in-Differences Estimates with Variation in Timing Are Biased When Treatment Effects Vary Over Time



Heterogeneity bias

The TWFE DiD estimator (when $N \to \infty$, T fixed):

$$\text{plim}_{N\to\infty} \hat{\beta}^{DD} = VWATT + VWCT - \Delta ATT \tag{11}$$

- VWATT is the "variance-weighted average treatment effect on the treated"
- VWCT is the "variance-weighted common trend". Different groups might not have the same underlying trend in outcome dynamics.
- ΔATT weighted sum of the change in treatment effects within each unit's post-period with respect to another unit's treatment timing.
  - Last term enters because of the comparison between later-earlier treated, only when the treatment effect is not stable across the same unit over time, otherwise it is 0.
  - Note: heterogeneity across cohorts is not the problem here, rather dynamics for a given treatment unit over time.

Event study settings

Example from Callaway Sant'Anna 2020

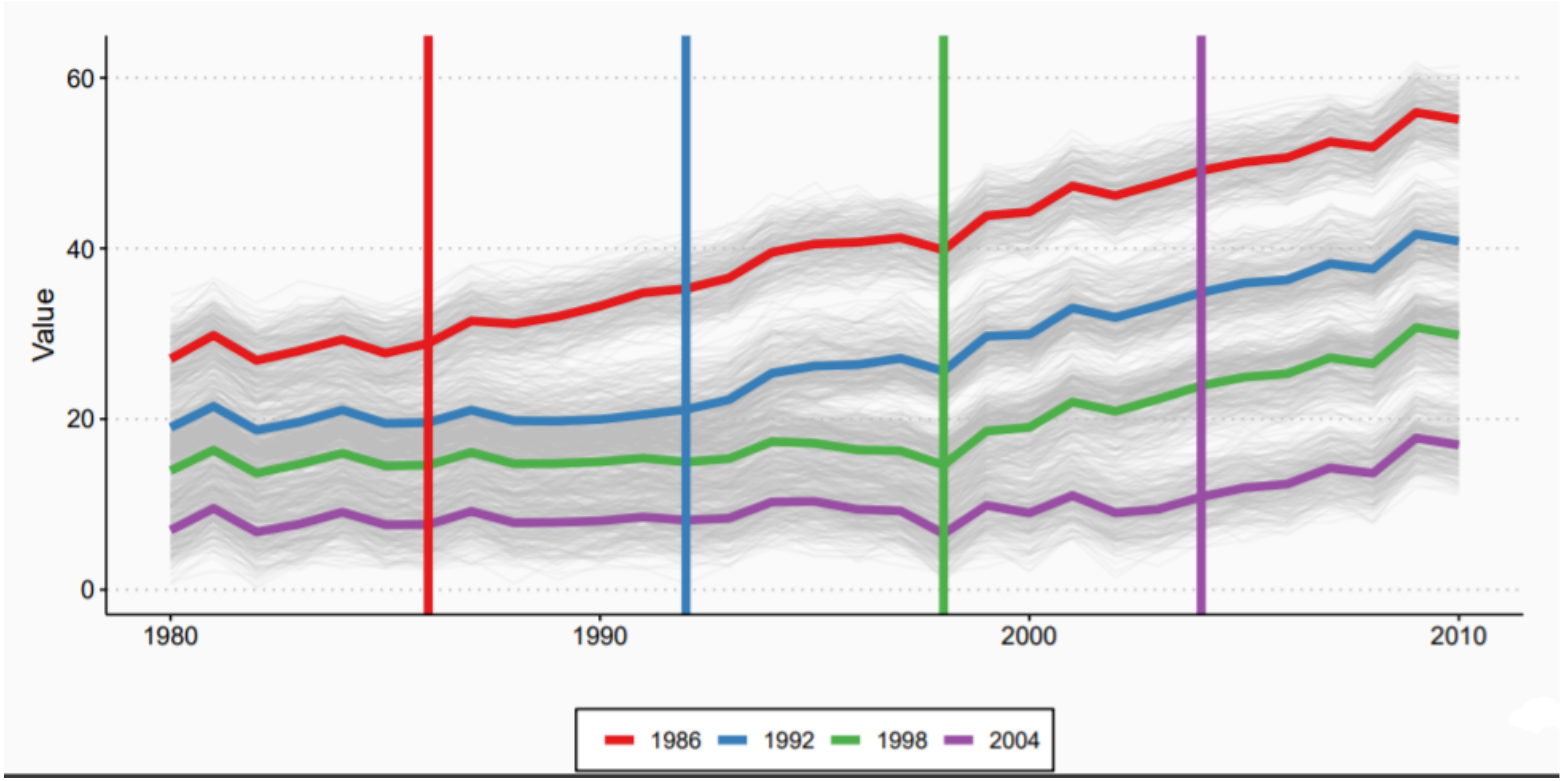


Figure 1: Simulation from Callaway Sant'Anna

Event study setting

Example

Event study setting takes the form:

$$Y_{i,t} = \alpha_i + \alpha_t + \gamma_k^{-K} D_{i,t}^{<-K} + \sum_{k=-K}^{-2} \gamma_k^{lead} D_{i,t}^{k} + \sum_{k=0}^{L} \gamma_k^{lags} D_{i,t}^{k} + \gamma_k^{L+} D_{i,t}^{>L} + \varepsilon_{i,t} \tag{12}$$

D_{it}^{l} indicator for being l time periods relative to i 's initial treatment (treatment $l = 0$), and α_i and α_t are unit and time fixed effects.

Event study settings

Example from Callaway Sant'Anna 2020

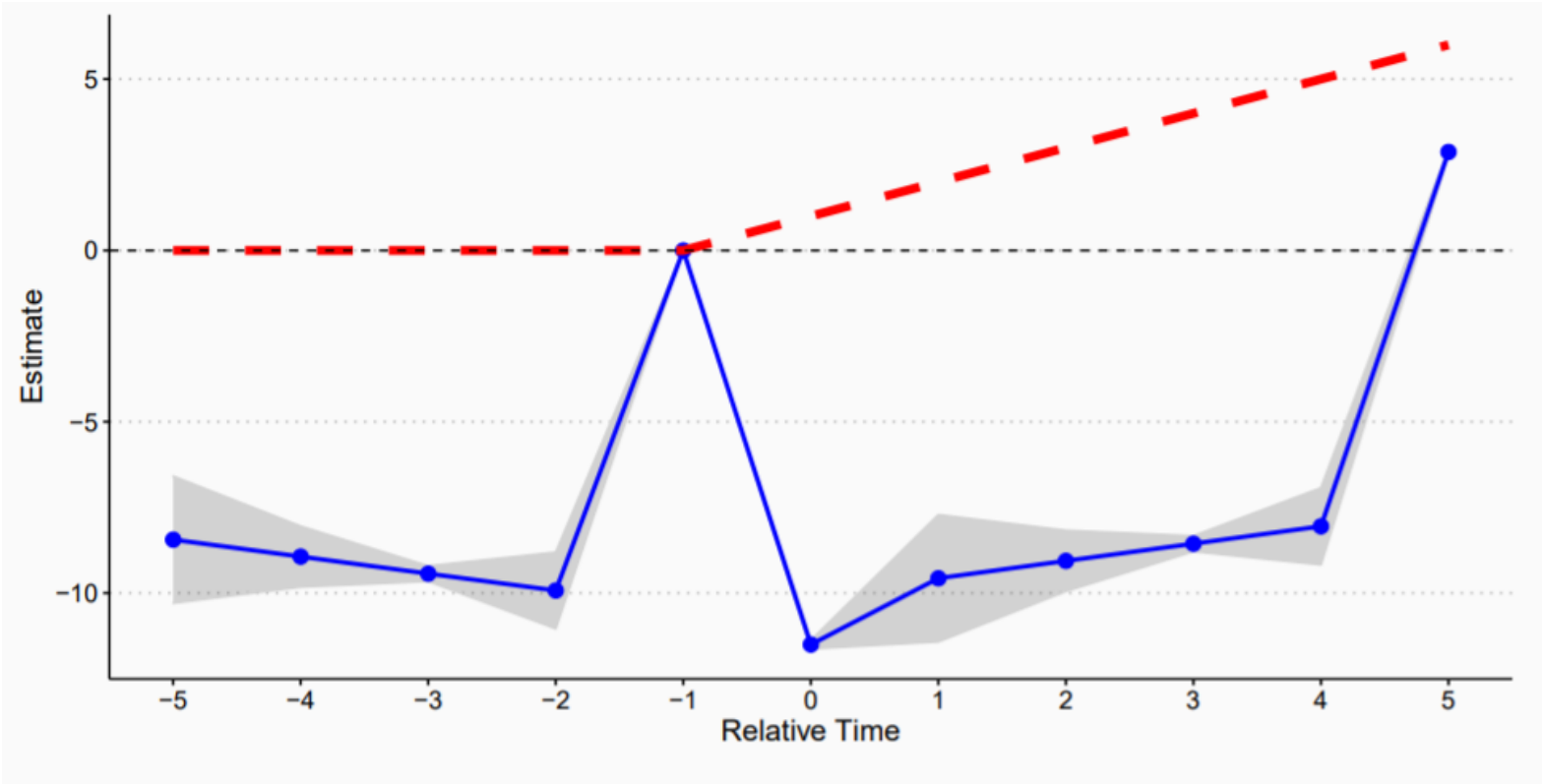


Figure 2: Event study - All leads and lags

Event studies

Abraham and Sun 2020

- In the previous case (CS simulation), treatment effects had the same shape across cohorts, so mainly we want to ensure we do not use earlier treated as controls.
- In general, we can not make sure treatment effect shape is homogeneous across groups.
- **Event-study breaks down when the treatment effects' shape (slope) is not constant across cohorts.**
  - In settings with variation in treatment timing across units, the coefficient on a given lead or lag can be contaminated by effects from other periods, and apparent pre-trends can arise solely from treatment effects heterogeneity.
- Solution:
  1. Cohort-specific treatment effects: dummies for each cohort relative-timeXtreatment-cohorts interaction (cohort-specific treatment dummies)
  2. Aggregation: weight by cohort size.

Abraham and Sun 2020

$$y_{it} = \alpha_i + \alpha_t + \sum_e \sum_{l \neq -1} \gamma_{e,l} \left(1\{E_i = e\} \cdot D_{i,t}^l \right) + \epsilon_{i,t} \qquad (13)$$

- E_i : cohort-specific indicators, determine whether cohort e entered treatment.
- D_{it}^l: relative time indicators, l periods from treatment.
- Omit t = -1, and last treated cohort.
- We will have relative year indicators for each treatment cohort, and arrive at Cohort-specific Average Treatment effects (a different "event study" for each cohort).
- Linear combination of the CATTs for each relative time period l, weighting by each cohort's relative share of the sample.

Callaway and Sant'Anna (2020)

Reweighting estimators

ATT for a specific group and time, comparisons are made between "similar units" (propensity scores).

- Groups are cohorts of units treated at the same time
- Calculate an ATE per group/time (only using un-treated units as controls - "long-differences")
- Does not restrict heterogeneous TE or changes over time in TE
- Allow for covariates for matching comparisons
- Provides ways to aggregate over these to get a single ATT
- Bootstrap procedure to conduct asymptotically valid inference, adjusted for autocorrelation and clustering
- Applicable to panel settings and repeated cross-sections.

Callaway and Sant'Anna (2020)

Setting - Group Treatment Effects

Assumptions

1. Sample: i.i.d (panel)
2. Parallel trends, conditional on covariates

$$E\left[Y_t^0 - Y_{t-1}^0 \mid X, G_g = 1\right] = \left[Y_t^0 - Y_{t-1}^0 \mid X, C = 1\right]$$

3. Irreversible treatment
4. Common support (propensity score)

Callaway and Sant'Anna (2020)

Setting - Group Treatment Effects

- We are trying to estimate:

$$ATT(g,t) = E\left[Y_t^1 - Y_t^0 \mid G_g = 1\right]$$

- Estimator:

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

- Weighted average of the "long difference" of the outcome variable, weights depending on the propensity score ($p_g(X)$, estimated by $\widehat{p}_g(X)$)
- We only use comparisons between g and its control.
- More weight to observations from the control group with similar characteristics to g
- Reweighting ensures covariate balance.

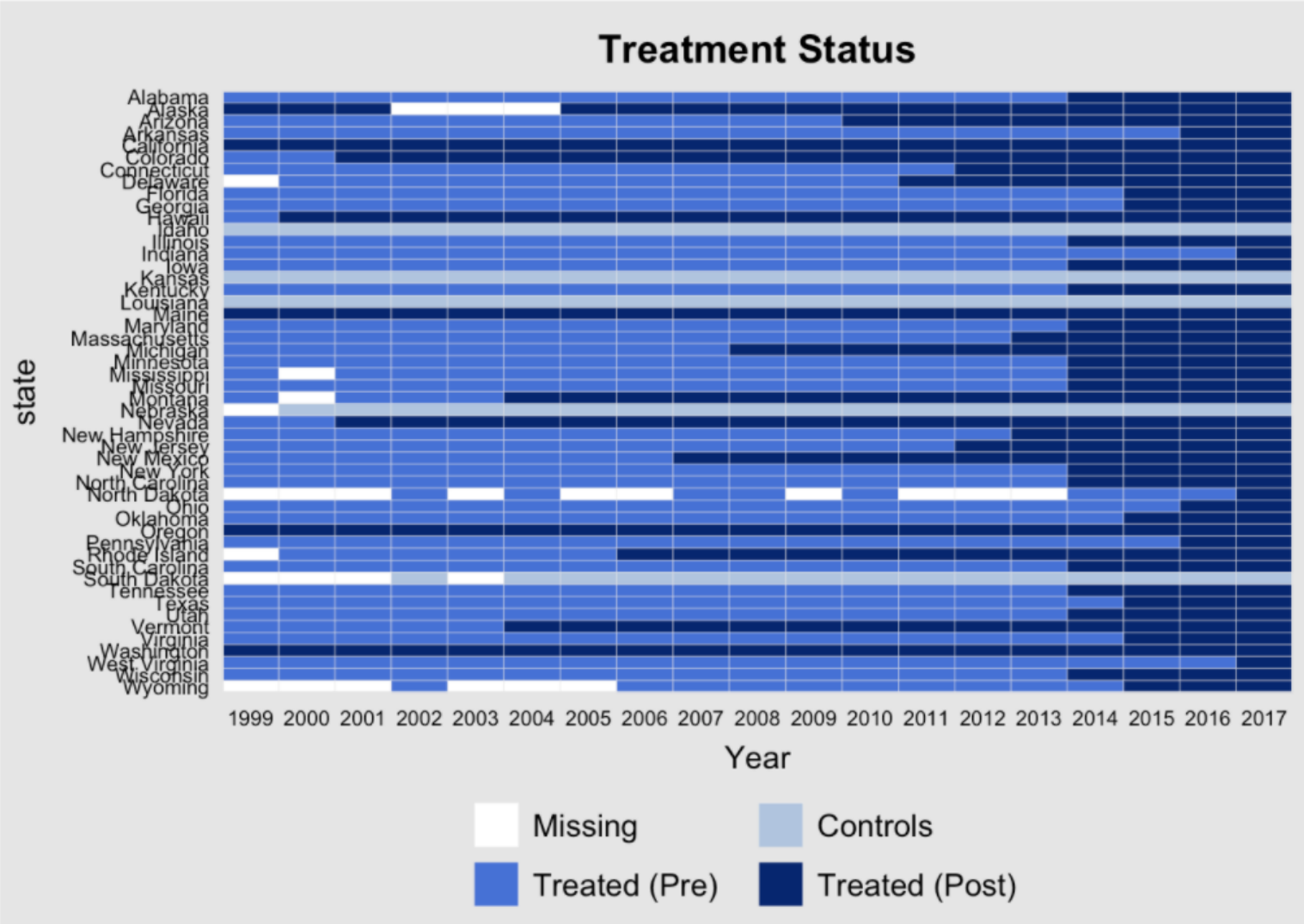
Application

What Can We Say About Medical Marijuana and Opioid Overdose Mortality?

- Does the adoption of legalized medical cannabis laws has a causal effect on opioid overdose mortality?
- Previous study (Bachhuber et al. 2014) shows a decrease in opioid overdose mortality ("deaths of despair").
- Shover, Davis, Gordon and Humphreys (PNAS) assess robustness of estimates, extending the original sample to 2017.
- *.." it is unlikely that medical cannabis, used by about 2.5% of the US population, has exerted large conflicting effects on opioid overdose mortality. A more plausible interpretation is that this association is spurious"*
- Could some of the staggered DiD issues be at play?

Example from Andrew Baker's website.

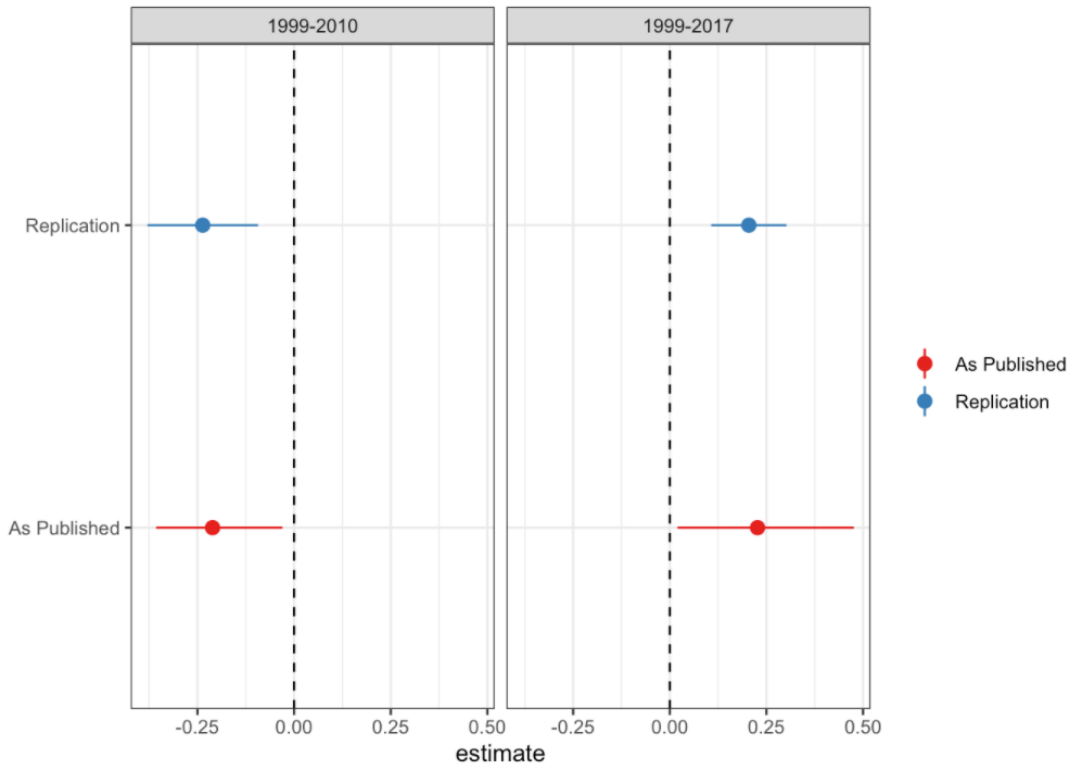
Treatment timing



Treatment effect

Andrew Baker's replication

Outcome: opioid overdose mortality.



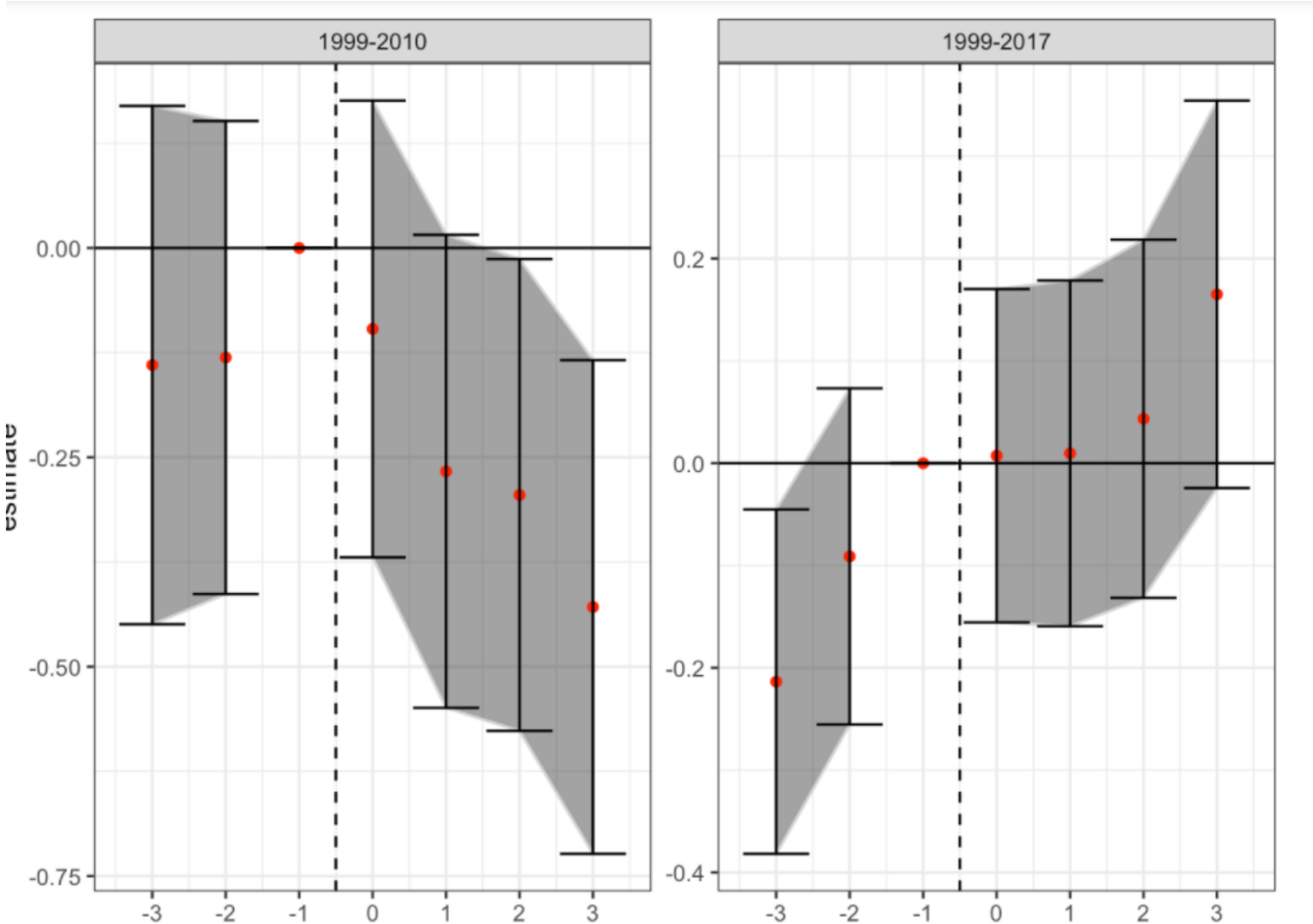
Estimation

$$y_{it} = \alpha_i + \alpha_t + \sum_{k=Pre,Post} \gamma_k + \sum_{-3}^{3} \gamma_k + \theta' X_{it} + \epsilon_{it}$$

where α_i and α_t are state and year fixed effects respectively, γ_k are the relevant time period indicators, and X_{it} is a matrix of state-year covariates. The event study DiD estimates for the two relevant time periods (1999-2010, and 19992017) are presented below.

Treatment effect - originally estimated

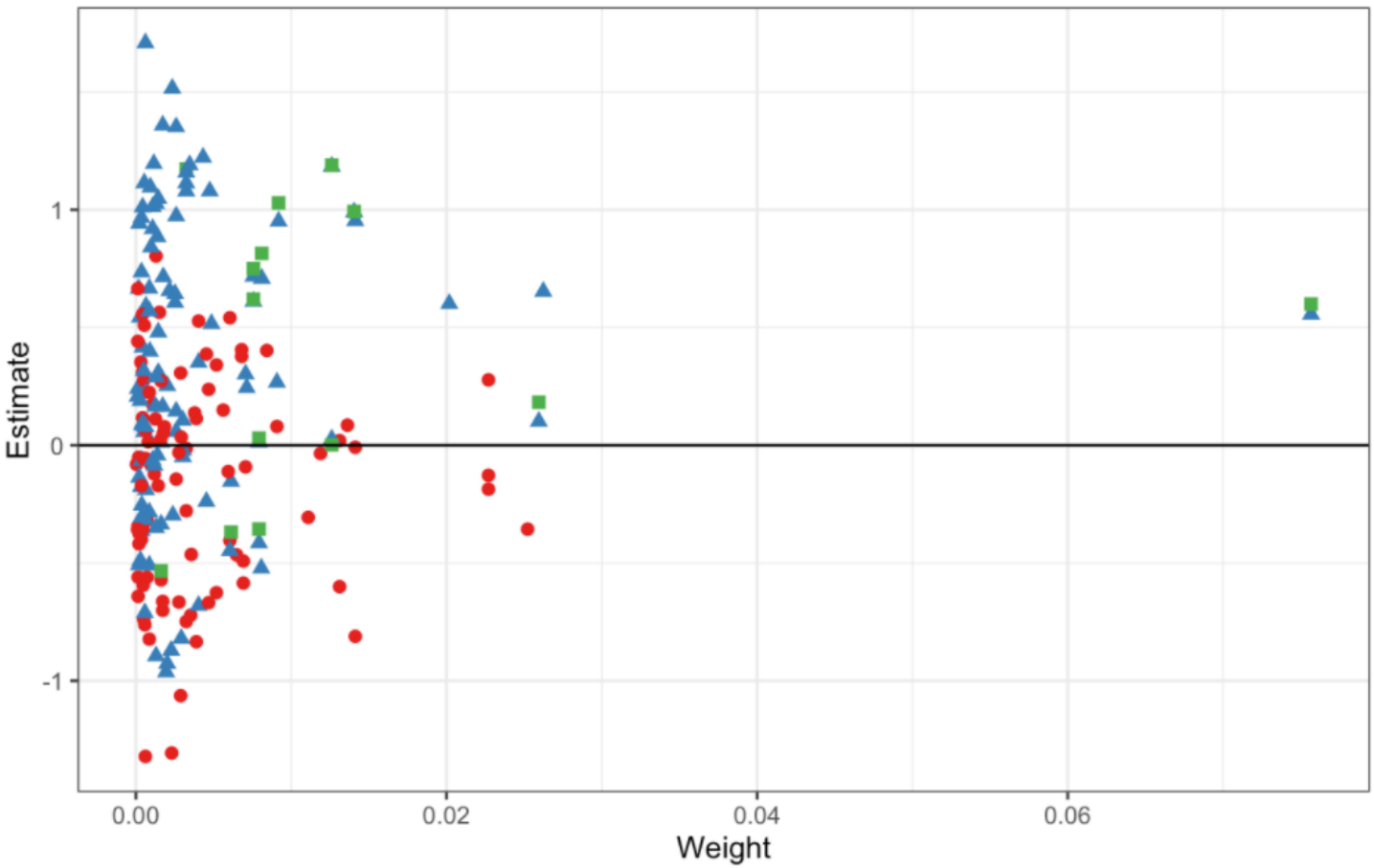
Divided in two groups: 1999-2010, 1999-2017



Step I. Diagnostics

Goodman-Bacon decomposition

- blue could be an issue - dynamic treatment effects which are subtracted out of the treatment effect estimate for the later-treated units.



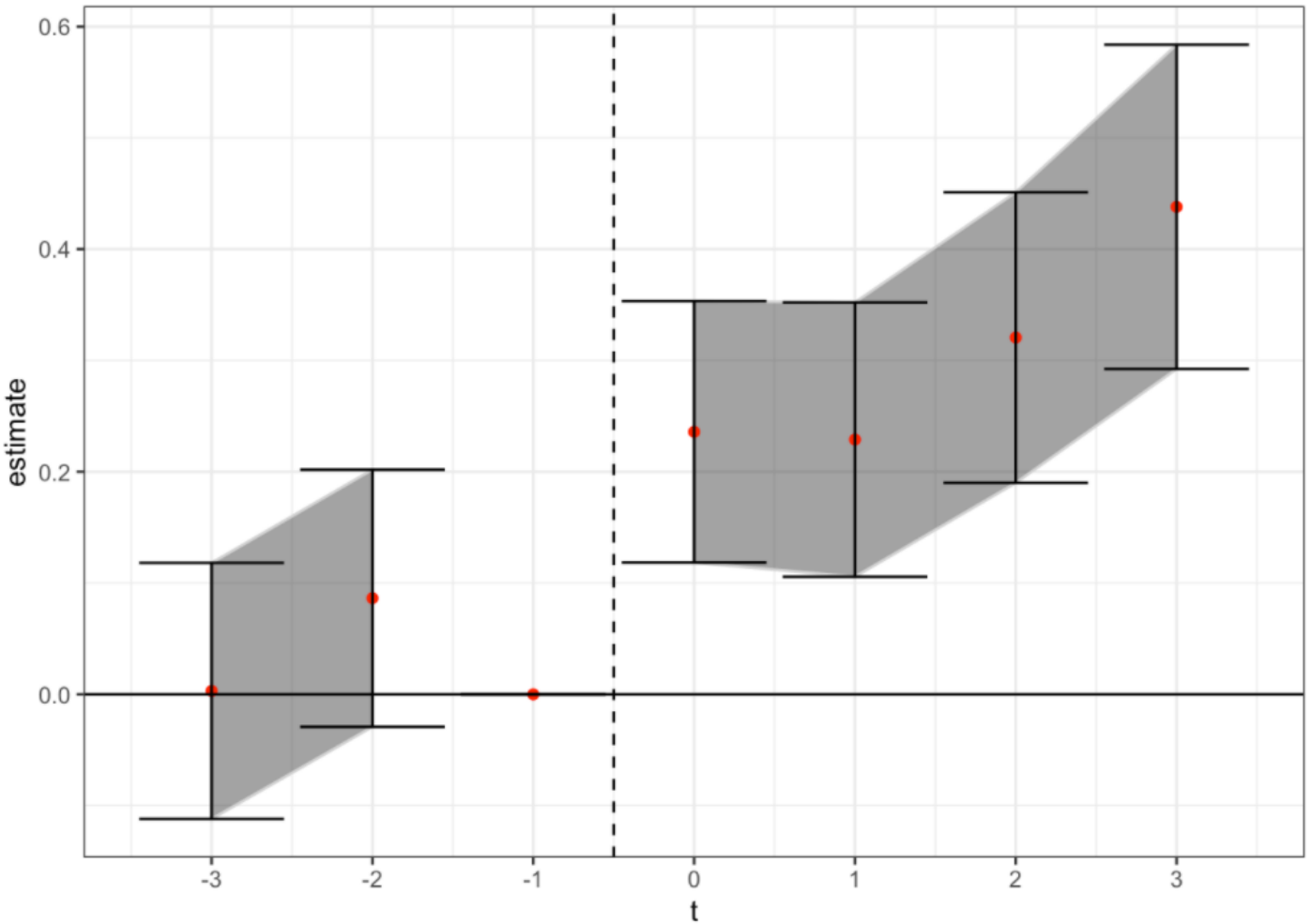
Step I. Diagnostics

Goodman-Bacon, applied by Andrew Baker

Type	Avg Estimate	N. of 2x2	Tot. Weight
Earlier- Later	-0.16	91	0.38
Later-Earlier	0.32	105	0.42
Treated-Untreated	0.44	14	0.20

Conclusion: Weight of problematic comparisons is high.
Standard staggered DiD likely a problem.

Step II. Application of Abraham and Sun 2020

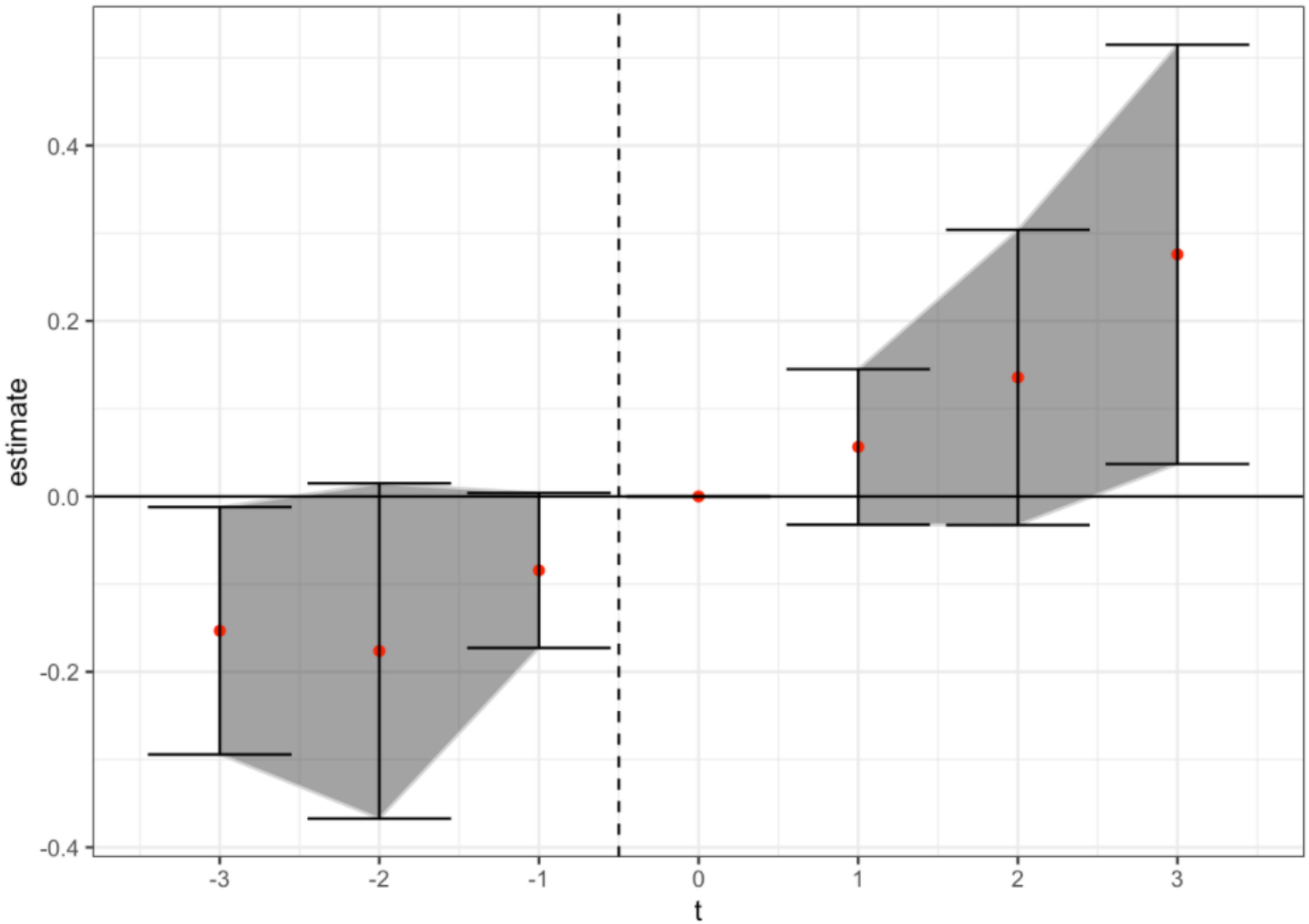


Step II. Application of Abraham and Sun 2020

- Now we are only adopting "valid" comparisons:
  - Later treated states take as comparisons i) even later treated states or ii) untreated.
  - N. of controls for later treated is low (we are making fewer comparisons)

Step II. Application of Callaway Sant'Anna 2020

Result



Positive effect, but may be starting before enactment of CL.

Conclusion

- Using two-way fixed effects estimators for staggered DiD designs may be problematic.
  - Weights + Dynamics.
- Goodman-Bacon shows that standard DiD is a weighted average between all possible 2X2 comparisons across periods and treated /untreated.
- Abraham and Sun (2020). Saturate the fixed effects structure to ensure that prior treated units do not enter within the test window as a control unit.
- Non-parametric correction by Callaway and Sant'Anna (2020) enables accounting for this, also exploiting covariates.
- Practical relevance: bias (from dynamics) can switch the sign of the estimate, weights not proportional to sample size.