DiD with variation in treatment timing

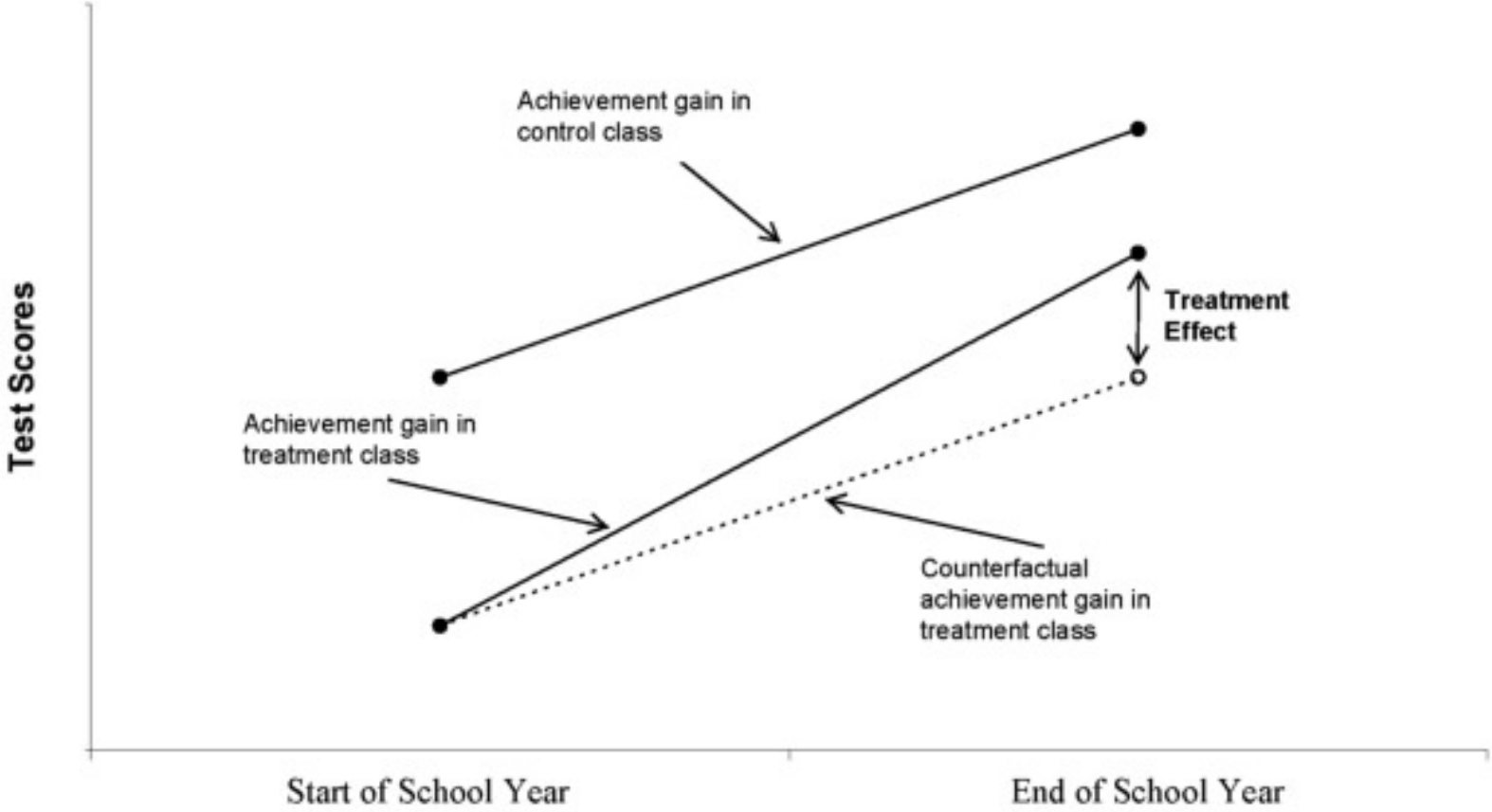
(Goodman-Bacon 2020)

Colin Aitken

April 11, 2022

Differences-in-Differences

Recall the classic differences-in-differences setup with units $i = 0, 1$ and time periods $t = 0, 1$:



Can recover treatment effect β^{DD} from the regression

$$Y_{it} = \alpha_i + \gamma_t + \beta^{DD} TREATED_{it} + \varepsilon_{it}$$

Two-Way Fixed Effects

Definition

The Two-Way Fixed Effect (TWFE) Estimator outputs the coefficient $\hat{\beta}^{DD}$ from the regression

$$Y_{it} = \alpha_i + \gamma_t + \beta^{DD} D_{it} + \varepsilon_{it}$$

(where we've abbreviated $TREATED_{it}$ to the more standard D_{it}.)

We just saw that with two units and two time periods (the "2×2 case"), TWFE is the standard diff-in-diff specification

Naive Idea (Assumed in like 50 years of empirical work)

We can use the same TWFE specification with more units and time periods, as long as we have parallel trends and treatment is absorbing (i.e. treated units stay treated forever)

What if this is misspecified? (i.e. heterogeneous treatment effects)

Outline of Today

- TWFE is a weighted average of "good", "bad", and "control" estimators.
  - These weights are all positive, but related to both group size and treatment timing.
  - Workaround: use these estimators directly and weight them however you like

Outline of Today

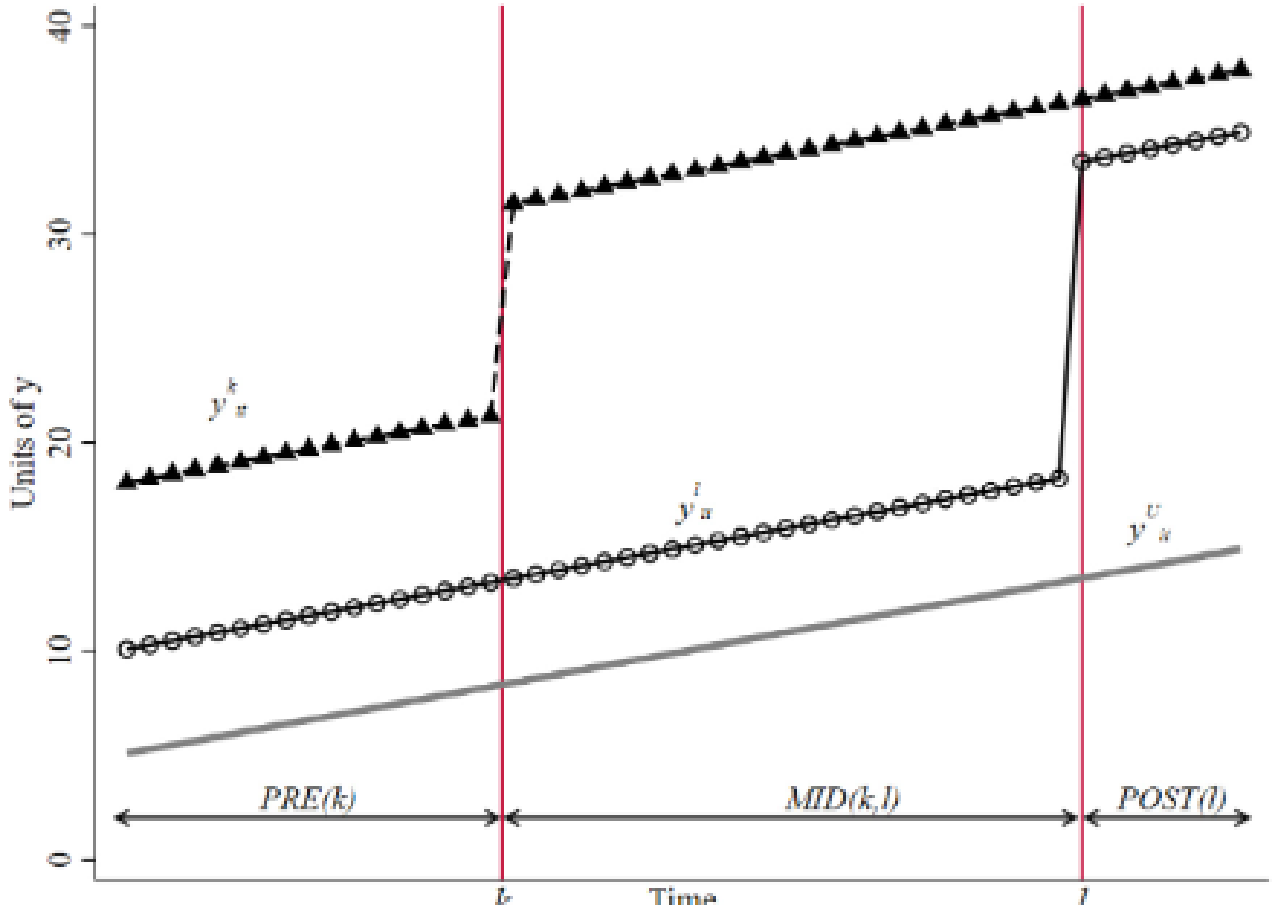
- TWFE is a weighted average of "good", "bad", and "control" estimators.
  - These weights are all positive, but related to both group size and treatment timing.
  - Workaround: use these estimators directly and weight them however you like
- "Bad" estimators use already-treated groups as a control group.
  - This introduces bias when treatment effects vary with time.
  - Can have the wrong sign in pretty reasonable situations.

Outline of Today

- TWFE is a weighted average of "good", "bad", and "control" estimators.
  - These weights are all positive, but related to both group size and treatment timing.
  - Workaround: use these estimators directly and weight them however you like
- "Bad" estimators use already-treated groups as a control group.
  - This introduces bias when treatment effects vary with time.
  - Can have the wrong sign in pretty reasonable situations.
- Time-varying covariates make the results of TWFE regression even harder to interpret
  - Should these be controlled for at all? (Post-treatment bias a real concern!)
  - Even without staggered adoption, introduce weird comparisons if population effects not actually linear.

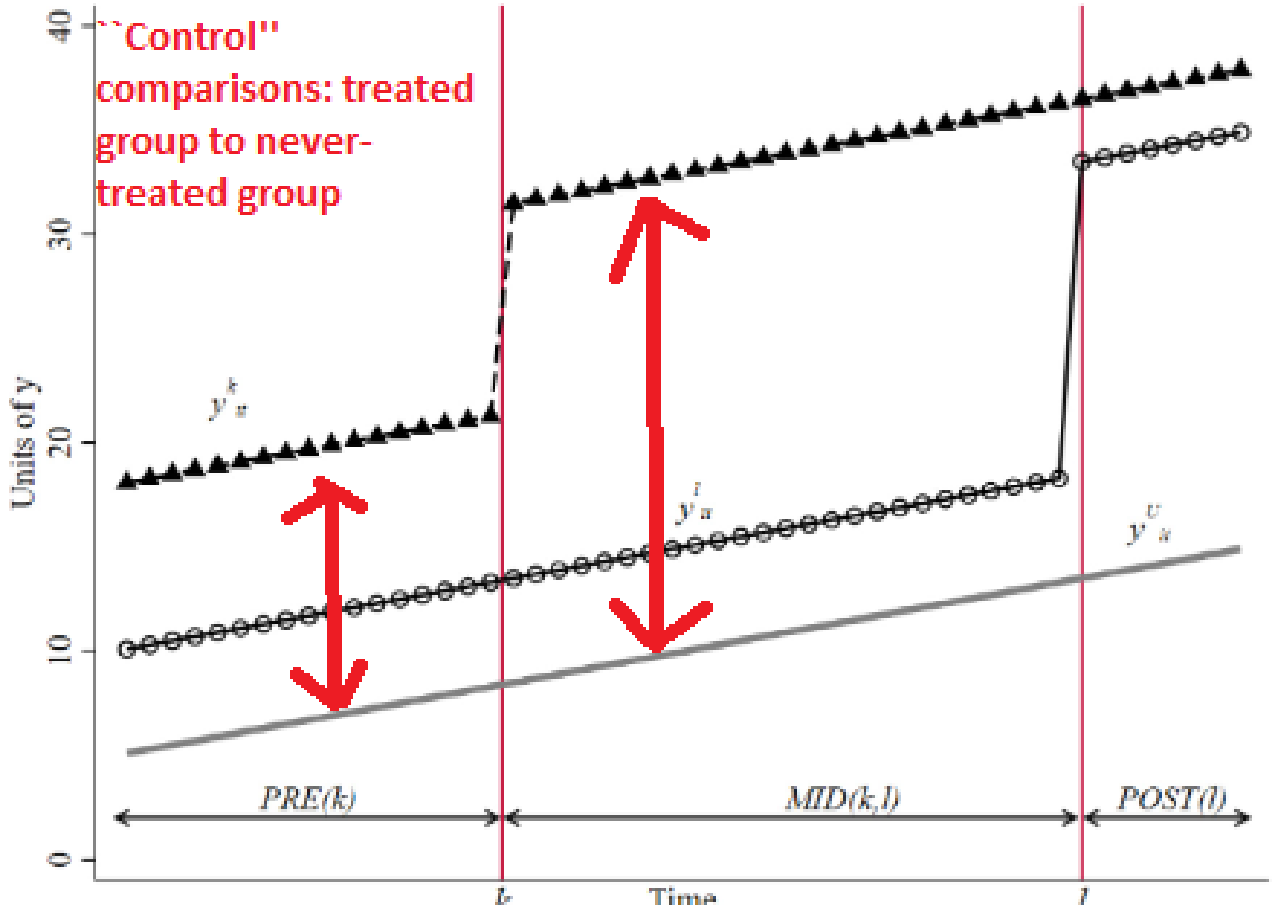
Three kinds of comparisons

With multiple treatment groups and staggered timing, a variety of "ordinary" differences-in-differences to compute



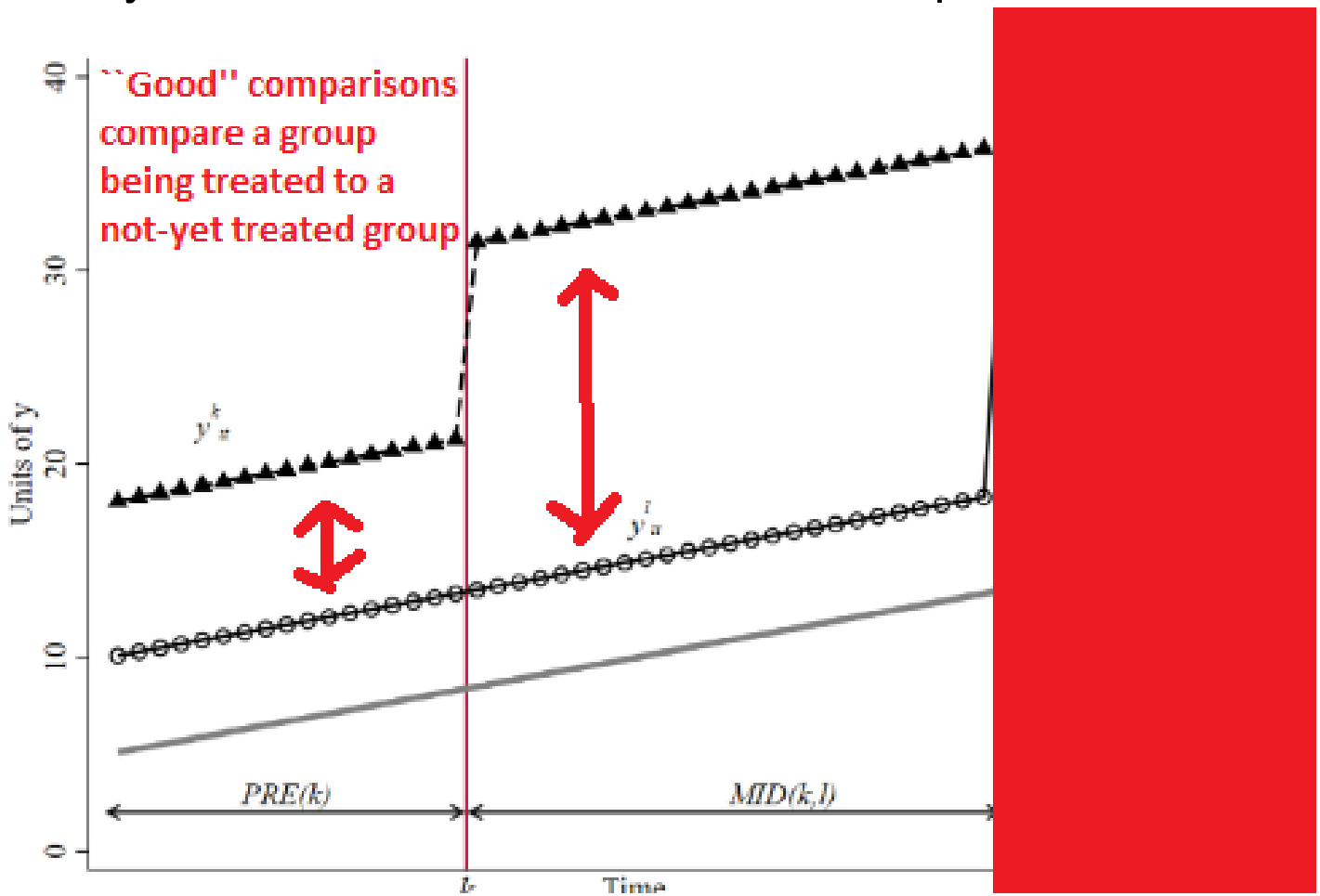
Three kinds of comparisons

With multiple treatment groups and staggered timing, a variety of "ordinary" differences-in-differences to compute



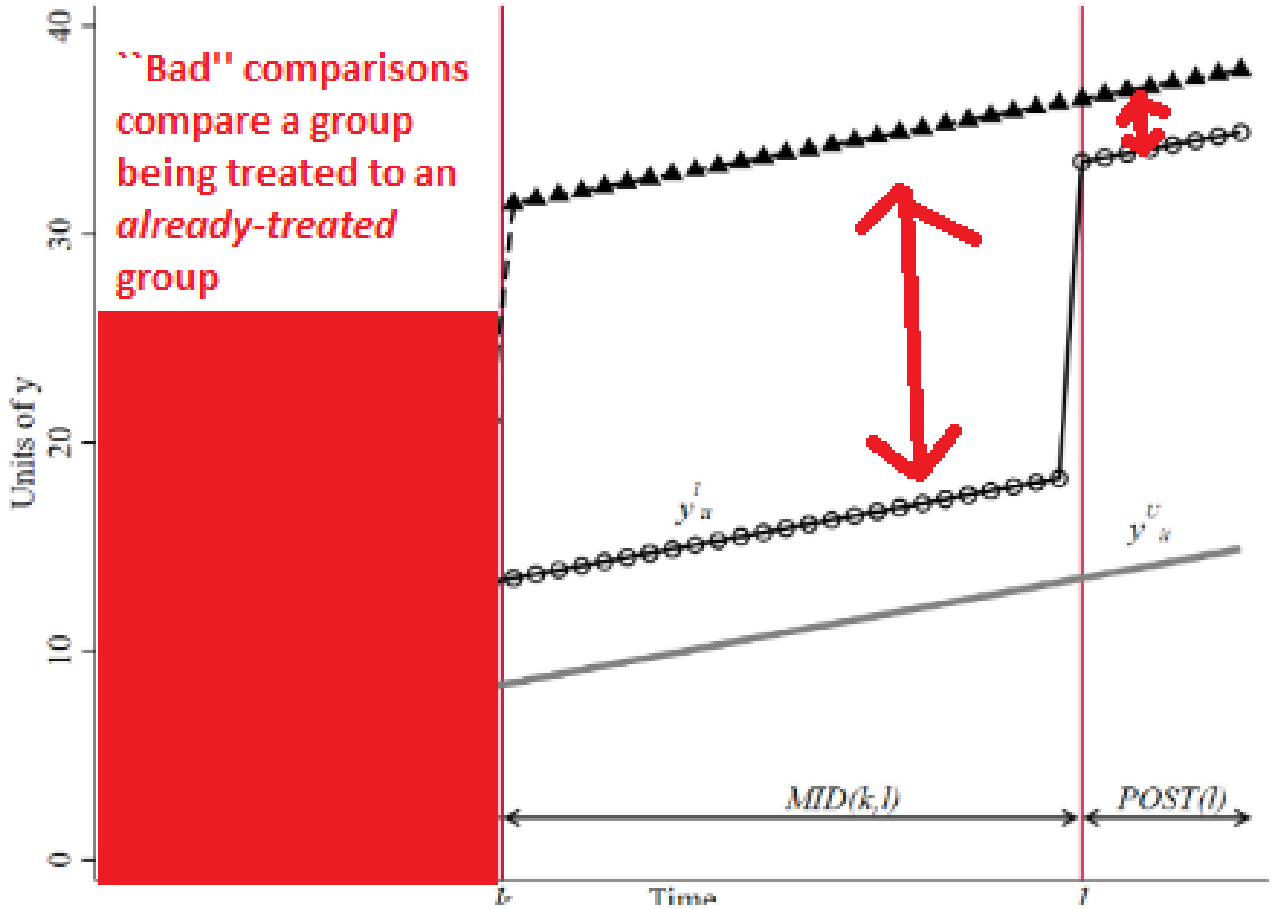
Three kinds of comparisons

With multiple treatment groups and staggered timing, a variety of "ordinary" differences-in-differences to compute



Three kinds of comparisons

With multiple treatment groups and staggered timing, a variety of "ordinary" differences-in-differences to compute



The Main Theorem (without controls)

Theorem

The two-way fixed effects estimator $\hat{\beta}^{DD}$ is a (positive) weighted average of the three types of "2×2 estimators":

$$\hat{\beta}^{DD} = \sum_{k \neq U} s_k^{control} \hat{\beta}_k^{control} + \sum_{k \neq U} \sum_{\ell > k} s_{k\ell}^{good} \hat{\beta}_{k\ell}^{good} + s_{k\ell}^{bad} \hat{\beta}_{k\ell}^{bad}$$

The weights s_* satisfy:

- They are all positive and sum to one.
- Comparisons between larger groups get more weight.
- (!!!) The *time at which a group is treated* significantly affects the weights

What are these weights?

Proposition

We can write:

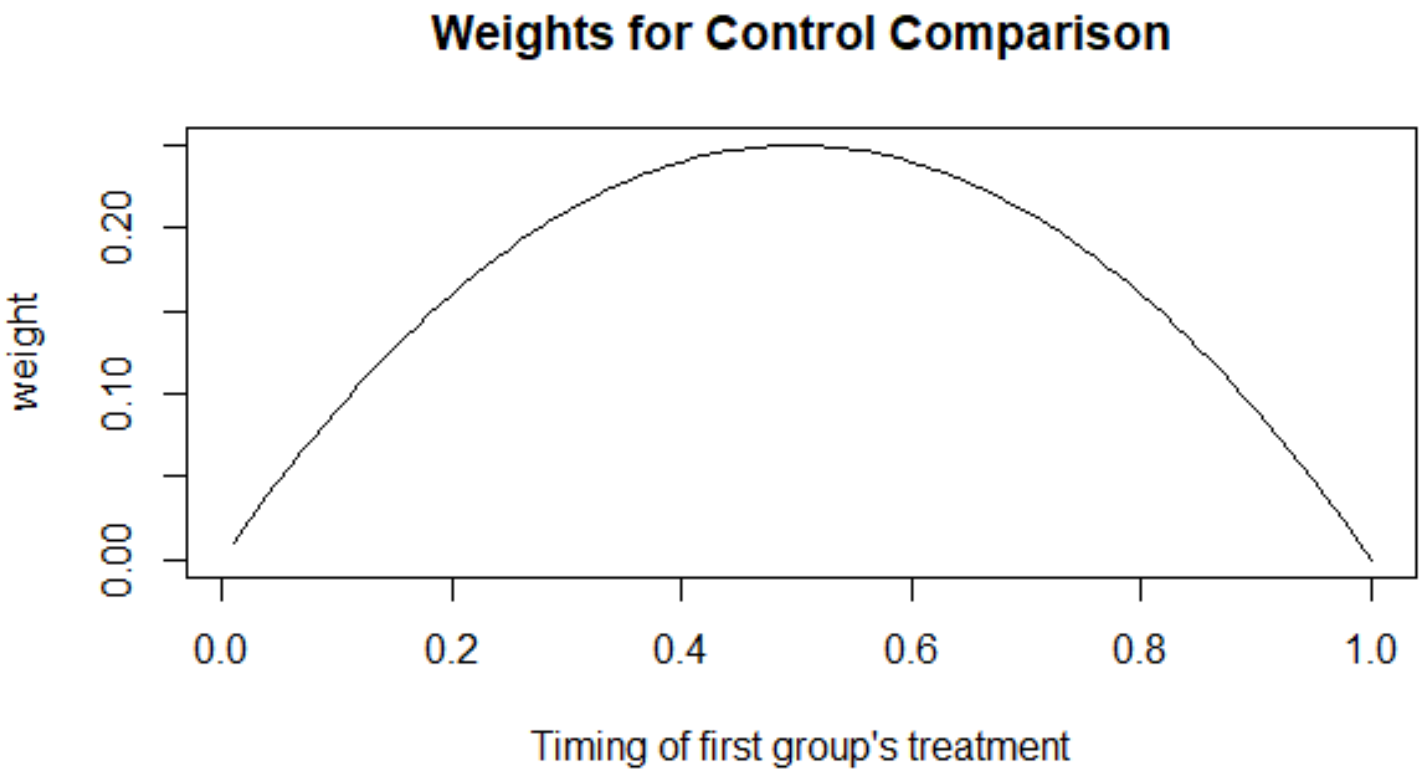
$$s_{k\ell}^{good} = \frac{\left(\Phi_{k\ell}^{good}\right)^2 \left(\hat{V}_{k\ell}^{D,good}\right)}{\hat{V}^D}$$

where:

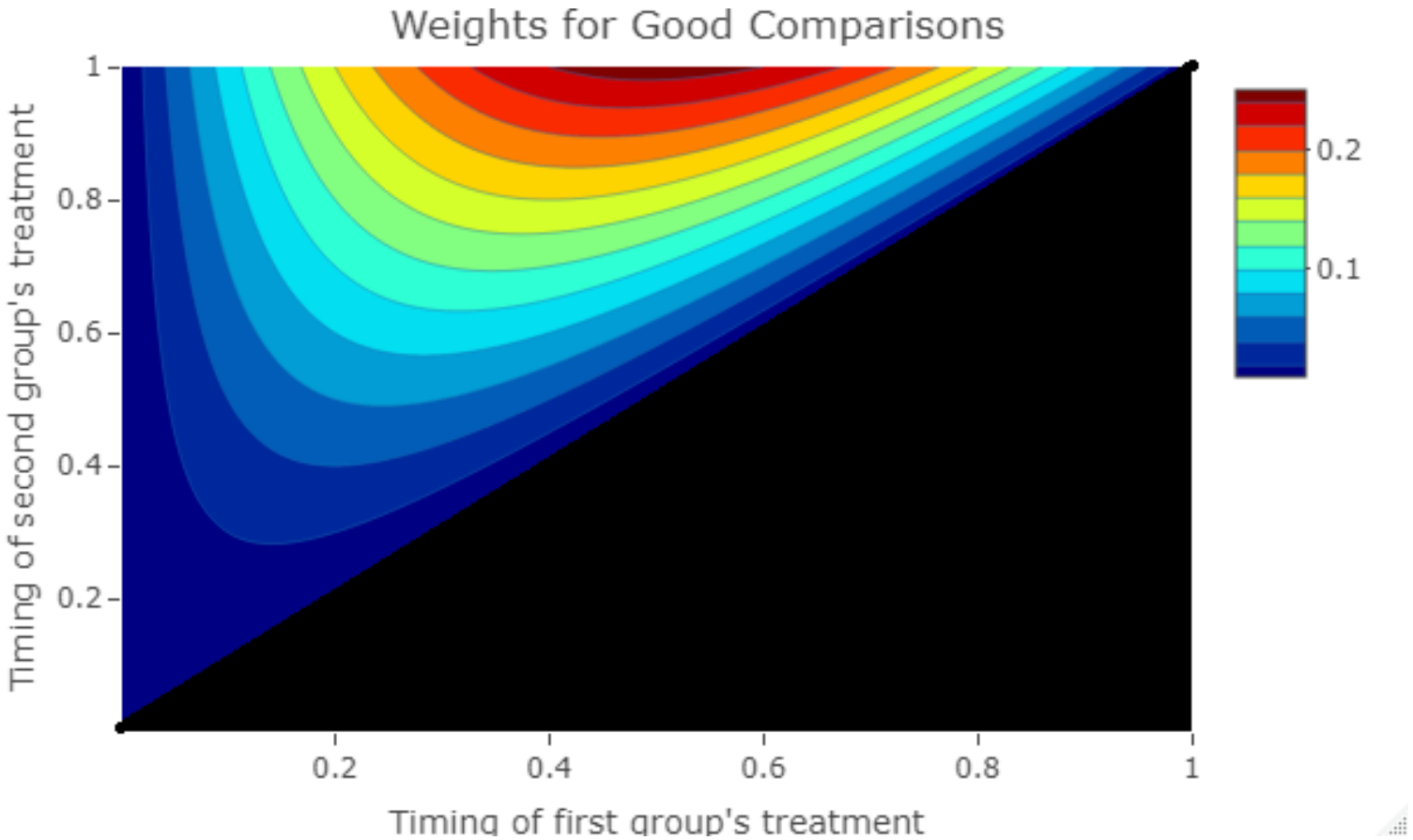
- $\Phi_{k\ell}^{good}$ *is the share of observations in the subsample used for this* 2×2 *comparison (as a fraction of the whole sample)*
- $\hat{V}_{k\ell}^{D,good}$ *is the variance of the fixed-effects adjusted* D *in the subsample used for this* 2×2 *comparison.*
- $\hat{V}^D$ *is the variance of the fixed-effects adjusted* D *across the entire sample*

The same decomposition holds for $s_{k\ell}^{\text{bad}}$ and s_k^{control}.

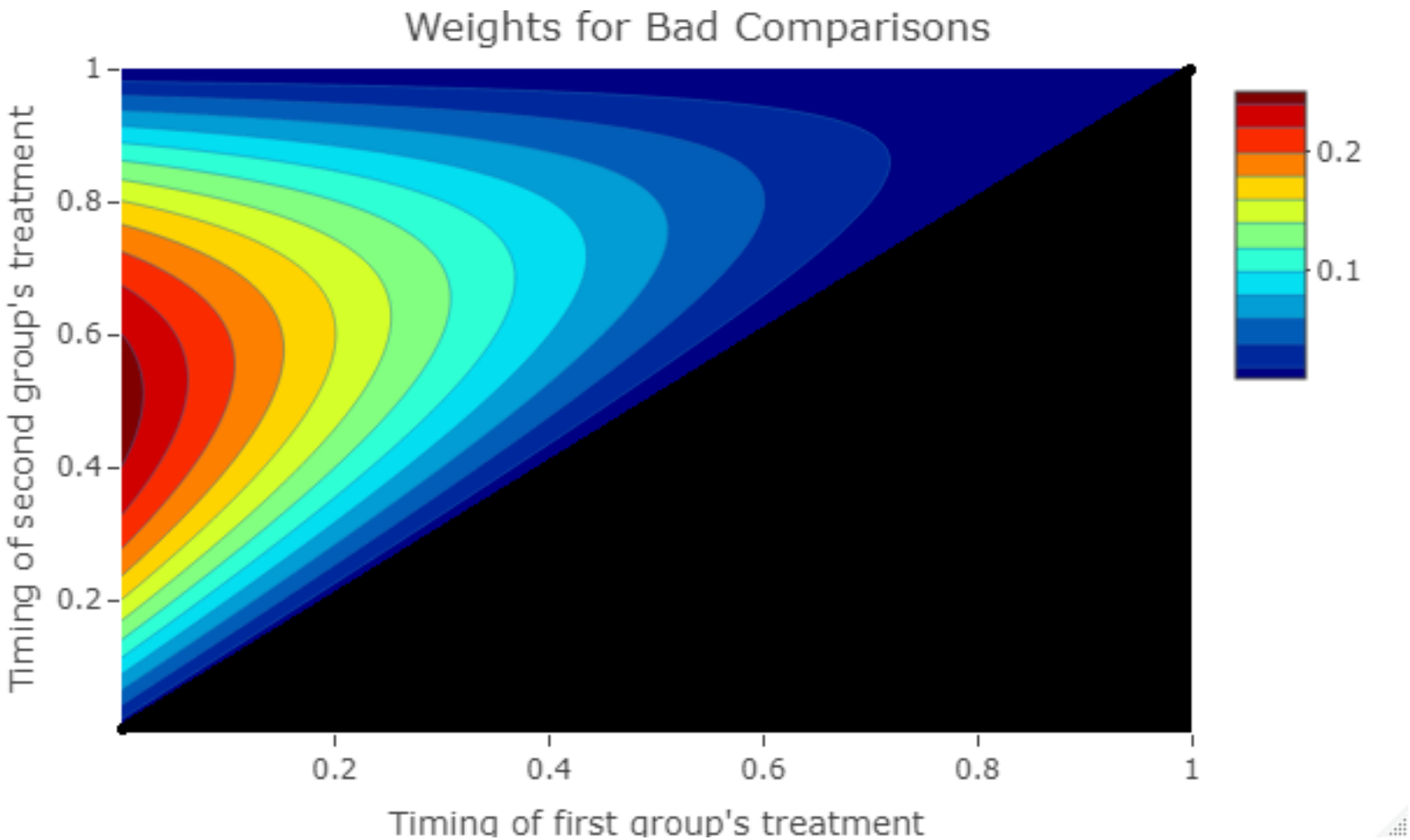
Weights for 2 × 2 Comparisons with the Control Group



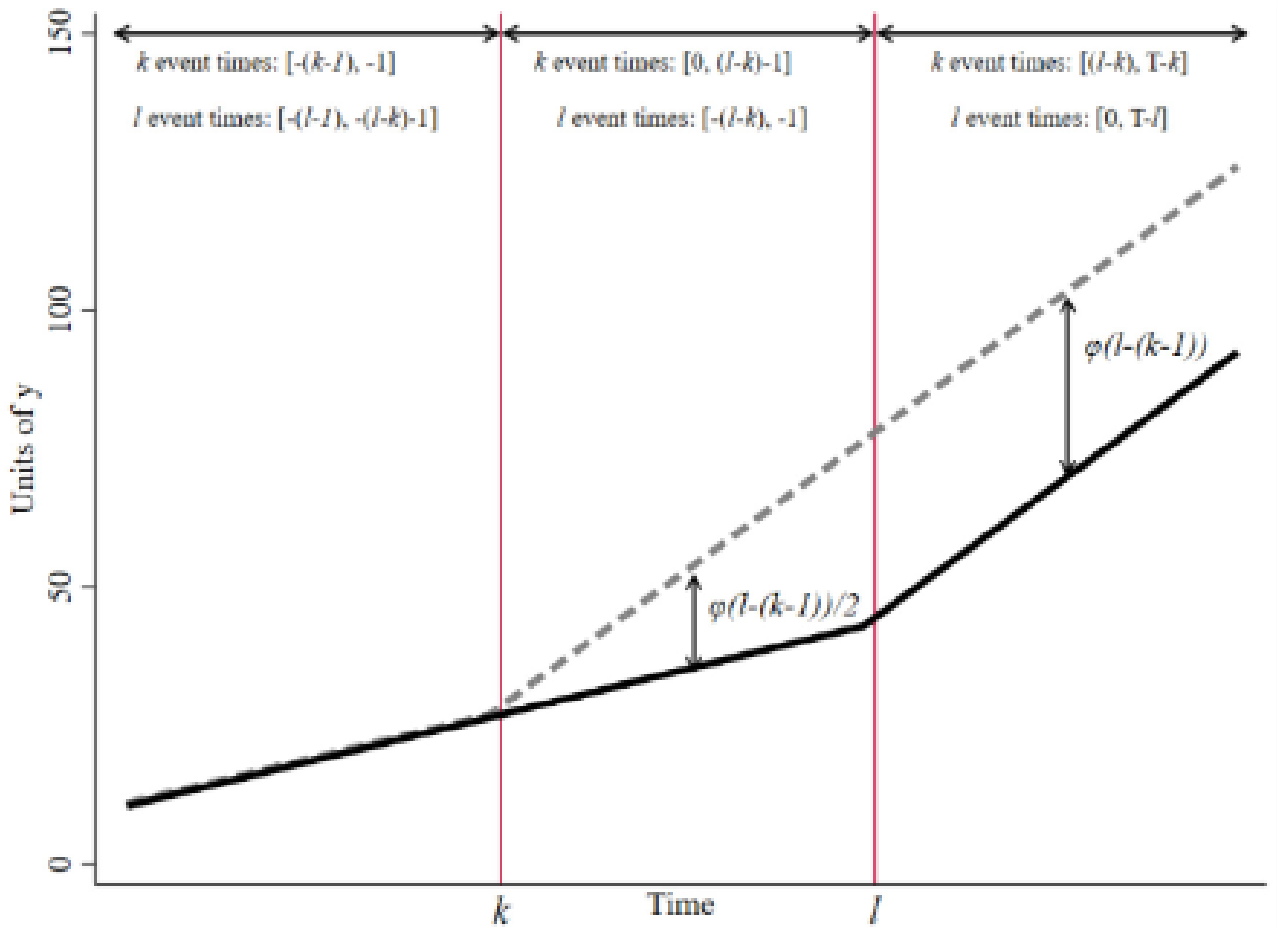
Weights for "Good" 2 × 2 Comparisons Between Treated Groups



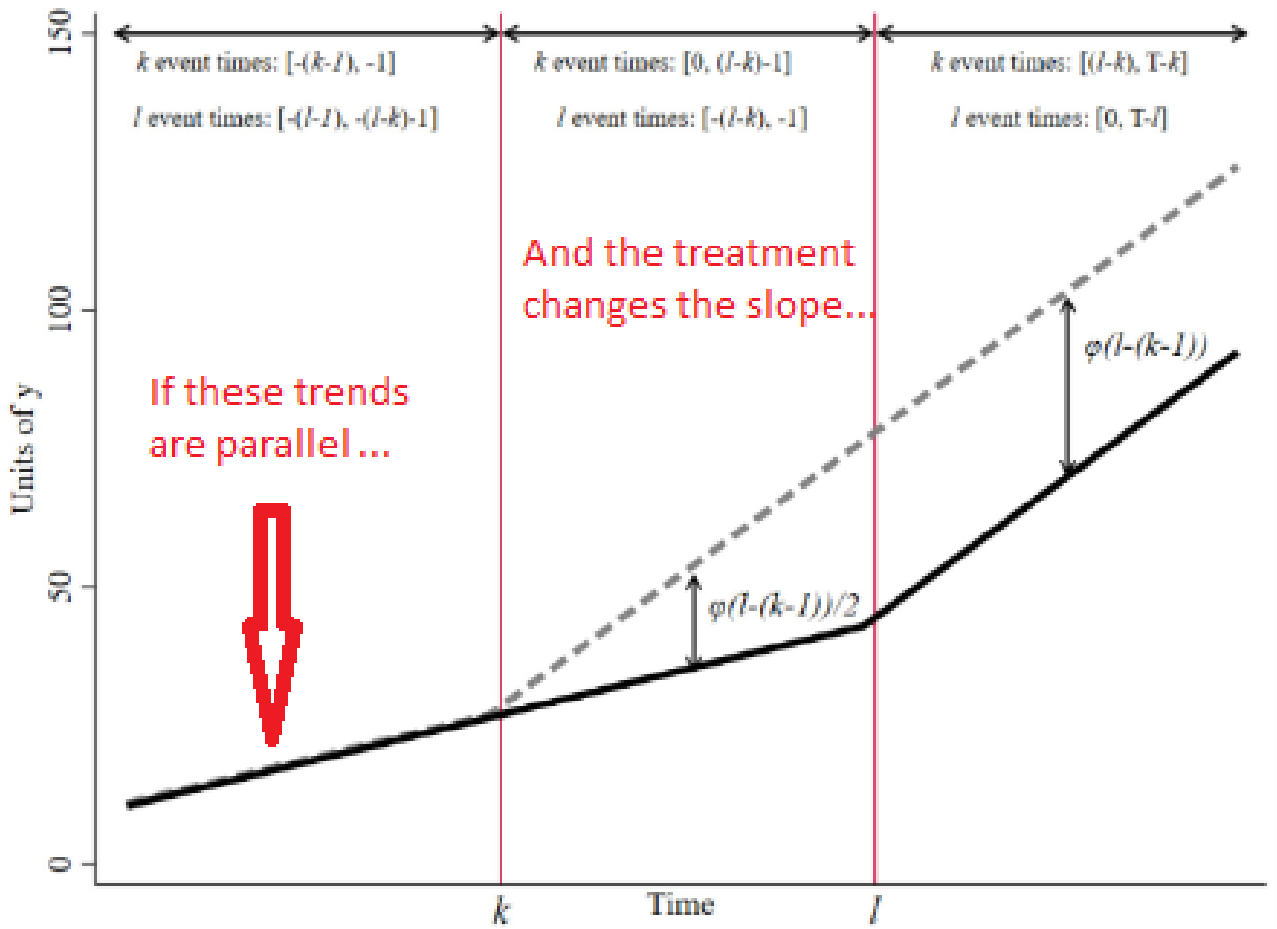
Weights for "Bad" 2 × 2 Comparisons Between Treated Groups



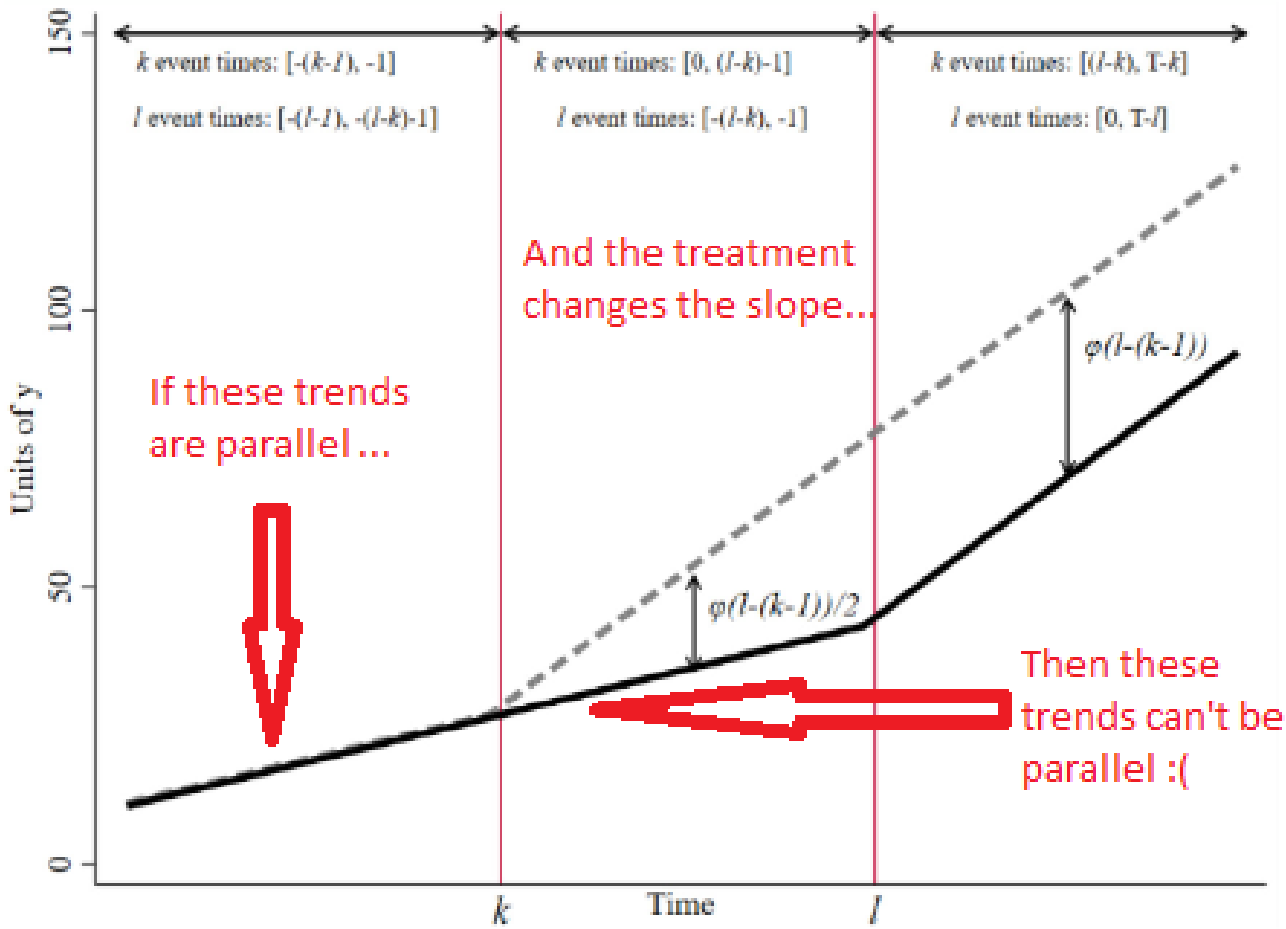
What makes the bad comparisons bad?



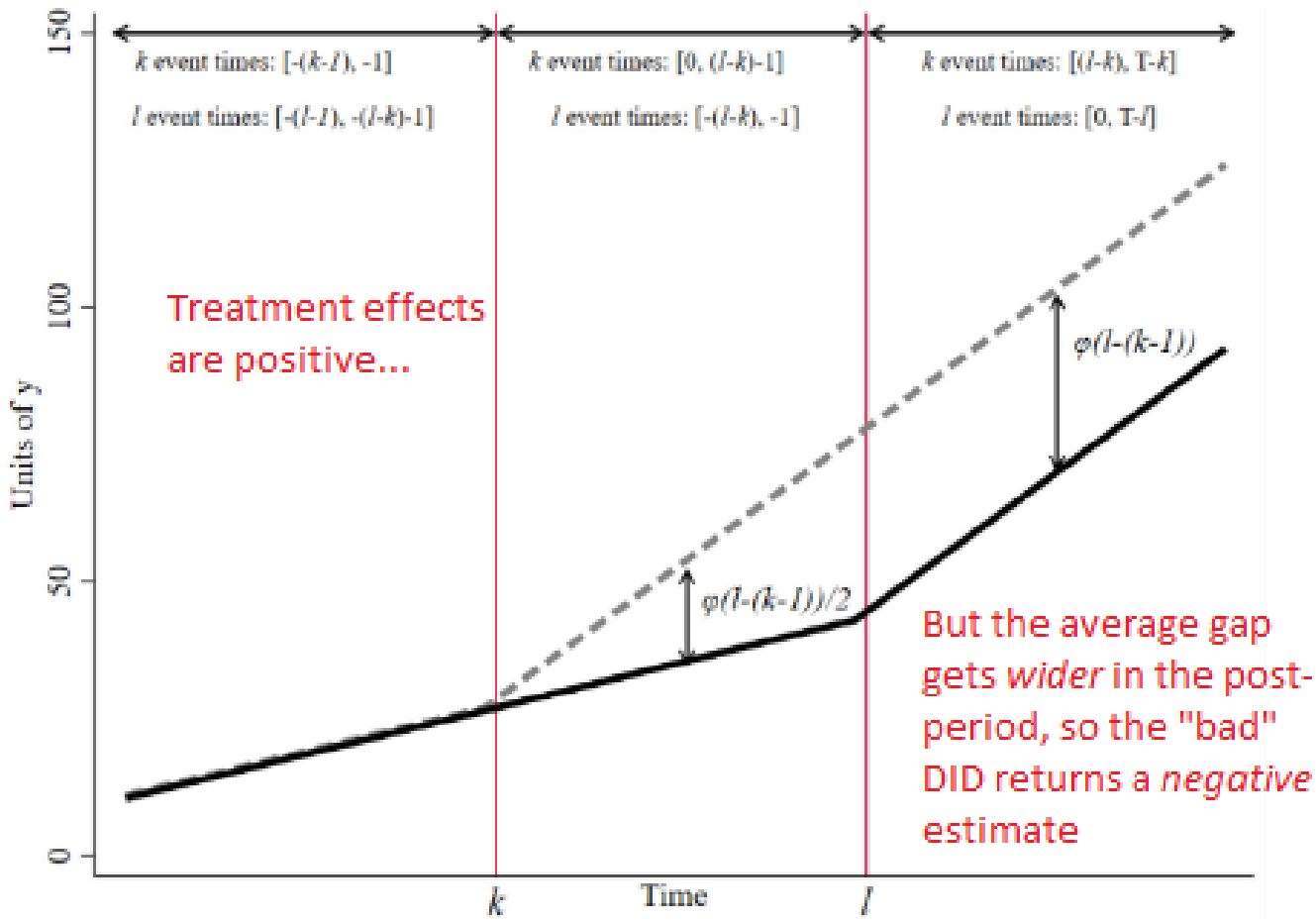
What makes the bad comparisons bad?



What makes the bad comparisons bad?



What makes the bad comparisons bad?



What are we estimating?

For groups k, ℓ, define

- $POST(k)$: the time period when k is treated
- $MID(k, \ell)$: the time period when k is treated but ℓ is not.

The closest thing we have to a policy-relevant parameter is the Variance-Weighted Average Treatment on the Treated:

$$VWATT = \sum_{k \neq U} \sigma_k^{\text{control}} ATT_k^{POST(k)} + \sum_{k \neq U} \sum_{\ell > k} \sigma_{k\ell}^{\text{good}} ATT_k^{MID(k,\ell)} + \sigma_{k\ell}^{\text{bad}} ATT_k^{POST(\ell)},$$

where the $\sigma_{k\ell}$ are population versions of the $s_{k\ell}$ and ATT_k^T denotes the average treatment effect on unit k over period T.

What are we estimating?

For groups k, ℓ, define

- $POST(k)$: the time period when k is treated
- $MID(k, \ell)$: the time period when k is treated but ℓ is not.

The closest thing we have to a policy-relevant parameter is the Variance-Weighted Average Treatment on the Treated:

$$VWATT = \sum_{k \neq U} \sigma_k^{\text{control}} ATT_k^{POST(k)} + \sum_{k \neq U} \sum_{\ell > k} \sigma_{k\ell}^{\text{good}} ATT_k^{MID(k,\ell)} + \sigma_{k\ell}^{\text{bad}} ATT_k^{POST(\ell)},$$

where the $\sigma_{k\ell}$ are population versions of the $s_{k\ell}$ and ATT_k^T denotes the average treatment effect on unit k over period T.

- (!!!) Different lengths of time. Not clear what we'd do with "the average of the policy's effect in one year in California and the policy's effect over fifty years in Iowa"

TWFE is a biased estimator of VWATT

Recall that the "bad" comparisons don't estimate their ATTs correctly if treatment effects change over time. This biases the whole estimator:

Theorem

Assume parallel trends on the $Y(0)$s. Then, the TWFE estimand converges in probability to

$$VWATT - \Delta ATT,$$

where ΔATT is an error term given by

$$\Delta ATT = \sum_{k \neq U} \sum_{\ell > k} \sigma^{bad}_{k,\ell} \left(ATT_k^{POST(\ell)} - ATT_k^{MID(k,\ell)} \right)$$

Weights in an Empirical Example

(I stole this example from Andrew Baker's slides on DiD)

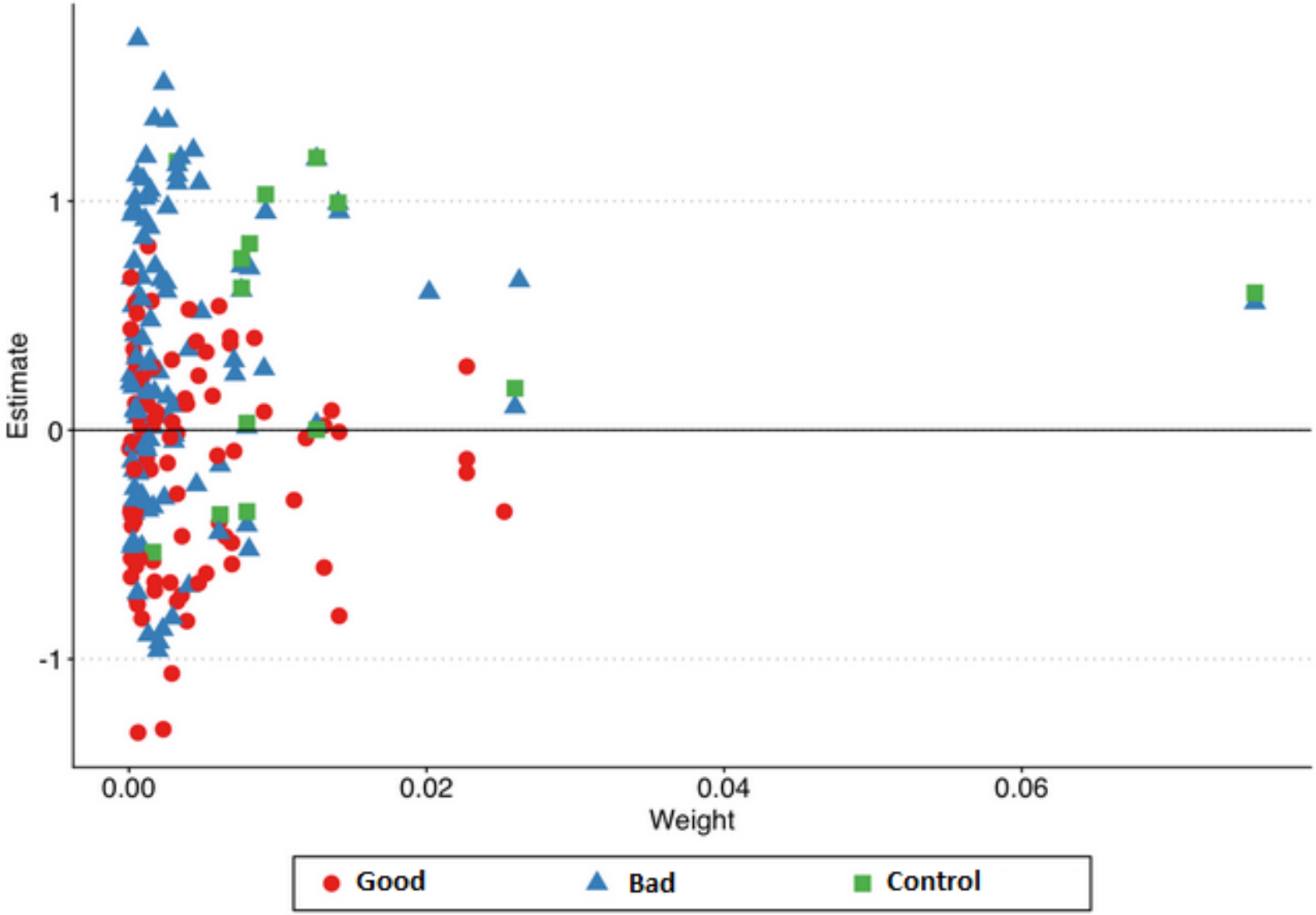
- Bachhuber et al. (2014) use TWFE, find that state-level medical marijuana laws decrease opioid deaths over 1999-2010.
- Shover et al (2020) expand data through 2017, find that the sign flips completely

Two competing explanations:

- Result is just fragile, or original was p-hacked
- Many states passed medical marijuana laws in the 2010-2017 period. A much larger fraction of the 2×2 comparisons in the TWFE are now "bad."

Weights in an Empirical Example

Positive sign in the 2020 study seems to be driven by "bad" estimates:



Covariates lead to more comparisons

In practice, people often run the regression

$$Y_{it} = \alpha_i + \gamma_t + \beta X_{it} + \beta^{DD} D_{it} + \varepsilon_{it},$$

with some (potentially time-varying) covariates X_{it}.

Theorem (Frisch–Waugh–Lovell aka "partialling")

This is equivalent to running the regression

$$\tilde{Y}_{it} = \beta^{DD} \tilde{D}_{it} + \tilde{\varepsilon}_{it},$$

where $\tilde{Y}$ is the residual of a regression on X_{it} with unit and time fixed effects.

Concern: two units with the same D_{it} can have different values of $\tilde{D}_{it}$, so the resulting value of β^{DD} is affected by hard-to-interpret *within-group comparisons.*

Covariates lead to more comparisons

Theorem

The treatment coefficient β^{DD} from a TWFE regression with linear controls can be written as

$$\beta^{DD} = \Omega\beta^{within} + (1 - \Omega)\beta^{between},$$

where $\beta^{between}$ comes from 2×2 regressions comparing groups with different treatment timings and β^{within} compares groups with the same treatment timing.

- If $\mathbb{E}[Y(d)|X = x]$ is in fact linear in x, this can be useful. Otherwise, difficult to interpret β^{within}.
- In an empirical example, Goodman-Bacon finds that including controls leads to a smaller (in magnitude) estimate. 73% of this change comes from new "within" comparisons, 22% from individual 2×2 comparisons, 5% from changes in weights.

Covariates: an example

Consider a collection of three groups, with a single continuous covariate. For simplicity, say Y depends on D but not t.

Group	D		X		Y	
	t = 0	t = 1	t = 0	t = 1	D = 0	D = 1
A	0	0	1	1	0	0
B	0	1	0	0	0	1
C	0	1	1	x_C	0	y_C

The ATT is $\frac{1+y_c}{2}$, and TWFE without covariates correctly identifies this. If we include X as a linear covariate, TWFE gives:

$$\beta_{dd} = \underbrace{\frac{1}{2} \cdot 1 + \frac{1}{2} \cdot y_C}_{\text{ATT}} - \underbrace{\frac{1}{2} \cdot \frac{x_c}{x_c - 2} \cdot (y_c - 1)}_{\text{“Within” comparison of B and C}}$$

Conclusion

- TWFE is a weighted average of "good", "bad", and "control" estimators.
  - These weights are all positive, but related to both group size and treatment timing.
  - Workaround: use these estimators directly and weight them however you like.
- "Bad" estimators use already-treated groups as a control group.
  - This introduces bias when treatment effects vary with time.
  - Can have the wrong sign in pretty reasonable situations.
- Time-varying covariates make the results of TWFE regression even harder to interpret
  - Should these be controlled for at all? (Post-treatment bias a real concern!)
  - Even without staggered adoption, introduce weird comparisons if population effects not actually linear.