Difference-in-Differences with a Continuous Treatment

Brantly Callaway

Andrew Goodman-Bacon

Pedro Sant'Anna

8/12/2021

DD Revolution

$$Y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{TWFE} D_{it} + u_{it}$$

Unit fixed effects

Time fixed effects

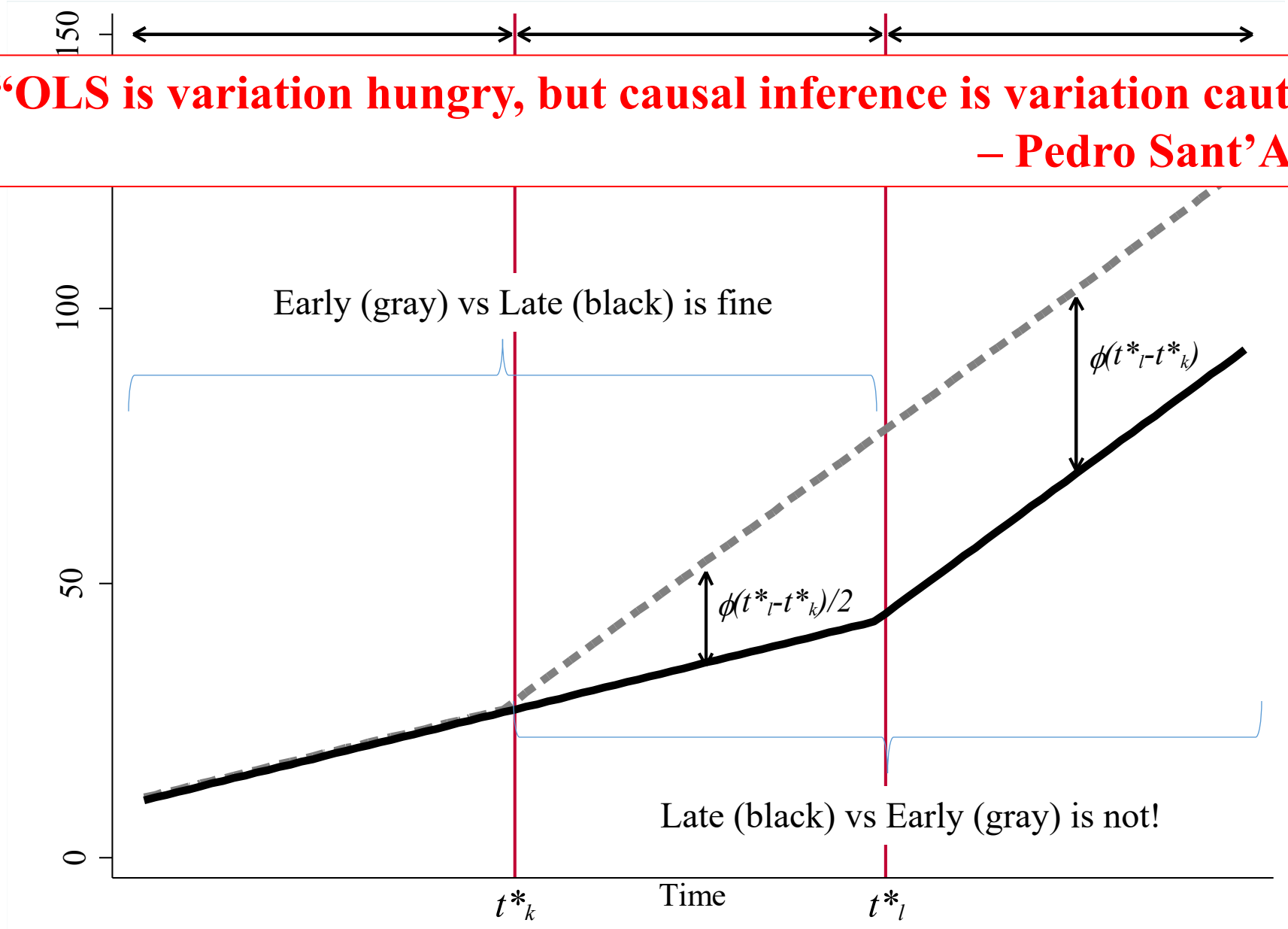
Treatment **dummy**

1. **With staggered treatment, $\hat{\beta}^{TWFE}$ is biased when treatment effects change over time** ***even under parallel trends in Y(0).***

Borusyak and Jaravel (2017), Strezhnev (2019), de Chaisemartin and D'Haultfouielle (2020), Sun and Abraham (2020), Goodman-Bacon (2021)

Bias from Time-Varying Effects

"OLS is variation hungry, but causal inference is variation cautious"
– Pedro Sant'Anna



DD Revolution

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

Unit fixed effects

Time fixed effects

Treatment **dummy**

1. With staggered treatment, $\hat{\beta}^{TWFE}$ is biased when treatment effects change over time *even under parallel trends in Y(0).*

2. **Alternative estimators that avoid those bad comparisons can estimate a causal effect *under parallel trends in Y(0) alone.***

Callaway and Sant'Anna (2020), Sun and Abraham (2020), Athey and Imbens (2020), Strezhnev (2019), de Chaisemartin and D'Haultfouielle (2020), Borusyak, Jaravel, and Speiss (2021)

What about continuous treatments?

Variation in treatment intensity (or "dose") is:

1. Common — *3/5 of my DD papers*
2. The only or maybe the "best" variation — *eg. small v large distances*

"Dose-response effects" are:

1. Evidence for internal validity — *Hill Criteria*
2. Of interest per se — *marginal vs average effects*

What do we know about continuous treatments?

Econometrics:

Fuzzy DD (dC&d'H 2018), TWFE in general (dC&d'H 2020): problems with heterogeneity

NL DD estimator (d'H, Hoderlein, Sasaki 2020)

Practice:

Just do TWFE

"the two-period regression estimator can be easily modified to allow for continuous, or at least non-binary, 'treatments'" (Wooldridge 2005)

"a second advantage of regression DD is that it facilitates the study of policies other than those that can be described by a dummy" (Angrist and Pischke 2008)

Continuous Treatments?

Hani Mansour @hnmansour · Dec 14, 2020
I remember seeing a paper about estimating an event study with a continuous variable but can't seem to track it. Any leads #EconTwitter @causalinf @agoodmanbacon?
7 8 27

David Burgherr @D_Burgherr · Mar 25, 2020
On this point, I am very curious how much the issues you point out with DD designs -- variance-weighting of treatment effects (TE) and bias in the case of time-varying TE -- matter for continuous treatment variables. Do you have any take or reference on that?

Thanks in advance!
2

Khoa Vu @KhoaVuUmn · Nov 16, 2020
#EconTwitter Question on DiD: I'm looking for a reference on what to do when the treatment variable is continuous and you suspect that the effect is nonlinear, e.g. medium exposure might have bigger effect than high exposure.
4 7 42

Ben Glasner @BenGlasner · Oct 29, 2020
Any recs on DID packages in r for multiple treatment periods with different timings and continuous treatment values? Think minimum wage changes over time? Something to compare TWFE against... #econtwitter
2

Michelle Spiegel @michspieg · Apr 22, 2020
I am writing a DiD paper with a continuous treatment. Any paper recommendations to help think about statistical power in this context? #EconTwitter #soctwitter #AcademicTwitter

Kait Sims @kaitmsims · Aug 25, 2020
#EconTwitter recommendations for event study/DiD papers with staggered treatment time, continuous treatment intensity, and where treatment can turn on and off more than once for the same individual?

3 4 4

Jason Baron @JasonBaron4 · Apr 21, 2020
I know there have been previous threads on the most recent DID papers, but does anyone know if there are any recent methodological papers specifically looking at DID implementation with a continuous treatment variable? @causalinf @jondr44
4 9 37

Adam Roberts @adamn_roberts · Mar 28
Question for DiD experts:
Is there a heterogenous treatment effects solution that works for continuous treatments? I'm specifically thinking about early childhood intervention papers that define treatment as "age 0-5" exposure to something like county food stamps availability.
1 7

Adam Roberts @adamn_roberts · Mar 28
This type of treatment has staggered timing and enough heterogeneity to make TWFE a poor approach but after diving into the new DiD literature I'm struggling to figure out the "correct" approach with a continuous treatment variable. Any thoughts @causalinf @Andrew__Baker ?
4

Nicholas Reynolds @nick_reynolds88 · Apr 28
Does anyone know of papers deriving what TWFE with continuous treatment and allowing for heterogenous treatment effects estimates?

My intuition is same "problems" found for staggered diff-in-diff would exist here ... but notation would explode so no one has written it out?
1 2 2

Peter Bergman @peterbergman_ · May 11, 2020
Seems like "dosage"/"intensity" diff-in-diff--where there aren't 2 groups but a continuous measure w/ varying intensity of treatment--requires potentially stronger identifying assumptions than DID for 2 groups. Is this discussed in any of the recent DID lit updates? cc @causalinf
6 5 47

Nick Hagerty @hagertynw · Mar 29
Conceptually it's not that distinct right? We're still trying to identify off similar shocks in different places at different times. I thought the main difference is that our variables are continuous treatments -- algebra is harder but papers prob. coming in next couple years
3 6

Michael Wiebe @michael_wiebe · Feb 9
Who's writing the @agoodmanbacon paper on diff-in-diff with a continuous treatment variable (instead of binary)?

#EconTwitter
1

Davide Proserpio @dade_us · Apr 12, 2020
Looking for recommendations about DD papers where the treatment is continuous, thanks! #EconTwitter
6 1 8

This Paper: Mainly Diagnosis

1. Define causal parameters
   a. ATT = effect of dose $D = d$ vs $D = 0$ (for group d)
   b. ACRT = effect of dose $D = d_j$ vs $D = d_{j-1}$ (for group d)

2. Identification of "building blocks"
   a. ATTs: just like 2x2 DD *Parallel Trends – good!*
   b. ACRTs: essentially need homogeneous effects *Strong Parallel Trends – bad?*

3. Two-Period TWFE Estimand
   a. "heterogeneity bias" and weights *Weighted average decomposition*

4. Staggered Timing TWFE Estimand *Weighted average decomposition*
   a. All the bias!

Upshot/Future Work

You can’t totally estimate your way out of these problems.

But you can state the identifying assumptions and try to justify them**

*(perhaps with alternative estimators to TWFE)

Causal Parameters

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:

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

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:

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

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:

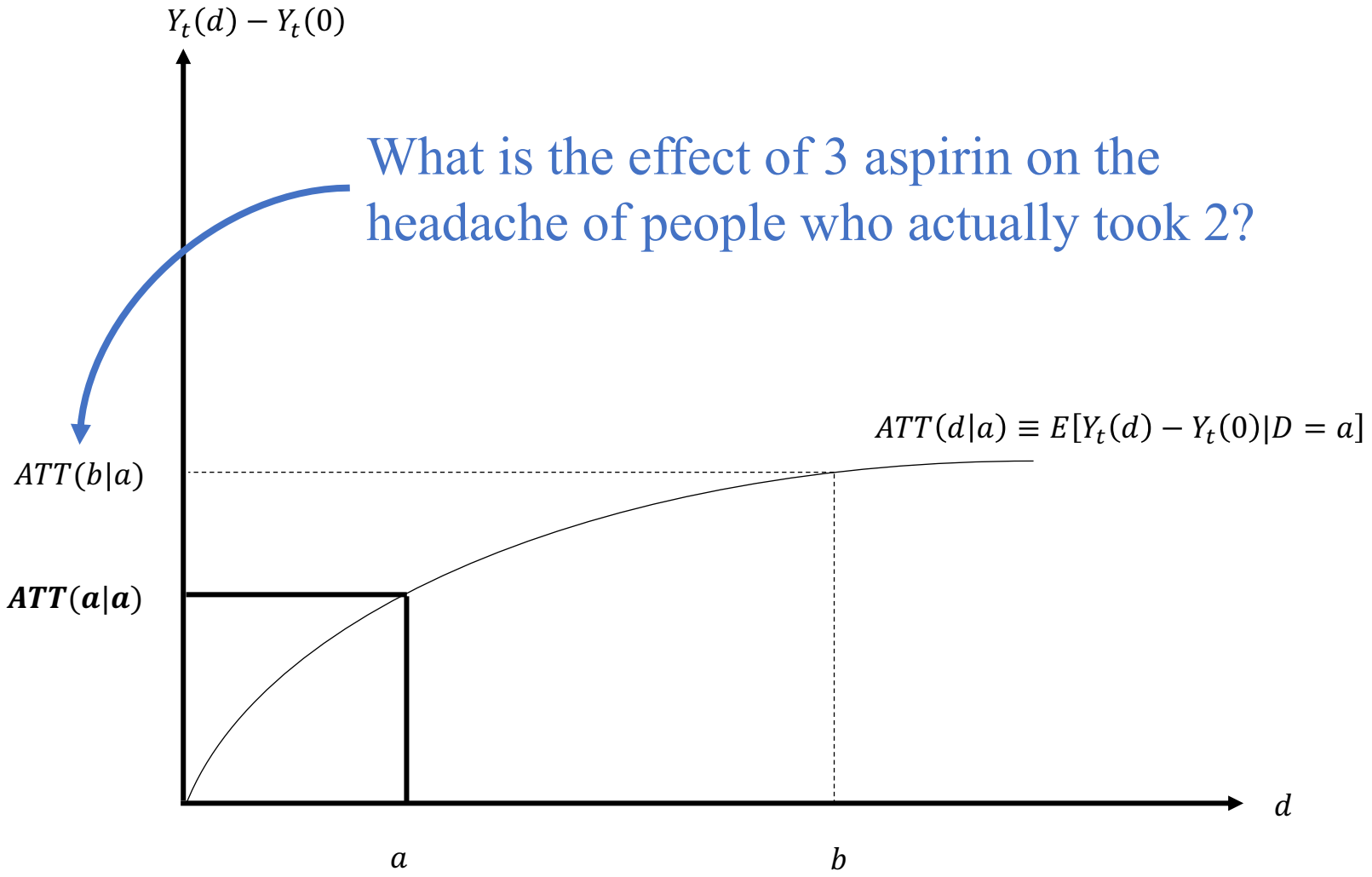
$$Y_{it}(d) - Y_{it}(0)$$

Average treatment effect on the treated in period t:

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

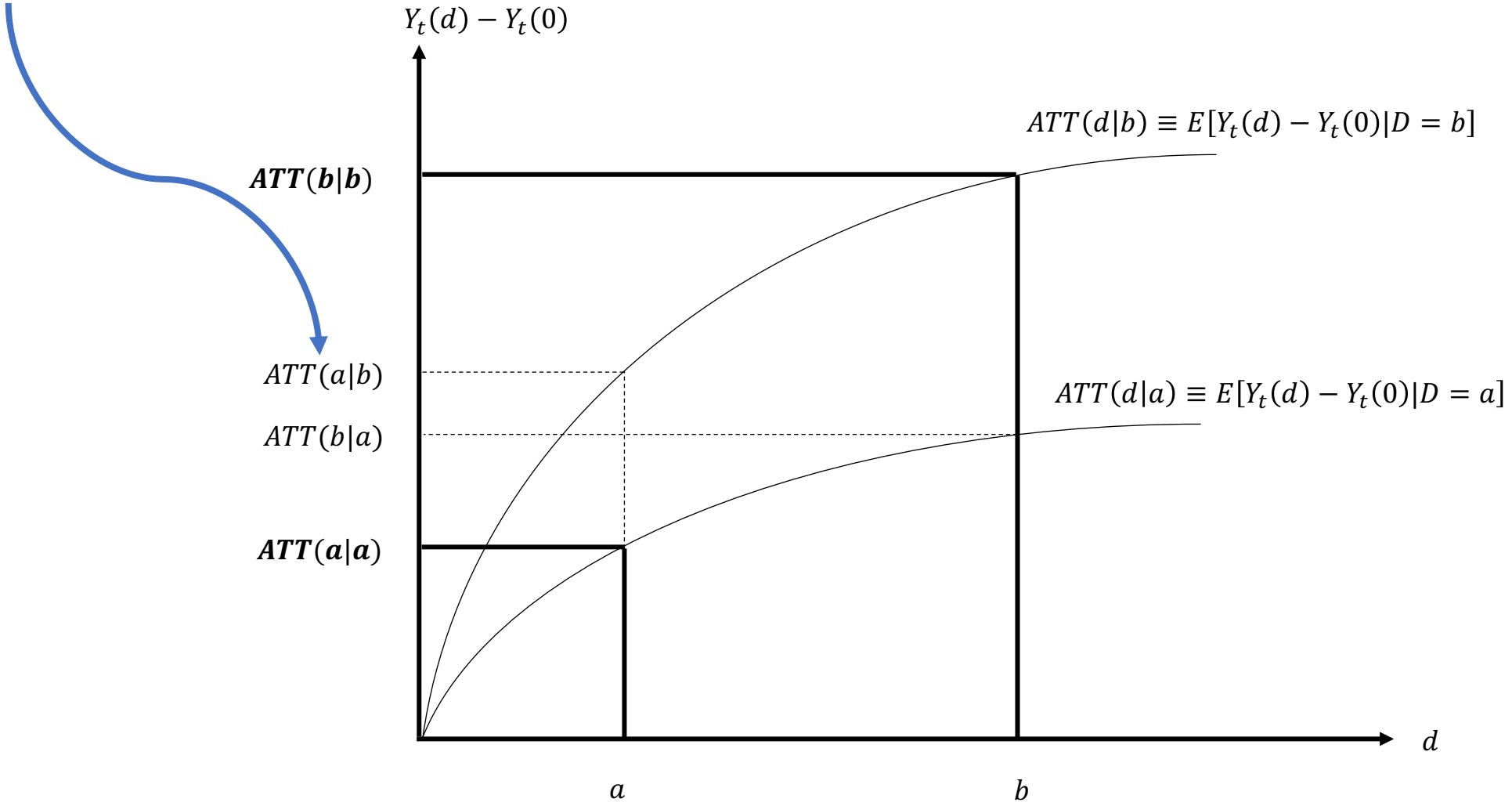
But this one "treatment" could be anything, any dose, any intensity.

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



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

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



Building Blocks: Slopes

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

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

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

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

When pressed, this is what we usually "say"

*Me conflating ATT per dose unit and ACR:

> *"After Medicaid, nonwhite child mortality fell by 1.4 percent (s.e. = 0.34; table 3) for each percentage point difference in initial AFDC rates."*

> *(later on I say "treatment effects")*

*Elasticity interpretation (Card 1992)

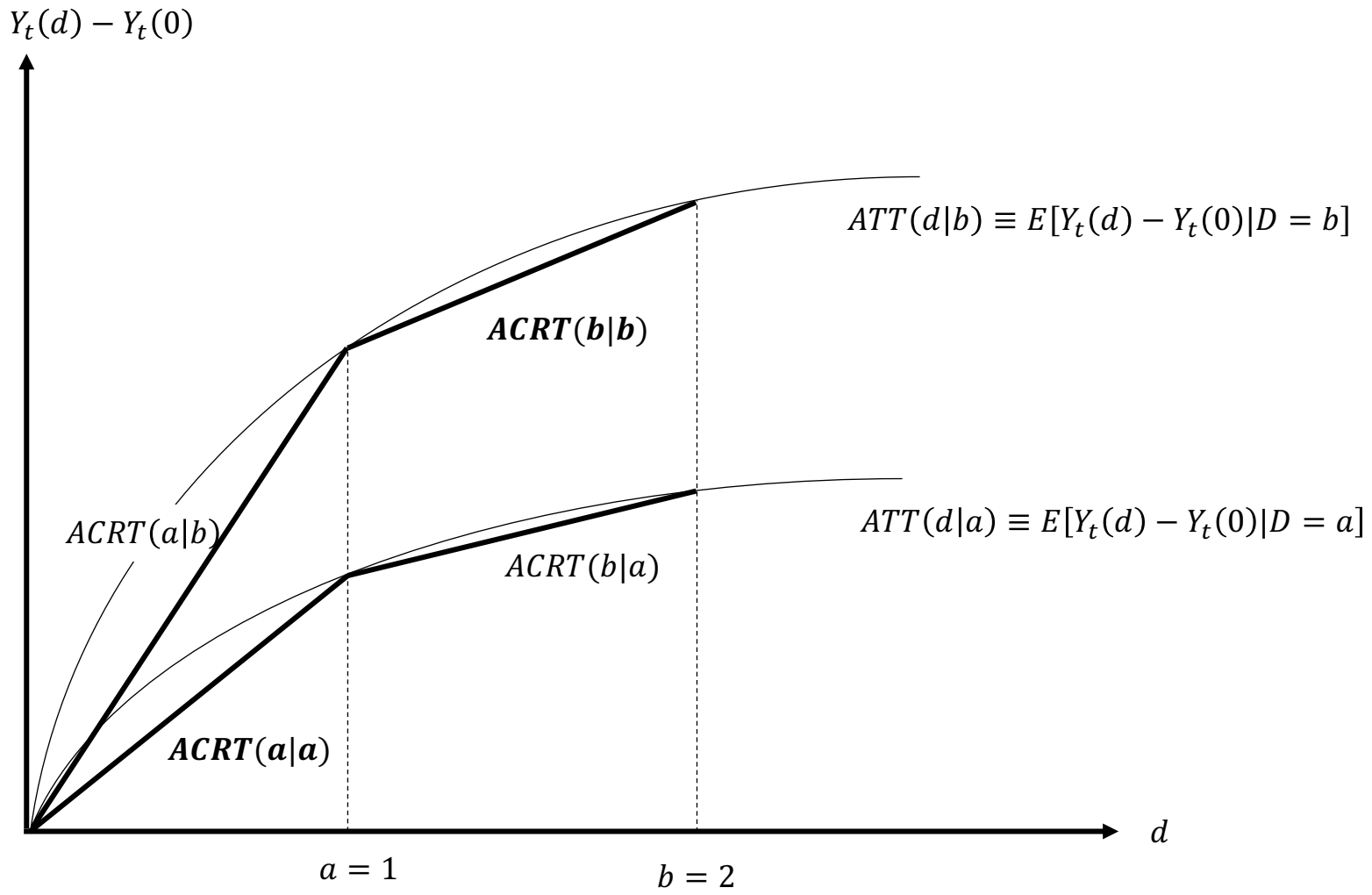
Building Blocks: Slopes

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

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

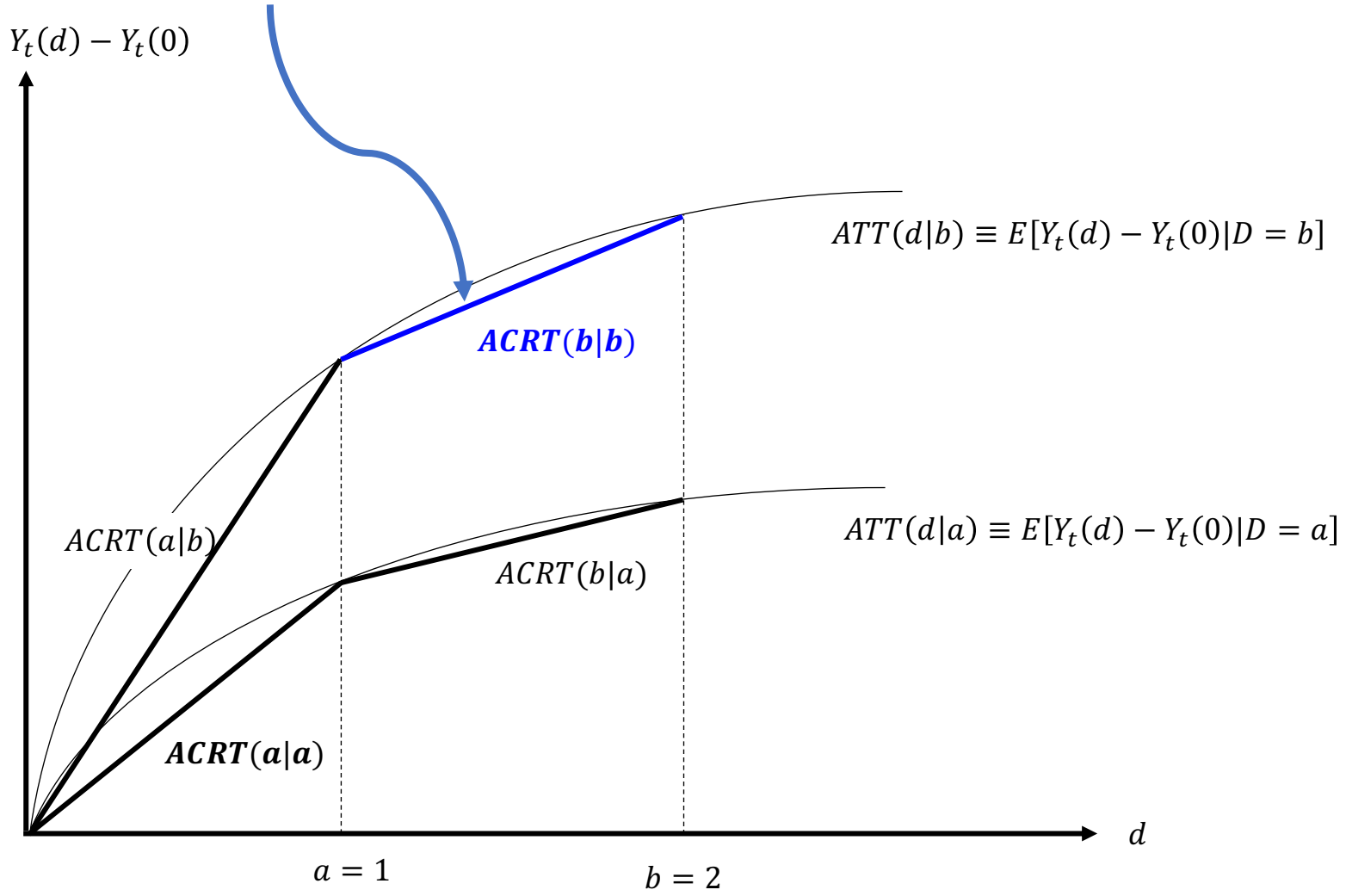
"the causal effect of the j^{th} dose increment for unit i"

Average Causal Responses on the Treated, Discrete: $\boldsymbol{ACRT(d_j|d_j)}$



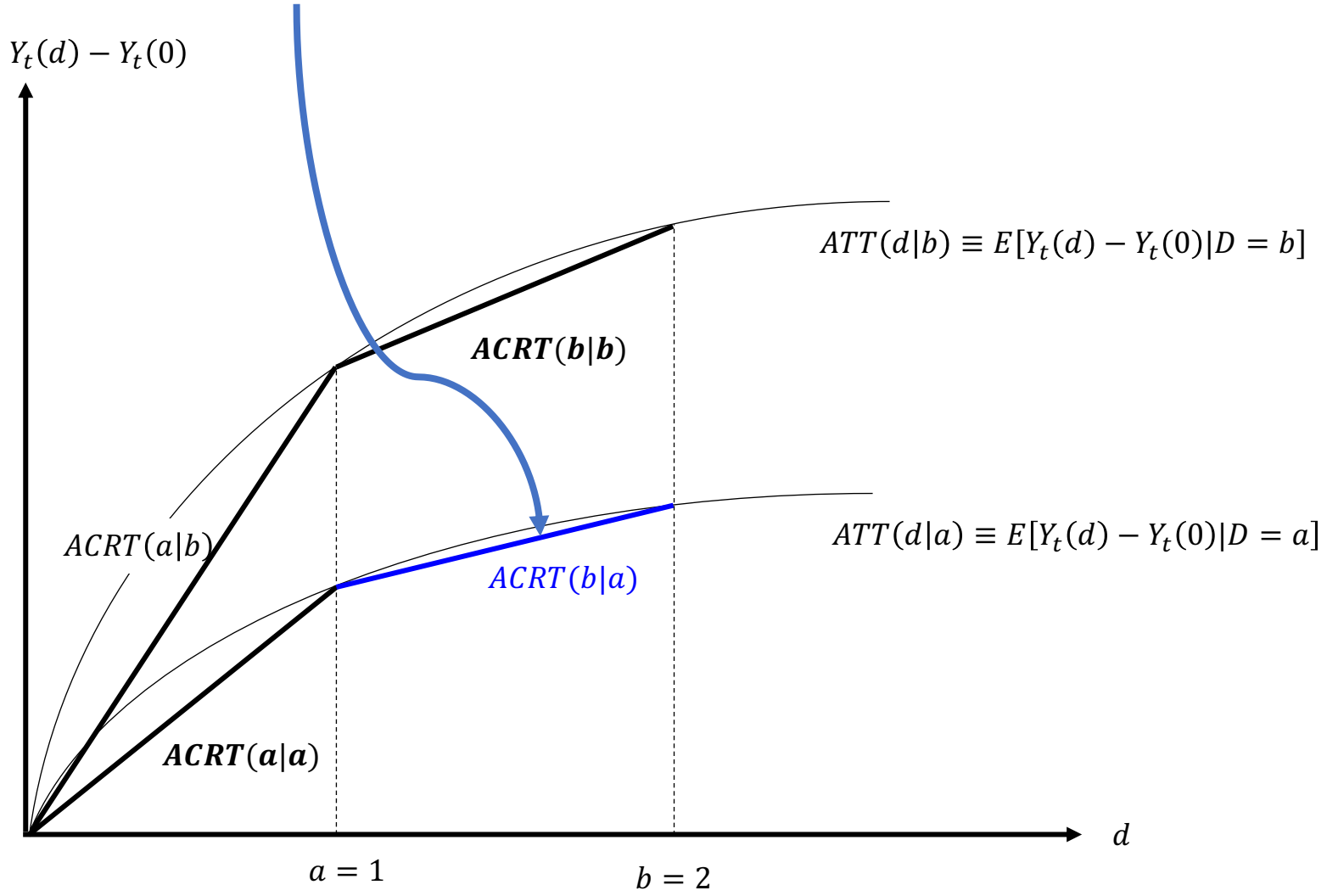
Average Causal Responses on the Treated, Discrete: $\boldsymbol{ACRT(d_j|d_j)}$

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



Average Causal Responses on the Treated, Discrete: $\boldsymbol{ACRT(d_j|d_j)}$

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



Building Blocks: ACRTs

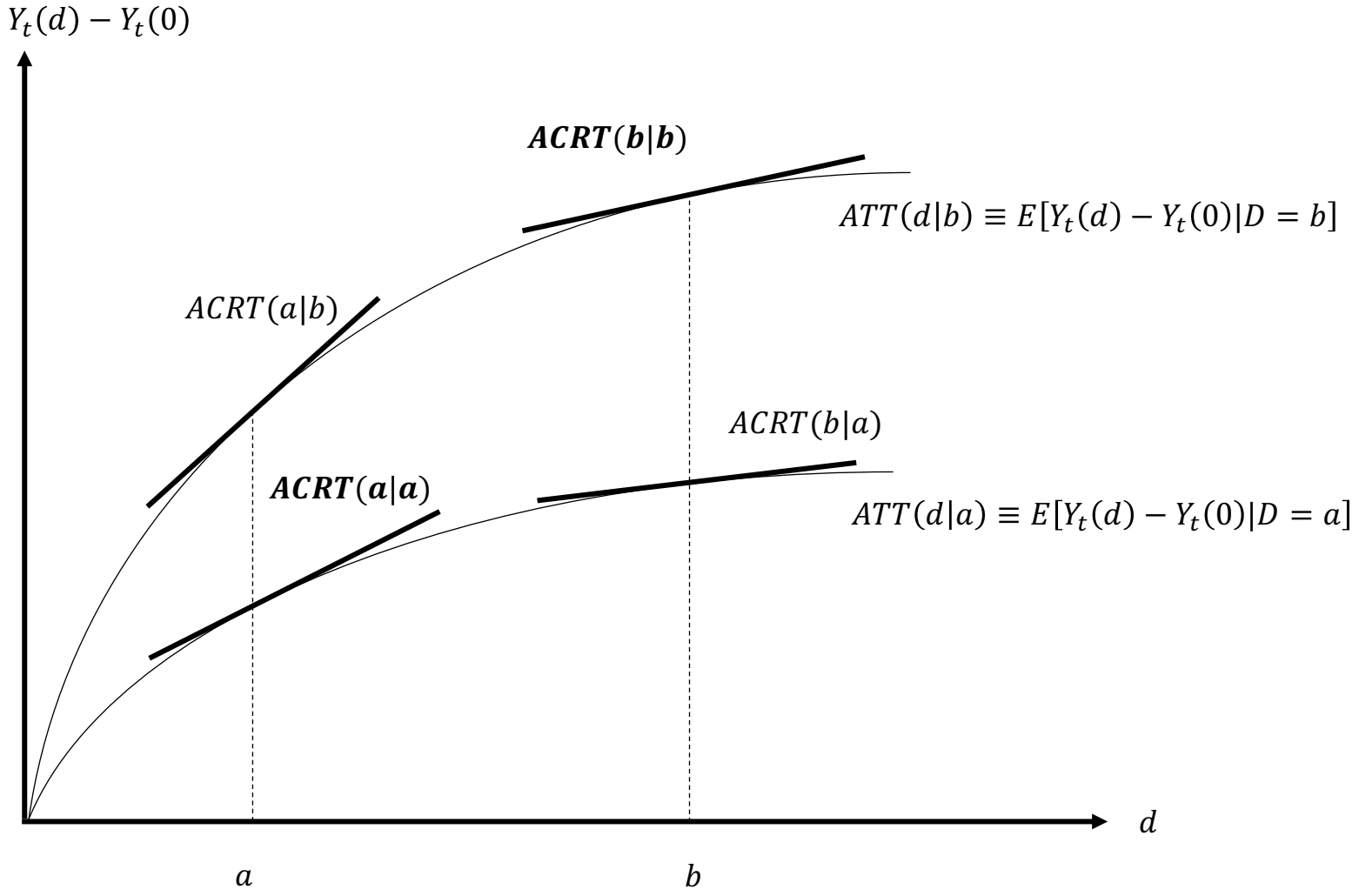
Discrete/multi-valued treatment:

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

Continuous treatment:

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

Average Causal Responses on the Treated, Continuous: $\boldsymbol{ACRT(d|d)}$



Comment 1: ATEs and ACRs

Average Treatment Effect:

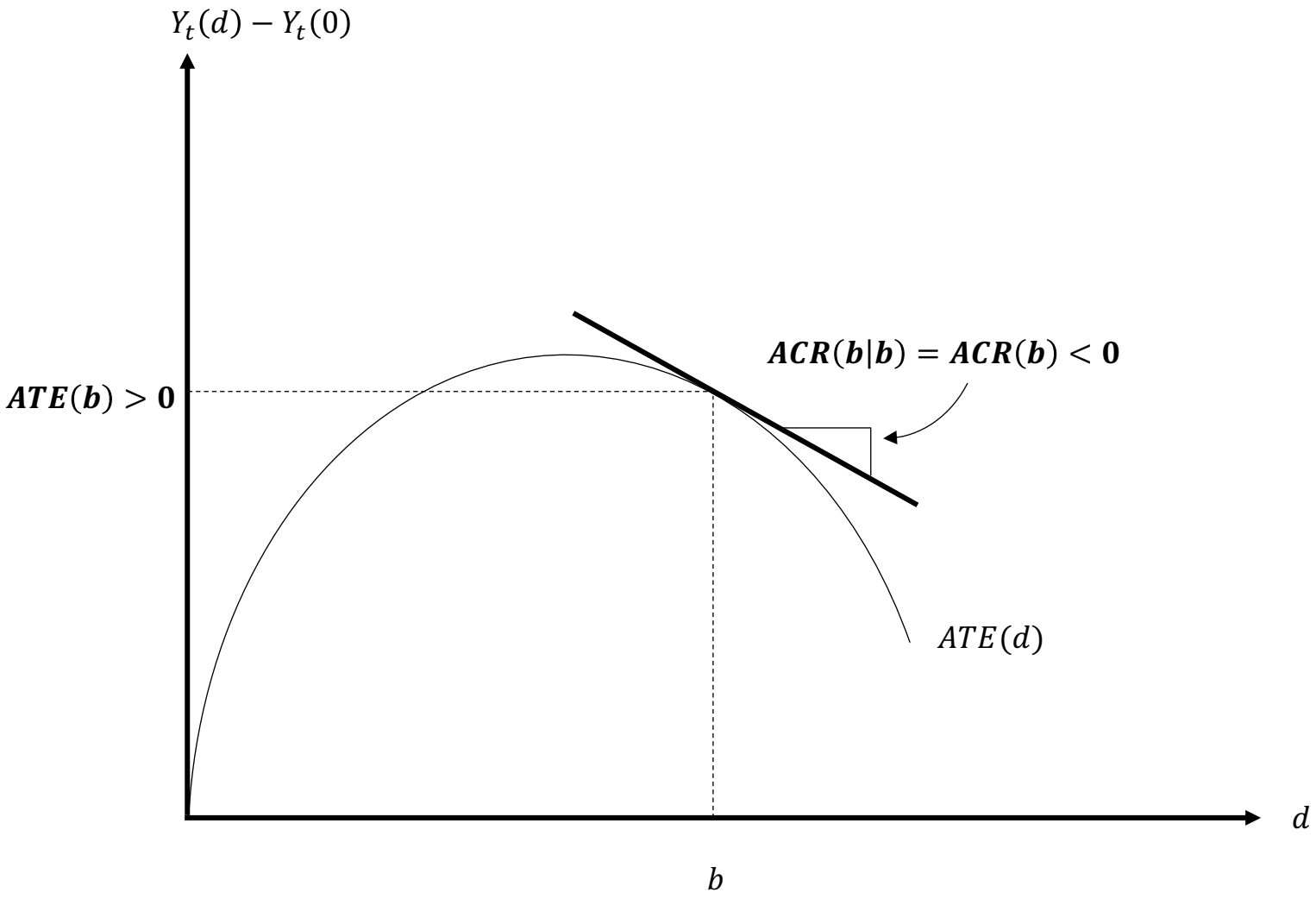
$$ATE(d) \equiv E[Y_{it}(d) - Y_{it}(0)]$$

Average Causal Response

$$ACRT(d_j) \equiv E\left[Y_{it}\left(d_j\right) - Y_{it}\left(d_{j-1}\right)\right]$$

$$ACRT(d) \equiv \frac{\partial E[Y_{it}(d)]}{\partial d}$$

Comment 2: The distinction matters!



Identification

Two Period Set Up

n units (i), 2 time periods ($t-1$ and t).

Assumption 1 (Random Sampling). *The observed data consists of* $\{Y_{it}, Y_{it-1}, D_i\}_{i=1}^{n}$ *which is independent and identically distributed.*

Assumption 2 (Support). *The support of* D, $D = \{0\} \cup \mathcal{D}_+$. *In addition,* $P(D = 0) > 0$ *and* $dF_D(d) > 0$ *for all* $d \in \mathcal{D}_+$. *No units are treated in period* $t-1$.

Assumption 3 (Observed Outcomes/No Anticipation). *For all units,*

$$Y_{it-1} = Y_{it-1}(0) \text{ and } Y_{it} = Y_{it}(D_i).$$

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 Ant.

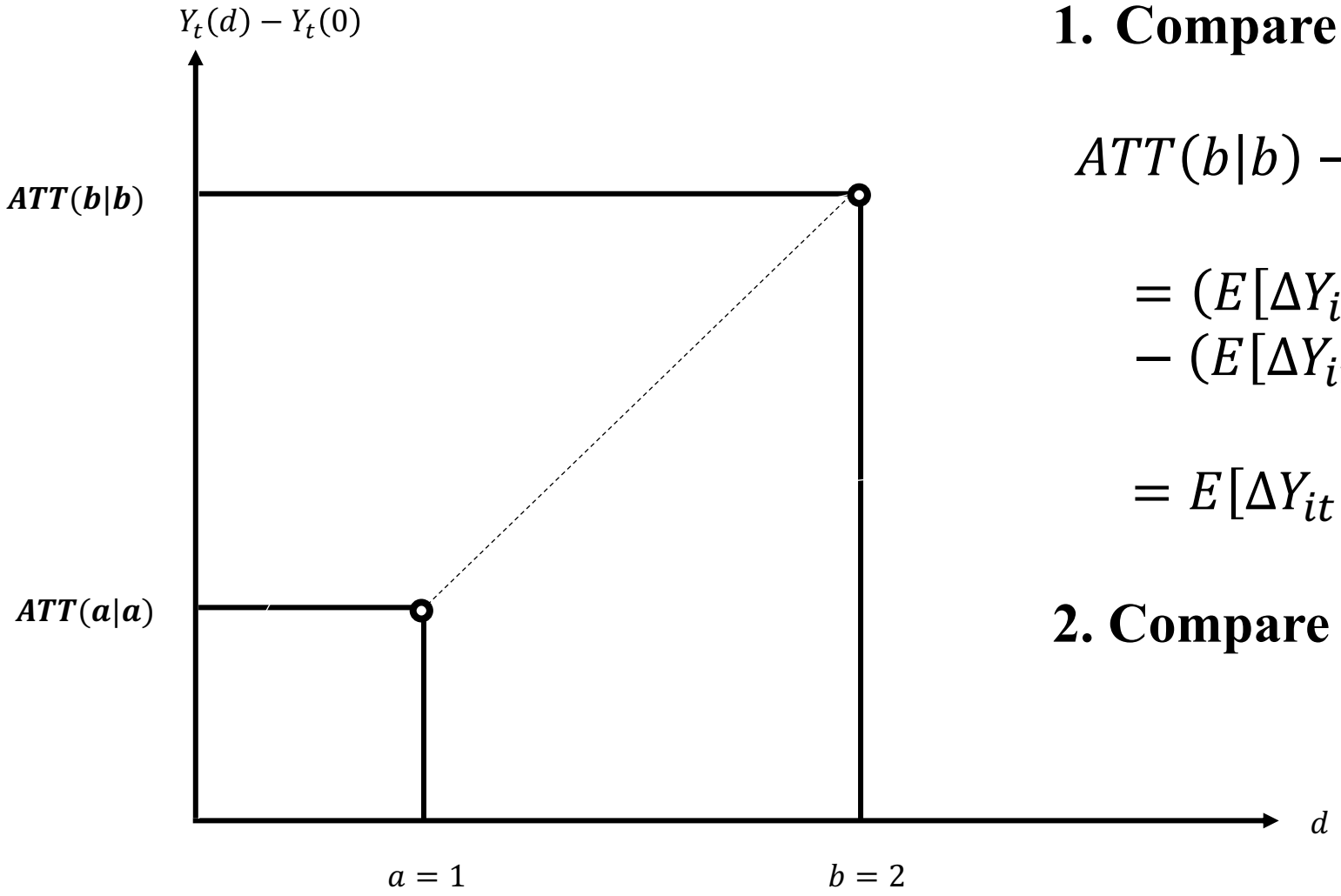
$$= E[Y_{it}(d) - Y_{it-1}(0)|D_i = d] - E[Y_{it}(d) - Y_{it-1}(0)|D_i = 0]$$

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

PT

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

Identifying ACRT



Two approaches:

1. Compare ATTs

$$ATT(b|b) - ATT(a|a) =$$

$$= (E[\Delta Y_{it}|D_i = a] - E[\Delta Y_{it}|D_i = 0]) - (E[\Delta Y_{it}|D_i = b] - E[\Delta Y_{it}|D_i = 0])$$

$$= E[\Delta Y_{it}|D_i = a] - E[\Delta Y_{it}|D_i = b]$$

2. Compare high-dose to low-dose

Identifying ACRT

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

(No Ant. + PT)

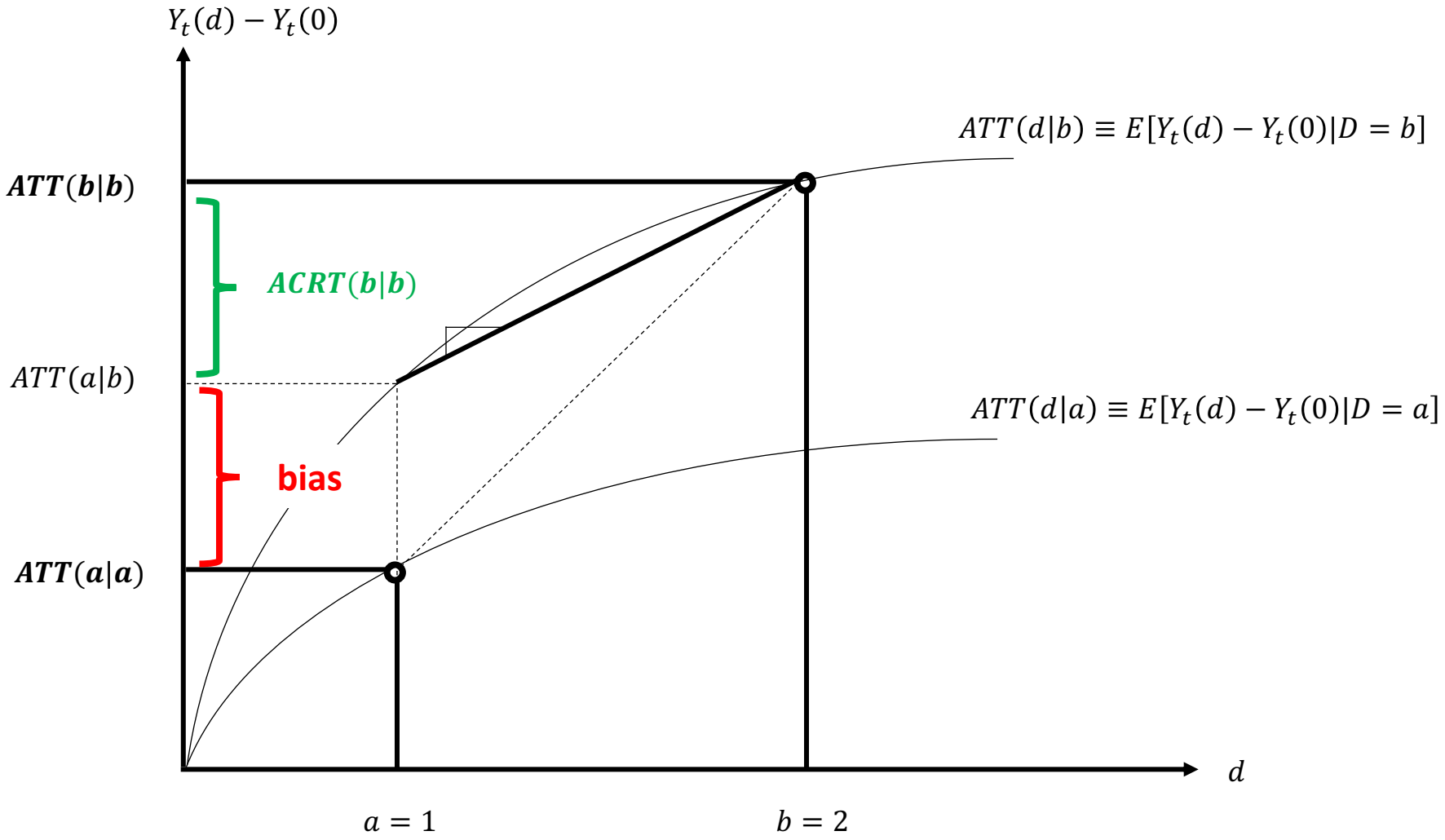
$$= \left(ATT(d_j|d_j) - ATT(d_{j-1}|d_j)\right) + \left(ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})\right)$$

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

Interpretable causal effect

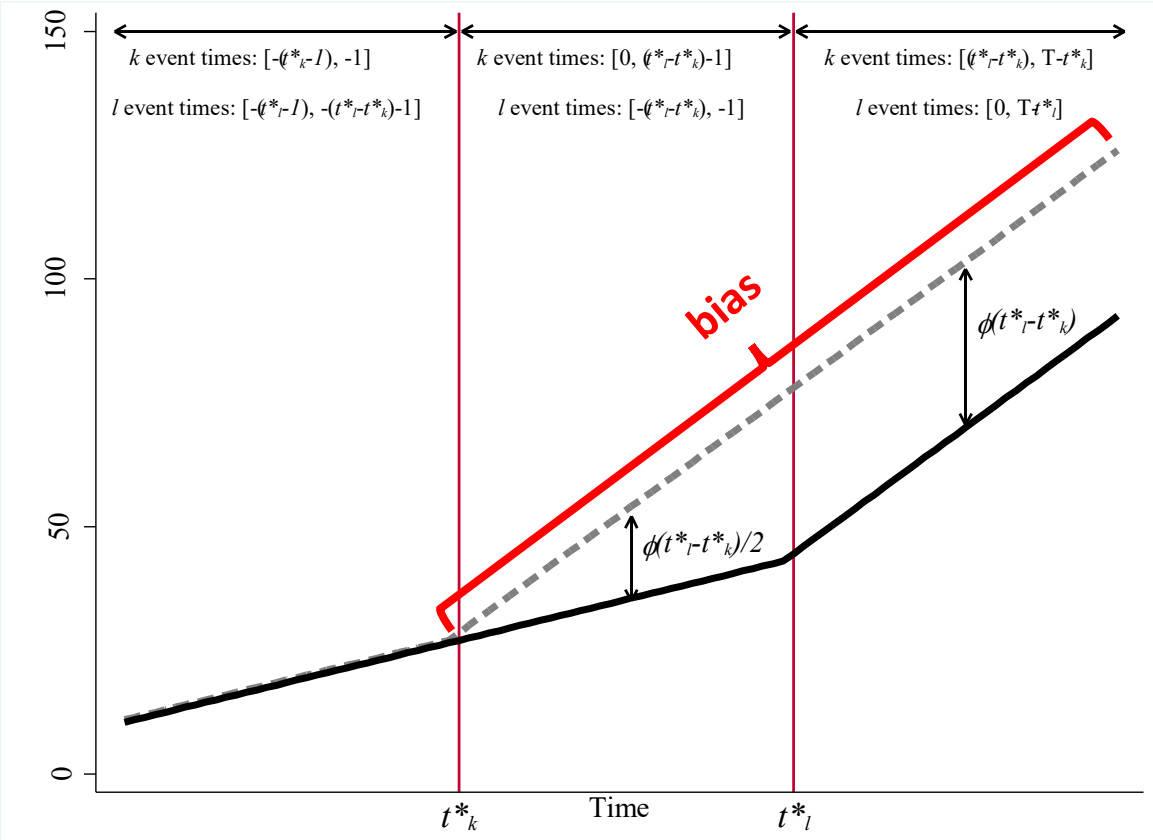
"Selection bias" (heterogeneous effects at d_{j-1})

Identifying ACRT

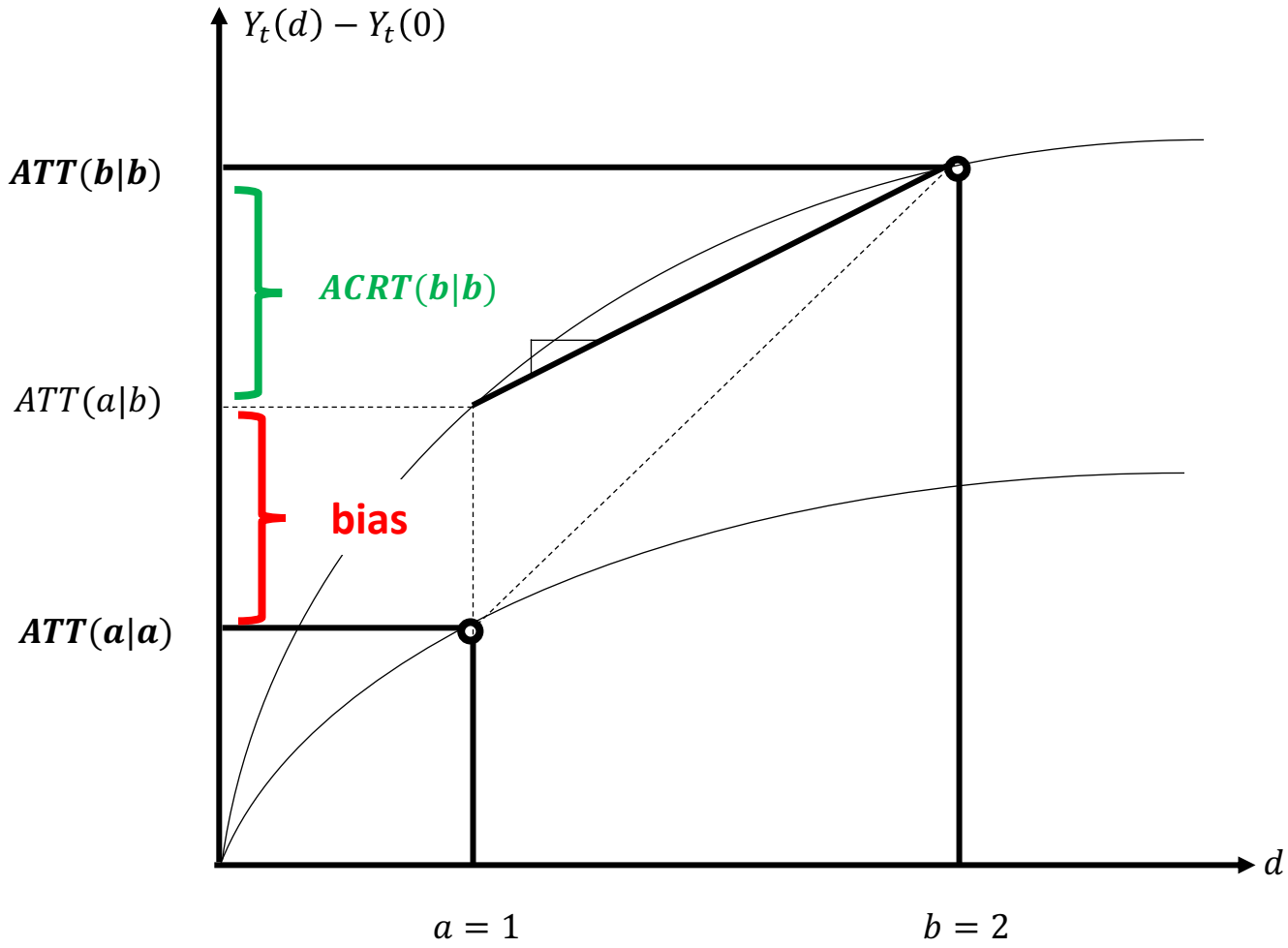


Treated comparison groups are the problem

a. Staggered Timing, Binary Treatment



b. One Treatment Time, Non-Binary Treatment



Strong Parallel Trends

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

$$\underbrace{E[Y_{it}(d) - Y_{it-1}(0)]}_{\text{Average change in PO from 0 to } d \text{ in the whole sample}} = \underbrace{E[Y_{it}(d) - Y_{it-1}(0)|D_i = d]}_{\text{Change in PO from 0 to } d \text{ for the } d \text{ group}}$$

Stronger but more intuitive assumption

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 do they change for treatment effect reasons at a given dose

We never say this

- Jackson, Johnson, and Persico (2019):

> *There are two identifying assumptions. First, counties that experienced larger or smaller increases in Head Start spending over time were not already on a trajectory of improving or deteriorating outcomes over time. Second, counties that saw larger or smaller increases in Head Start spending did not also undergo other unobserved changes that would also affect outcomes.*

- Me and Lucie Schmidt (2020):

> *Internal validity of our design requires that no other important determinants of changing safety net outcomes correlate with the level of APTD generosity in the specific way that SSI did.*

…almost never

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

Gabriel Chodorow-Reich

Plamen T. Nenov

Alp Simsek

AMERICAN ECONOMIC REVIEW
VOL. 111, NO. 5, MAY 2021
(pp. 1613-57)

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.

Comment 1: SPT→ATE and ACR

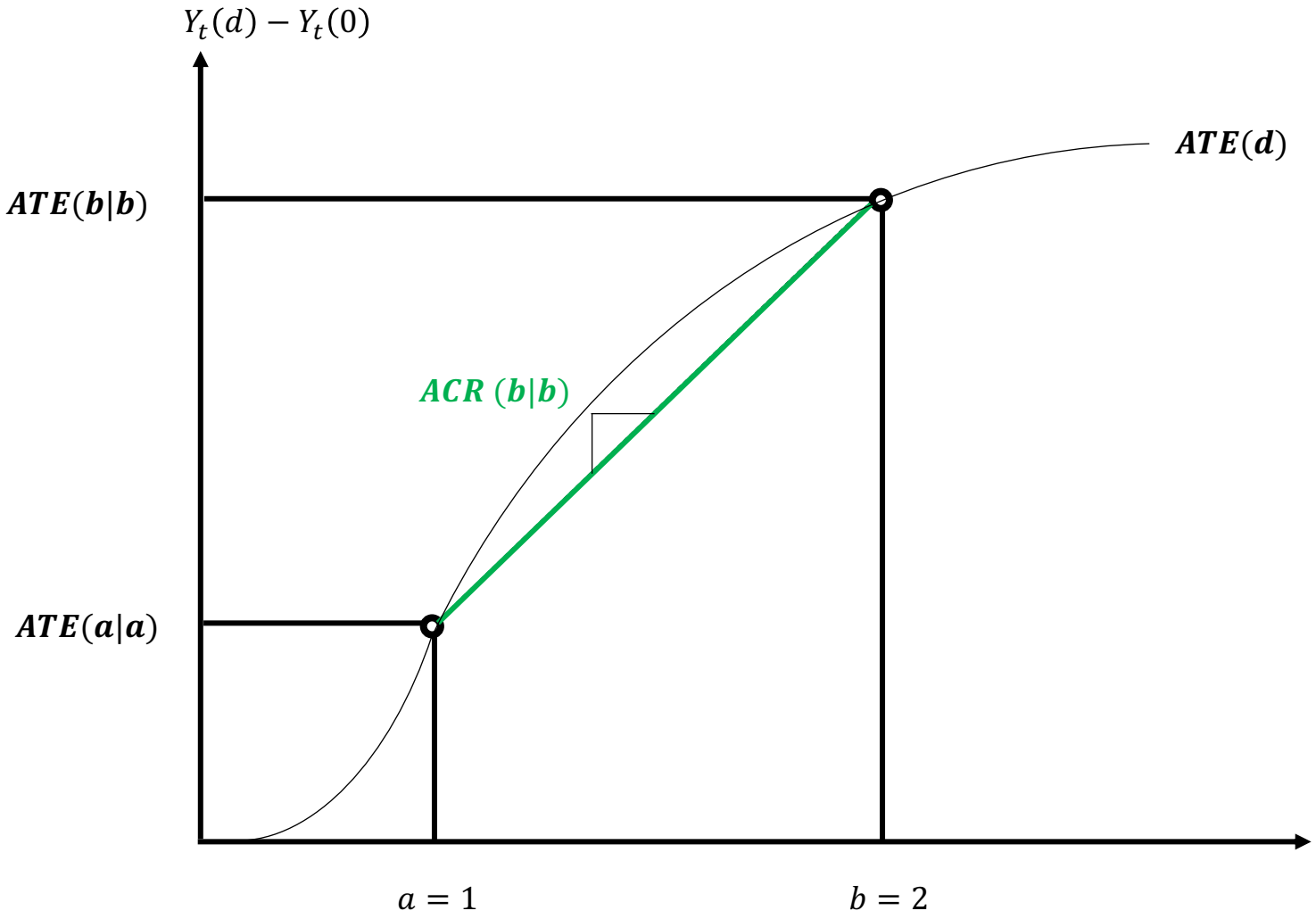
Under PT + homogeneity:

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

And

$$ATE(d_j) - ATE(d_{j-1}) = E[\Delta Y_{it}|D_i = d_j] - E[\Delta Y_{it}|D_i = d_{j-1}] = \boldsymbol{ACR}(\boldsymbol{d})$$

Identifying ACR



Good:

Shows what assumption you need to ID ACR.

Bad:

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

inching closer to IV independence assumption/quasi-randomness

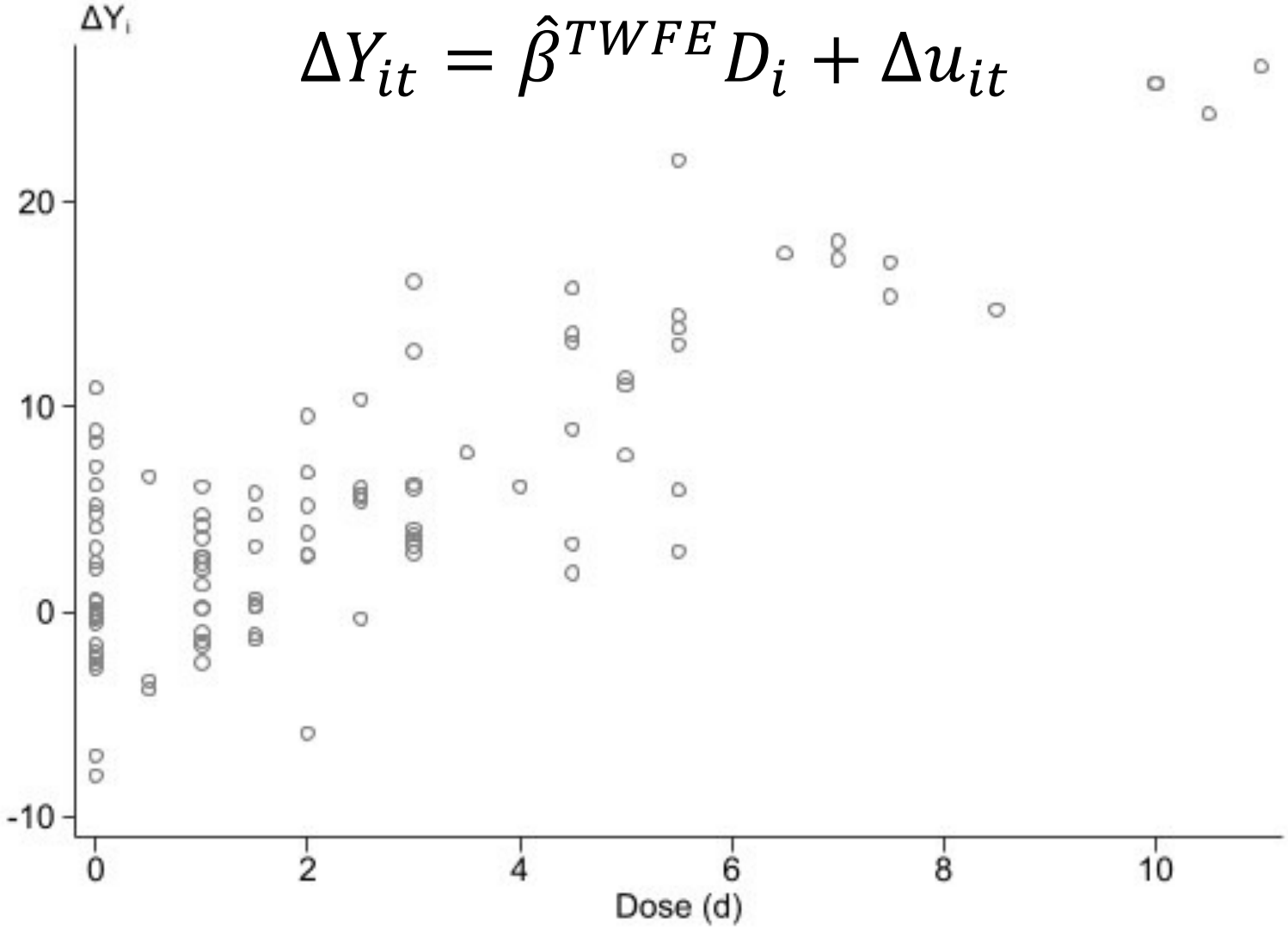
TWFE: 2 Periods

Two-way fixed effects

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

Two-way fixed effects

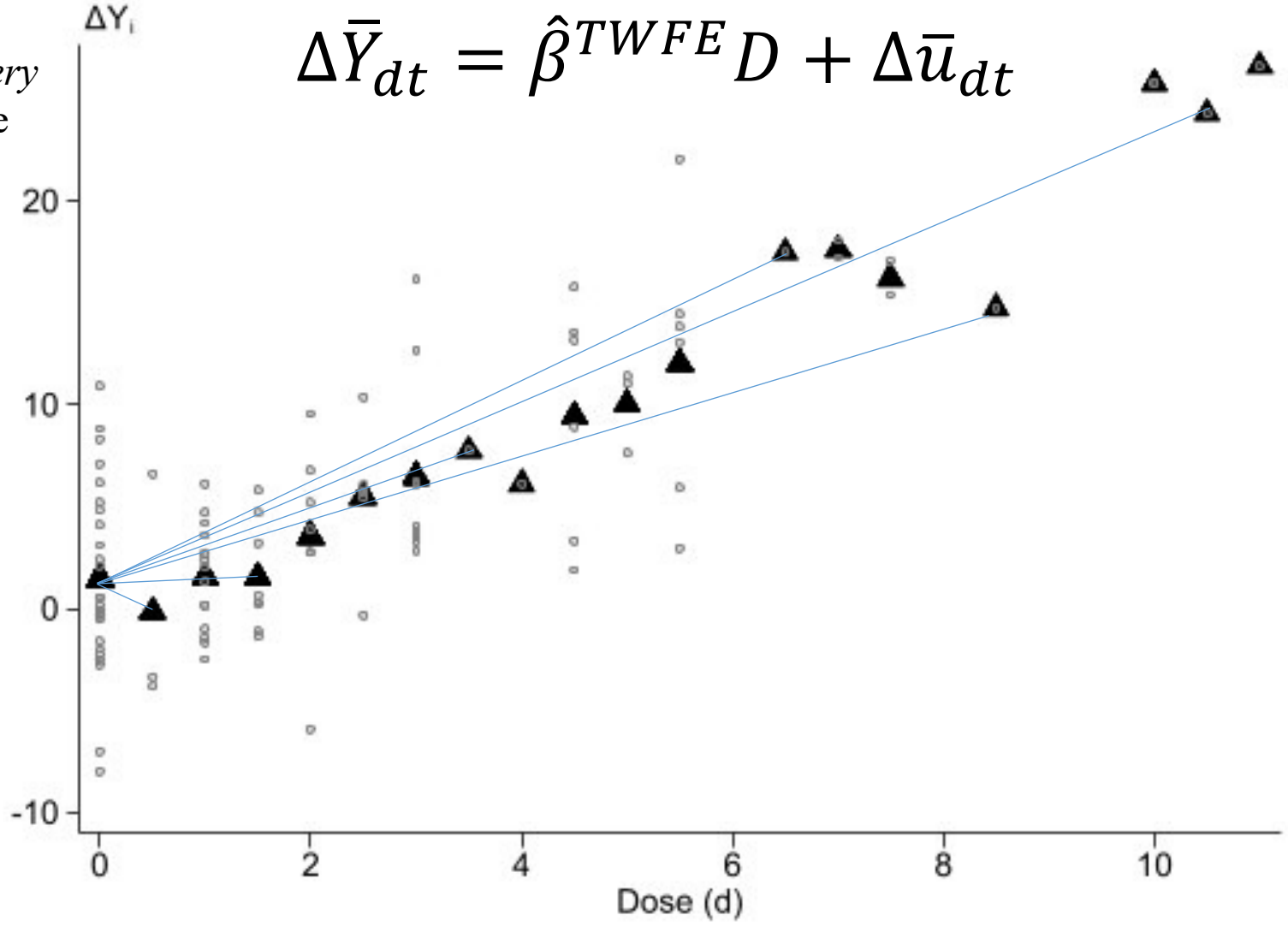
$$\Delta Y_{it} = \hat{\beta}^{TWFE} D_i + \Delta u_{it}$$



Two-way fixed effects

$\hat{\beta}^{TWFE}$ = average of *every* pairwise slope

$$\Delta\bar{Y}_{dt} = \hat{\beta}^{TWFE} D + \Delta\bar{u}_{dt}$$

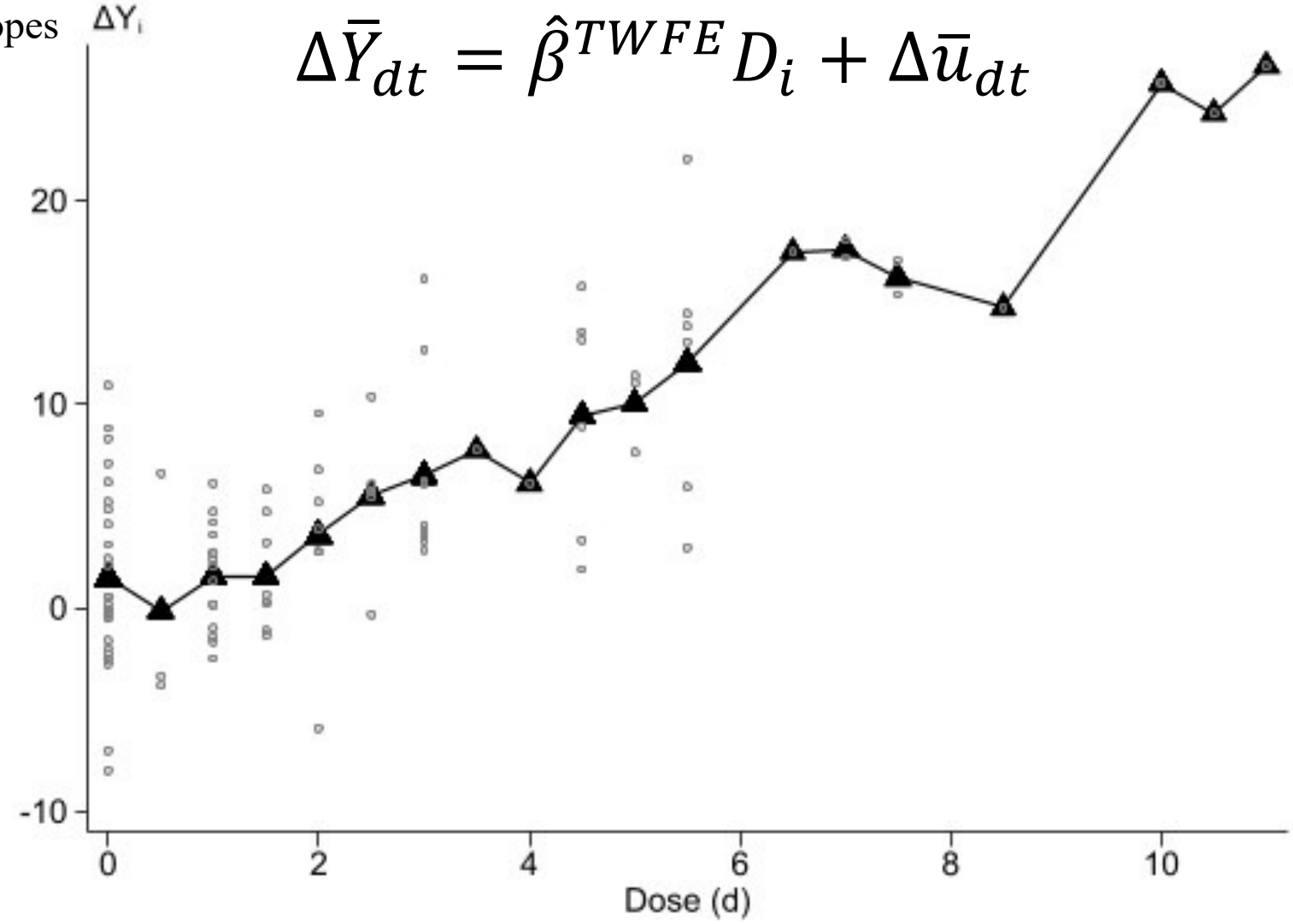


Two-way fixed effects

Easier (Yitzhaki 1996):

$\hat{\beta}^{TWFE}$ = weighted average of *adjacent* slopes

$$\Delta\bar{Y}_{dt} = \hat{\beta}^{TWFE} D_i + \Delta\bar{u}_{dt}$$



Two-way fixed effects, discrete

Adjacent slopes: "Wald DDs"

$$\beta^{TWFE} = \sum_{d_j \in \mathcal{D}_+} w_1(d_j) \frac{E\left[\Delta Y_t \middle| D = d_j\right] - E\left[\Delta Y_t \middle| D = d_{j-1}\right]}{d_j - d_{j-1}}$$

Where:

$$w_1(d_j) = \frac{\left(E\left[D \middle| D > d_j\right] - E[D]\right) P\left(D > d_j\right)}{V(D)}$$

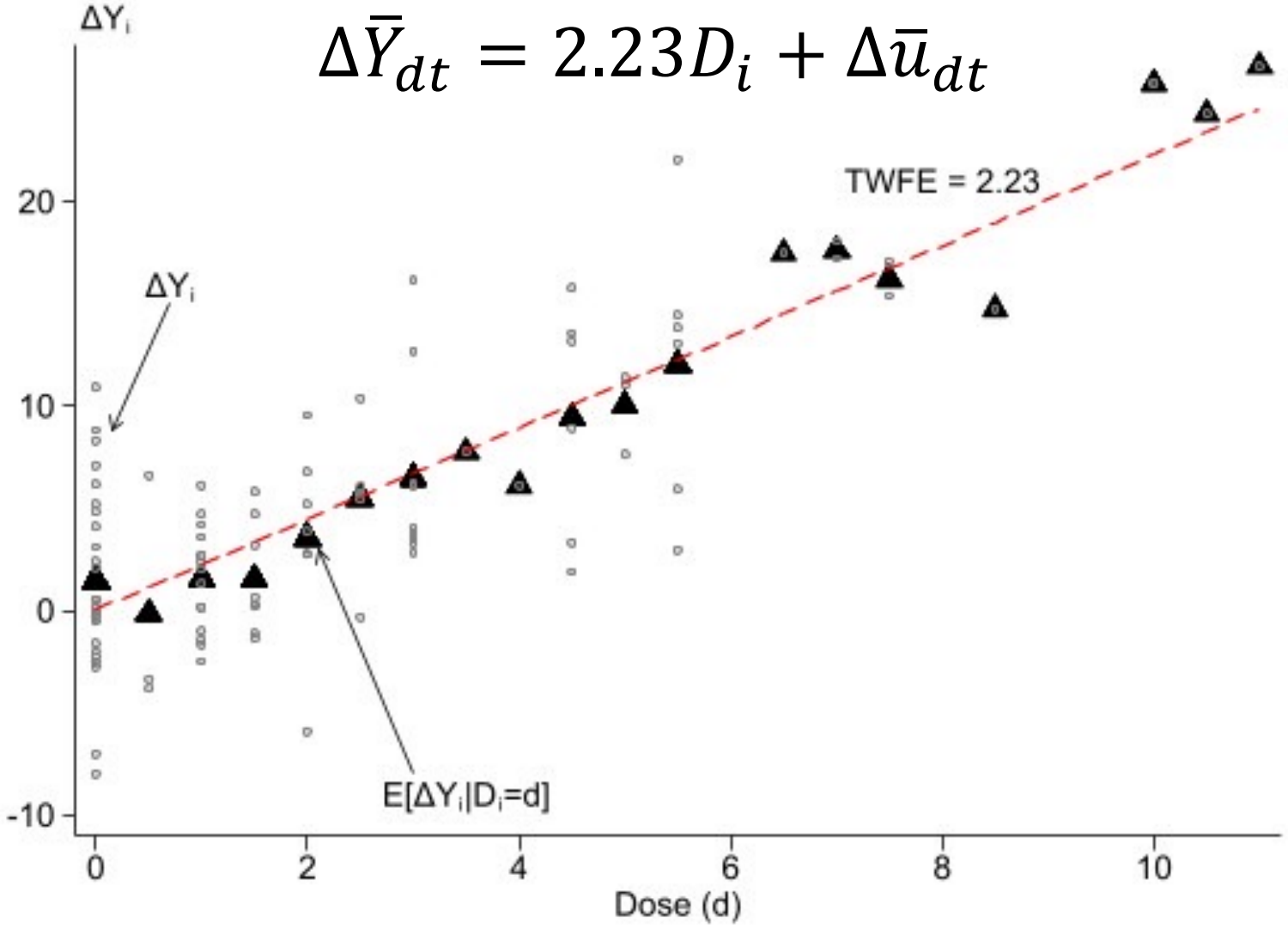
and that are hump-shaped and centered on E[D]

And $w_1(d_j) \geq 0$ and $\sum_{d_j \in \mathcal{D}_+} w_1(d_j) = 1.$

Weights are non-negative and sum to one.

Two-way fixed effects

$$\Delta\bar{Y}_{dt} = 2.23D_i + \Delta\bar{u}_{dt}$$



How should we interpret 2.23?

$$2.23 = \sum_{d_j \in \mathcal{D}_+} w_1(d_j)\left[ACRT(d_j|d_j) + \left(ACR(d_{j-1}|d_j) - ACR(d_{j-1}|d_{j-1})\right)\right]$$

$$= \sum_{d_j \in \mathcal{D}_+} w_1(d_j) ACRT(d_j|d_j) + \sum_{d_j \in \mathcal{D}_+} w_1(d_j)\left(ACR(d_{j-1}|d_j) - ACR(d_{j-1}|d_{j-1})\right)$$

How should we interpret 2.23?

$$2.23 = \sum_{d_j \in \mathcal{D}_+} w_1(d_j)\left[ACRT(d_j|d_j) + \left(ACR(d_{j-1}|d_j) - ACR(d_{j-1}|d_{j-1})\right)\right]$$

$$= \sum_{d_j \in \mathcal{D}_+} w_1(d_j) ACRT(d_j|d_j) + \sum_{d_j \in \mathcal{D}_+} w_1(d_j)\left(ACR(d_{j-1}|d_j) - ACR(d_{j-1}|d_{j-1})\right)$$

Problem 1:

combines a causal parameter with treatment effect heterogeneity.

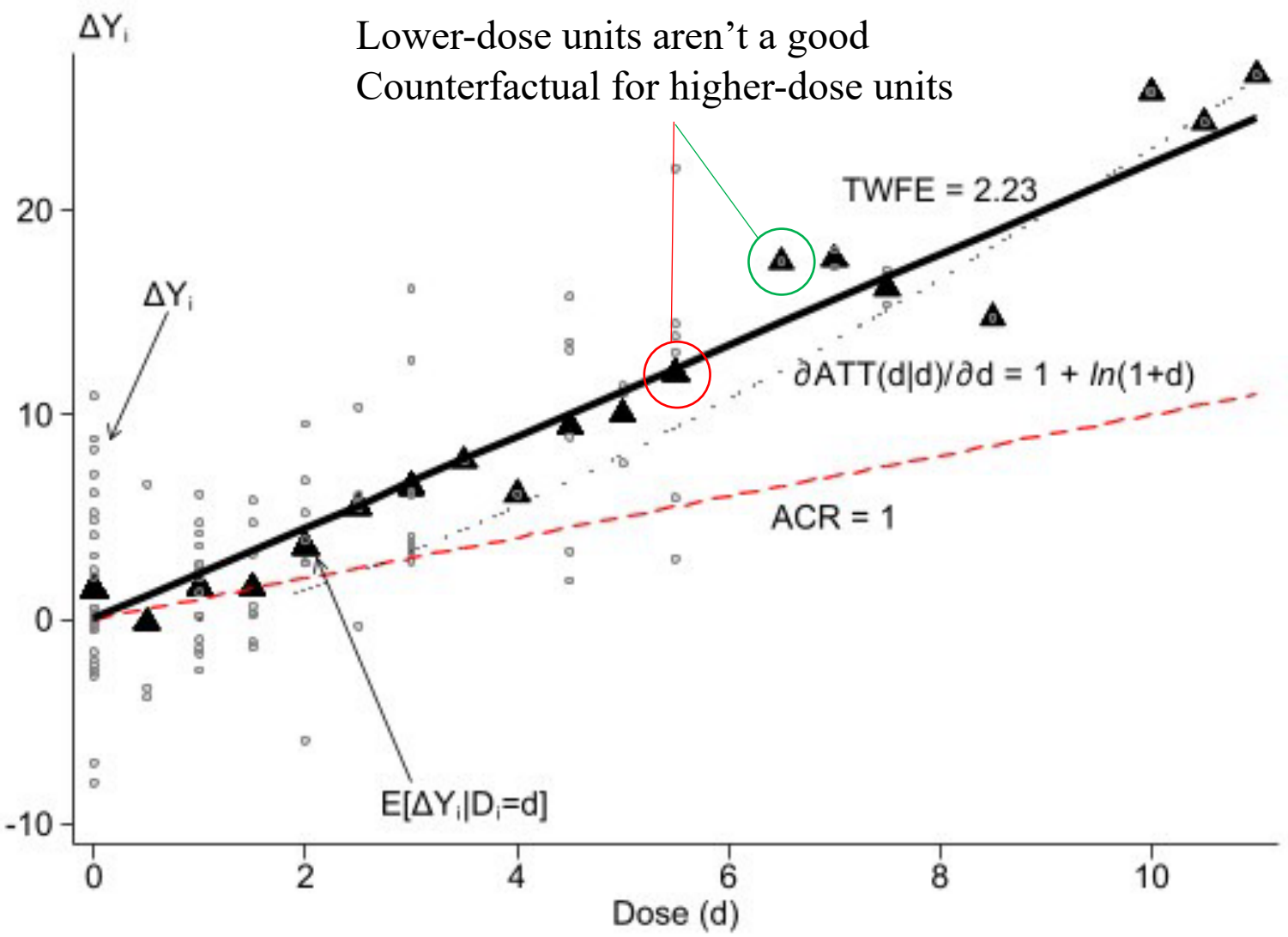
TWFE Bias

$ATT(D|d)$ in this example is:

$$\underbrace{(1+d_i)}_{\text{Shifted by the amount of each group's actual dose}}\underbrace{\ln(1+D_i)}_{\text{Concave ATT function for each group}}$$

cov(heterogeneity,dose)>0

→ $\boldsymbol{ACRT(d|d)=1}$ **for everyone**



Hard to Fix


Staggered timing fix is to use
an untreated control group
You can’t get ACRT with
an untreated control group
ASSUMPTIONS

How should we interpret 2.23?

$$2.23 = \sum_{d_j \in \mathcal{D}_+} w_1(d_j)\left[ACRT(d_j|d_j) + \left(ACR(d_{j-1}|d_j) - ACR(d_{j-1}|d_{j-1})\right)\right]$$

$$= \sum_{d_j \in \mathcal{D}_+} w_1(d_j) ACRT(d_j|d_j) + \sum_{d_j \in \mathcal{D}_+} w_1(d_j)\left(ACR(d_{j-1}|d_j) - ACR(d_{j-1}|d_{j-1})\right)$$

=0 under SPT

Problem 2:

Even in the causal estimand, the weights are probably not what you want (OLS chooses them)

TWFE Weighting

Let $ACRT^*$ be an average of $ACRT(d|d)$ weighted by $f_{D|D>0}(d)$

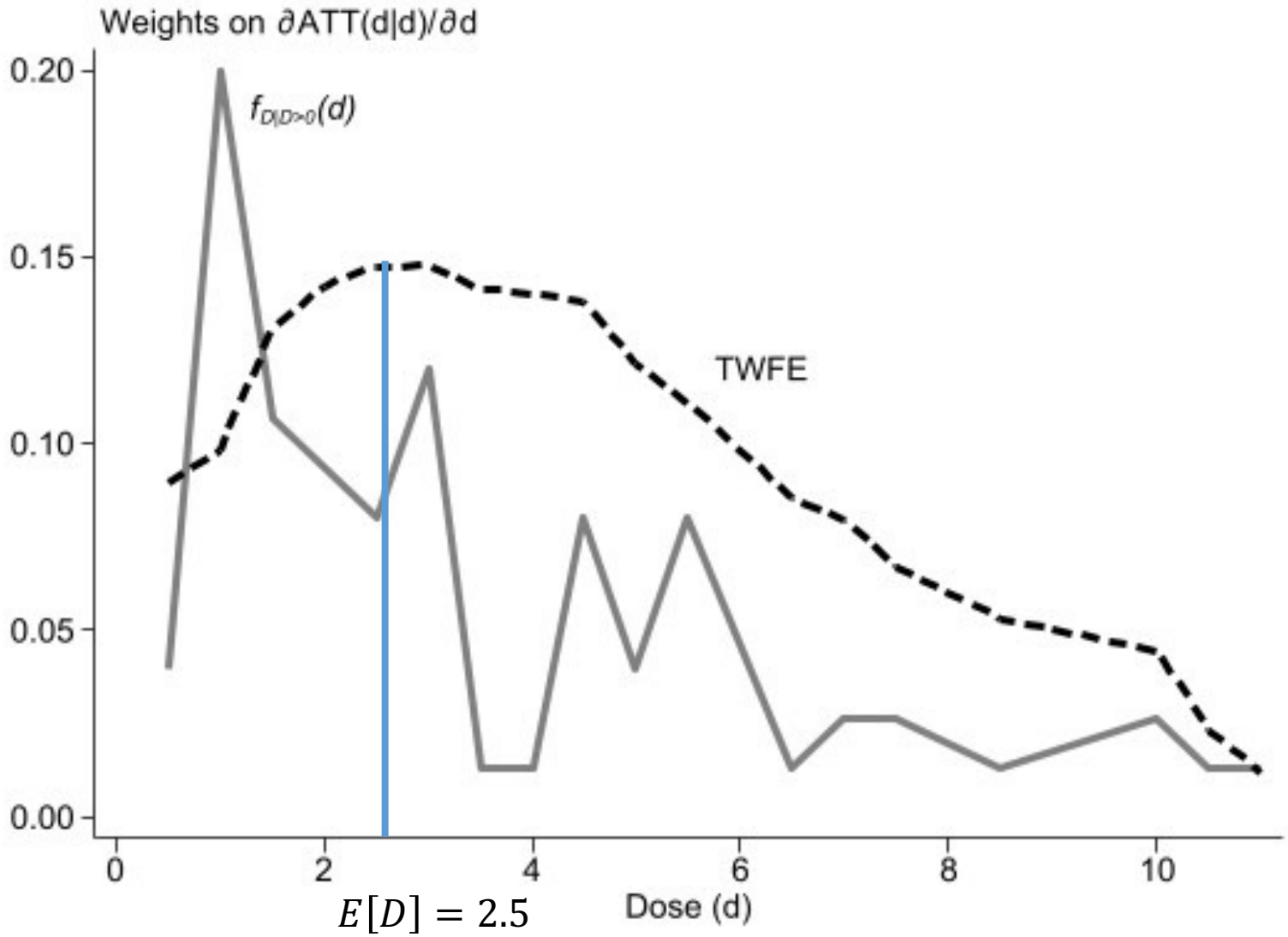
Here $f_D = \exp(\lambda)$ w/ $\lambda = \frac{1}{3}$

TWFE weights are:

$$w_1(d) = \lambda d f_D(d)$$

Difference in weighting is:

$$f_D(d) - w_1(d) = f_{D(d)}\lambda\left(\frac{1}{\lambda} - d\right)$$



TWFE Weighting

Exponential dose:

TWFE under-weights the lowest doses over-weights the highest doses.

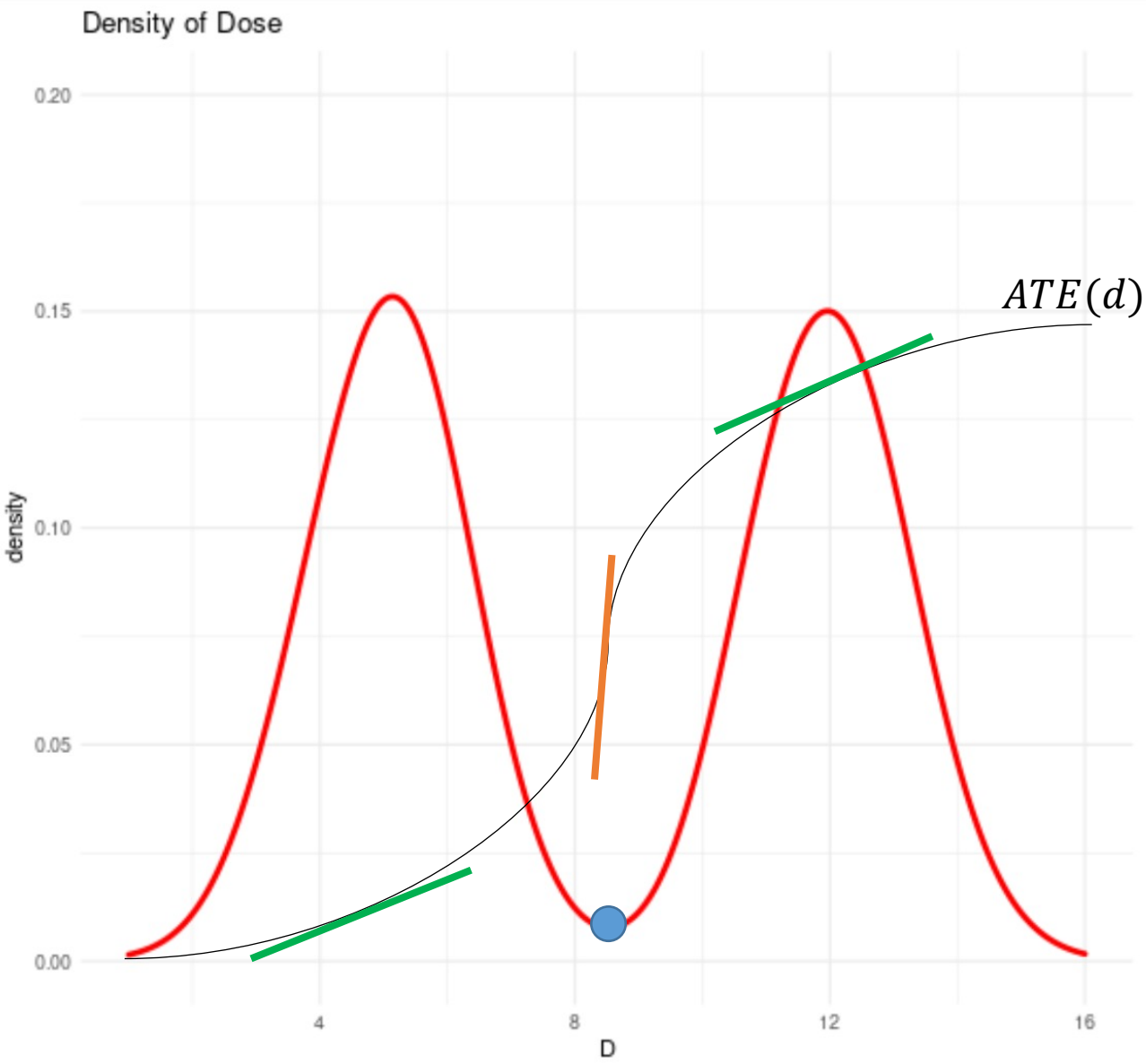
Uniform dose:

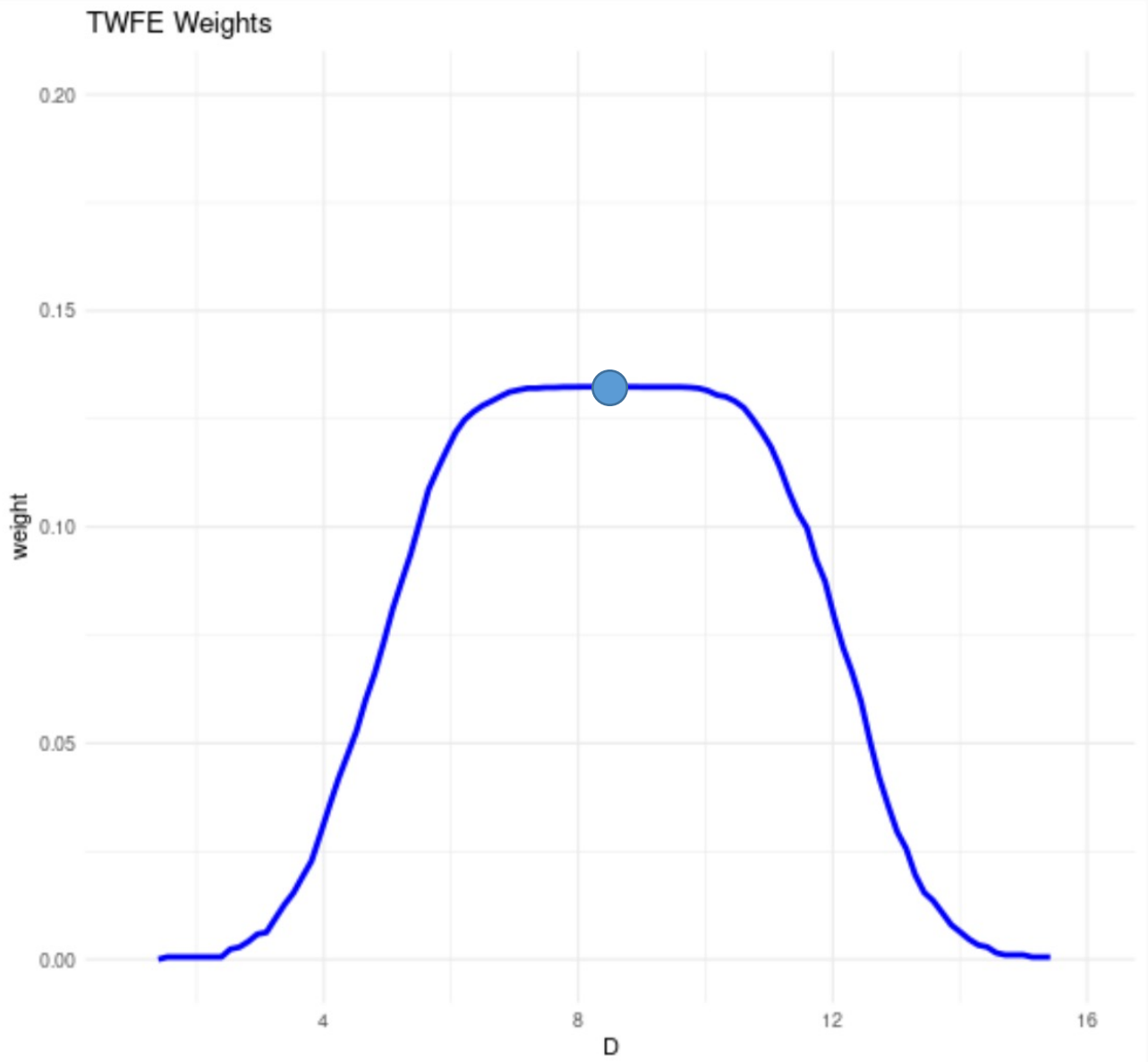
TWFE underweights the highest and lowest evenly.

Normal dose:

TWFE weights equal $f_D(d)$!

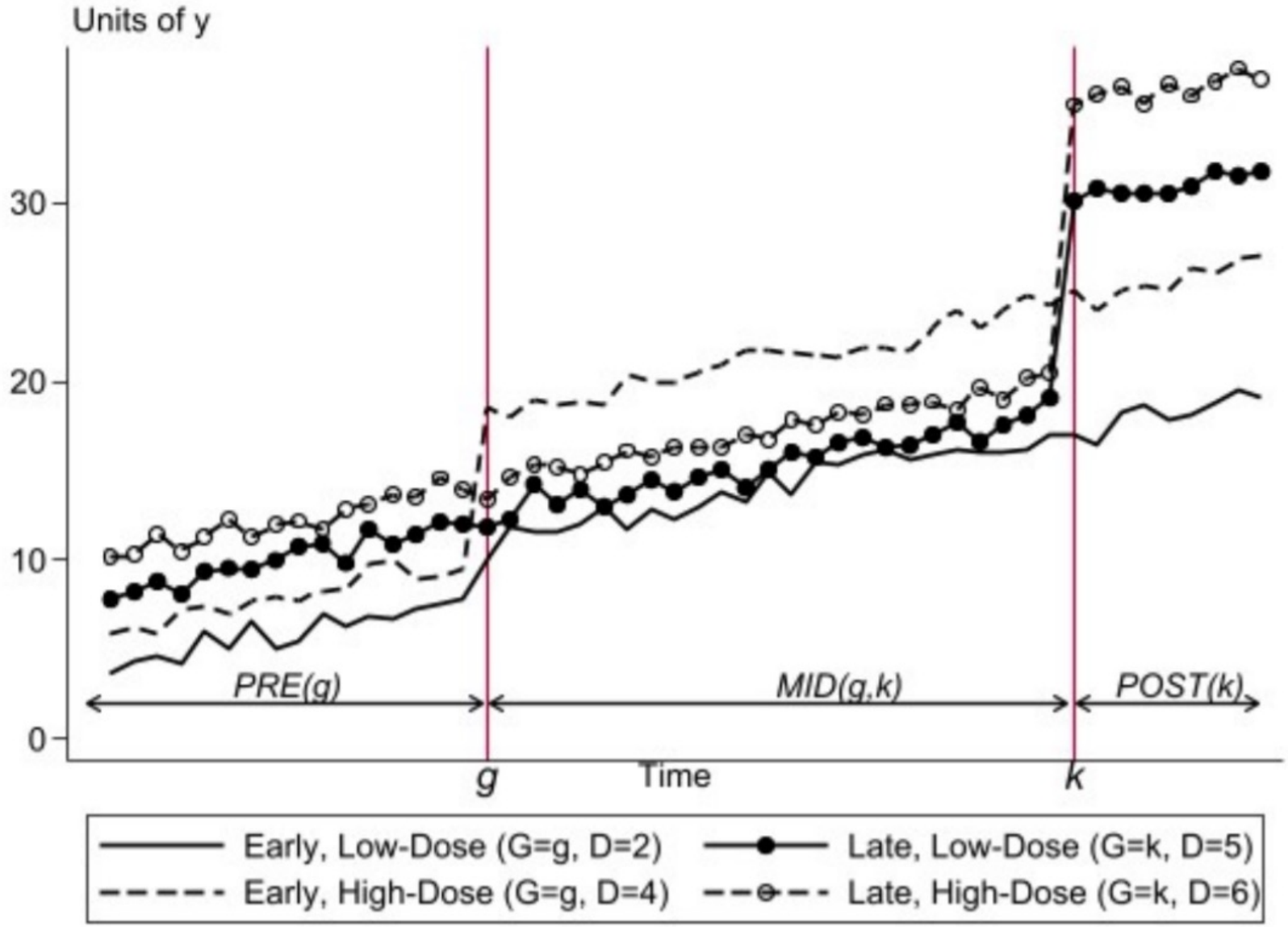
TWFE Weighting





TWFE: Multiple Periods

Rough Sketch



Units are defined by *when* they are treated (G) and *how much* they are treated (D).

Treatment effects are defined by dose, timing, and time period:

$$ATT(g,t,d|g,d)$$

$$ACRT(g,t,d|g,d)$$

Rough Sketch

$$Y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{TWFE} D_i \mathbf{1}\{t \geq G_i\} + u_{it}$$

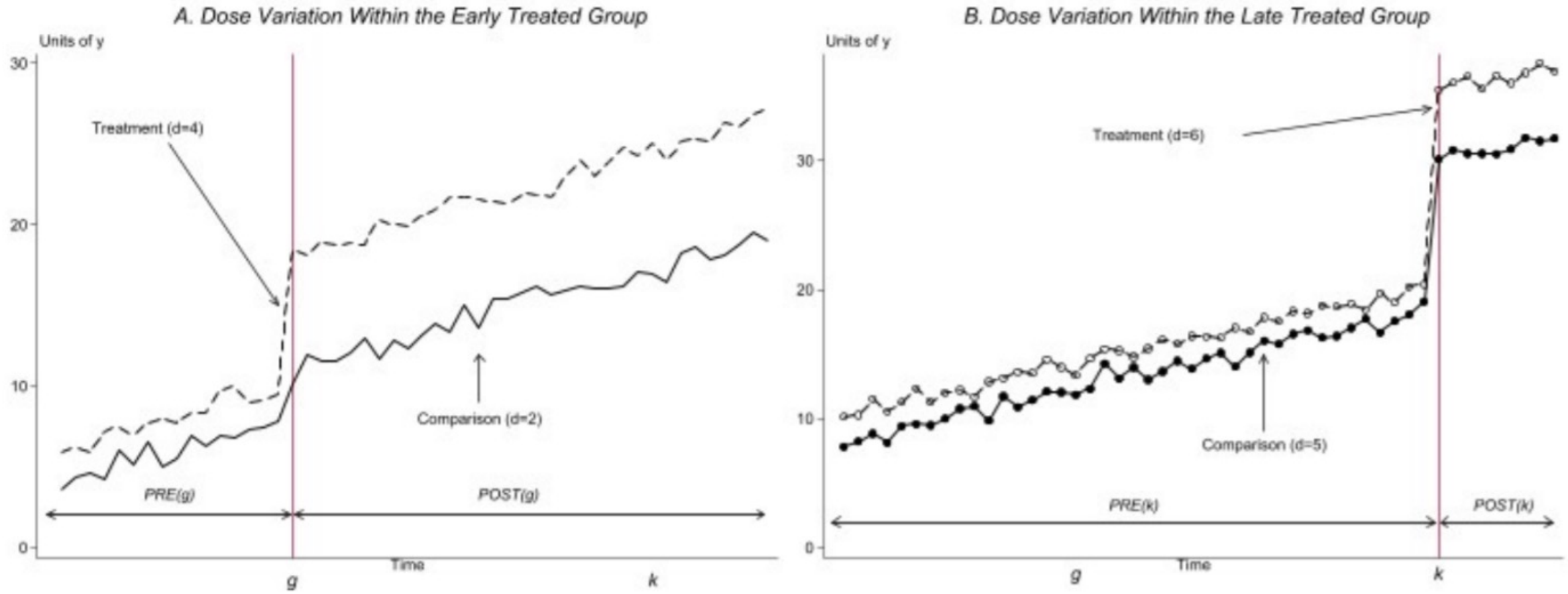
Varies across units with different doses

and

across units treated at different times.

Our approach: write $\hat{\beta}^{TWFE}$ as a weighted average of terms that only use *one* source of variation at a time.

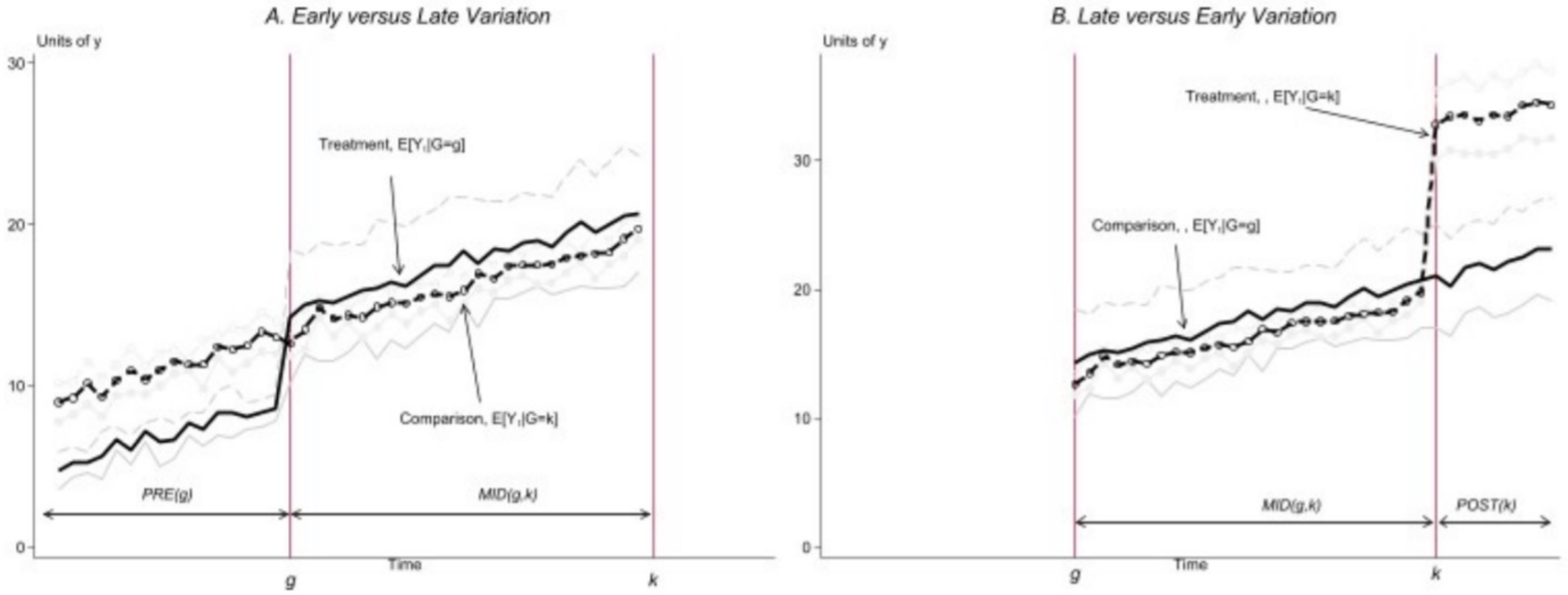
Dose Variation within Timing Group



This is exactly what we've been talking about.

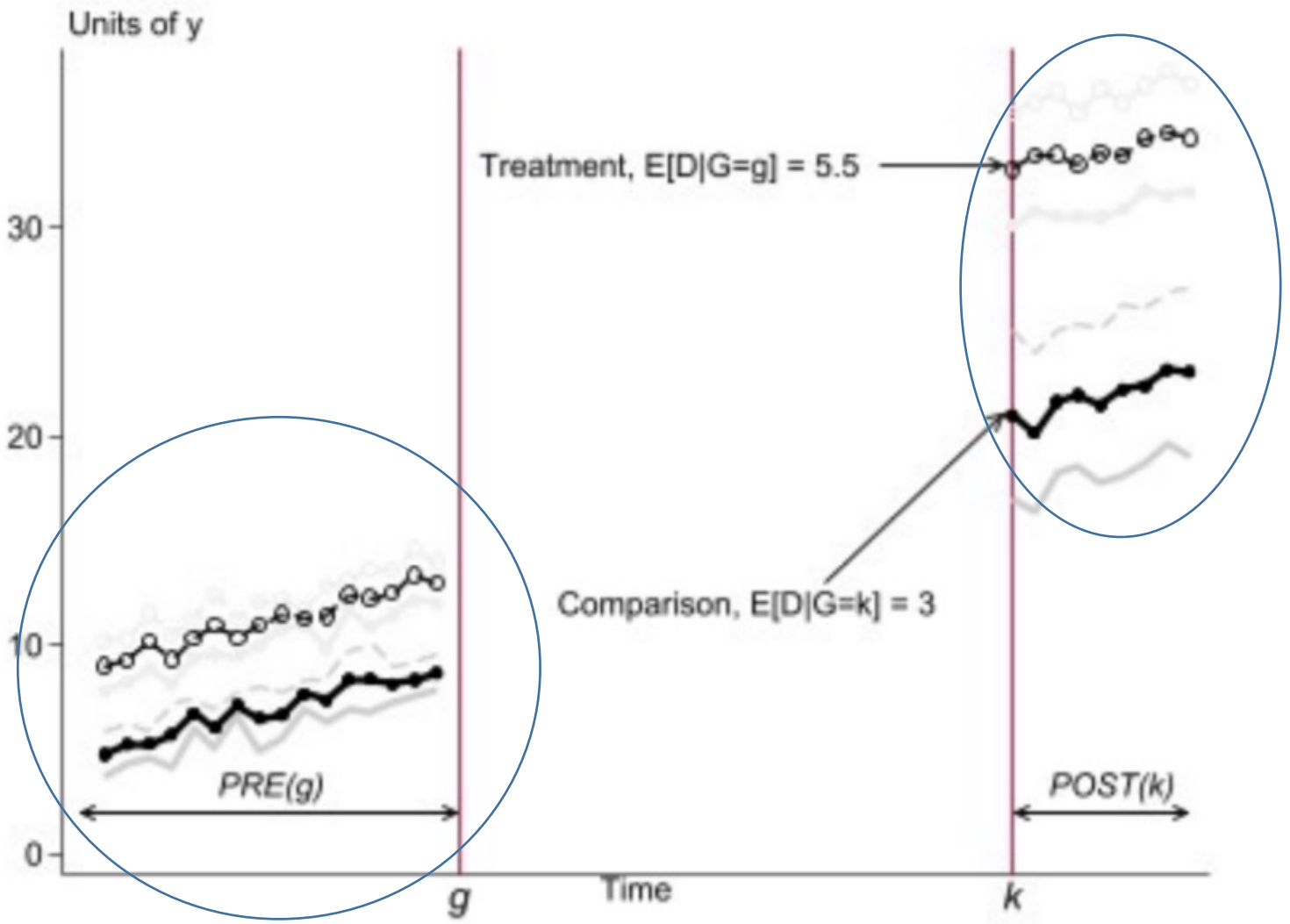
These yield ACR parameters plus bias from heterogeneous ATT functions.

Timing Variation



This is Goodman-Bacon (2021) and what people have been talking about for the last three years. These yield ATT parameters and the second one is biased by time-varying ATTs.

Dose Variation on Long Differences



Once both timing groups are treated they may be treated with different intensities.

Compare the higher to the lower intensity group here, subtract off the pre-period diff when neither group is treated.

This is biased by *any* kind of cross-group heterogeneity!

Verdict on Staggered + Continuous TWFE

$$Y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{TWFE} D_i \mathbf{1}\{t \geq G_i\} + u_{it}$$

- TWFE in this context is biased by:

  Heterogeneous ATT functions across groups

  Any kind of time-varying ATT

- It conflates ATT parameters and ACR parameters

- It has the typical OLS weighting issues.

Upshot

When you compare to treated units, treatment effect heterogeneity can be a source of bias.

This is inherent in continuous designs.

There is no obvious estimator that gets back to a PT world.

Understand the stronger assumptions you need and justify them.

Thank you!

…and I’m sorry.

andrew@goodman-bacon.com

DD Revolution

2x2 DD (Snow 1855) →
OLS 2x2 DD (~Ashenfelter and Card 1985) →
TWFE staggered timing (Jacobson, LaLonde, and Sullivan 1993) →
IPW (Abadie 2005)/**distributions** (Athey and Imbens 2006)

"old" stuff

LaPorte and Windmeijer (2004)
Wooldridge (2005)
Meer and West (2016)

the trickle

Borusyak and Jaravel (2017), Strezhnev (2019),
Athey and Imbens (2019), Imai and Kim (2020),
de Chaisemartin and D'Haultfouielle (2020),
Goodman-Bacon (2021), Callaway and Sant'Anna (2020),
Gardner (2021), Borusyak, Jaravel and Speiss (2021)

THE FLOOD

DD Revolution

2x2 DD (Snow 1855) →
OLS 2x2 DD (~Ashenfelter and Card 1985) →
TWFE staggered timing (Jacobson, LaLonde, and Sullivan 1993) →
IPW (Abadie 2005)/**distributions** (Athey and Imbens 2006)

} "old" stuff

LaPorte and Windmeijer (2004)
Wooldridge (2005)
Meer and West (2016)

} the trickle

TWFE bias with staggered timing + time-varying treatment effects

Borusyak and Jaravel (2017), Strezhnev (2019),
Athey and Imbens (2019), Imai and Kim (2020),
Sun and Abraham (2020) de Chaisemartin and
D'Haultfouielle (2020), Goodman-Bacon (2021),
Callaway and Sant'Anna (2020),
Gardner (2021), Borusyak, Jaravel and Speiss (2021)

} THE FLOOD

DD Revolution

2x2 DD (Snow 1855) →
OLS 2x2 DD (~Ashenfelter and Card 1985) →
TWFE staggered timing (Jacobson, LaLonde, and Sullivan 1993) →
IPW (Abadie 2005)/**distributions** (Athey and Imbens 2006)

} "old" stuff

LaPorte and Windmeijer (2004)
Wooldridge (2005)
Meer and West (2016)

} the trickle

Borusyak and Jaravel (2017), Strezhnev (2019),
Athey and Imbens (2019), Imai and Kim (2020),
Sun and Abraham (2020) de Chaisemartin and
D'Haultfouielle (2020), Goodman-Bacon (2021),
Callaway and Sant'Anna (2020),
Gardner (2021), Borusyak, Jaravel and Speiss (2021)

} THE FLOOD

Alternative estimators to OLS that avoid the bias

Two-way fixed effects, continuous

$$\beta^{TWFE} = \int_{d_L}^{d_U} w_1(l) \frac{\partial E[\Delta Y_t | D = l]}{\partial l} dl \; + w_0 \frac{\left[E[\Delta Y_t | D = d_L] - E[\Delta Y_t | D = 0]\right]}{d_L}$$

Where:

$$w_1(l) = \frac{(E[D|D > l] - E[D])P(D > l)}{V(D)}$$

$$w_0 = \frac{(E[D|D > 0] - E[D])P(D > 0)d_L}{V(D)}$$

And $w_1(l) \geq 0 \; \forall l \in \mathcal{D}$, $w_0 > 0$, and $\int_{d_L}^{d_U} w_1(l)\, dl + w_0 = 1$.