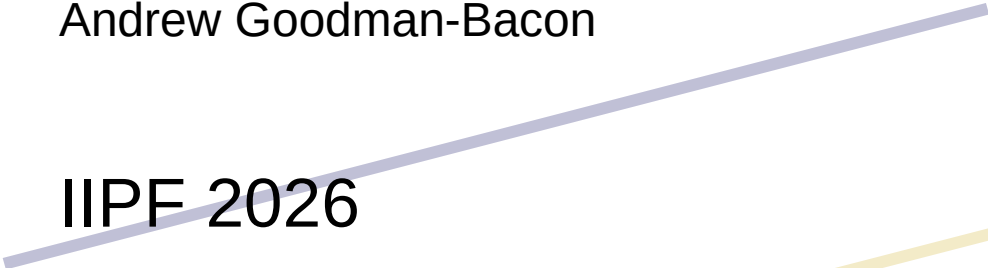
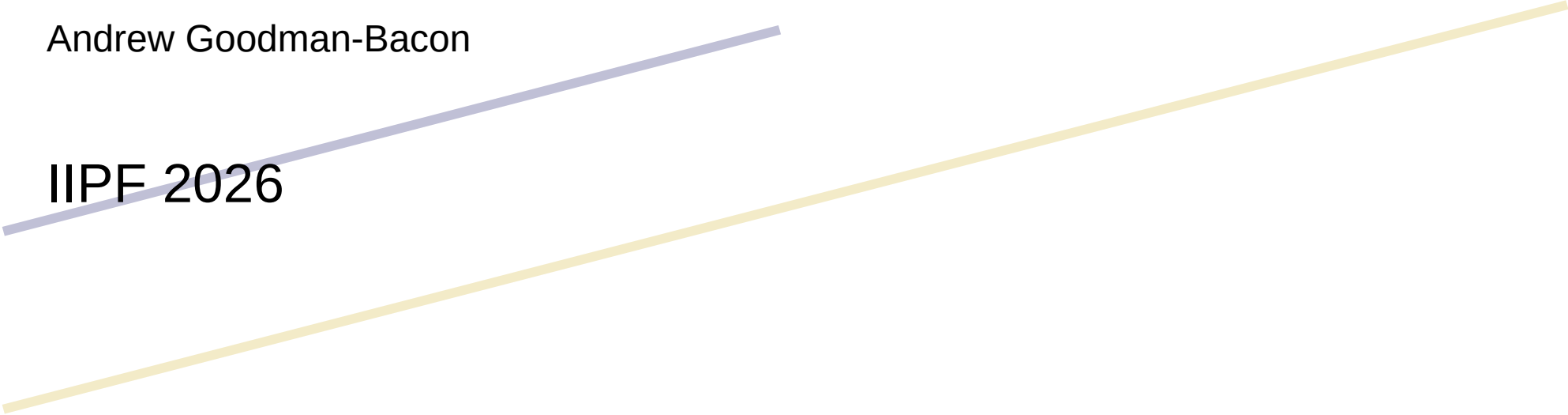


New Developments in Difference-in-Differences: *Basics* + *Covariates*

Andrew Goodman-Bacon



IIPF 2026



What is difference-in-differences?

Difference 1

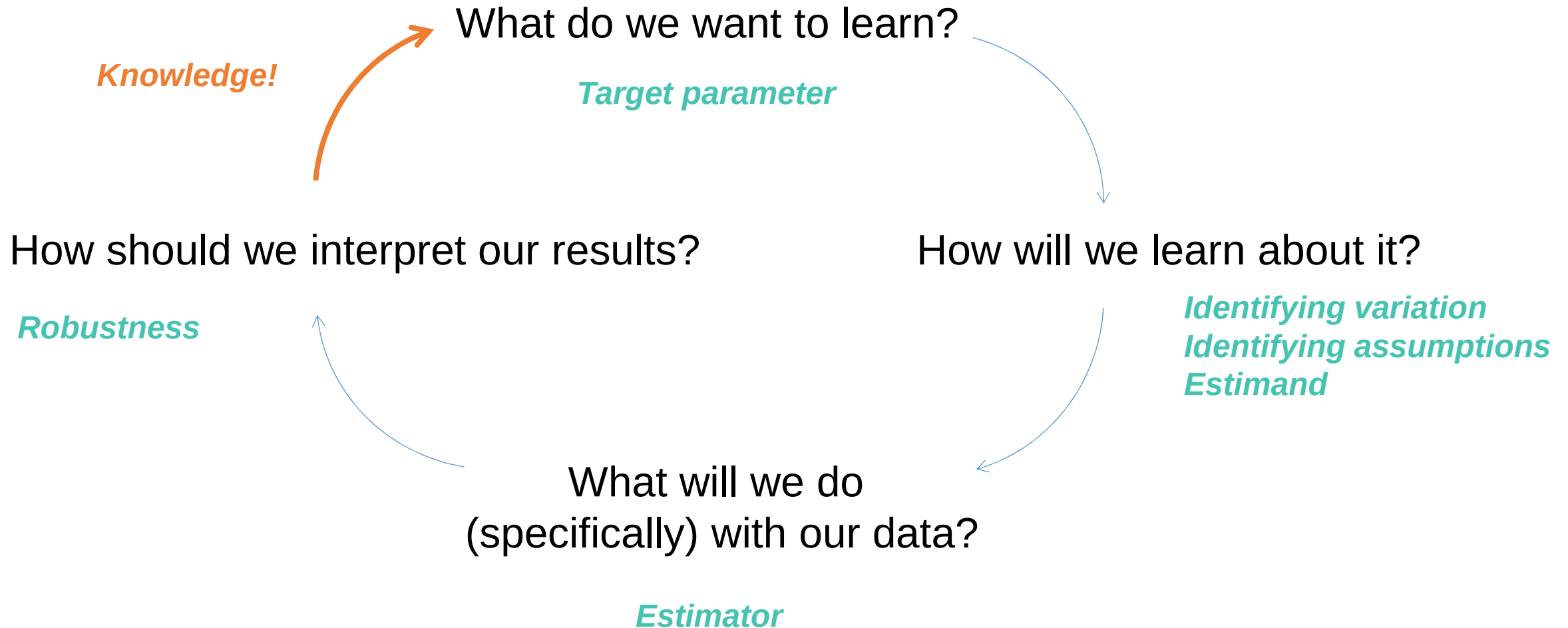
A comparison of outcome changes between two periods and between two groups with different changes in treatment status.

Difference 2

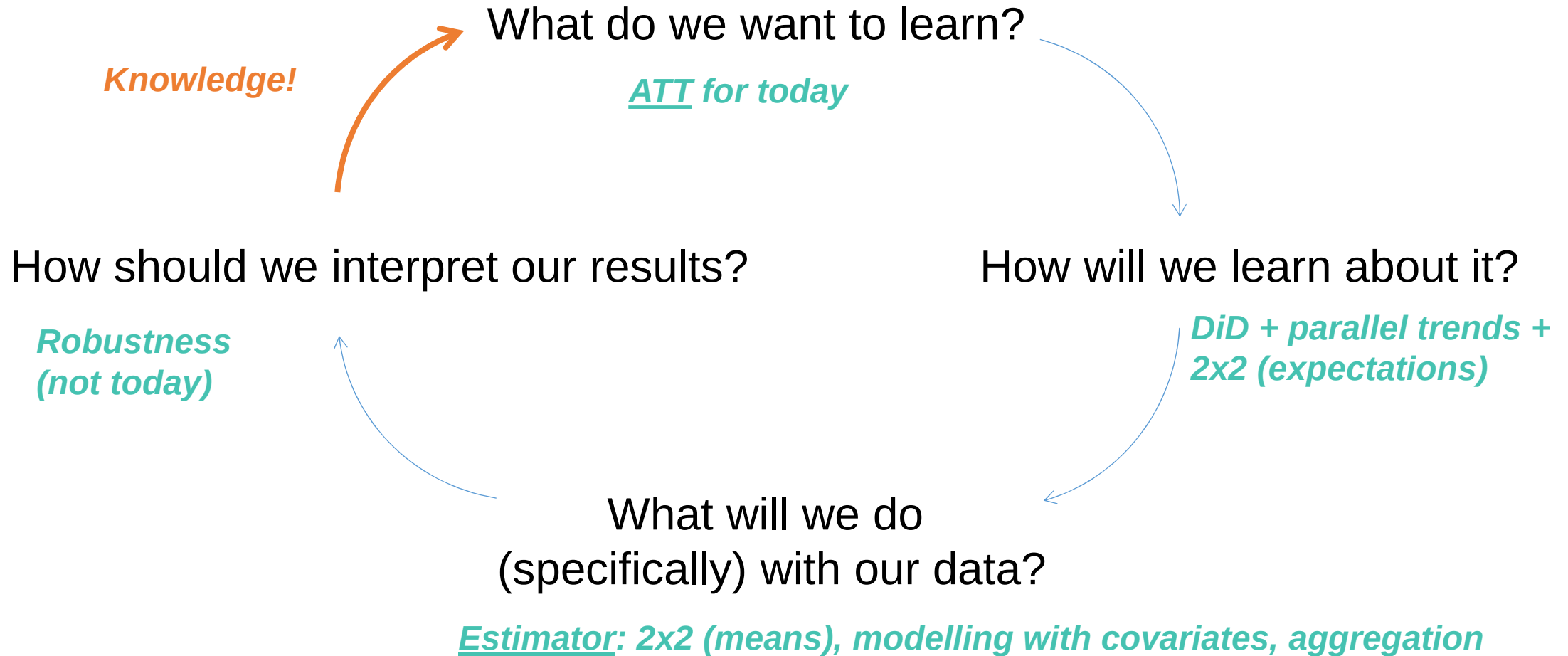
Broad goals

1. Understand a “forward engineering” causal inference approach
2. See it applied to several difference-in-differences designs
3. Confidence when things are slightly different in *your* work

A causal inference blueprint



A difference-in-differences blueprint



Course Plan

1. 2x2 DiD (average treatment effects on the treated: **ATT**)

1. Basic building block (two periods, two groups) ←
2. Conditioning on covariates ←

2. GxT DiD (staggered binary treatment: **aggregated ATT**)

1. Break down into building blocks ←
2. **Aggregation**

3. dX2 DiD (continuous treatment: **ATT or ACR**)

1. Break down into (new) building blocks ←
2. **Aggregation**

- 
1. Target parameter
 2. Identification
+ estimation
 3. Reporting

What do you want to know?

The Effects of a Large-Scale Mental Health Reform: Evidence from Brazil[†]

By MATEUS DIAS AND LUIZ FELIPE FONTES*

Despite broad consensus that access to mental health care is currently inadequate, debate over the best methods to provide such care continues. In the second half of the twentieth century, many countries transitioned from a model of mental health care centered on psychiatric hospitals to one based on community care. However, the effects of this change are far from clear.

General set up

Units:

$$i = 1, \dots, N$$

Time:

$$t = \{0, 1\}$$

Treatment status:

$$D_{i,t} = 1 \text{ if } i \text{ is treated in } t$$

Treatment path:

$$\mathbf{D}_i = \{D_{i,0}, D_{i,1}\}$$

Assume $D_{i,0} = 0$ for all $i \rightarrow$ only two paths: $\{0,1\}$ and $\{0,0\}$

Treatment group:

$$D_i = 1 \text{ if } i \text{ is "ever" treated}$$

$$D_i \equiv 1\{\mathbf{D}_i = \{0,1\}\}$$

Treatment time:

$$g = 1$$

$$D_{it} \equiv D_i \times 1\{t \geq g\}$$

Outcome:

$$Y_{i,t}$$

The Effects of a Large-Scale Mental Health Reform: Evidence from Brazil[†]

Psychosocial Care Centers

Centros de Atenção Psicossocial-

By MATEUS DIAS AND LUIZ FELIPE FONTES*

g D_i i Y_{it}
We assess the effects of introducing a CAPS in a municipality on homicides, :

Brazilian set up

Units:

$$i = 1, \dots, N$$

Time (years):

$$t = \{2005, 2007\}$$

Treatment status:

$$D_{i,t} = 1 \text{ if } i \text{ has a CAPS in } t$$

Treatment path:

$$\mathbf{D}_i = \{D_{i,2005}, D_{i,2007}\}$$

Assume no CAPS in 2005
→ start in 2006 or not

Treatment group:

$$D_i = 1 \text{ if } i \text{ gets a CAPS in 2006}$$

$$D_i \equiv 1\{\mathbf{D}_i = \{0, 1\}\}$$

Treatment time:

$$g = 2006$$

$$D_{it} \equiv D_i \times 1\{t \geq 2006\}$$

Outcome:

Homicides per 10,000 in municipality i in year t

Potential and observed outcomes

$Y_{i,t}(D_i)$ outcome of unit i in period t under treatment D_i

$$Y_{i,t} = \begin{cases} Y_{i,t}(1) & \text{if } D_i = 1 \text{ and } t = 2007 \\ Y_{i,t}(0) & \text{if } D_i = 0 \text{ or } t = 2005 \end{cases}$$

	$D_i = 0$	$D_i = 1$
$t = 2005$	$Y_{i,t}(0)$	$Y_{i,t}(0)$
$t = 2007$	$Y_{i,t}(0)$	$Y_{i,t}(1)$

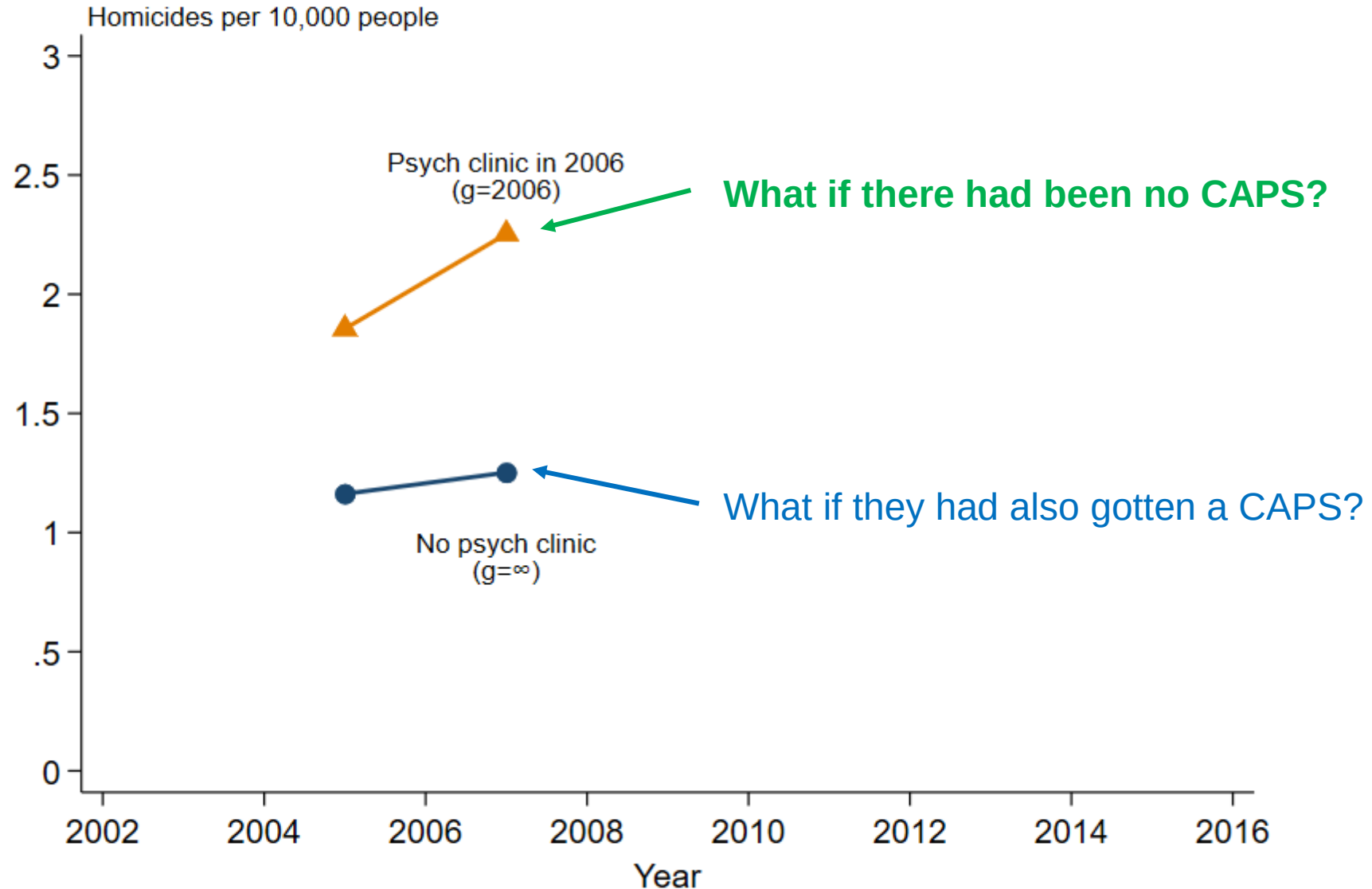
No Anticipation:

We would see this homicide rate even if there had been no CAPS

Stable Unit Treatment Value Assumption (SUTVA):

Only a municipality's *own* treatment status matters (no spillovers!)

Observed outcomes and causal questions



Causal effects and target parameters

But **today** we choose to care about (*target*) “on the treated” effects

Causal effects for untreated units exist and make sense!

	$D_i = 0$	$D_i = 1$
$t = 2005$	$Y_{i,t}(0)$	$Y_{i,t}(0)$
$t = 2007$	$Y_{i,t}(1) - Y_{i,t}(0)$	$Y_{i,t}(1) - Y_{i,t}(0)$

And we often talk about effects for the whole population (RCTs!)

But what “effect”?

The Effects of a Large-Scale Mental Health Reform: Evidence from Brazil[†]

Psychosocial Care Centers

Centros de Atenção Psicossocial-

By MATEUS DIAS AND LUIZ FELIPE FONTES*

We assess the **effects** of introducing a CAPS in a municipality on ... homicides, following the criminology literature that reports a correlation between deinstitutionalization and community-based care with violent crimes (Lamb 2015).

Average Treatment Effect on the Treated

$$ATT = E[Y_{i,t}(1) - Y_{i,t}(0) | D_i = 1]$$

This is a **choice**!

“Want” vs “Can”

But what “effect”?

The Effects of a Large-Scale Mental Health Reform: Evidence from Brazil[†]

Psychosocial Care Centers

Centros de Atenção Psicossocial-

By MATEUS DIAS AND LUIZ FELIPE FONTES*

We assess the **effects** of introducing a CAPS in a municipality on ... homicides, following the criminology literature that reports a correlation between deinstitutionalization and community-based care with violent crimes (Lamb 2015).

Average Treatment Effect on the Treated

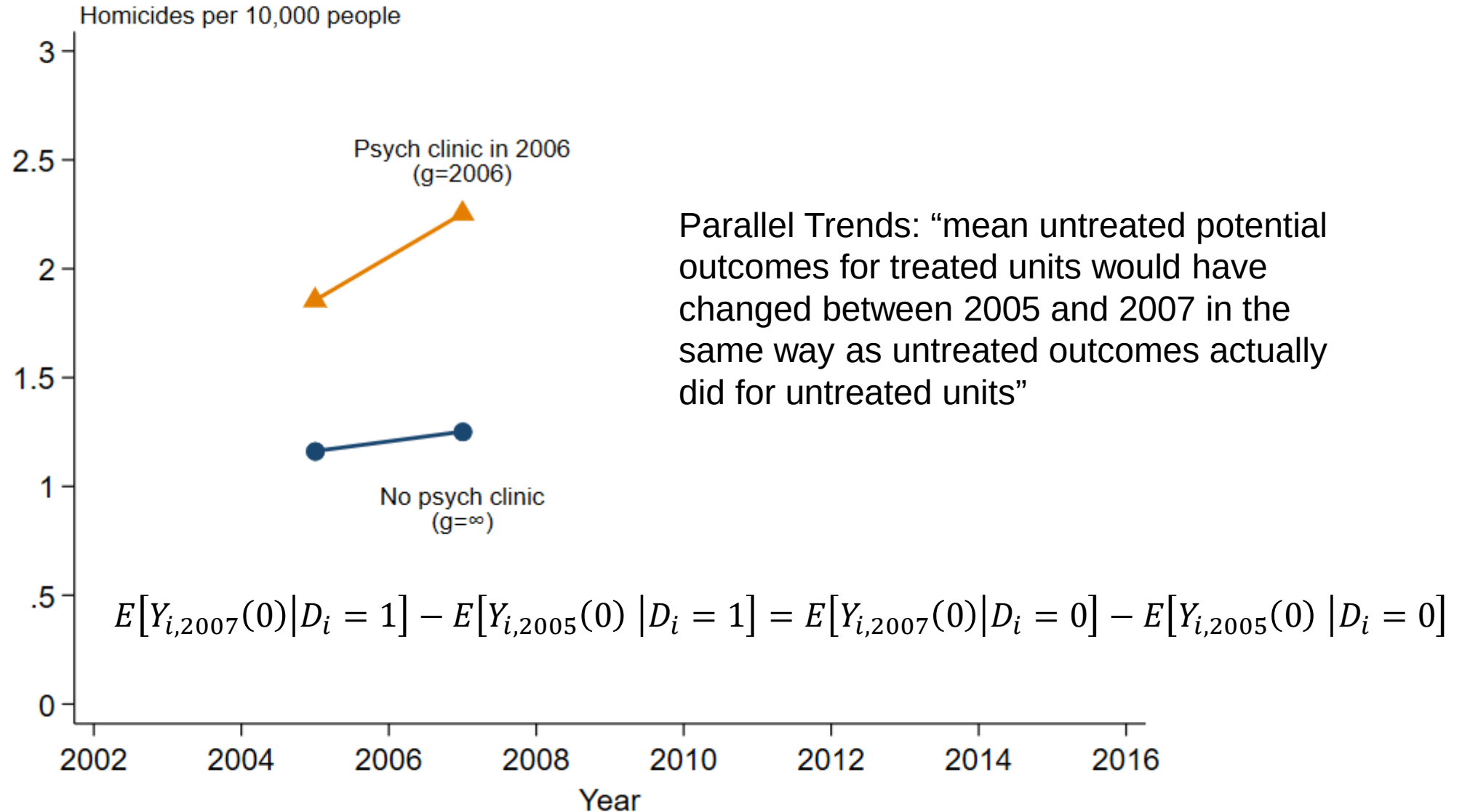
$$ATT = E[Y_{i,t}(1) - Y_{i,t}(0) | D_i = 1]$$

ATT Counterfactual:

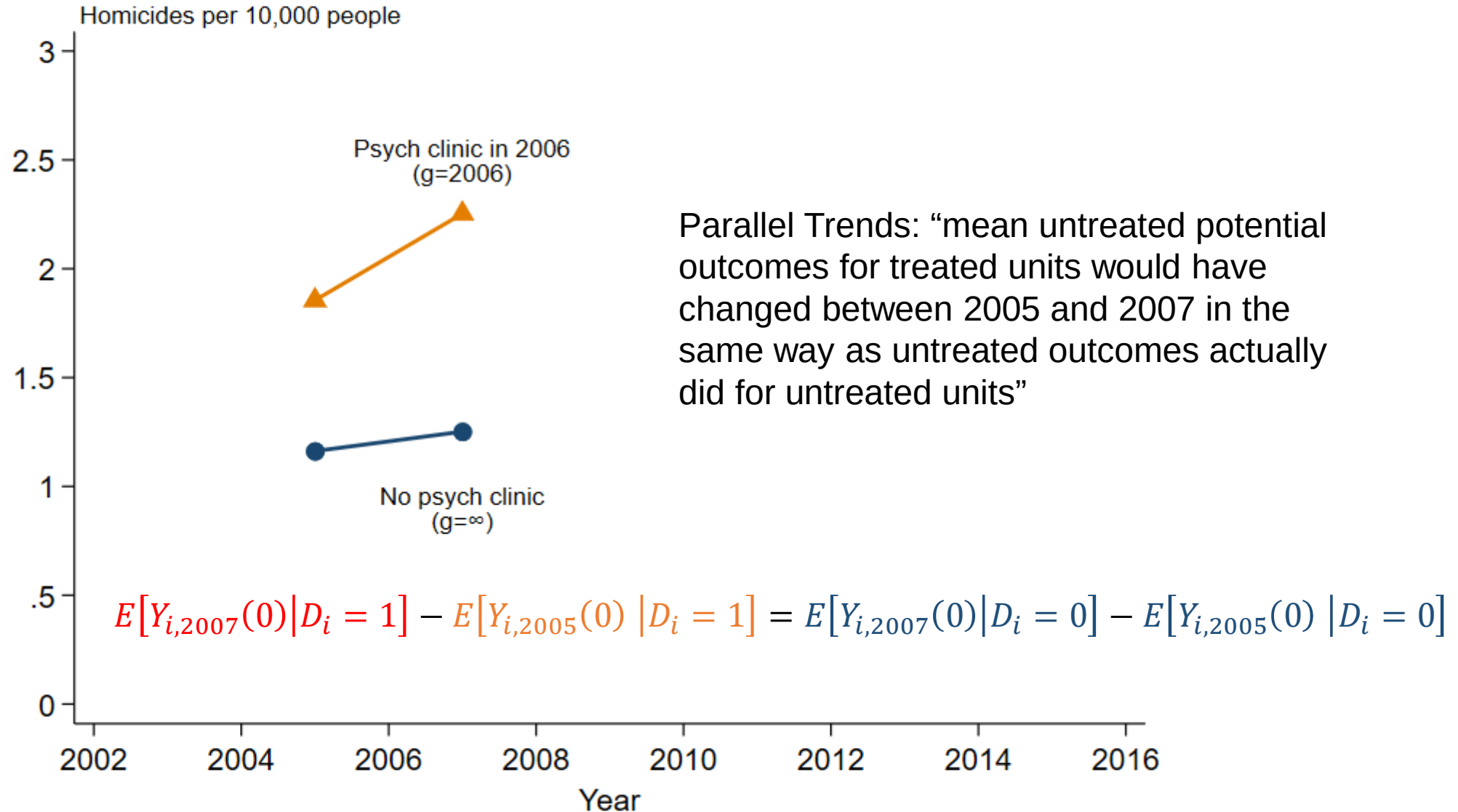
What would mean outcomes among treated units have been had they not been treated?

What are you willing to believe?

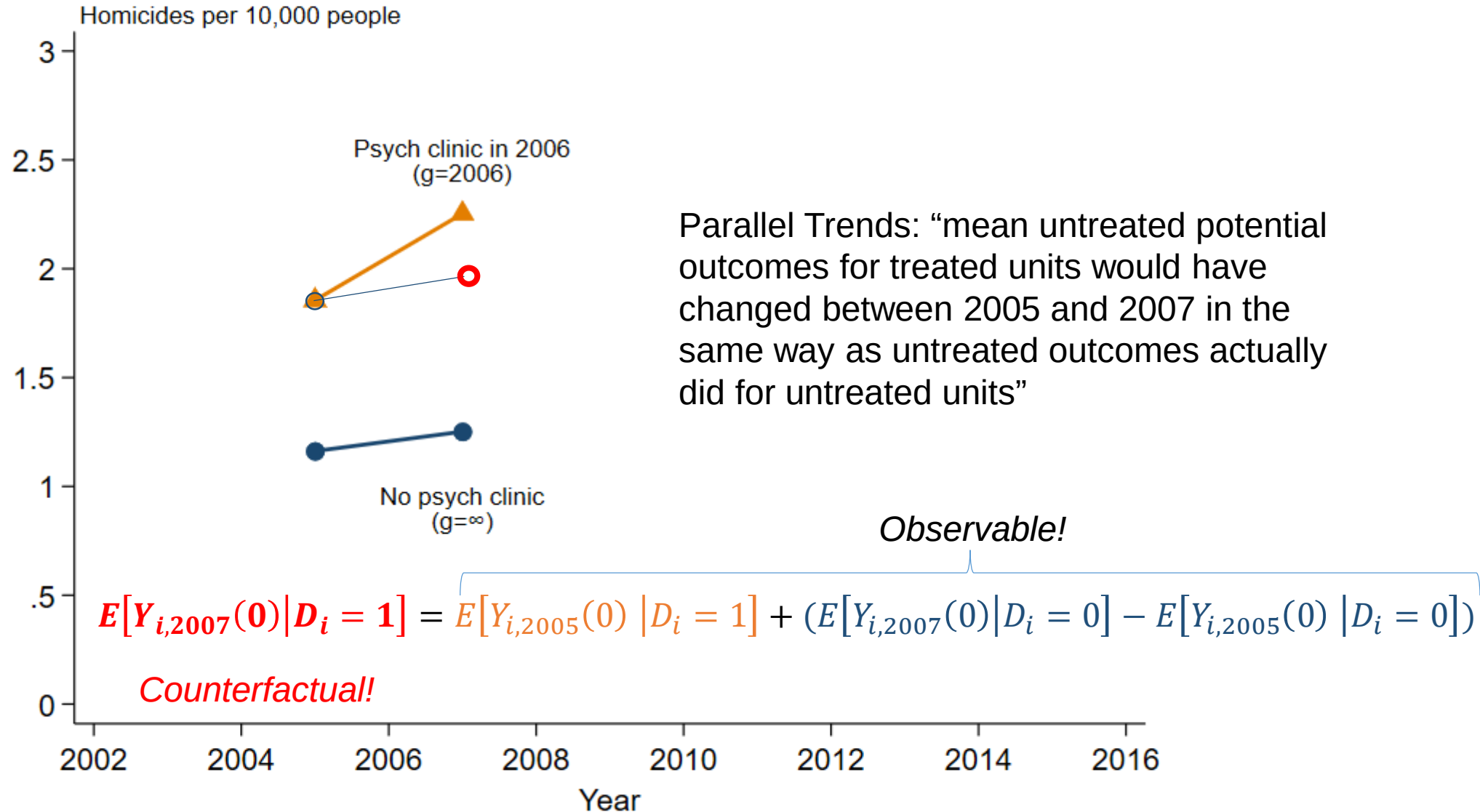
“The” parallel trends assumption (2x2)



“The” parallel trends assumption (2x2)

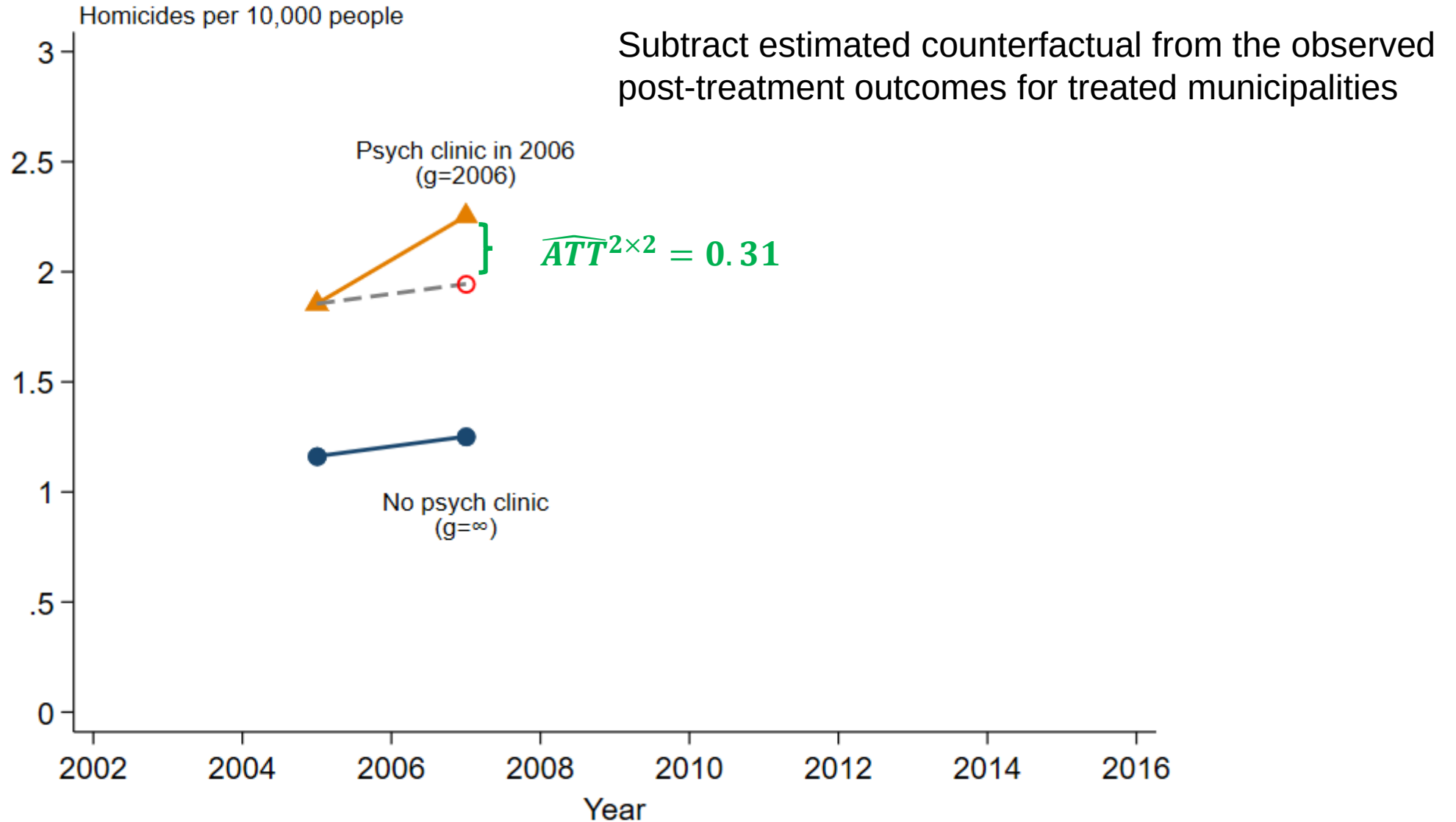


“The” parallel trends assumption (2x2)



What should you actually do?

The 2x2 DiD estimand



The 2x2 DiD estimand/estimator

Define target parameter (ATT)

$$ATT = E[Y_{i,2007}(1)|D_i = 1] - E[Y_{i,2007}(0)|D_i = 1]$$

Impose identifying assumption (PT)

$$E[Y_{i,2007}(0)|D_i = 1] - E[Y_{i,2005}(0)|D_i = 1] = E[Y_{i,2007}(0)|D_i = 0] - E[Y_{i,2005}(0)|D_i = 0]$$

Get the DiD estimand (2x2)

$$ATT^{2 \times 2} = E[\Delta Y_{i,2007}|D_i = 1] - E[\Delta Y_{i,2007}(0)|D_i = 0]$$

Population objects → sample analogues → DiD estimator (2x2):

$$\widehat{ATT}^{2 \times 2} = \left(\bar{Y}_{2007}^{D_i=1} - \bar{Y}_{2005}^{D_i=1} \right) - \left(\bar{Y}_{2007}^{D_i=0} - \bar{Y}_{2005}^{D_i=0} \right)$$

Where $\bar{Y}_t^{D_i=d} \equiv \frac{\sum_i 1\{D_i=d\}Y_{it}}{\sum_i 1\{D_i=d\}}$

Two ways to get $\widehat{ATT}^{2 \times 2}$

$$\widehat{ATT}^{2 \times 2} = (\bar{Y}_{2007}^{D_i=1} - \bar{Y}_{2005}^{D_i=1}) - (\bar{Y}_{2007}^{D_i=0} - \bar{Y}_{2005}^{D_i=0})$$

1. By hand: easy, clear, forces you to look at the data, but no standard errors.
2. With OLS: also easy, regression standard errors.

$$Y_{it} = \hat{\alpha}_0 + \hat{\alpha}_1 D_i + \hat{\alpha}_2 1\{t = 2007\} + \hat{\beta}^{2 \times 2} D_i \times 1\{t = 2007\} + u_{it}$$

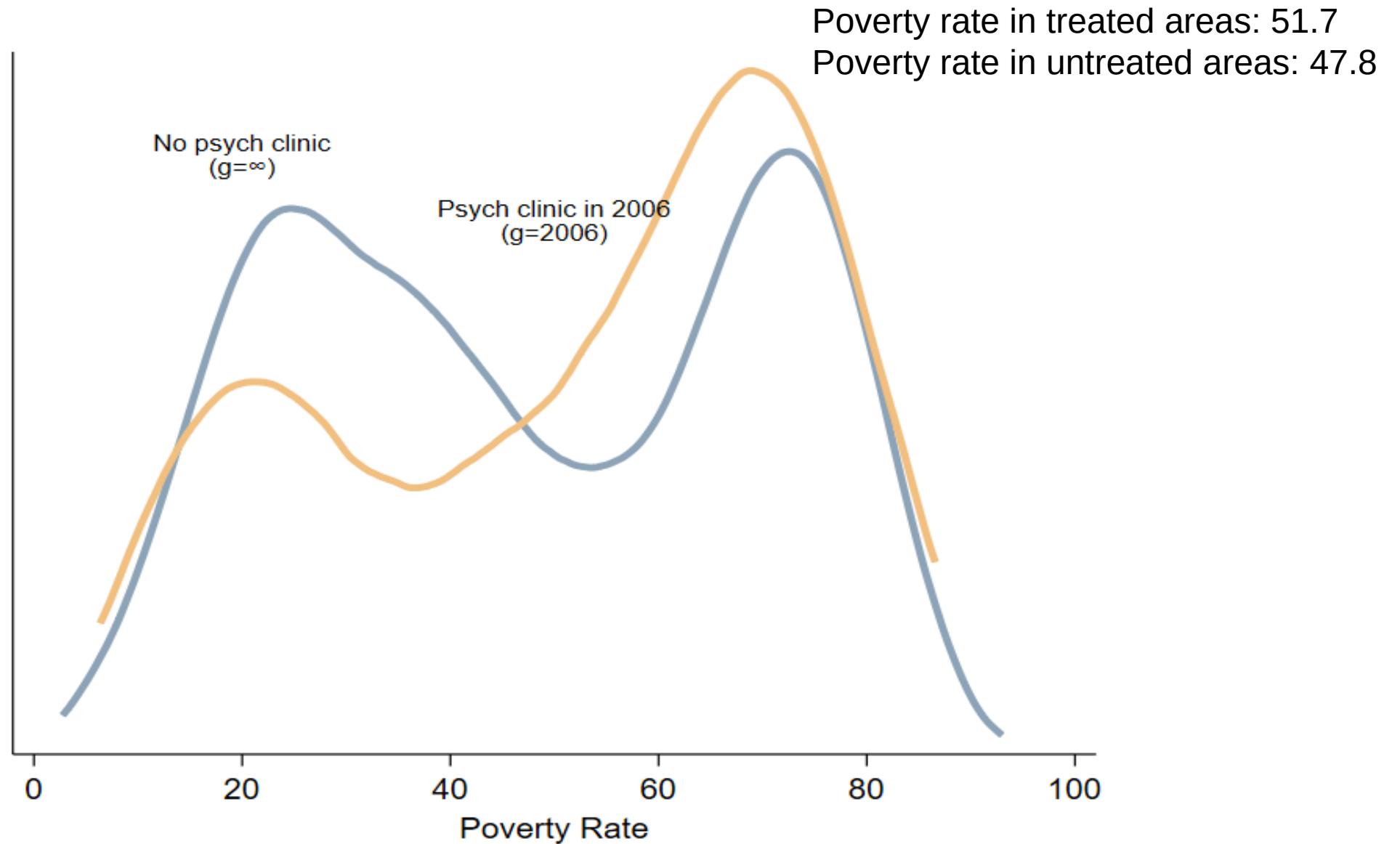
$$\hat{\beta}^{2 \times 2} = (\bar{Y}_{2007}^{D_i=1} - \bar{Y}_{2005}^{D_i=1}) - (\bar{Y}_{2007}^{D_i=0} - \bar{Y}_{2005}^{D_i=0})$$

But is it OK to alter this regression? Multiple periods? Complex treatments? Covariates?

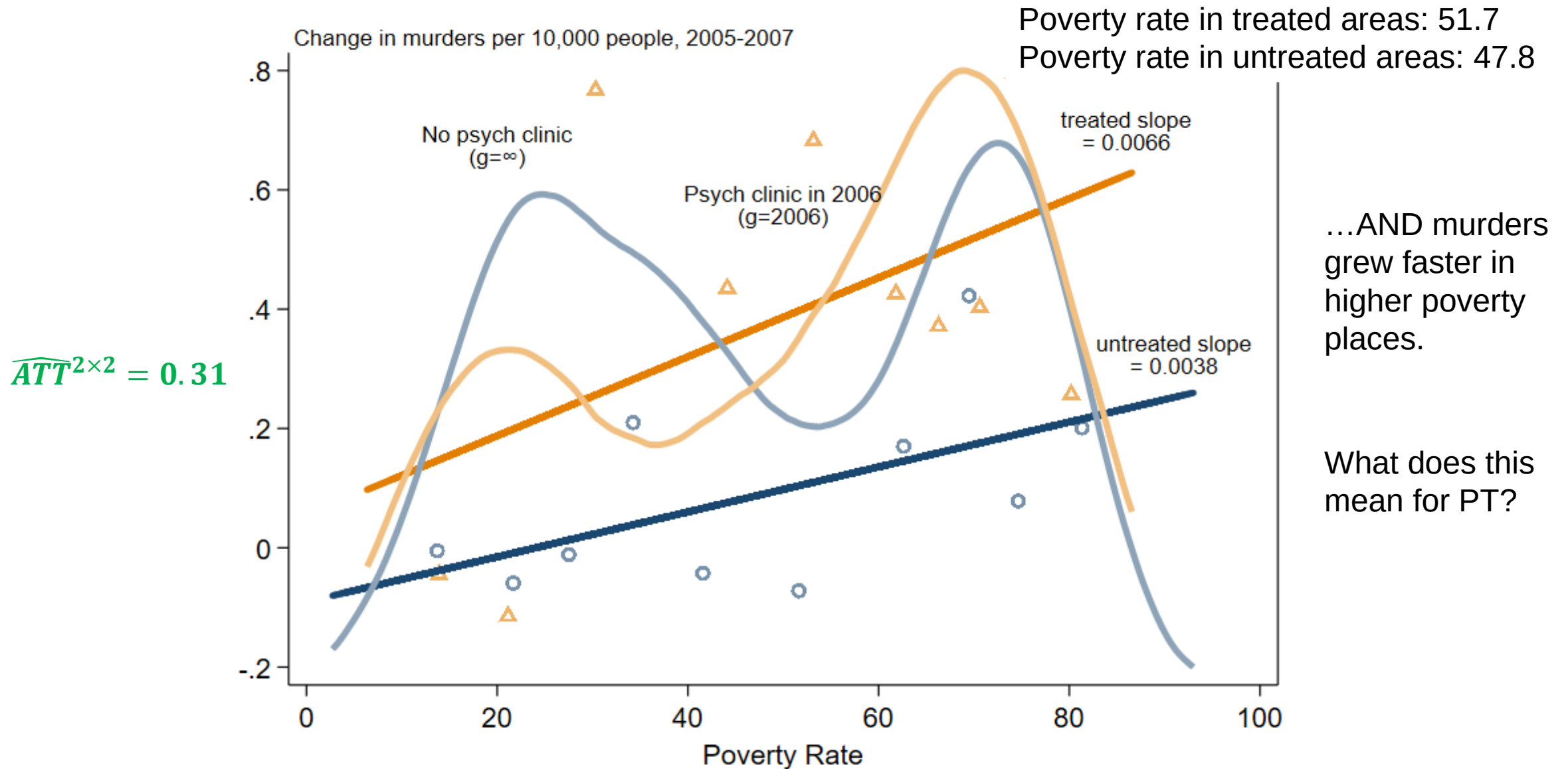
NO (in general)

Covariates (OLS)

Imbalance in poverty



Imbalance in poverty



Covariate subgroups underneath PT

Imagine that there were just “poor” ($X_i = 1$) and “not poor” ($X_i = 0$) municipalities. Then (by L.I.E.) PT is:

$$E[\Delta Y(0)|D_i = 1] = E[\Delta Y(0)|D_i = 0]$$

$$\begin{aligned} & E[\Delta Y(0)|D_i = 1, X_i = 1]P(X_i|D_i = 1) + \\ & E[\Delta Y(0)|D_i = 1, X_i = 0](1 - P(X_i|D_i = 1)) \end{aligned} = \begin{aligned} & E[\Delta Y(0)|D_i = 0, X_i = 1]P(X_i|D_i = 0) + \\ & E[\Delta Y(0)|D_i = 0, X_i = 0](1 - P(X_i|D_i = 0)) \end{aligned}$$

Covariate and treatment specific trends (“relevance”)

Treatment-specific weights (imbalance)

Covariate subgroups underneath PT

Imagine that there were just “poor” ($X_i = 1$) and “not poor” ($X_i = 0$) municipalities. Then (by L.I.E.) PT is:

$$E[\Delta Y(0)|D_i = 1] = E[\Delta Y(0)|D_i = 0]$$

$$\begin{aligned} & E[\Delta Y(0)|D_i = 1, X_i = 1]P(X_i|D_i = 1) + \\ & E[\Delta Y(0)|D_i = 1, X_i = 0](1 - P(X_i|D_i = 1)) \end{aligned} \neq \begin{aligned} & E[\Delta Y(0)|D_i = 0, X_i = 1]P(X_i|D_i = 0) + \\ & E[\Delta Y(0)|D_i = 0, X_i = 0](1 - P(X_i|D_i = 0)) \end{aligned}$$

Unless covariate-specific trends are the equal across groups
($E[\Delta Y(0)|D_i, X_i] = E[\Delta Y(0)|X_i = 1]$) AND X is balanced ($P(X_i|D_i) = P(X_i)$), PT fails.

So what should we do?

What does “controlling” mean?

ATT counterfactual:

$$E[\Delta Y(0) | D_i = 1]$$

...it is all about using untreated units to “look like” treated units.

But OLS does this:

estimated using *all* data → “overadjustment”
when treatment effects correlated with X !*

$$Y_{it} = \theta_t + \eta_i + \gamma X_{it} + \beta D_{it} + u_{it}$$

We want methods that use X but identify ATT!

* See Sloczynski (ReStat 2022) for full details on this.

Target parameter

$$ATT = E[Y_{i,1}(1) - Y_{i,1}(0) | D_i = 1]$$

Same as without covariates (back to using $t = \{0,1\}$).

Identifying assumption: Conditional PT

$$ATT = E[Y_{i,1}(1) - Y_{i,1}(0) | D_i = 1]$$

A weaker assumption is PT for units with a given X (**CPT**):*

$$E[Y_{i,1}(0) - Y_{i,0}(0) | D_i = 1, X_i] = E[Y_{i,1}(0) - Y_{i,0}(0) | D_i = 0, X_i]$$

**Same covariate-specific trends—one ingredient from our rough PT diagnosis!*

Identifying assumption: SO

Strong overlap: $0 \leq P(D_i = 1|X_i) < 1$

“There are untreated units at every value of X.”

Why do we need it?

- Without SO, either $E[\Delta Y(0)|D_i = 1, X_i]$ or $E[\Delta Y(0)|D_i = 0, X]$ are not defined!
- There are (modeling) ways around this.

Identification under CPT (Estimand)

Conditioning on covariates gets right to the difference between estimands and estimators. See:

$$\begin{aligned} ATT &= E[Y_{i,1}(1)|D_i = 1] - E[Y_{i,1}(0) | D_i = 1] \\ &= E[Y_{i,1}(1)|D_i = 1] - E[E[Y_{i,1}(0)|X, D_i = 1]|D_i = 1] \\ &= E[Y_{i,1}(1)|D_i = 1] - E[E[Y_{i,0}(0)|X, D_i = 1] + E[\Delta Y_i(0)|X, D_i = 0]|D_i = 1] \\ &= E[\Delta Y_i | D_i = 1] - E[E[\Delta Y_i | X, D_i = 0] | D_i = 1] \end{aligned}$$

Annotations:

- Law of iterated expectations (points to the inner expectation in the second line)
- Conditional parallel trends (points to the term $E[\Delta Y_i(0)|X, D_i = 0]$ in the third line)
- Sub observed outcomes for potential outcomes (points to the term $E[\Delta Y_i | X, D_i = 0]$ in the fourth line)

This is a conditional DiD ESTIMAND: all observable population quantities

What should you actually do? (Estimator)

$$E[\Delta Y_i | D_i = 1] - E[E[\Delta Y_i | X, D_i = 0] | D_i = 1]$$

What can you do with this?

- 1. Categorical covariates $\rightarrow$ non-parametric estimator*

What should you actually do? (Estimator)

$$E[\Delta Y_i | D_i = 1] - E[E[\Delta Y_i | X, D_i = 0] | D_i = 1]$$

What can you do with this?

1. Categorical covariates → non-parametric estimator

*Just get mean outcome change for untreated units at each X ($E[\Delta Y_i | X, D_i = 0]$),
Then weight them by distribution of X for treated units ($E[\cdot | D_i = 1]$)**

Challenge: dimensionality/sample size

**or calculate a 2x2 DiD at each value of X and average those*

What should you actually do? (Estimator)

$$E[\Delta Y_i | D_i = 1] - E[E[\Delta Y_i | X, D_i = 0] | D_i = 1]$$

What can you do with this?

1. *Categorical covariates → non-parametric estimator*

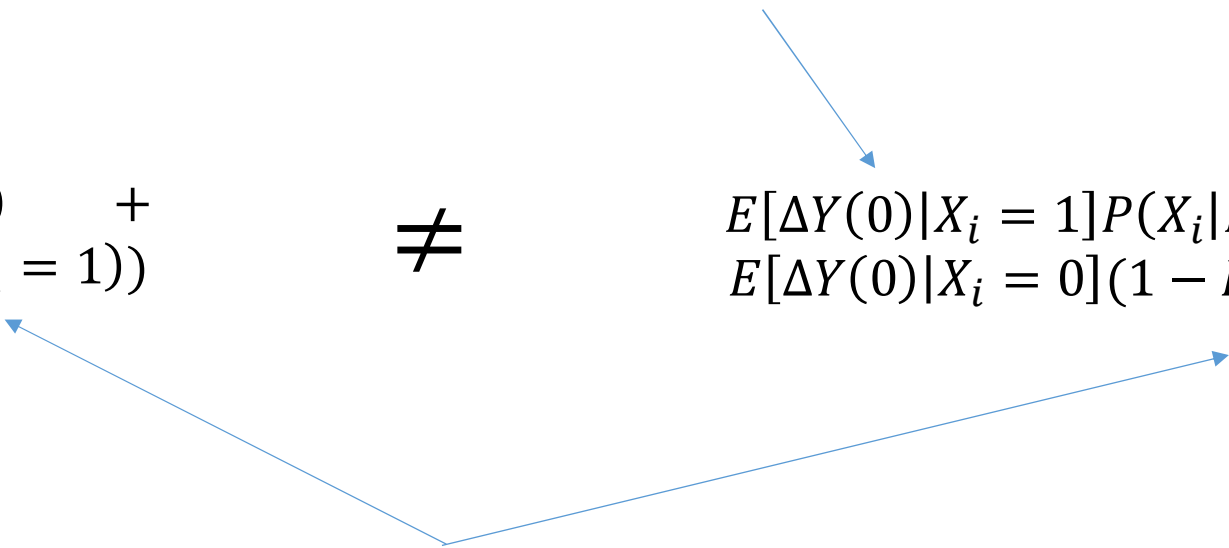
2. *Continuous covariates → modelling **choices**:*

1. *Why? Calculating $E[\Delta Y_i | X, D_i = 0]$ may be impossible. Estimand has no feasible analog in the sample.*

Option 1: fix the weights

Intuition:

Use the real untreated conditional outcome changes...

$$\begin{aligned} & E[\Delta Y(0)|X_i = 1]P(X_i|D_i = 1) + \\ & E[\Delta Y(0)|X_i = 0](1 - P(X_i|D_i = 1)) \end{aligned} \quad \neq \quad \begin{aligned} & E[\Delta Y(0)|X_i = 1]P(X_i|D_i = 0) + \\ & E[\Delta Y(0)|X_i = 0](1 - P(X_i|D_i = 0)) \end{aligned}$$


But find a (feasible) way to fix the weights to achieve balance

Inverse Propensity Score Weighting (IPW)

Model conditional probability of treatment: $p(X_i) \equiv E[D_i|X_i]$

Dimension
reduction!

Why does this object help? Because of a theorem!

$$E[E[\Delta Y_i | X, D_i = 0] | D_i = 1] = E\left[\frac{p(X_i)}{1 - p(X_i)} \Delta Y_i \middle| D_i = 0\right] E\left[\frac{p(X_i)}{1 - p(X_i)} \middle| D_i = 0\right]^{-1}$$

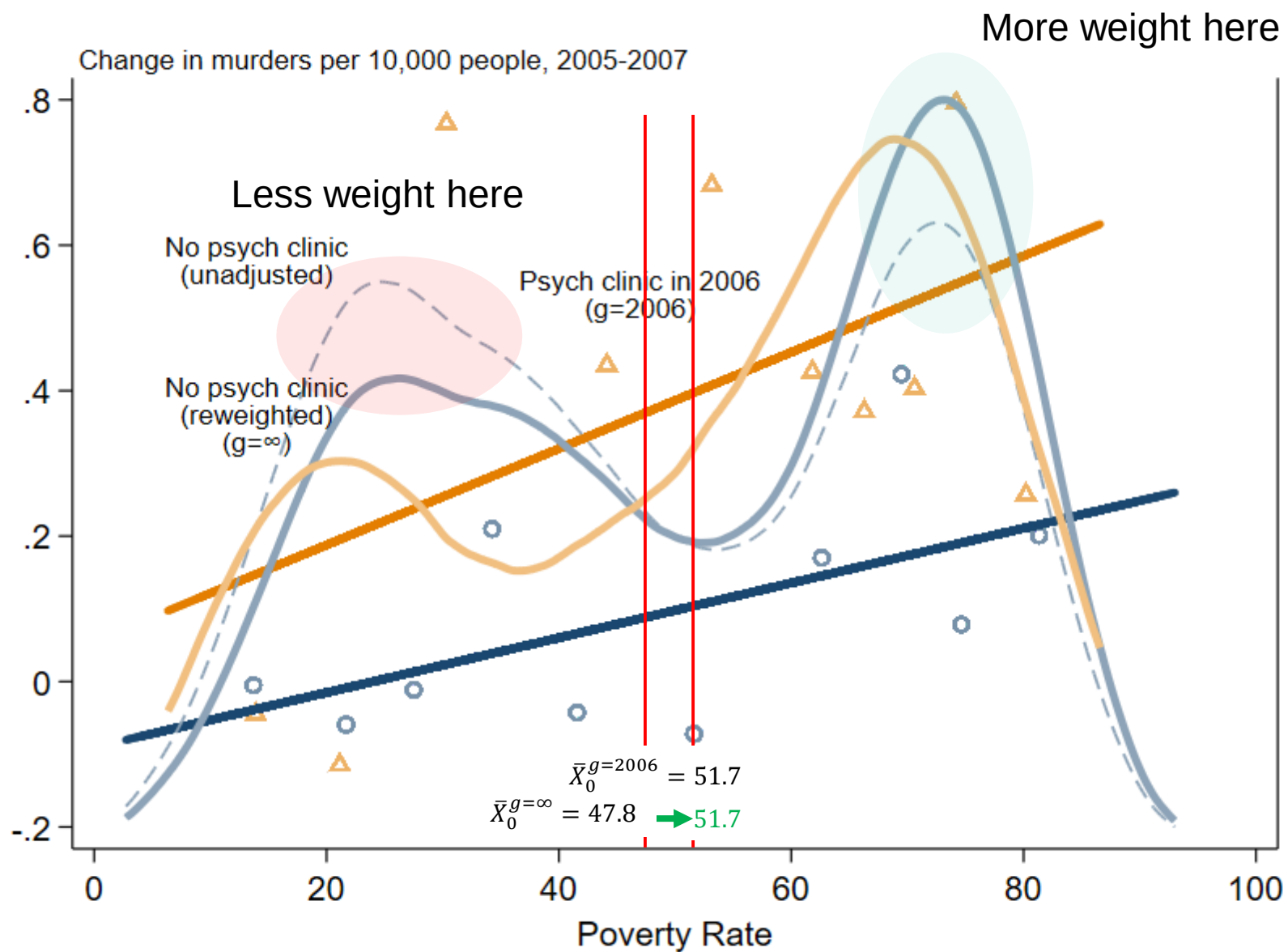
In practice: Use logit/probit to get an estimated p-score, $\hat{\pi}(X_i)$, and form weights:

$$\hat{w}(X_i) \equiv \frac{\hat{\pi}(X_i)}{1 - \hat{\pi}(X_i)} \left(\frac{1}{N_0} \sum_i 1\{D_i = 0\} \frac{\hat{\pi}(X_i)}{1 - \hat{\pi}(X_i)} \right)^{-1}$$

Then, get counterfactual by averaging ΔY_i for untreated units using $\hat{w}_i(X_i)$, form ATT

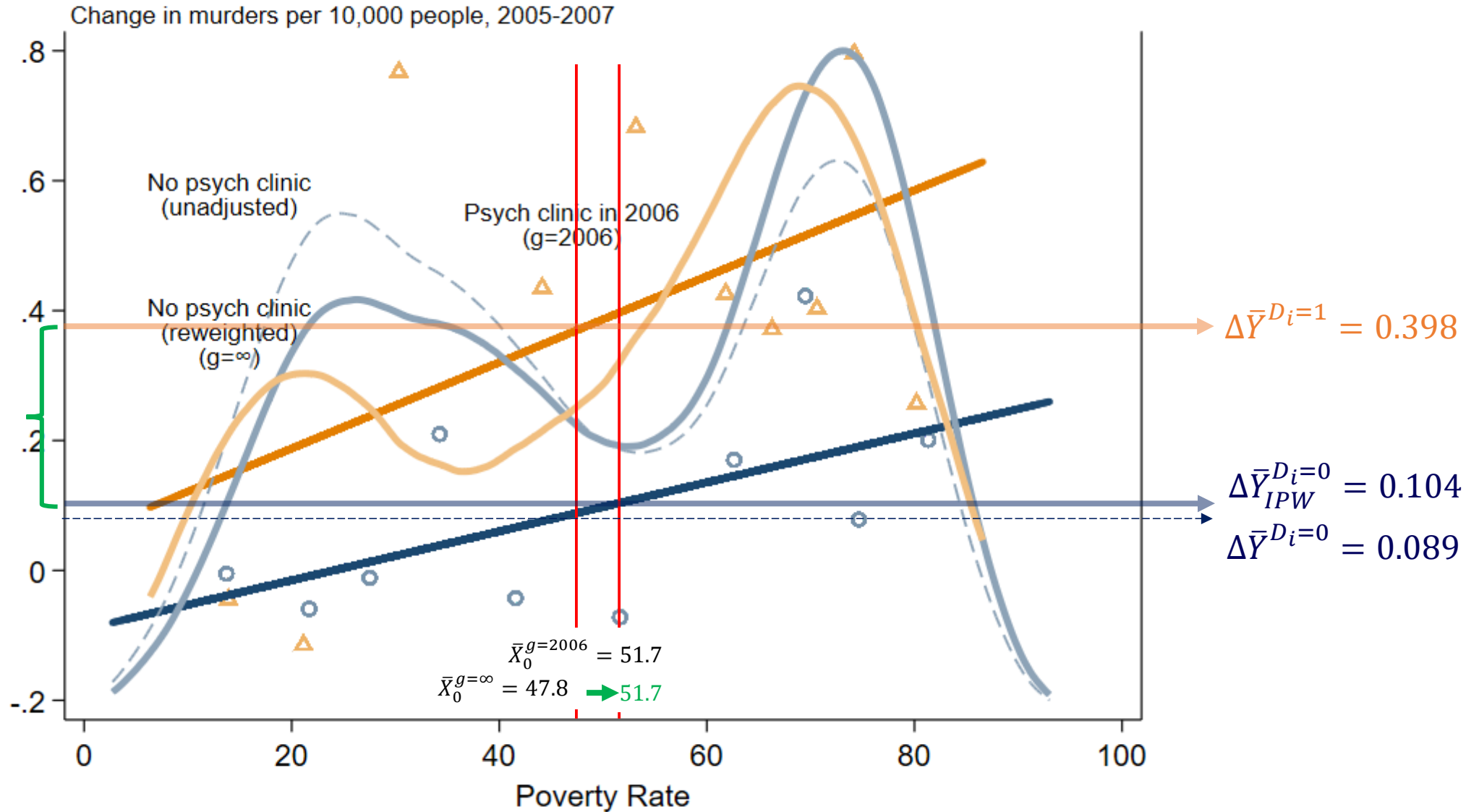
$$\widehat{ATT}^{IPW} = \underbrace{\frac{1}{N_1} \sum_i 1\{D_i = 1\} \Delta Y_i}_{\text{Untouched treatment mean}} - \underbrace{\sum_i 1\{D_i = 0\} \hat{w}(X_i) \Delta Y_i}_{\text{Rewighted control mean}}$$

$\widehat{ATT}^{IPW}$



$$\widehat{ATT}^{IPW}$$

$$\widehat{ATT}^{IPW} = 0.294$$



Option 2: understand the relationship between X_i and $\Delta Y(0)$

Intuition:

Understand the conditional (on X_i) mean of $\Delta Y(0)$ in the untreated group

$$E[\Delta Y(0)|X_i = 1]P(X_i|D_i = 1) + E[\Delta Y(0)|X_i = 0](1 - P(X_i|D_i = 1))$$

$\neq$

$$E[\Delta Y(0)|X_i = 1]P(X_i|D_i = 0) + E[\Delta Y(0)|X_i = 0](1 - P(X_i|D_i = 0))$$

Then apply that relationship to the observed X_i in the treatment group

Regression Adjustment (RA)

Dimension
reduction!

Model untreated outcome trends as some function, $\mu_{\Delta}(X_i; \theta) = E[\Delta Y_i | X_i, D_i = 0]$

Why does this help? Because under CPT, it fills in counterfactual trends for $D_i = 1$

In practice: Pick a functional form for the untreated outcome model $\mu_{\Delta}(X_i; \theta)$

Linear: $\mu_{\Delta}(X_i; \theta) = X_i \theta$; non-linear: $\mu_{\Delta}(X_i; \theta) = G(X_i \theta)$; non-parametric/ML/etc

Use untreated observations to estimate that model's parameters; $\hat{\theta}$

Apply the estimated model to treated units' X to predict counterfactuals:

$$\widehat{ATT}^{RA} = \frac{1}{N_1} \sum_i \mathbf{1}\{D_i = 1\} \left(\Delta Y_i - \mu_{\Delta}(X_i; \hat{\theta}) \right)$$

Average only of treated obs

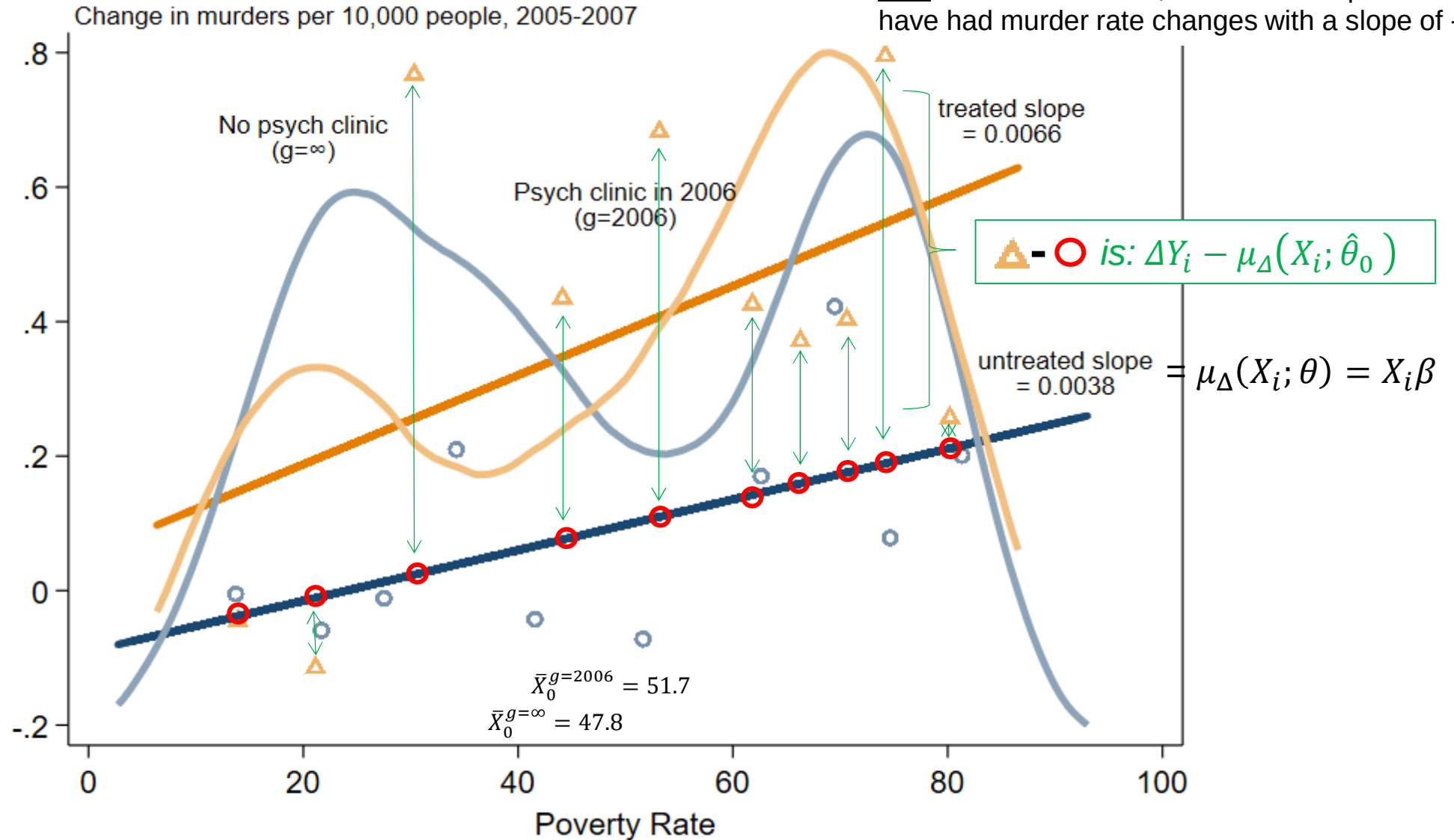
Controls contributed here

$$\widehat{ATT}^{RA}$$

Assumptions

Correct model: Untreated outcome trends really are a linear function of size with a slope of -0.0038.

CPT: Absent treatment, treated municipalities would have had murder rate changes with a slope of -0.0038



$$\widehat{ATT}^{RA} = 0.293$$

Doubly Robust Estimator (DR)

Combine *both* IPW and RA into a *doubly robust* estimator:

$$\widehat{ATT}^{DR} = \frac{1}{N} \sum_i (\hat{w}_{D=1} - \hat{w}_{D=0}(X_i)) (\Delta Y_i - \mu_{\Delta}(X_i; \hat{\theta}_0))$$
$$\hat{w}_{D=1} \equiv \frac{\omega D}{\frac{1}{N} \sum_i \omega D}, \quad \hat{w}_{D=0}(X_i) \equiv \frac{\omega(1-D) \frac{\hat{\pi}_i(X_i)}{1-\hat{\pi}_i(X_i)}}{\frac{1}{N} \sum_i \omega(1-D) \frac{\hat{\pi}_i(X_i)}{1-\hat{\pi}_i(X_i)}}$$

Identifies ATT if either the p-score model OR the outcome model is correct

“Doubly Robust”

Combine *both* IPW and RA into a *doubly robust* estimator:

- **Double robustness is a property** not a specific estimator (eg. Sloczynski and Wooldridge 2018) and there are many ways to do it (ie. Oaxaca-Blinder is doubly robust! [Kline 2011])
- `teffects/drdid` (Stata), `did` (R)

If time: Medicaid expansion example

Table 2: Simple 2×2 DiD

	Unweighted Averages			Weighted Averages		
	Expansion	No Expansion	Gap/DiD	Expansion	No Expansion	Gap/DiD
2013	419.2	474.0	-54.8	322.7	376.4	-53.7
2014	428.5	483.1	-54.7	326.5	382.7	-56.2
<i>Trend/DiD</i>	9.3	9.1	0.1	3.7	6.3	-2.6

Means-tested public health insurance in the US. Some states expanded coverage to certain low-income adults without children starting in 2014.

This table reports average county-level mortality rates (deaths among adults aged 20-64 per 100,000 adults) in 2013 (top row) and 2014 (middle row) in states that expanded adult Medicaid eligibility in 2014 (columns 1 and 4, 978 counties) and states that have not expanded by 2019 (columns 2 and 5, 1,222 counties). The first three columns present unweighted averages and the second three columns present population-weighted averages. Columns 1, 2, 4, and 5 in the third row show time trends in mortality between 2013 and 2014 for each group of states. The first two rows of columns 3 and 6 show the cross-sectional gap in mortality between expansion and non-expansion states in 2013 and 2014. The entries in red text in the bottom row show the simple 2×2 difference-in-differences estimates without weights (column 3) and with them (column 6)

If time: Medicaid expansion example

Table 4: Covariate Balance Statistics

Variable	Unweighted			Weighted		
	Non-Adopt	Adopt	Norm. Diff.	Non-Adopt	Adopt	Norm. Diff.
2013 Covariate Levels						
% Female	49.43	49.33	-0.03	50.48	50.07	-0.24
% White	81.64	90.48	0.59	77.91	79.54	0.11
% Hispanic	9.64	8.23	-0.10	17.01	18.86	0.11
Unemployment Rate	7.61	8.01	0.16	7.00	8.01	0.50
Poverty Rate	19.28	16.53	-0.42	17.24	15.29	-0.37
Median Income	43.04	47.97	0.43	49.31	57.86	0.68
2014 - 2013 Covariate Differences						
% Female	-0.02	-0.02	0.00	0.02	0.01	-0.09
% White	-0.21	-0.21	0.01	-0.32	-0.33	-0.04
% Hispanic	0.20	0.21	0.04	0.25	0.33	0.29
Unemployment Rate	-1.16	-1.30	-0.21	-1.08	-1.36	-0.55
Poverty Rate	-0.55	-0.28	0.14	-0.41	-0.35	0.05
Median Income	0.98	1.11	0.06	1.10	1.74	0.32

$$= \frac{\bar{X}_{\omega,T} - \bar{X}_{\omega,C}}{\sqrt{(S_{\omega,T}^2 + S_{\omega,C}^2)/2}},$$

This table reports the covariate balance between 978 counties in states that expanded Medicaid in 2014 and 1,222 counties in states that did not expand by 2019. In the top panel, we report the averages and standardized differences of each variable, measured in 2013, by adoption status. All variables are measured in percentage values, except for median household income, which is measured in thousands of U.S. dollars. In the bottom panel we report the average and standardized differences of the county-level long differences between 2014 and 2013 of each variable. We report both weighted and unweighted measures of the averages to correspond to the different estimation methods of including covariates in a 2×2 setting.

If time: Medicaid expansion example

Table 6: Outcome Regression and Propensity Score Models

	Unweighted		Weighted	
	Regression	Propensity Score	Regression	Propensity Score
	(1) OLS	(2) Logit	(3) OLS	(4) Logit
Constant	-20.91 (69.85)	-10.00*** (1.35)	-4.62 (44.98)	-8.17*** (0.01)
% Female	0.04 (0.82)	-0.04** (0.02)	-0.09 (0.68)	-0.19*** (0.00)
% White	0.15 (0.22)	0.06*** (0.00)	0.20* (0.11)	0.04*** (0.00)
% Hispanic	-0.08 (0.20)	-0.02*** (0.00)	-0.08 (0.08)	-0.02*** (0.00)
Unemployment Rate	1.14 (1.56)	0.32*** (0.03)	0.88 (0.99)	0.68*** (0.00)
Poverty Rate	0.21 (0.98)	0.03* (0.02)	-0.13 (0.53)	0.11*** (0.00)
Median Income	0.09 (0.52)	0.08*** (0.01)	-0.05 (0.24)	0.15*** (0.00)

Mean predicted change in mortality:

$$\sum_i \frac{1\{G_i = 2014\} \mu_{\Delta}(X_i; \hat{\theta}_0)}{\sum_i 1\{G_i = 2014\}}$$

= 7.2 deaths per 100,000

This table reports the outcome regression propensity score models that enter into the estimator from Sant'Anna and Zhao (2020) and Callaway and Sant'Anna (2021). The first two columns report the results for unweighted regressions and the second two report results from weighted regression models. The regression model predicts changes in the outcome variable (mortality rates) as a function of 2013 covariate values for the 1,222 counties that do not expand Medicaid in 2014. The propensity score model uses data on all 2,200 counties for 2013 and estimates a logit model of an expansion indicator variable on the 2013 covariate levels. Standard errors (in parentheses) are clustered at the county level.

$$\rightarrow \widehat{ATT}^{RA} = 3.7 - 7.2 = -3.5$$

If time: Medicaid expansion example

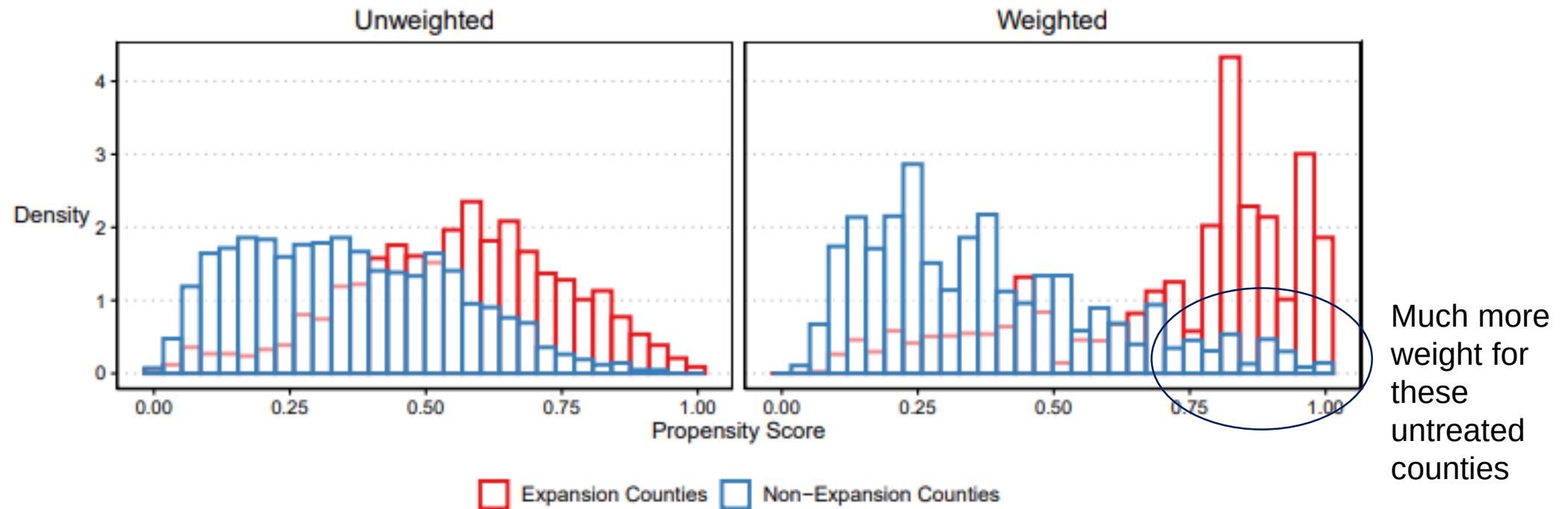
Table 6: Outcome Regression and Propensity Score Models

	Unweighted		Weighted	
	Regression	Propensity Score	Regression	Propensity Score
	(1) OLS	(2) Logit	(3) OLS	(4) Logit
Constant	-20.91 (69.85)	-10.00*** (1.35)	-4.62 (44.98)	-8.17*** (0.01)
% Female	0.04 (0.82)	-0.04** (0.02)	-0.09 (0.68)	-0.19*** (0.00)
% White	0.15 (0.22)	0.06*** (0.00)	0.20* (0.11)	0.04*** (0.00)
% Hispanic	-0.08 (0.20)	-0.02*** (0.00)	-0.08 (0.08)	-0.02*** (0.00)
Unemployment Rate	1.14 (1.56)	0.32*** (0.03)	0.88 (0.99)	0.68*** (0.00)
Poverty Rate	0.21 (0.98)	0.03* (0.02)	-0.13 (0.53)	0.11*** (0.00)
Median Income	0.09 (0.52)	0.08*** (0.01)	-0.05 (0.24)	0.15*** (0.00)

This table reports the outcome regression propensity score models that enter into the estimator from Sant'Anna and Zhao (2020) and Callaway and Sant'Anna (2021). The first two columns report the results for unweighted regressions and the second two report results from weighted regression models. The regression model predicts changes in the outcome variable (mortality rates) as a function of 2013 covariate values for the 1,222 counties that do not expand Medicaid in 2014. The propensity score model uses data on all 2,200 counties for 2013 and estimates a logit model of an expansion indicator variable on the 2013 covariate levels. Standard errors (in parentheses) are clustered at the county level.

If time: Medicaid expansion example

Figure 1: Distribution of Propensity Scores



Notes: This figure shows the distribution of propensity scores using both the weighted and unweighted logit propensity score estimates in our 2 × 2 Medicaid example.

Reweighted change in mortality among untreated counties: $\sum_i \frac{1\{G_i=\infty\}w_i(x)\Delta Y_i}{\sum_i 1\{G_i=\infty\}w_i(x)} = 7.5$ deaths per 100,000

$$\rightarrow \widehat{ATT}^{IPW} = 3.7 - 7.5 = -3.8$$

If time: Medicaid expansion example

Table 7: DiD Estimates with Covariates

	Unweighted			Weighted		
	Regression	IPW	Doubly Robust	Regression	IPW	Doubly Robust
Medicaid Expansion	-1.62 (4.73)	-0.86 (4.76)	-1.23 (4.61)	-3.46 (2.29)	-3.84 (3.22)	-3.76 (3.59)

This table reports the 2×2 DiD estimate comparing 978 counties in states that expand Medicaid in 2014 to 1,222 counties in states that did not expand Medicaid by 2019, adjusting for 2013 covariate values using the methodologies discussed in Sant'Anna and Zhao (2020) and Callaway and Sant'Anna (2021). The first column reports results using regression adjustment, the second column uses inverse probability weighting based on a propensity score model using the included covariates, and the third column uses the doubly robust combination of the two approaches. Standard errors (in parentheses) are clustered at the county level.

If time: Medicaid expansion example

Table 5: Regression 2×2 DiD with Covariates

	Unweighted			Weighted		
	No Covs	$X_{i,t=2013}$	$X_{i,t}$	No Covs	$X_{i,t=2013}$	$X_{i,t}$
	(1)	(2)	(3)	(4)	(5)	(6)
Medicaid Expansion	0.12 (3.75)	-2.35 (4.29)	-0.49 (3.83)	-2.56* (1.49)	-2.56 (1.78)	-1.37 (1.62)

This table reports the regression 2×2 DiD estimate comparing 978 counties that expanded Medicaid in 2014 with 1,222 counties that did not expand Medicaid by 2019, adjusting for covariates (percent female, percent white, percent hispanic, the unemployment rate, the poverty rate, and median household income). Columns 1-3 report unweighted regression results, while columns 4-6 weight by county population aged 20-64 in 2013. Columns 1 and 4 report results for expansion states without covariates, columns 2 and 5 adjust for the baseline levels of the covariates in 2013, and columns 3 and 6 control for the time-varying covariate values in 2014 and 2013. Standard errors (in parentheses) are clustered at the county level.



New Developments in Difference-in-Differences: *Staggered Timing*

Andrew Goodman-Bacon

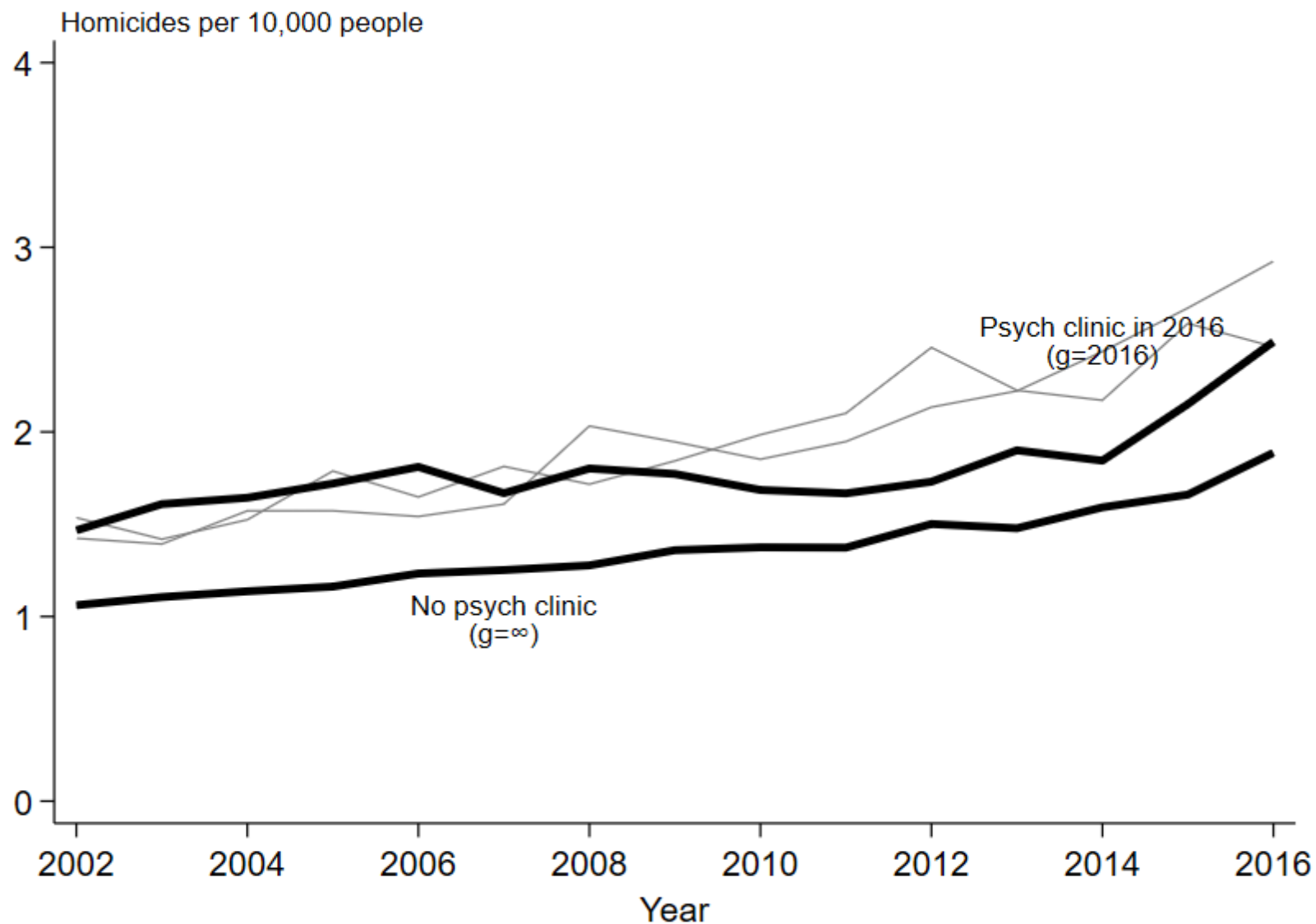
IIPF, 2026

The Effects of a Large-Scale Mental Health Reform: Evidence from Brazil[†]

By MATEUS DIAS AND LUIZ FELIPE FONTES*

Our empirical strategy exploits the rollout of CAPS across Brazilian municipalities in a difference-in-differences (DID) framework.

Staggered roll-out, 2003-2016



Two-way fixed effects

Suppose g takes many values; ie. treatment timing varies. This equation doesn't work anymore:

$$y_{it} = \hat{\alpha}_0 + \hat{\alpha}_1 \cancel{D_i} + \hat{\alpha}_2 \cancel{1\{t = 2007\}} + \hat{\beta}^{2x2} \cancel{D_i \times 1\{t = 2007\}} + u_{it}$$

Not just one “treatment group” anymore

“Post” differs across units now

$$y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{DiD} D_{it} + u_{it}$$

Unit fixed effects

Time fixed effects

Treatment dummy

What is $\hat{\beta}^{DiD}$?

What is $\hat{\beta}^{DiD}$?

$$y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{DiD} D_{it} + u_{it}$$

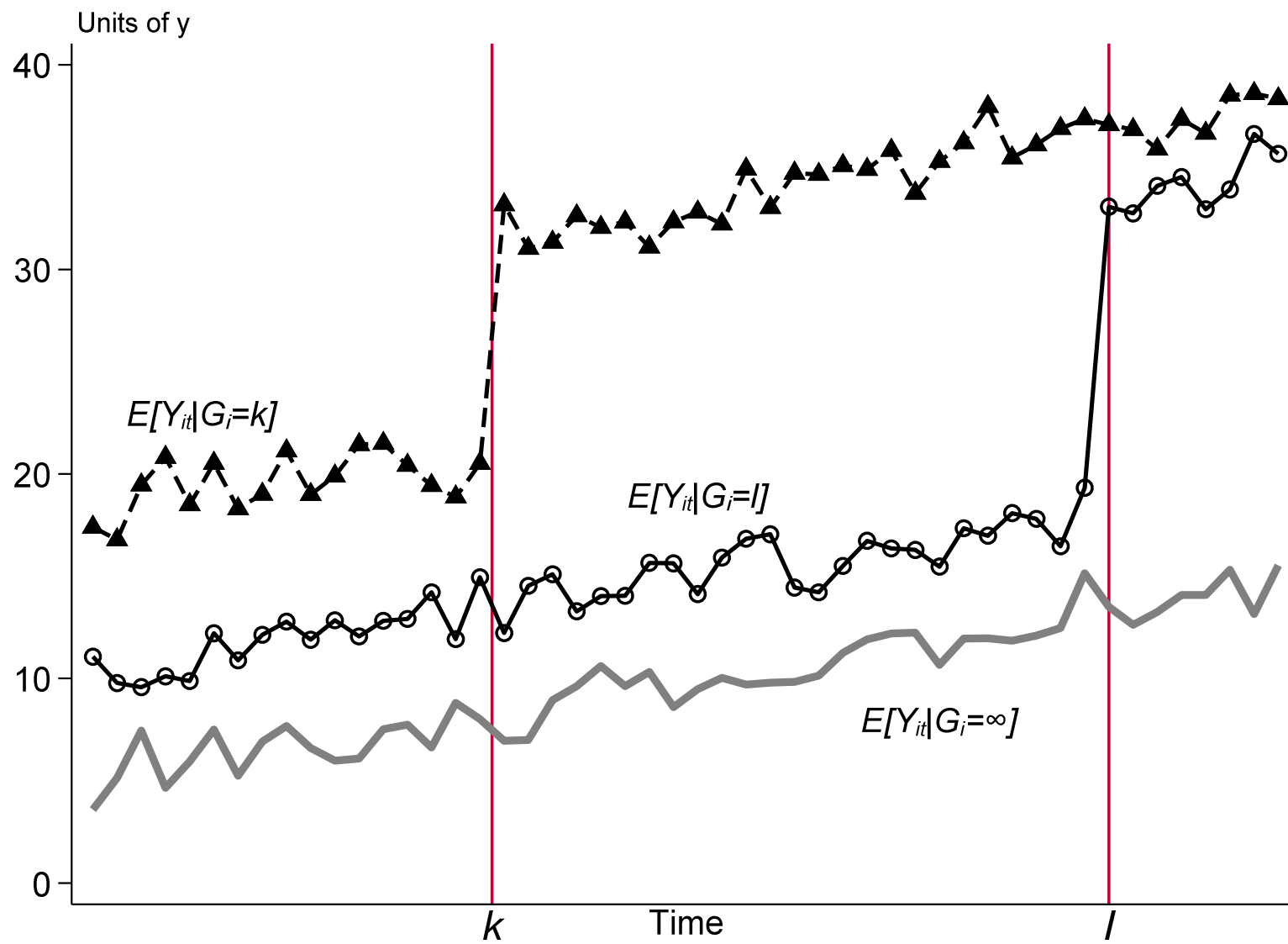
1. Partial out fixed effects (Frisch-Waugh):

$$\tilde{D}_{it} = (D_{it} - \bar{D}_i) - (\bar{D}_t - \bar{\bar{D}})$$

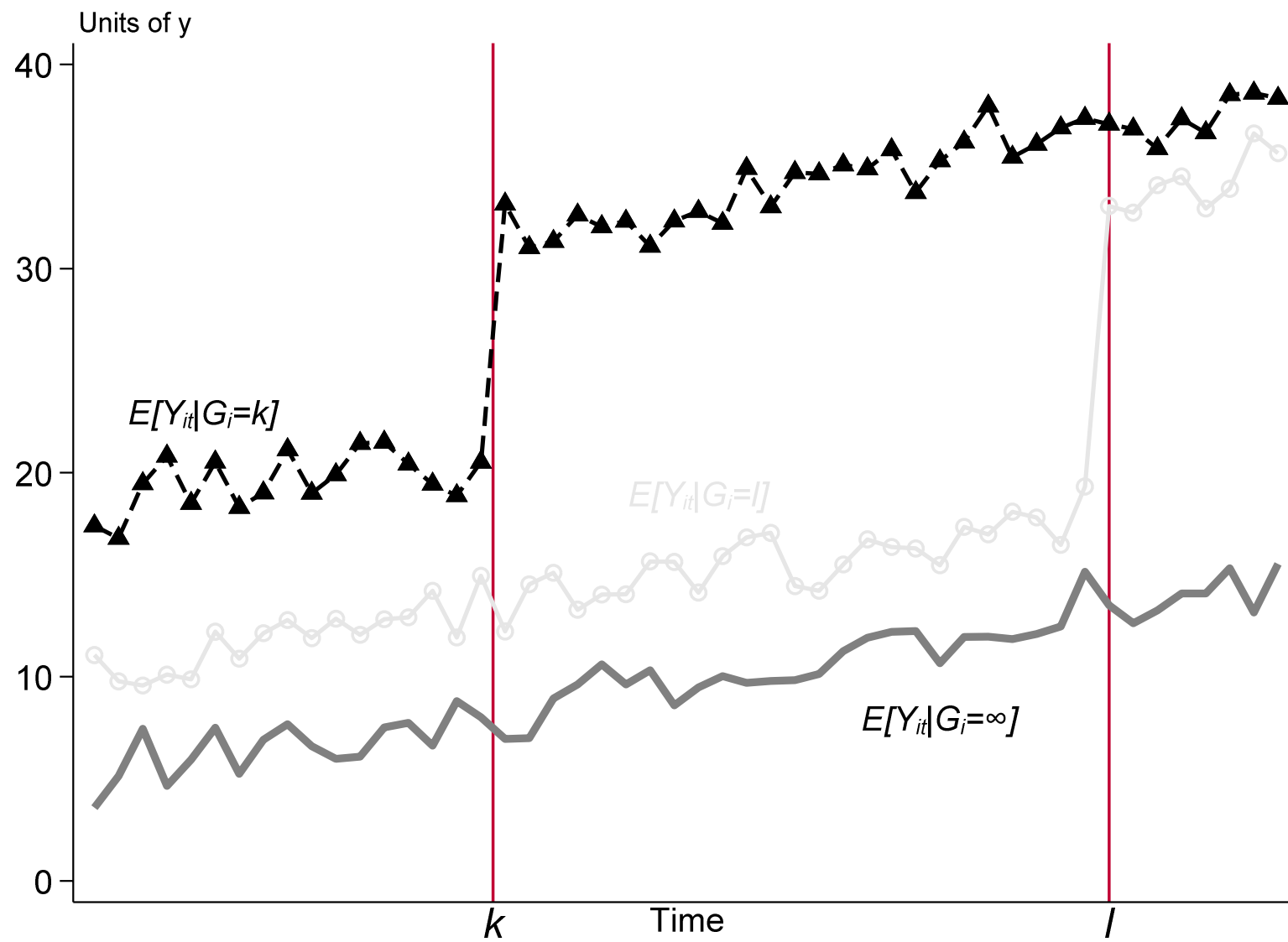
2. Calculate univariate coefficient by brute force:

$$\hat{\beta}^{DiD} = \frac{\widehat{cov}(\tilde{D}_{it}, y_{it})}{\widehat{V}(\tilde{D}_{it})} = \frac{\frac{1}{NT} \sum_i \sum_t (y_{it} - \bar{\bar{y}}) ((D_{it} - \bar{D}_i) - (\bar{D}_t - \bar{\bar{D}}))}{\frac{1}{NT} \sum_i \sum_t ((D_{it} - \bar{D}_i) - (\bar{D}_t - \bar{\bar{D}}))}$$

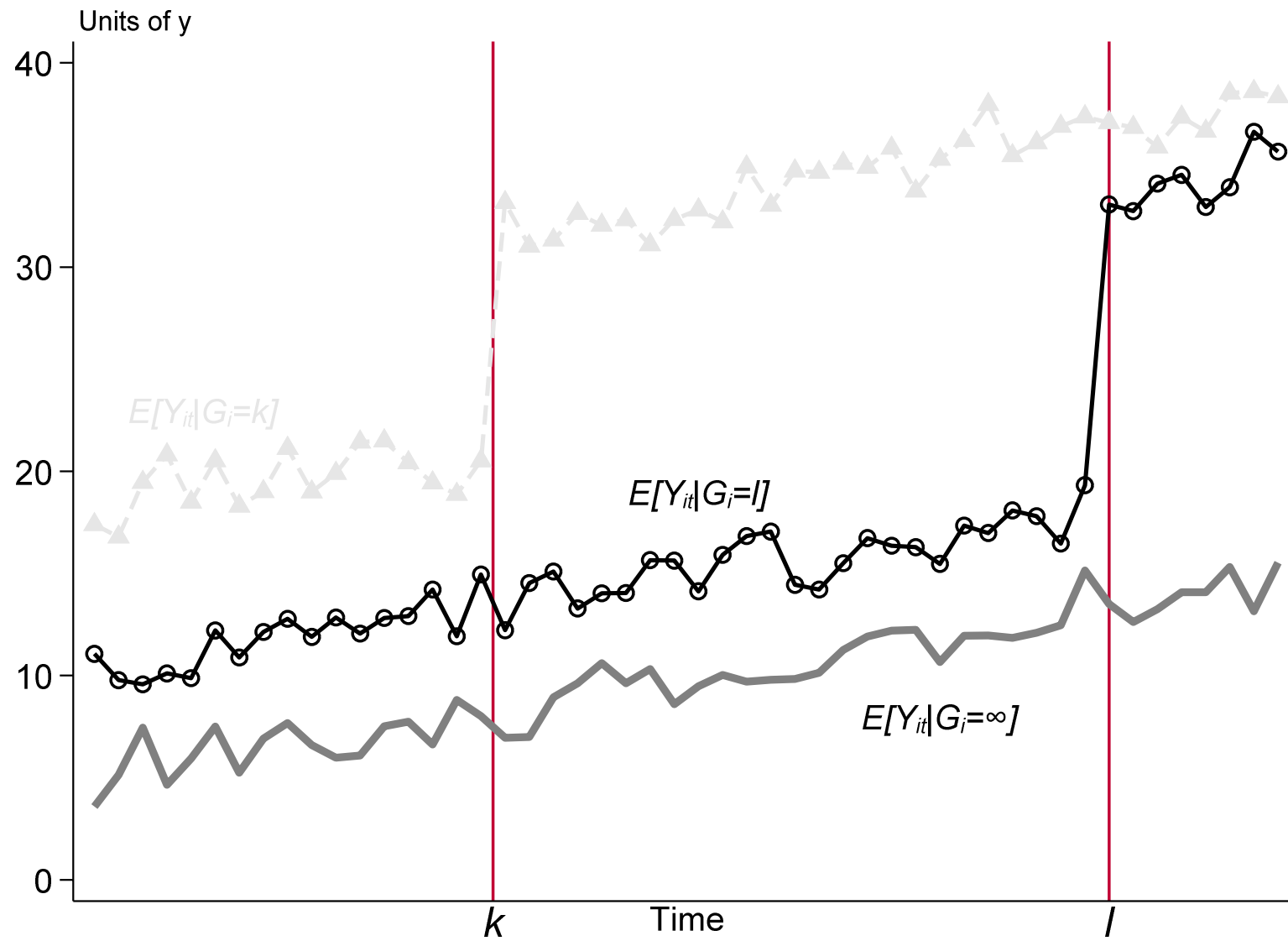
$$\hat{\beta}^{DiD?}$$



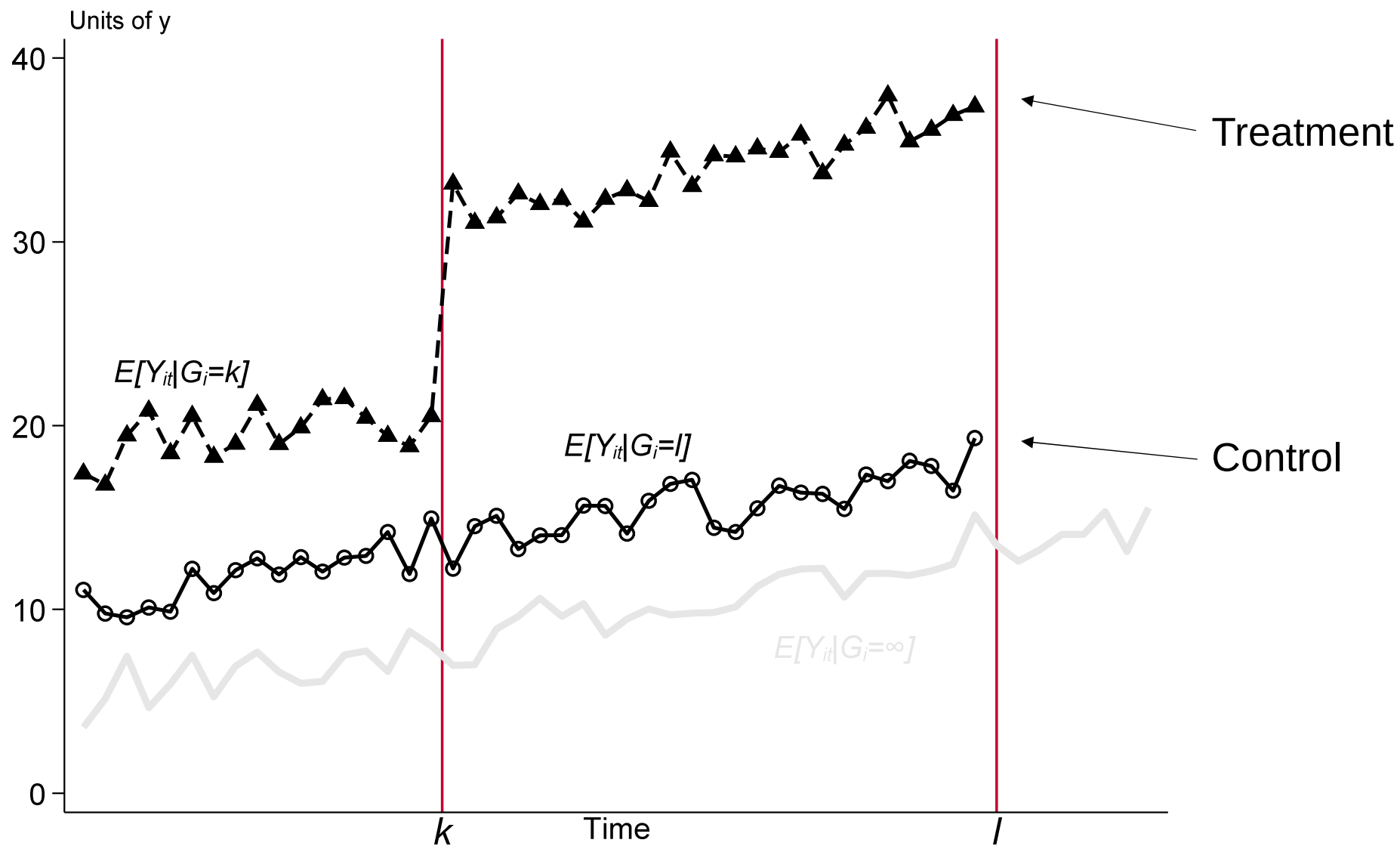
$$\hat{\beta}_{k\infty}^{DiD}$$



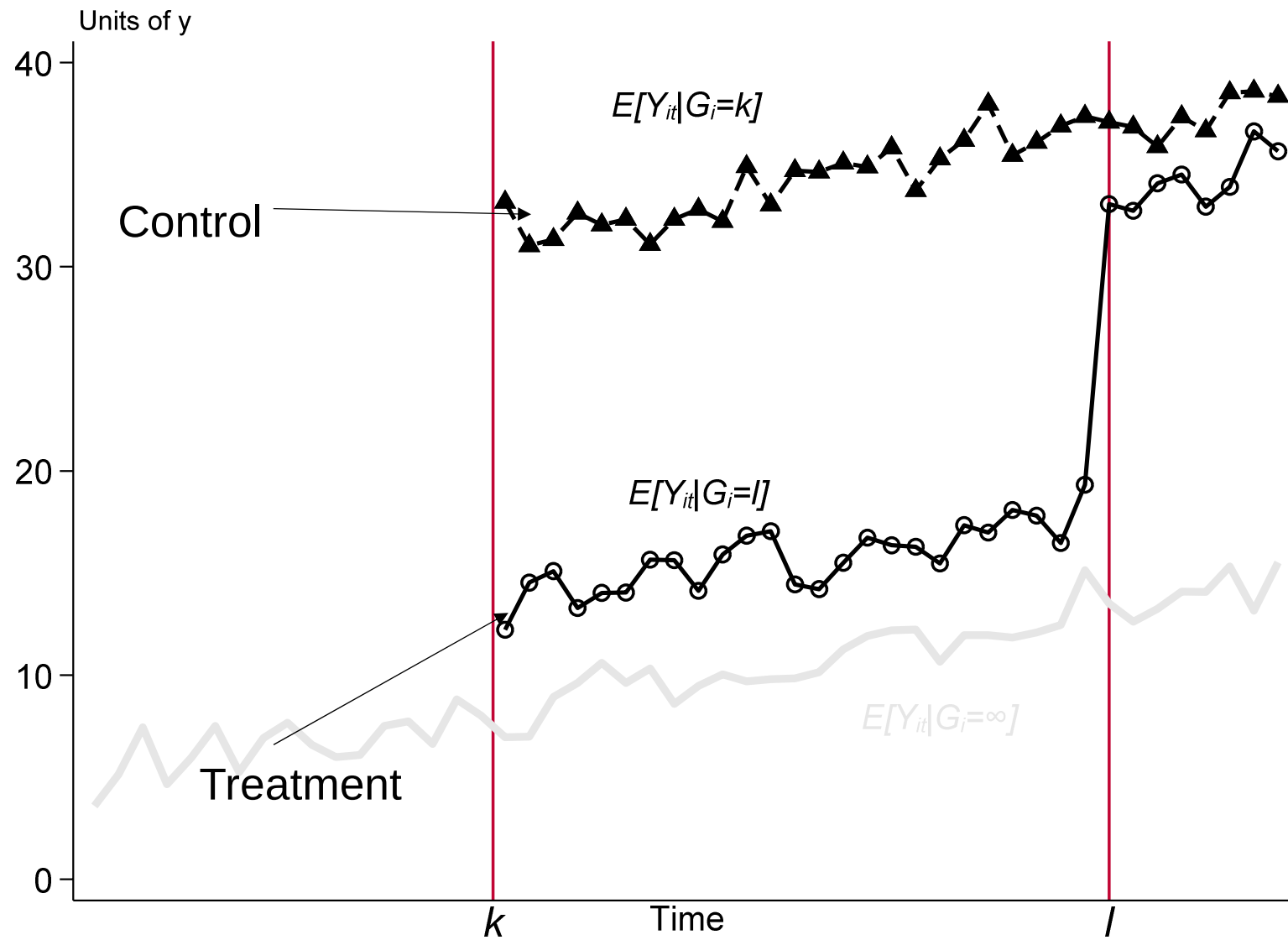
$$\hat{\beta}_{\ell\infty}^{DiD}$$



$$\hat{\beta}_{k\ell}^{DiD,k}$$



$$\hat{\beta}_{k\ell}^{DiD,\ell}$$



What is $\hat{\beta}^{DiD}$?

$$y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{DiD} D_{it} + u_{it}$$

For three groups:

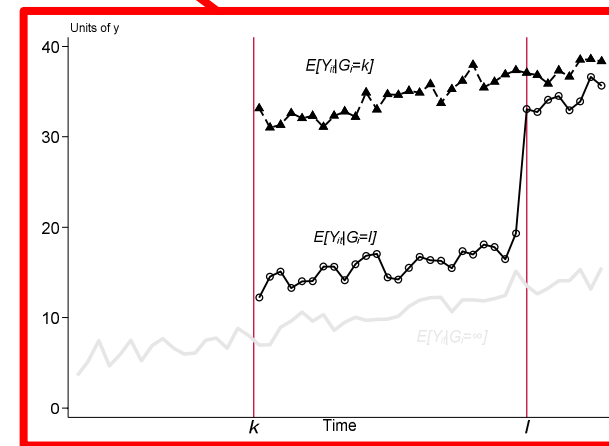
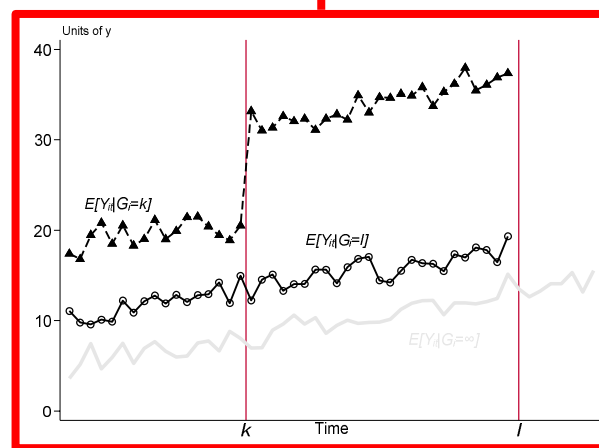
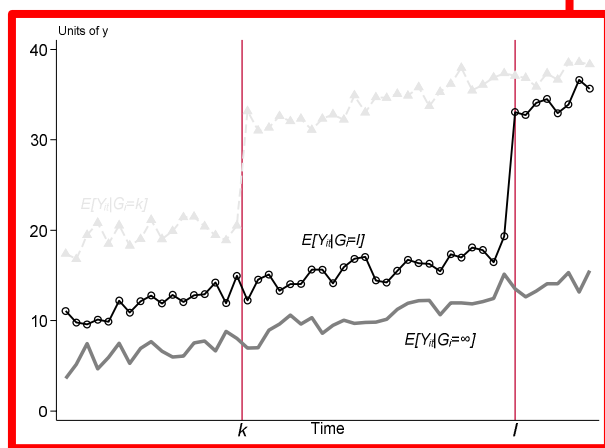
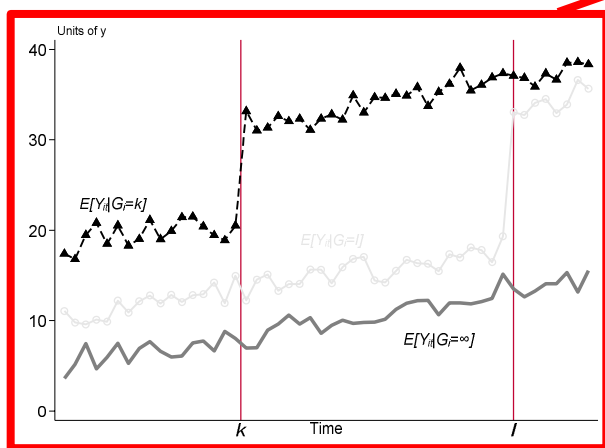
$$\hat{\beta}^{DiD} = s_{k\infty} \hat{\beta}_{k\infty}^{DiD} + s_{\ell\infty} \hat{\beta}_{\ell\infty}^{DiD} + [s_{k\ell}^k \hat{\beta}_{k\ell}^{DiD,k} + s_{k\ell}^\ell \hat{\beta}_{k\ell}^{DiD,\ell}]$$

What is $\hat{\beta}^{DiD}$?

$$y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{DiD} D_{it} + u_{it}$$

For three groups:

$$\hat{\beta}^{DiD} = s_{k\infty} \hat{\beta}_{k\infty}^{DiD} + s_{\ell\infty} \hat{\beta}_{\ell\infty}^{DiD} + [s_{k\ell}^k \hat{\beta}_{k\ell}^{DiD,k} + s_{k\ell}^\ell \hat{\beta}_{k\ell}^{DiD,\ell}]$$



What is $\hat{\beta}^{DiD}$?

$$y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{DiD} D_{it} + u_{it}$$

For three groups:

$$\hat{\beta}^{DiD} = s_{k\infty} \hat{\beta}_{k\infty}^{DiD} + s_{\ell\infty} \hat{\beta}_{\ell\infty}^{DiD} + [s_{k\ell}^k \hat{\beta}_{k\ell}^{DiD,k} + s_{k\ell}^\ell \hat{\beta}_{k\ell}^{DiD,\ell}]$$

$$s_{k\infty} = \frac{(n_k + n_\infty)^2}{V(\tilde{D}_{it})}$$

Sample size²

$$s_{k\ell}^k = \frac{((n_k + n_\ell)(1 - \bar{D}_\ell))^2}{V(\tilde{D}_{it})}$$

$$s_{k\ell}^\ell = \frac{((n_k + n_\ell)\bar{D}_k)^2}{V(\tilde{D}_{it})}$$

What is $\hat{\beta}^{DiD}$?

$$y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{DiD} D_{it} + u_{it}$$

For three groups:

$$\hat{\beta}^{DiD} = s_{k\infty} \hat{\beta}_{k\infty}^{DiD} + s_{\ell\infty} \hat{\beta}_{\ell\infty}^{DiD} + [s_{k\ell}^k \hat{\beta}_{k\ell}^{DiD,k} + s_{k\ell}^\ell \hat{\beta}_{k\ell}^{DiD,\ell}]$$

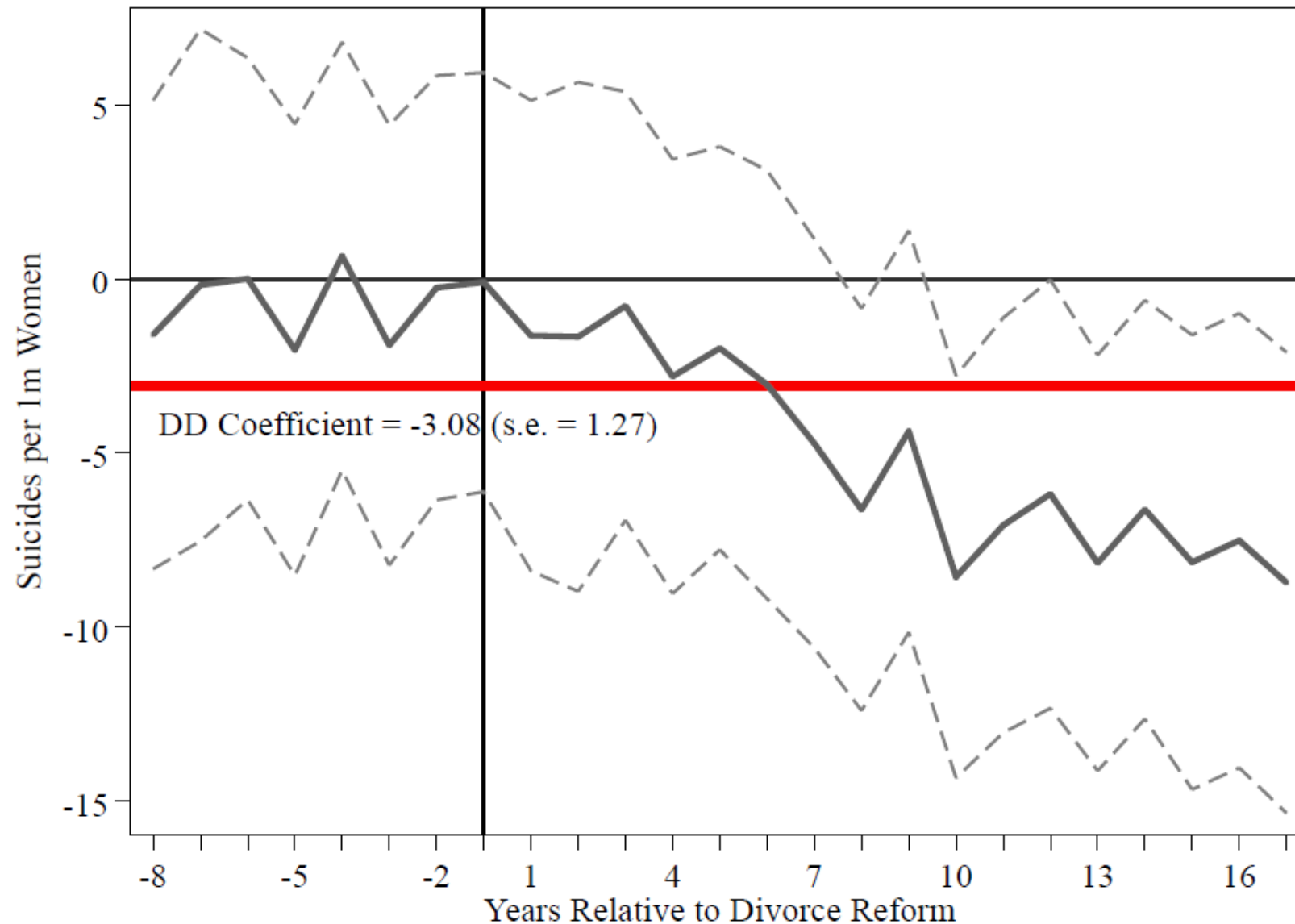
$$s_{k\infty} = \frac{(n_k + n_\infty)^2 n_{k\infty} (1 - n_{k\infty}) \bar{D}_k (1 - \bar{D}_k)}{V(\tilde{D}_{it})}$$

$$s_{k\ell}^k = \frac{((n_k + n_\ell)(1 - \bar{D}_\ell))^2 n_{k\ell} (1 - n_{k\ell}) \frac{\bar{D}_k - \bar{D}_\ell}{1 - \bar{D}_\ell} \frac{1 - \bar{D}_k}{1 - \bar{D}_\ell}}{V(\tilde{D}_{it})}$$

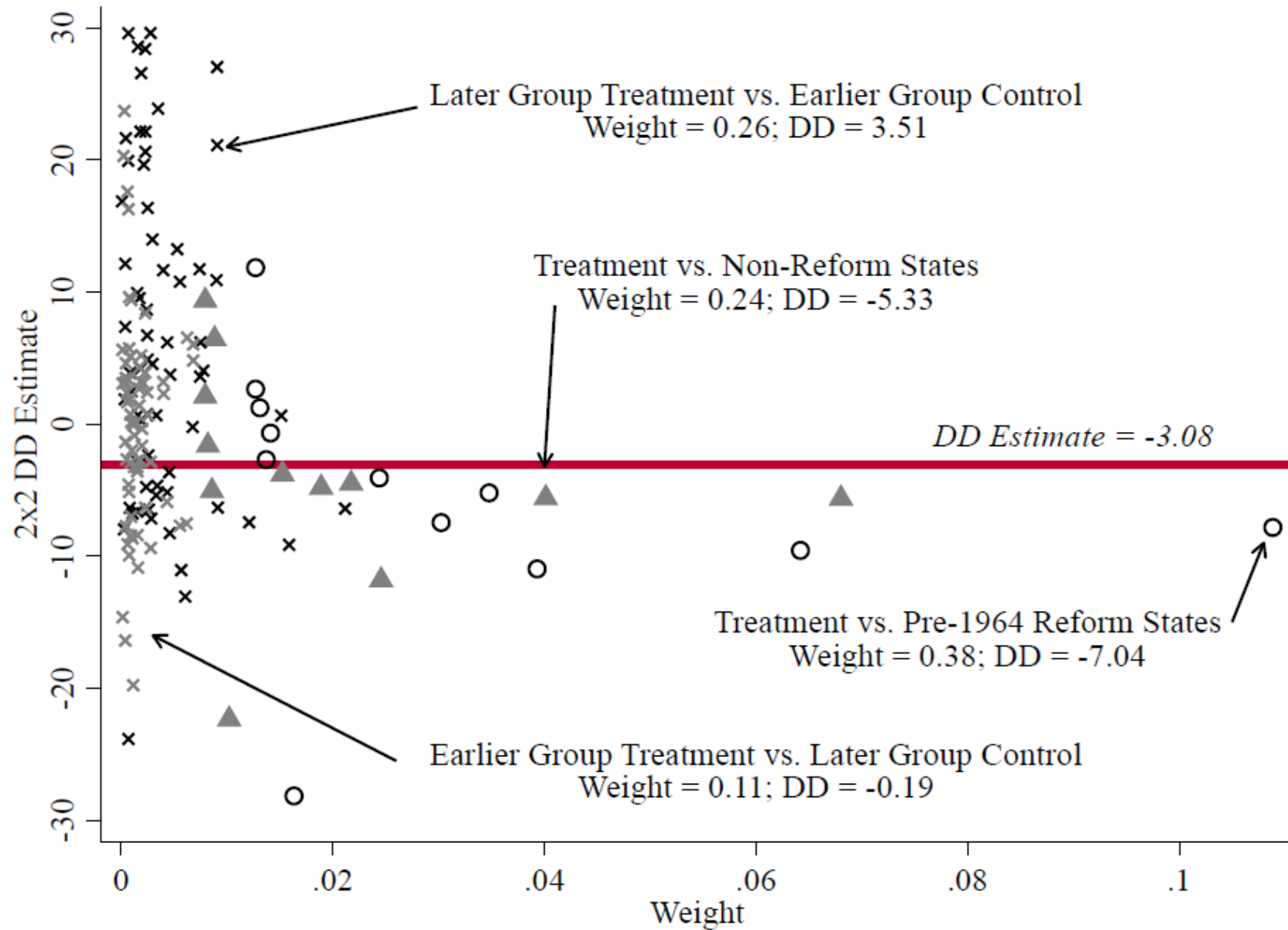
$$s_{k\ell}^\ell = \frac{((n_k + n_\ell) \bar{D}_k)^2 n_{k\ell} (1 - n_{k\ell}) \frac{\bar{D}_k - \bar{D}_\ell}{\bar{D}_k} \frac{\bar{D}_\ell}{\bar{D}_k}}{V(\tilde{D}_{it})}$$

Subsample
variance of
treatment

Example: Stevenson and Wolfers (2006)



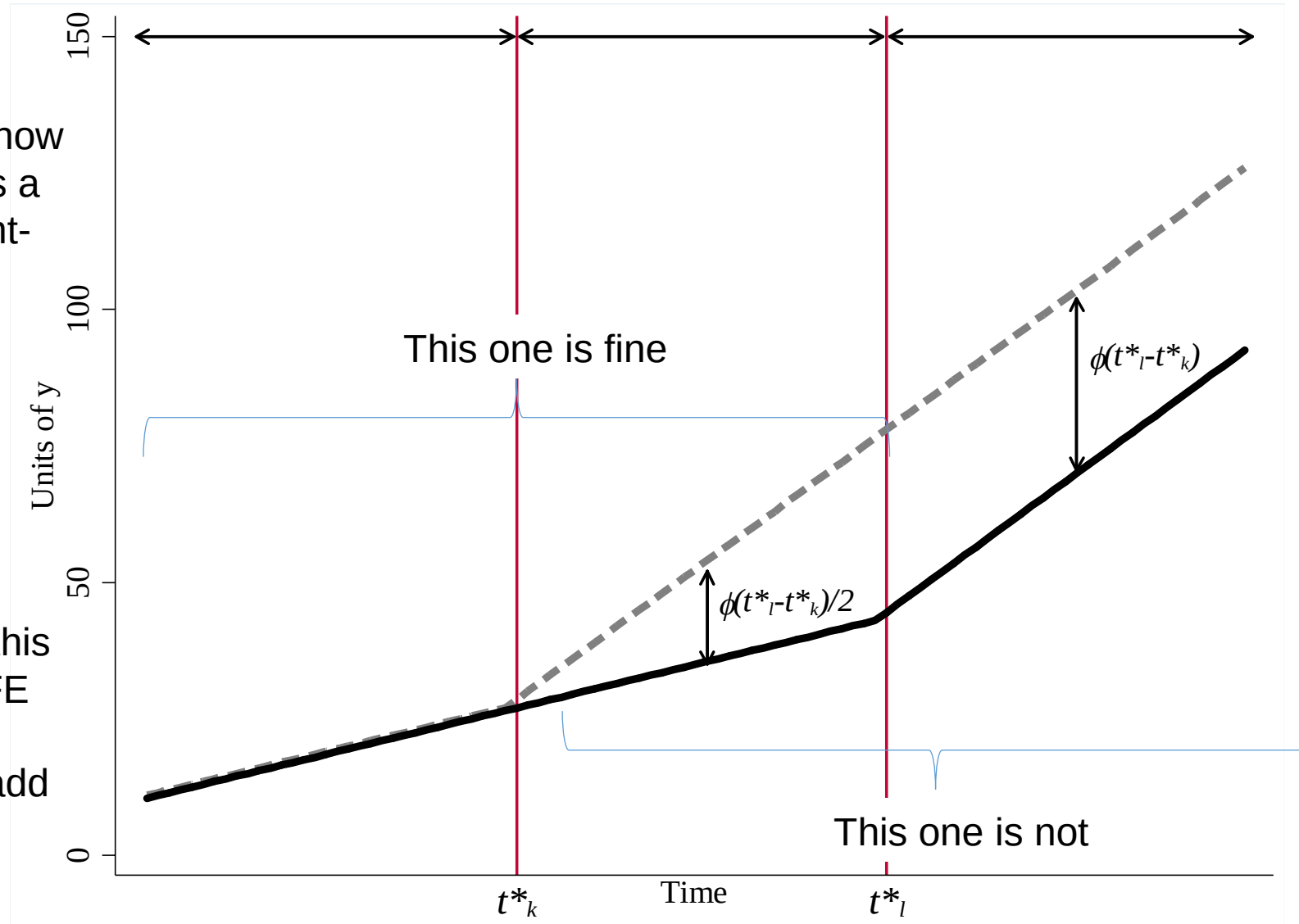
bacondecomp



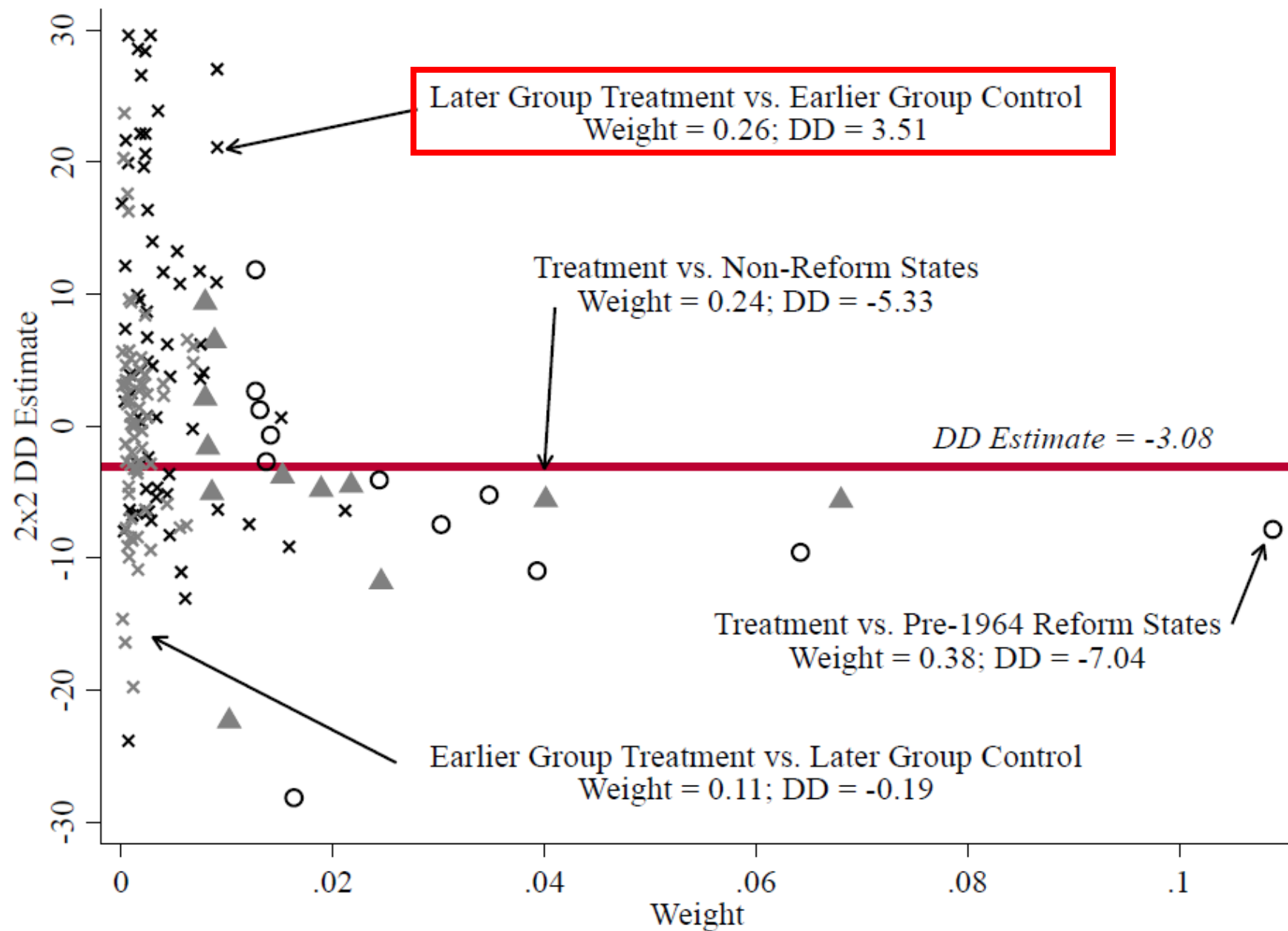
Biggest TWFE Problem: Time-Varying Homogeneous Effects

**Sun and Abraham show that a version of this is a problem even for event-study specifications.

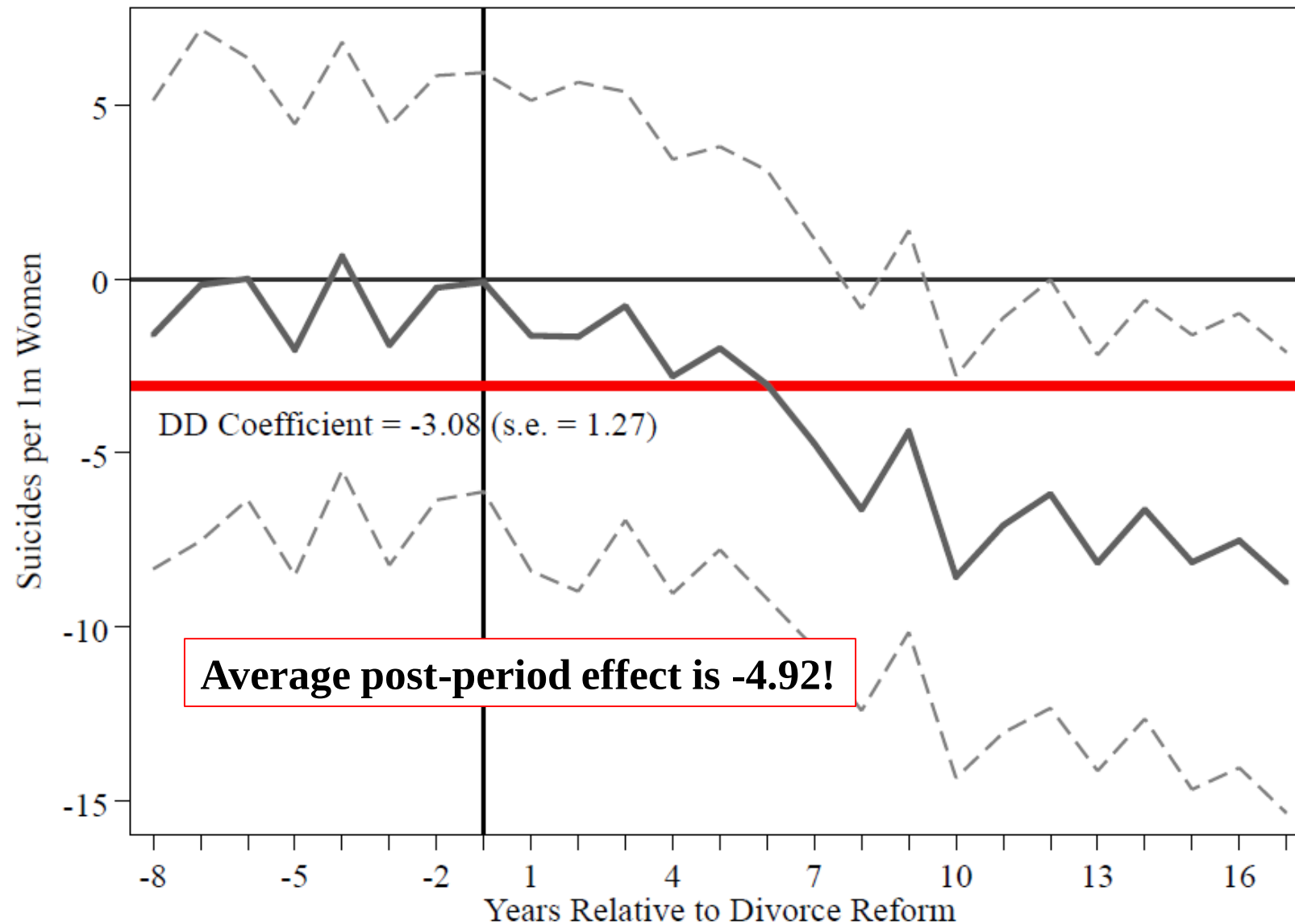
**Wooldridge argues this is not a failure of TWFE per se, but of the specification (ie. just add more flexible terms)



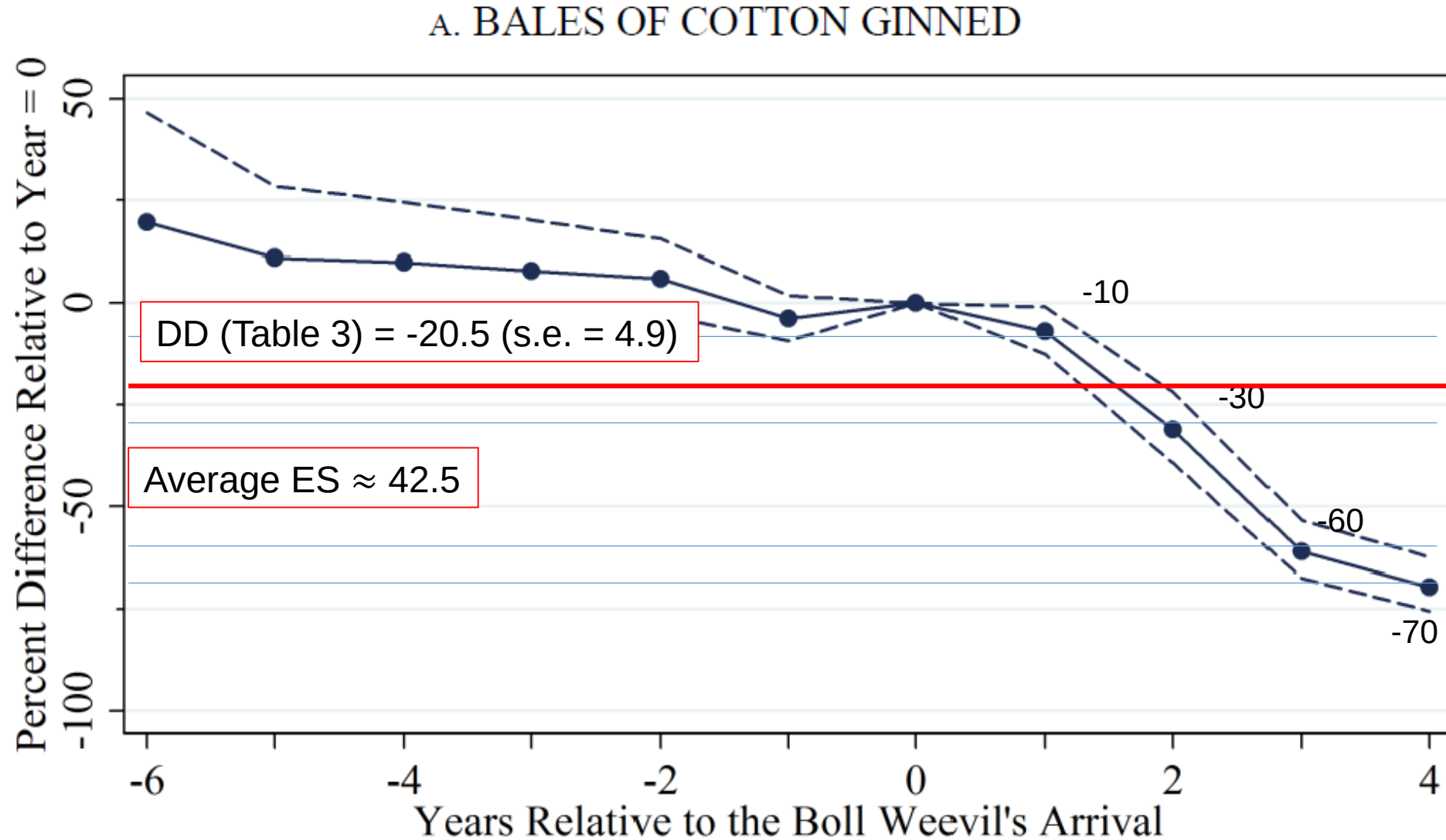
Heterogeneity bias in bacondecomp



Seeing it in the event-study: SW (2006)

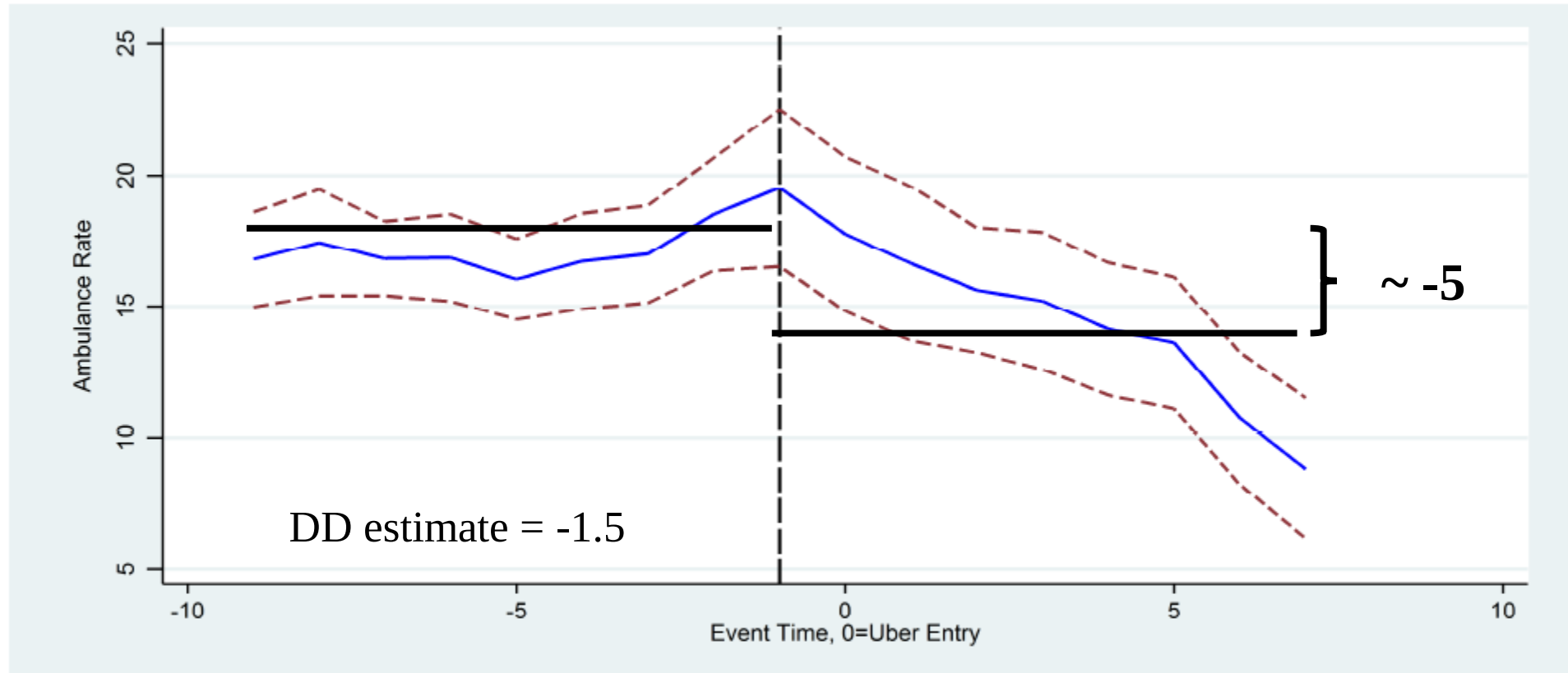


Seeing it in the event-study: Baker (2015)



Seeing it in the event-study: Moskatek and Slusky (2017)

Figure 2: Event Study



Notes: The dashed red lines show the 95% confidence interval. The vertical dashed black line shows the last quarter when Uber was not available at least some of the time.

But: don't do bacondecomp

- It isn't a robustness check of anything
- It doesn't work (that well) with covariates
- It sows confusion about “negative weights”
- **It diagnoses a problem with many easy fixes**
- (It is very pedagogically valuable, though!)

The Effects of a Large-Scale Mental Health Reform: Evidence from Brazil[†]

By MATEUS DIAS AND LUIZ FELIPE FONTES*

Our empirical strategy exploits the rollout of CAPS across Brazilian municipalities in a difference-in-differences (DID) framework. Following De Chaisemartin

What does “ATT” mean when there are multiple treated groups?

General set up

- Units:
 - Time (years):
-

$$i = 1, \dots, N$$

$$t = \{1, \dots, T\}$$

- Treatment status:
- Treatment path:
- Treatment group:

$$D_{i,t} = 1 \text{ if } i \text{ is treated in } t$$

$$\mathbf{D}_i = \{D_{i,1}, \dots, D_{i,T}\}$$

Assume $D_{i,0} = 0$ for all i

$$G_i = g \text{ if } i \text{ is treated at time } g$$

$$G_i = \infty \text{ if } i \text{ is "never" treated}$$

$$g \in \{2, \dots, T\} \cup \infty$$

$$D_{it} \equiv 1\{t \geq G_i\}$$

Treatment time:

Outcome:

$$Y_{i,t}$$

Potential and observed outcomes

- $Y_{i,t}(g)$ outcome of unit i in period t if treated at time g

$$Y_{i,t} = \sum_g 1\{G_i = g\}Y_{i,t}(g)$$

	$G_i = \infty$	$G_i = 2$	$G_i = g$	$G_i = T$
$t = 1$	$Y_{i,t}(\infty)$	$Y_{i,t}(\infty)$	$Y_{i,t}(\infty)$	$Y_{i,t}(\infty)$
$t = 2$	$Y_{i,t}(\infty)$	$Y_{i,t}(2)$	$Y_{i,t}(\infty)$	$Y_{i,t}(\infty)$
$\vdots$				
$t = T - 1$	$Y_{i,t}(\infty)$	$Y_{i,t}(2)$	$Y_{i,t}(g)$	$Y_{i,t}(\infty)$
$t = T$	$Y_{i,t}(\infty)$	$Y_{i,t}(2)$	$Y_{i,t}(g)$	$Y_{i,t}(T)$

Potential and observed outcomes

- $Y_{i,t}(g)$ outcome of unit i in period t if treated at time g

$$Y_{i,t} = \sum_g 1\{G_i = g\} Y_{i,t}(g)$$

	$G_i = \infty$	$G_i = 2$	$G_i = g$	$G_i = T$
$t = 1$	$Y_{i,t}(\infty)$	$Y_{i,t}(\infty)$	$Y_{i,t}(\infty)$	$Y_{i,t}(\infty)$
$t = 2$	$Y_{i,t}(\infty)$	$Y_{i,t}(2)$	$Y_{i,t}(\infty)$	$Y_{i,t}(\infty)$
$\vdots$				
$t = T - 1$	$Y_{i,t}(\infty)$	$Y_{i,t}(2)$	$Y_{i,t}(g)$	$Y_{i,t}(\infty)$
$t = T$	$Y_{i,t}(\infty)$	$Y_{i,t}(2)$	$Y_{i,t}(g)$	$Y_{i,t}(T)$

No Anticipation:

We would see
this homicide rate
even if there had
been no CAPS

Group-Time ATTs: $ATT(g, t)$

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

Being averaged (again,
this is a choice!)

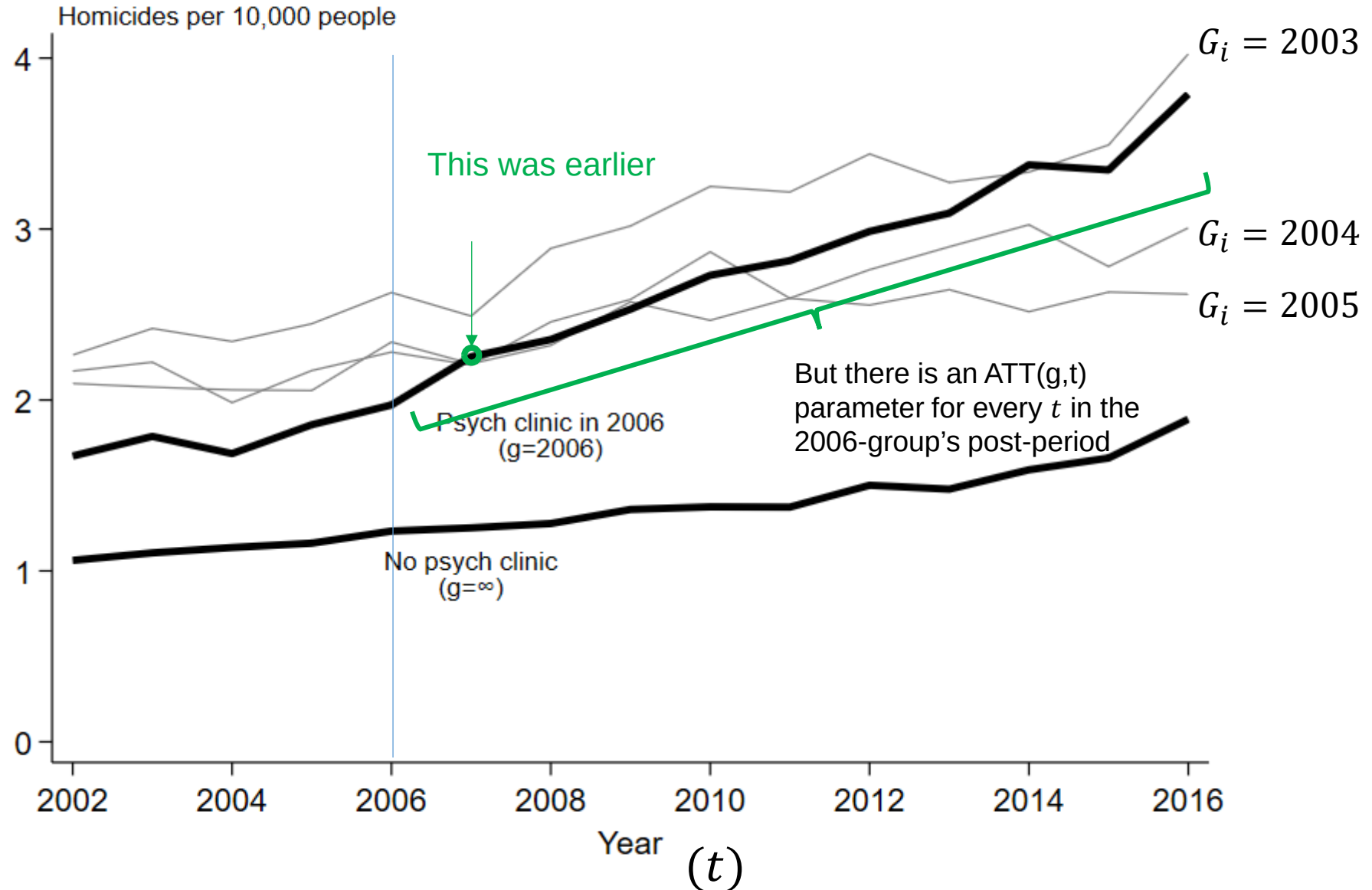
The causal effect on unit
 i in period t of having
been treated at time g

Only over the units who really
did get treated at time g

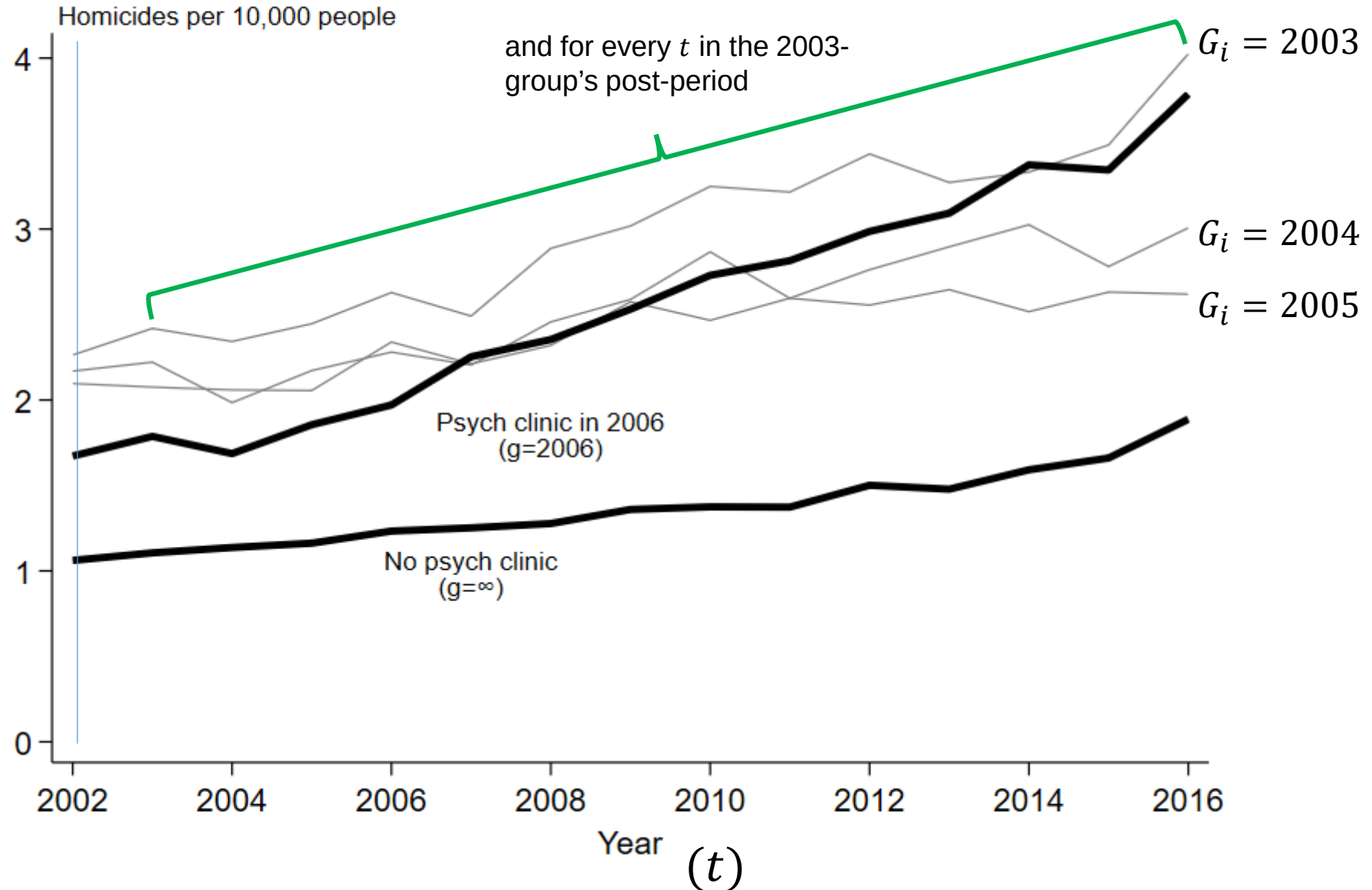
“Building block parameter”:

A simple DiD type of ATT parameter with one treatment (time g)
and one “post” period (time t), averaged over one group (group g)

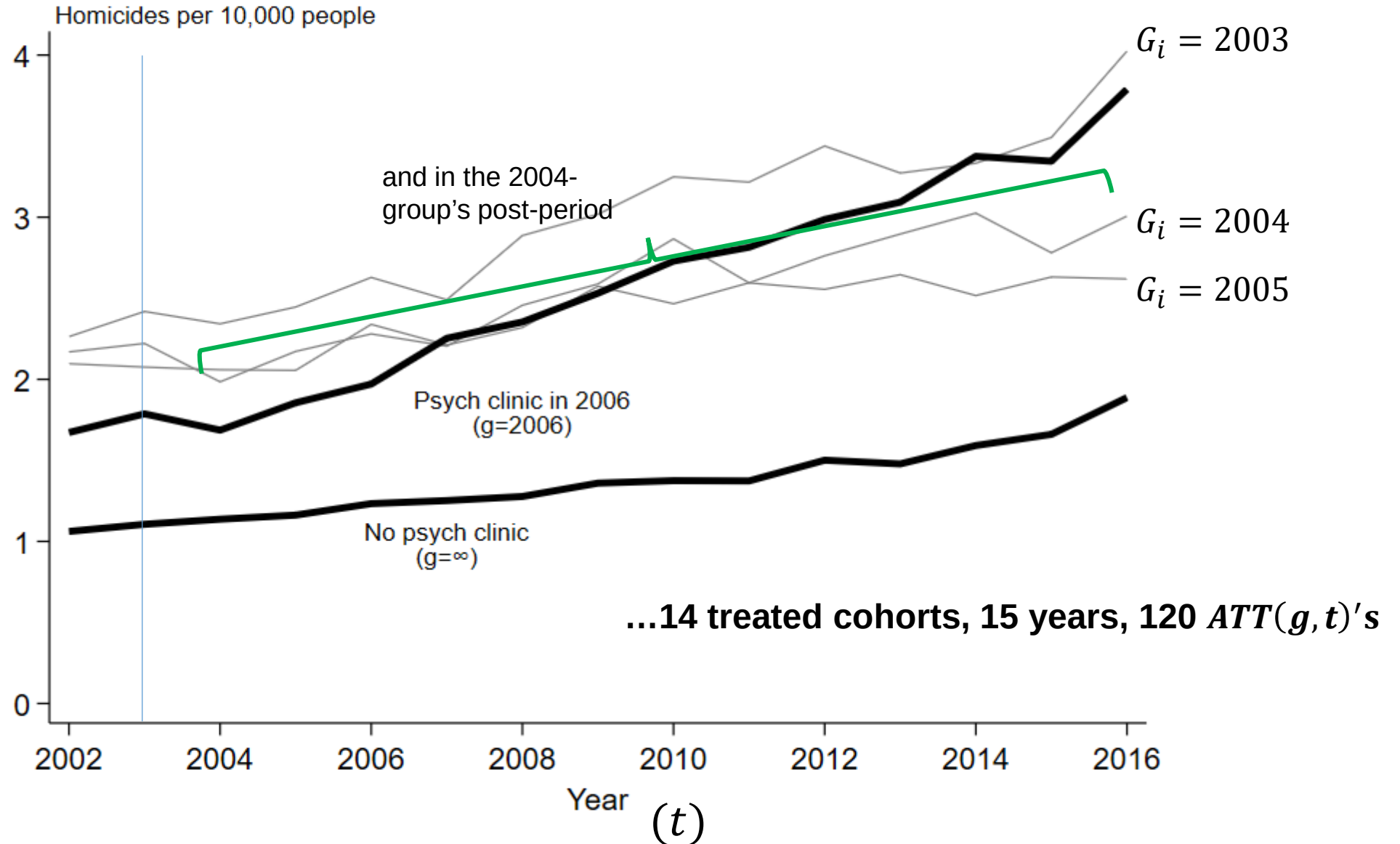
Group-Time ATTs: $ATT(g, t)$



Group-Time ATTs: $ATT(g, t)$



Group-Time ATTs: $ATT(g, t)$



All DiD analyses center on “building blocks”

1. That is where identification lives and dies (next section)
2. That is how what you should think about estimating (section after that)
3. You **choose** up front what combination of building blocks you want.

Group-Time ATTs: $ATT(g, t)$

Imagine (for simplicity) that this were the roll-out and the data ended in 2008

G_i

2006

$ATT(2006, 2006)$ $ATT(2006, 2007)$ $ATT(2006, 2008)$

2007

$ATT(2007, 2007)$ $ATT(2007, 2008)$

2008

$ATT(2008, 2008)$



Sample period (t)

Group-Time ATTs: $ATT(g, t)$

G_i These are the “instantaneous”
effects for each timing group

2006 $ATT(2006, 2006)$ $ATT(2006, 2007)$ $ATT(2006, 2008)$

2007 $ATT(2007, 2007)$ $ATT(2007, 2008)$

2008 $ATT(2008, 2008)$



Sample period (t)

Group-Time ATTs: $ATT(g, t)$

These are the “second-year” effects for each timing group, *but you can only hope to estimate them for $G=2006$ and $G=2007$!*

G_i

These are the “instantaneous” effects for each timing group

2006

$ATT(2006, 2006)$ **$ATT(2006, 2007)$** $ATT(2006, 2008)$

2007

$ATT(2007, 2007)$ **$ATT(2007, 2008)$**

2008

$ATT(2008, 2008)$



Sample period (t)

Group-Time ATTs: $ATT(g, t)$

G_i

These are the “instantaneous” effects for each timing group

These are the “second-year” effects for each timing group, *but you can only hope to estimate them for $G=2006$ and $G=2007$!*

The “longest run” effect can only be estimated for the earliest group!

2006

$ATT(2006, 2006)$ $ATT(2006, 2007)$ **$ATT(2006, 2008)$**

2007

$ATT(2007, 2007)$ $ATT(2007, 2008)$

2008

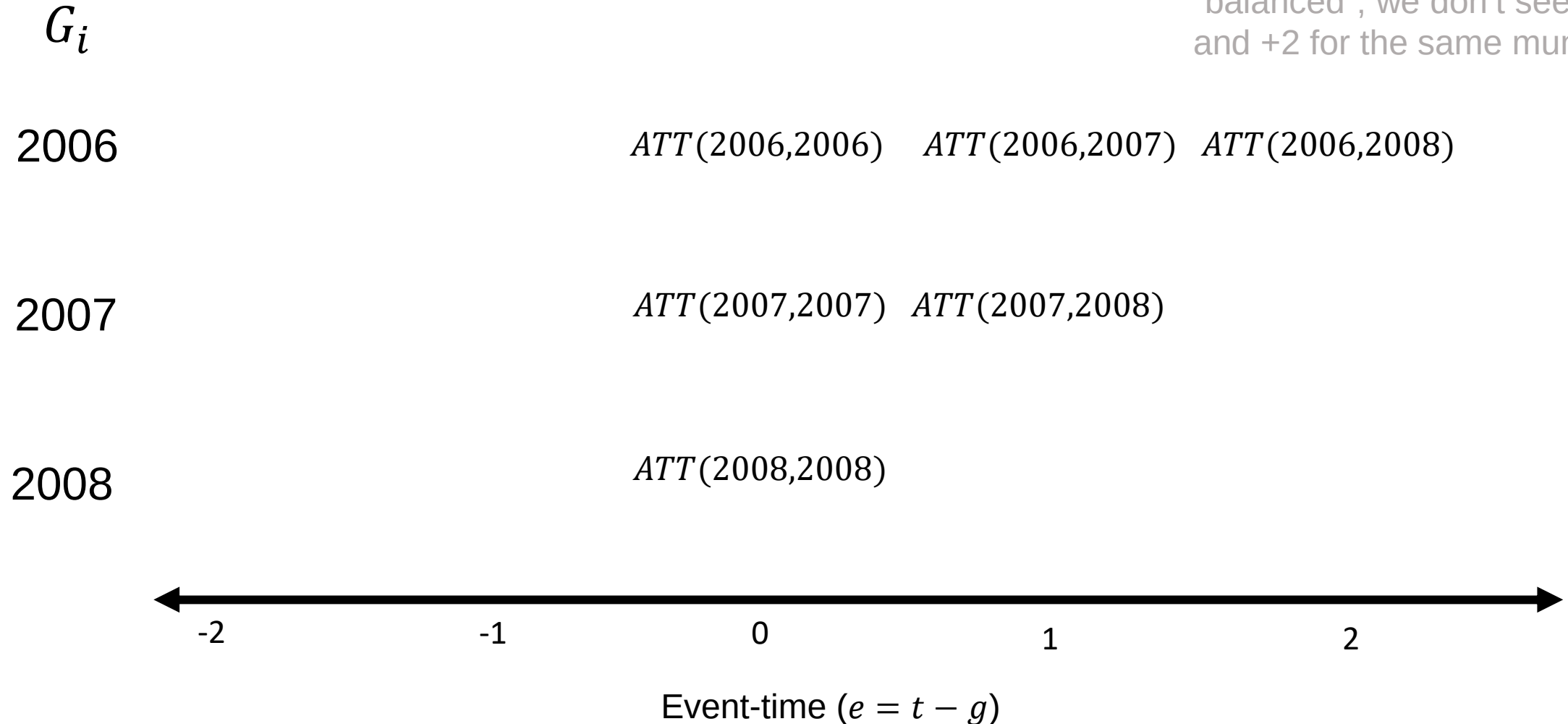
$ATT(2008, 2008)$



Sample period (t)

Event-Time ATTs: $ATT(e)$

Note: These event-times are not “balanced”; we don’t see +0, +1, and +2 for the same municipalities!



Overall ATT: $\overline{ATT}$

G_i

Or you can average them all, but note that groups contribute different numbers of parameters!

Aggregation matters and it starts with the target parameter you **choose!**

2006

$ATT(2006,2006)$ $ATT(2006,2007)$ $ATT(2006,2008)$

2007

$ATT(2007,2007)$ $ATT(2007,2008)$

2008

$ATT(2008,2008)$



Event-time ($e = t - g$)

The Effects of a Large-Scale Mental Health Reform: Evidence from Brazil[†]

By MATEUS DIAS AND LUIZ FELIPE FONTES*

ATT(0)

The average first-year treatment effect of CAPS clinics on homicide rates among treated municipalities.

$\overline{ATT}$

The average treatment effect in treated years in treated municipalities.

Identifying the building blocks

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

$$E[Y_{it}(g) | G_i = g] - E[Y_{it}(\infty) | G_i = g]$$

Can identify just like a 2x2:

Periods t and $g - 1$

Groups $G_i = g$ and $G_i \in \mathcal{G}^{comp}$ ← more in a second

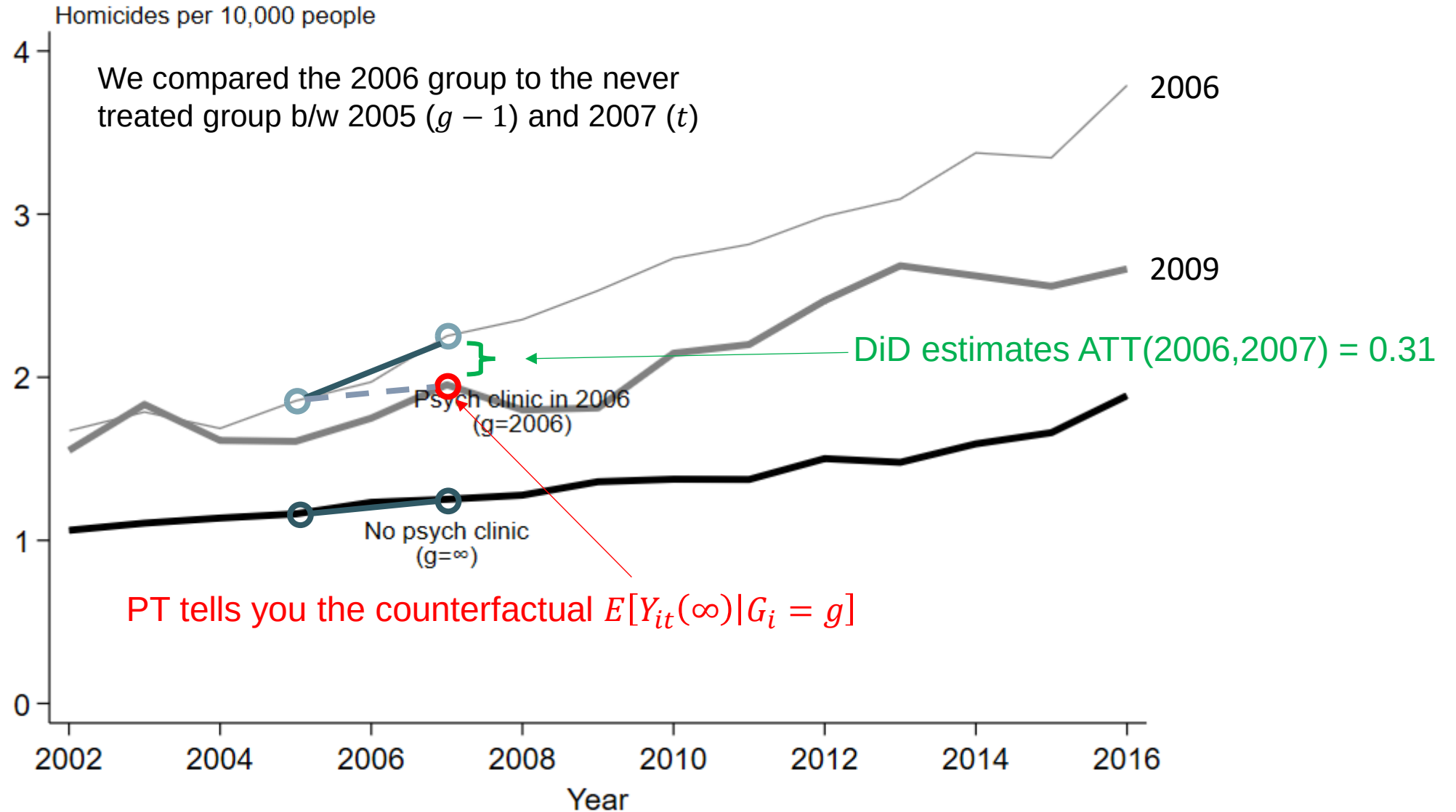
+ parallel trends!

$$\underbrace{E[Y_{it}(\infty) | G_i \in \mathcal{G}^{comp}] - E[Y_{i,g-1}(\infty) | G_i \in \mathcal{G}^{comp}]}_{\text{Observed change in mean outcomes for comparison units from } g - 1 \text{ to } t} = \underbrace{E[Y_{it}(\infty) | G_i = g] - E[Y_{i,g-1}(\infty) | G_i = g]}_{\text{Unobserved change in mean untreated potential outcomes for units in group } g \text{ from } g - 1 \text{ to } t}$$

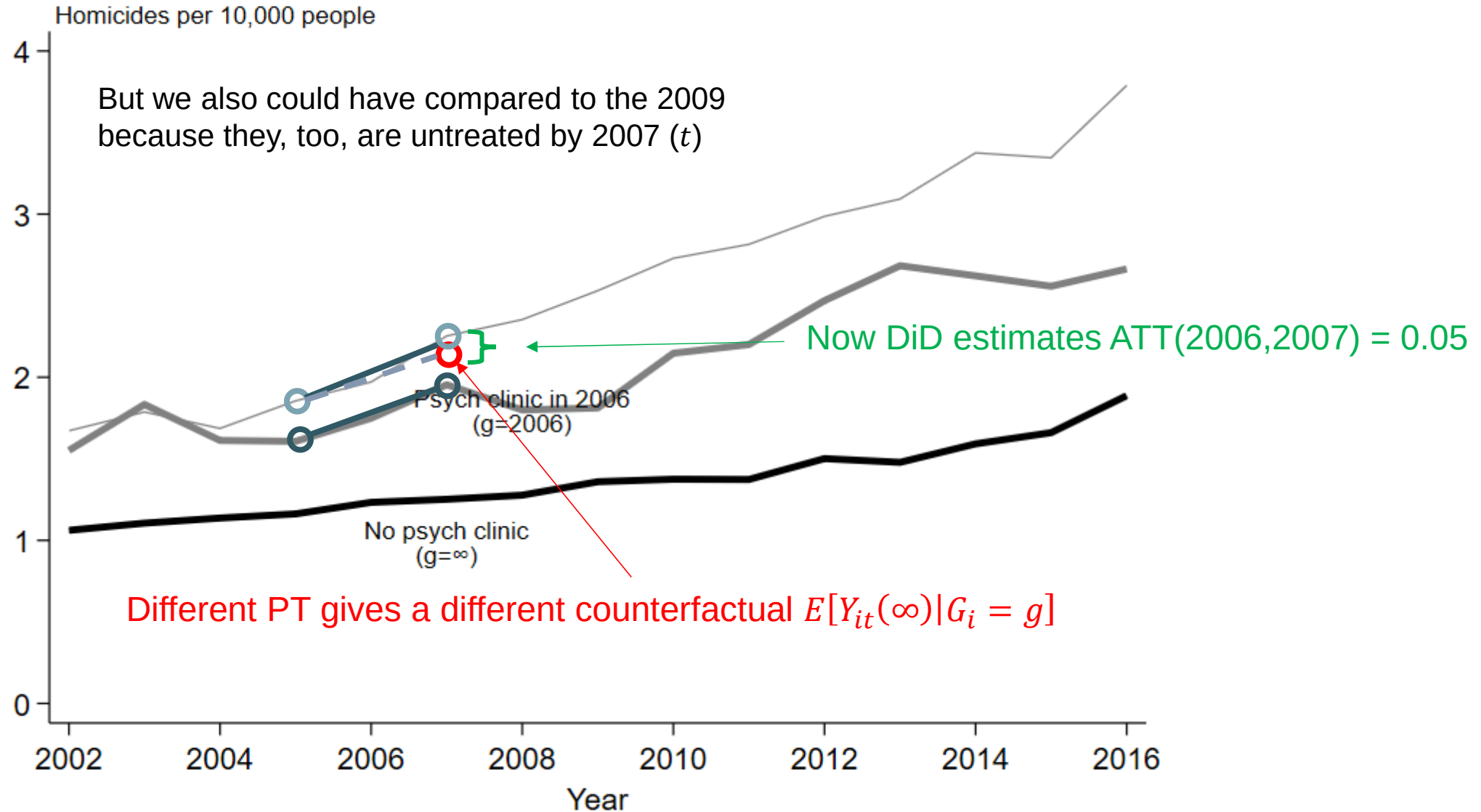
Observed change in mean outcomes
for comparison units from $g - 1$ to t

Unobserved change in mean
untreated potential outcomes
for units in group g from $g - 1$ to t

Parallel Trends for $ATT(2006, 2007)$



Parallel Trends for $ATT(2006, 2007)$



Many PTs in a staggered design

Now you have **choices** about comparison units in $\mathcal{G}^{comp}$:

$$E[Y_{it}(\infty)|G_i \in \mathcal{G}^{comp}] - E[Y_{i\textcolor{brown}{g}-1}(\infty)|G_i \in \mathcal{G}^{comp}] = \textcolor{red}{E[Y_{it}(\infty)|G_i = \textcolor{brown}{g}]} - E[Y_{i\textcolor{brown}{g}-1}(\infty)|G_i = \textcolor{brown}{g}]$$

Never-treated units:

$$\mathcal{G}^{nev} = \{g = \infty\} \quad \text{Same comparison group for all } ATT(g, t)!$$

Not-yet-treated units:

$$\mathcal{G}^{ny} = \{g: t < g < \infty\} \quad \text{Shrinking comparison t grows.}$$

All untreated units at time t:

$$\mathcal{G}^{un} = \mathcal{G}^{ny} \cup \mathcal{G}^{nev} = \{g: t < g\} \quad \text{Biggest comparison group at time } t.$$

Last treated group:

$$\mathcal{G}^{max} = \{g = g^{max}\} \quad \text{Sun and Abraham}$$

Choice of comparison group and “willingness to believe”

You need a reason to believe.

- Why did treatment happen? Did someone chose it? What constrained them? What were they maximizing? What did they know?
- Seminal example: [Ashenfelter and Card \(1985\)](#); recent discussion [Ghanem, Sant'Anna, and Wuthrich\(2022\)](#)

Choice of comparison group and “willingness to believe”

You need a reason to believe.

- Why did treatment happen? Did someone chose it? What constrained them? What were they maximizing? What did they know?
- Seminal example: [Ashenfelter and Card \(1985\)](#); recent discussion [Ghanem, Sant’Anna, and Wuthrich\(2022\)](#)

“Parallel trends”: mean $Y(\infty)$ would have changed identically in T/C groups.

- Levels *can* be different, but sometimes that makes PT untenable
- Not typically invariant to transformations ([Roth & Sant’Anna 2023](#))

Choice of comparison group and “willingness to believe”

You need a reason to believe.

- Why did treatment happen? Did someone chose it? What constrained them? What were they maximizing? What did they know?
- Seminal example: [Ashenfelter and Card \(1985\)](#); recent discussion [Ghanem, Sant’Anna, and Wuthrich\(2022\)](#)

“Parallel trends”: mean $Y(\infty)$ would have changed identically in T/C groups.

- Levels *can* be different, but sometimes that makes PT untenable
- Not typically invariant to transformations ([Roth & Sant’Anna 2023](#))

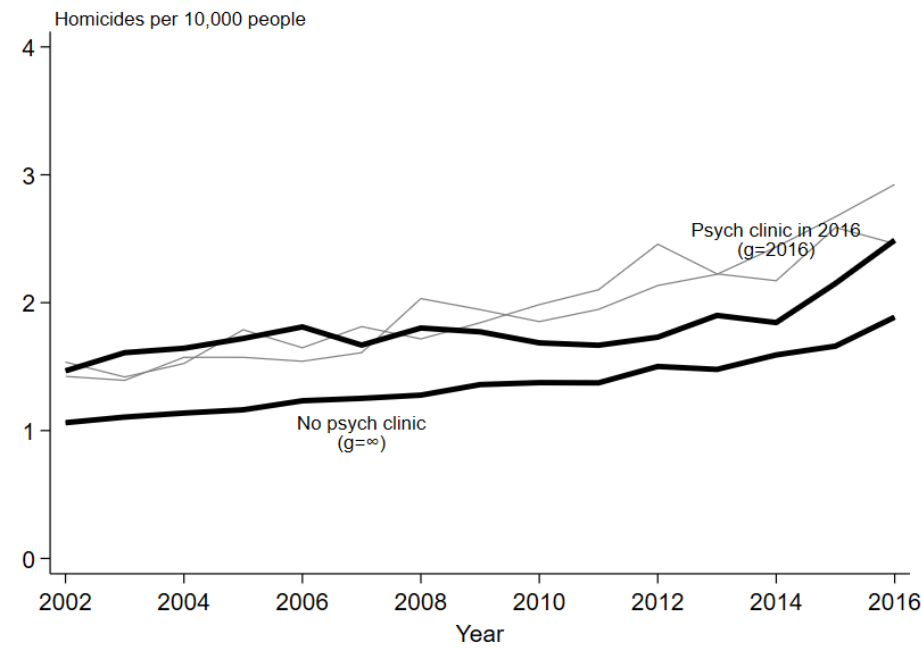
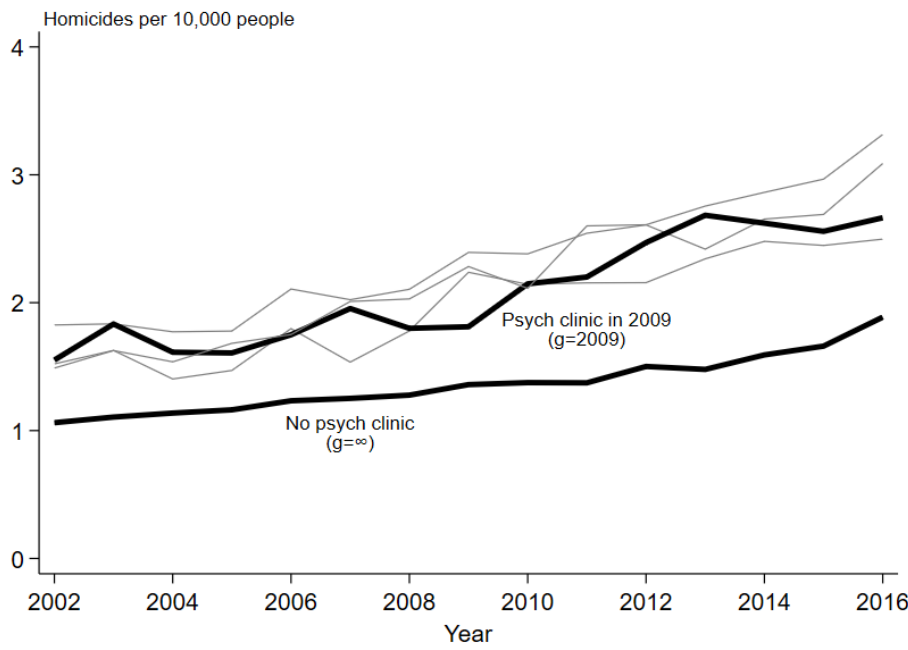
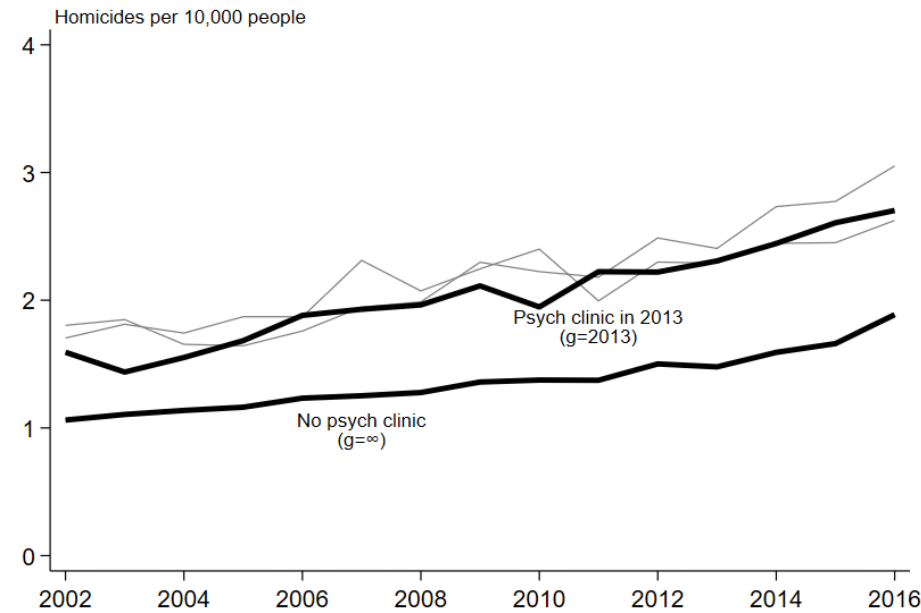
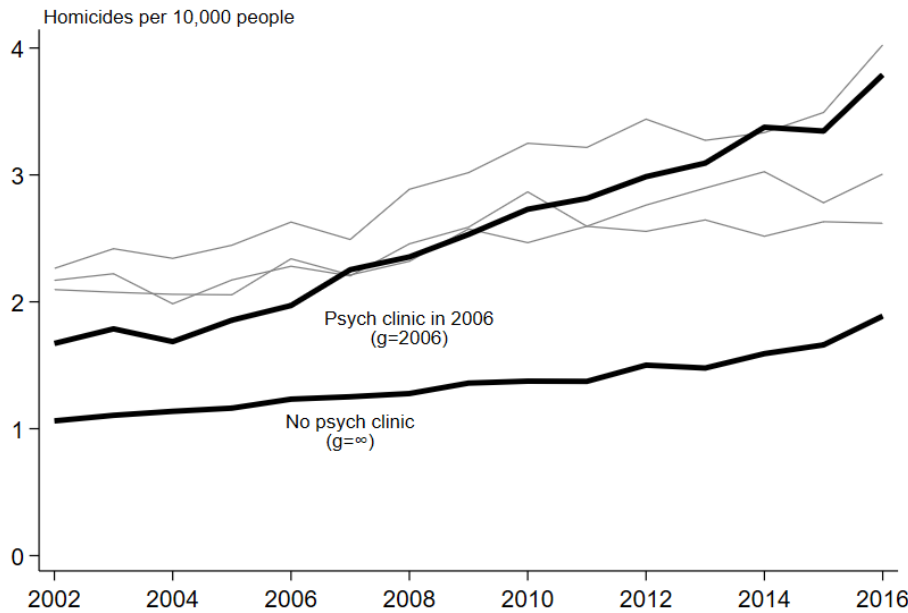
“Bias stability”: selection bias yesterday is the same as it would be today

- Just because a confounder is constant doesn’t mean it’s not a potential PT violation
- Alternative assumptions imply different estimators (ie. AR(1)→LDV; [Ding and Li 2019](#))

What should you actually do?

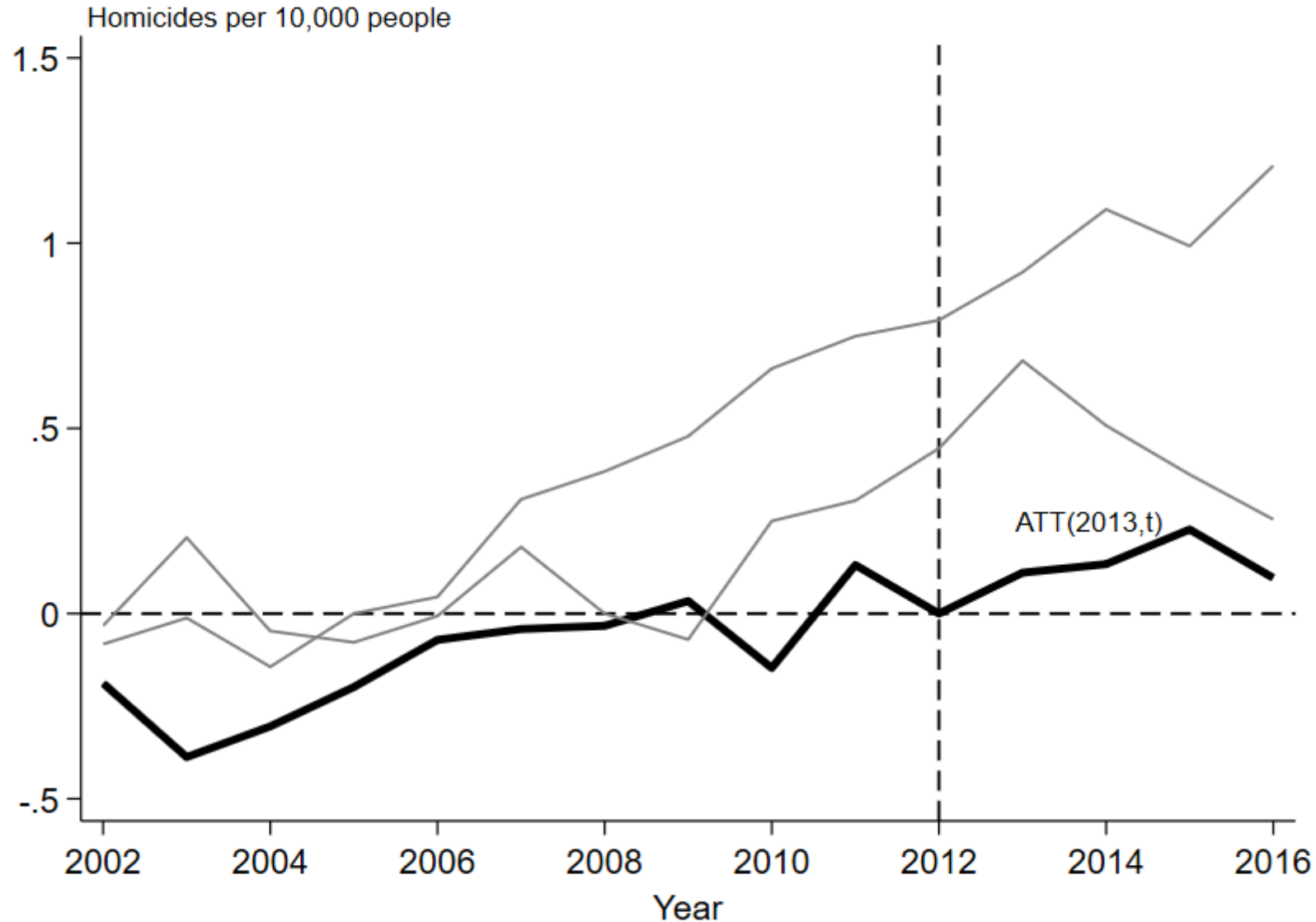
Estimation is “simple”!
Choose a PT
assumption/comparison
group

Go cohort-by-cohort and
estimate $ATT(g,t)$ s by
simple 2x2's

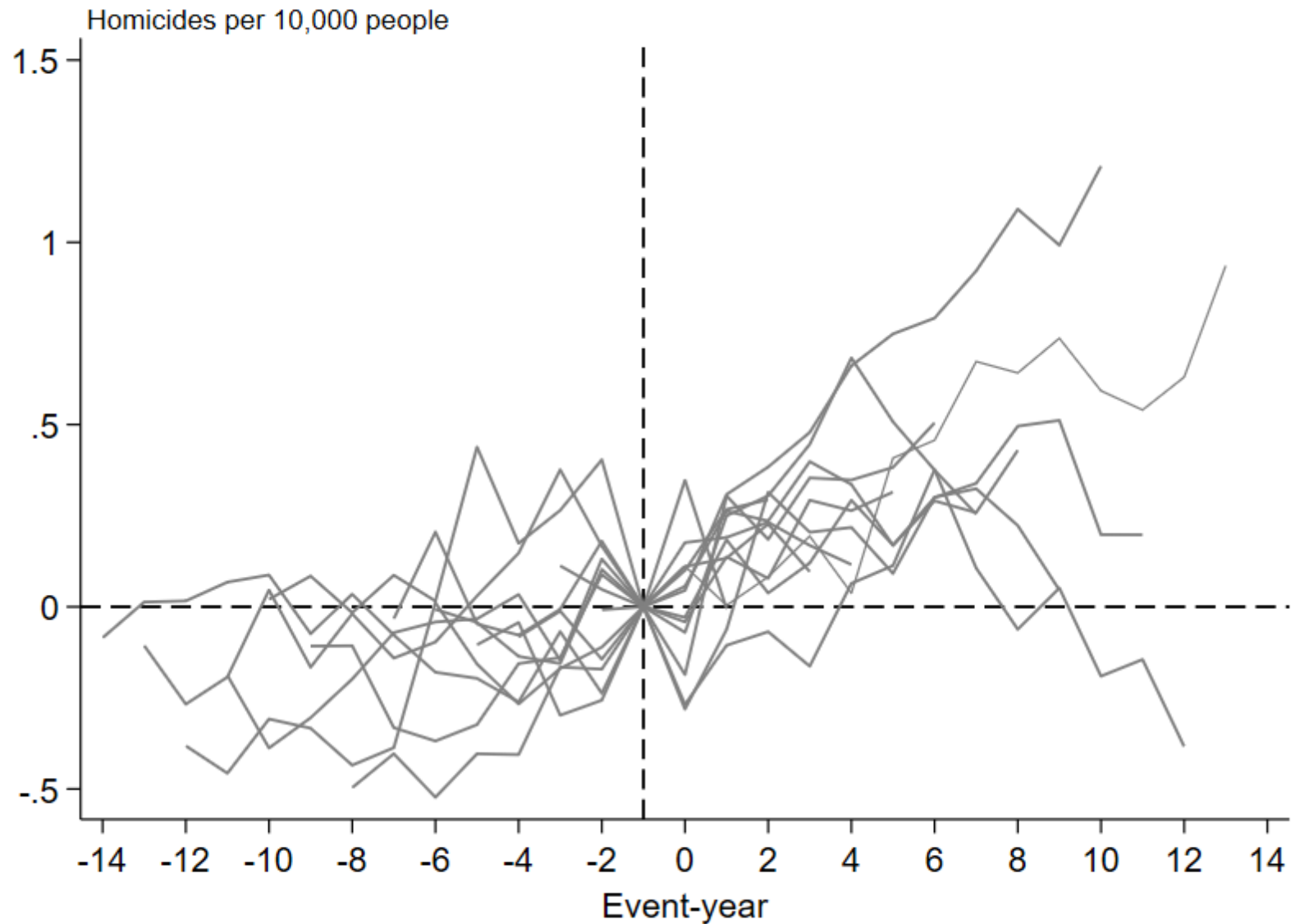


$\widehat{ATT}(g, t)$ plotted against t

Never-treated
comparison group:
 $\mathcal{G}^{comp} = \mathcal{G}^{nev}$



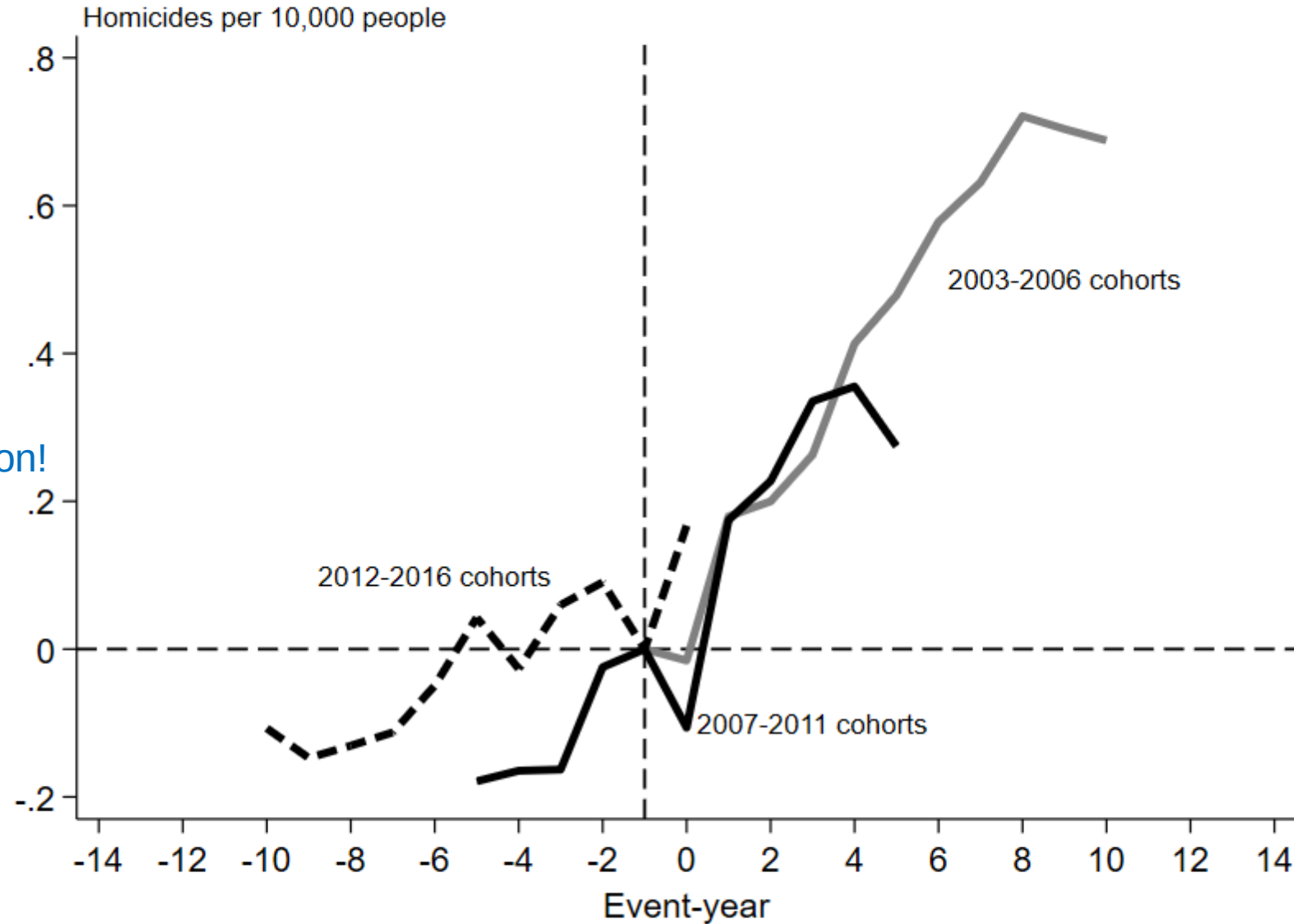
$\widehat{ATT}(g, t)$ plotted against $e = t - g$



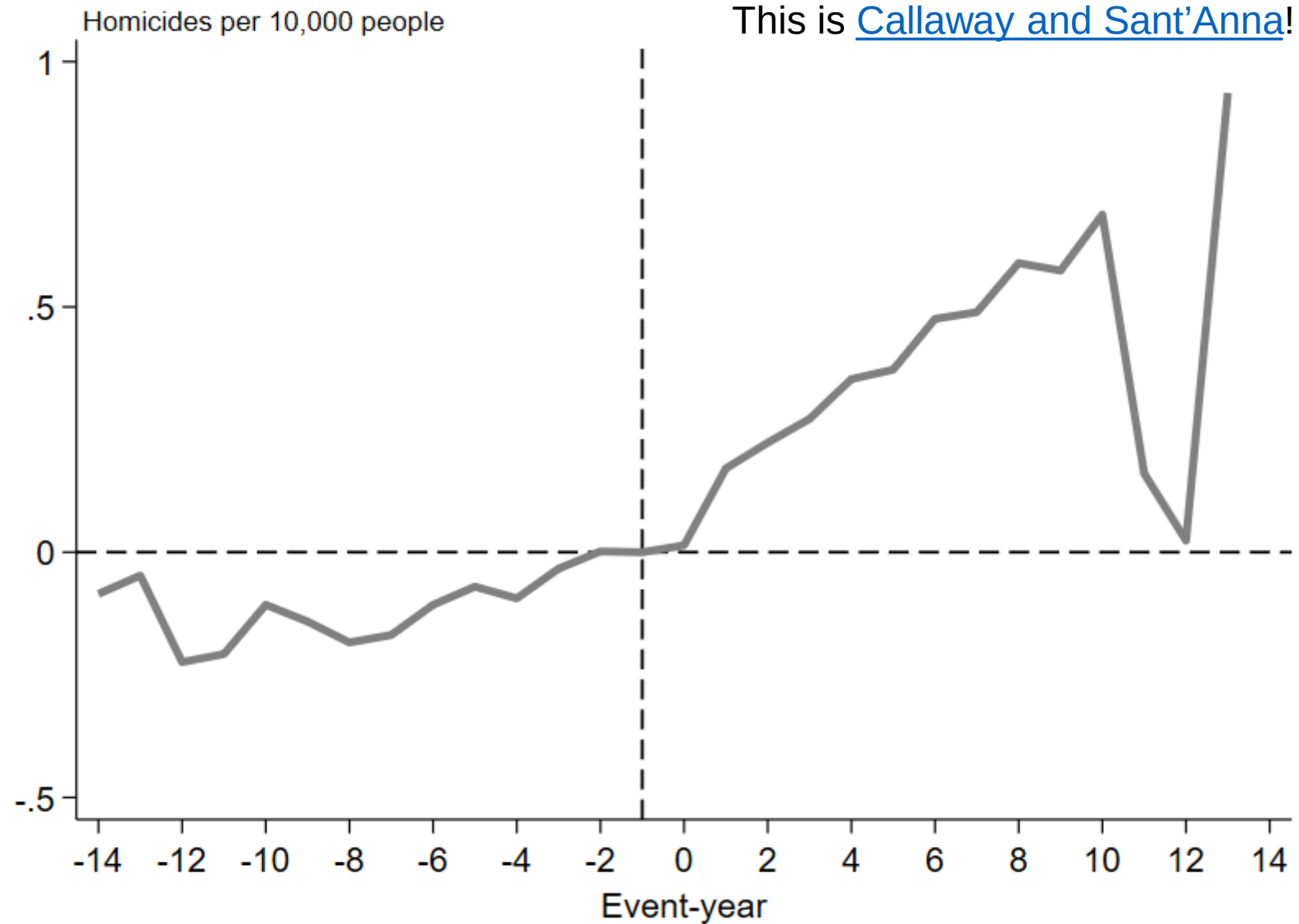
$\sum_{g=g_a}^{g_b} \widehat{w_g} \widehat{ATT}(g, t)$ plotted against $e = t - g$

Share of treated municipalities in cohort g

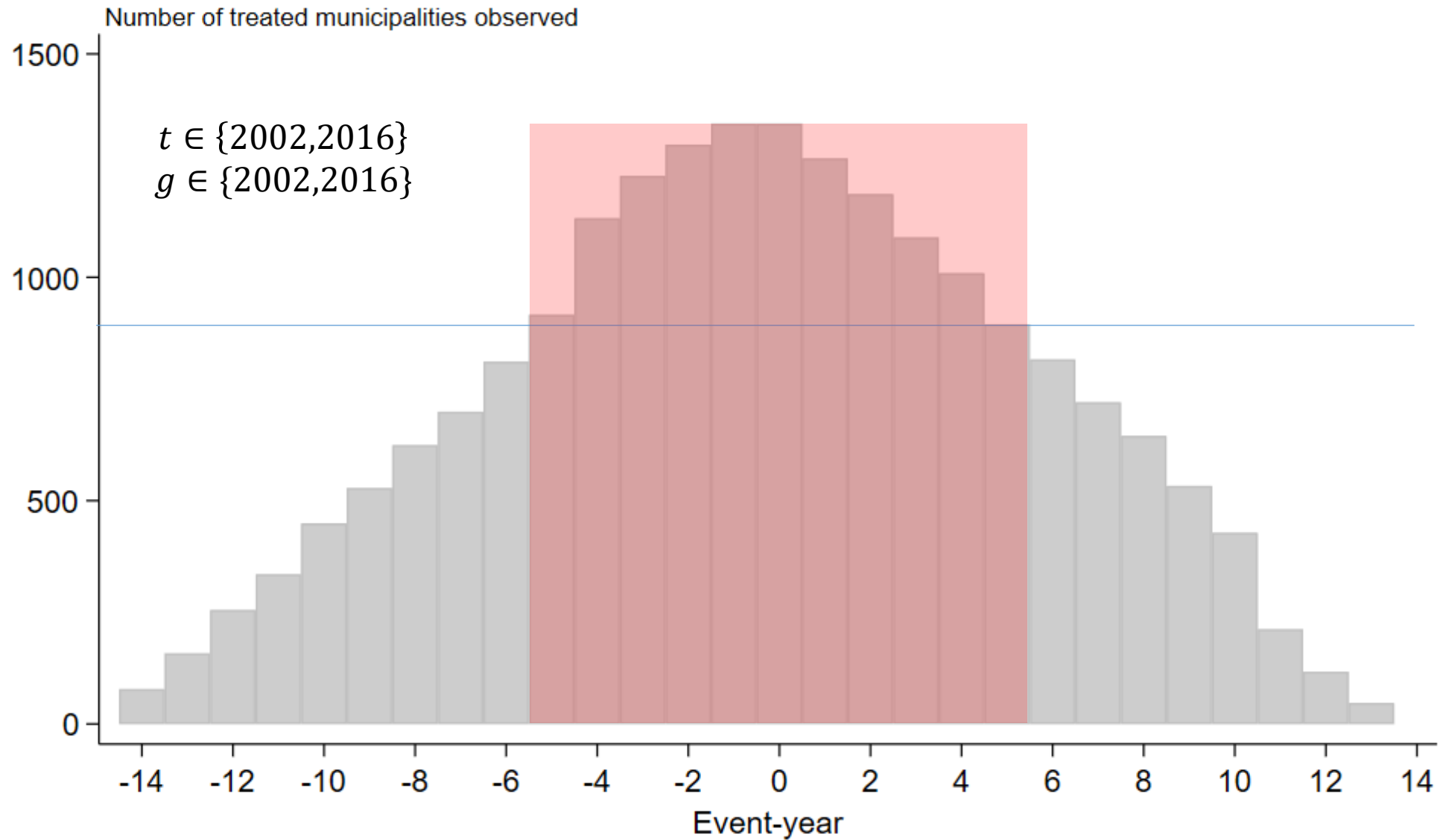
Another role for weighting: aggregation!



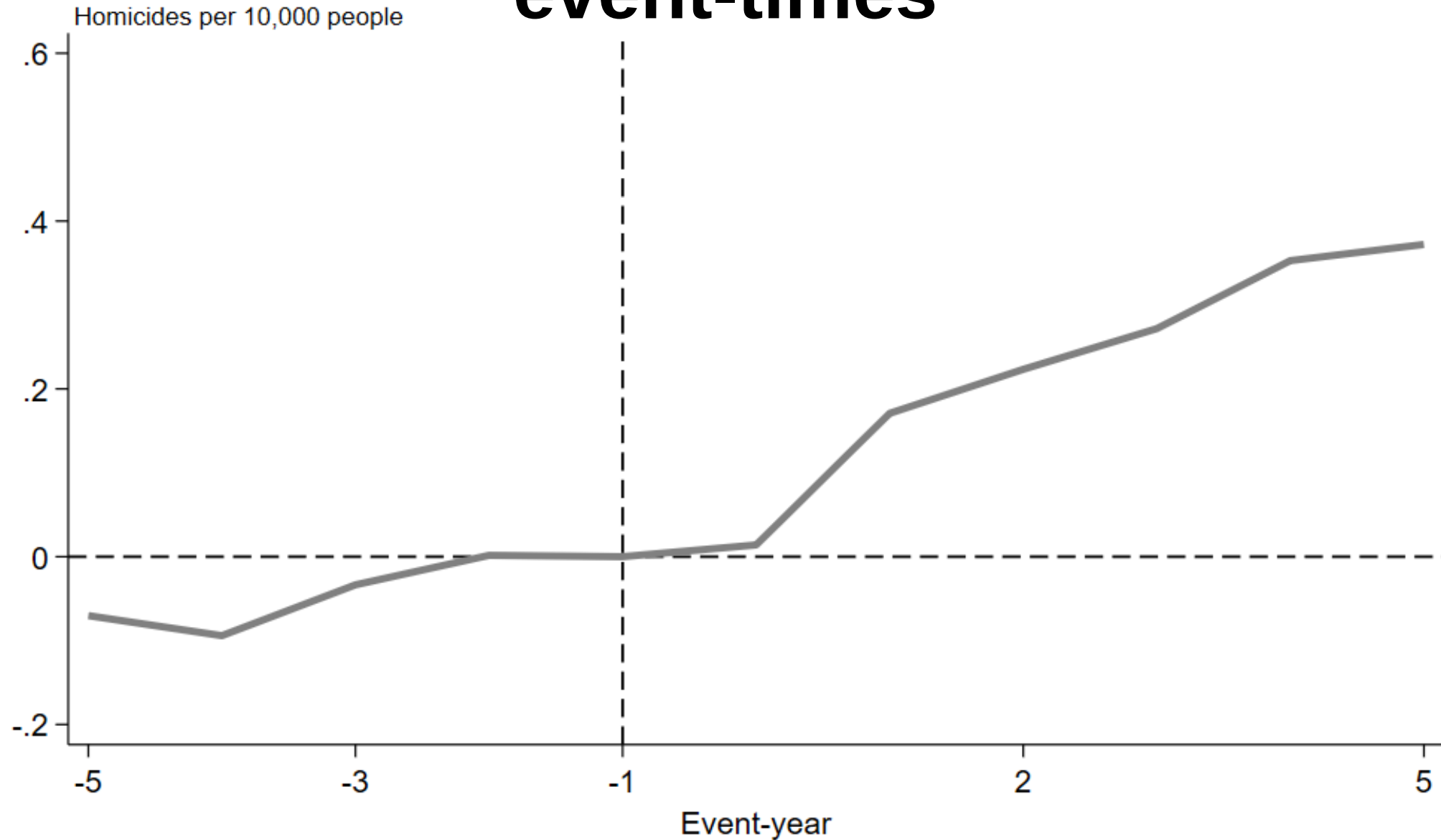
$\widehat{ATT}(e)$ plotted against e



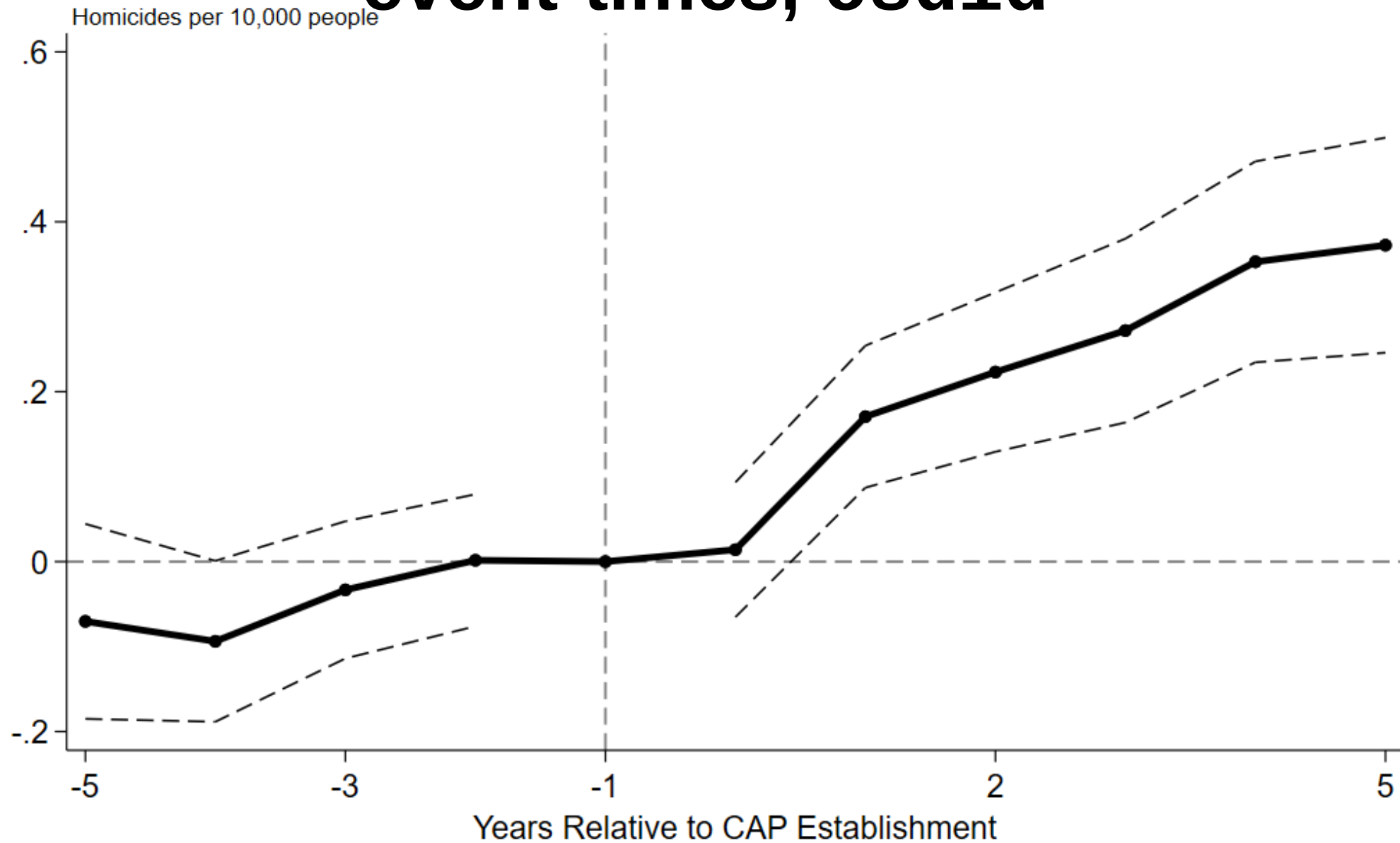
Event-time (im)balance



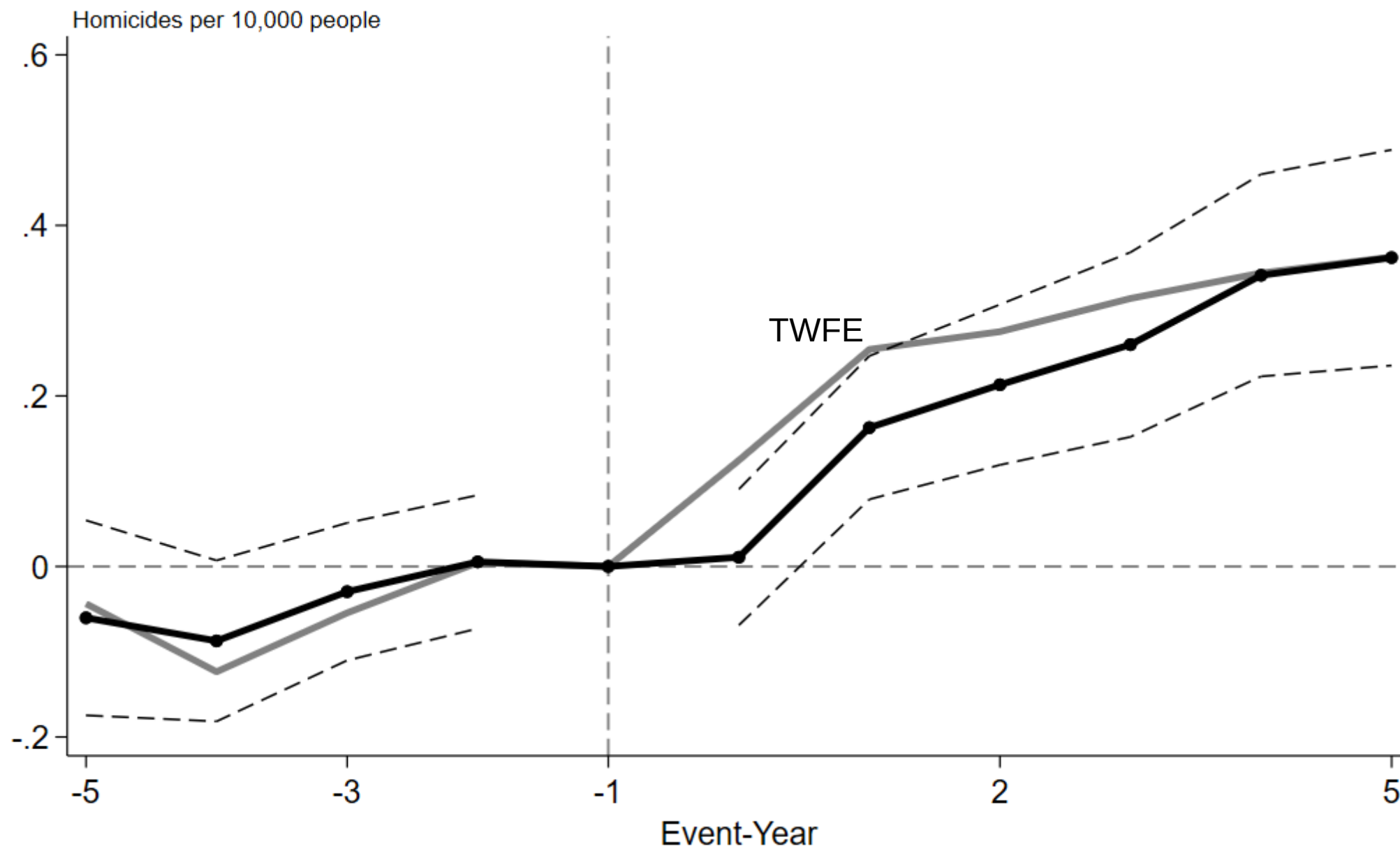
$\widehat{ATT}(e)$ plotted against e , “balanced” event-times



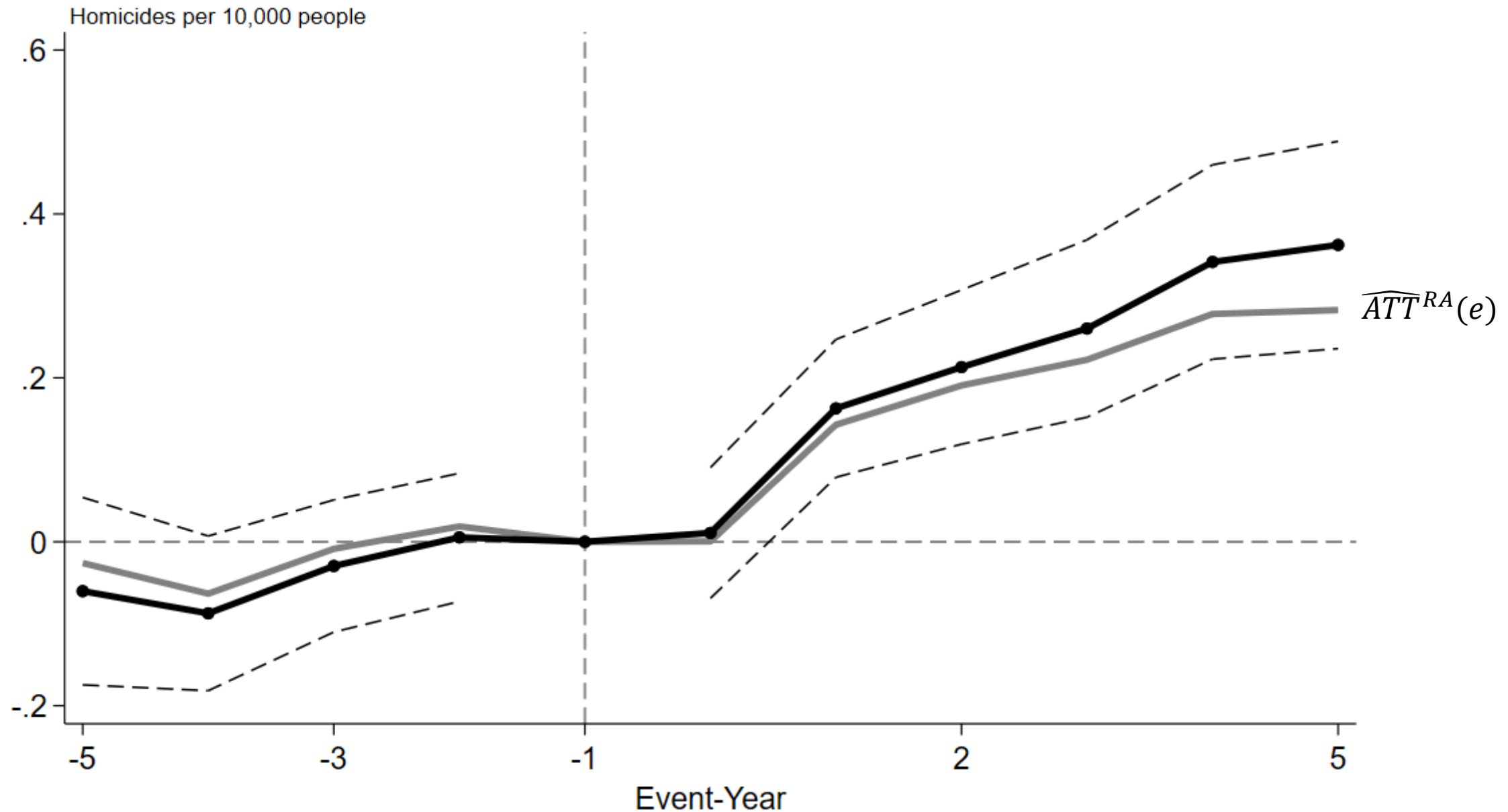
$\widehat{ATT}(e)$ plotted against e , “balanced” event-times, csdid



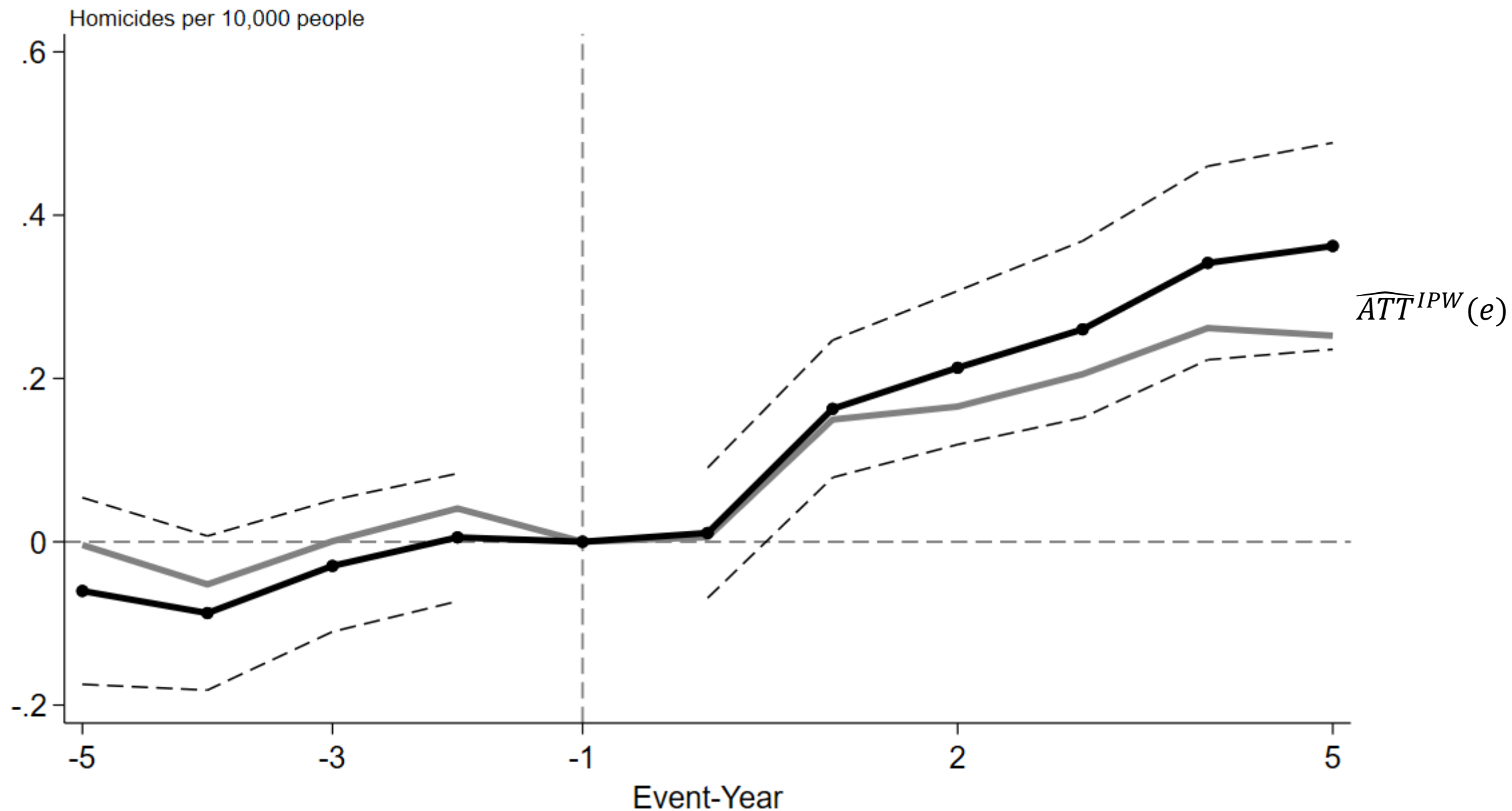
Comparing to TWFE



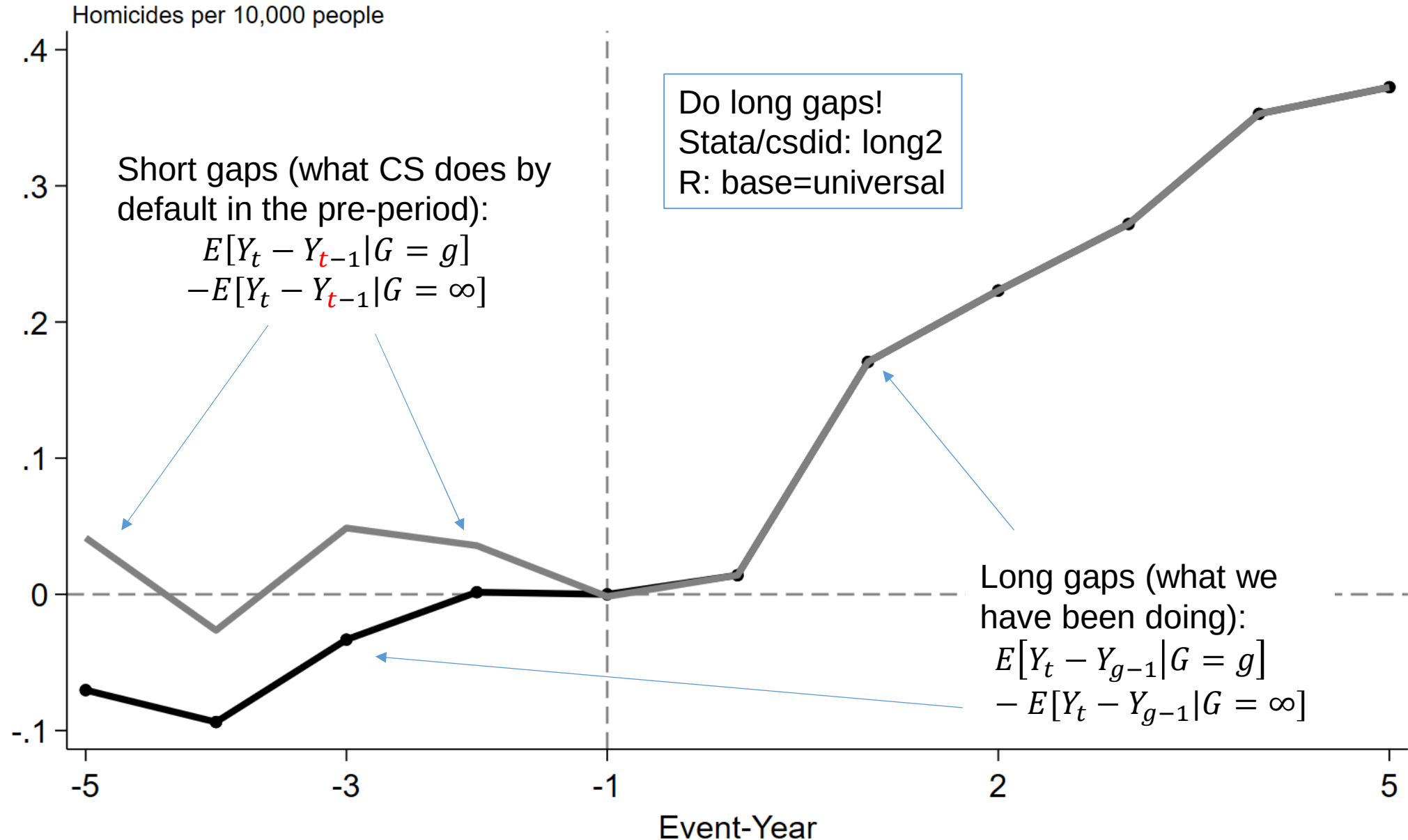
Adding covariates (RA estimator)



Adding covariates (IPW estimator)



Short gaps (no) vs long gaps (yes)

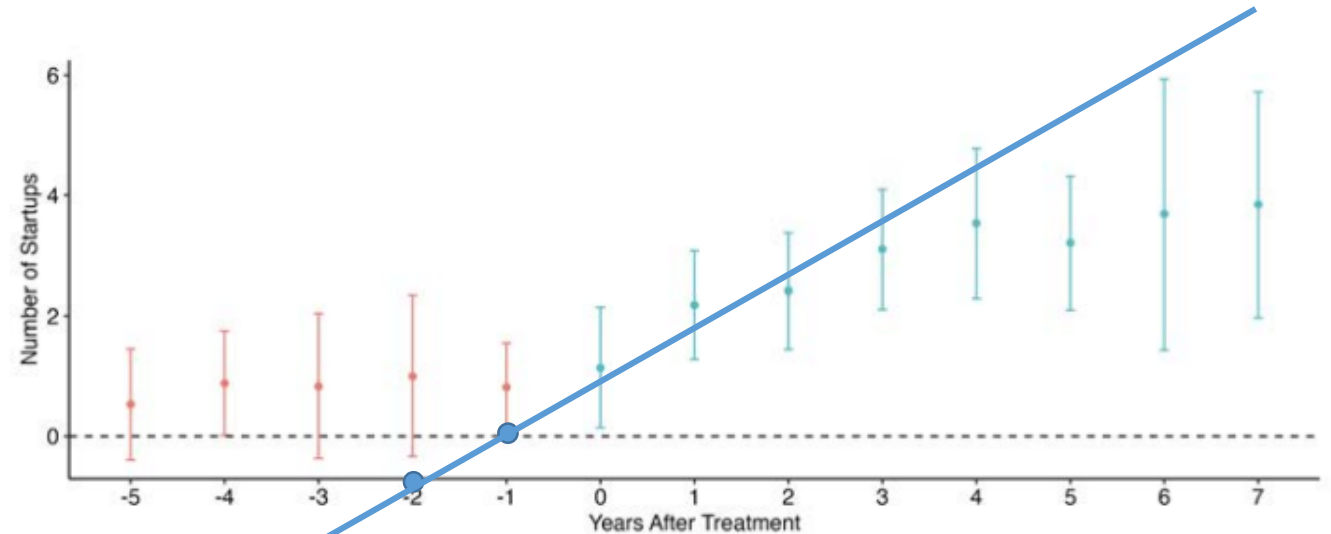


Accidental short gaps in the wild

THIRD PLACES AND NEIGHBORHOOD ENTREPRENEURSHIP:
EVIDENCE FROM STARBUCKS CAFÉS

Jinkyong Choi
Jorge Guzman
Mario L. Small

C. All First Starbucks



Note: This figure shows the impact over time of the entry of the first Starbucks into census tracts that previously did not have coffee shops, using each of analytical samples defined in section 2. 95 percent confidence intervals are reported. Panel A compares census tracts that received a Starbucks café to those initially targeted by Starbucks for entry but ultimately rejected for reasons external to the company. Panel B compares tracts that received a Magic Johnson Starbucks café to those that did not but that were matched to resemble the distribution of treated tract closely through a matching procedure. Panel C compares tracts that received their first Starbucks café to all tracts that remained without a Starbucks café for our study period. Each figure reports the marginal effects employing the difference-in-differences methodology as per Callaway and Sant’Anna (2021).

Another example: free lawyers, more divorce?

Cunningham and Goodman-Bacon (2024)

- **Context:** Federally funded legal clinics providing advice and representation to (some) low-income communities.
- **Treatment:** Year in which county i received its first “LSP” grant (1965-1975)
- **Outcome:** county-by-year panel of divorce rates (~3,100 counties, 30 years from 1959-1988)
- **Target parameter:** average treatment effects of “LSP” on county-level divorce rates in treated counties.**

***Not targeting the effect of LSP availability on, say, eligible or “served” people.*

Building *ATT*(1966,1969)

```
8 use "$datatemp/cgb_lsp_analysis_data" if inrange(year,1960,1988) & sd, clear
9
10 *time variable is year: year
11 *unit variable is county (Federal Information Processing System; FIPS) code: fips
12 *treatment variable is year of first legal services program grant:
13 recode legal (.=0), gen(G)
14 *outcome variable is county-level divorce rate: drate
15 *continuous covariates from 1960 are: share urban, share nonwhite, share aged 0-4,
16 *median family income, poverty rate (<$3k), high-income share (>$10k), share with more than HS education
17 local x = "_60pcturban _60pctnonwhit _60pct04years _59medfaminc _pct59inclt3k _pct59incmt10k _60pctmt12schl"

26 *Let's focus on the G==1966 cohort and a never-treated comparison group and how
27 *control strategies will for estimating ATT(1966,1969) (as an example)
28 preserve
29 *all we need for this is two years, two groups, and differences in the outcome
30 keep if inlist(year,1965,1969) ← Periods g-1 and t
31 keep if inlist(G,0,1966) ← Groups G=1966 and G=∞
32 egen drate0 = total(drate*(year==1965)), by(fips) ← Calculate ΔY
33 gen ddrate = drate-drate0
34 keep if year==1969 ← Just need one period now
```

$\widehat{ATT}^{IPW}$ (1966, 1969)

```

37 *****
38 *****IPW*****
39 *****
40 *estimate propensity scores:
41 logit G `x'
42 predict pscore, p

```

```
. logit G `x'
```

```

Iteration 0:  log likelihood = -404.44396
Iteration 1:  log likelihood = -278.64994
Iteration 2:  log likelihood = -215.25416
Iteration 3:  log likelihood = -207.08312
Iteration 4:  log likelihood = -206.74557
Iteration 5:  log likelihood = -206.74477
Iteration 6:  log likelihood = -206.74477

```

```

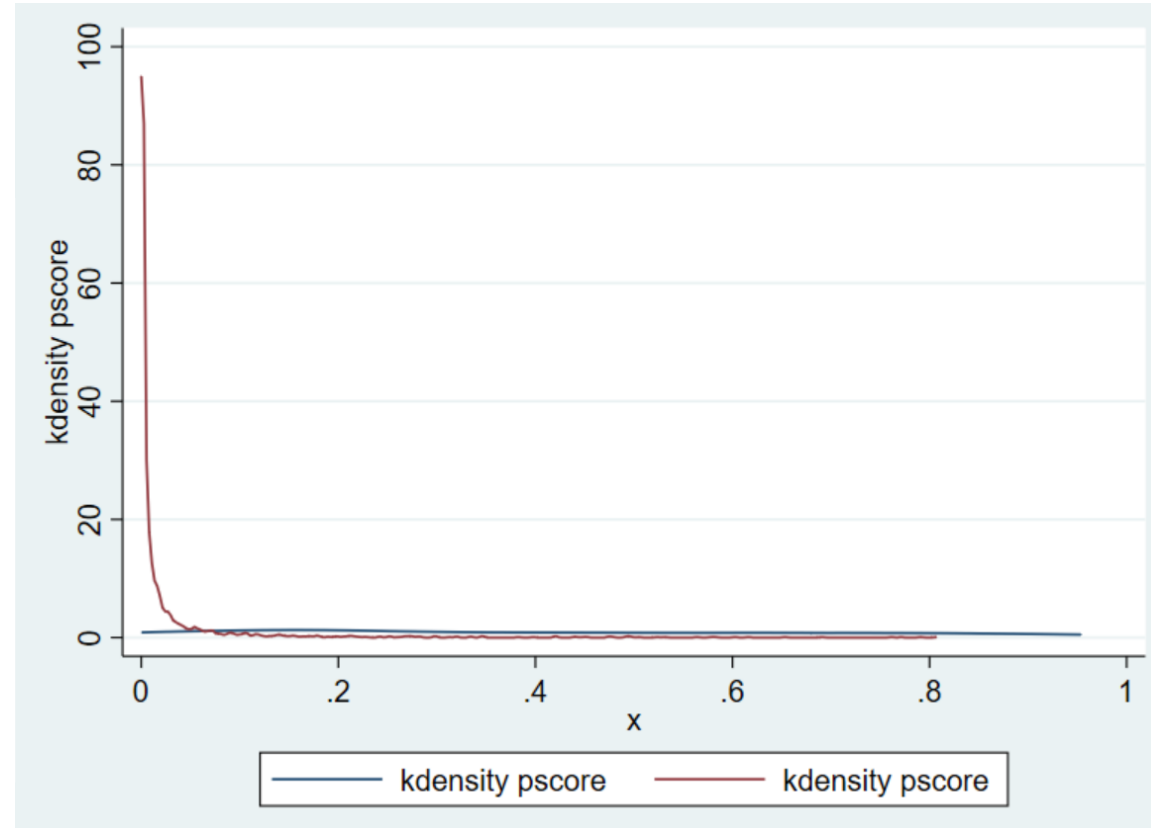
Logistic regression              Number of obs   =      2,434
                                LR chi2(7)         =      395.40
                                Prob > chi2         =      0.0000
Log likelihood = -206.74477      Pseudo R2       =      0.4888

```

G	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_60pcturban	.078914	.0093967	8.40	0.000	.0604968	.0973312
_60pctnonwhit	.0312022	.0172953	1.80	0.071	-.002696	.0651004
_60pct04years	-.2632151	.0979293	-2.69	0.007	-.455153	-.0712771
_59medfaminc	-.0005443	.0008692	-0.63	0.531	-.002248	.0011594
_pct59inclt3k	-.1099146	.0634671	-1.73	0.083	-.2343078	.0144786
_pct59incmt10k	.1399152	.0819349	1.71	0.088	-.0206743	.3005047
_60pctmt12schl	-.0850336	.0231876	-3.67	0.000	-.1304804	-.0395867
_cons	2.287712	5.372309	0.43	0.670	-8.24182	12.81724

$\widehat{ATT}^{IPW}$ (1966, 1969)

```
37 *****  
38 *****IPW*****  
39 *****  
40 *estimate propensity scores:  
41 logit G `x'  
42 predict pscore, p  
43 *do we have overlap? kind of...  
44 twoway kdensity pscore if G==1966 || kdensity pscore if G==0
```



$\widehat{ATT}^{IPW}$ (1966, 1969)

```
37 *****
38 *****IPW*****
39 *****
40 *estimate propensity scores:
41 logit G `x'
42 predict pscore, p
43 *do we have overlap? kind of...
44 twoway kdensity pscore if G==1966 || kdensity pscore if G==0
45 *to get a sense of these, just check the share above/below say pscore=0.1
46 gen lowp = pscore<.1
47 tab G lowp, row
48 *a very coarse p-score matching estimator would therefore use 75 treated counties
49 *with "high" pcores and 138 comparison counties with "high" pcores
```

```
. tab G lowp, row
```

Key			
frequency			
row percentage			
RECODE of legal (Year of first LSP grant)	lowp		Total
	0	1	
0	138 5.90	2,200 94.10	2,338 100.00
1966	75 78.13	21 21.88	96 100.00
Total	213 8.75	2,221 91.25	2,434 100.00

$\widehat{ATT}^{IPW}$ (1966, 1969)

```

37 *****
38 *****IPW*****
39 *****
40 *estimate propensity scores:
41 logit G `x'
42 predict pscore, p
43 *do we have overlap? kind of...
44 twoway kdensity pscore if G==1966 || kdensity pscore if G==0
45 *to get a sense of these, just check the share above/below say pscore=0.1
46 gen lowp = pscore<.1
47 tab G lowp, row
48 *a very coarse p-score matching estimator would therefore use 75 treated counties
49 *with "high" p-scores and 138 comparison counties with "high" p-scores
50
51 *calculate IPW weights: this convex function obliterates most of the very small p-scores
52 gen ipw = pscore/(1-pscore)*(G==0) + (G==1966)
53 egen tipw = total(ipw), by(G)
54 replace ipw = ipw/tipw
55 twoway kdensity ipw if G==1966 || kdensity ipw if G==0
56 *lets see how much weight is being given to the bulk of the comparison counties
57 sum ipw if G==0, det
58 gen lowipw = ipw<r(p75) & G==0
59 sum ipw if lowipw & G==0
60 di r(mean)*r(N)
61 *just 5% of the total comparison group weight is on the bottom 3/4 of comparison
62 *counties<--lots of imbalance to undo!
63

```

```

. *lets see how much weight is being given to the bulk of the comparison
. sum ipw if G==0, det

```

		ipw			
	Percentiles	Smallest			
1%	2.97e-07	1.31e-07			
5%	7.16e-07	1.38e-07			
10%	1.07e-06	1.42e-07	Obs		2,338
25%	2.87e-06	1.49e-07	Sum of Wgt.		2,338
50%	.0000255		Mean		.0004277
		Largest	Std. Dev.		.0022308
75%	.0001511	.0358977			
90%	.000647	.0378819	Variance		4.98e-06
95%	.0014769	.042536	Skewness		12.99134
99%	.0088927	.0471366	Kurtosis		216.2869

```

. gen lowipw = ipw<r(p75) & G==0
. sum ipw if lowipw & G==0

```

Variable	Obs	Mean	Std. Dev.	Min	Max
ipw	1,753	.0000269	.0000363	1.31e-07	.0001501

```

. di r(mean)*r(N)
.04714288

```

$\widehat{ATT}^{IPW}$ (1966, 1969)

```

37 *****
38 *****IPW*****
39 *****
40 *estimate propensity scores:
41 logit G `x'
42 predict pscore, p
43 *do we have overlap? kind of...
44 twoway kdensity pscore if G==1966 || kdensity pscore if G==0
45 *to get a sense of these, just check the share above/below say pscore=0.1
46 gen lowp = pscore<.1
47 tab G lowp, row
48 *a very coarse p-score matching estimator would therefore use 75 treated counties
49 *with "high" p-scores and 138 comparison counties with "high" p-scores
50
51 *calculate IPW weights: this convex function obliterates most of the very small p-scores
52 gen ipw = pscore/(1-pscore)*(G==0) + (G==1966)
53 egen tipw = total(ipw), by(G)
54 replace ipw = ipw/tipw
55 twoway kdensity ipw if G==1966 || kdensity ipw if G==0
56 *lets see how much weight is being given to the bulk of the comparison counties
57 sum ipw if G==0, det
58 gen lowipw = ipw<r(p75) & G==0
59 sum ipw if lowipw & G==0
60 di r(mean)*r(N)
61 *just 5% of the total comparison group weight is on the bottom 3/4 of comparison
62 *counties--lots of imbalance to undo!
63
64 *now we can obtain a (standardized) IPW estimate of ATT(1966,1969) with a
65 *never-treated comparison group
66 gen treat = G==1966
67 reg ddrate treat [aw=ipw], robust

```

```

. reg ddrate treat [aw=ipw], robust
(sum of wgt is 2.000000053121852)

```

Linear regression

Number of obs	=	2,434
F(1, 2432)	=	8.65
Prob > F	=	0.0033
R-squared	=	0.0289
Root MSE	=	2.9111

ddrate	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
treat	1.004219	.3414443	2.94	0.003	.3346675	1.673771
_cons	1.473025	.2418168	6.09	0.000	.9988364	1.947213

$\widehat{ATT}^{IPW}$ (1966, 1969)

```

37 *****
38 *****IPW*****
39 *****
40 *estimate propensity scores:
41 logit G `x'
42 predict pscore, p
43 *do we have overlap? kind of...
44 twoway kdensity pscore if G==1966 || kdensity pscore if G==0
45 *to get a sense of these, just check the share above/below say pscore=0.1
46 gen lowp = pscore<.1
47 tab G lowp, row
48 *a very coarse p-score matching estimator would therefore use 75 treated counties
49 *with "high" p-scores and 138 comparison counties with "high" p-scores
50
51 *calculate IPW weights: this convex function obliterates most of the very small p-scores
52 gen ipw = pscore/(1-pscore)*(G==0) + (G==1966)
53 egen tipw = total(ipw), by(G)
54 replace ipw = ipw/tipw
55 twoway kdensity ipw if G==1966 || kdensity ipw if G==0
56 *lets see how much weight is being given to the bulk of the comparison counties
57 sum ipw if G==0, det
58 gen lowipw = ipw<r(p75) & G==0
59 sum ipw if lowipw & G==0
60 di r(mean)*r(N)
61 *just 5% of the total comparison group weight is on the bottom 3/4 of comparison
62 *counties<--lots of imbalance to undo!
63
64 *now we can obtain a (standardized) IPW estimate of ATT(1966,1969) with a
65 *never-treated comparison group
66 gen treat = G==1966
67 reg ddrate treat [aw=ipw], robust
68
69 *verify this is how teffects does it
70 teffects ipw (ddrate) (treat `x'), atet

```

```

. *verify this is how teffects does it
. teffects ipw (ddrate) (treat `x'), atet

```

```

Iteration 0:  EE criterion = 4.542e-23
Iteration 1:  EE criterion = 5.986e-31

```

```

Treatment-effects estimation      Number of obs      =      2,434
Estimator      : inverse-probability weights
Outcome model  : weighted mean
Treatment model: logit

```

	ddrate	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
ATET	treat (1 vs 0)	1.004219	.3318128	3.03	0.002	.3538781	1.65456
POmean	treat 0	1.473025	.235857	6.25	0.000	1.010753	1.935296

$\widehat{ATT}^{RA}(1966, 1969)$

```

73 *****
74 *****RA/OR*****
75 *****
76 *estimate a linear outcome model for the change in divorce rates in untreate
77 reg ddrate `x' if G==0
78 predict ra, xb

```

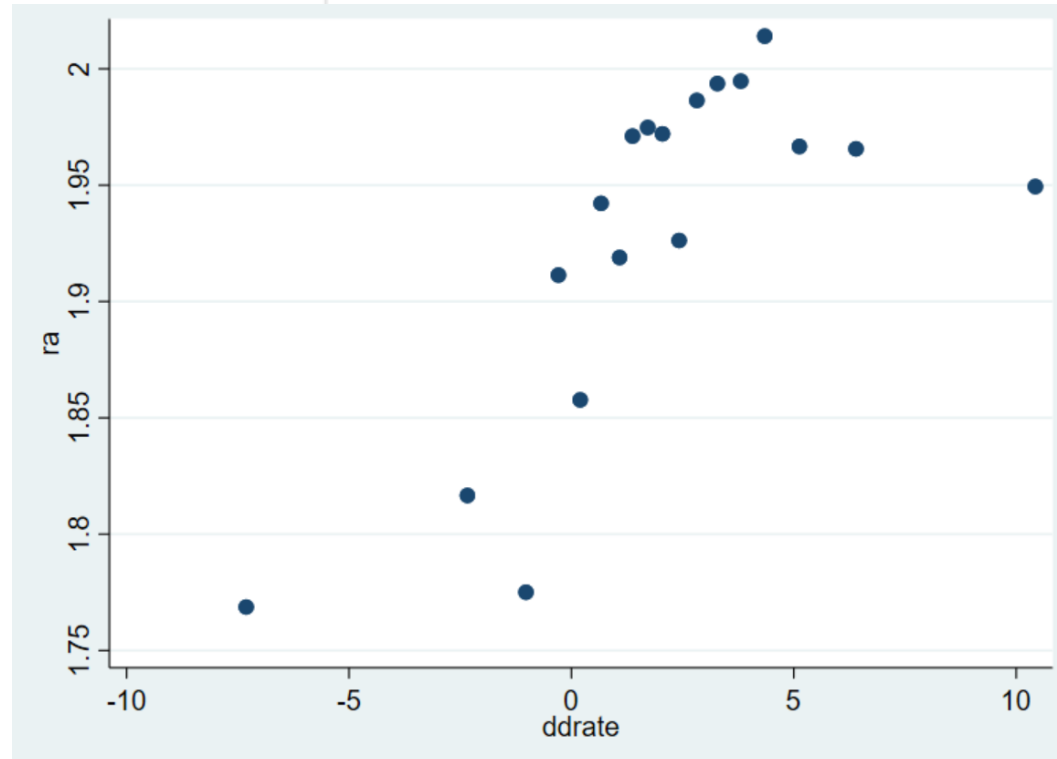
```
. reg ddrate `x' if G==0
```

Source	SS	df	MS	Number of obs	=	2,338
Model	441.839126	7	63.1198752	F(7, 2330)	=	3.70
Residual	39770.5326	2,330	17.0688981	Prob > F	=	0.0006
				R-squared	=	0.0110
				Adj R-squared	=	0.0080
Total	40212.3718	2,337	17.2068343	Root MSE	=	4.1315

ddrate	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_60pcturban	.003115	.0039925	0.78	0.435	-.0047141	.0109441
_60pctnonwhit	.0069781	.0071768	0.97	0.331	-.0070954	.0210516
_60pct04years	-.0920228	.0566324	-1.62	0.104	-.2030779	.0190324
_59medfaminc	.0013614	.0004856	2.80	0.005	.0004092	.0023136
_pct59incl3k	.0509874	.0333883	1.53	0.127	-.0144864	.1164612
_pct59incmt10k	-.1976191	.0456728	-4.33	0.000	-.2871827	-.1080555
_60pctmt12schl	.0024528	.0136704	0.18	0.858	-.0243546	.0292602
_cons	-3.25865	3.106376	-1.05	0.294	-9.350199	2.832899

$\widehat{ATT}^{RA}(1966, 1969)$

```
73 *****  
74 *****RA/OR*****  
75 *****  
76 *estimate a linear outcome model for the change in divorce rates in untreated counties  
77 reg ddrate `x' if G==0  
78 predict ra, xb  
79  
80 *how well does this linear model appear to fit comparison county divorce trends?  
81 binsreg ra ddrate if G==0
```



$\widehat{ATT}^{RA}(1966, 1969)$

```
73 *****
74 *****RA/OR*****
75 *****
76 *estimate a linear outcome model for the change in divorce rates in untreated counties
77 reg ddrate `x' if G==0
78 predict ra, xb
79
80 *how well does this linear model appear to fit comparison county divorce trends?
81 binsreg ra ddrate if G==0
82
83 gen diff = ddrate-ra if G==1966
84 sum diff if G==1966
85
```

```
. sum diff if G==1966
```

Variable	Obs	Mean	Std. Dev.	Min	Max
diff	96	.5213176	2.39558	-12.25087	7.709376

$\widehat{ATT}^{RA}$ (1966, 1969)

```

73 *****
74 *****RA/OR*****
75 *****
76 *estimate a linear outcome model for the change in divorce rates in untreated counties
77 reg ddrate `x' if G==0
78 predict ra, xb
79
80 *how well does this linear model appear to fit comparison county divorce trends?
81 binsreg ra ddrate if G==0
82
83 gen diff = ddrate-ra if G==1966
84 sum diff if G==1966
85
86 *verify this is how teffects does it
87 teffects ra (ddrate `x') (treat), atet
88 restore
89

```

```

Iteration 0:  EE criterion = 9.431e-27
Iteration 1:  EE criterion = 1.472e-31

```

```

Treatment-effects estimation      Number of obs      =      2,434
Estimator      : regression adjustment
Outcome model   : linear
Treatment model : none

```

ddrate		Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
ATET	treat						
	(1 vs 0)	.5213176	.3359418	1.55	0.121	-.1371161	1.179751
POmean	treat						
	0	1.955926	.2346329	8.34	0.000	1.496054	2.415798

CS estimates $\widehat{ATT}(g, t)$

```

159 *So how do we USE the staggering? We already explored conditional estimators
160 *using a never-treated comparison group, but could also consider not-yet-treated.
161 *Here we do it both ways for the RA/OR estimator:
162 csdid drate `x', ivar(fips) time(year) gvar(G) method(reg) long2 never
163 estimates save $datatemp/cs_drate_never_ra, replace
164 csdid drate `x', ivar(fips) time(year) gvar(G) method(reg) long2 notyet
165

```

$\widehat{ATT}^{RA}(1966, 1969)$

11 treatment groups, 29 periods (30 years minus one normalization) = **319 $\widehat{ATT}(g, t)$ s**

$\widehat{ATT}(1967, t), \widehat{ATT}(1968, t), \dots, \widehat{ATT}(1975, t)$

$\widehat{ATT}(1965, t)$

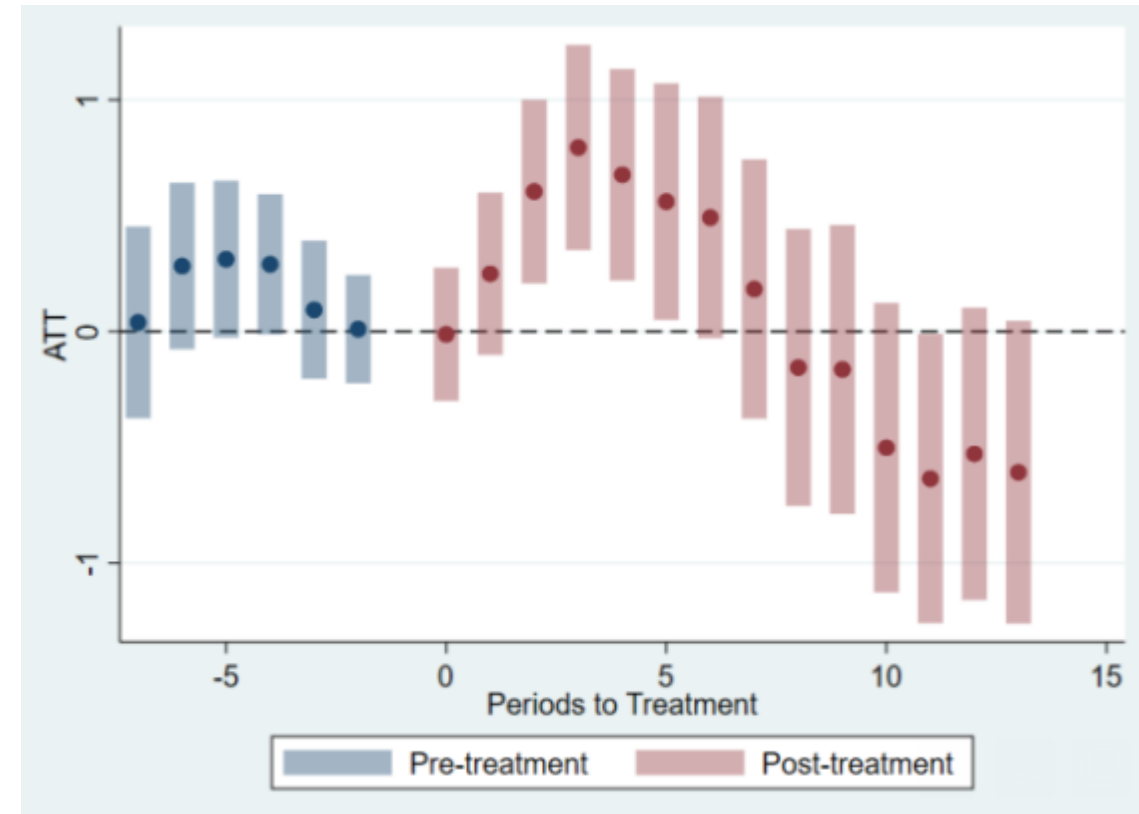
$\widehat{ATT}(1966, t)$

csdid drate `x', ivar(fips) time(year) gvar(G) method(reg) long2 never										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										
.....										

Aggregating CS estimates

```
170 *how do you then actually produce a useful figure from that ATT(g,t) output
171 *lets choose the RA/OR estimator with the never treated comparison group
172 estimates use $datatemp/cs_drate_never_ra
173 estat event, window(-6,13)
174 csdid_plot
```

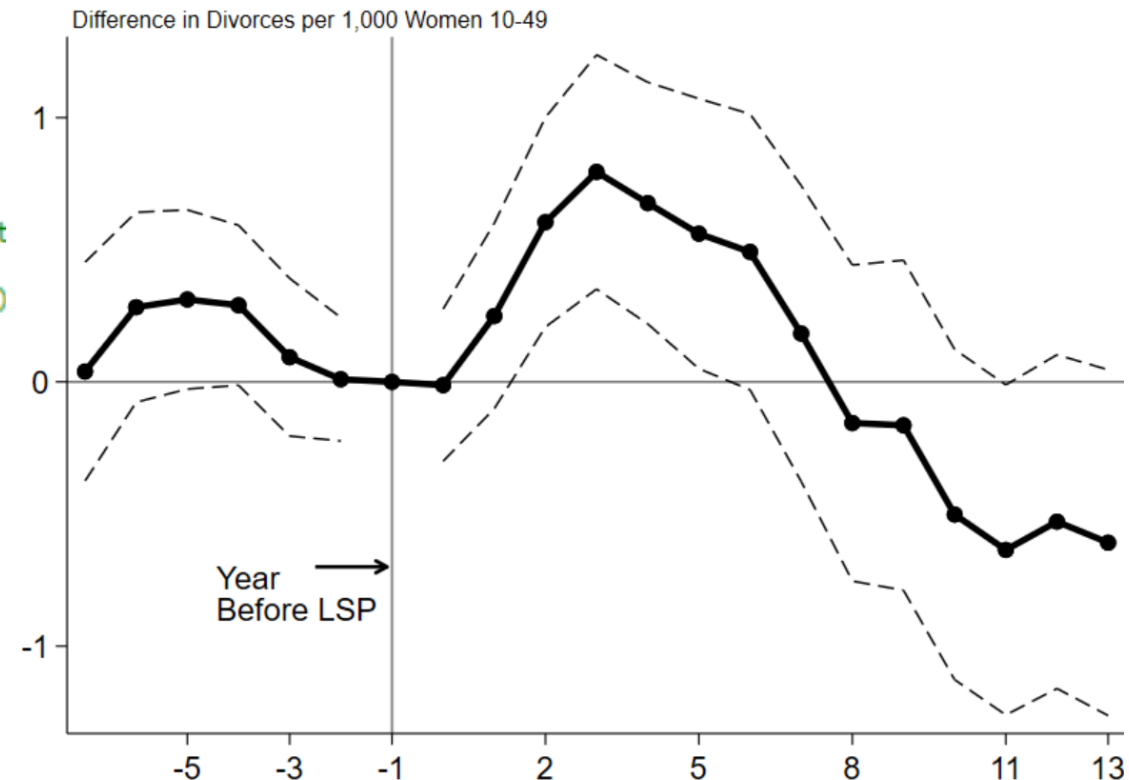
$\widehat{ATT}(e)$ weights $\widehat{ATT}(g, g + e)$
across cohorts according to
number of treated units



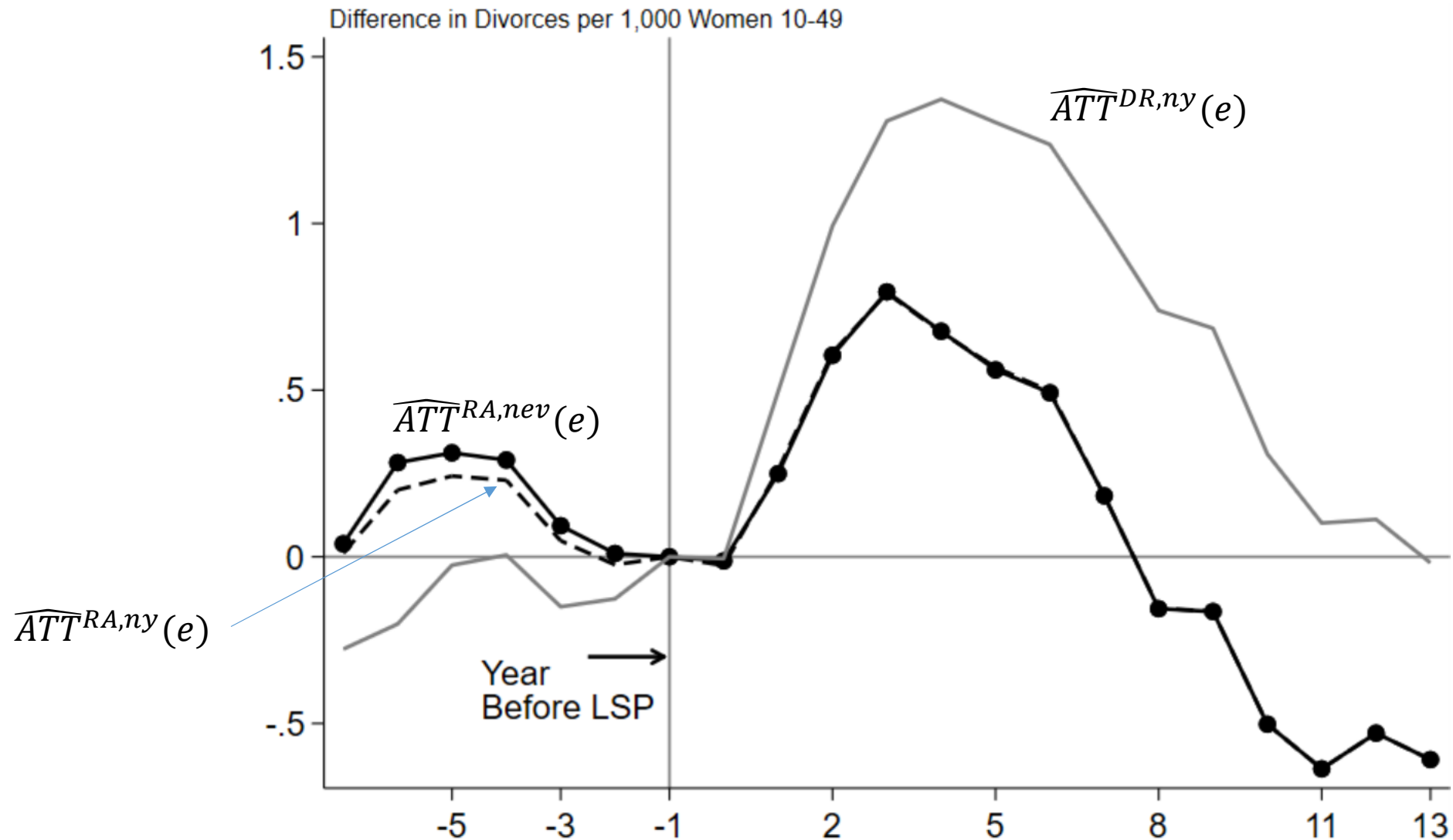
Aggregating CS estimates

```
170 *how do you then actually produce a useful figure from that ATT(g,t) output
171 *lets choose the RA/OR estimator with the never treated comparison group
172 estimates use $datatemp/cs_drate_never_ra
173 estat event, window(-6,13)
174 csdid_plot
```

```
177 estimates use $datatemp/cs_drate_never_ra
178 estat event, window(-6,13) post
179 *keep output in a Stata file to manipulate and prepare for graphing
180 regsave
181 *drop the first two obs that CS creates which are the aggregated pre and post estimat
182 keep if substr(var,1,1)=="T"
183 *extract the event-times and set pre-period times to negative (I call event-time exp)
184 destring var, ignore("T" "m" "p") gen(exp)
185 gen pre = substr(var,2,1)=="m"
186 replace exp = -exp if pre
187 *add a row and a zero for the omitted period.
188 count
189 local obs = r(N) + 1
190 set obs `obs'
191 replace exp = -1 in `obs'
192 for var coef*: replace X = 0 in `obs'
193 *construct CIs
194 gen ci_lower = coef-1.96*stderr
195 gen ci_upper = coef+1.96*stderr
196 *organize dataset for plotting
197 keep if coef<.
198 keep coef stderr ci_* exp
199 order exp coef stderr ci_*
200 sort exp
```



Comparing CS estimates



Other paths through the blueprint:

What do you want to know/what can you know? (Parameter of Interest)

What are you willing to believe? (*Parallel Trends for all $t < g$*)

What should you actually do? (Estimand/Estimator)

Were your beliefs plausible? (Internal Validity)

Borusyak, Jaravel, Spiess/Gardner/Wooldridge:

PT relative to ALL pre-treatment periods is stronger so it buys you more

Other paths through the blueprint:

What do you want to know/what can you know? (Parameter of Interest)

What are you willing to believe? (*Parallel Trends for all $t < g$*)

What should you actually do? (regression/imputation)

Were your beliefs plausible? (Internal Validity)

Borusyak, Jaravel, Spiess/Gardner/Wooldridge (did_imputation):

Compare to mean pre-treatment outcome, not just g . *Can be* more efficient (Harmon 2022).

But if you want efficiency, just get it with GMM ([Chen, Sant'Anna, Xie 2025](#))

(Don't) “Try all the new estimators”

Social Media and Mental Health[†]

By LUCA BRAGHERI, RO'EE LEVY, AND ALEXEY MAKARIN*

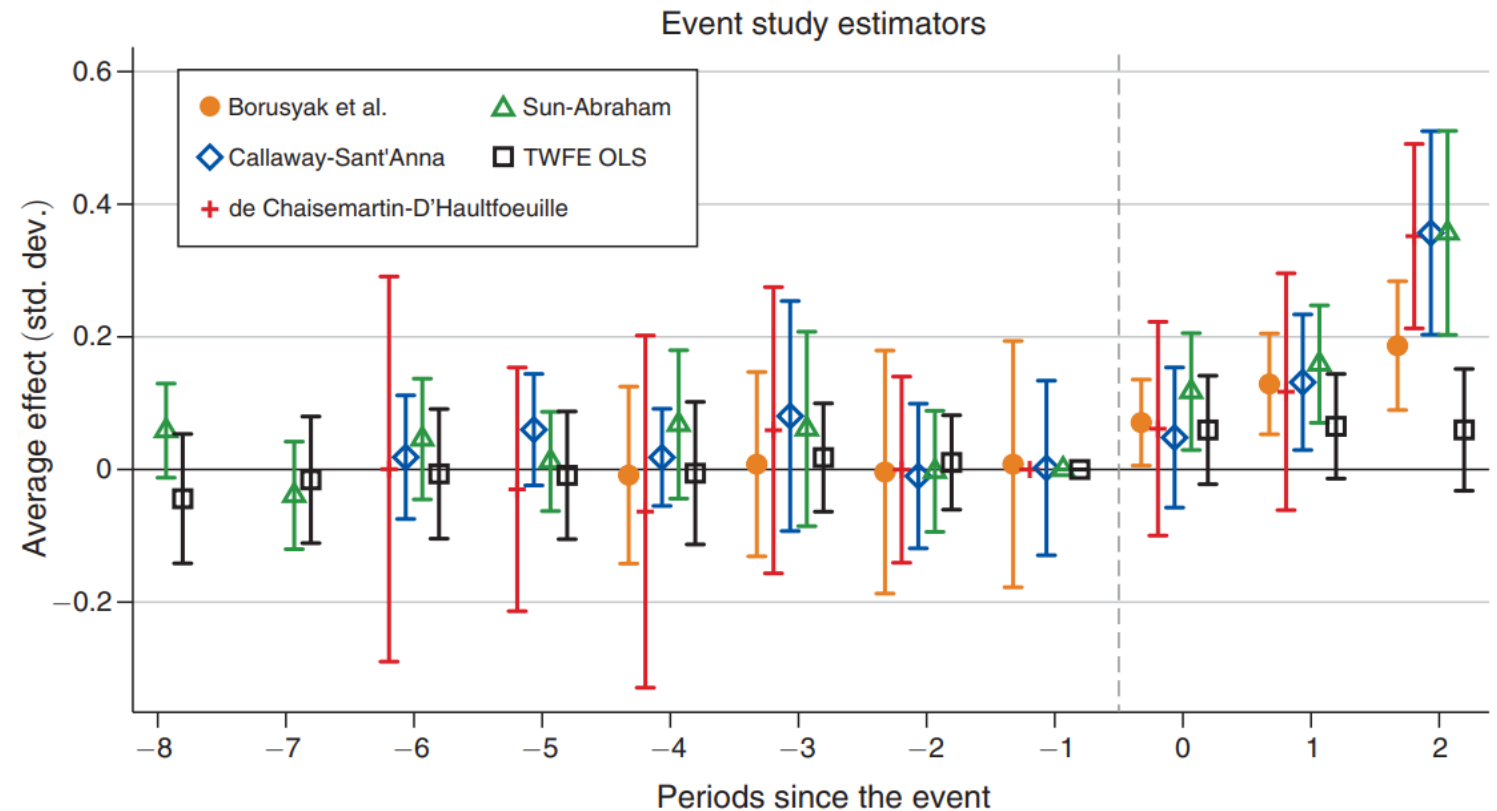


FIGURE 2. EFFECTS OF FACEBOOK ON THE INDEX OF POOR MENTAL HEALTH BASED ON DISTANCE TO/FROM FACEBOOK INTRODUCTION

Return to the blueprint:

What do you want to know/what can you know?

What are you willing to believe?

What should you actually do?

Were your beliefs plausible?

What did you learn?



When you go this direction, many questions start clear and remain clear, *even when the design gets more complex*:

What is your goal?
What do you assume?
What are you doing?
What evidence is there that your approach is good?

Return to the blueprint:

What do you want to know/what can you know?

What are you willing to believe?

What should you actually do?

Were your beliefs plausible?

What did you learn?

But when you start with
a familiar *estimator*:

You sometimes don't
get what you want

Sometimes you don't
get anything causal

Unless you have extra
assumptions

Another example: cheaper hip replacements?

Original Investigation

Association Between a Bundled Payment Program for Lower Extremity Joint Replacement and Patient Outcomes Among Medicare Advantage Beneficiaries

Amanda Sutherland, PhD; Emily Boudreau, PhD; Andy Bowe, MPH; Qian Huang, MPH; Joshua M. Liao, MD, MSc; Matthew Flagg, MBA; Deborah Cousins, MSPH; Dana Drzayich Antol, MS; William H. Shrank, MD, MSHS; Brian W. Powers, MD, MBA; Amol S. Navathe, MD, PhD

DESIGN, SETTING, AND PARTICIPANTS Cross-sectional study of 23 034 LEJR surgical episodes that emulated a stepped-wedge design by using the time-varying, geographically staggered rollout of the bundled payment program from January 1, 2012, to September 30, 2019. Episode-level multivariable regression models were estimated within practice to compare changes before and after program participation, using episodes at physician practices that had not yet begun participating in the program during a given time period (but would go on to do so) as the control. Data analyses were performed from July 1, 2021, to June 30, 2022.

What does “ATT” mean when there are multiple treated groups?

Group-Time ATTs: $ATT(g, t)$

Figure 2. Unadjusted Spending by Physician Practice Year of Entry (2012-2019)

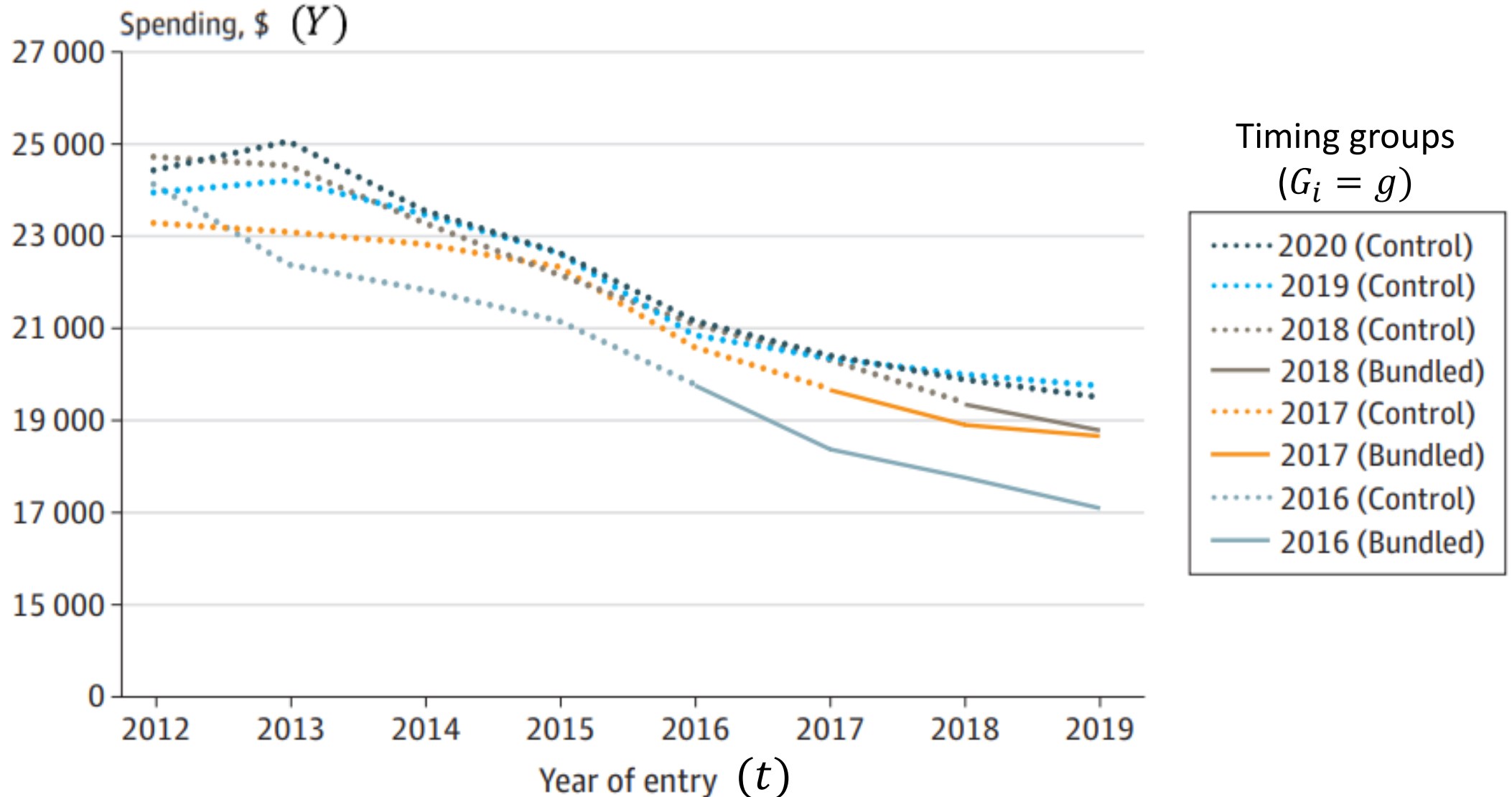


Figure 2: log Mean outcomes by g and t

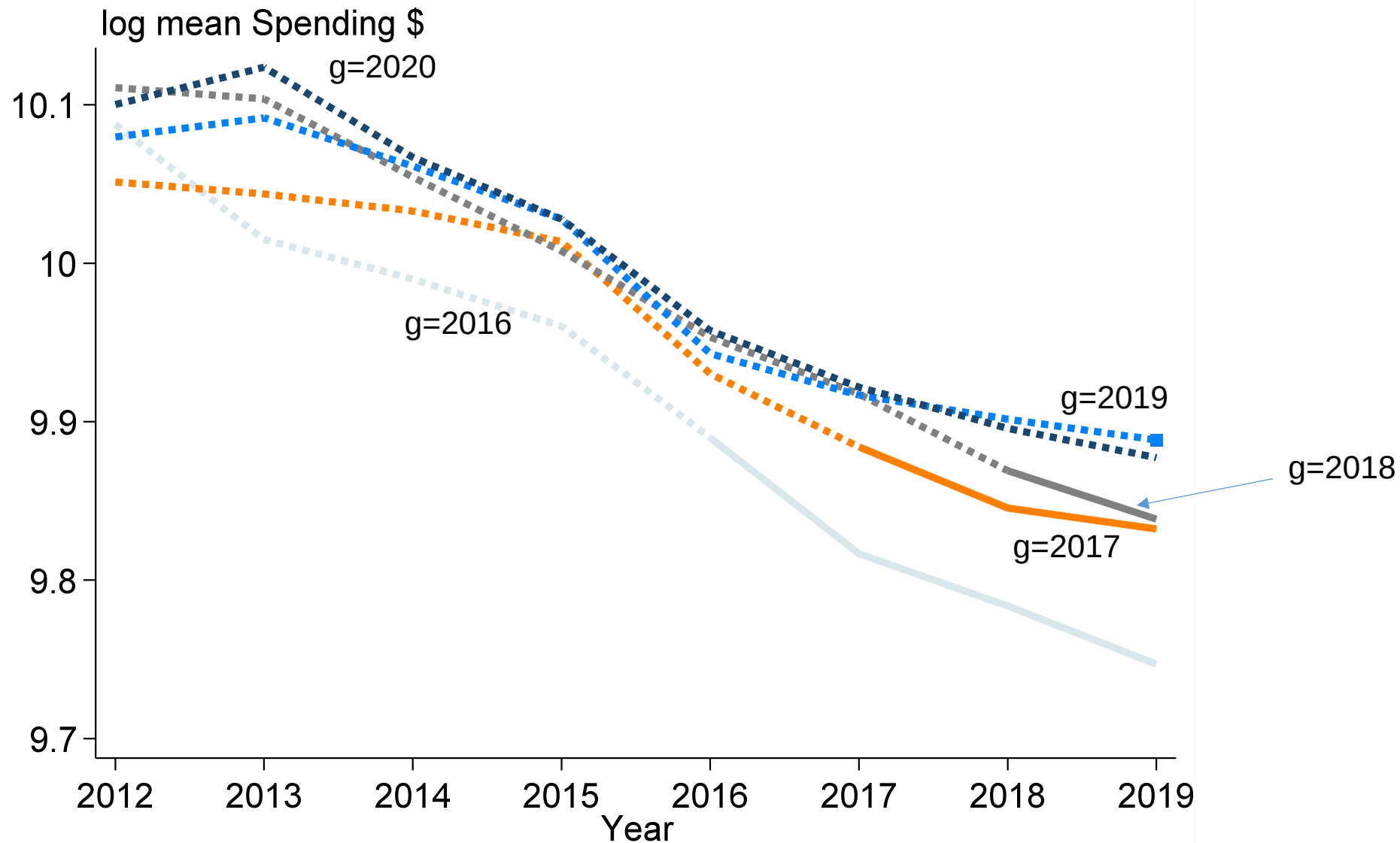


Figure 2: log Mean outcomes by g and t

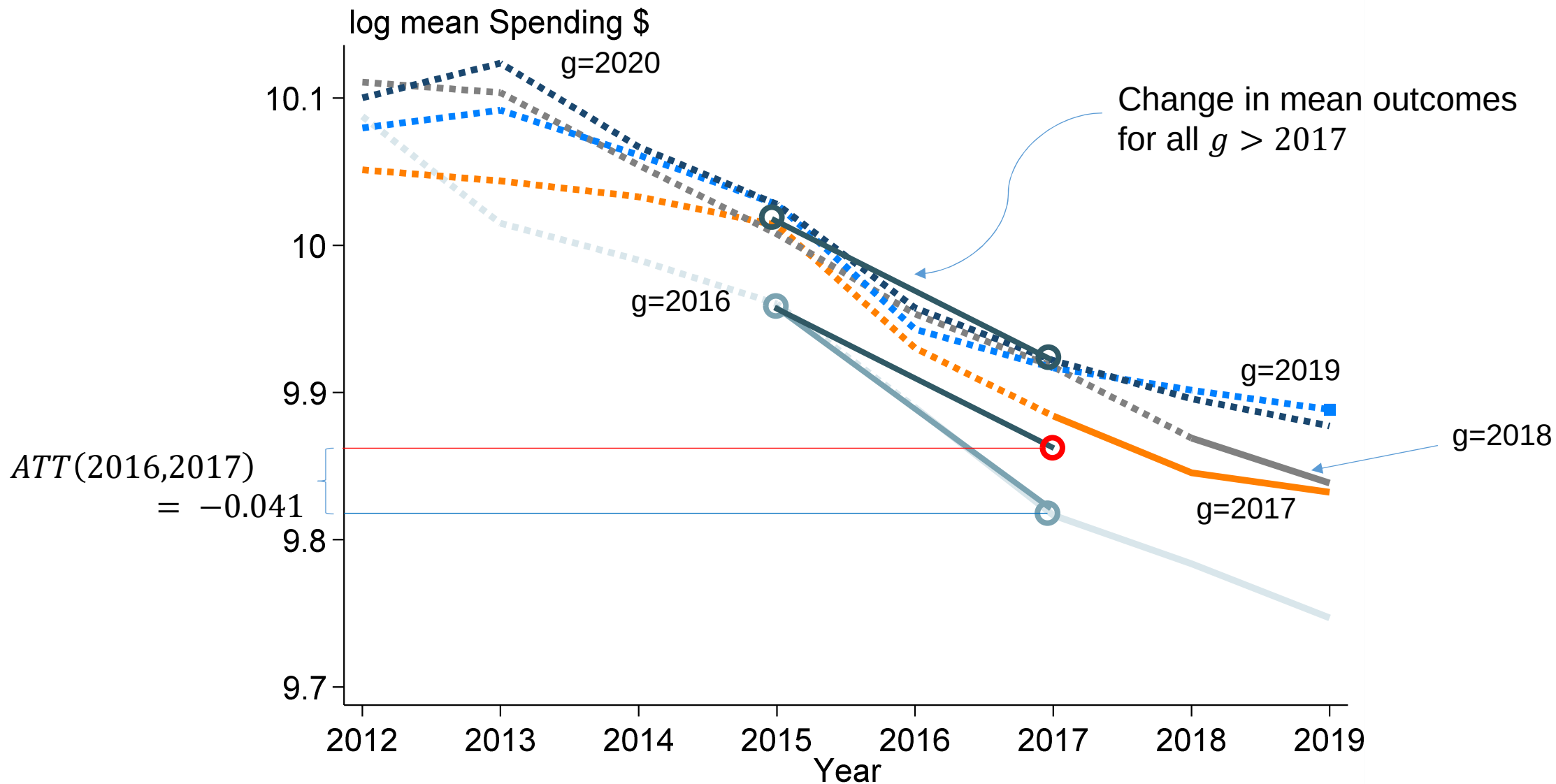


Figure 2: log Mean outcomes by g and t

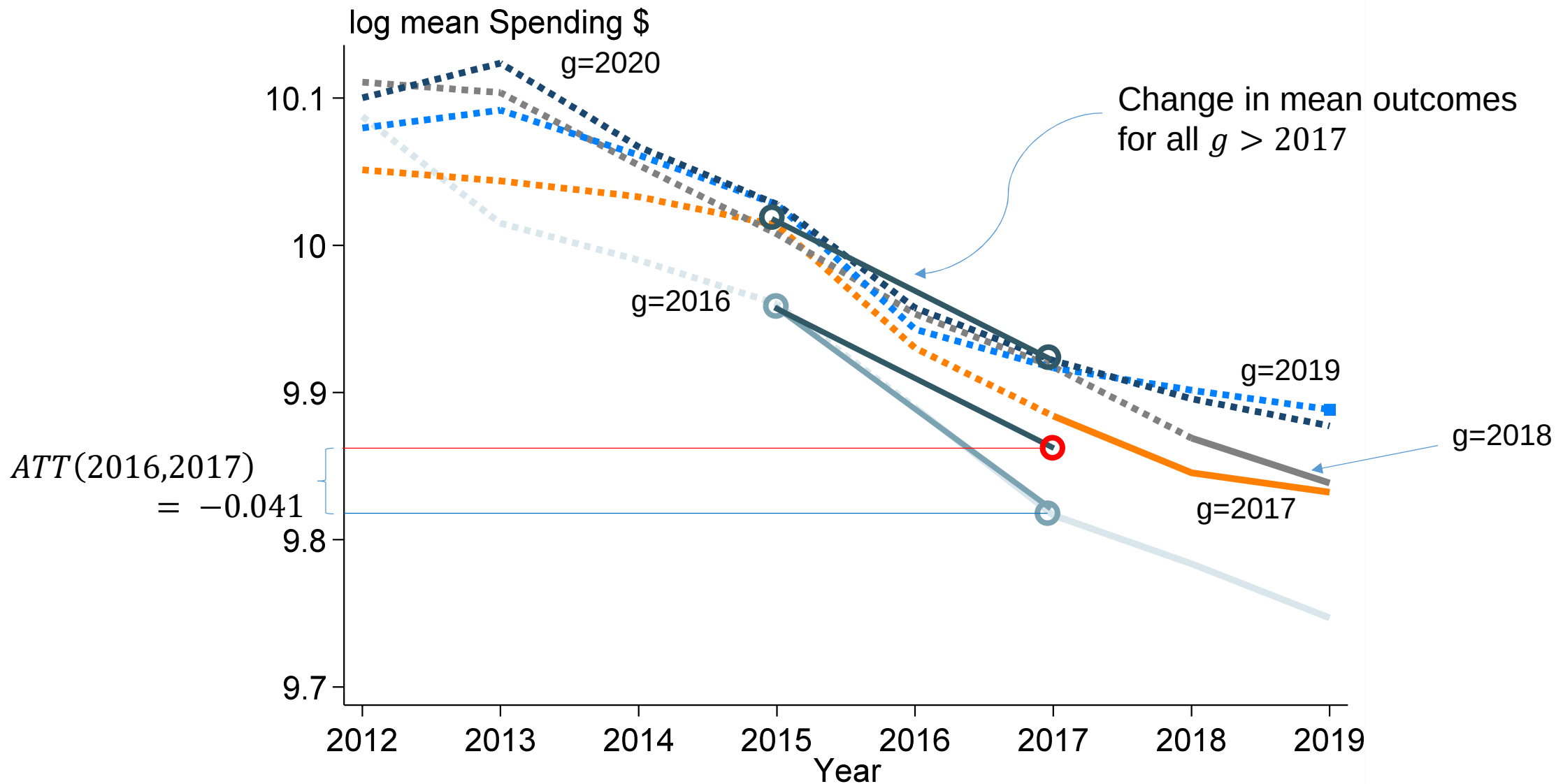
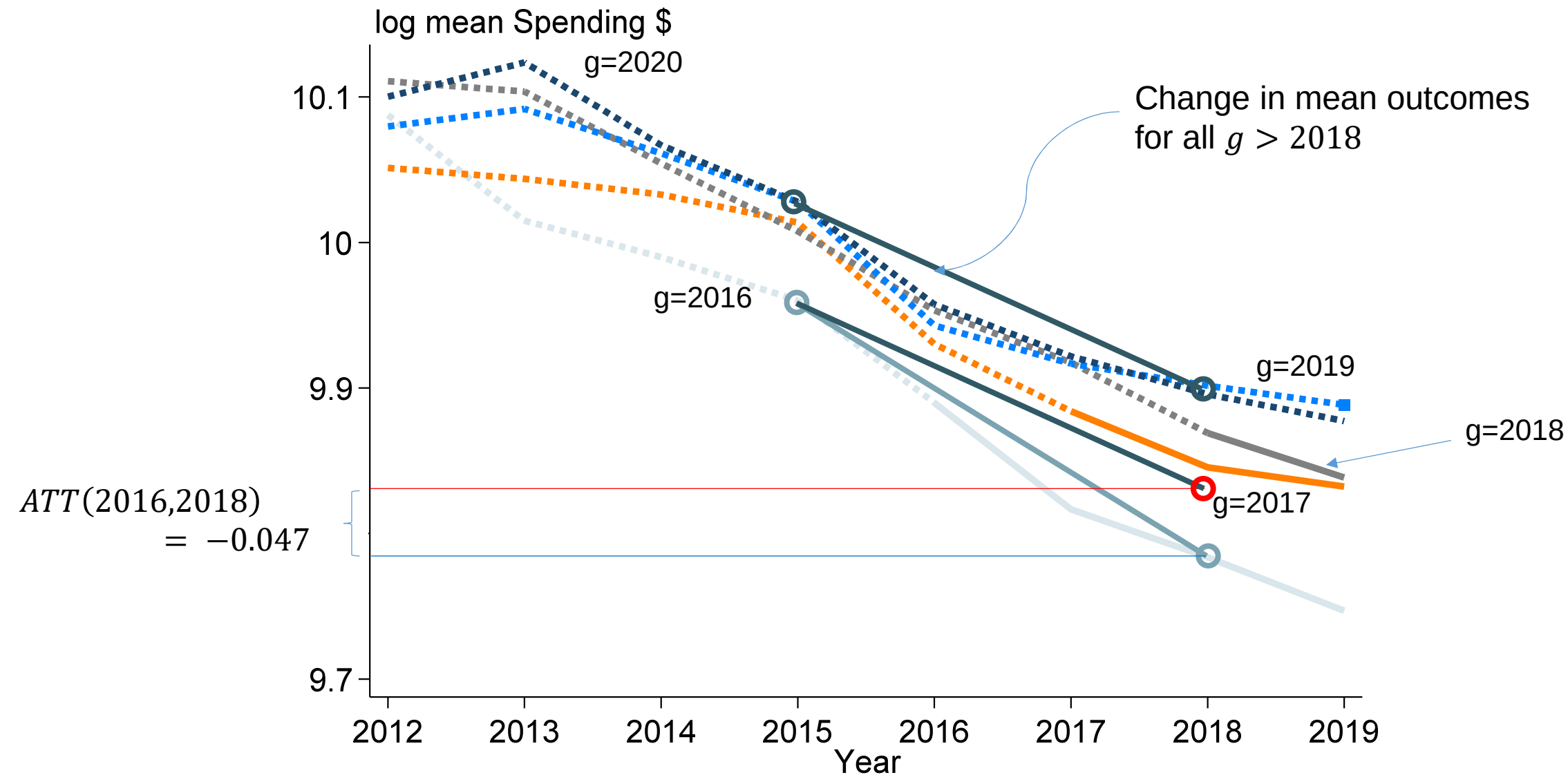
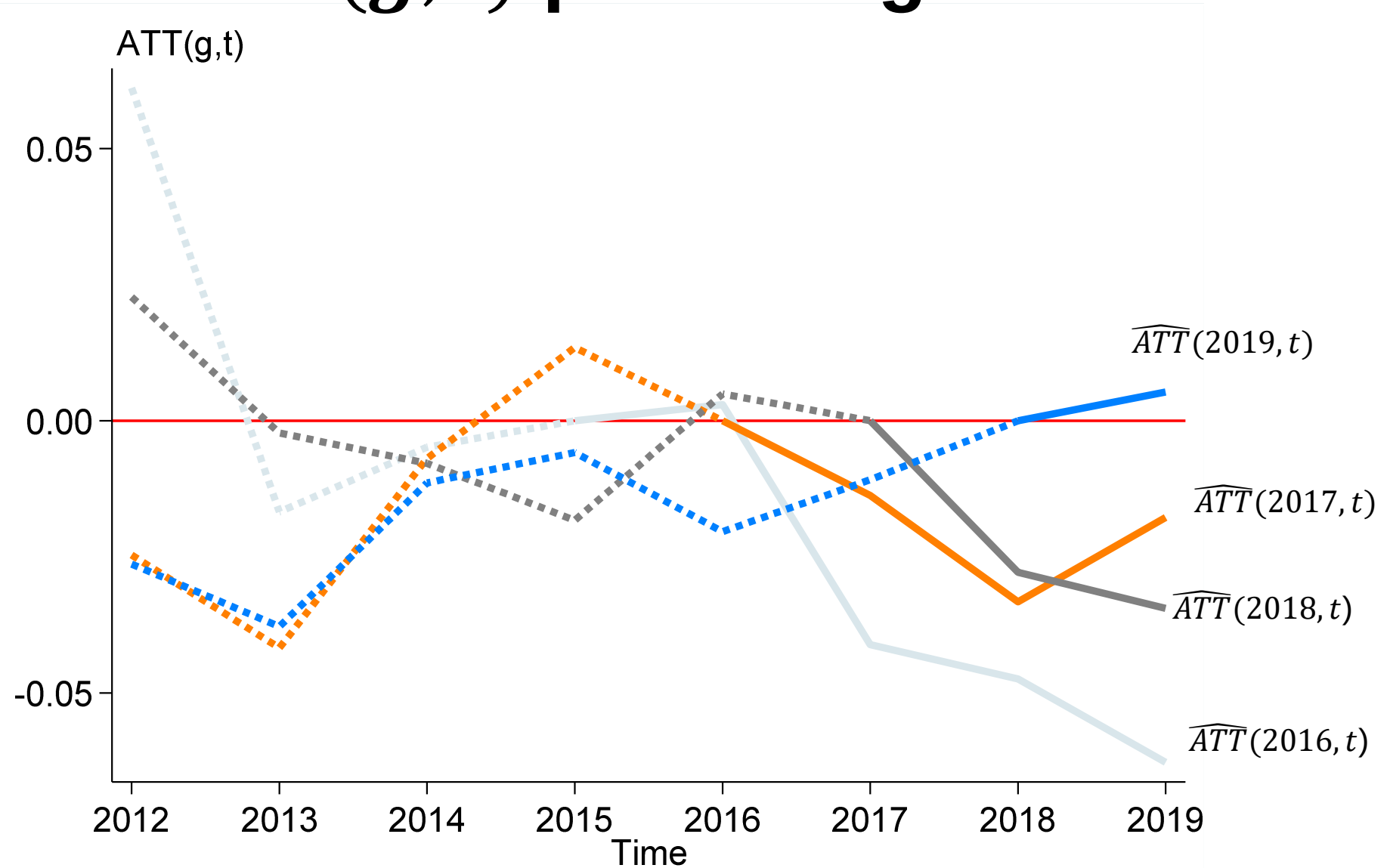


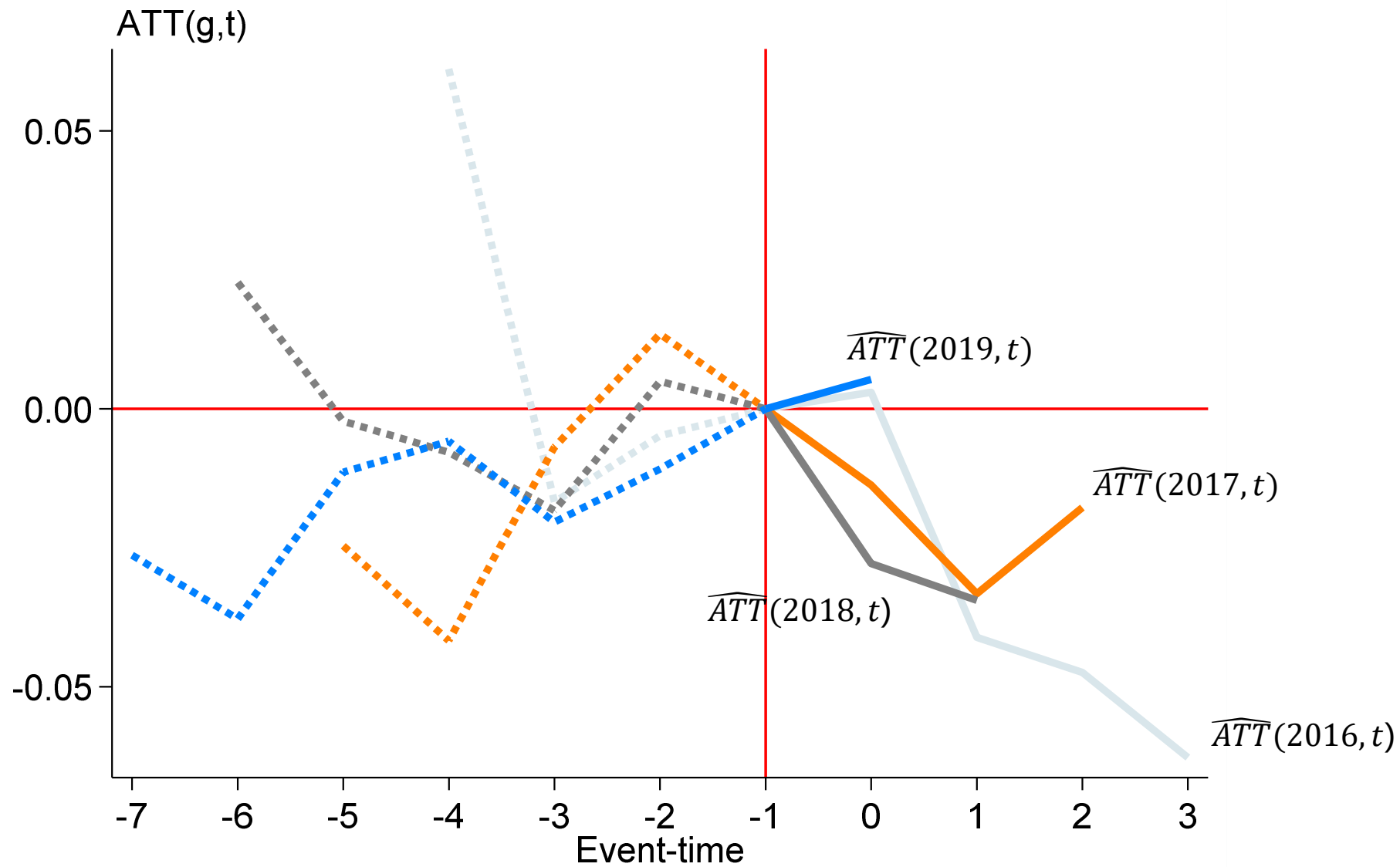
Figure 2: log Mean outcomes by g and t



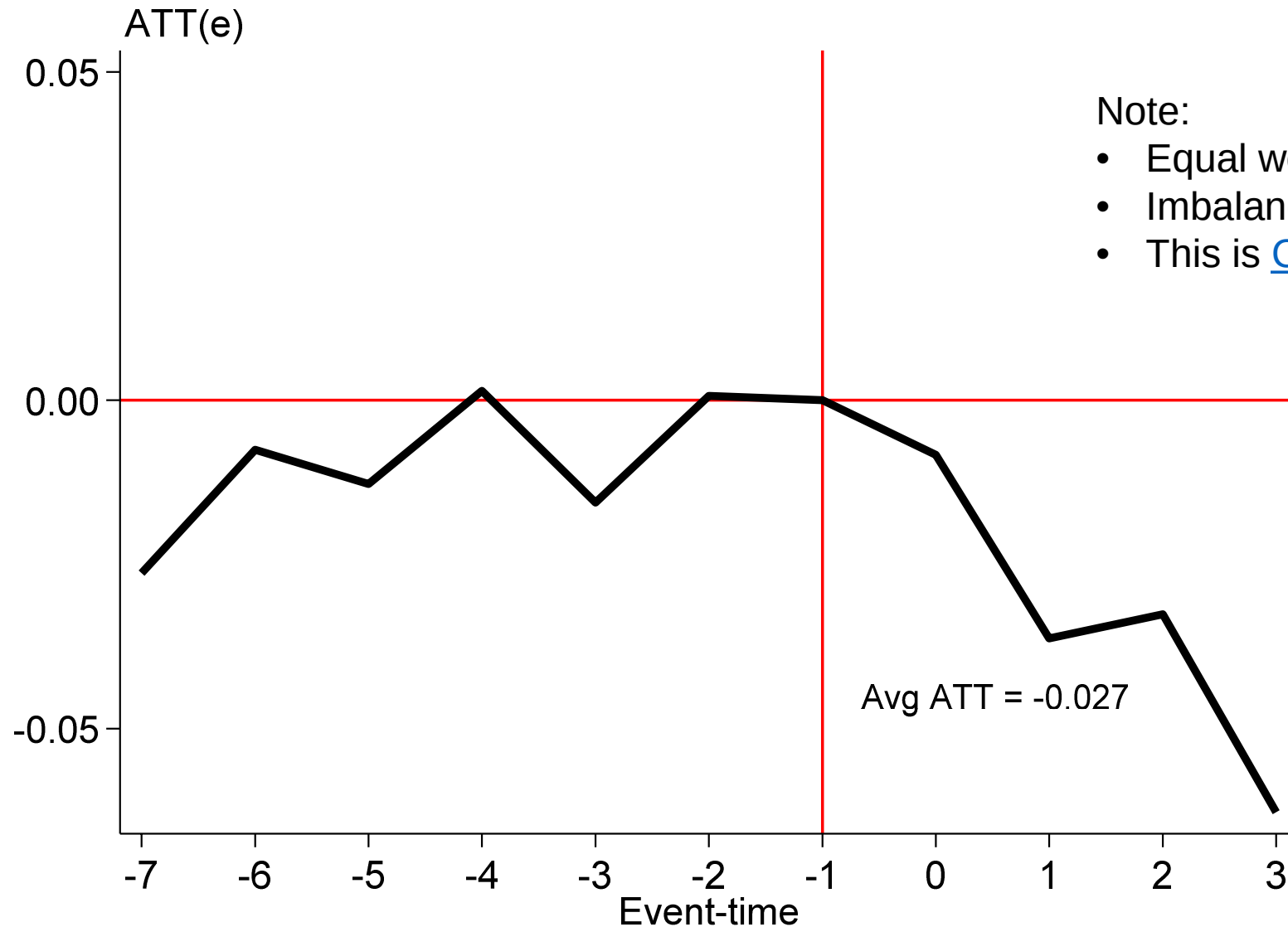
$\widehat{ATT}(g, t)$ plotted against t



$\widehat{ATT}(g, t)$ plotted against $e = t - g$



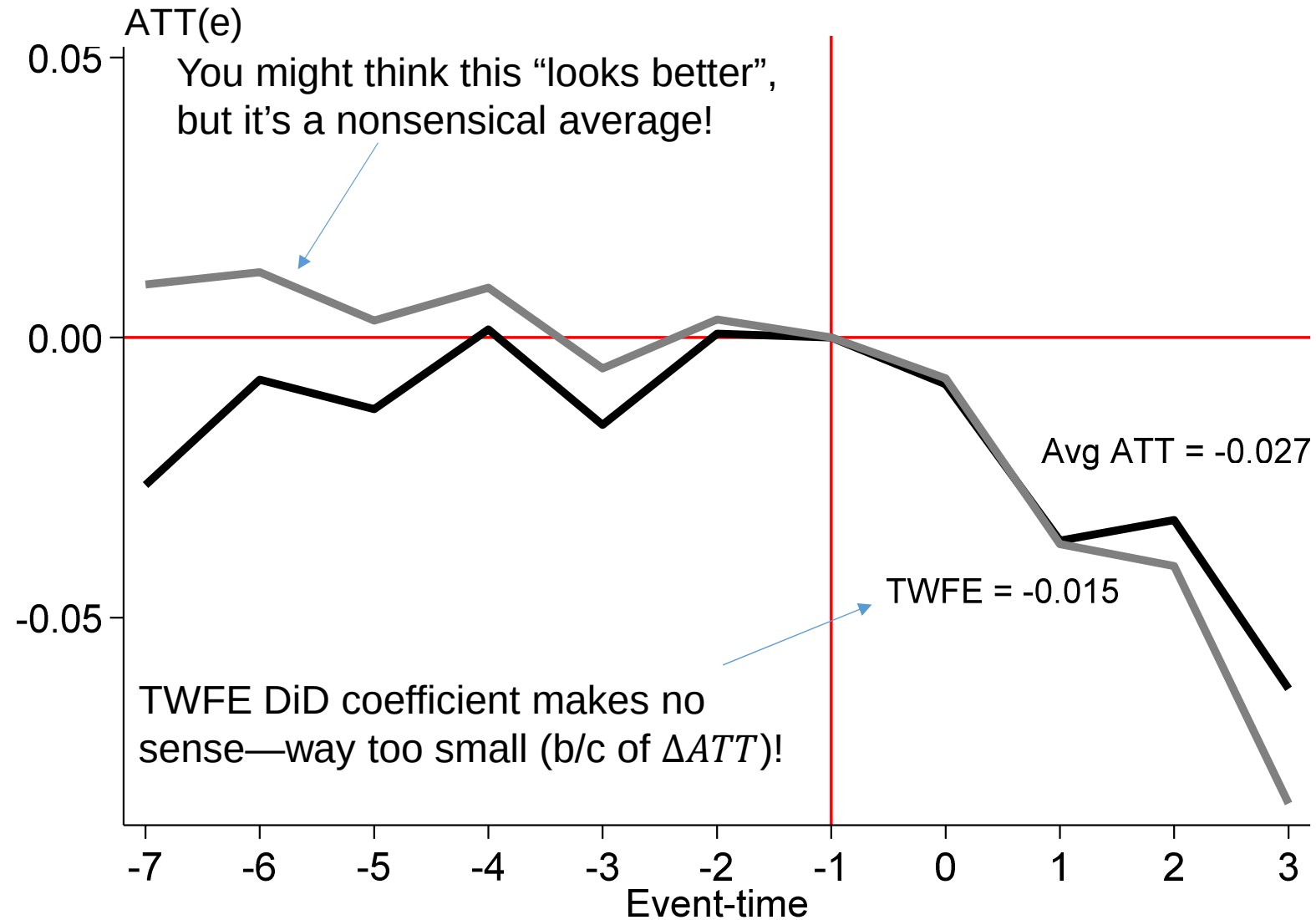
$\widehat{ATT}(e)$ plotted against e

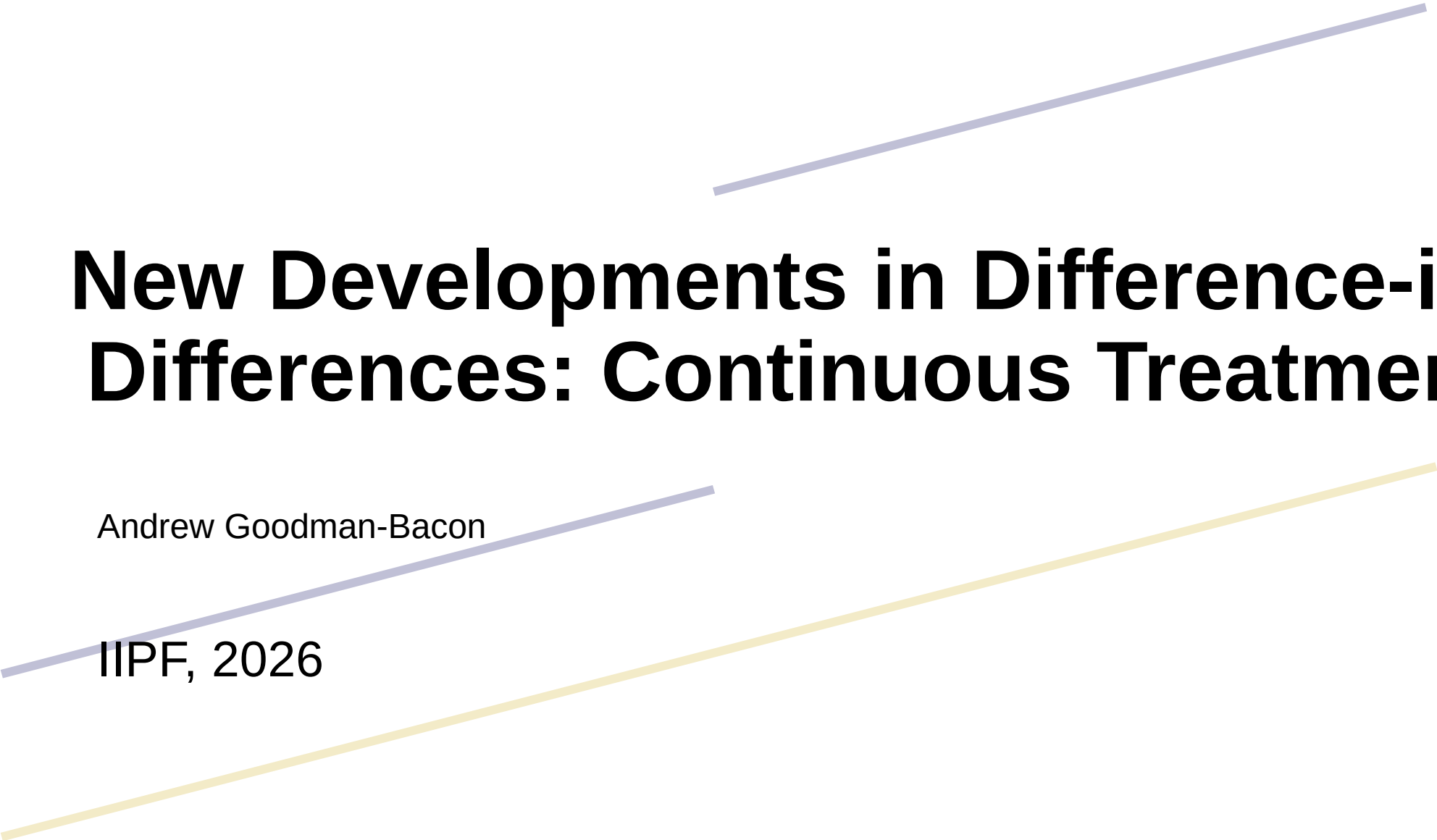


Note:

- Equal weights across groups
- Imbalanced event-times!
- This is [Callaway and Sant'Anna!](#)

Comparing to TWFE





New Developments in Difference-in-Differences: Continuous Treatment

Andrew Goodman-Bacon

IIPF, 2026

What about continuous treatments?

Variation in treatment intensity/dose/level is:

1. Common
2. Often the only variation
3. Or maybe the “best” variation

3/5 of my DD papers!

Think federal policies!

*Think “near vs. far”
instead of “shock vs.
no shock”*

What about continuous treatments?

Variation in treatment intensity/dose/level is:

1. Common
2. Often the only variation
3. Or maybe the “best” variation

“Dose-response effects” are:

1. Evidence for internal validity

Minimum Wages and Employment:
A Case Study of the Fast-Food Industry
in New Jersey and Pennsylvania

By DAVID CARD AND ALAN B. KRUEGER*

$$(1b) \quad \Delta E_i = a' + \mathbf{b}'\mathbf{X}_i + c'GAP_i + \varepsilon'_i$$

where ΔE_i is the change in employment from wave 1 to wave 2 at store i , $\mathbf{X}_i$ is a set of characteristics of store i , and NJ_i is a dummy variable that equals 1 for stores in New Jersey. GAP_i is an alternative measure of the impact of the minimum wage at store i based on the initial wage at that store

What about continuous treatments?

Variation in treatment intensity/dose/level is:

1. Common
2. Often the only variation
3. Or maybe the “best” variation

Minimum Wages and Employment:
A Case Study of the Fast-Food Industry
in New Jersey and Pennsylvania

By DAVID CARD AND ALAN B. KRUEGER*

“Dose-response effects” are:

1. Evidence for internal validity
2. Of interest per se

Stock Market Wealth and the Real Economy:
A Local Labor Market Approach[†]

By GABRIEL CHODOROW-REICH, PLAMEN T. NENOV, AND ALP SIMSEK*

Interpreted through this model, our empirical estimates map into a household-level annual MPC of 3.2 cents per dollar of stock wealth

What do you want to know?

Building Blocks: Levels

Canonical “causal effect” of treatment (D) on unit i in period t :

$$Y_{it}(1) - Y_{it}(0)$$

Average treatment effect on the treated in period t :

$$E[Y_{it}(1) - Y_{it}(0) | D_i = 1]$$

Building Blocks: Levels

Canonical “causal effect” of treatment (D) on unit i in period t :

$$Y_{it}(\textcolor{red}{1}) - Y_{it}(0)$$

Average treatment effect on the treated in period t :

$$E[Y_{it}(\textcolor{red}{1}) - Y_{it}(0) | D_i = \textcolor{red}{1}]$$

But this one “treatment” could be anything, any dose, any intensity.

Building Blocks: Levels

Canonical “causal effect” of treatment (D) on unit i in period t :

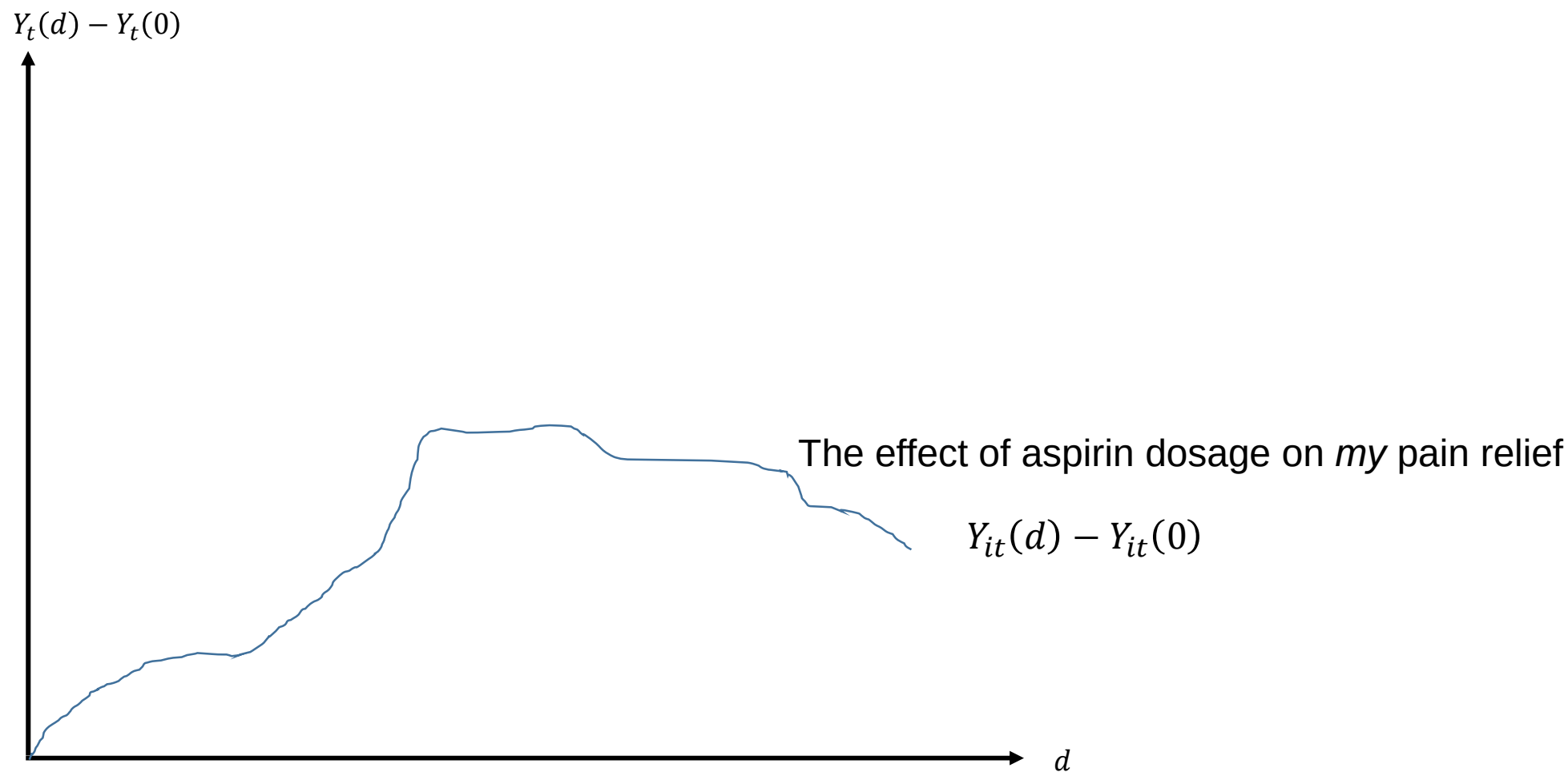
$$Y_{it}(\textcolor{red}{d}) - Y_{it}(0)$$

Average treatment effect on the treated in period t :

$$ATT(d|d) \equiv E[Y_{it}(\textcolor{red}{d}) - Y_{it}(0) | D_i = \textcolor{red}{d}]$$

But this one “treatment” could be anything, any dose, any intensity.

“Treatment Effect” function: $Y_{it}(d) - Y_{it}(0)$



“Treatment Effect” function: $Y_{it}(d) - Y_{it}(0)$

$Y_t(d) - Y_t(0)$

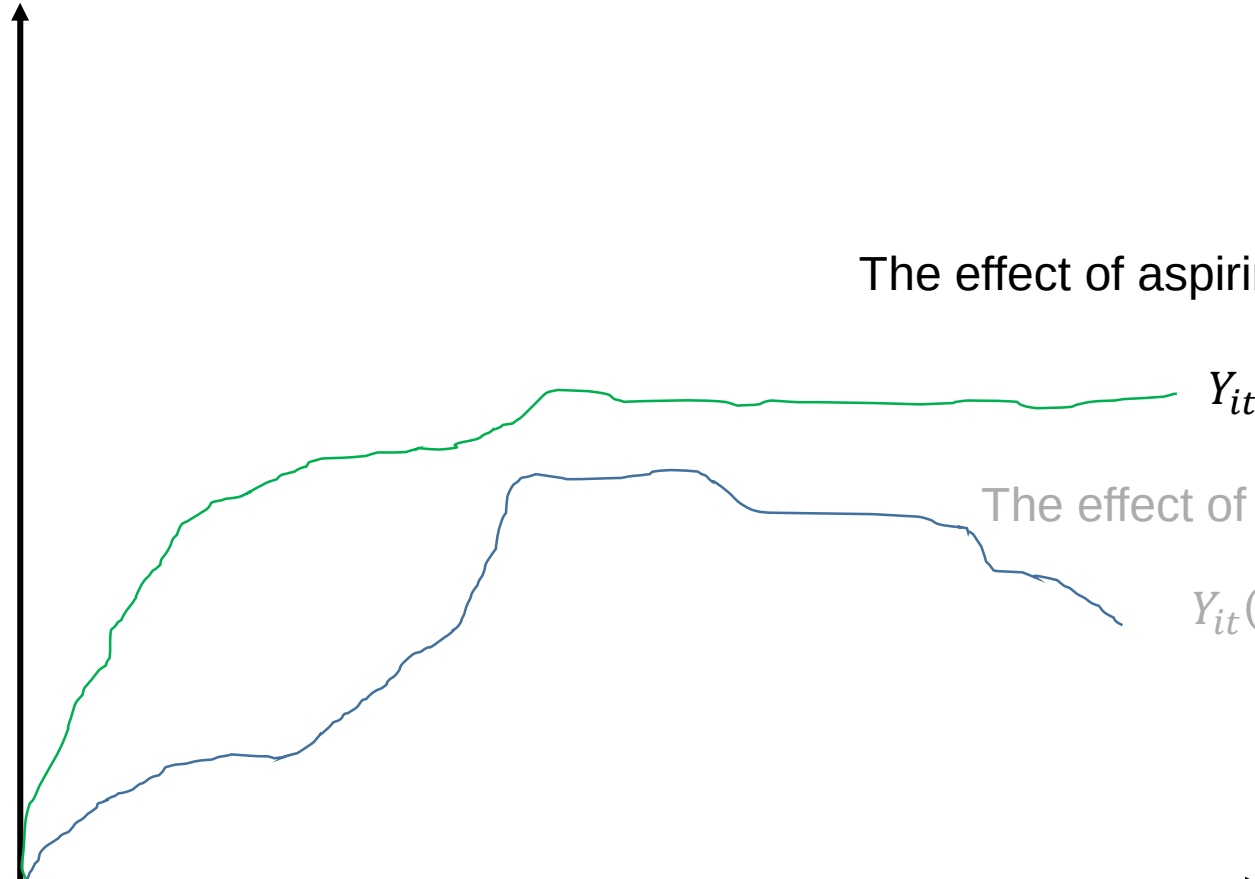
The effect of aspirin dosage on *your* pain relief

$Y_{it}(d) - Y_{it}(0)$

The effect of aspirin dosage on *my* pain relief

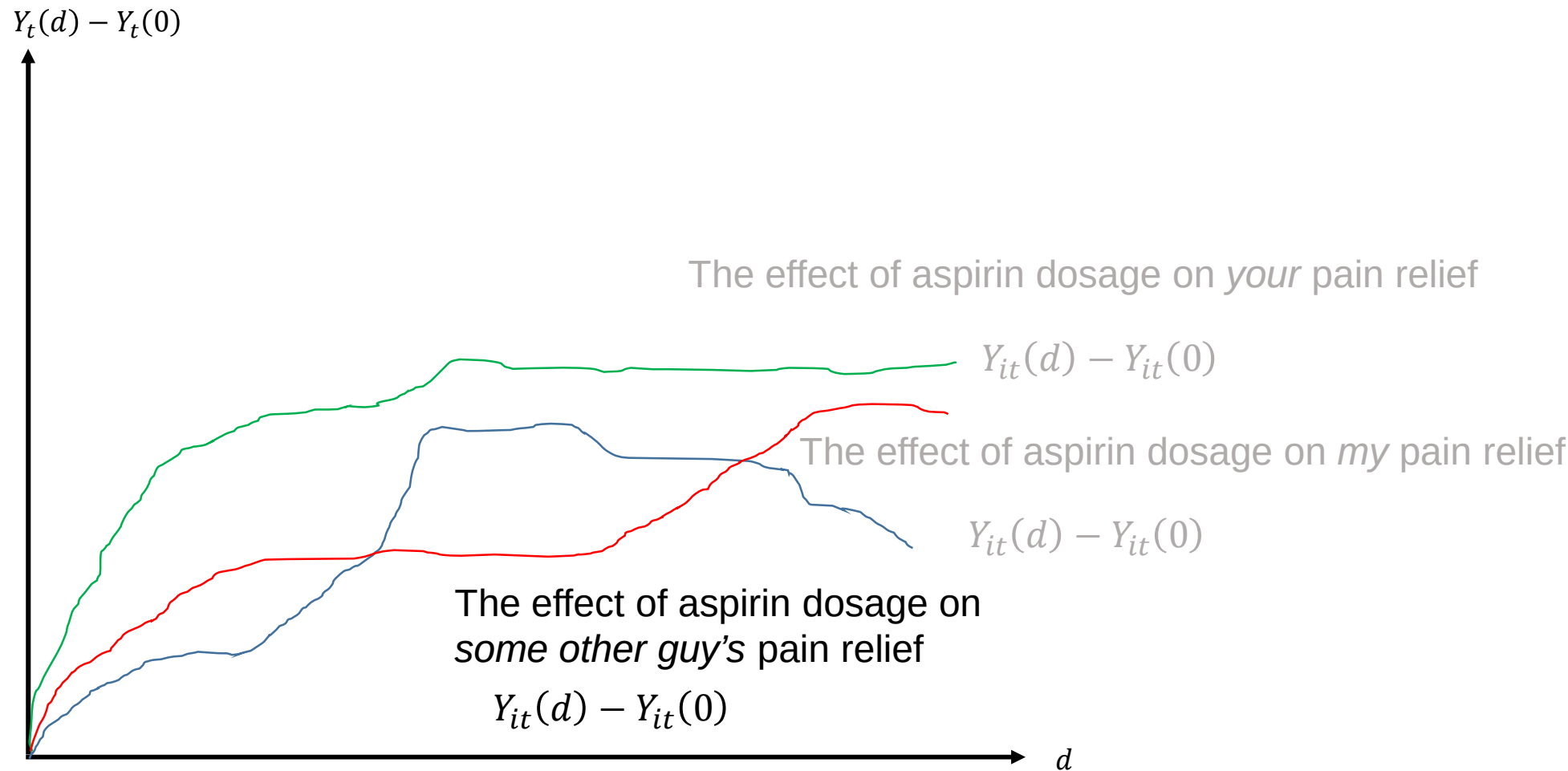
$Y_{it}(d) - Y_{it}(0)$

d

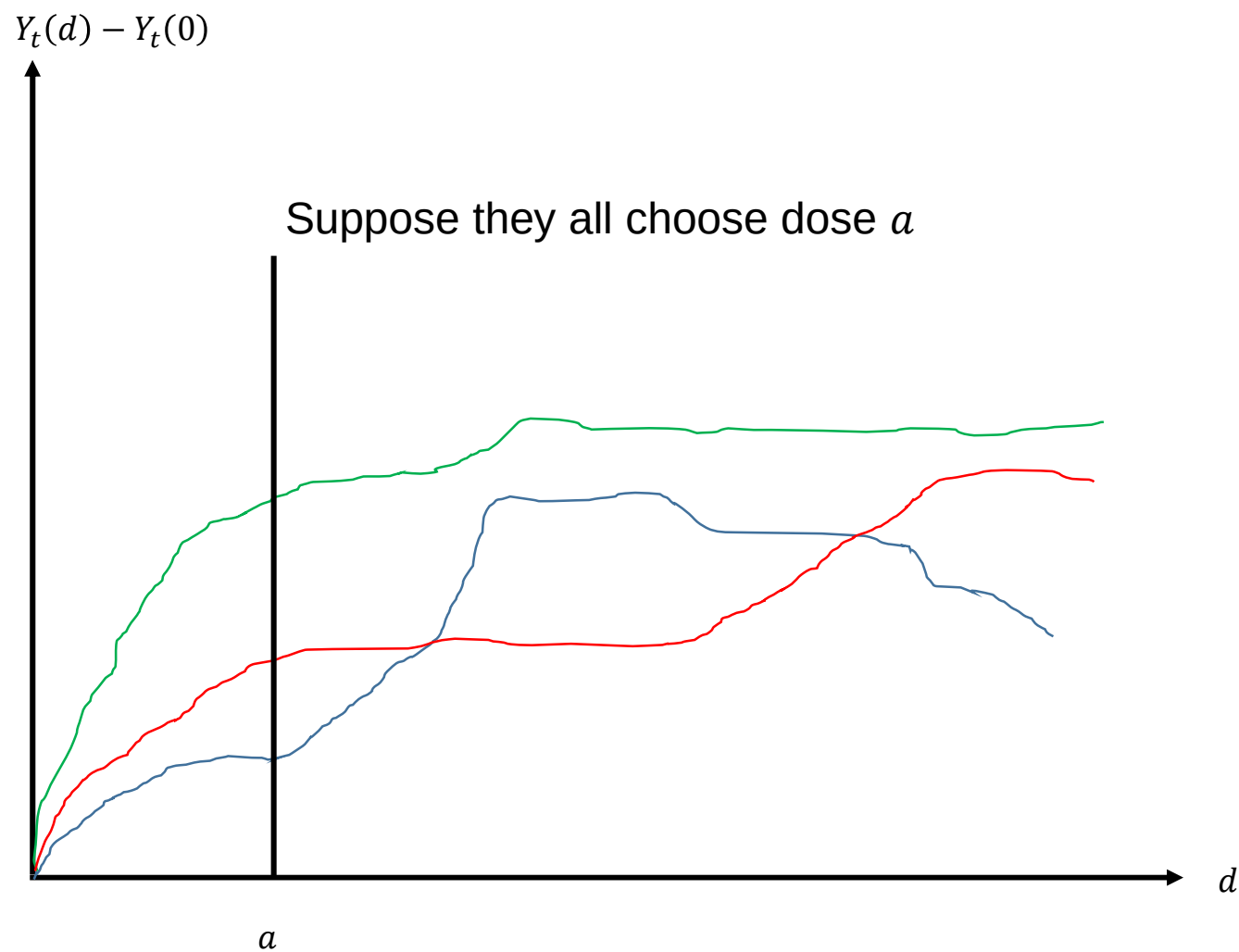


The graph illustrates the treatment effect of aspirin dosage on pain relief for two individuals. The vertical axis represents the treatment effect, $Y_t(d) - Y_t(0)$, and the horizontal axis represents the dosage, d . The green curve, labeled 'The effect of aspirin dosage on *your* pain relief', shows a higher treatment effect than the blue curve, labeled 'The effect of aspirin dosage on *my* pain relief'. Both curves start at the origin, rise to a peak, and then decline. The green curve is consistently higher than the blue curve, indicating that the treatment effect is greater for the first individual.

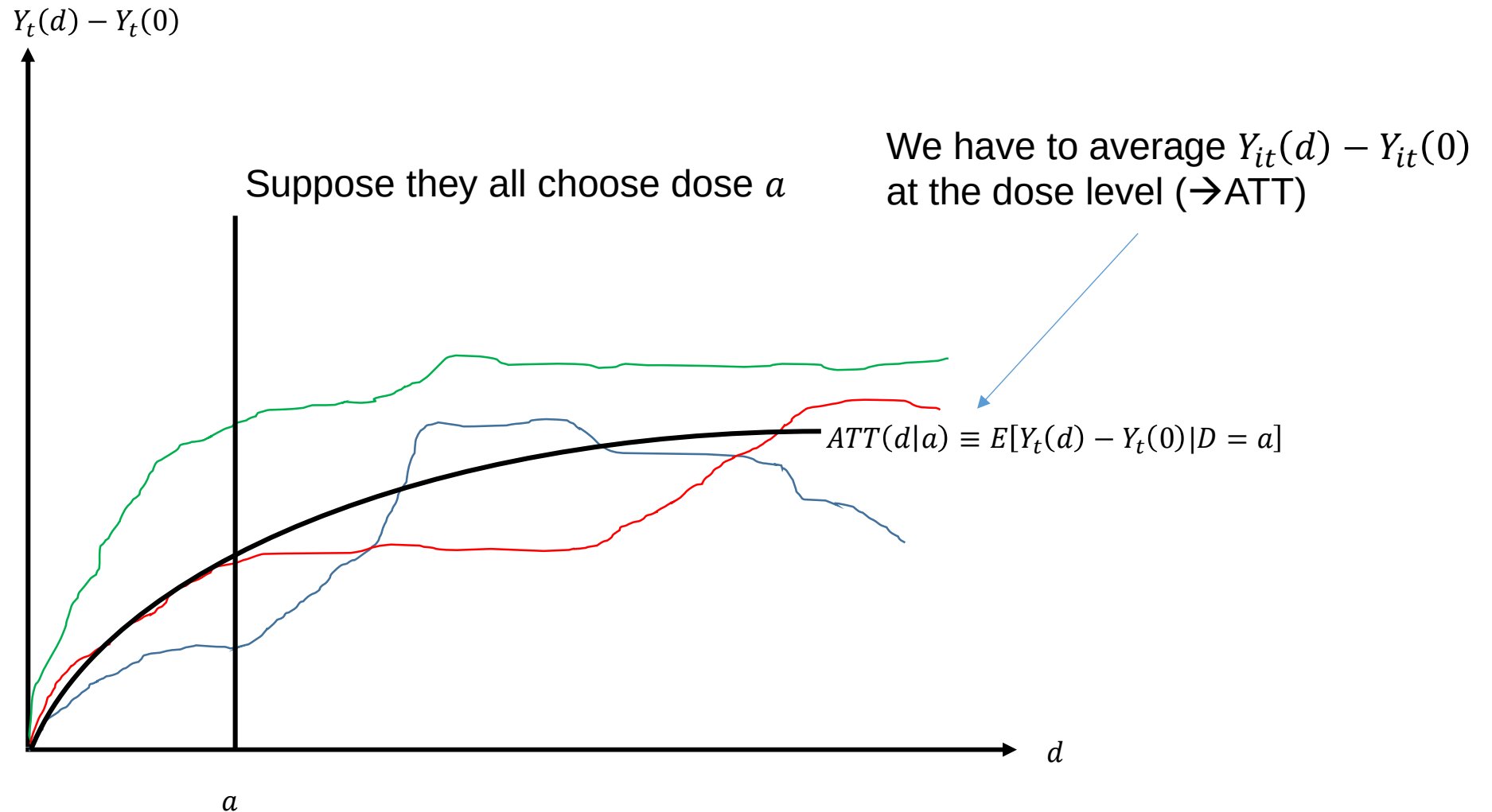
“Treatment Effect” function: $Y_{it}(d) - Y_{it}(0)$



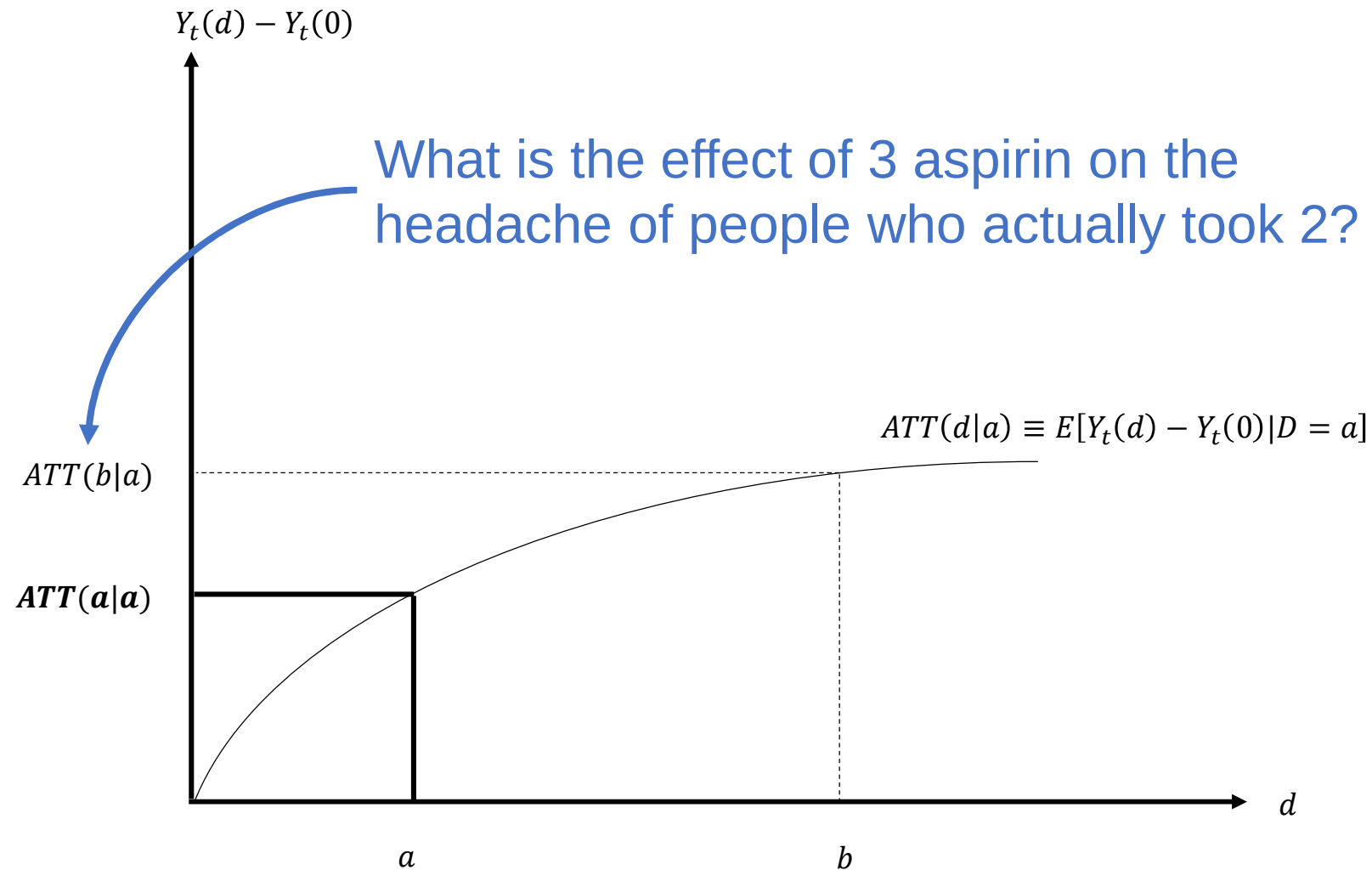
“Treatment Effect” function: $Y_{it}(d) - Y_{it}(0)$



“Treatment Effect” function: $Y_{it}(d) - Y_{it}(0)$



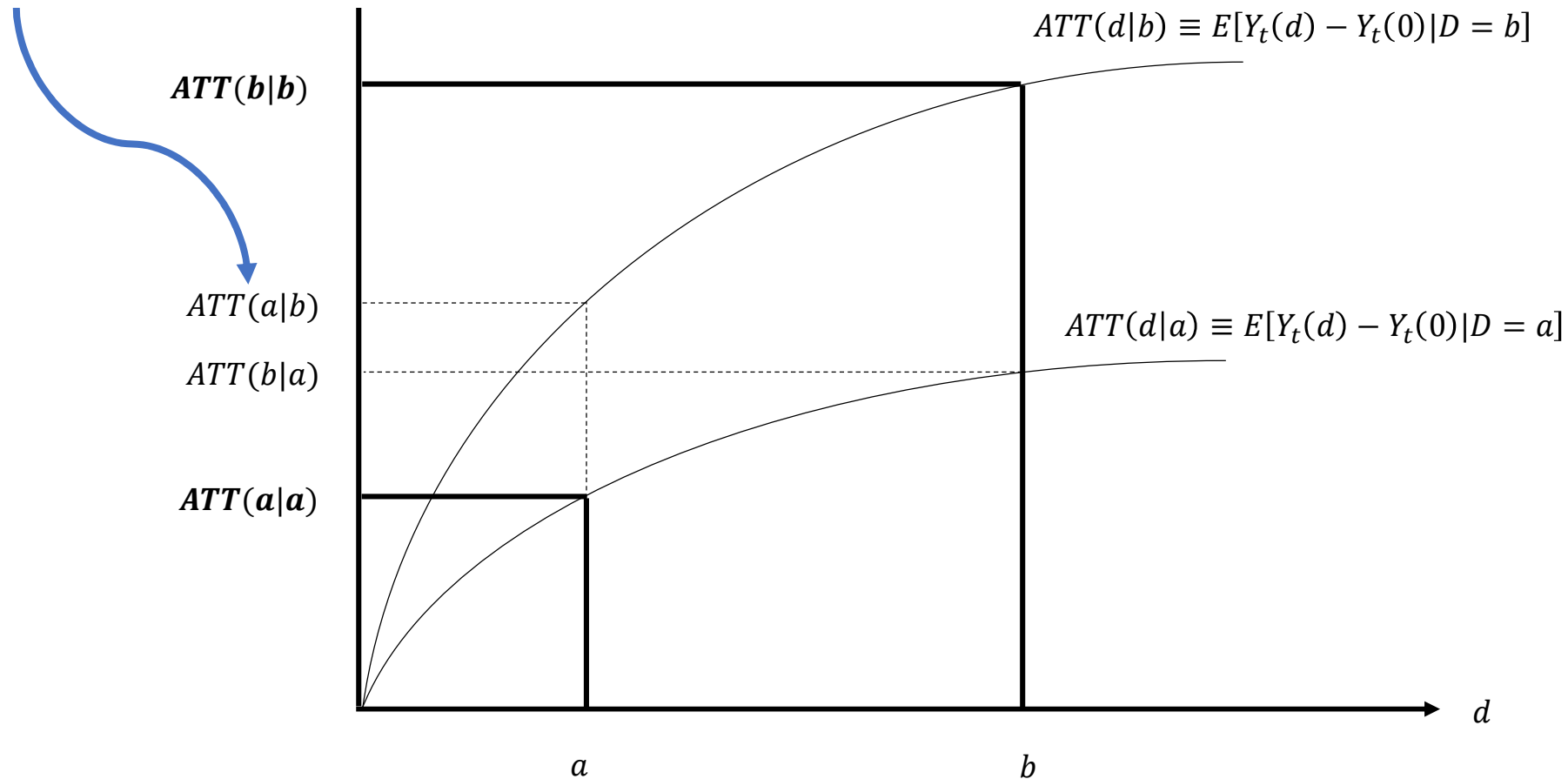
Average Treatment Effects on the Treated: $ATT(d|d)$



Average Treatment Effects on the Treated: $ATT(d|d)$

$$Y_t(d) - Y_t(0)$$

What is the effect of 2 aspirin on the headache of people who actually took 3?



Building Blocks: Slopes

$$ATT(d|d) \equiv E[Y_{it}(d) - Y_{it}(0) | D_i = d]$$

No treatment is not the only counterfactual!

→ The effect of one treatment level versus an “adjacent” treatment level:

**Two-Stage Least Squares Estimation
of Average Causal Effects in Models
With Variable Treatment Intensity**

Joshua D. ANGRIST and Guido W. IMBENS*

Building Blocks: Slopes

$$ATT(d|d) \equiv E[Y_{it}(d) - Y_{it}(0) | D_i = d]$$

No treatment is not the only counterfactual!

→ The effect of one treatment level versus an “adjacent” treatment level:

We refer to the parameter β as the **average causal response** (ACR). This parameter captures a weighted average of causal responses to a unit change in treatment, for those whose treatment status is affected by the instrument. The weight

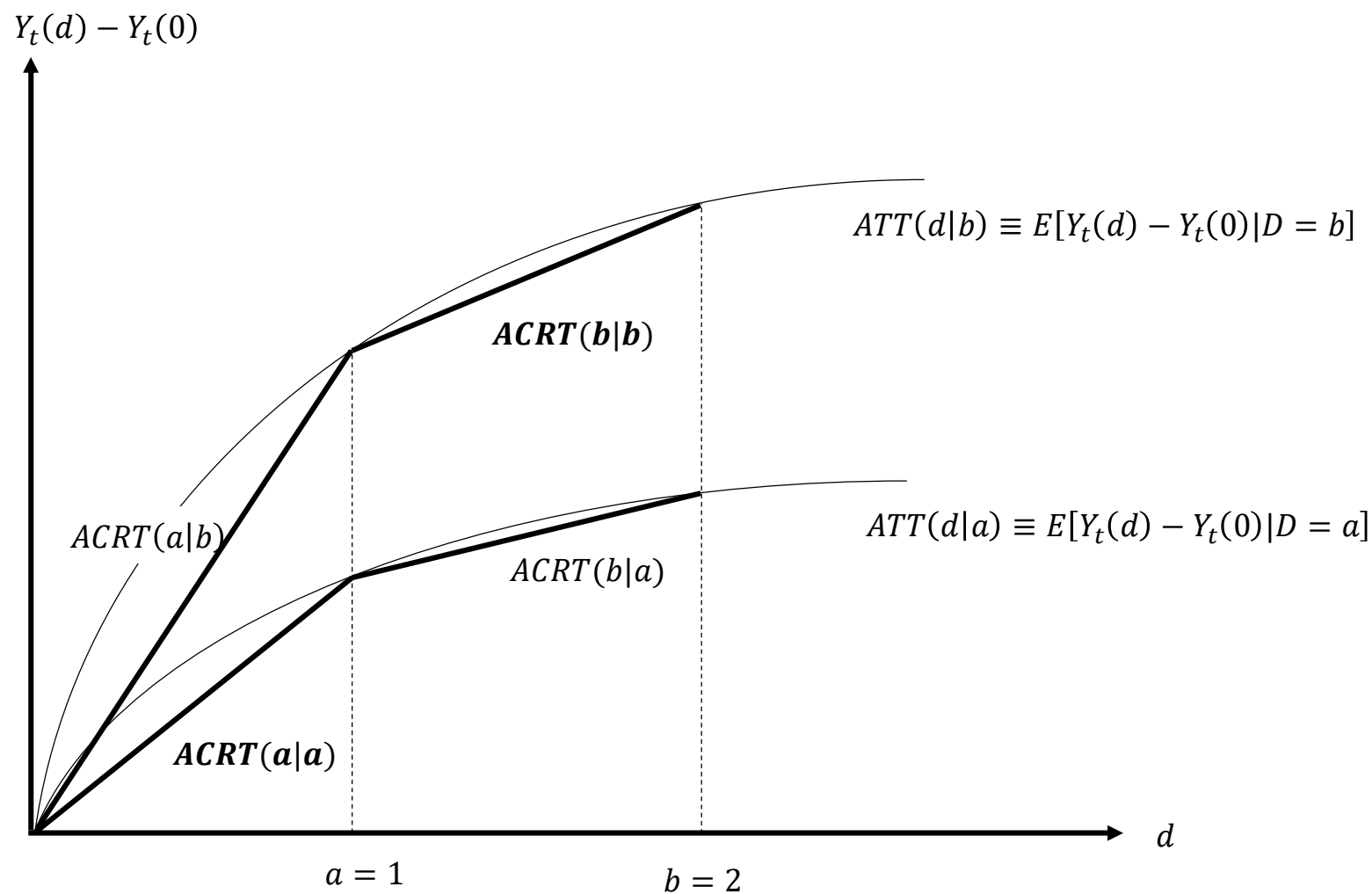
Building Blocks: Slopes

Discrete/multi-valued treatment: *“for units that actually got d_j ”*

$$ACRT(d_j|d_j) \equiv E \left[\underbrace{Y_{it}(d_j) - Y_{it}(d_{j-1})}_{\text{causal effect of the } j^{\text{th}} \text{ dose increment for unit } i} \middle| D_i = d_j \right]$$

“the causal effect of the j^{th} dose increment for unit i ”

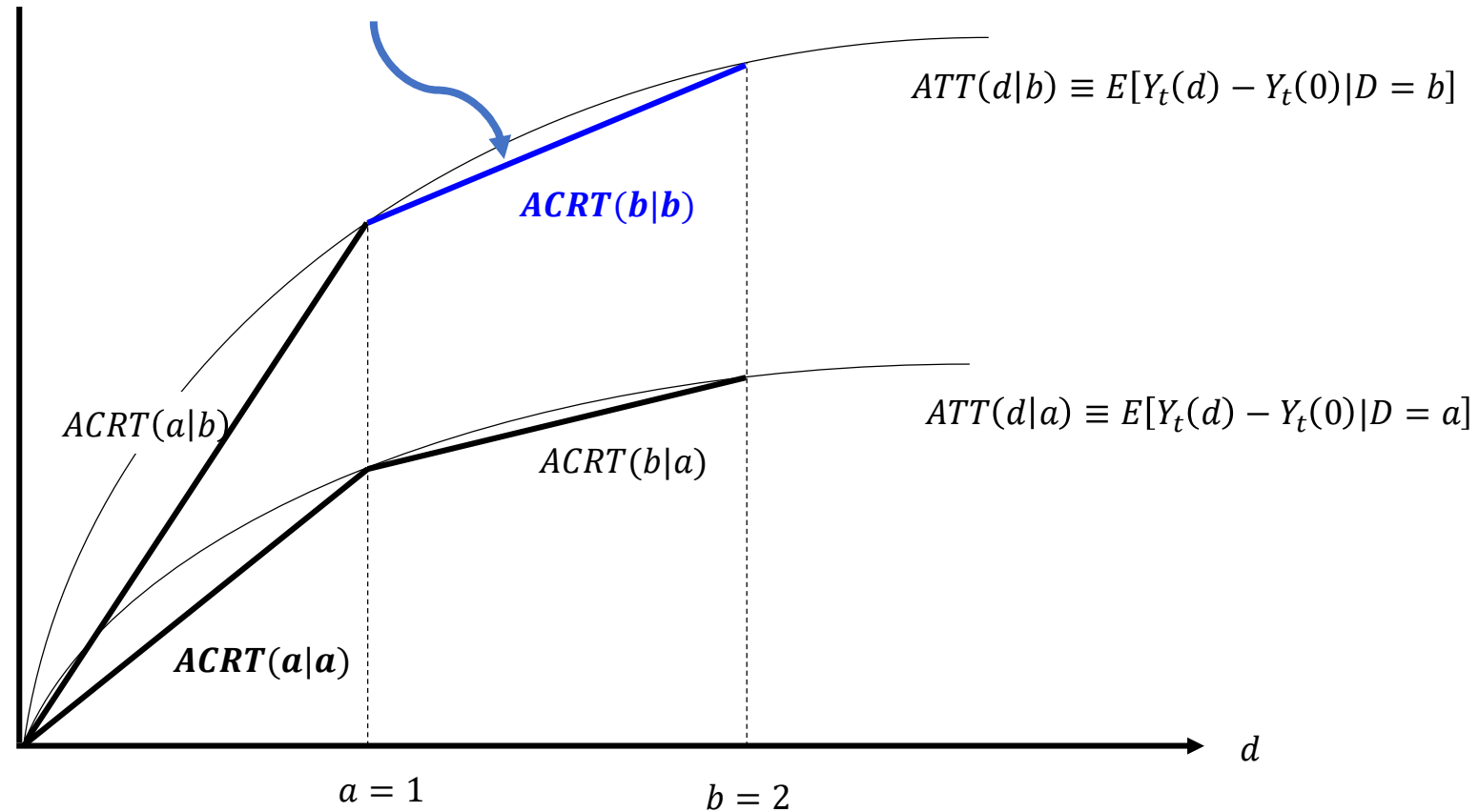
Average Causal Responses on the Treated, Discrete: $ACRT(d_j|d_j)$



Average Causal Responses on the Treated, Discrete: $ACRT(d_j|d_j)$

$$Y_t(d) - Y_t(0)$$

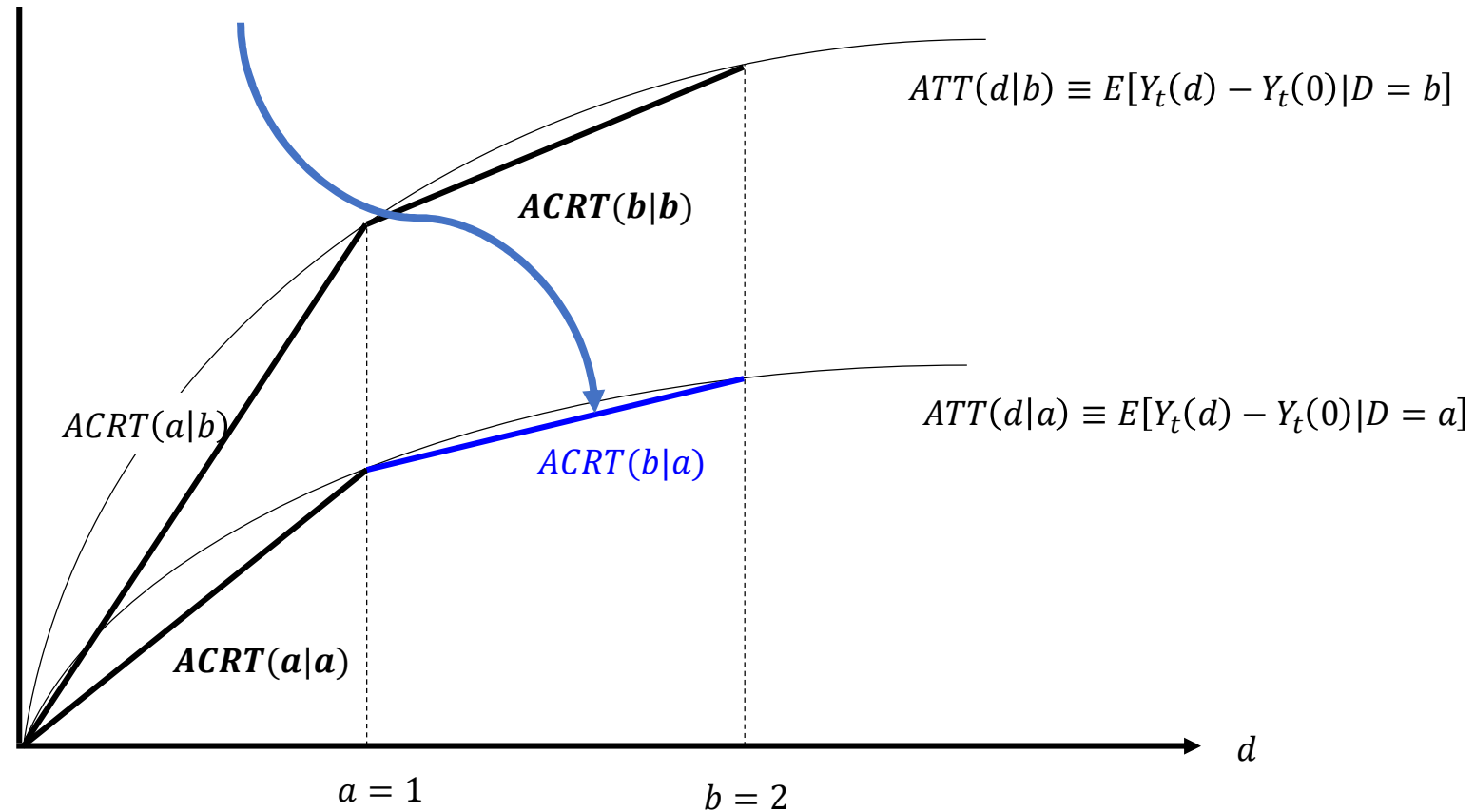
What is the effect of the 2nd aspirin on the headache of people who actually took 2?



Average Causal Responses on the Treated, Discrete: $ACRT(d_j|d_j)$

$$Y_t(d) - Y_t(0)$$

What is the effect of the 2nd aspirin on the headache of people who actually took 1?



Building Blocks: ACRTs

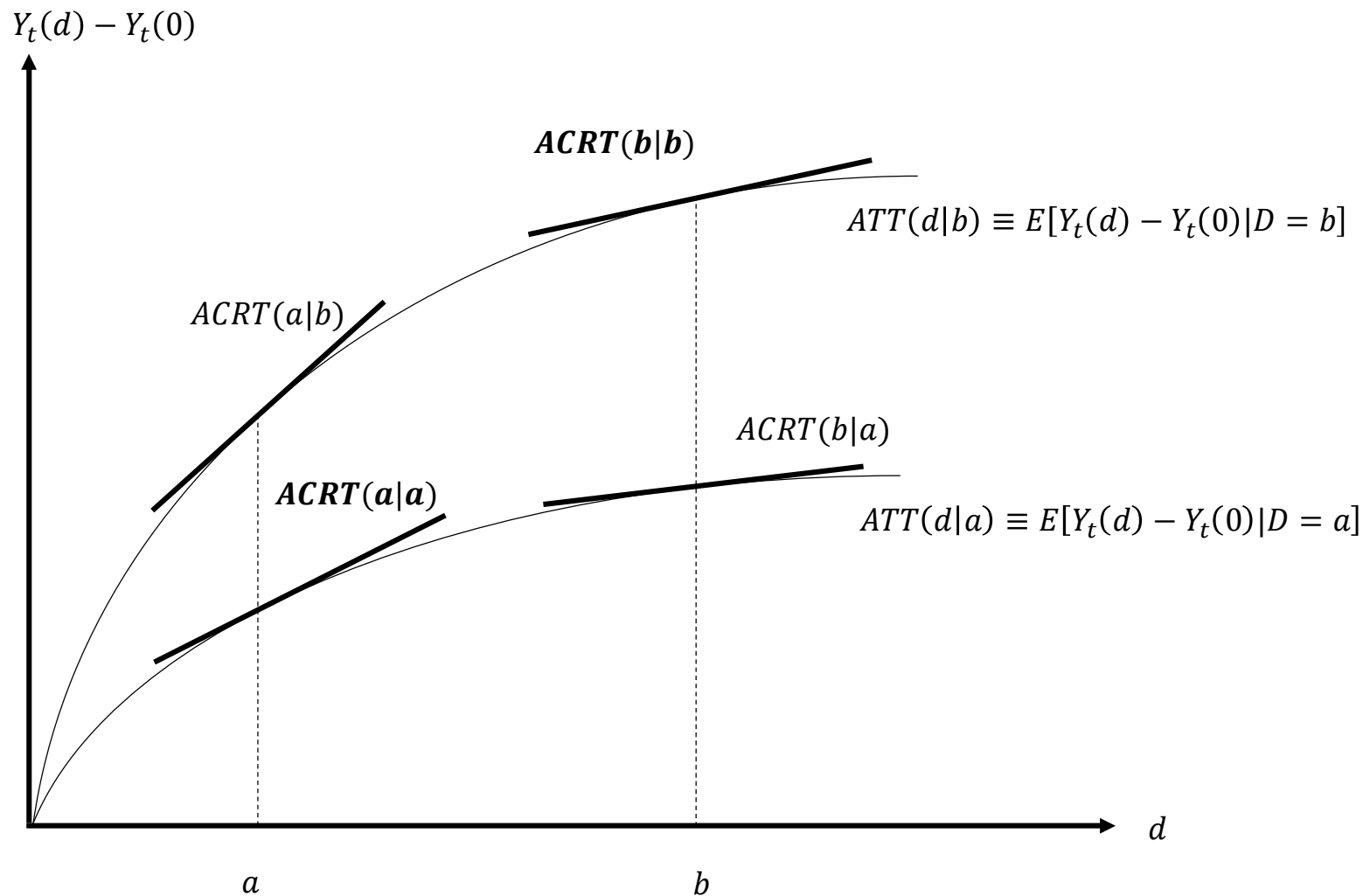
Discrete/multi-valued treatment:

$$ACRT(d_j|d_j) \equiv E[Y_{it}(d_j) - Y_{it}(d_{j-1}) | D_i = d_j]$$

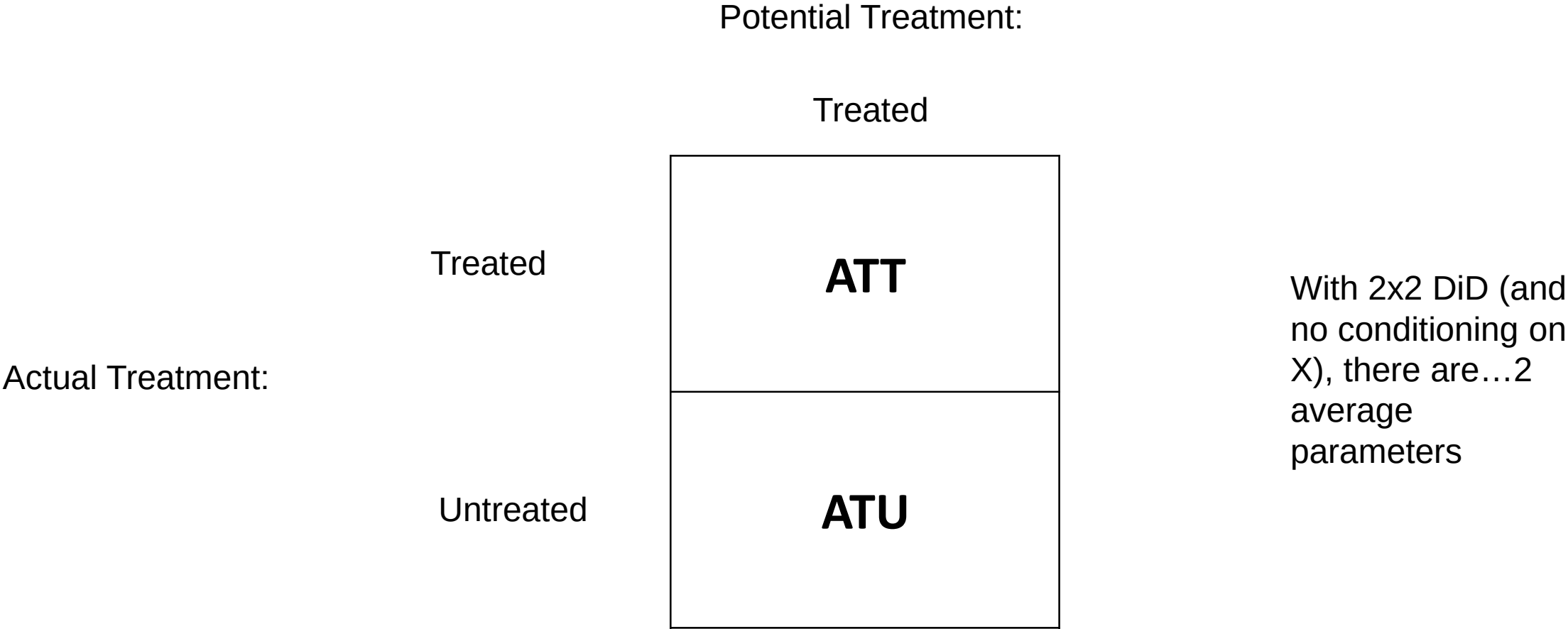
Continuous treatment:

$$ACRT(d|d) \equiv \left. \frac{\partial E[Y_{it}(l) | D_i = d]}{\partial l} \right|_{l=d}$$

Average Causal Responses on the Treated, Continuous: $ACRT(d|d)$



Comment 1: This is a lot of causal parameters



Comment 1: This is a lot of causal parameters

		Potential Treatment:					
		d_1	d_2	d_3	...	d_{J-1}	d_J
Actual Treatment:	d_1	$ATT(\mathbf{d}_1 \mathbf{d}_1)$	$ATT(d_2 d_1)$	$ATT(d_3 d_1)$		$ATT(d_{J-1} d_1)$	$ATT(d_J d_1)$
	d_2	$ATT(d_1 d_2)$	$ATT(\mathbf{d}_2 \mathbf{d}_2)$	$ATT(d_3 d_2)$		$ATT(d_{J-1} d_2)$	$ATT(d_J d_2)$
	d_3	$ATT(d_1 d_3)$	$ATT(d_2 d_3)$	$ATT(\mathbf{d}_3 \mathbf{d}_3)$		$ATT(d_{J-1} d_3)$	$ATT(d_J d_3)$
	$\vdots$				$\ddots$		
	d_{J-1}	$ATT(d_1 d_{J-1})$	$ATT(d_2 d_{J-1})$	$ATT(d_3 d_{J-1})$		$ATT(\mathbf{d}_{J-1} \mathbf{d}_{J-1})$	$ATT(d_J d_{J-1})$
	d_J	$ATT(d_1 d_J)$	$ATT(d_2 d_J)$	$ATT(d_3 d_J)$		$ATT(d_{J-1} d_J)$	$ATT(\mathbf{d}_J \mathbf{d}_J)$

Comment 1: This is a lot of causal parameters

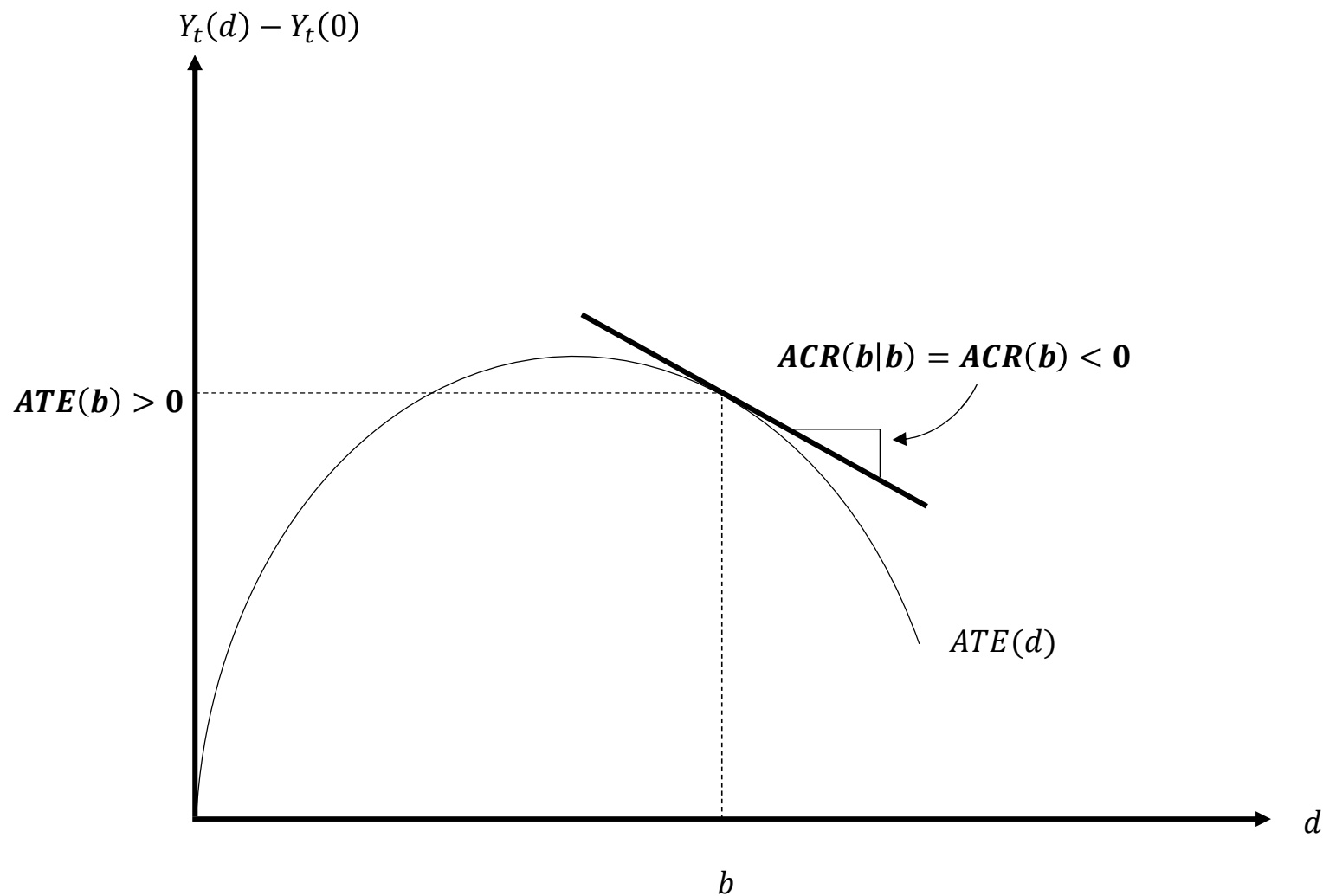
Potential Treatment:

Actual Treatment:

	d_1	d_2	d_3	...	d_{J-1}	d_J
d_1	$ACRT(d_2 d_1)$ $ATT(d_1 d_1)$	$ACRT(d_3 d_1)$ $ATT(d_2 d_1)$	$ATT(d_3 d_1)$		$ACRT(d_J d_1)$ $ATT(d_{J-1} d_1)$	$ATT(d_J d_1)$
d_2	$ACRT(d_2 d_2)$ $ATT(d_1 d_2)$	$ACRT(d_3 d_2)$ $ATT(d_2 d_2)$	$ATT(d_3 d_2)$		$ACRT(d_J d_2)$ $ATT(d_{J-1} d_2)$	$ATT(d_J d_2)$
d_3	$ACRT(d_2 d_3)$ $ATT(d_1 d_3)$	$ACRT(d_3 d_3)$ $ATT(d_2 d_3)$	$ATT(d_3 d_3)$		$ACRT(d_J d_3)$ $ATT(d_{J-1} d_3)$	$ATT(d_J d_3)$
$\vdots$				$\ddots$		
d_{J-1}	$ACRT(d_2 d_{J-1})$ $ATT(d_1 d_{J-1})$	$ACRT(d_3 d_{J-1})$ $ATT(d_2 d_{J-1})$	$ATT(d_3 d_{J-1})$		$ACRT(d_J d_{J-1})$ $ATT(d_{J-1} d_{J-1})$	$ATT(d_J d_{J-1})$
d_J	$ACRT(d_2 d_J)$ $ATT(d_1 d_J)$	$ACRT(d_3 d_J)$ $ATT(d_2 d_J)$	$ATT(d_3 d_J)$		$ACRT(d_J d_J)$ $ATT(d_{J-1} d_J)$	$ATT(d_J d_J)$

And if D is continuous these are functions!

Comment 2: The distinction matters!



What are you willing to believe?

Parallel Trends

For all $d \in \mathcal{D}$:

$$E[Y_{it}(0) - Y_{it-1}(0)] = E[Y_{it}(0) - Y_{it-1}(0) | D_i = d]$$

Extension of binary parallel trends (when $D \in \{0,1\}$).

Identifying ATT

This is just like traditional 2x2 DD



$$E[\Delta Y_{it} | D_i = d] - E[\Delta Y_{it} | D_i = 0]$$

No Anticipation

$$= E[Y_{it}(d) - Y_{it-1}(0) | D_i = d] - E[Y_{it}(0) - Y_{it-1}(0) | D_i = 0]$$

PT

$$= E[Y_{it}(d) - Y_{it}(0) | D_i = d]$$

$$= \mathbf{ATT}(d|d)$$

Identifying ACRT

(No Ant. + PT)

“Compare the effect of d_j to the effect of d_{j-1} ”

$$ATT(d_j | d_j) - ATT(d_{j-1} | d_{j-1})$$

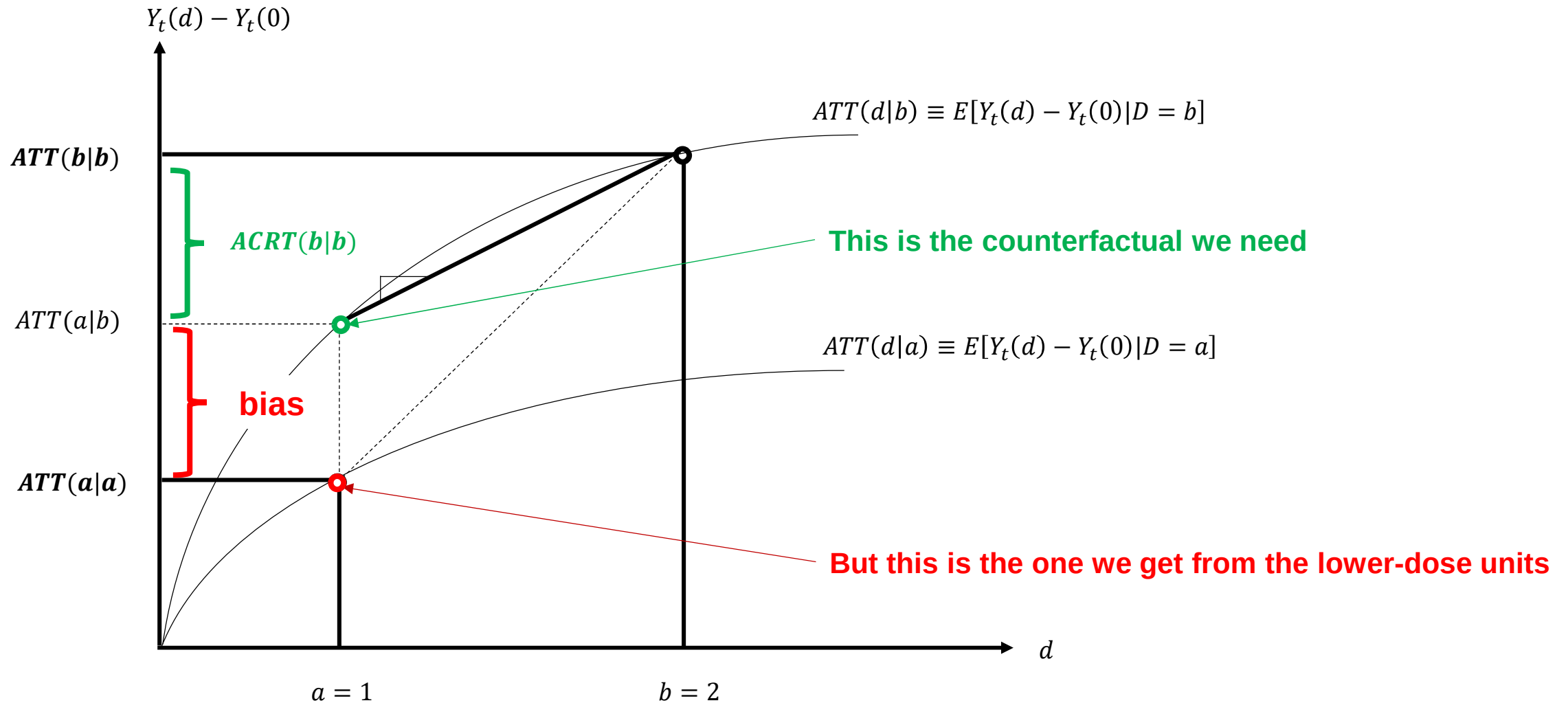
$$= \left(ATT(d_j | d_j) - ATT(d_{j-1} | d_j) \right) + \left(ATT(d_{j-1} | d_j) - ATT(d_{j-1} | d_{j-1}) \right)$$

$$= \underbrace{ACRT(d_j | d_j)}_{\text{Interpretable causal effect}} + \underbrace{\left(ATT(d_{j-1} | d_j) - ATT(d_{j-1} | d_{j-1}) \right)}_{\text{“Selection bias” (heterogeneous effects at } d_{j-1})}$$

Interpretable causal effect

“Selection bias” (heterogeneous effects at d_{j-1})

Identifying ACRT

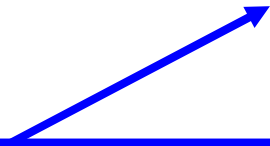


Strong Parallel Trends

PT + homogenous treatment effect functions:

$$E[Y_{it}(0) - Y_{it-1}(0)] = E[Y_{it}(0) - Y_{it-1}(0) | D_i = d] \quad \forall d$$

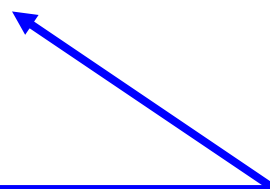
Groups don't change differentially
for non-treatment-related reasons



+

$$ATT(d|d) = ATT(d|d') \quad \forall d, d'$$

Nor would they have change for treatment
effect reasons had they gotten the lower dose



We (almost) never say this

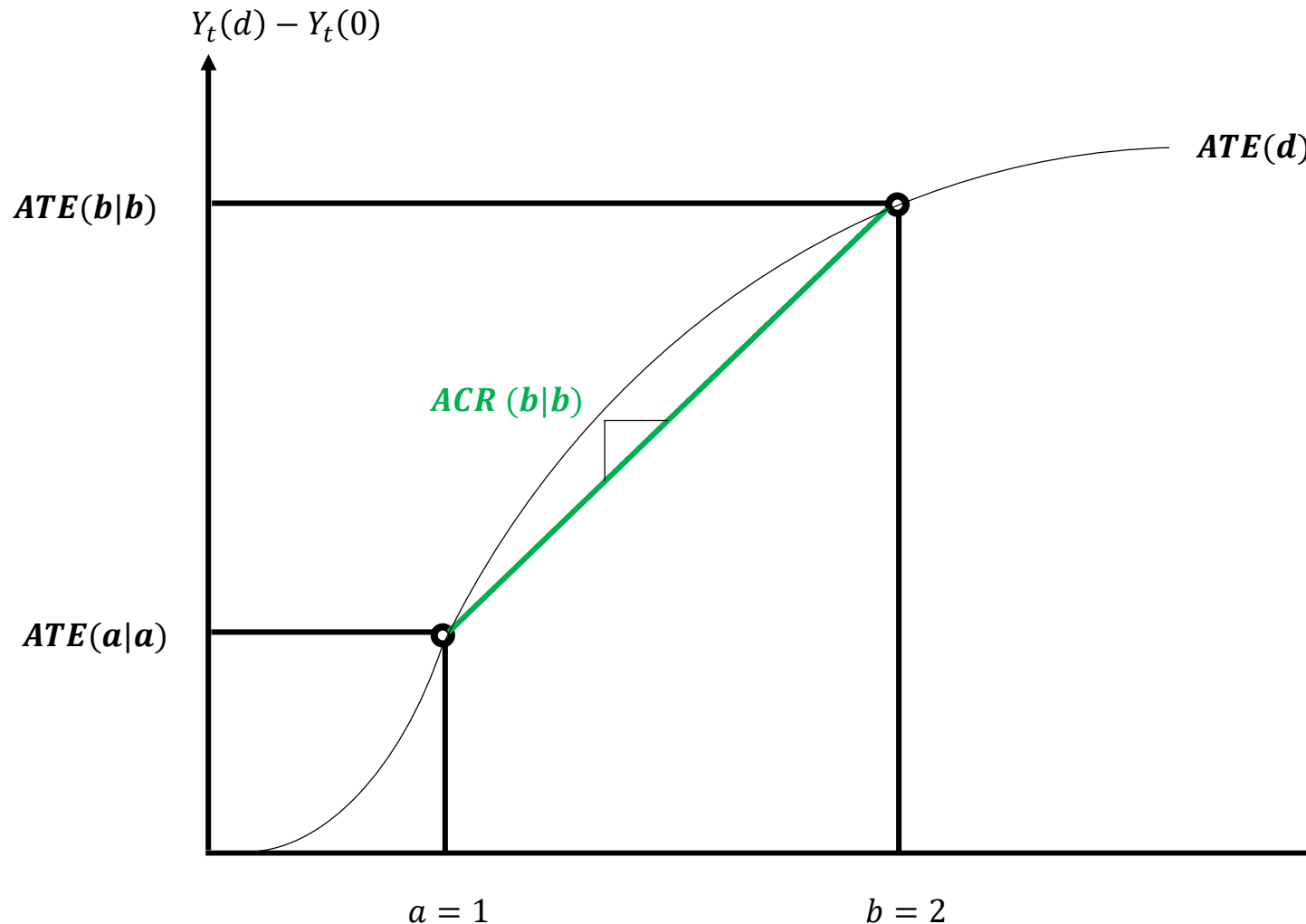
Stock Market Wealth and the Real Economy: A Local Labor Market Approach[†]

By GABRIEL CHODOROW-REICH, PLAMEN T. NENOV, AND ALP SIMSEK*

The possibility of heterogeneous MPCs also has implications for the interpretation of our baseline coefficients. In general, when treatment effects are correlated with the regressor, the OLS coefficient in a specification without treatment effect heterogeneity need not lie in the convex hull of the individual treatment effects; intuitively, if low wealth areas have high MPCs and high wealth areas have low MPCs, an increase in the stock market could induce the same change in spending in both low and high wealth areas.

*also Saez, Slemrod, Giertz 2012 JEL.

Identifying ACR under SPT: “same curve”



Good:

Identifies $ACR(d)$ and $ATE(d)$

(...not $ATT(d|d)$ and $ACRT(d|d)$!)

Bad:

We were never willing to assume
SPT in binary designs—too strong!

inching closer to IV independence
assumption/quasi-randomness

TWFE: 2 Periods

Acemoglu and Finkelstein (2008)

Medicare prospective payment system (PPS; 1984) lowered the relative price of capital more so for hospitals with more Medicare patients. What happens to K/L?

$$Y_{it} = \alpha_i + \alpha_t + \overbrace{\hat{\beta}^{TWFE}}^{1.13} D_i \times POST_t + u_{it}$$

← it goes up!

D_i = pre-PPS Medicare share of inpatient days

Acemoglu and Finkelstein (2008)

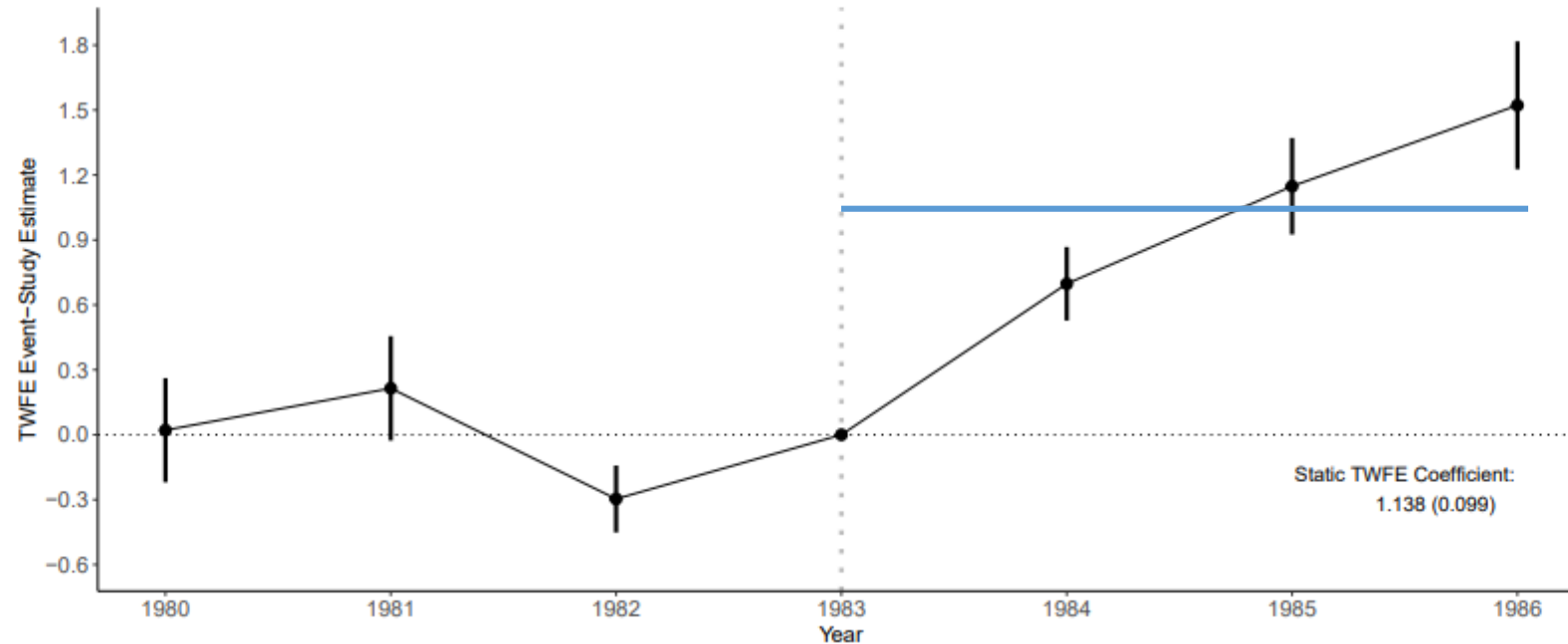
Medicare prospective payment system (PPS; 1984) lowered the relative price of capital more so for hospitals with more Medicare patients. What happens to K/L?

$$\Delta Y_{it} = \overset{1.13}{\hat{\beta}^{TWFE}} D_i + \Delta u_{it}$$

D_i = pre-PPS Medicare share of inpatient days

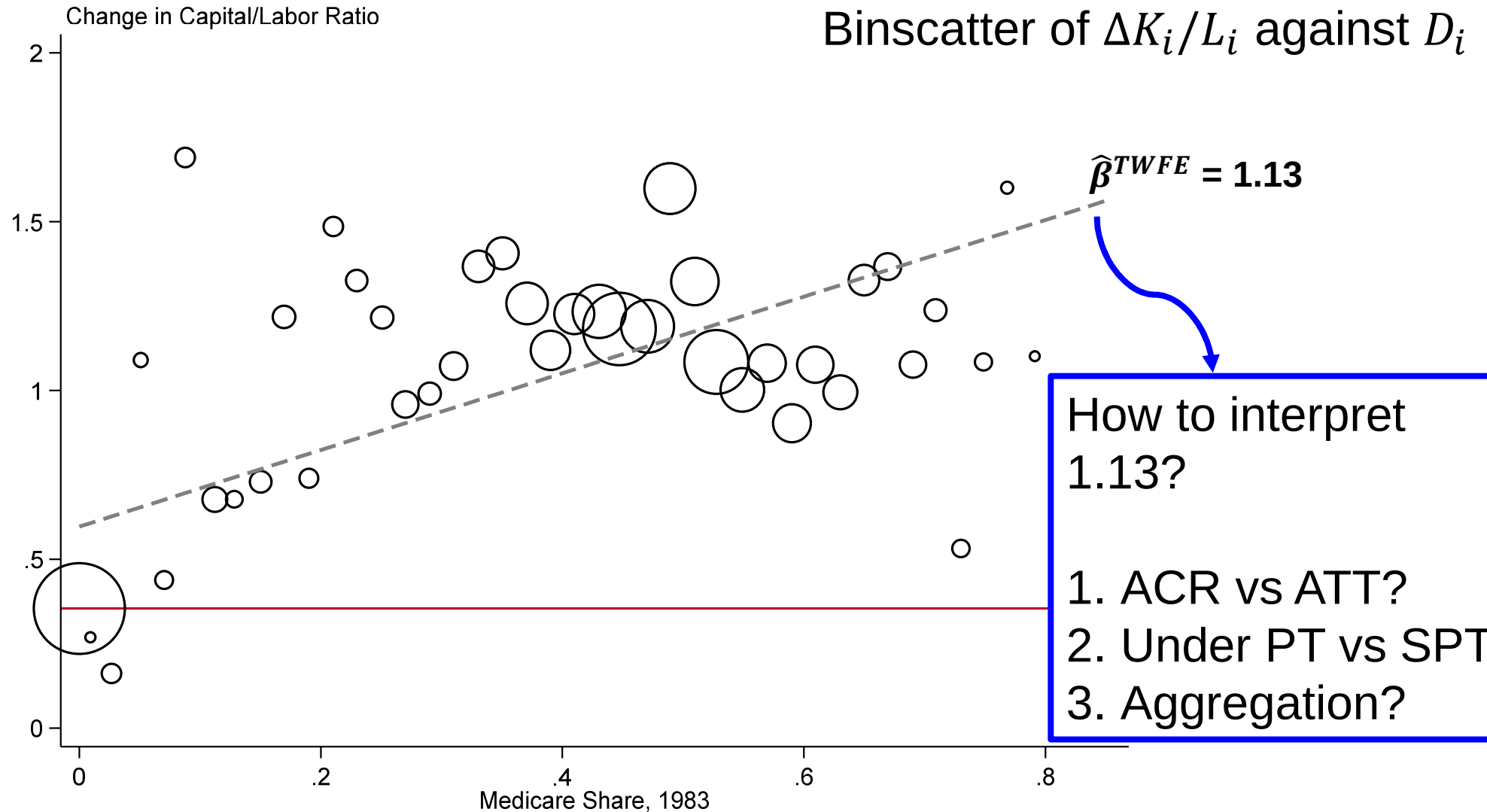
Acemoglu and Finkelstein (2008)

Figure 1: Two-Way Fixed Effects Event-Study Estimates of the Effect of Medicare's Reimbursement Reform on Hospital Input Mix

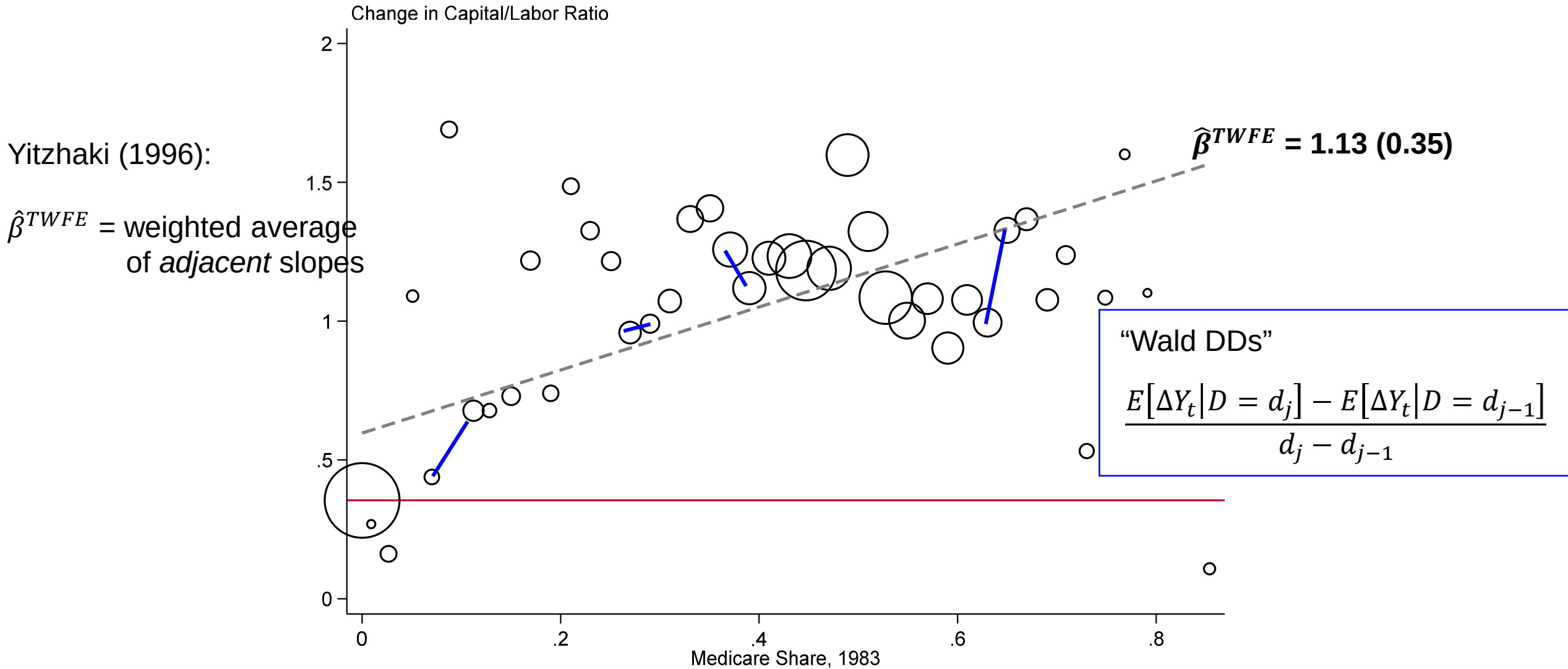


Notes: The figure plots TWFE event-study coefficients and their 95% confidence intervals from regressions with hospital fixed effects, year fixed effects, and the 1983 Medicare inpatient share (m_i) interacted with either a dummy for years after 1983 or the year dummies. The outcome variable is the depreciation share of total operating expenses, a measure of hospitals' capital/labor ratio. The data cover the years 1980-1986 and come from the American Hospital Association's annual survey (American Hospital Association, 1986). We dropped 860 hospitals (out of 6741) that have missing data for the outcome. We also report the static TWFE coefficient and standard errors associated with (1.1). All standard errors are clustered at the hospital level.

A simple scatter plot



A simple scatter plot



Two-way fixed effects, discrete (ACR)

$$\hat{\beta}^{TWFE} = \sum_{d_j \in \mathcal{D}_+} w^{acr}(d_j) \frac{E[\Delta Y_t | D = d_j] - E[\Delta Y_t | D = d_{j-1}]}{d_j - d_{j-1}}$$

Where:

$$w^{acr}(d_j) = \frac{(E[D | D > d_j] - E[D])P(D > d_j)}{V(D)}$$

and that are hump-shaped and centered on $E[D]$

And $w^{acr}(d_j) \geq 0$ and $\sum_{d_j \in \mathcal{D}_+} w^{acr}(d_j) = 1$. Weights are non-negative and sum to one.

How should we interpret 1.13?*

$$\begin{aligned} 1.13 &= \sum_{d_j \in \mathcal{D}_+} w^{acr}(d_j) [ACRT(d_j|d_j) + (ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1}))] \\ &= \sum_{d_j \in \mathcal{D}_+} w^{acr}(d_j) ACRT(d_j|d_j) + \sum_{d_j \in \mathcal{D}_+} w^{acr}(d_j) (ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})) \end{aligned}$$

*Under PT and in terms of ACRT parameters.

How should we interpret 1.13?*

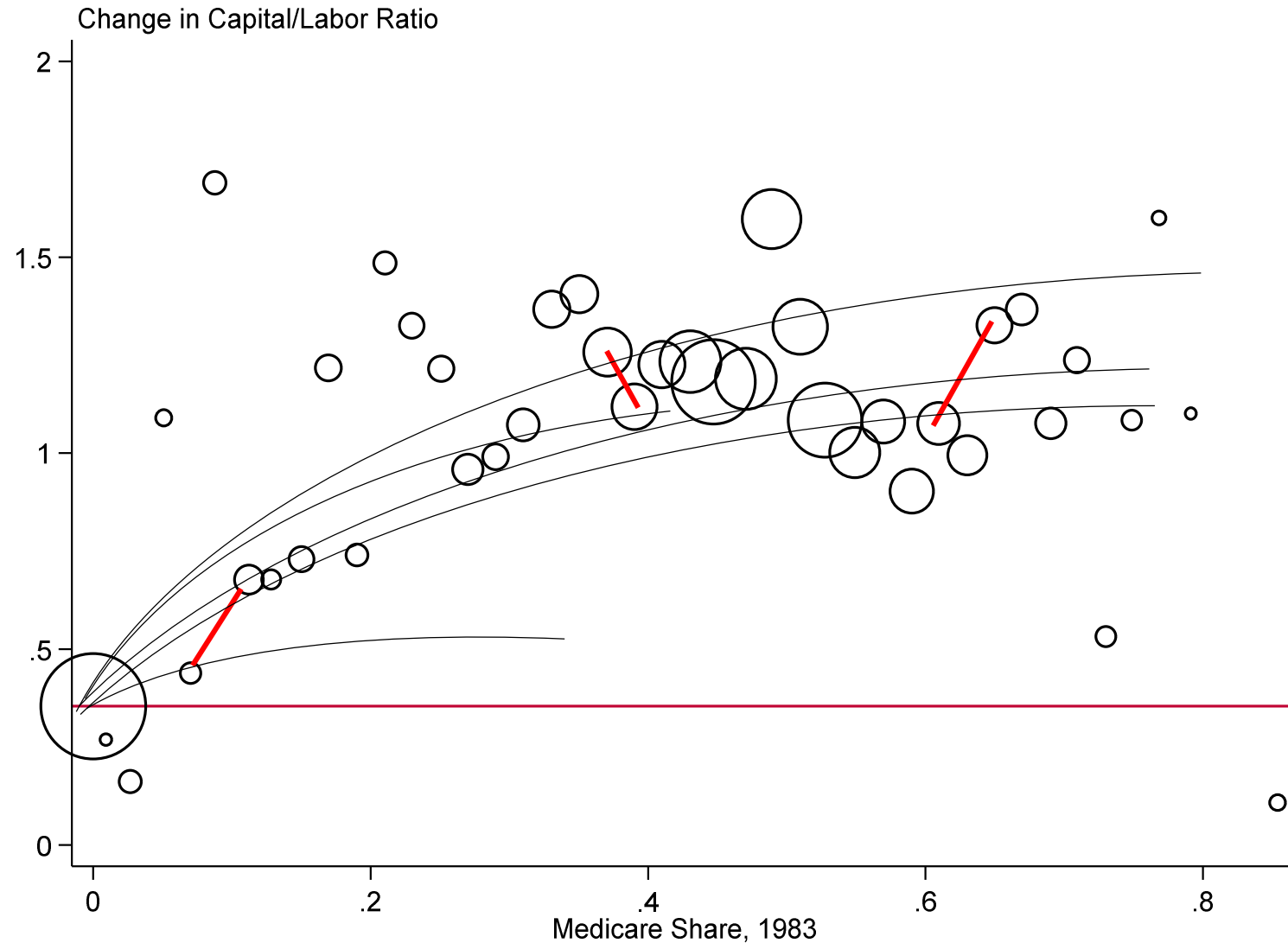
$$\begin{aligned} 1.13 &= \sum_{d_j \in \mathcal{D}_+} w^{acr}(d_j) [ACRT(d_j|d_j) + (ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1}))] \\ &= \sum_{d_j \in \mathcal{D}_+} w^{acr}(d_j) ACRT(d_j|d_j) + \sum_{d_j \in \mathcal{D}_+} w^{acr}(d_j) (ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})) \end{aligned}$$

Problem 1:

combines a “weakly” causal parameter with treatment effect heterogeneity.

*Under PT and in terms of ACRT parameters.

Without SPT, each $\widehat{ACRT}(d|d)$ is biased



With SPT, $w^{acr}(d_j)$ -weighted average of ACR(d)

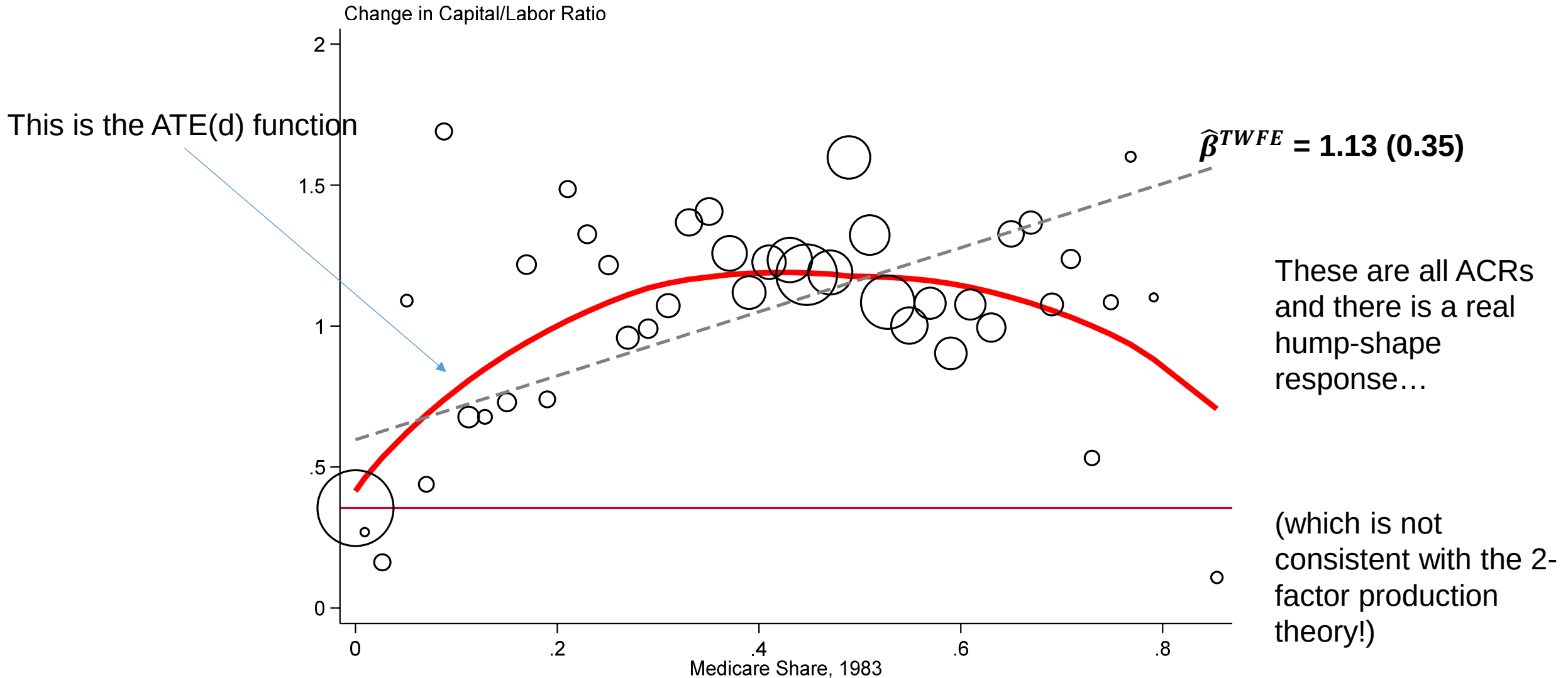
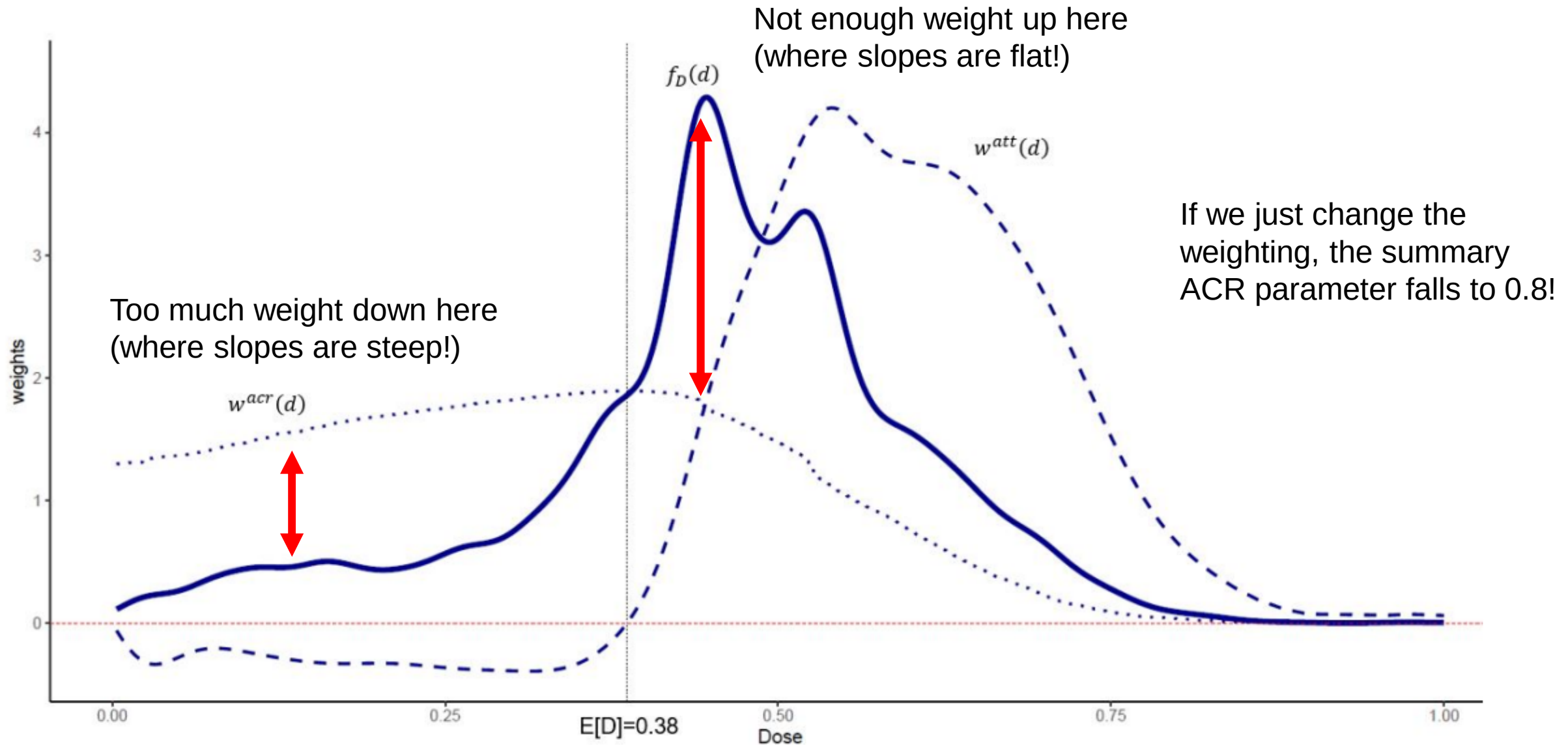
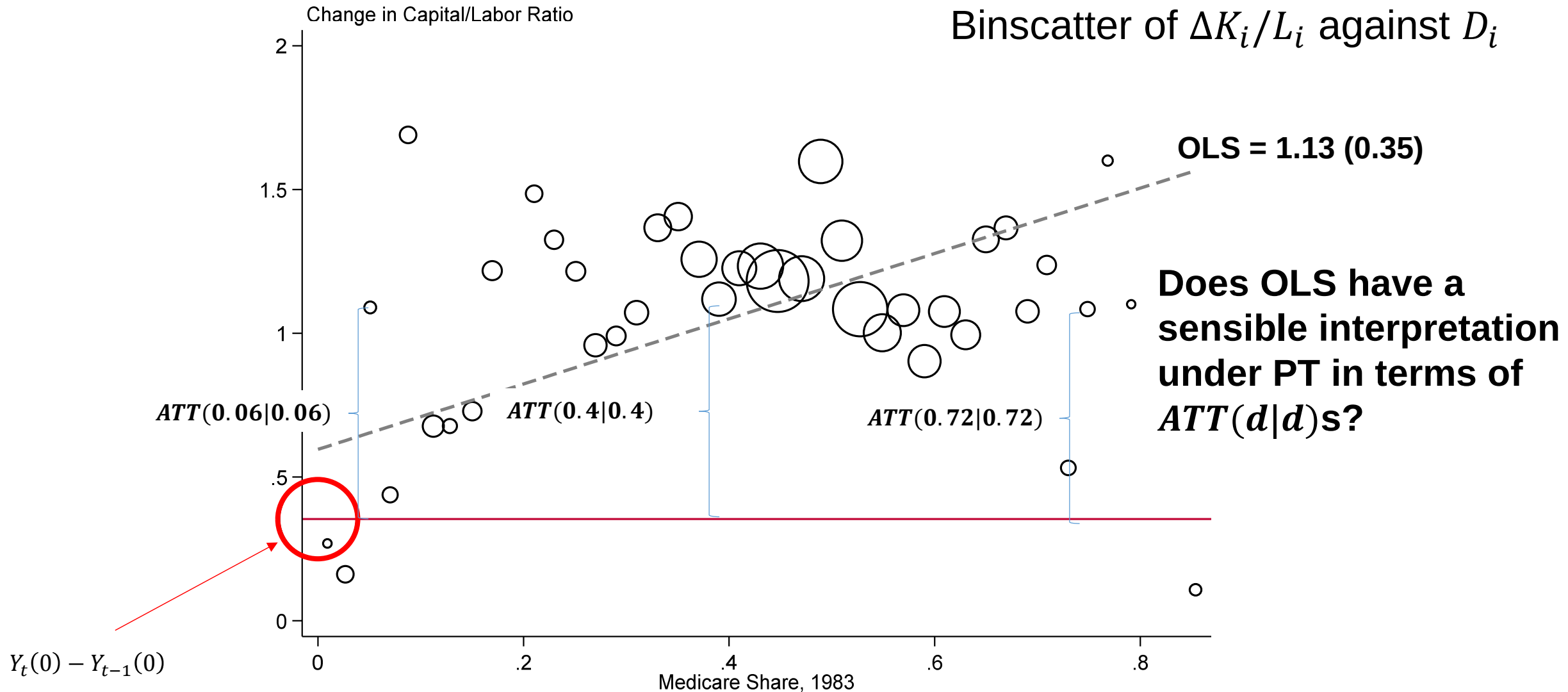


Figure 6: Weighting Schemes: ATT, ACR, and Dose Distribution for AF



The simple scatter plot again



Two-way fixed effects, discrete (ATT)

$$\hat{\beta}^{TWFE} = \sum_{d_j \in \mathcal{D}_+} w^{att}(d_j) \left[E[\Delta Y_t | D = d_j] - E[\Delta Y_t | D = 0] \right]$$

Where:

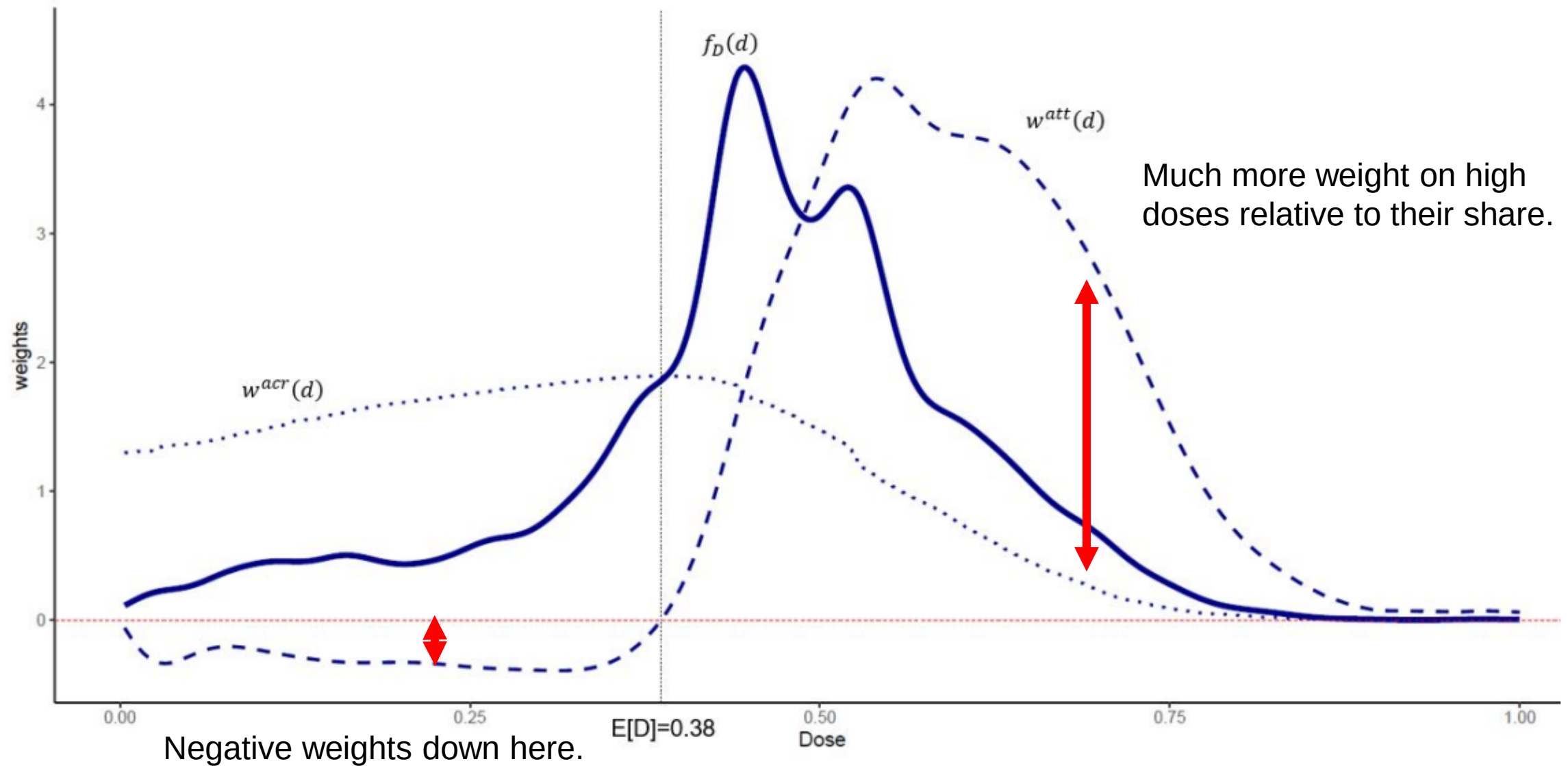
$$w^{att}(d_j) = \frac{(d_j - E[D])P(D = d_j)}{V(D)}$$

Equal zero at $E[D]$,
Over-weight extreme doses

And $w^{att}(d_j) > 0$ if $d_j > E[D]$ and $\sum_{d_j \in \mathcal{D}_+} w^{att}(d_j) = 0$.

Weights are
negative for below
average doses
and sum to zero.

Figure 6: Weighting Schemes: ATT, ACR, and Dose Distribution for AF



TWFE binarizes and rescales

$$1.13 = \frac{E[w_1^{bin}(D)\Delta Y|D > E[D]] - E[w_0^{bin}(D)\Delta Y|D < E[D]]}{E[w_0^{bin}(D)D|D > E[D]] - E[w_0^{bin}(D)D|D < E[D]]}$$

Below-mean dose control group

Scales by a dose difference (producing a slope)

Where:

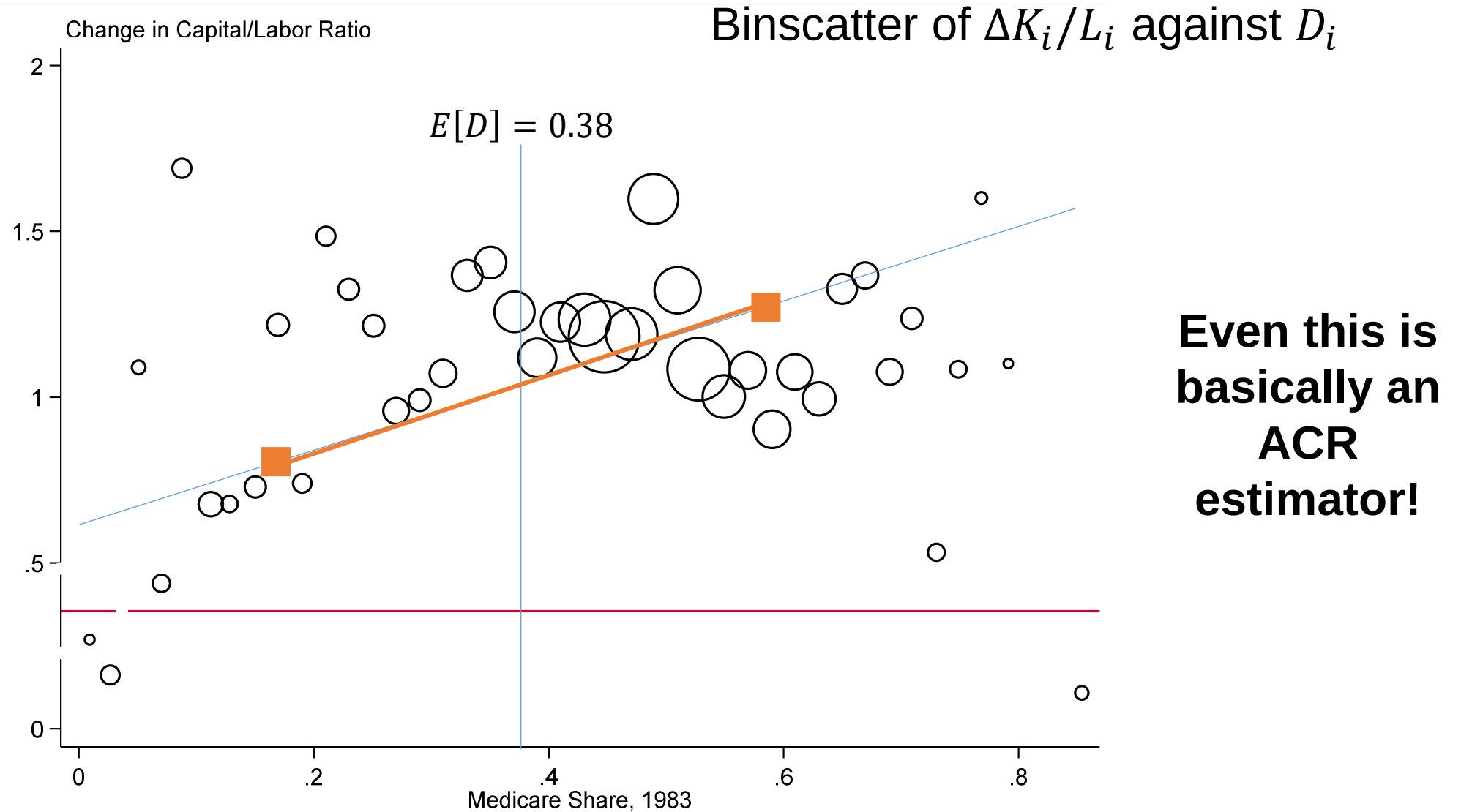
$$w_1^{bin}(d) \equiv \frac{|d - E[D]|}{E[|D - E[D]| | D > E[D]]}$$

Weights by dose size

$$w_0^{bin}(d) \equiv \frac{|d - E[D]|}{E[|D - E[D]| | D < E[D]]}$$

And $E[w_1^{bin}(d)|D > E[D]] = E[w_0^{bin}(d)|D < E[D]]$

Negative weights in ATT interpretation



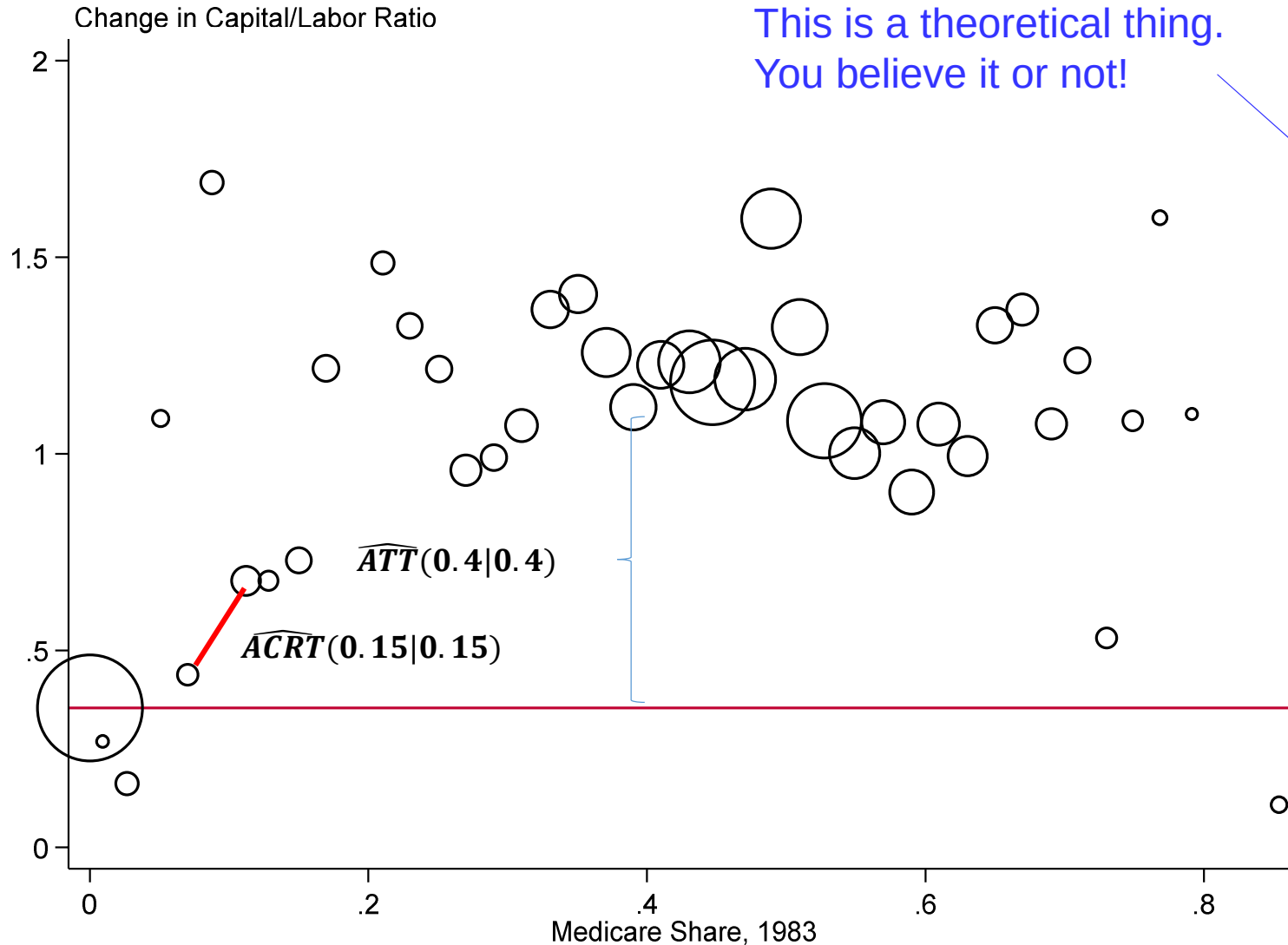
Non-Parametric Estimation

(No way around the assumptions, but good methods for estimating functions and aggregating them)

A quick recap

This is just your data. Can't do anything about that!

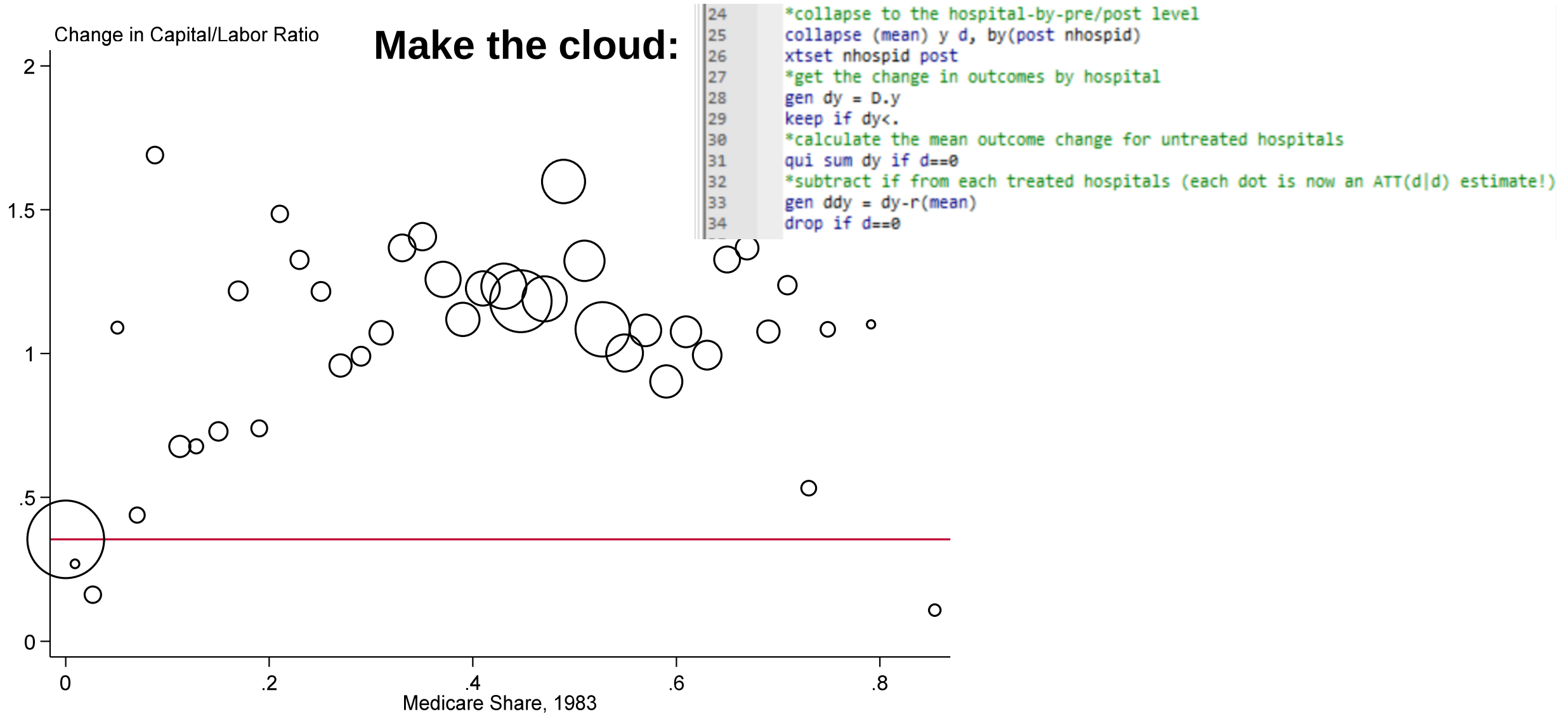
This is a theoretical thing.
You believe it or not!



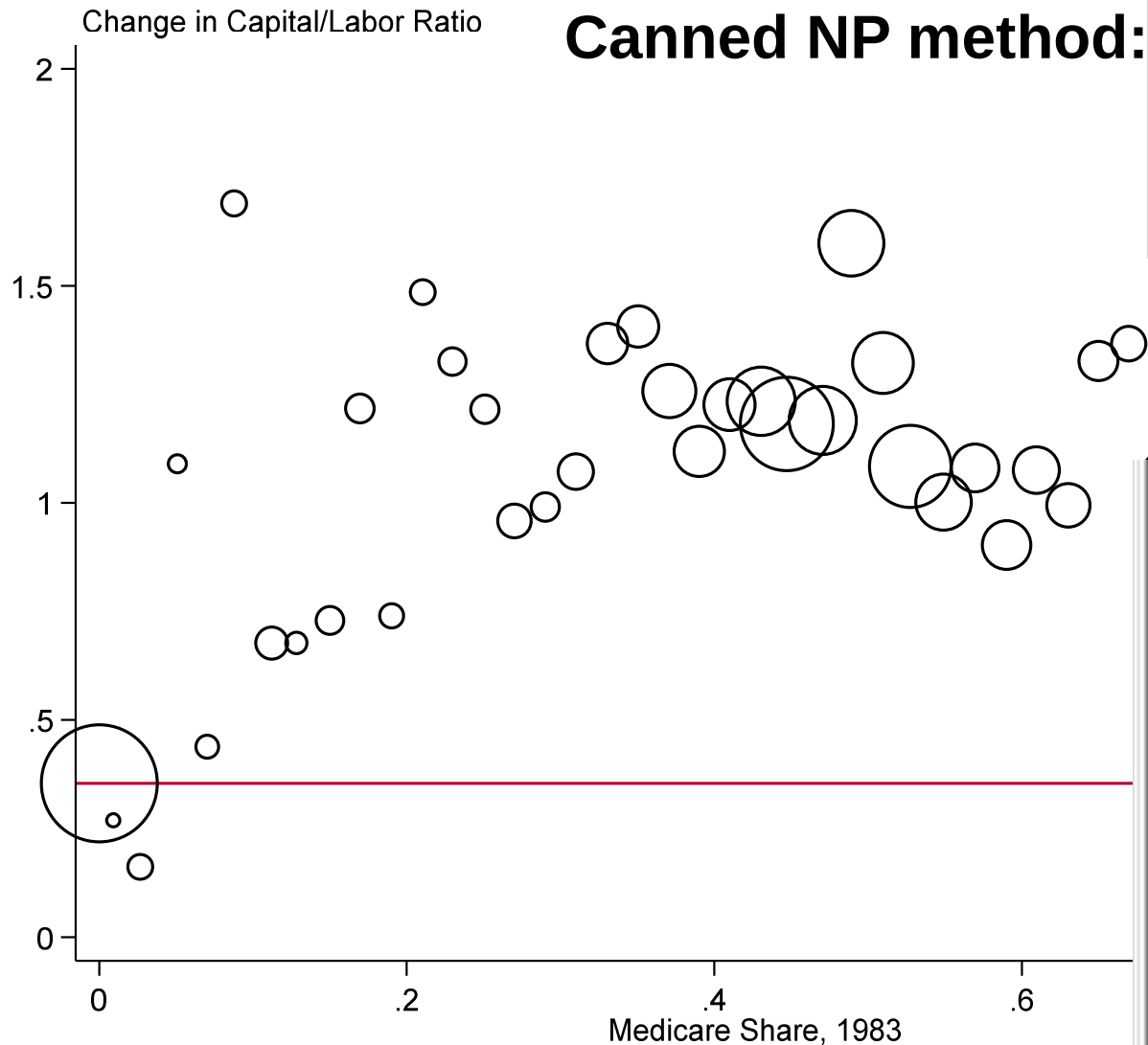
1. **Lots of raw info:** this scatter plot represents many 2x2 ATT(d|d) estimators and ACRT(d) estimators
2. **Causal identification:** PT $\rightarrow$ ATT(d|d), SPT $\rightarrow$ ACR(d)
3. **Estimation failure:** OLS weights behave strangely, no matter how you look at it.

OK, so summarizing the cloud with the best-fitting line isn't great, what to do instead?

Summarize the cloud nonparametrically



Summarize the cloud nonparametrically

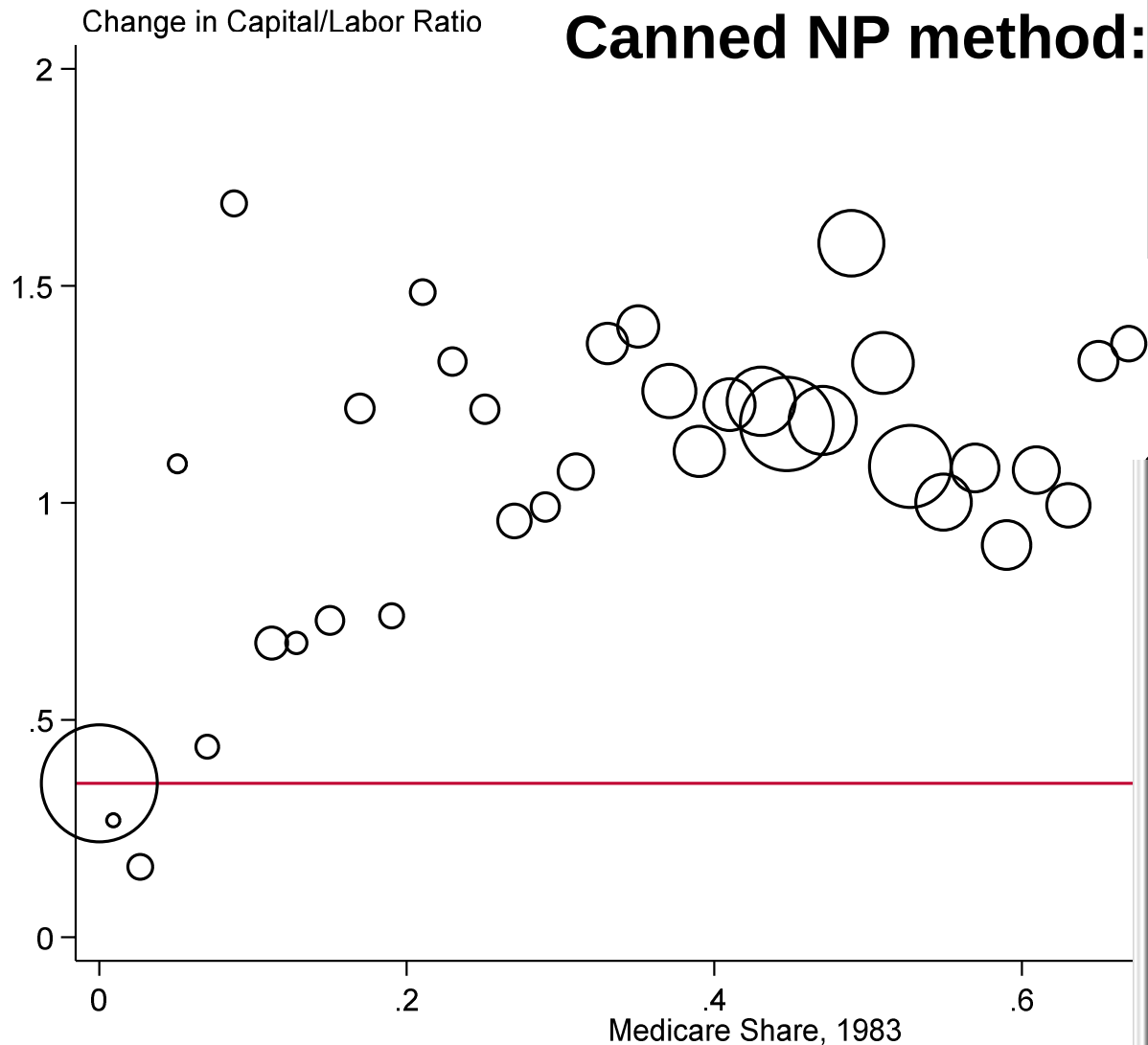


```
36 *fit non-parametric regressions through the resulting dots
37 npregress kernel ddy d
38 predict attd_kernel
39 npregress series ddy d, vce(robust) knots(1) bspline
40 predict attd_bspline
41 npregress series ddy d, vce(robust) polynomial(3)
42 predict attd_poly
```

Making the splines by hand (boring and error prone):

```
44 *do it by hand
45 cap drop dspline
46 mkspline dspline = d , cubic nk(5)
47 reg ddy dspline*
48 predict attd_byhand, xb
49
50 *do it really by hand--these don't quite get the
51 *points to line up which is an important and deeply boring feature of splines
52 egen bins = cut(d), group(2)
53 xi i.bins, pref(k) noomit
54 foreach var of varlist K*{
55     cap drop rd
56     qui sum d if `var'
57     gen rd = (d-r(min))/(r(max)-r(min))
58     gen `var'd = `var'*rd
59     gen `var'd2 = `var'*rd^2
60     gen `var'd3 = `var'*rd^3
61 }
62 reg ddy K*
63 predict attd_reallybyhand, xb
```

Summarize the cloud nonparametrically

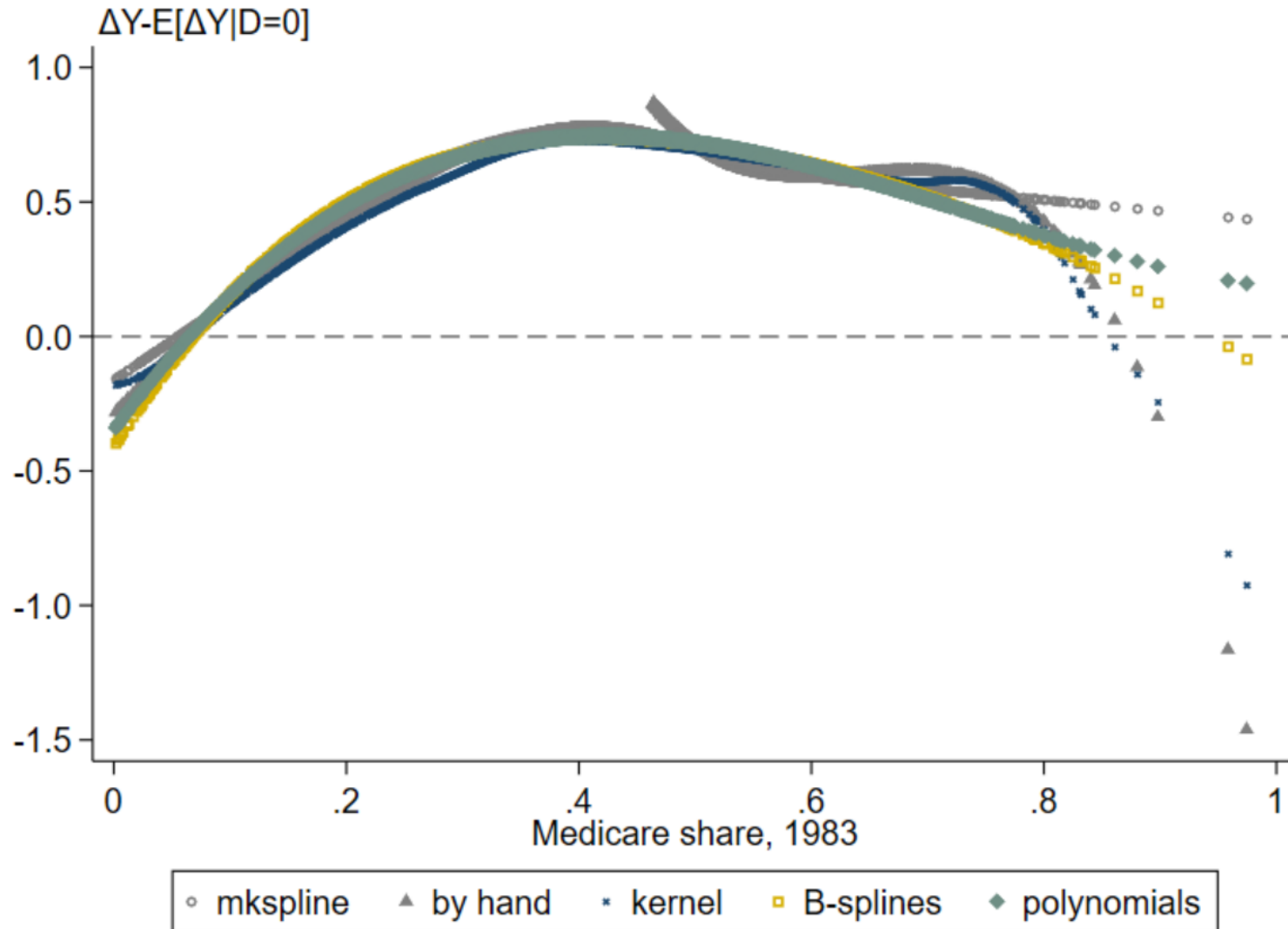


```
36 *fit non-parametric regressions through the resulting dots
37 npregress kernel ddy d
38 predict attd_kernel
39 npregress series ddy d, vce(robust) knots(1) bspline
40 predict attd_bspline
41 npregress series ddy d, vce(robust) polynomial(3)
42 predict attd_poly
```

Making the splines by hand (boring and error prone):

```
44 *do it by hand
45 cap drop dspline
46 mkspline dspline = d , cubic nk(5)
47 reg ddy dspline*
48 predict attd_byhand, xb
49
50 *do it really by hand--these don't quite get the
51 *points to line up which is an important and deeply boring feature of splines
52 egen bins = cut(d), group(2)
53 xi i.bins, pref(k) noomit
54 foreach var of varlist K*{
55     cap drop rd
56     qui sum d if `var'
57     gen rd = (d-r(min))/(r(max)-r(min))
58     gen `var'd = `var'*rd
59     gen `var'd2 = `var'*rd^2
60     gen `var'd3 = `var'*rd^3
61 }
62 reg ddy K*
63 predict attd_reallybyhand, xb
```

Summarize the cloud nonparametrically

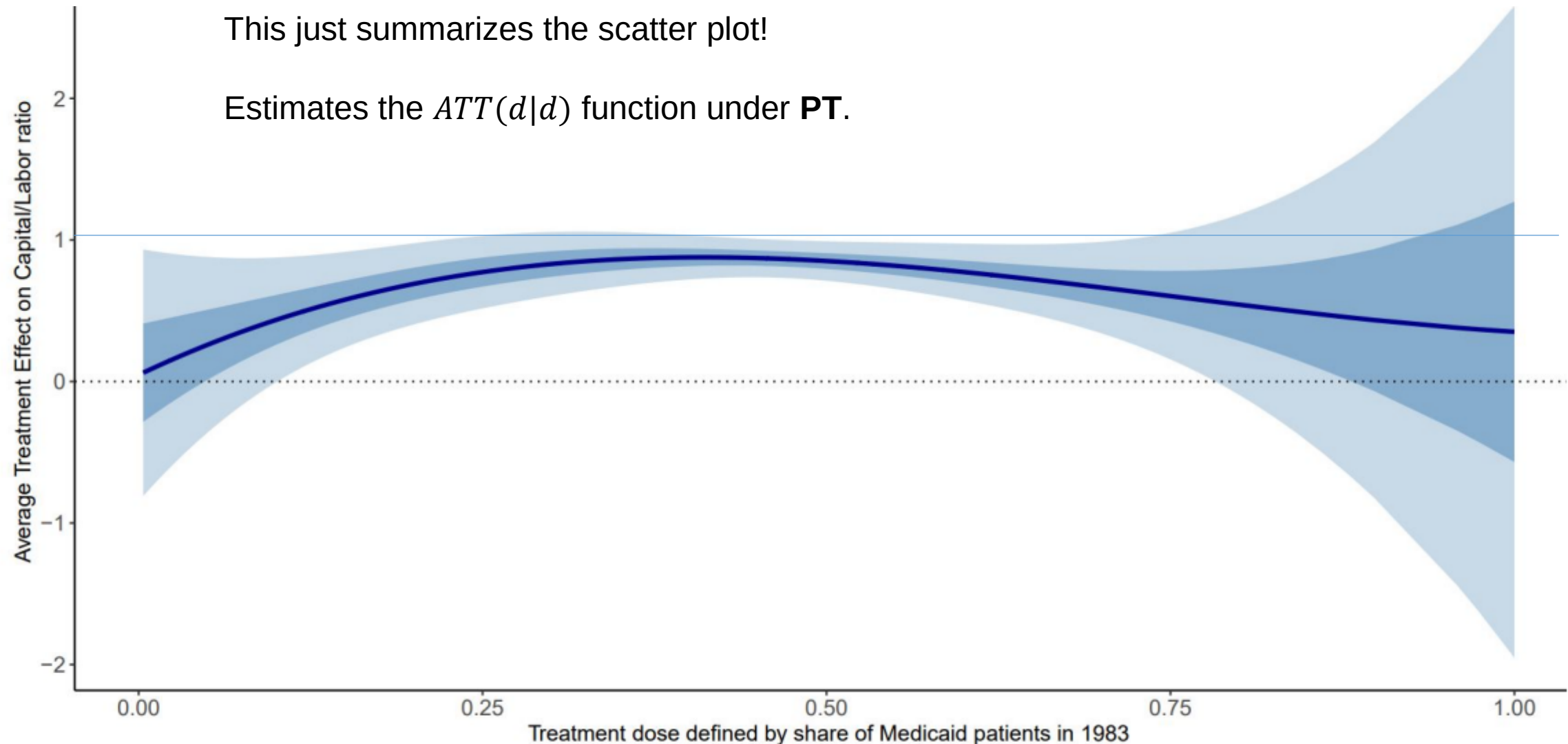


They all show the same thing (hump-shaped relationship) with some noise at the endpoint

One issue with NP methods is picking parameters (how many knots? What polynomial order? What kernel?)

So...

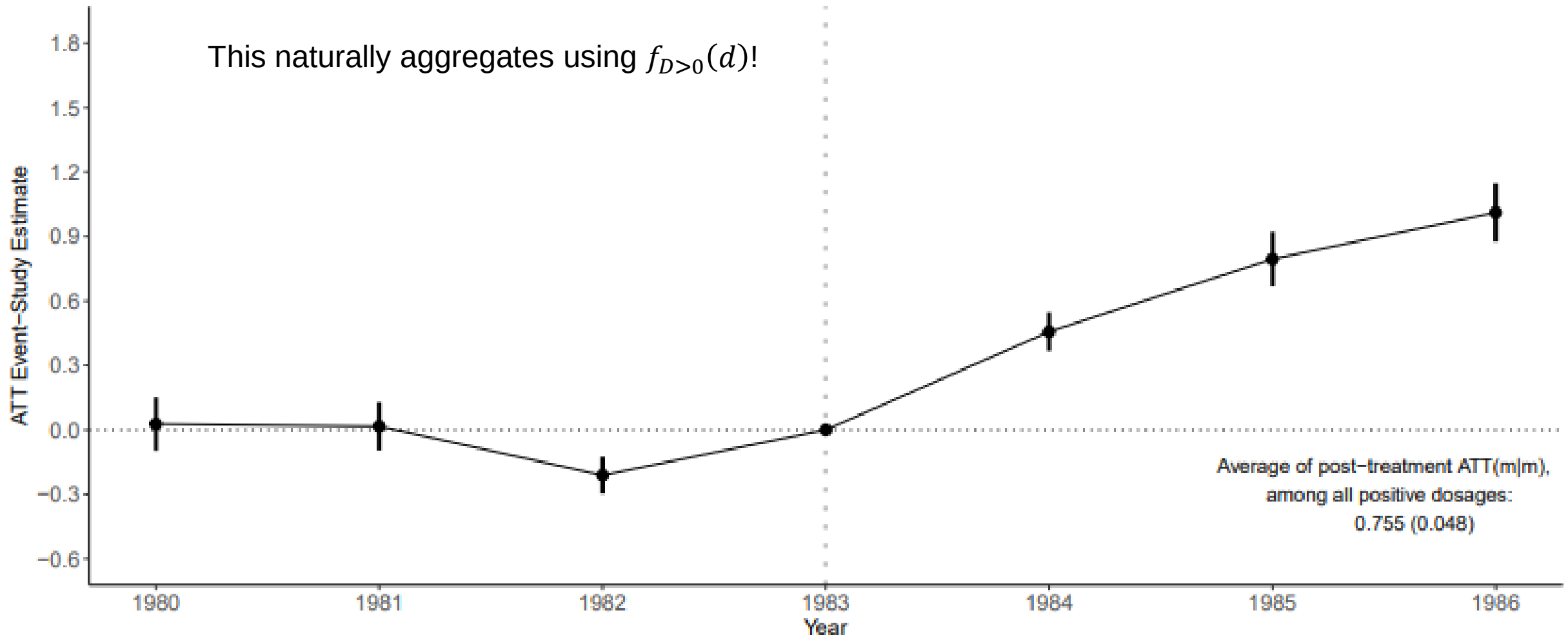
Non-Parametric Estimates of $ATT(d|d)$ for Medicare PPS



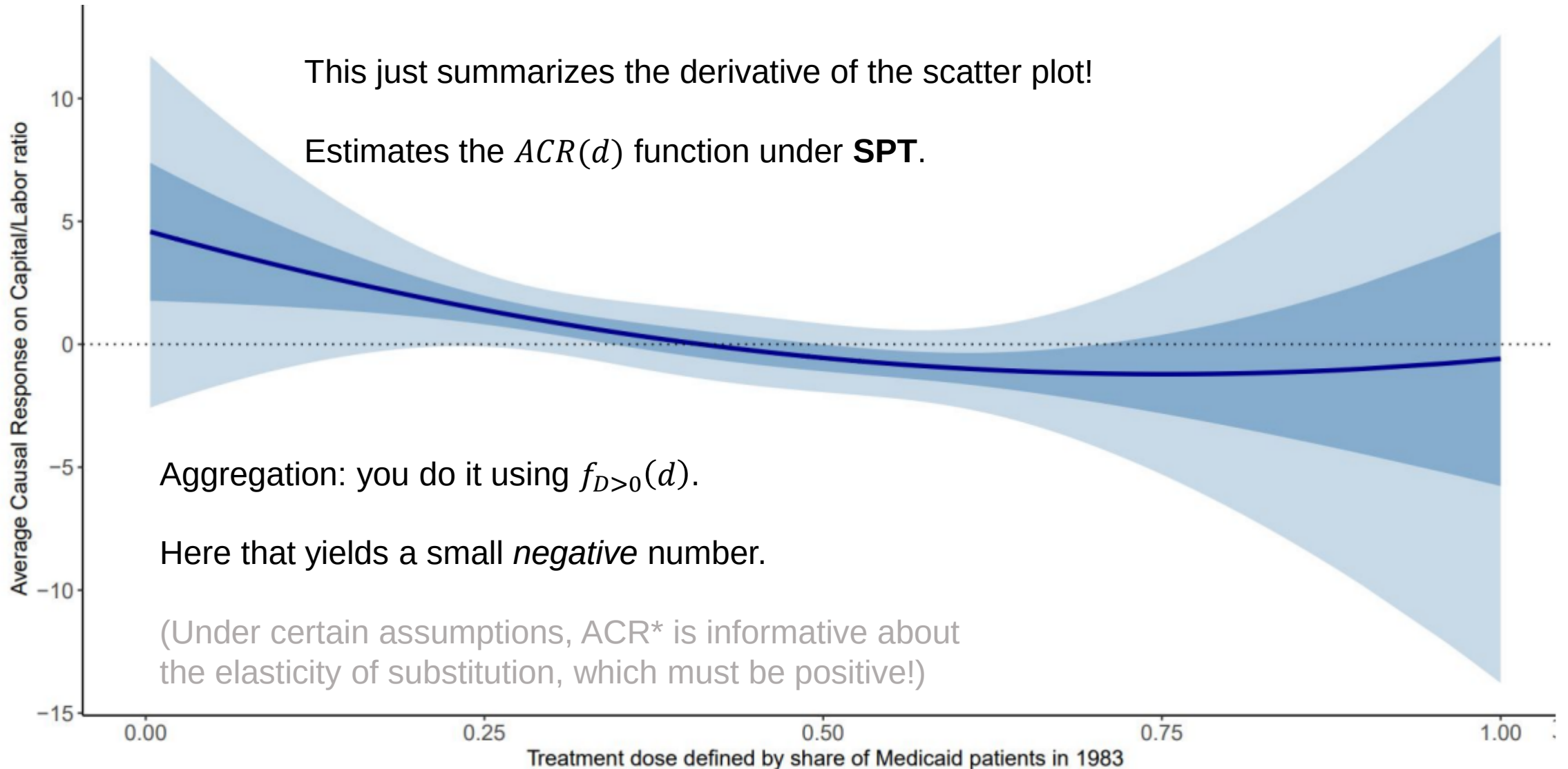
Aggregating: Event-Study Estimates of ATT_t^* for Medicare PPS

Use a binary variable $1\{D_i > 0\}$ and do a 2XT event-study:

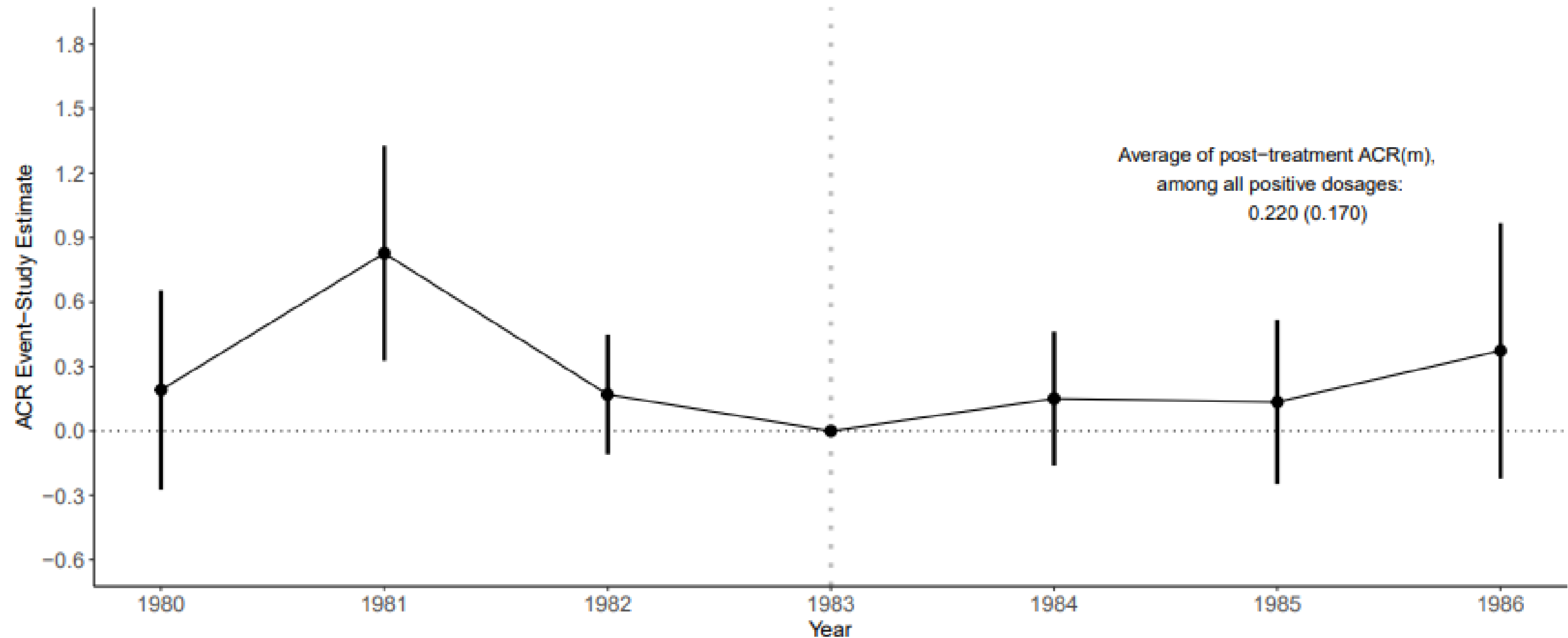
This naturally aggregates using $f_{D>0}(d)$!



Non-Parametric Estimates of $ACR(d)$ for Medicare PPS



Aggregating: Event-Study Estimates of ACR_t^* for Medicare PPS



What can be a “treatment”?

To begin, consider the profit maximization problem for a hospital with Medicare inpatient share M . We follow AF and assume a production function, $F_t(L, K)$, that is homothetic in labor (L), and capital (K). Market wages and rental rates are normalized by the output price, and Medicare subsidies mean that net input prices are $(1 - s_{L,t}M)w$ and $(1 - s_{K,t}M)r$. Firms consider the following profit maximization problem:

$$\max_{L,K} F_t(L, K) - (1 - s_{L,t}M)wL - (1 - s_{K,t}M)rK.$$

The solution to this problem generates factor demands and a capital-labor ratio that is only a function of the input price ratio, $k_t^* \left(\frac{(1-s_{L,t}M)w}{(1-s_{K,t}M)r} \right)$. We write the subsidy ratio, $\frac{(1-s_{L,t}M)}{(1-s_{K,t}M)}$ as $1 + S_t(M) = 1 + \frac{(s_{K,t}-s_{L,t})M}{1-s_{K,t}M}$. This reflects the fact that hospitals with no Medicare patients ($M = 0$), and all hospitals before PPS (when $s_{K,t=1} = s_{L,t=1} = s$) face no relative price distortion. PPS set $s_{L,t} = 0$ in 1983, making $S_{t=2}(M) = \frac{s_{K,t=2}M}{1-s_{K,t=2}M}$.

This structure allows us to define the capital-labor ratio potential outcomes in terms of Medicare inpatient share M :

$$\begin{aligned} Y_{t=1} &= Y_{t=1}(0) = k_{t=1}^* \left(\frac{w}{r} \right) \\ Y_{t=2} &= Y_{t=2}(M) = k_{t=2}^* \left((1 + S_{t=2}(M)) \frac{w}{r} \right) \end{aligned}$$

M is fixed:
“Characteristic”
not a
“treatment”
(Holland 1986)

Subsidy ratio is
manipulable
(sensible as a
“treatment”)

Causal effects of *subsidies*
not M per se

Shorthand in terms of M

Policy maps
 $M \rightarrow$ prices

Staggered *and* Continuous Treatments

What does fracking do for local economies?

“Shale plays”: geologic formations with natural gas (maybe) embedded in the rock

Prospectivity/suitability: ease with which that gas can be extracted.

Shale play (1)	Play first frac year (3)
Woodford-Anadarko	2008
Marcellus	2008
Woodford-Ardmore	2007
Fayetteville	2005
Woodford-Arkoma	2006
Barnett	2001
Permian, all plays	2005
Haynesville	2008
Bakken	2007

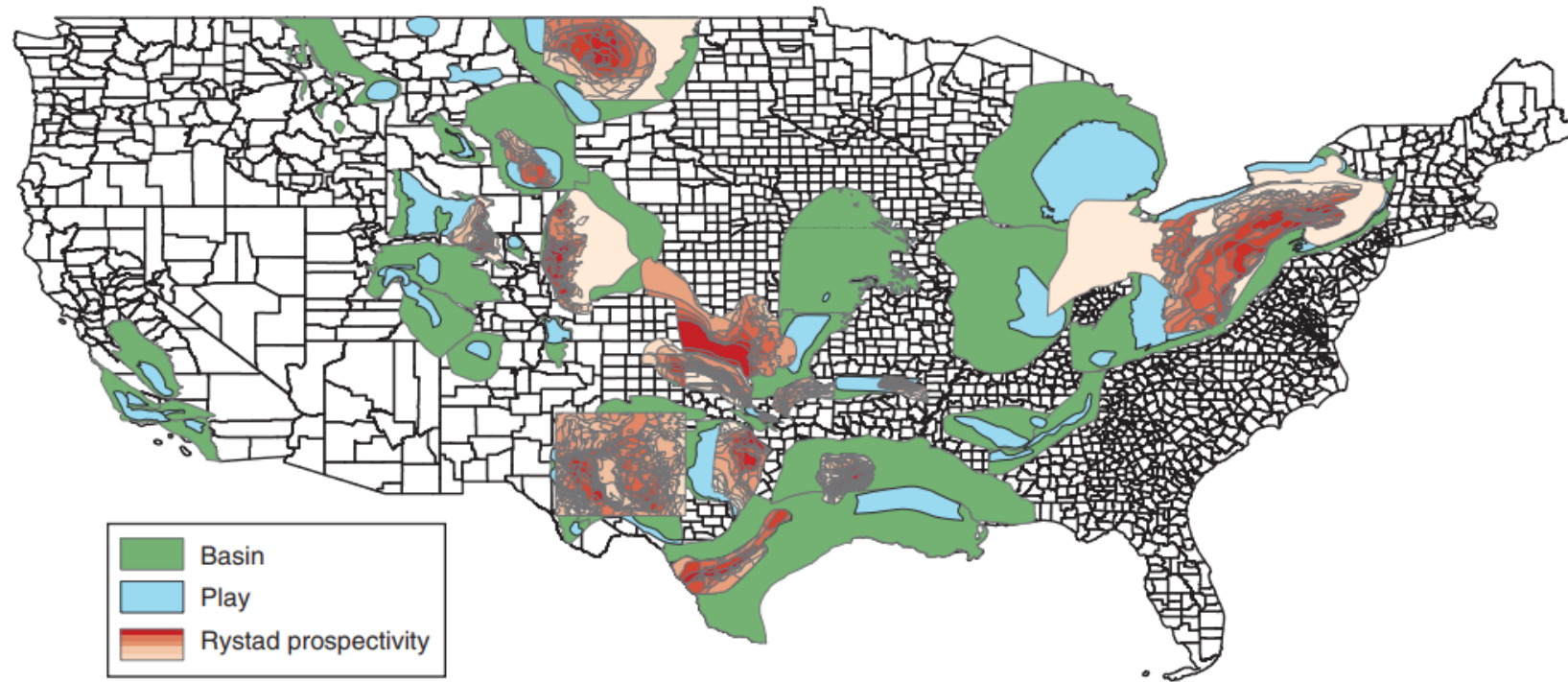


FIGURE 2. SHALE BASINS, PLAYS, AND PROSPECTIVITY SCORES

Staggered/Continuous DiD and choices

Continuous dose D_i begins at time G_i (multiple periods $T > 2$):

a huge number of parameters!

$$E[Y_t(g, d) - Y_t(\infty, 0) | G = g, D = d] \equiv ATT(g, d, t | g, d)$$

“The average effect in period t of having “ d ” suitability starting in period g versus no-fracking”

$$\left. \frac{\partial E[Y_t(g, l) | G = g, D = d]}{\partial l} \right|_{l=d} \equiv ACRT(g, d, t | g, d)$$

“The average effect in period t of ‘ d ’ suitability vs ‘ $d-1$ ’ starting in period g ?”

A functional parameter
for every G
and every t !

Estimation choices

Brute force approach:

Non-parametrically estimate $ATT(g, d, t|g, d)$ (under PT) or $ATE(g, d, t|g)$ and $ACR(g, d, t|g)$ (under Strong PT) then average.

This is basically how binary staggered estimators work

Note 1: What comparison group to use to get $E[Y_t(\infty, 0)|G = g, D = d]$?

$D = 0$ and $G = g$? All $D = 0$? $D = d$ and $G > t$? All $G > t$?

Choice about PT and estimator

Note 2: Small groups.

NP estimators won't perform when n_g is small.

Note 3: Calculating the weights.

Requires objects like $P(G = g|D = d)$ which may not be well-behaved (0 or 1) or require additional NP estimation.

Practical limits on the choices you can make

Aggregation choices

Brute force output: $\widehat{ATT}(g, d, t|g, d)$, $\widehat{ATE}(g, d, t|g)$ or $\widehat{ACR}(g, d, t|g)$

There is no “right” aggregation; you choose:

Dynamic Q: “How do the average effects of fracking evolve over time? (Does the boom persist?)”

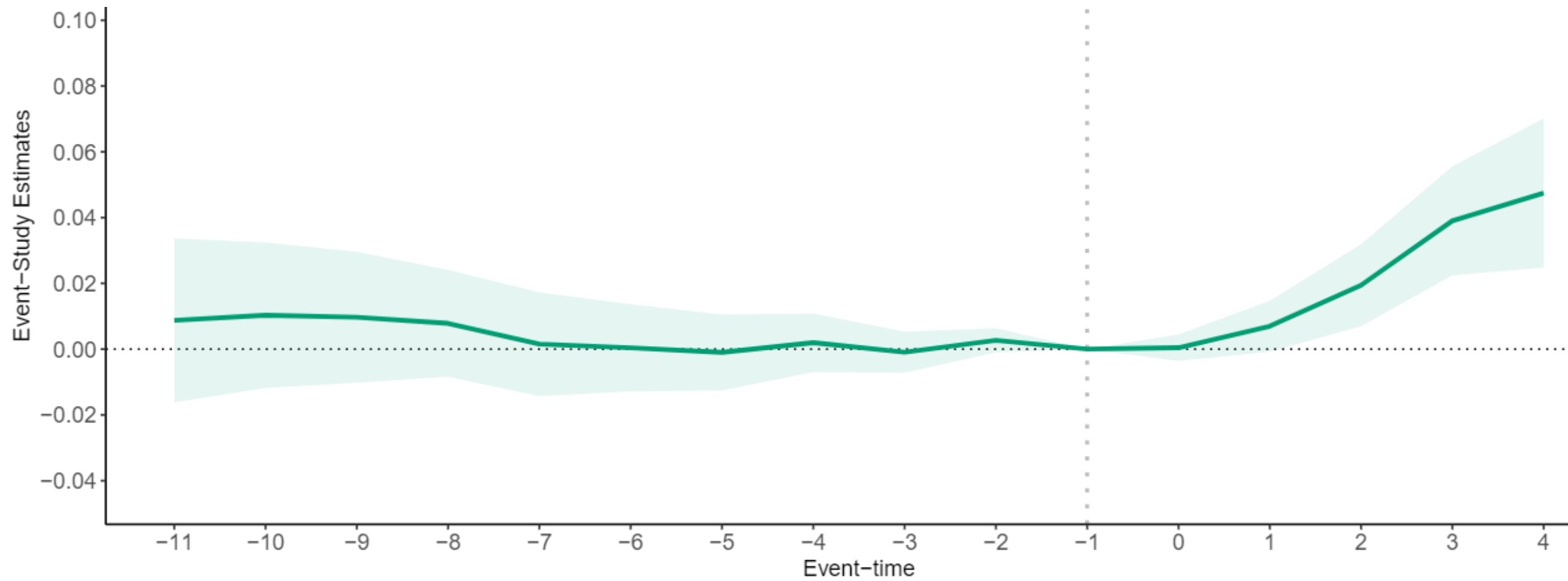
Balanced dynamic parameter(s): $ATT^*(e) \equiv \sum_g P(G = g|D > 0) ATT^*(g, g + e|g)$

2. Then average across g (note event-time balance and weights)

1. Average across dose distribution for each g :

$$ATT^*(g, g + e|g) \equiv E[ATT(g, d, g + e|g, d)|G = g]$$

$$\widehat{ATT}^*(e)$$



Aggregation choices

Brute force output: $\widehat{ATT}(g, d, t|g, d)$, $\widehat{ATE}(g, d, t|g)$ or $\widehat{ACR}(g, d, t|g)$

There is no “right” aggregation; you choose:

Dynamic Q: “How do the average effects of fracking evolve over time? (Does the boom persist?)”

Balanced dynamic parameter(s): $ATT^*(e) \equiv \sum_g P(G = g|D > 0)ATT^*(g, g + e|g)$

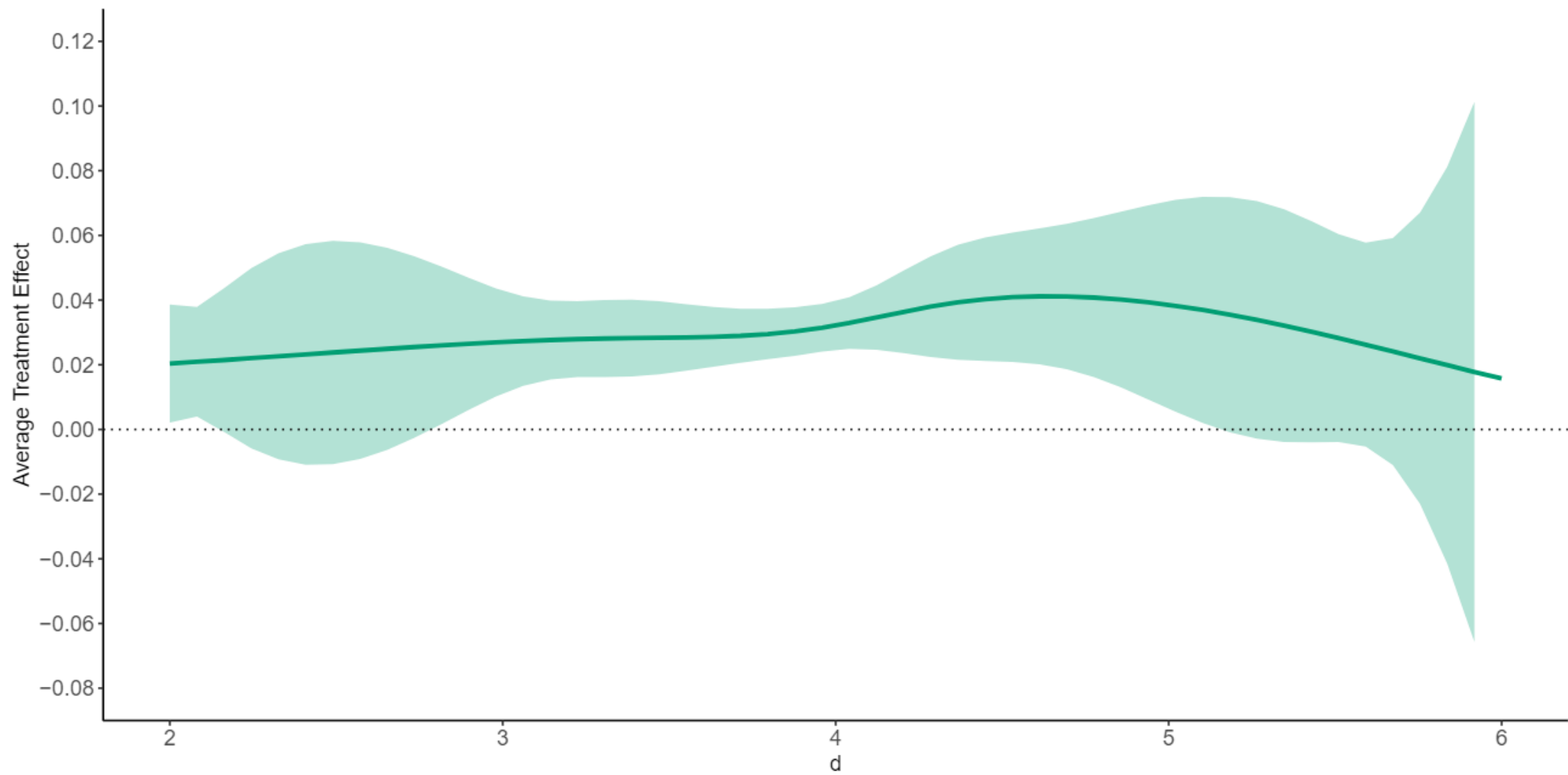
Dose-response Q: “Does fracking matter most where the best drilling happens?”

Dose-response parameter(s): $ATT^*(d|d) \equiv \sum_g P(G = g|D > 0)ATT^*(g, d |g, d)$

2. Then average across g

1. Average over “some” post-periods for each g : $ATT^*(g, d |g, d) \equiv \sum_{t=g}^T \frac{ATT(g, d, t|g, d)}{T-(g-1)}$

$$\widehat{ATT}^*(d|d)$$



Intermediate aggregation choices

Dynamic dose-response Q: “How do the average effects of fracking evolve over time for ‘high’ and ‘low’ suitability areas?”

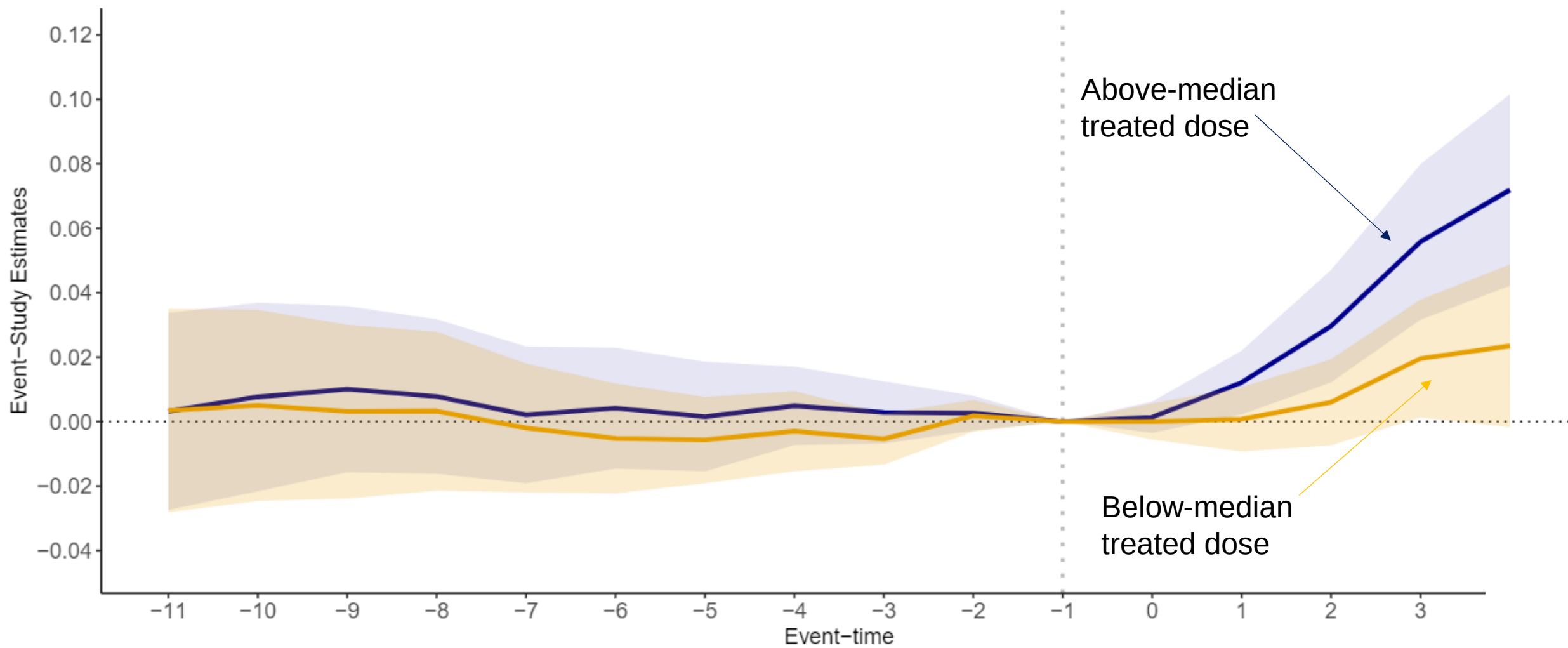
Full dynamic/partial dose-response parameter(s):

$$ATT^*(e, h) \equiv \sum_g P(G = g | D \in h) E[ATT(g, d, g + e | g, d) | G = g, D > h]$$

$$ATT^*(e, l) \equiv \sum_g P(G = g | D \in l) E[ATT(g, d, g + e | g, d) | G = g, D \leq h]$$

Just aggregate over *part* of the dose distribution

$$\widehat{ATT}^*(h, e) \text{ and } \widehat{ATT}^*(l, e)$$



Intermediate aggregation choices

Dynamic dose-response Q: “Does fracking matter most where the best drilling happens in the shorter-run vs the longer-run?”

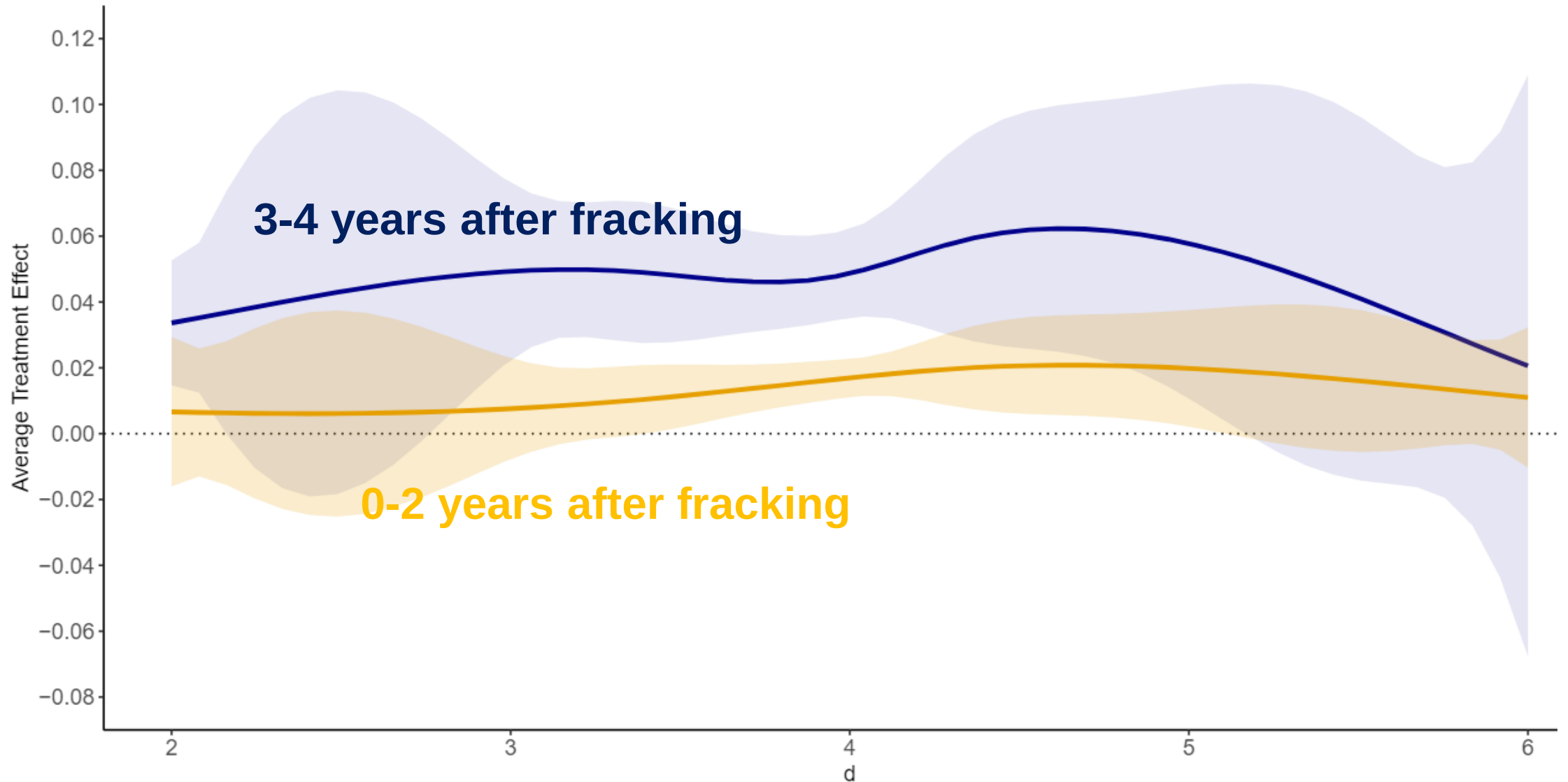
Partially dynamic/full dose-response parameter(s):

$$ATT^*(d, SR|d) \equiv \sum_g P(G = g|D > 0) \left[\sum_{t=g}^{g+SR} \frac{ATT(g, d, t|g, d)}{\#\{SR\}} \right]$$

$$ATT^*(d, LR|d) \equiv \sum_g P(G = g|D > 0) \left[\sum_{t=g+SR+1}^{g+LR} \frac{ATT(g, d, t|g, d)}{\#\{LR\}} \right]$$

Average over *parts* of the post-period (SR and LR)

$$\widehat{ATT}^*(d, SR|d) \text{ and } \widehat{ATT}^*(d, LR|d)$$



Aggregation **choices** affect estimation **choices**

$$\begin{aligned} ATT^*(h, e) &\equiv \sum_g P(G = g | D > h) E[ATT(g, d, g + e | g, d) | G = g, D > h] \\ &= \sum_g P(G = g | D > h) \int_{x=h}^{\infty} E[Y_{g+e} - Y_{g-1} | G = g, D = x] - E[Y_{g+e} - Y_{g-1} | G \in \mathcal{G}^{comp}, D \in \mathcal{D}^{comp}] f_{D>h|g}(x) dx \\ &= \sum_g P(G = g | D > h) \left[\int_{x=h}^{\infty} E[Y_{g+e} - Y_{g-1} | G = g, D = x] f_{D>h|g}(x) dx - E[Y_{g+e} - Y_{g-1} | G \in \mathcal{G}^{comp}, D \in \mathcal{D}^{comp}] \right] \\ &= \sum_g P(G = g | D > h) \left[E[Y_{g+e} - Y_{g-1} | G = g, D > h] - E[Y_{g+e} - Y_{g-1} | G \in \mathcal{G}^{comp}, D \in \mathcal{D}^{comp}] \right] \end{aligned}$$

This is a binary DiD estimand for group g using $D > h$ as the treatment group.

Can estimate with any binary DiD tool; no need for brute force NP estimation!

Aggregation **choices** affect estimation **choices**

$$ATT^*(d, E|d)$$

$$\equiv \sum_g P(G = g|D > 0) \left[\sum_{t=g}^{g+E} \frac{E[Y_t - Y_{g-1}|G = g, D = d] - E[Y_t - Y_{g-1}|G \in \mathcal{G}^{comp}, D \in \mathcal{D}^{comp}]}{E + 1} \right]$$

$$\equiv \sum_g P(G = g|D > 0) \left[E[\bar{Y}_t^E - Y_{g-1}|G = g, D = d] - E[\bar{Y}_t^E - Y_{g-1}|G \in \mathcal{G}^{comp}, D \in \mathcal{D}^{comp}] \right]$$

This is a single ATT(d|d) estimand for outcomes averaged between g and $g + E$. Can estimate with a single NP procedure (CGBS 2024)!

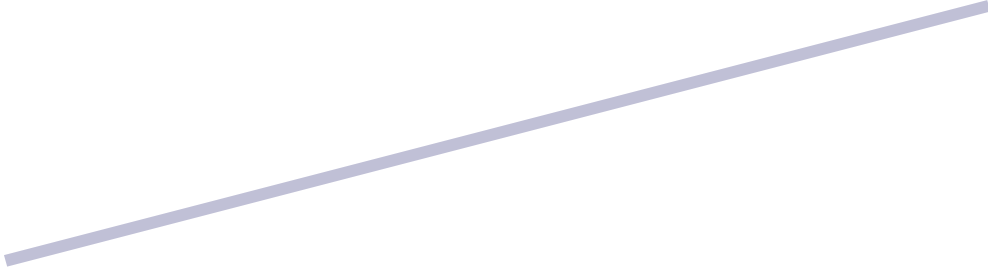
The role of choices in complex DiD designs

The design largely defines the “building blocks”, perhaps many.

The building blocks largely define the assumptions, which yield estimators.

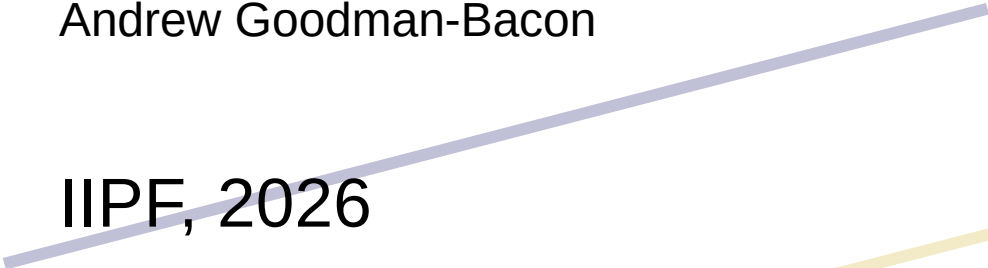
Researchers *questions* drive aggregation **choices**.

Those aggregation **choices** can feedback and simplify estimation **choices**.

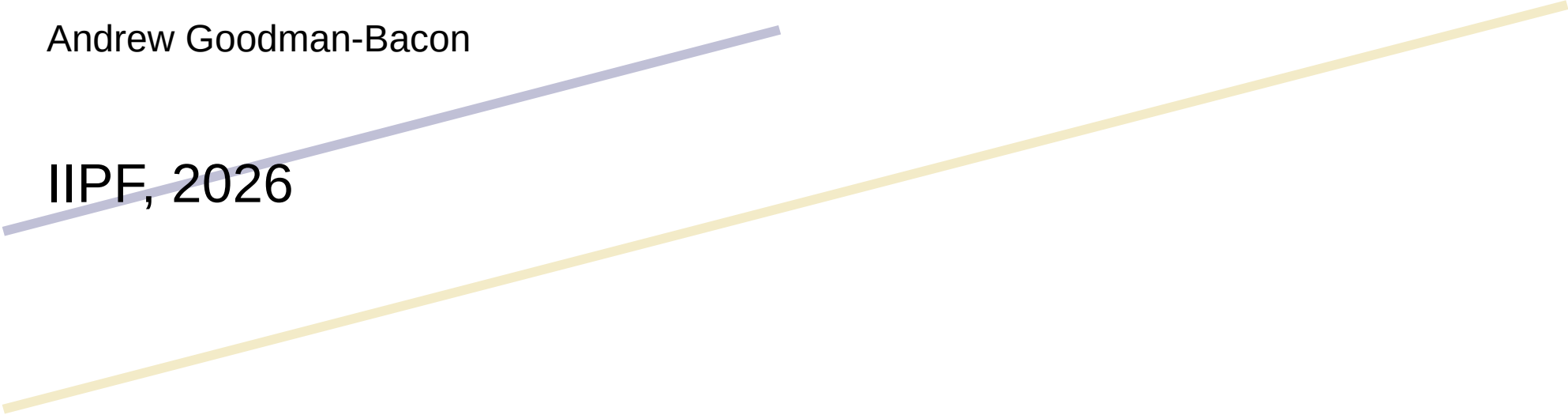


New Developments in Difference-in-Differences: *Extras* + *Extensions*

Andrew Goodman-Bacon



IIPF, 2026



Today: a smorgasbord of additional DiD details

1. Starting from a non-zero dose (dcDH et al 2025)

Starting from a non-zero dose

Set up: $t = \{1, 2\}$,
treatment: $(D_{i1}, D_{i2}) = (d_1, d_2)$, neither of which is zero

Target parameter:

$$ACR(d_1, d_2 | d_1, d_2) \equiv E[Y_{i2}(d_2) - Y_{i2}(d_1) | D_{i1} = d_1, D_{i2} = d_2]$$

Identifying assumption:

$$\forall d_1 \quad E[\Delta Y(d_1) | D_{i1} = d_1, D_{i2}] = E[\Delta Y(d_1) | D_{i1} = d_1]$$

What does PT in a treated PO mean?

Equal change in “ d_1 ” outcomes for changers and switchers who start at d_1 :

$$\begin{aligned} & E[Y_2(d_1) - Y_1(d_1) | D_{i1} = d_1, D_{i2} = d_2] \\ &= E[Y_2(d_1) - Y_1(d_1) | D_{i1} = d_1, D_{i2} = d_1] \end{aligned}$$

What does PT in a treated PO mean?

Equal change in “ d_1 ” outcomes for changers and switchers who start at d_1 :

Level effect of d_1 in period 2

Level effect of d_1 in period 2

$$E[Y_2(d_1) - Y_2(0) + Y_2(0) - Y_1(d_1) + Y_1(0) - Y_1(0) | D_{i1} = d_1, D_{i2} = d_2]$$

$$= E[Y_2(d_1) - Y_2(0) + Y_2(0) - Y_1(d_1) + Y_1(0) - Y_1(0) | D_{i1} = d_1, D_{i2} = d_1]$$

Level effect of d_1 in period 2

Level effect of d_1 in period 2

What does PT in a treated PO mean?

Equal change in “ d_1 ” outcomes for changers and switchers who start at d_1 :

Level effect of d_1 in period 2

Level effect of d_1 in period 2

$$E[Y_2(d_1) - Y_2(0) + Y_2(0) - Y_1(d_1) + Y_1(0) - Y_1(0) | D_{i1} = d_1, D_{i2} = d_2]$$

$$= E[Y_2(d_1) - Y_2(0) + Y_2(0) - Y_1(d_1) + Y_1(0) - Y_1(0) | D_{i1} = d_1, D_{i2} = d_1]$$

Level effect of d_1 in period 2

Level effect of d_1 in period 2

Traditional parallel trends!

What does PT in a treated PO mean?

Equal change in “ d_1 ” outcomes for changers and switchers who start at d_1 :

$$\begin{aligned} & E[Y_2(d_1) - Y_1(d_1) | D_{i1} = d_1, D_{i2} = d_2] \\ &= E[Y_2(d_1) - Y_1(d_1) | D_{i1} = d_1, D_{i2} = d_1] \end{aligned}$$

Therefore equivalent to:

Dynamic effects of d_1 for the (d_1, d_2) group

Dynamic effects of d_1 for the (d_1, d_1) group

$$\begin{aligned} & \Delta ATT(d_1 | d_1, d_2) - \Delta ATT(d_1 | d_1, d_1) \\ &= \underbrace{E[Y_2(0) - Y_1(0) | D_{i1} = d_1, D_{i2} = d_2] - E[Y_2(0) - Y_1(0) | D_{i1} = d_1, D_{i2} = d_1]}_{\text{Parallel trends!}} \end{aligned}$$

What will the estimand be?

Target parameter:

$$ACR(d_1, d_2 | d_1, d_2) \equiv E[Y_{i2}(d_2) - Y_{i2}(d_1) | D_{i1} = d_1, D_{i2} = d_2]$$

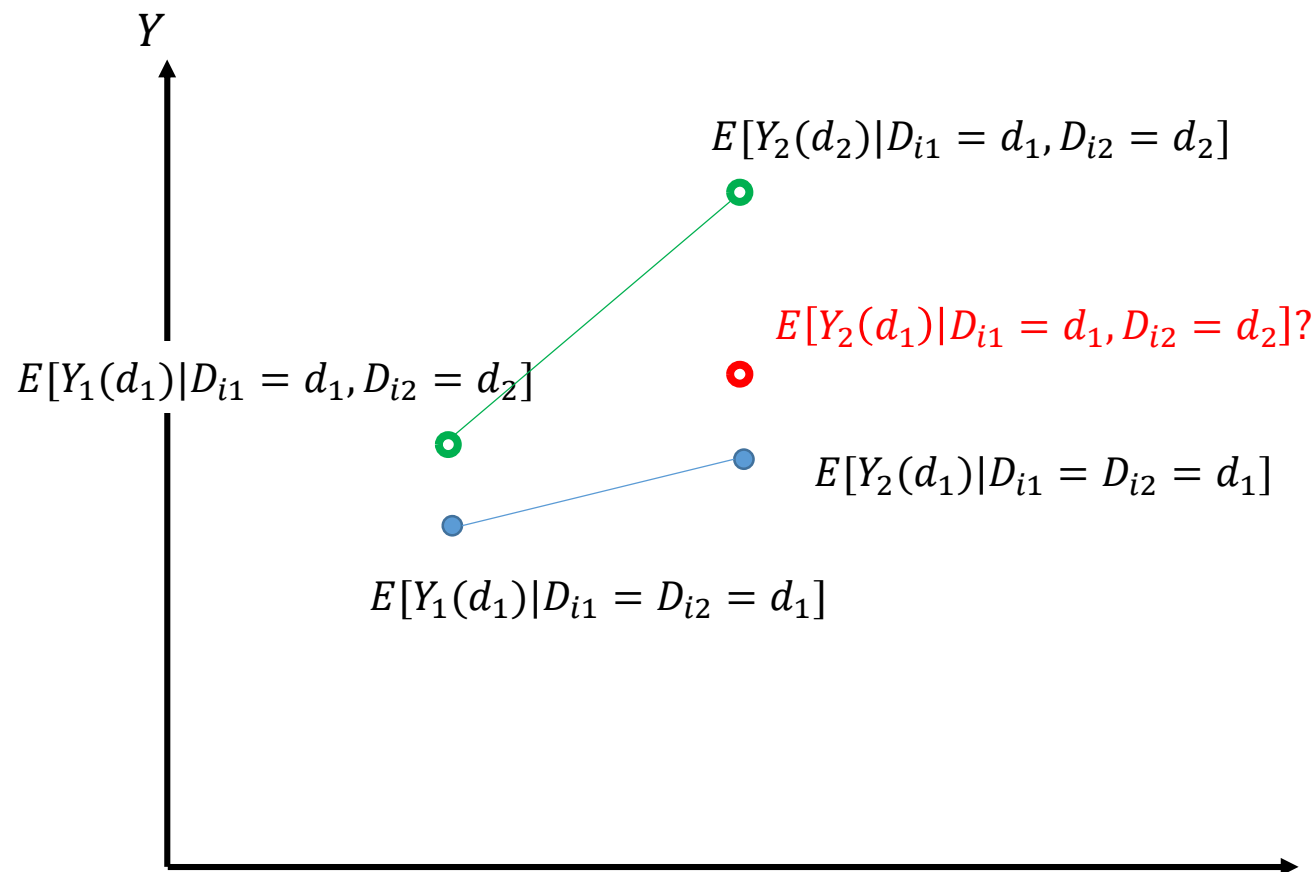
Identifying assumption:

$$\forall d_1 \quad E[\Delta Y(d_1) | D_{i1} = d_1, D_{i2}] = E[\Delta Y(d_1) | D_{i1} = d_1]$$

Estimand?

$$E[\Delta Y | D_{i1} = d_1, D_{i2} = d_2] - E[\Delta Y | D_{i1} = D_{i2} = d_1]$$

Illustration



What determines where $Y_2(d_1)$ would have been for treated units?

1. Untreated trends
2. Treatment effect *dynamics*

What is the distinction with SPT in the case with $D_{i1} = 0$?

Comparing to other stayers?

Set up: $(D_{i1}, D_{i2}) = (d_1, d_2)$, neither of which is zero

Target parameter:

$$ACR(d_1, d_2 | d_1, d_2) \equiv E[Y_{i2}(d_2) - Y_{i2}(d_1) | D_{i1} = d_1, D_{i2} = d_2]$$

Identifying assumption(s):

$$\forall d_1 \quad E[\Delta Y(d_1) | D_{i1} = d_1] = E[\Delta Y(d_1)]$$

$$\forall d_1, d_2 \quad E[\Delta Y(d_1)] = E[\Delta Y(d_2)]$$

What do these PT assumptions mean?

Identifying assumption(s):

Everyone's PO trend at d_1 is the same
(but we don't see anyone but the d_1 stayers at d_1)

$$\forall d_1 \quad E[\Delta Y(d_1) | D_{i1} = d_1] = E[\Delta Y(d_1)]$$

$$\forall d_1, d_2 \quad E[\Delta Y(d_1)] = E[\Delta Y(d_2)]$$

So...we also need everyone's PO trend at every dose is the same.

“To compare a 2→5 switcher to a 3 stayer, $E[\Delta Y(2)]$ and $E[\Delta Y(3)]$ should be equal”

→ So all potential outcome trends must be the same!

Today: a smorgasbord of additional DiD details

1. Starting from a non-zero dose (dCDH et al 2025) (same dynamics across groups)
2. Complicated treatment paths (dcDH 2024)

Complicated treatment paths

Set up: $\mathbf{D}_i = (d_{i1}, d_{i2}, \dots, d_{iT}), Y_i(\mathbf{D}_i) = Y_i(d_{i1}, d_{i2}, \dots, d_{iT})$

(Robins 1986)

$$G_i = \min\{t: t \geq 2 \text{ s.t. } d_{it} \neq d_{it-1}\}$$

First treatment change

Target parameter 1:

$$ATT(\mathbf{D}_i, t) \equiv E[Y_{it}(\overbrace{d_{i1}, \dots, d_{it}}^{\text{Treatment path up to } t}) - Y_{it}(\overbrace{d_{i1}, \dots, d_{i1}}^{\text{Status quo treatment path up to } t}) | \mathbf{D}_i]$$

Identifying assumption: The group whose path really is D

For $\mathbf{D}_{i1} = \mathbf{D}'_{i1} = d_1$

$$E[\Delta Y(d_1, \dots, d_1) | \mathbf{D}_i] = E[\Delta Y(d_1, \dots, d_1) | \mathbf{D}'_i]$$

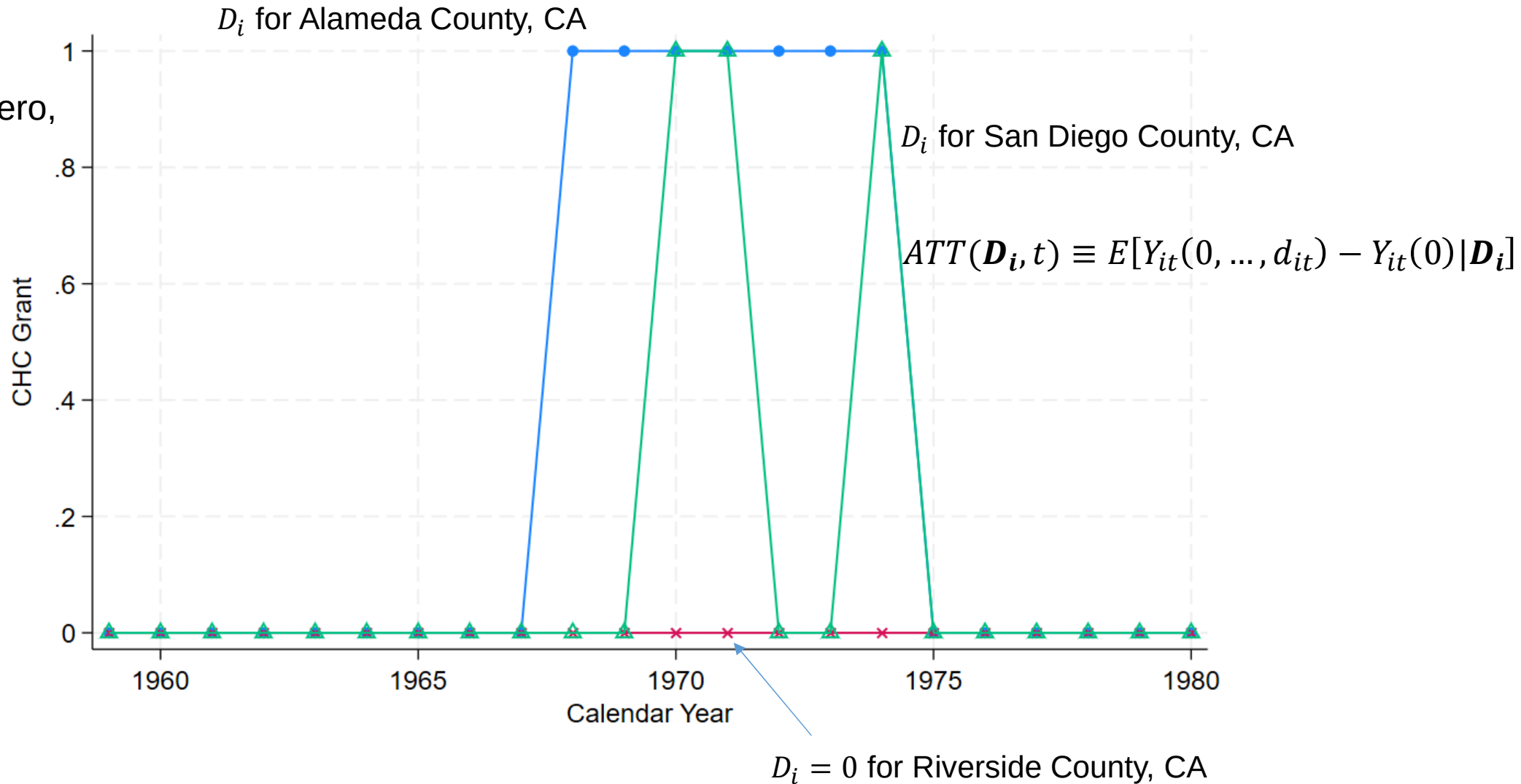
Notation break,

Bailey and Goodman-Bacon: grants for Community Health Centers

Baseline value is zero,

Different G_i s and
different paths

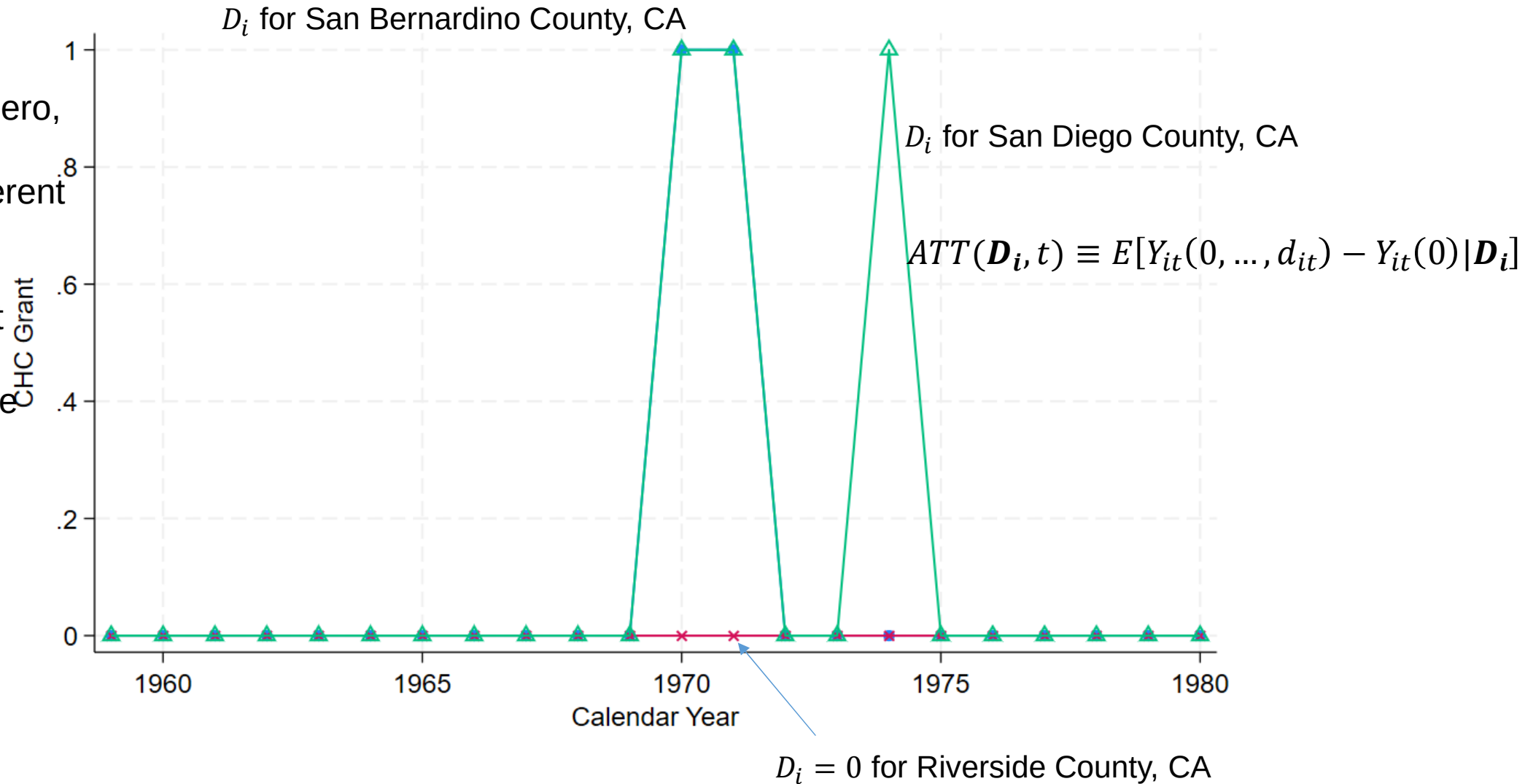
These are different
“groups”



Notation break,

Bailey and Goodman-Bacon: grants for Community Health Centers

Baseline value is zero,
Same G_i s and different
paths
These are different
“groups” despite
starting at the same
time



Complicated treatment paths

Target parameter 1:

$$ATT(\mathbf{D}_i, t) \equiv E[Y_{it}(d_{i1}, \dots, d_{it}) - Y_{it}(d_{i1}, \dots, d_{i1}) | \mathbf{D}_i]$$

Identifying assumption:

For $\mathbf{D}_{i1} = \mathbf{D}'_{i1} = d_1$

$$E[\Delta Y(d_1, \dots, d_1) | \mathbf{D}_i] = E[\Delta Y(d_1, \dots, d_1) | \mathbf{D}'_i]$$

Change in mean outcomes for units who start at d_1 and stay there until t

Estimand?

$$E[Y_{it} - Y_{ig-1} | \mathbf{D}_i] - E[Y_{it} - Y_{ig-1} | \mathbf{D}_i = (d_1, \dots, d_1)]$$

Change in mean outcomes for specific treatment path, starting at d_1

How it incorporates dynamic effects

Effect up to t sums lagged effects:

$$E[Y_{i3}(0,1,2) - Y_{i3}(0,0,0) | \mathbf{D}_i = (0,1,2)] =$$

$$E[Y_{i3}(0,1,2) - Y_{i3}(0,1,0) | \mathbf{D}_i = (0,1,2)] + E[Y_{i3}(0,1,0) - Y_{i3}(0,0,0) | \mathbf{D}_i = (0,1,2)]$$

The effect of the whole path	(0,1,2) vs (0,0,0)
is the effect of adding the most recent	
change conditional on the past path,	(0,1,2) vs (0,1,0)
plus the effect of the first change alone	(0,1,0) vs (0,0,0)

How it incorporates dynamic effects

Scaled Effect up to t averages lagged effects weighted by Δd :

$$\frac{ATT((0,1,2), 3)}{3} = \frac{E[Y_{i3}(0,1,2) - Y_{i3}(0,0,0)|\mathbf{D}_i = (0,1,2)]}{3} =$$
$$\frac{2}{3} \frac{E[Y_{i3}(0,1,2) - Y_{i3}(0,1,0)|\mathbf{D}_i = (0,1,2)]}{2} + \frac{1}{3} \frac{E[Y_{i3}(0,1,0) - Y_{i3}(0,0,0)|\mathbf{D}_i = (0,1,2)]}{1}$$

Dose size matters, too, for scaled parameters.

Aggregates by group: binarize and staggerize

Target parameter 2: Average effect at time t of starting with dose d_1 and first switching in time g

***kind of like an intention to treat*

Identifying assumption: same

This is exactly CS' conditional estimator with the period 1 dose as the covariate

Estimand?

Use timing of first change

$$E[Y_{it} - Y_{ig-1} | G_i = g, D_{i1} = d_1] - E[Y_{it} - Y_{ig-1} | G_i > t, D_{i1} = d_1]$$

Conditional on initial dose

Scaling the group-time-effects

Target parameter 3: An average slope type parameter. ATT per dose unit.

Identifying assumption: same

This is CS again, but using cumulative value of D as the outcome

Estimand?

$$\frac{E[Y_{it} - Y_{ig-1} | G_i = g, D_{i1} = d_1] - E[Y_{it} - Y_{ig-1} | G_i > t, D_{i1} = d_1]}{E[\sum_{\tau=g}^t D_{i\tau} | G_i = g, D_{i1} = d_1] - (t - g + 1)d_1}$$

Average cumulative treatment in group g, minus cumulative treatment between g-1 and t had they stayed at their initial value

Two notions of weighted averages

Scaled Effect in “binarized” groups (size p gets a 3 unit change, size $(1-p)$ gets a 4 units change; mean change is $3p + 4(1 - p)$):

Average “binarized” effect scaled by mean cumulative dose size:

$$ATT(g, t) = p \frac{ATT((0,1,2), 3)}{3p + 4(1 - p)} + (1 - p) \frac{ATT((0,2,2), 3)}{3p + 4(1 - p)} =$$
$$\frac{3p}{3p + 4(1 - p)} \frac{ATT((0,1,2), 3)}{3} + \frac{4(1 - p)}{3p + 4(1 - p)} \frac{ATT((0,2,2), 3)}{4}$$

For many g , you’d do this across the g as well.

In practice (kind of)

Bailey and Goodman-Bacon

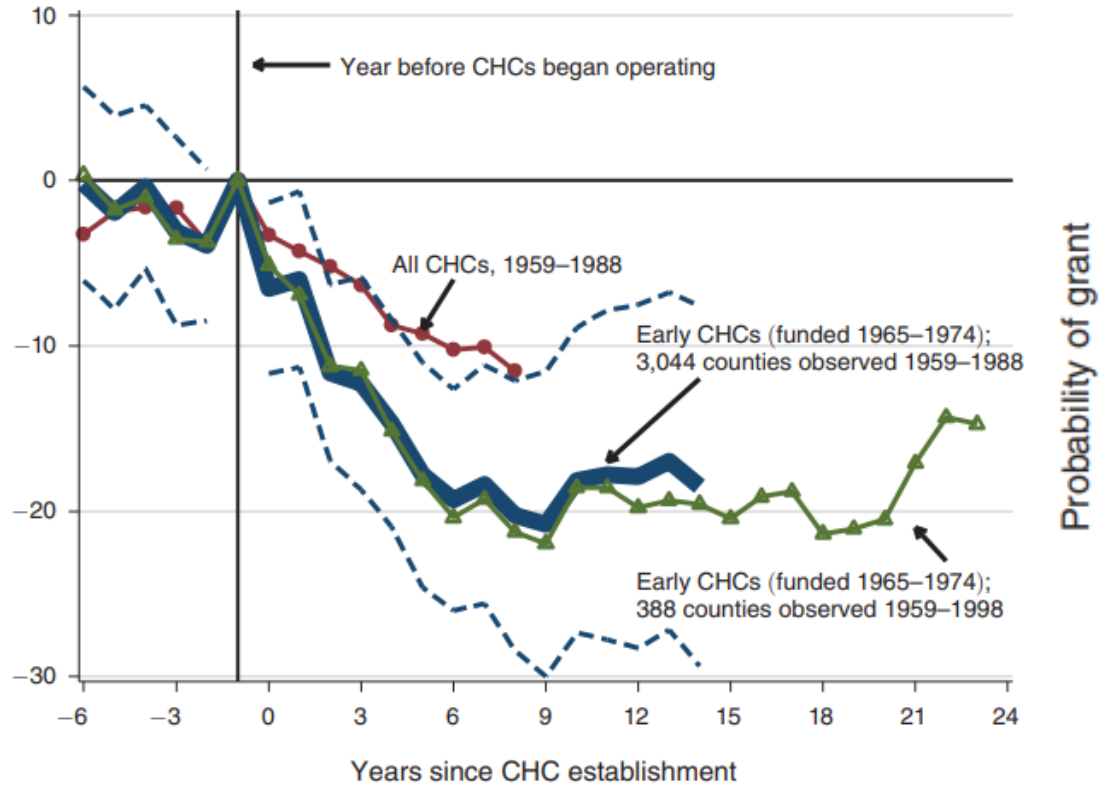
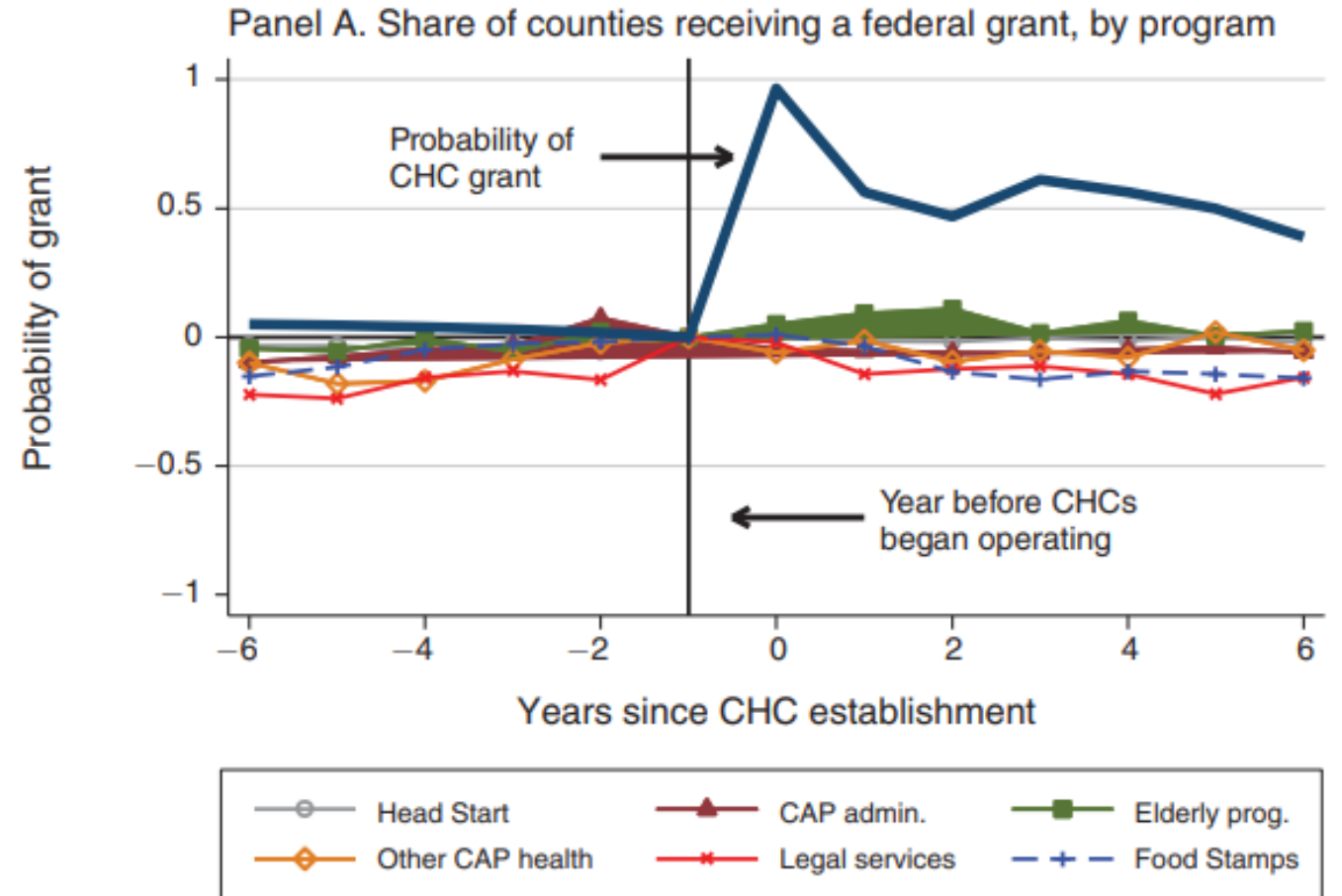


FIGURE 5. THE RELATIONSHIP BETWEEN COMMUNITY HEALTH CENTERS AND MORTALITY RATES



*Not cumulative

Today: a smorgasbord of additional DiD details

1. Starting from a non-zero dose (dCDH et al 2025) (same dynamics across groups)
2. Complicated treatment paths (dcDH 2024) (binarize, staggerize, scale by change in D)
3. Turn off designs (Kim and Lee)

Turn off designs

Set up:

never treated:	$\mathbf{D}_i = (0,0)$
always treated:	$\mathbf{D}_i = (1,1)$
turn off:	$\mathbf{D}_i = (1,0)$

Target parameter 1:

$$ATT(1) \equiv E[Y_{i1}(1) - Y_{i1}(0) | \mathbf{D}_i = (1,0)]$$

Identifying assumption(s):

PT: $E[\Delta Y(0,0) | \mathbf{D}_i = (1,0)] = E[\Delta Y(0,0) | \mathbf{D}_i = (0,0)]$

No *carryover* : $Y_2(1,0) = Y_2(0,0)$

Turn off designs

Target parameter 1:

$$ATT(1) \equiv E[Y_{i1}(1) - Y_{i1}(0) | \mathbf{D}_i = (1,0)]$$

Identifying assumption:

PT: $E[\Delta Y(0,0) | \mathbf{D}_i = (1,0)] = E[\Delta Y(0,0) | \mathbf{D}_i = (0,0)]$

No *carryover* : $Y_2(1,0) = Y_2(0,0)$

What is the estimand? Reverse DiD

$$E[Y_{i1} - Y_{i2} | \mathbf{D}_i = (1,0)] - E[Y_{i1} - Y_{i2} | \mathbf{D}_i = (0,0)]$$

Turn off designs

Target parameter 2:

$$ATT(2) \equiv E[Y_{i1}(1) - Y_{i1}(0) | \mathbf{D}_i = (1,0)]$$

Identifying assumption:

Similar implication: PT plus same treatment effect dynamics

PT: $E[\Delta Y(1,1) | \mathbf{D}_i = (1,0)] = E[\Delta Y(1,1) | \mathbf{D}_i = (1,1)]$

No *carryover* : $Y_2(1,0) = Y_2(0,0)$

What is the estimand? Reverse DiD w/ already treated control

$$E[Y_{i1} - Y_{i2} | \mathbf{D}_i = (1,0)] - E[Y_{i1} - Y_{i2} | \mathbf{D}_i = (1,1)]$$

Today: a smorgasbord of additional DiD details

1. Starting from a non-zero dose (dCDH et al 2025) (same dynamics across groups)
2. Complicated treatment paths (dcDH 2024) (binarize, staggerize, scale by change in D)
3. Turn off designs (Kim and Lee) (no carryover + no dynamics)
4. Selection and PT (Ghanem, Sant'Anna, and Wuthrich 2024)

Justifying parallel trends

Other paths through the blueprint:

What do you want to know/what can you know? (Parameter of Interest)

What are you willing to believe?	(<i>Parallel Trends</i>)
----------------------------------	----------------------------

What should you actually do? (Estimand/Estimator)

Were your beliefs plausible? (Internal Validity)

What behavior generates PT?

What does selection mean?

...while the new papers [in the DiD literature] clarify very well the statistical assumptions needed for estimation, effective use of these methods also requires being able to understand what the threats to these assumptions are in different contexts, and to make a plausible rhetorical argument as to why we should think the assumptions hold.

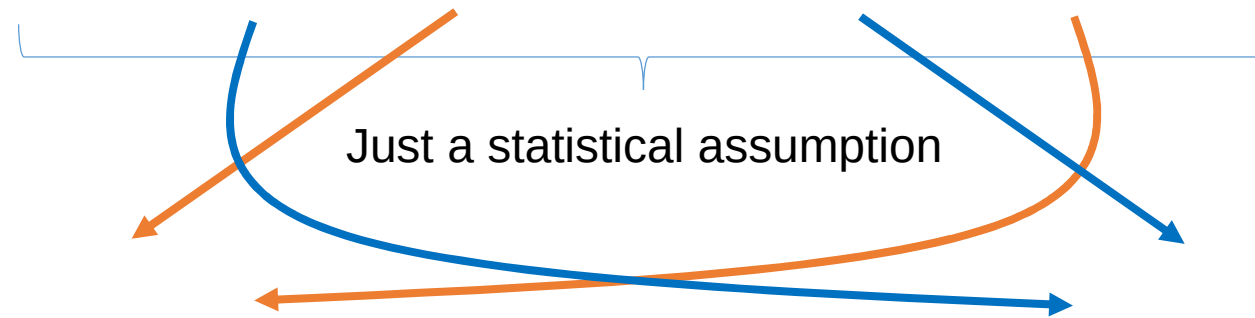
— David McKenzie, *World Bank Development Impact Blog* (McKenzie, 2022)

Can we build a mathematical argument to support (or refute) the rhetoric?

PT:

$$E[\Delta Y_i(\infty) | G_i = g] - E[\Delta Y_i(\infty) | G_i = \infty]$$

Just a statistical assumption



But someone chooses G : who? why?

...and is that choice related to trends in $Y_i(\infty)$?

(in fact, there are others you could make just as easily!)

What does selection mean?

Ashenfelter and Card (1985):

A. Simple Models

Suppose that earnings of the i^{th} individual in period t , y_{it} , follow a simple components-of-variance scheme:

$$y_{it} = \omega_i + d_t + D_{it}\beta + \epsilon_{it} \quad (1)$$

where ω_i is a permanent component, d_t is an economy-wide component, D_{it} is a dummy variable for participation in training during period t ,

1. Extreme (and convenient!) non-selection
→ simple mean difference

It is obvious from (1) that if assignment to training is independent of ω_i and ϵ_{it} , then a simple post-training difference in earnings between trainees and controls will estimate the training effect β .

What does selection mean?

Ashenfelter and Card (1985):

A. Simple Models

Suppose that earnings of the i^{th} individual in period t , y_{it} , follow a simple components-of-variance scheme:

$$y_{it} = \omega_i + d_t + D_{it}\beta + \epsilon_{it} \quad (1)$$

where ω_i is a permanent component, d_t is an economy-wide component, D_{it} is a dummy variable for participation in training during period τ ,

1. Extreme (and convenient!) non-selection
→ simple mean difference
2. Mild (yet still convenient) selection on the FE → DiD

To accommodate this fact, suppose that participation in training in period τ is governed by the magnitude of the permanent component of earnings, with

$$D_{it} = 1 \quad \text{for } t > \tau \text{ if and only if } \omega_i < \bar{y},$$

In this case, a simple estimate of the training effect is obtained from a comparison of the change in earnings for the trainees between some pre-training period ($\tau - j$) and the post-training period ($\tau + 1$) relative to the change in earnings for the control group over the same period. This “difference-in-differences” estimator provides an unbiased estimate of the training effect because

$$\begin{aligned} E(y_{i\tau+1} - y_{i\tau-j} | D_{i\tau+1} = 1) \\ - E(y_{i\tau+1} - y_{i\tau-j}) = \beta(1 - p) \end{aligned} \quad (2)$$

What does selection mean?

Ashenfelter and Card (1985):

A. Simple Models

Suppose that earnings of the i^{th} individual in period t , y_{it} , follow a simple components-of-variance scheme:

$$y_{it} = \omega_i + d_t + D_{it}\beta + \epsilon_{it} \quad (1)$$

where ω_i is a permanent component, d_t is an economy-wide component, D_{it} is a dummy variable for participation in training during period τ ,

1. Extreme (and convenient!) non-selection
→ simple mean difference
2. Mild (yet still convenient) selection on the FE → DiD
3. Slightly more involved selection + time series assumptions → DiD

Suppose, for example, that selection is based on the rule

$$D_{i\tau+1} = \begin{cases} 1 & \text{if } y_{i\tau-k} \leq \bar{y} \\ 0 & \text{if } y_{i\tau-k} > \bar{y}. \end{cases}$$

Here, selection into training is based on actual earnings in the k^{th} period prior to the advent of training. Provided, however, that the transitory error in earnings ϵ_{it} is taken to be serially uncorrelated, it remains the case that equation (2) continues to hold for $j > k$. That is, the difference-

What does selection mean?

Ashenfelter and Card (1985):

A. Simple Models

Suppose that earnings of the i^{th} individual in period t , y_{it} , follow a simple components-of-variance scheme:

$$y_{it} = \omega_i + d_t + D_{it}\beta + \epsilon_{it} \quad (1)$$

where ω_i is a permanent component, d_t is an economy-wide component, D_{it} is a dummy variable for participation in training during period τ ,

1. Extreme (and convenient!) non-selection
→ simple mean difference
2. Mild (yet still convenient) selection on the FE → DiD
3. Slightly more involved selection + time series assumptions → DiD
4. Same selection + different time series assumptions → slightly convoluted DiD

that if the transitory component ϵ_{it} is serially correlated, then trainee earnings will be abnormally low in periods adjacent to $\tau - k$, returning to their permanent level only as the transitory shock wears off.

so long as the earnings process is covariance stationary. It follows immediately that the symmetric difference-in-differences

$$E(y_{i\tau+1} - y_{i\tau-2k-1} | D_{i\tau+1} = 1) - E(y_{i\tau+1} - y_{i\tau-2k-1}) = \beta(1 - p)$$

What does selection mean?

Ashenfelter and Card (1985):

A. Simple Models

Suppose that earnings of the i^{th} individual in period t , y_{it} , follow a simple components-of-variance scheme:

$$y_{it} = \omega_i + d_t + D_{it}\beta + \epsilon_{it} \quad (1)$$

where ω_i is a permanent component, d_t is an economy-wide component, D_{it} is a dummy variable for participation in training during period t ,

1. Extreme (and convenient!) non-selection
→ simple mean difference
2. Mild (yet still convenient) selection on the FE → DiD
3. Slightly more involved selection + time series assumptions → DiD
4. Same selection + different time series assumptions → slightly convoluted DiD

Can we update this (PO instead of regression errors) and generalize the selection/time-series relationship?

YES!

What does selection mean?

Ghanem, Sant'Anna, and Wuthrich (2024):

Two periods ($t = \{1, 2\}$), units (i) choose treatment in period 2
($G_i = 1$ if treated, 0 otherwise)

Selection is: $G_i = g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \nu_i, \eta_{i1}, \eta_{i2})$

What does the selector
know about untreated PO?

Untreated PO is: $Y_{it}(\infty) = \xi_t(\alpha_i, \varepsilon_{it})$

Maybe they choose in period 1
and know their FE + their shock?

α_i : "I have a genetic risk factor, I'd
better get a check up."

$\alpha_i + \varepsilon_{i1}$: "I feel terrible today, I'd
better get a check up"

ε_{i1} : "I feel worse than usual today,
I'd better get a check up"

What does selection mean?

Ghanem, Sant'Anna, and Wuthrich (2024):

Two periods ($t = \{1, 2\}$), units (i) choose treatment in period 2 ($G_i = 1$ if treated, 0 otherwise)

Selection is: $G_i = g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \nu_i, \eta_{i1}, \eta_{i2})$

What does the selector
know about untreated PO?

Untreated PO is: $Y_{it}(\infty) = \xi_t(\alpha_i, \varepsilon_{it})$

Maybe they know the post-period shock before they choose treatment?

ε_{i2} : "I feel a migraine coming on tomorrow, I'd better get a check up."

What does selection mean?

[Ghanem, Sant'Anna, and Wuthrich \(2024\):](#)

Two periods ($t = \{1,2\}$), units (i) choose treatment in period 2
($G_i = 1$ if treated, 0 otherwise)

Example:

$$G_i = g(\alpha_i) = 1\{\alpha_i < \bar{y}\}$$

A&C #2:

Treatment just cuts the FE distribution in two.

Shocks are just bouncing around however.

Untreated PO is:

$$Y_{it}(\infty) = \xi_t(\alpha_i, \varepsilon_{it})$$

What does selection mean?

[Ghanem, Sant'Anna, and Wuthrich \(2024\)](#):

Two periods ($t = \{1, 2\}$), units (i) choose treatment in period 2
($G_i = 1$ if treated, 0 otherwise)

Example: $G_i = g(\alpha_i, \varepsilon_{i1}) = 1\{\xi_1(\alpha_i, \varepsilon_{i1}) < \bar{y}\}$

A&C #3/#4:

Treatment just cuts the distribution of $Y_{i1}(\infty)$ in two.

Untreated PO is: $Y_{it}(\infty) = \xi_t(\alpha_i, \varepsilon_{it})$

Shocks are just bouncing around however. But ε_{i2} and G_i are linked by the role of ε_{i1} .

What does selection mean?

[Ghanem, Sant'Anna, and Wuthrich \(2024\)](#):

Two periods ($t = \{1, 2\}$), units (i) choose treatment in period 2
($G_i = 1$ if treated, 0 otherwise)

Example:

$$G_i = g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})$$

A&C didn't even touch it!

Treatment may now pick out people with different ε_{i1} and ε_{i2}

...and therefore maybe different $\varepsilon_{i2} - \varepsilon_{i1}$

...and therefore maybe different $\Delta Y_i(\infty)$

Untreated PO is:

$$Y_{it}(\infty) = \xi_t(\alpha_i, \varepsilon_{it})$$

→PT violation

What does selection mean?

[Ghanem, Sant'Anna, and Wuthrich \(2024\)](#):

Two periods ($t = \{1,2\}$), units (i) choose treatment in period 2
($G_i = 1$ if treated, 0 otherwise)

Selection is: $G_i = g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \boxed{v_i, \eta_{i1}, \eta_{i2}})$

What does the selector
know about “other things”?

Untreated PO is: $Y_{it}(\infty) = \xi_t(\alpha_i, \varepsilon_{it})$

Could be costs of treatment:

v_i : “the doctor’s office is far away”

η_{i1} : “scheduling website is down today”

η_{i2} : “road construction tomorrow”

What does selection mean?

Ghanem, Sant'Anna, and Wuthrich (2024):

Two periods ($t = \{1,2\}$), units (i) choose treatment in period 2 ($G_i = 1$ if treated, 0 otherwise)

Selection is: $G_i = g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \boxed{v_i, \eta_{i1}, \eta_{i2}})$

What does the selector
know about “other things”?

Could be anything else
that's either

Fixed (v_i)
Idiosyncratic to period 1 (η_{i1})
Idiosyncratic to period 2 (η_{i2})

Untreated PO is: $Y_{it}(\infty) = \xi_t(\alpha_i, \varepsilon_{it})$

What does selection mean?

[Ghanem, Sant'Anna, and Wuthrich \(2024\)](#):

Two periods ($t = \{1, 2\}$), units (i) choose treatment in period 2 ($G_i = 1$ if treated, 0 otherwise)

Selection is: $G_i = g(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2}, \nu_i, \eta_{i1}, \eta_{i2})$

What does the selector
know about “other things”?

Untreated PO is: $Y_{it}(\infty) = \xi_t(\alpha_i, \varepsilon_{it})$

Generally need:

Something “random” in here
(in principle).

These things can't be just
proxies for what determines
 $Y_{it}(\infty)$: $(\alpha_i, \varepsilon_{i1}, \varepsilon_{i2})$

Example of the analysis:

If selection into treatment depends on the lagged outcome:

$$G_i = g(\alpha_i, \varepsilon_{i1})$$

Then:

$$E[\varepsilon_{i2} | G_i, \alpha_i, \varepsilon_{i1}] = E[\varepsilon_{i2} | \alpha_i, \varepsilon_{i1}]$$

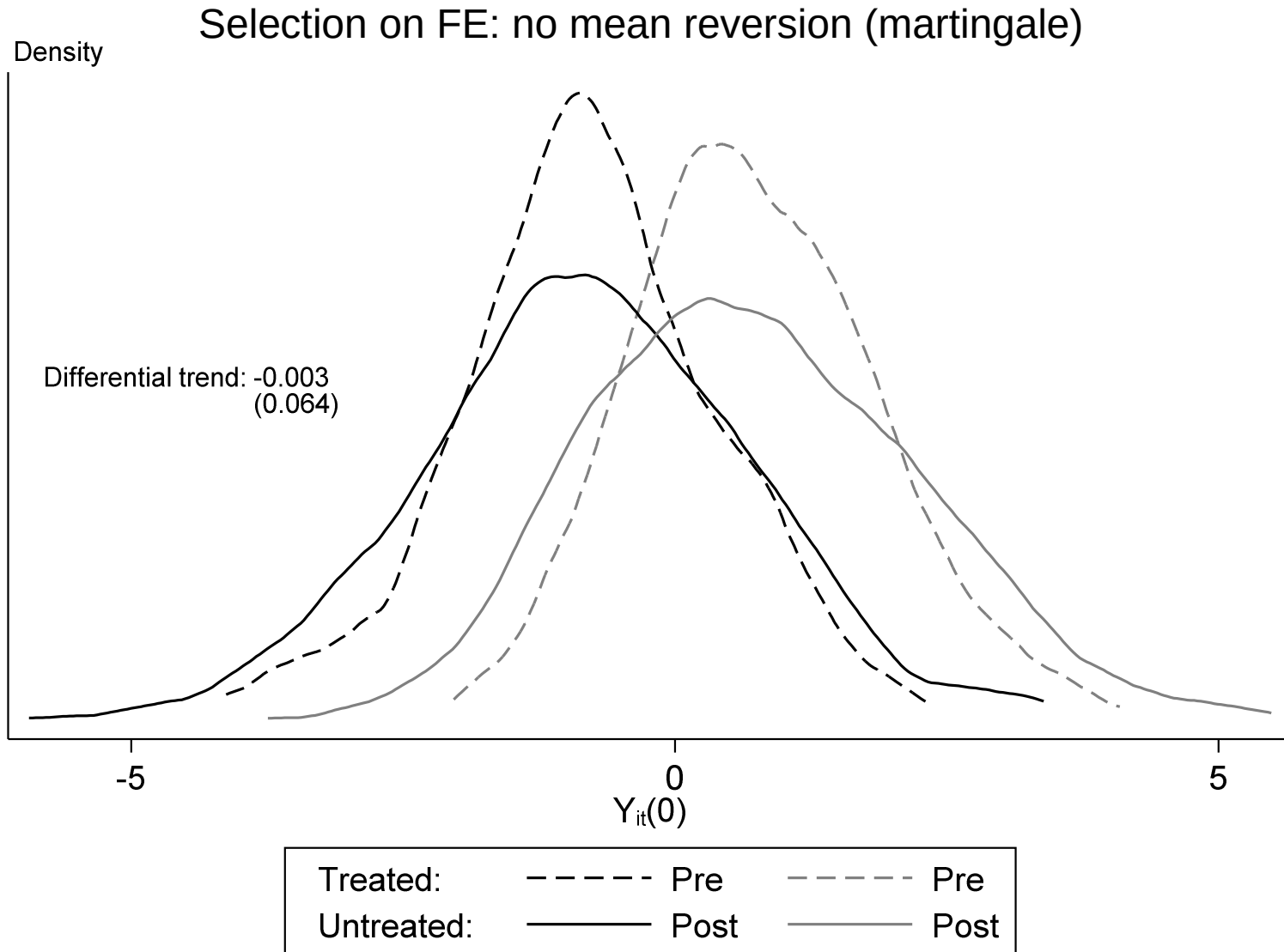
If, in addition, $E[\varepsilon_{i2} | \alpha_i, \varepsilon_{i1}] = \varepsilon_{i1}$, then:

$$E[\varepsilon_{i2} - \varepsilon_{i1} | G_i, \alpha_i, \varepsilon_{i1}] = 0$$

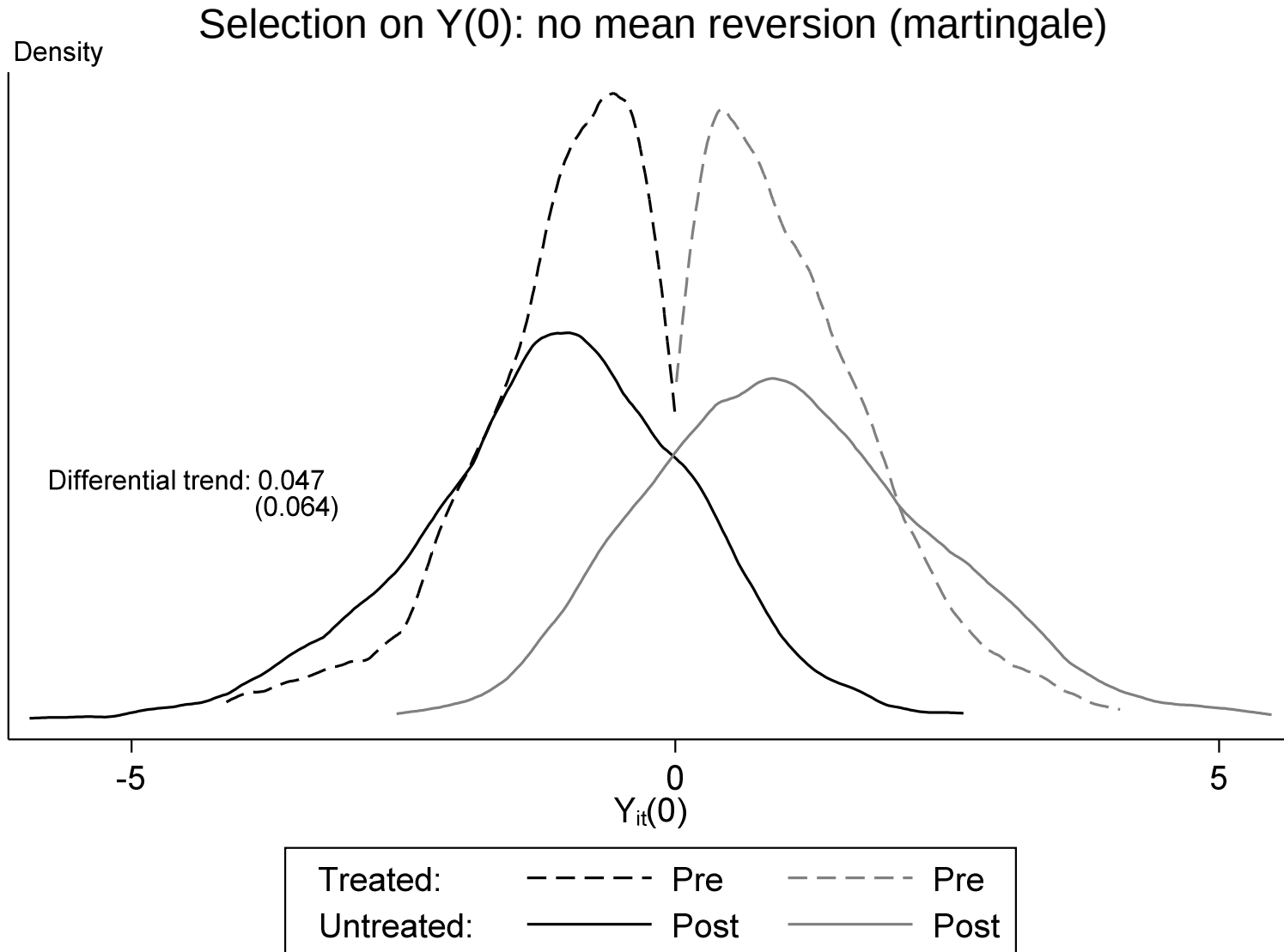
Which is Parallel Trends in a separable version of $Y_{it}(\infty) = \xi_t(\alpha_i, \varepsilon_{it})$.

...but its easier to see in examples.

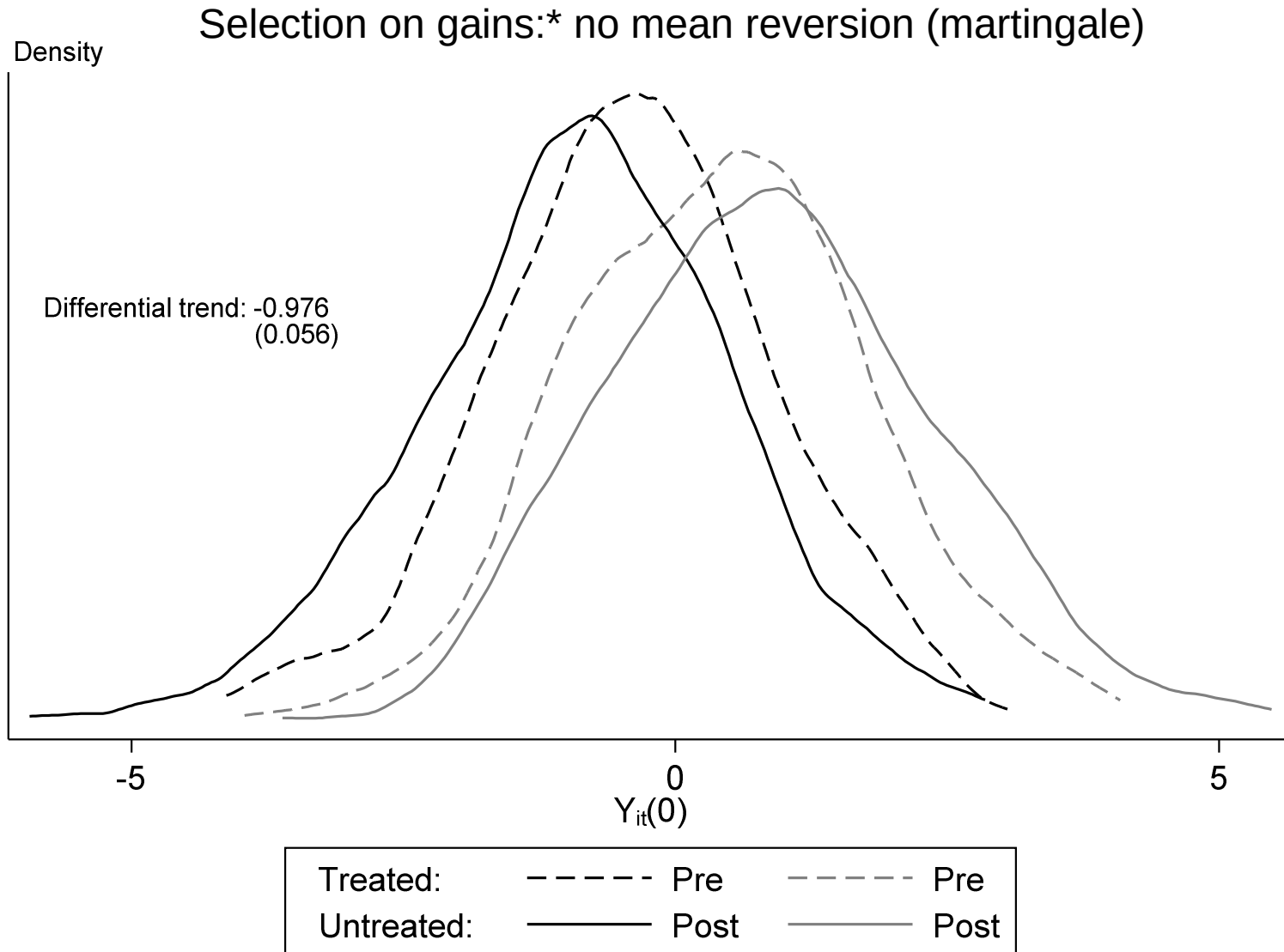
Tradeoff: selection vs time-series



Tradeoff: selection vs time-series

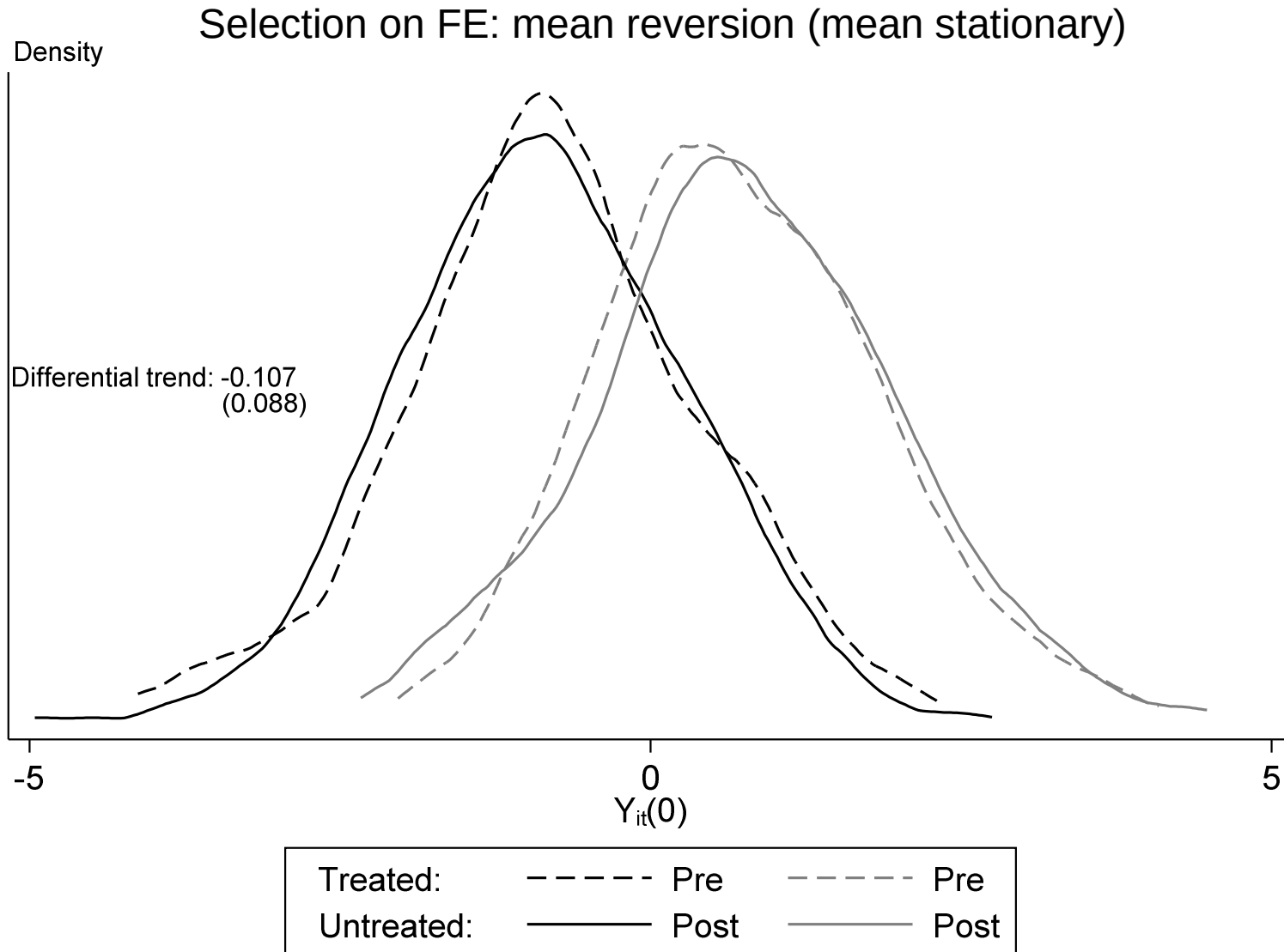


Tradeoff: selection vs time-series

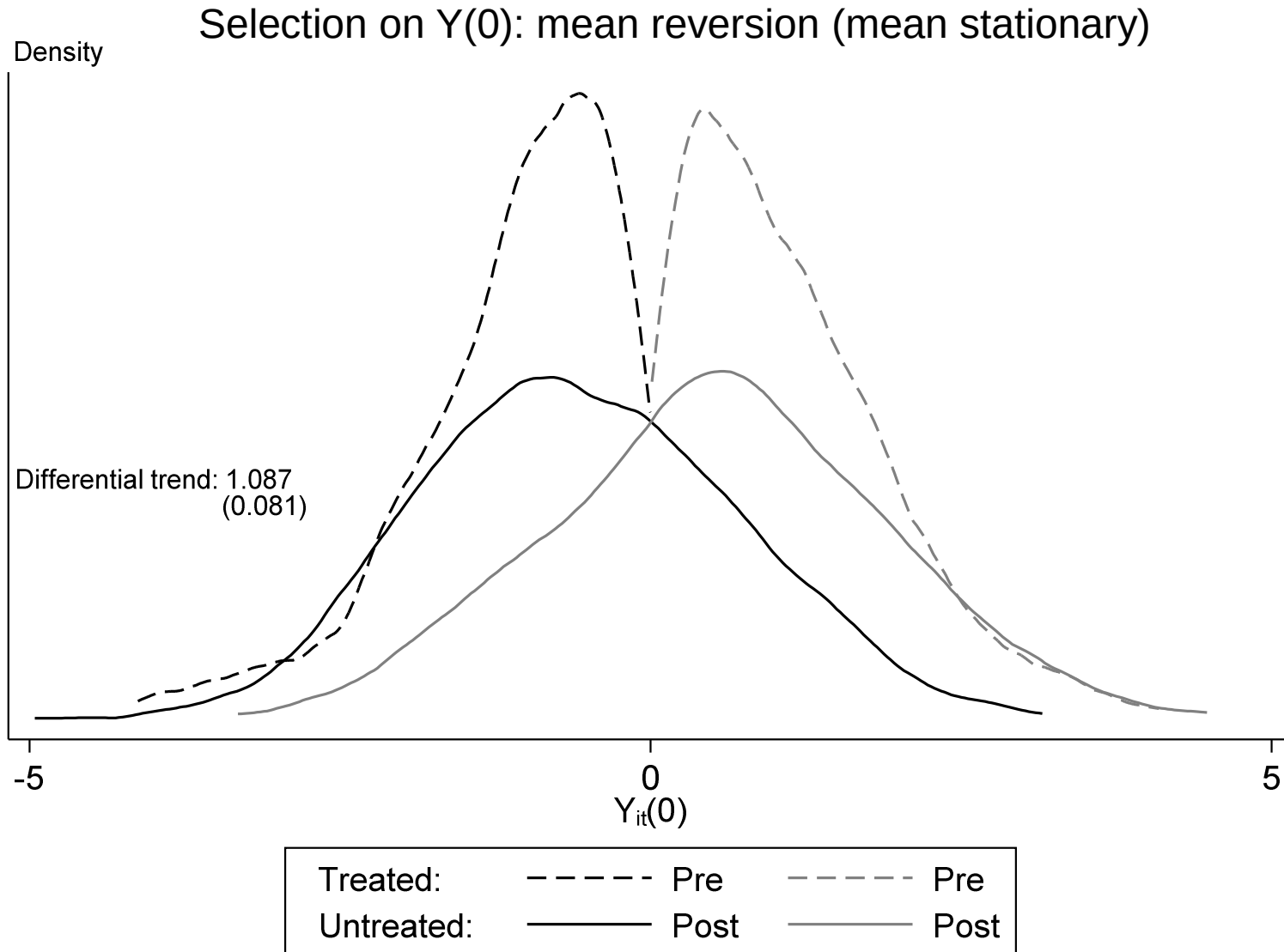


* $Y(1)$ negatively correlated with $Y(0)$

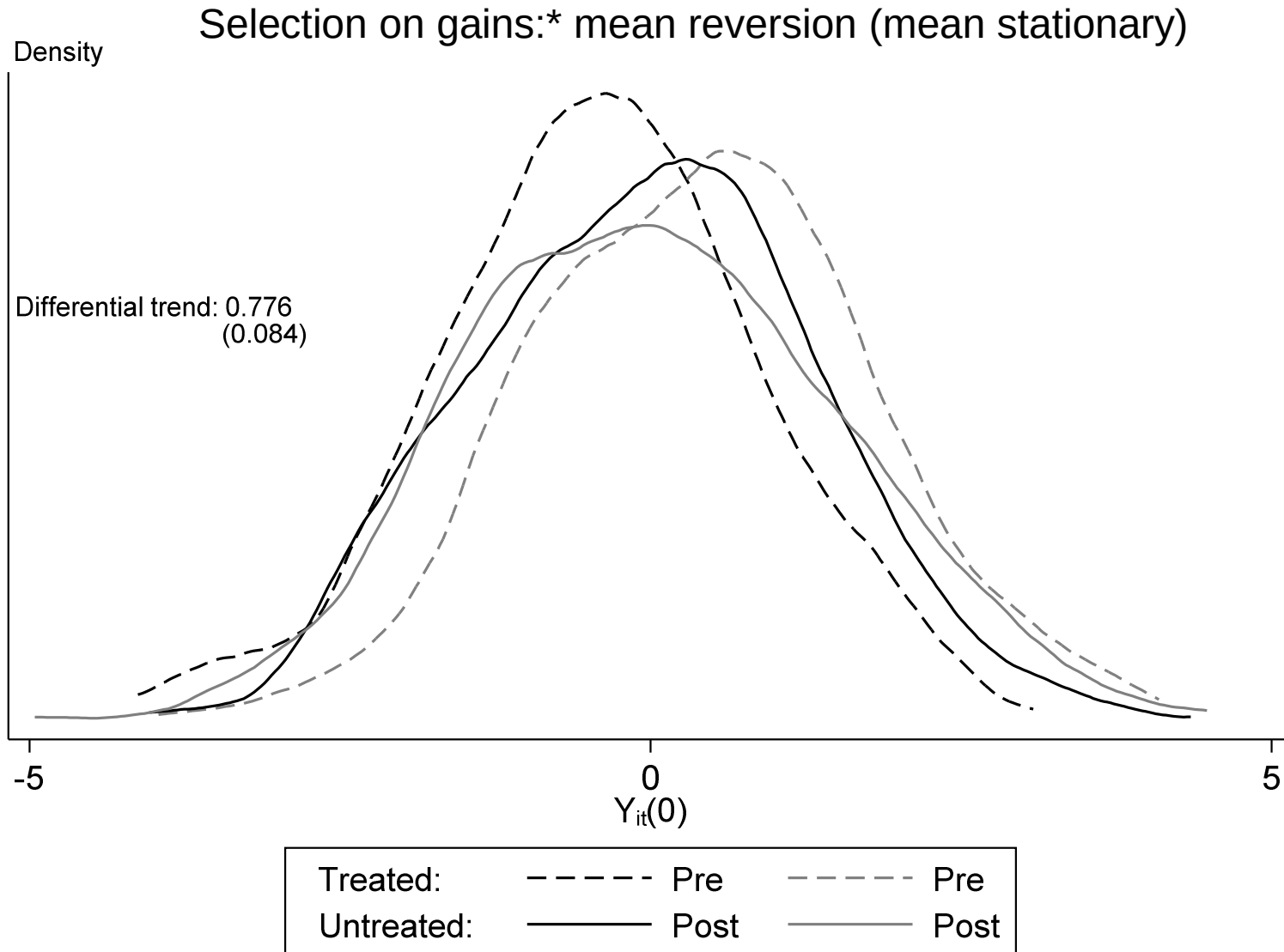
Tradeoff: selection vs time-series



Tradeoff: selection vs time-series



Tradeoff: selection vs time-series



* $Y(1)$ negatively correlated with $Y(0)$

Other paths through the blueprint:

What do you want to know/what can you know? (Parameter of Interest)

What are you willing to believe? (*Parallel Trends*)

What should you actually do? (Estimand/Estimator)

Were your beliefs plausible? (Internal Validity)

What if the answer is no?

Other paths through the blueprint:

What do you want to know/what can you know? (Parameter of Interest)

What are you willing to believe? *(Need a new assumption!)*

What should you actually do? (Estimand/Estimator)

Were your beliefs plausible? (Internal Validity)

What if the answer is no?

Other paths through the blueprint:

What do you want to know/what can you know? (Parameter of Interest)

What are you willing to believe? (*bounded PT violations*)

What should you actually do? (Estimand/Estimator)

Were your beliefs plausible? (Internal Validity)

Rambachan and Roth (honest did):

what if PT doesn't hold, but PT doesn't fail too badly?

Other paths through the blueprint:

What do you want to know/what can you know? (Parameter of Interest)

What are you willing to believe? (*bounded PT violations*)

What should you actually do? (inference on a set)

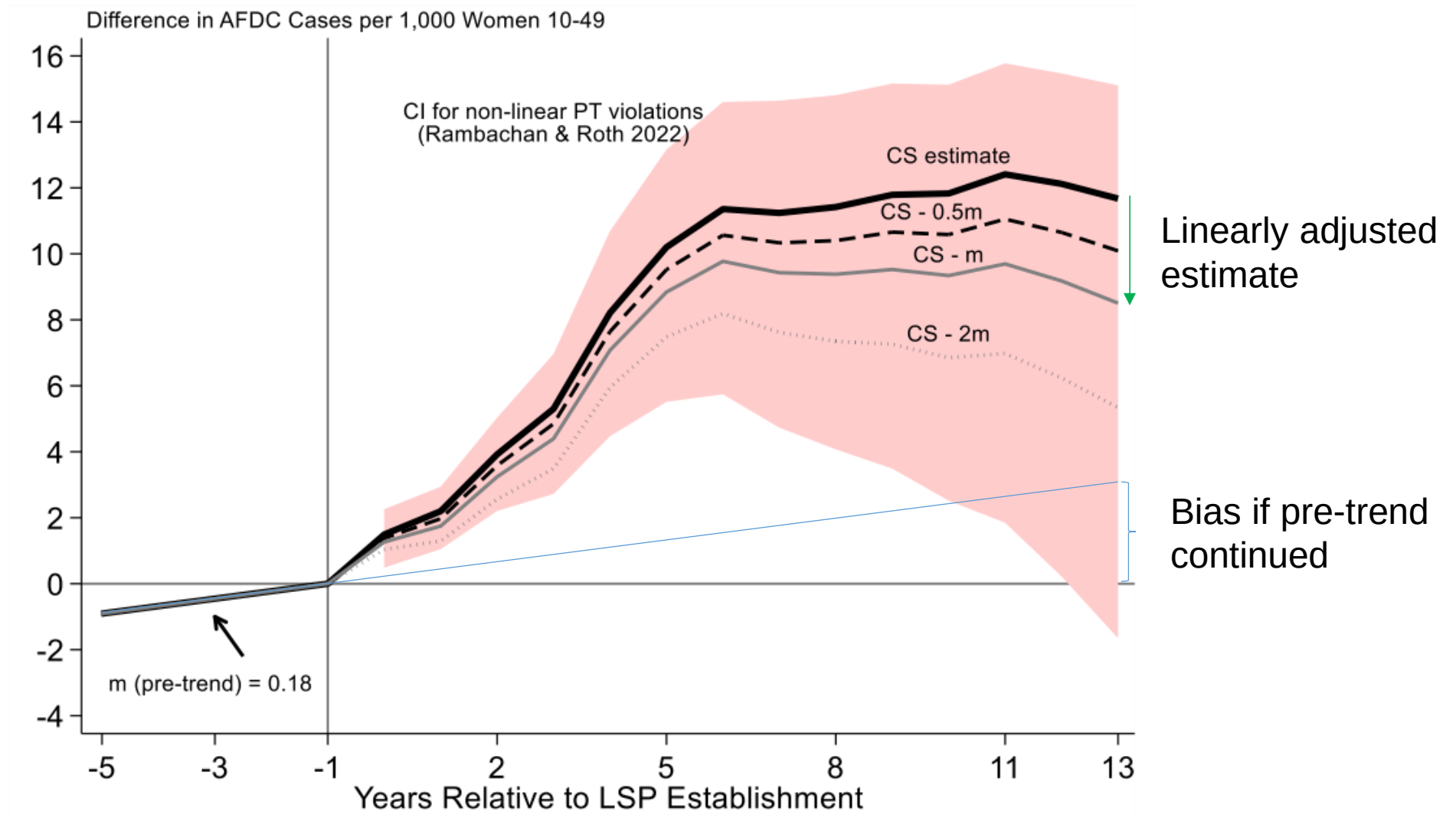
Were your beliefs plausible? (Internal Validity)

More plausible than PT b/c PT is nested!

Rambachan and Roth:

Construct C.I.s around a set of identified ATTs, given some PT violation

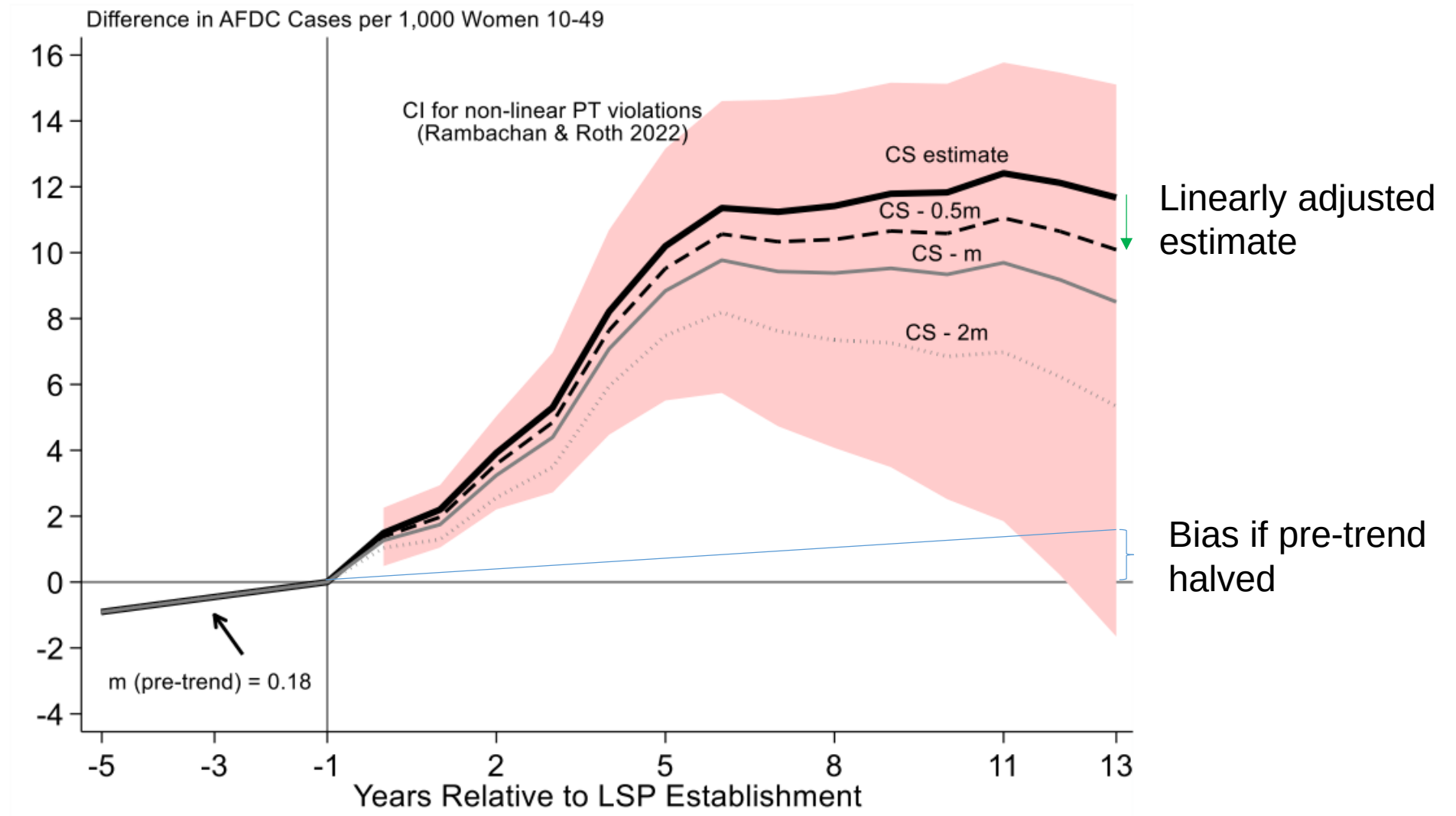
Bounding under PT violations



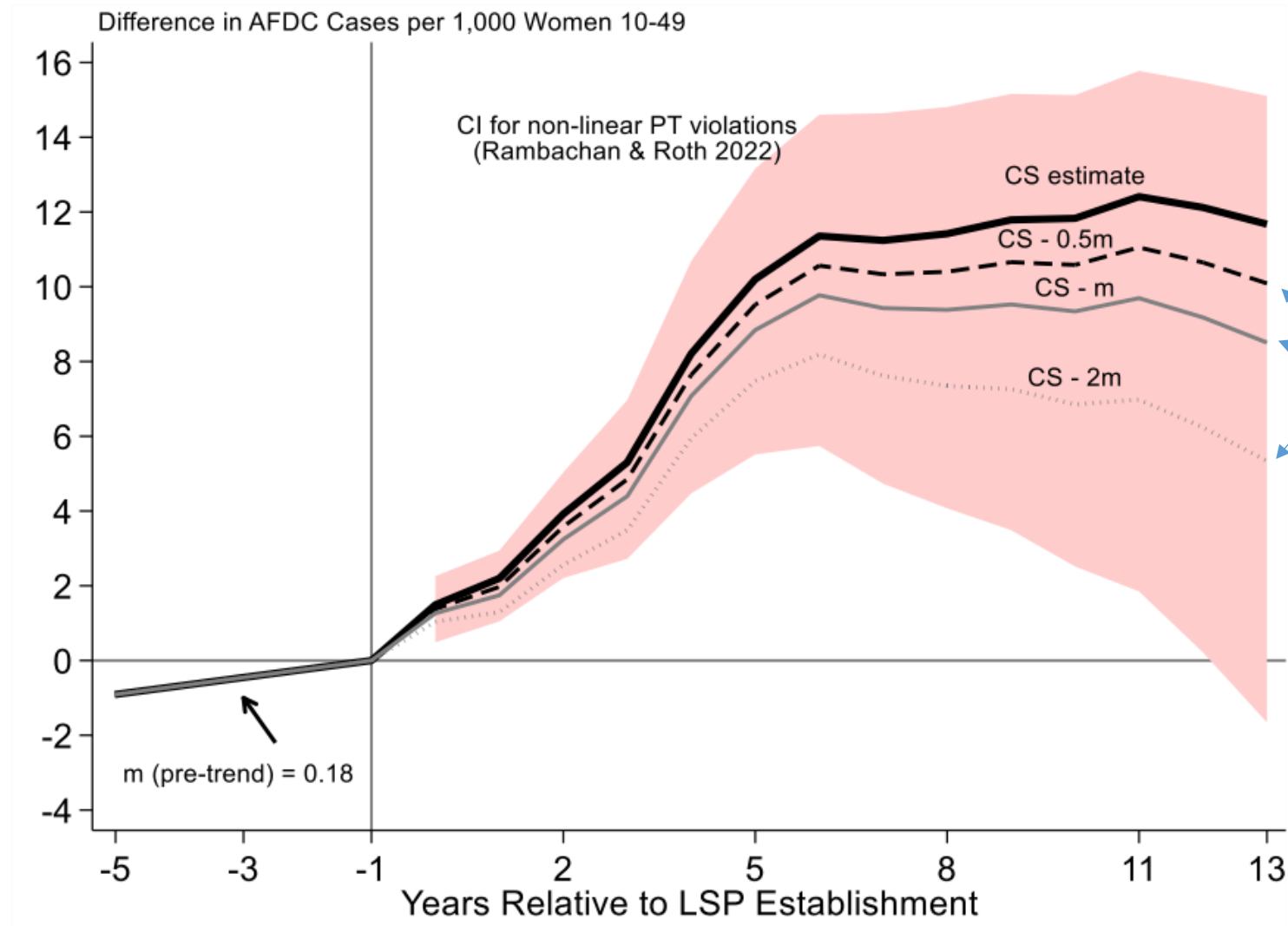
Bounding under PT violations



Bounding under PT violations



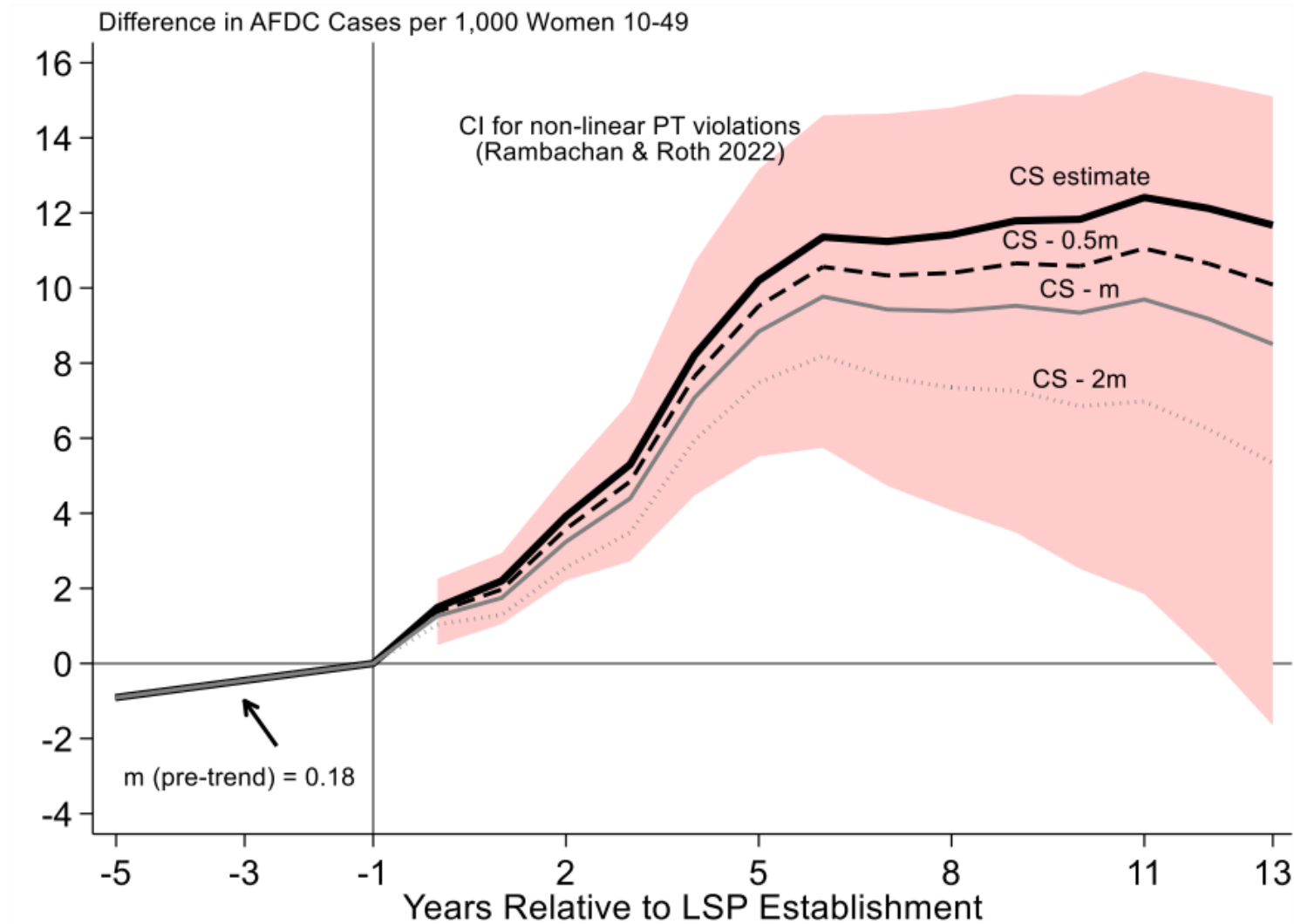
Bounding under PT violations



Limits:

- ad hoc
- linear violations
- no inference

Bounding under PT violations



Rambachan & Roth:

Confidence interval
around an *identified*
set under PT violations
that *get worse* by *m*
each year

Even more extensions...

Random treatment timing (Roth and Sant'Anna)

PT & functional form (Roth and Sant'Anna)

Bayesian version of bounding (Chen and Roth)

Spatial spillovers (Butts 2023)

Time-varying X's (Caetano et al 2024)

Compositional change (Hong 2013, Murto 2019, Sant'Anna and Zu 2025)

Triple Differences (Ortiz-Villavicencia and Sant'Anna 2025)

Efficient DiD (Chen ,Sant'Anna, Xie 2025)

A final thought

