Causal Inference

Lecture #2

Juraj Medzihorsky

SGIA & RMC
Durham University

23 January 2024

Common Misunderstandings about Effects

Unlearning Bad Habits

Effects

Many different effects, today we will work with these again

Average Treatment Effect

$$ATE = \delta = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

Average Treatment Effect on the Treated

$$ATT = \mathbb{E}[Y(1) - Y(0)|D = 1] = \mathbb{E}[Y(1)|D = 1] - \mathbb{E}[Y(0)|D = 1]$$

Individual Treatment Effect

$$\tau_i = Y_i(1) - Y_i(0)$$

Many different ITE distributions are consistent with the same ATE. Any set of $Y(0)$ and $Y(1)$ marginals can be combined with any pattern of association between $Y(0)$ and $Y(1)$.

Example:

	$Y(0)=0$	$Y(0)=1$	
$Y(1)=1$	9	11	20
$Y(1)=0$	1	19	20
	10	30	

	$Y(0)=0$	$Y(0)=1$	
$Y(1)=1$	2	18	20
$Y(1)=0$	8	12	20
	10	30	

The ATE is $50\% - 75\% = -25\%$, but in the first case 9 units gain and 19 units lose, in the other 2 units gain and 12 units lose.

Even if the average effect is negative, it can still positively affect some units, and vice vera.

Another example:

	$Y(0)=0$	$Y(0)=1$	
$Y(1)=1$	0	25	25
$Y(1)=0$	25	0	25
	25	25	

	$Y(0)=0$	$Y(0)=1$	
$Y(1)=1$	20	5	25
$Y(1)=0$	5	20	25
	25	25	

The ATE is $50\% - 50\% = 0\%$, but in the first case no units gain or lose, and in the other 40% loses and 40% gains.

Just because the treatment has no effect on average, it doesn't mean it has no effects whatsoever.

Neyman's null: $\mathbb{E}[\tau_i] = 0$

Fisher's null: $\forall i \ \tau_i = 0$

In practice, these can be very different scenarios. Knowing the ATE or ATT may not be enough to decide.

Selection on Observables

- Suppose

$$D \longrightarrow Y$$
$$\{Y(0), Y(1)\} \not\perp\!\!\!\perp D \text{ i.e. } \mathbb{P}[D = d|Y(0), Y(1)] \neq \mathbb{P}[D = d]$$

that is, there is selection into treatment, and also

$$\boldsymbol{X} \longrightarrow D$$
$$\boldsymbol{X} \longrightarrow Y$$

so

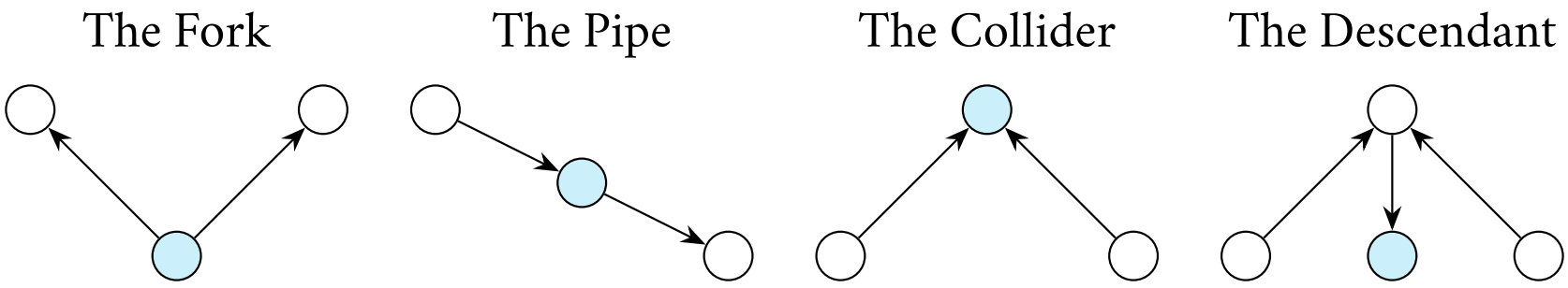
$$\{Y(0), Y(1)\} \perp\!\!\!\perp D|\boldsymbol{X}$$

- The big problem: We have to be able to observe and measure all of the X variables that affect selection into treatment and variation in outcomes.

Identification Assumptions

1. *Selection on Observables*
   - *ATE Version:* $\{Y(1), Y(0)\} \perp\!\!\!\perp D|X$
     - *There exists a set of observable covariates, X, such that after controlling for X, potential outcomes are independent of treatment status*
   - *ATT Version:* $Y(0) \perp\!\!\!\perp D|X$
     - *There exists a set of observable covariates, X, such that after controlling for X, counterfactual outcomes for treated units and observed outcomes for untreated units are independent of treatment status*
2. *Common Support*
   - *ATE Version:* $0 < \mathbb{P}[D = 1|X = x] < 1$
     - *For each value of X, there is a positive probability of observing both treated and untreated units*
   - *ATT Version:* $\mathbb{P}[D = 1|X = x] < 1$, *with* $\mathbb{P}[D = 1] > 0$
     - *For each value of X observed for a treated unit, there should exist an untreated unit with the same value of X*

DAGs: The Four Basic Patterns

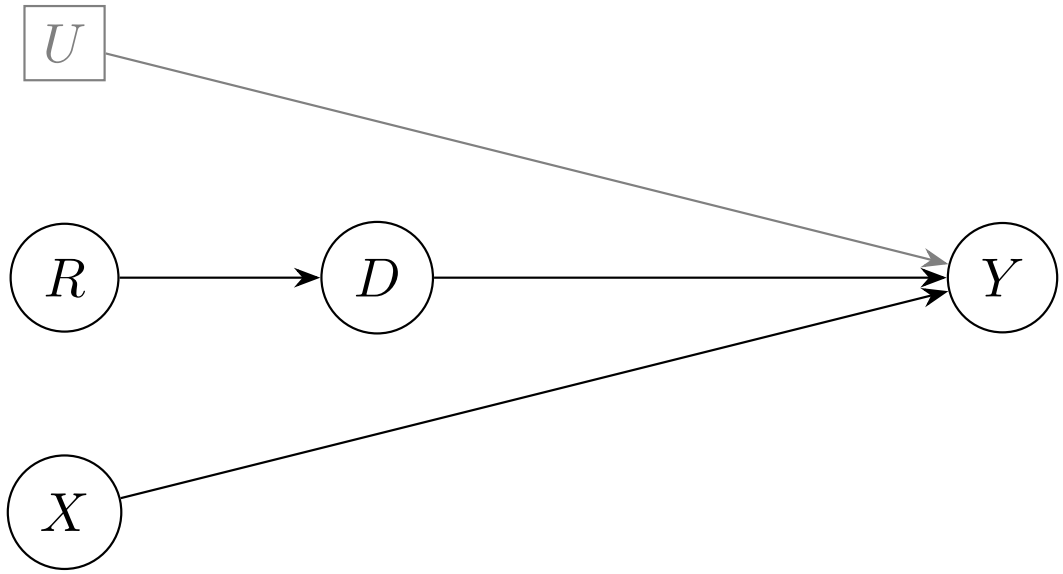


DAGs: Paths

A **path** is a connection between two nodes, directly or passing through other nodes.

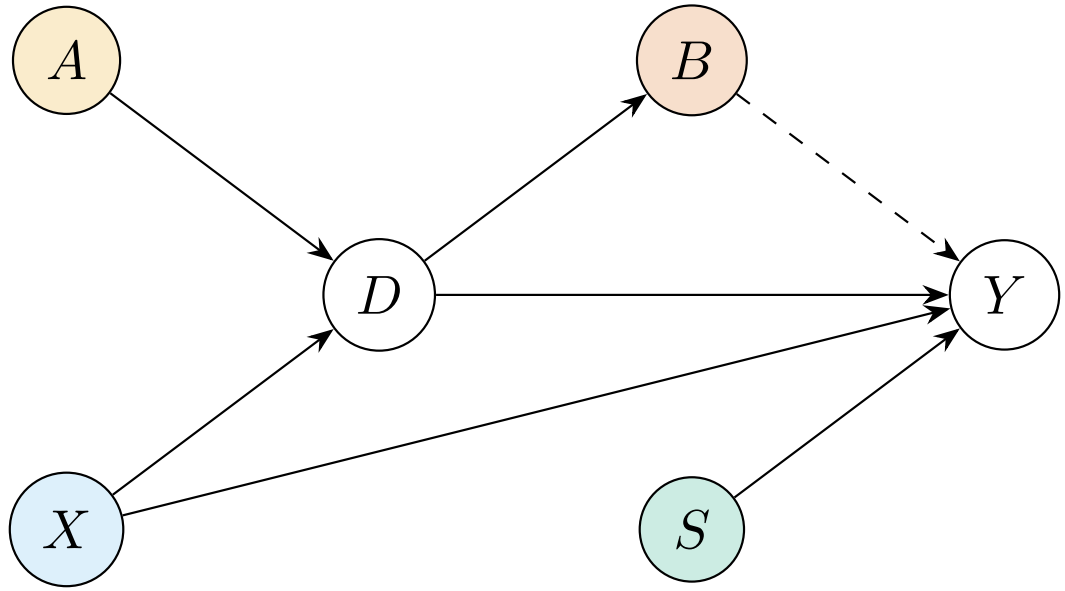
Front door	All the arrows point away from D.
Back door	At least one of the arrows points towards D.
Open	Variation in all variables along the path and no variation in any colliders on that path.
Closed	At least one variable with no variation, or a collider with variation.

DAGs: Randomized Experiments



Thanks to the randomization $R \to D$, there is only one path from D to Y: $D \to Y$. Thus, there is no need to adjust for X (observed) or U (unobserved).

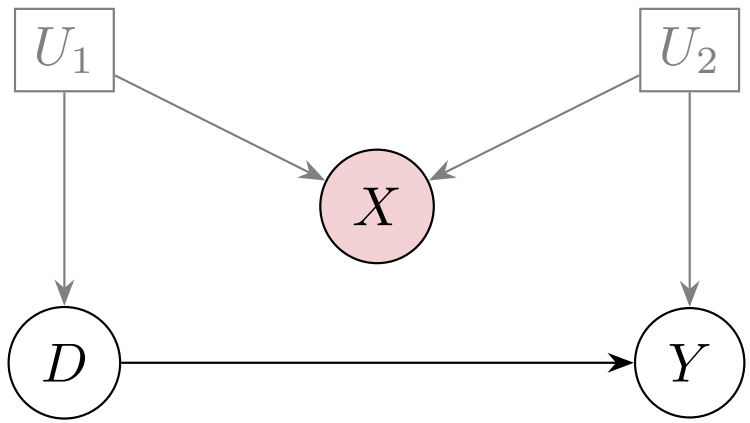
DAGs: Confounding



With regards to the effect $D \rightarrow Y$

- X is a confounder, ignoring it can introduce *omitted variable bias* because it leaves open the back door path $D \leftarrow X \rightarrow Y$
- B is not, controlling for it can cause *post-treatment bias* because it closes a front door path $D \rightarrow B \rightarrow Y$
- S is not either, controlling for it is *neutral*, as no paths from D to Y pass through it, but may be good for precision
- A is not either, however it can *amplify bias* if X is omitted and bad for precision otherwise

DAGs: Trouble with Colliders



With regards to the effect $D \to Y$

- X is a collider on the back door path $D \leftarrow U_1 \to X \leftarrow U_2 \to Y$ and controlling for it would introduce the risk of *M-bias*

Forthcoming, Journal Sociological Methods and Research.

TECHNICAL REPORT
R-493
Revised, March 2022

A Crash Course in Good and Bad Controls

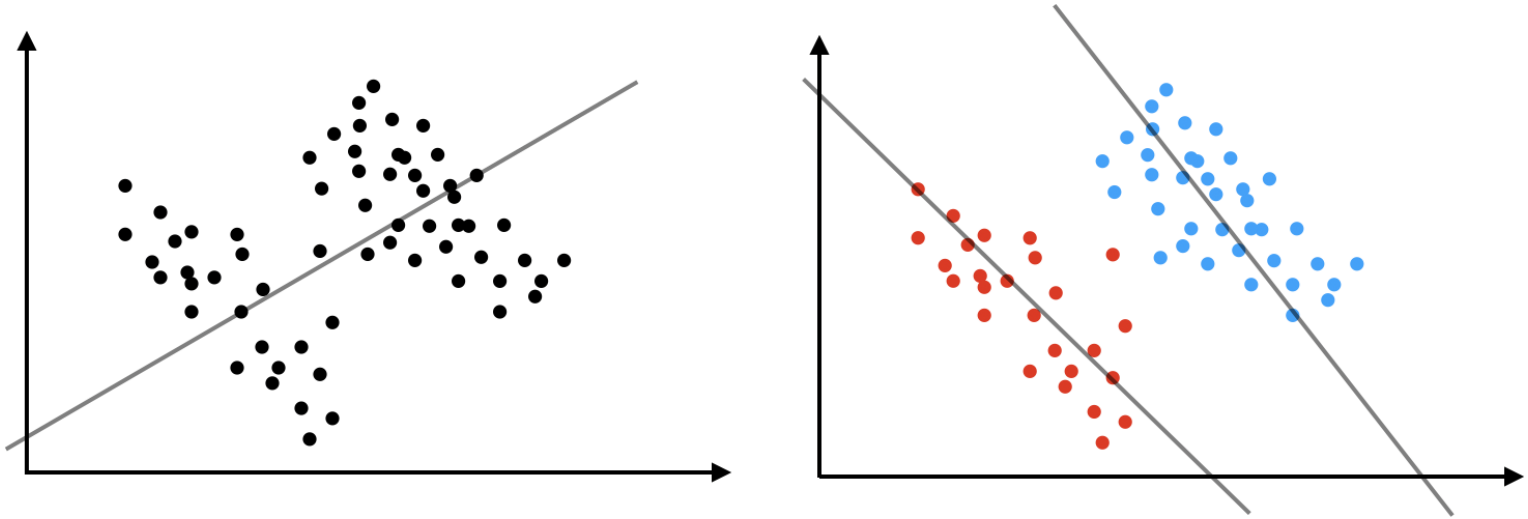
Carlos Cinelli* Andrew Forney† Judea Pearl ‡

March 21, 2022

Abstract

Many students of statistics and econometrics express frustration with the way a problem known as "bad control" is treated in the traditional literature. The issue arises when the addition of a variable to a regression equation produces an unintended discrepancy between the regression coefficient and the effect that the coefficient is intended to represent. Avoiding such discrepancies presents a challenge to all analysts in the data intensive sciences. This note describes graphical tools for understanding, visualizing, and resolving the problem through a series of illustrative examples. By making this "crash course" accessible to instructors and practitioners, we hope to avail these tools to a broader community of scientists concerned with the causal interpretation of regression models.

Simpson's Paradox



Source: https://bit.ly/3DFMd9B

Simpson's paradox: unconditional (left) and conditional (right) association between X and Y has a different direction.

A problem for causal inference: Requires us to chose whether the conditional or the unconditional association better represents the effect of interest.

(a)

Combined	E	$\neg E$		Recovery Rate
drug (C)	20	20	40	50%
no-drug ($\neg C$)	16	24	40	40%
	36	44	80	

(b)

Males	E	$\neg E$		Recovery Rate
drug (C)	18	12	30	60%
no-drug ($\neg C$)	7	3	10	70%
	25	15	40	

(c)

Females	E	$\neg E$		Recovery Rate
drug (C)	2	8	10	20%
no-drug ($\neg C$)	9	21	30	30%
	11	29	40	

Pearl (2014) *Understanding Simpson's Paradox*

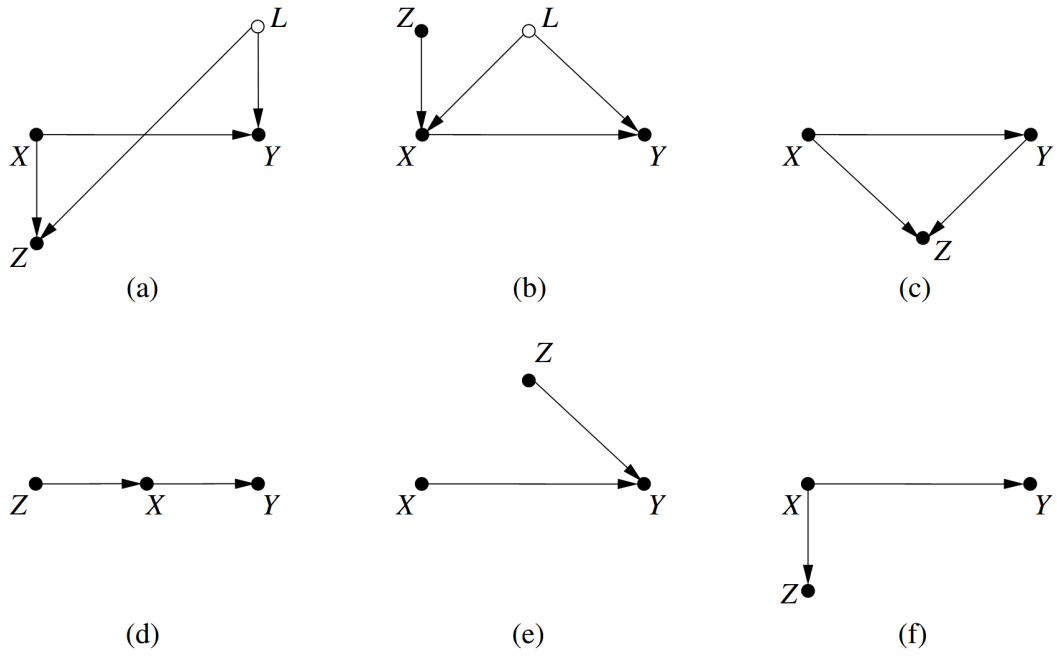


Figure 2: Simpson reversal can be realized in models (a), (b), and (c) but not in (d), (e), or (f).

Pearl (2014) *Understanding Simpson's Paradox*

Controlling: Stratification

Conceptually the simplest way of conditioning on a confounder

Compute for all $x \in \mathcal{X}$

$$\hat{\delta}_x = \mathbb{E}[Y|D = 1, X = x] - \mathbb{E}[Y|D = 0, X = x]$$

and compute their weighted average

$$\hat{\delta} = \sum_x \hat{\delta}_x \, \mathbb{P}[X = x]$$

Controlling: Stratification

The **naive** estimator

$$\mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0]$$

removes the D margin, weighting each unit with

$$w_i = \frac{1}{\mathbb{P}[D = D_i]}, \quad \sum_i w_i = 2, \quad \hat{\delta} = \frac{\mathbb{E}[w_i Y_i]}{\mathbb{E}[w_i]} = 0.5 \sum_i w_i Y_i$$

The **stratification** estimator

$$\sum_x \mathbb{P}[X = x]\big(\mathbb{E}[Y|D=1, X=x] - \mathbb{E}[Y|D=0, X=x]\big)$$

removes the $D|X$ margin, but preserves the X margin, weighting each unit with

$$w_i = \frac{1}{\mathbb{P}[D = D_i|X = X_i]}$$

Controlling: Stratification

(a)

Combined	E	$\neg E$		Recovery Rate
drug (C)	20	20	40	50%
no-drug ($\neg C$)	16	24	40	40%
	36	44	80	

(b)

Males	E	$\neg E$		Recovery Rate
drug (C)	18	12	30	60%
no-drug ($\neg C$)	7	3	10	70%
	25	15	40	

(c)

Females	E	$\neg E$		Recovery Rate
drug (C)	2	8	10	20%
no-drug ($\neg C$)	9	21	30	30%
	11	29	40	

- Naive estimate

$$50\% - 40\% = 10\%$$

- Stratified estimate

$$(60\%-70\%)\times 0.5+(20\%-30\%)\times 0.5 = -10\%\times 0.5-10\%\times 0.5 = -10\%$$

Both the naive and the stratified estimator are special cases of **Inverse Propensity Weighting** (IPW), which uses weights

$$w_i = \frac{1}{p_i}$$

Propensity scores for $\mathcal{D} \in \{0, 1\}$ are

$$\text{if } D_i = 1 \text{ then } p_i = \mathbb{P}[D_i = 1], \quad \text{if } D_i = 0 \text{ then } p_i = 1 - \mathbb{P}[D_i = 1].$$

The stratified estimator is non-parametric in the sense that each $x \in \mathcal{X}$ is treated separately.

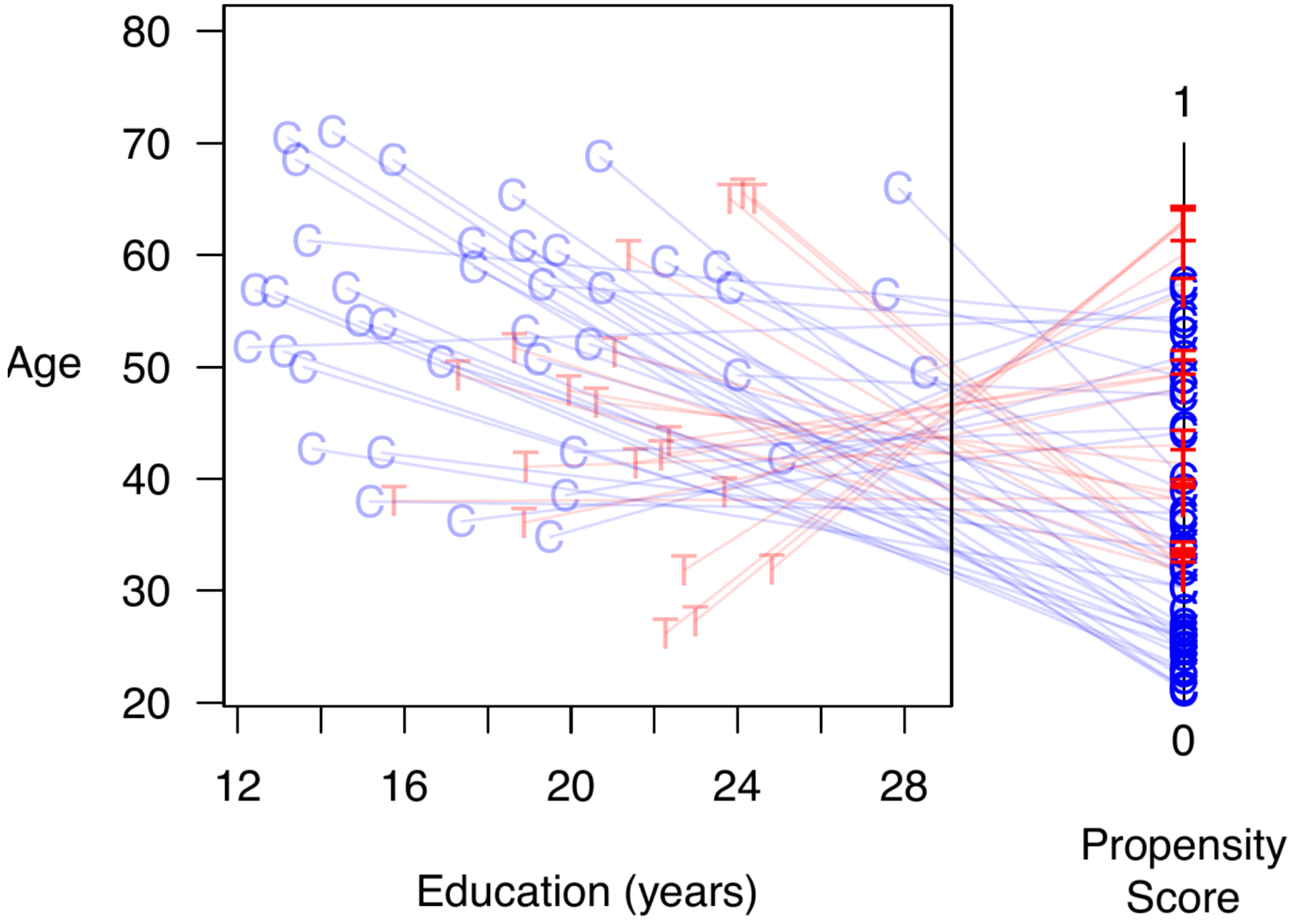
Controlling: Propensity Score Matching

Different use of propensity scores.

In a nutshell

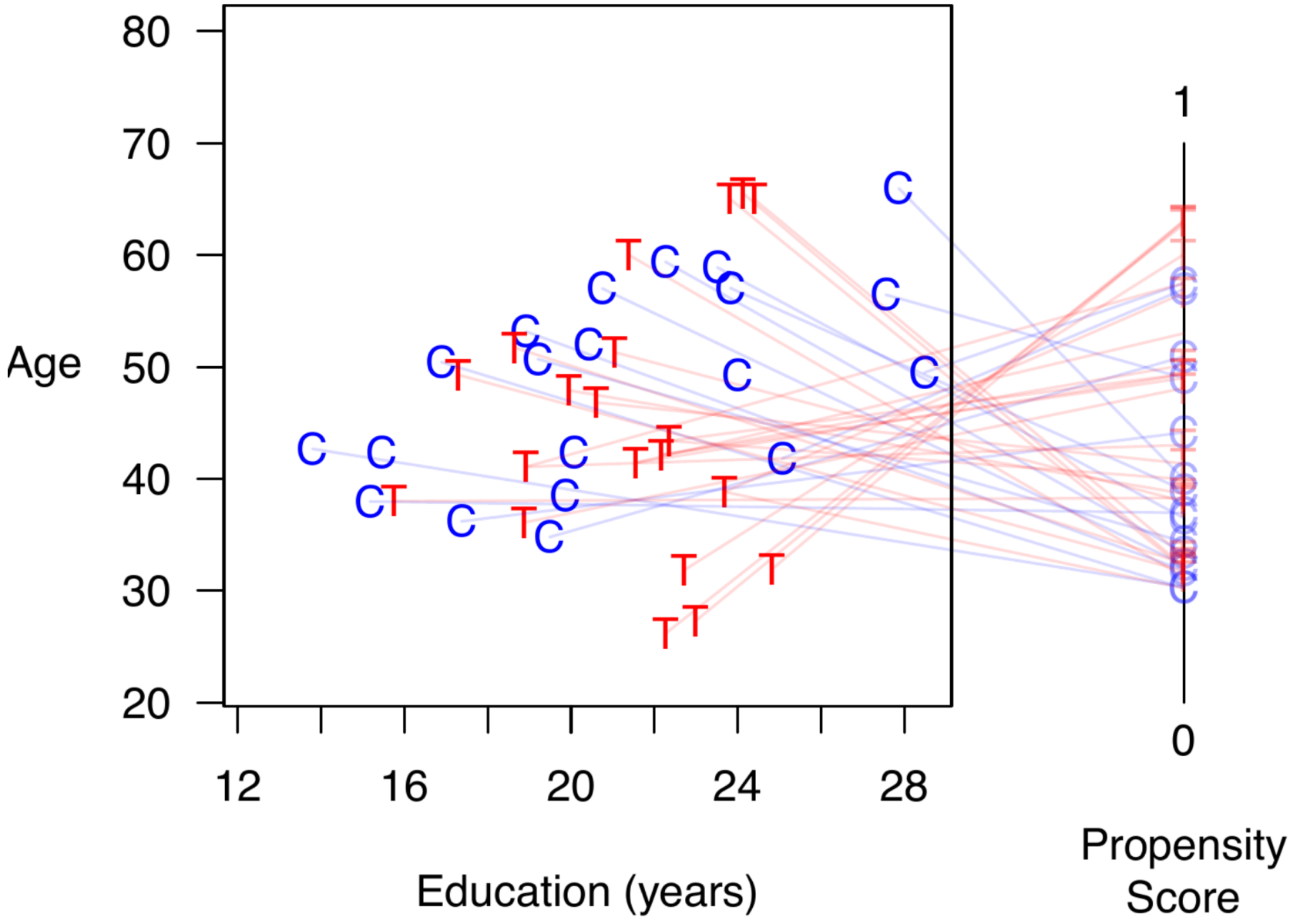
- Compress the potentially high-dimensional $\boldsymbol{X}$ onto a one-dimensional propensity score $\{p_i\}$.
- Construct a control group for the treated units by finding in the control pool the units that most resemble them in terms of the propensity scores.

Propensity Score Matching: Example



Source: Gary King

Propensity Score Matching: Example



Source: Gary King

Controlling: Matching

Propensity Score Matching is just one of many types of matching.

We've already seen Exact Matching, Coarsened Exact Matching, or Mahalanobis Distance Matching.

They all work with some divergence

$$d\big(g(\boldsymbol{X}_i), g(\boldsymbol{X}_j)\big)$$

where

- $\{i, j\}$ is a pair of units
- $d(.)$ a divergence
- $g(.)$ a transformation of $\boldsymbol{X}$

Many choices of $d()$, $g()$ and other settings that are not always easy to navigate.

But transparent regarding each unit's contribution to the estimate of the average effect.

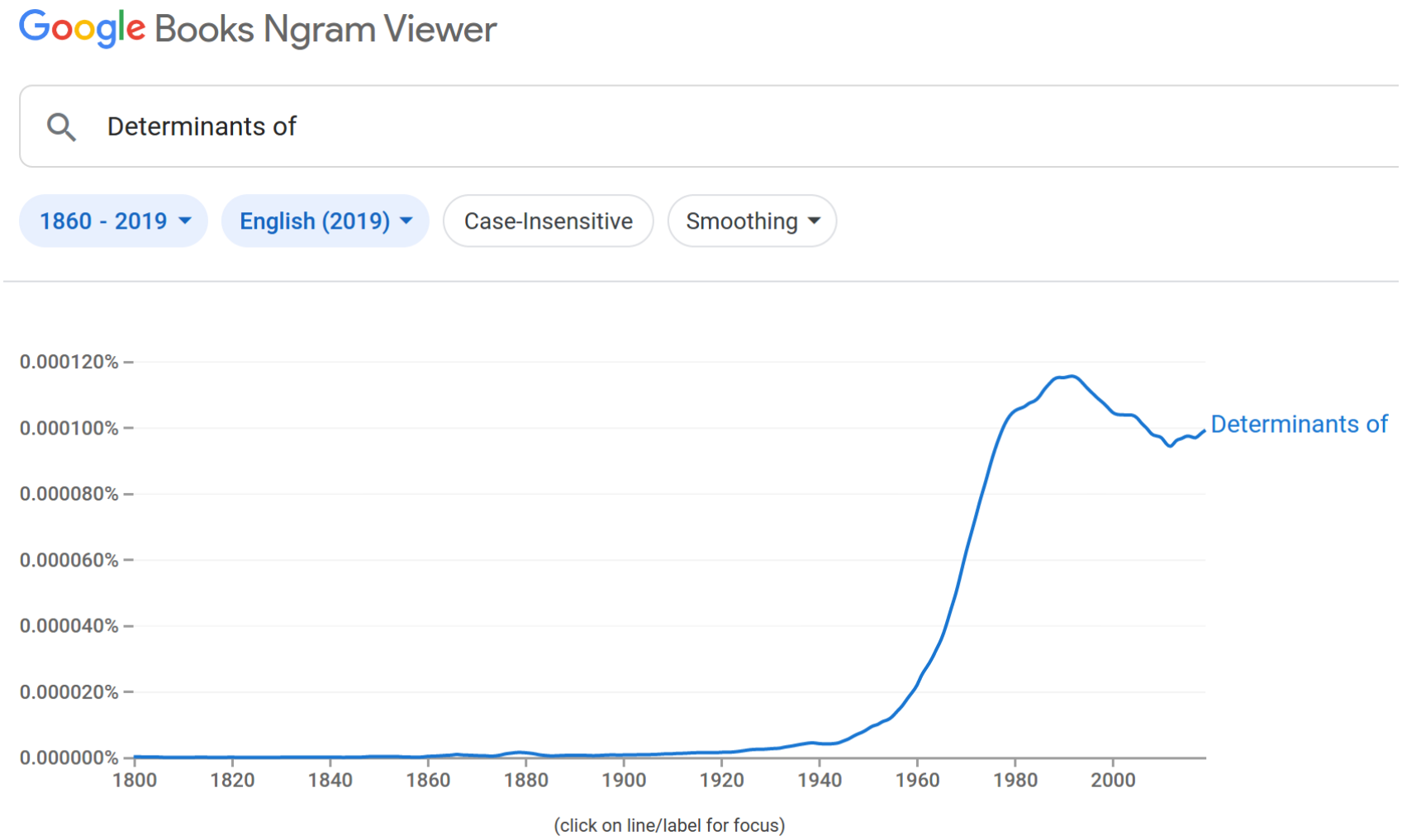
Controlling: Multiple Regression

Regression Adjustment

1. Define a model of the conditional distribution of the outcome given the treatment additional observables (confounders, covariates) and parameters to be estimated (e.g., coefficients, slopes etc)
2. Estimate the model's parameters by optimising some objective function (e.g. sum of squared residuals, likelihood function of the conditional distribution of the outcome, etc.)
3. Interpret one or more coefficients as effect estimate(s)

Given how accessible and widespread it is, it is no surprise that many/most applications fail to follow best practices in some way

Regression Adjustment: The "Determinants of ..." Problem



Regression Adjustment: The "Determinants of ..." Problem

- In some fields, the usual practice is to put multiple variables on the right-hand-side, and then interpret all of their coefficients as effect estimates
- This *almost never works*
- Even if we have only two variables, D and X, if X is pre-treatment w.r.t. D, then D is post-treatment w.r.t. X
- Controlling for post-treatment covariates will not give a good estimate of the treatment effect, at best only of the "direct controlled effect"

Regression Adjustment: The "Determinants of ..." Problem

- In some fields, the usual practice is to put multiple variables on the right-hand-side, and then interpret all of their coefficients as effect estimates
- This *almost never works*
- Even if we have only two variables, D and X, if X is pre-treatment w.r.t. D, then D is post-treatment w.r.t. X
- Controlling for post-treatment covariates will not give a good estimate of the treatment effect, at best only of the "direct controlled effect"
- Sadly, this means many "Determinants of ..." studies are making a negative contribution
- Some fields struggle with recognising this, as projects and careers are built on this mistake

Regression Adjustment: OLS

- Ordinary Least Squares is equivalent to linear regression with gaussian residuals

$$
\begin{aligned}
y_i &= \alpha + \beta d_i + \gamma_1 z_{i1} \cdots + \epsilon_i, \\
\epsilon_i &\sim Normal(0, \sigma)
\end{aligned}
$$

- Most popular regression technique for causal inference in observational studies
- Computationally relatively inexpensive ...

  unless you're dealing with really large data and many variables
- Despite decades of use, we are still learning what exactly does regression do when we use it for causal inference

Linear Regression: A Quick Anatomy

Consider the simplest setup

$$Y_i = \underbrace{\alpha + \beta X_i}_{\substack{\text{deterministic} \\ \text{component}}} + \underbrace{\epsilon_i, \quad \epsilon_i \sim R(.)}_{\substack{\text{stochastic} \\ \text{component}}}$$

The residual distribution $R(.)$ defines the objective:

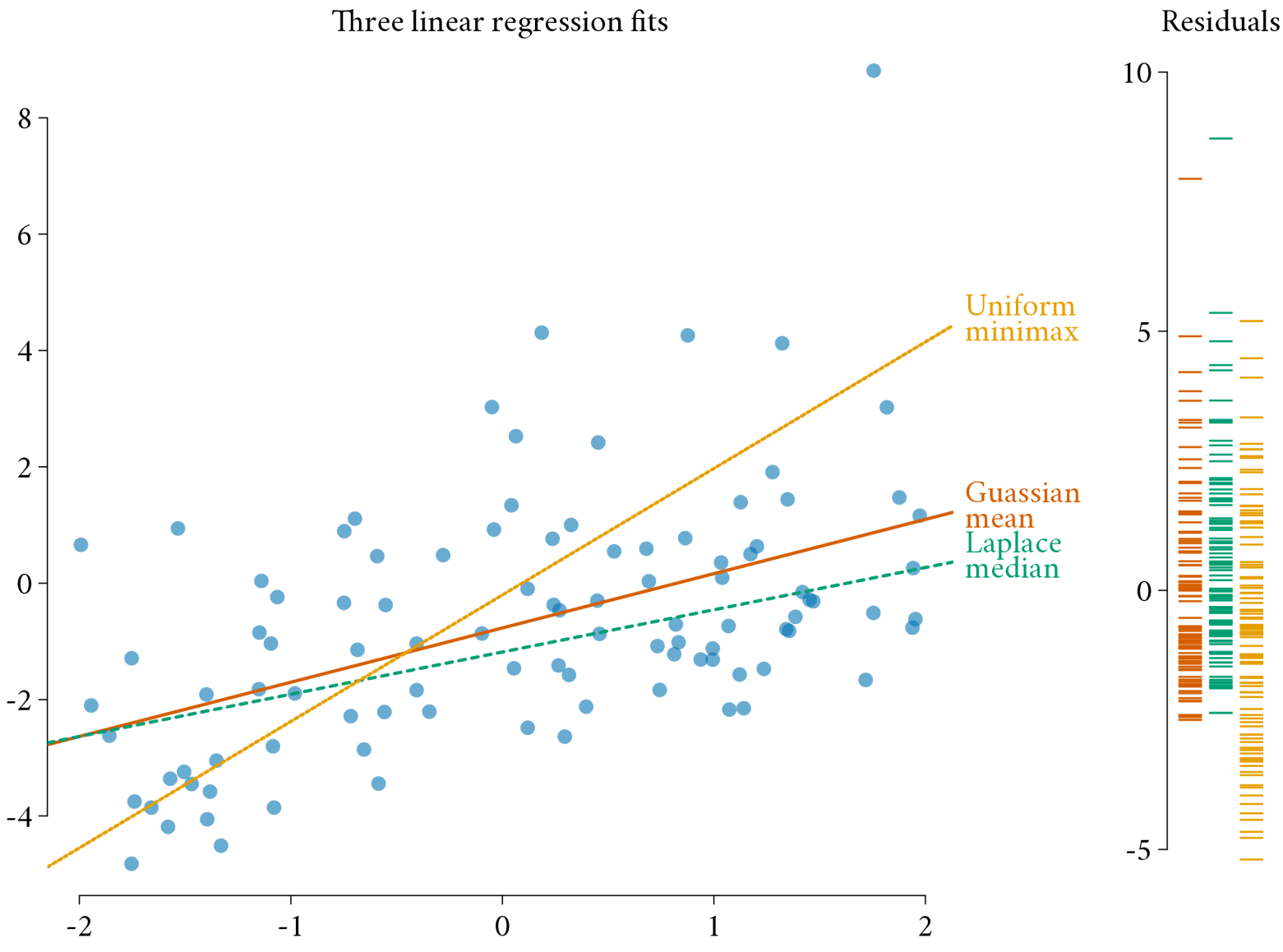
$R(.)$	Minimizes	Conditional
$Normal(0, \sigma)$	$\sum_i \epsilon_i^2$	Mean
$Laplace(0, \lambda)$	$\sum_i \lvert\epsilon_i\rvert$	Median
$Uniform(-\nu, \nu)$	$\max_i \lvert\epsilon_i\rvert$	Midpoint

There are many other options.

Despite its name, $R^2 = 1 - \frac{v(\epsilon)}{v(Y)}$ is variance *captured* rather than causally explained, a measure of the relative size of the deterministic component compared to the stochastic component.

Good fit and causal identification are two different things.

Linear Regression: A Quick Anatomy



Frisch–Waugh–Lovell Theorem

Under the simplest multiple regression setup

$$Y_i = \alpha + \delta D_i + \beta X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma)$$

We can get the exact same coefficient estimate of δ in a regression of two kinds of residuals:

$$\begin{aligned}
Y_i &= \gamma_{10} + \gamma_{11} X_i + \epsilon_i^Y, \quad \epsilon_i^Y \sim N(0, \sigma_Y) \\
D_i &= \gamma_{20} + \gamma_{21} X_i + \epsilon_i^D, \quad \epsilon_i^D \sim N(0, \sigma_D) \\
\epsilon_i^Y &= \delta \epsilon_i^D + \epsilon_i \\
\delta &= \frac{\epsilon_i^Y - \epsilon_i}{\epsilon_i^D}
\end{aligned}$$

OLS Coefficients as Effect Estimates

The Aronow-Samii (2016) *multiple regression weights*:

In multiple linear regression with Gaussian residuals

$$Y_i = \alpha + \delta D_i + \beta X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma)$$

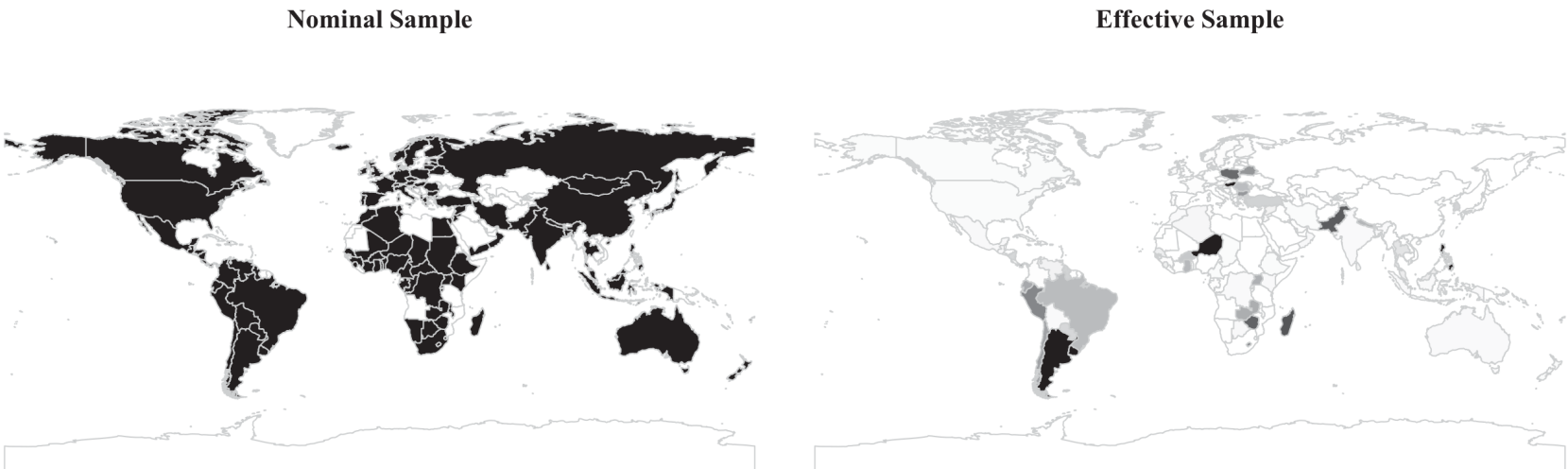
the estimate $\hat{\delta}$ converges to a weighted average of the individual treatment effects $\{\tau_i\}$

$$\hat{\delta} \xrightarrow{p} \frac{\mathbb{E}[\omega_i \tau_i]}{\mathbb{E}[\omega_i]}$$

where

$$\omega_i = \big(D_i - \mathbb{E}[D_i|X_i]\big)^2 = (\epsilon_i^D)^2$$

FIGURE 1 Example of nominal and effective samples from Jensen (2003)



Note: On the left, the shading shows countries in the nominal sample for Jensen (2003) estimate of the effects of regime type on FDI. On the right, darker shading indicates that a country contributes more to the effective sample, based on the panel specification used in estimation.

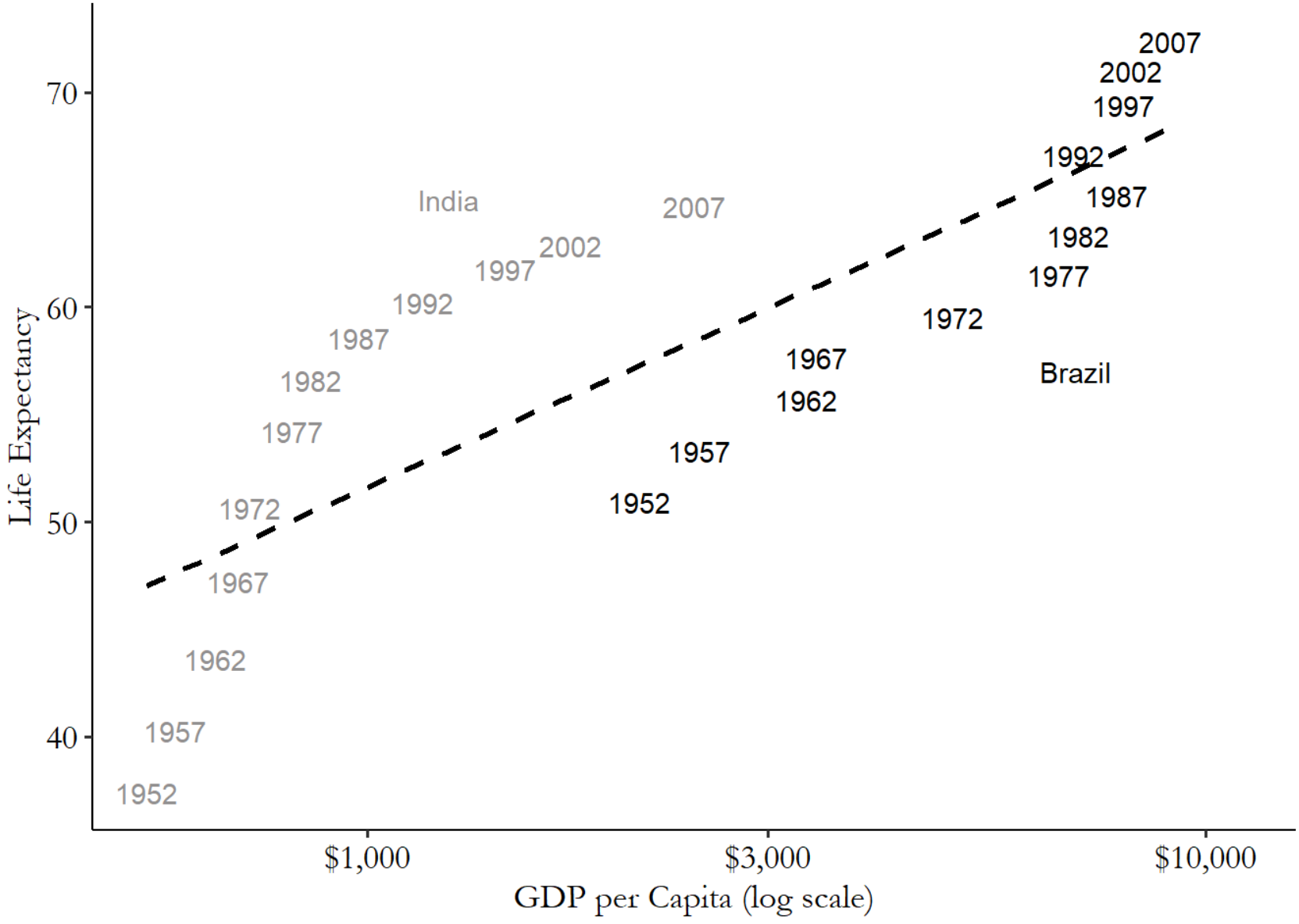
Source: Aronow and Samii (2016)

Just because we have the habit of casually calling every other parameter an 'effect', it does not mean it will be a good estimate of some causal effect.

So-called Fixed Effects

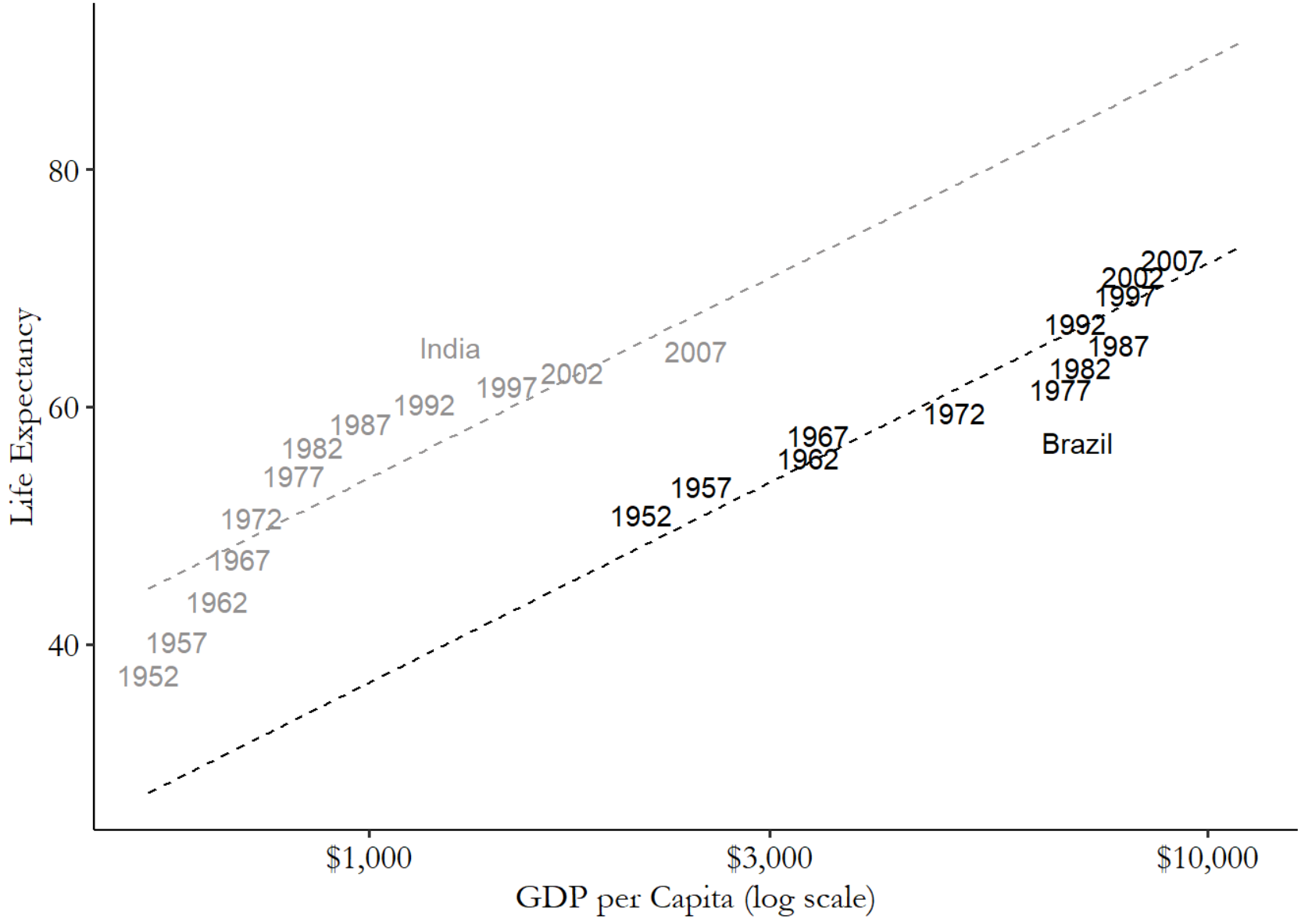
- Panel data: same units at multiple time-points
- Concern with unobserved confounding that is time-invariant
- Add unit-specific intercepts ("effects") into the regression model
- This makes the estimates based only on in-time variation

Fixed Effects are Just (Unpenalized) Intercepts



Source: *The Effect*

Fixed Effects are Just (Unpenalized) Intercepts



Source: *The Effect*

Difference-in-Differences

Basic Diff-in-Diff is straightforward. Suppose

- There are two groups, one received the treatment ($D = 1$), the other the control ($D = 0$).
- Both groups are observed before ($T = 0$) and after ($T = 1$) the treated received $D = 1$.
- The treated would change the same way as the control group had they not been treated,
- Confounding is both group- and time-invariant, i.e. it works the same in all four groups defined by treatment/contol and pre/post.

Then we can compute the difference in pre-post differences as

$$\begin{aligned}\widehat{ATT} &= \big(\mathbb{E}[Y|\widehat{D=1,T=1}] - \widehat{\mathbb{E}}[Y|D=1,T=0]\big) - \\ &\quad \big(\widehat{\mathbb{E}}[Y|D=0,T=1] - \widehat{\mathbb{E}}[Y|D=0,T=0]\big)\end{aligned}$$

Diff-in-Diff: Parallel Trends

To estimate

$$\widehat{ATT}_{DiD} = \widehat{\mathbb{E}}[Y(1)|D=1,T=1] - \widehat{\mathbb{E}}[Y(0)|D=1,T=1]$$

DiD imputes the counterfactual for the treated group in the after period

$$\widehat{\mathbb{E}}[Y(0)|D=1,T=1] =$$
$$\widehat{\mathbb{E}}[Y(0)|D=1,T=0] + \widehat{\mathbb{E}}[Y(0)|D=0,T=1] - \widehat{\mathbb{E}}[Y(0)|D=0,T=0]$$

Since this is all additive, time parallels in $Y(0)$

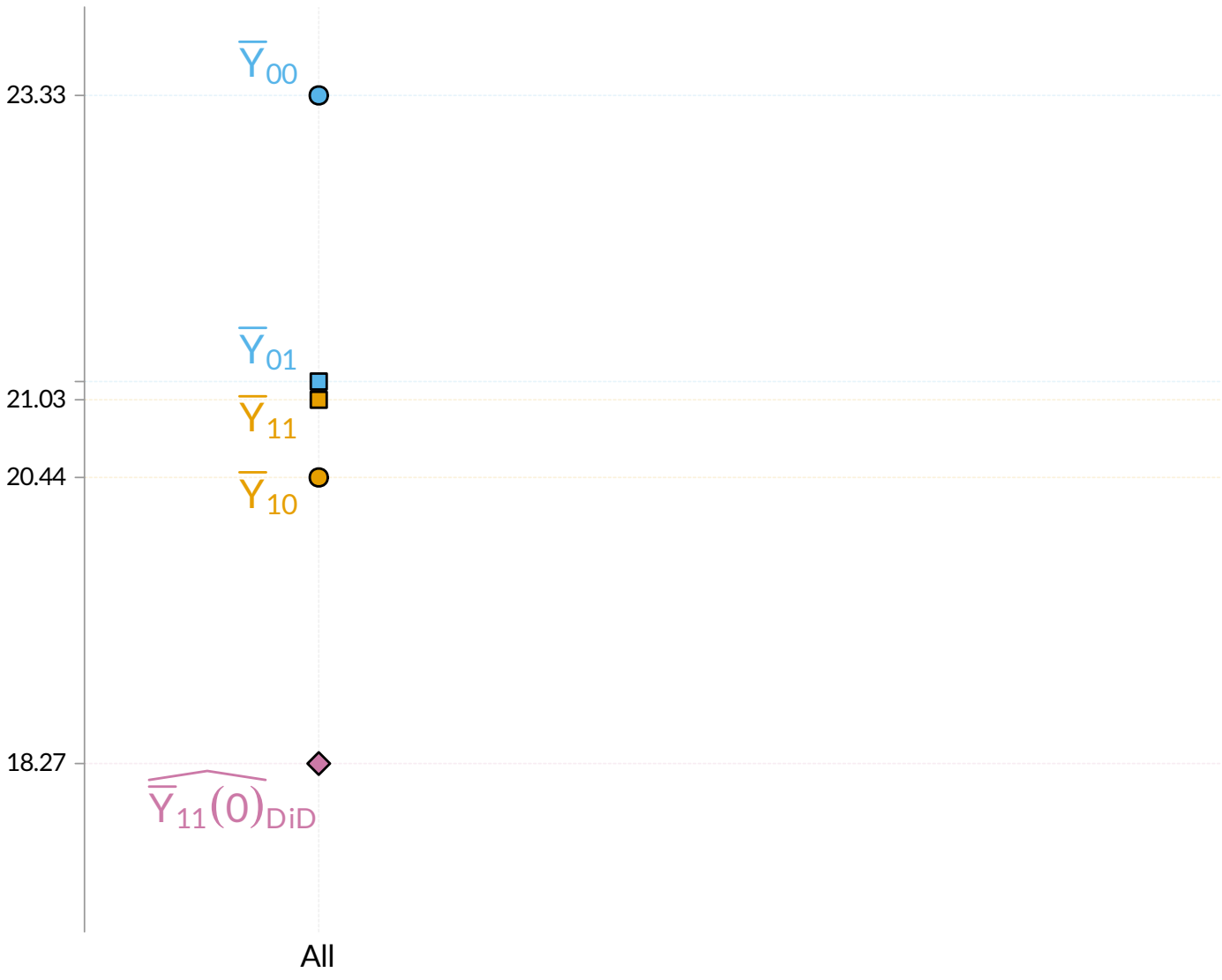
$$\mathbb{E}[Y(0)|D=1,T=1] - \mathbb{E}[Y(0)|D=1,T=0] =$$
$$\mathbb{E}[Y(0)|D=0,T=1] - \mathbb{E}[Y(0)|D=0,T=0]$$

and group parallels in $Y(0)$

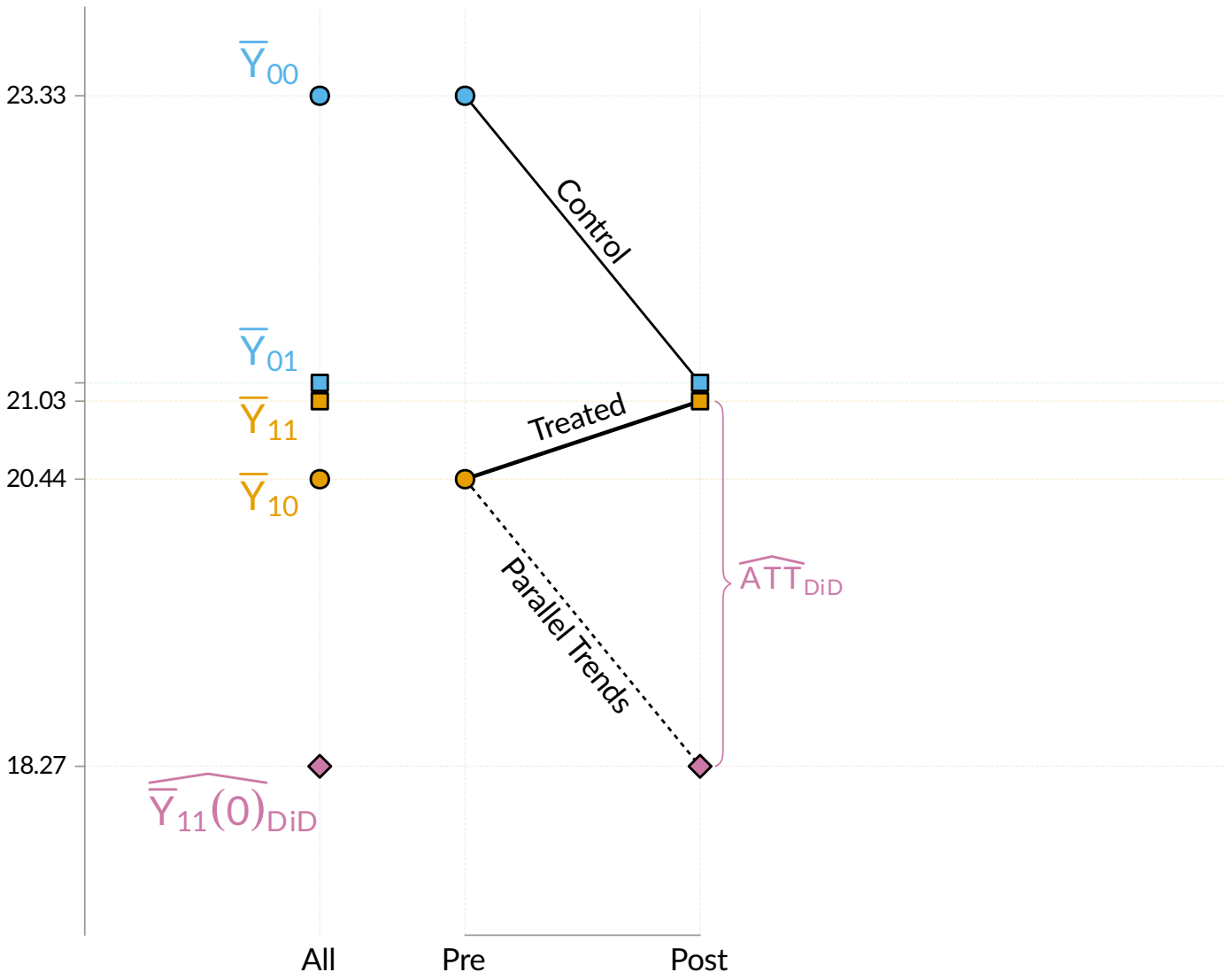
$$\mathbb{E}[Y(0)|D=1,T=1] - \mathbb{E}[Y(0)|D=0,T=1] =$$
$$\mathbb{E}[Y(0)|D=1,T=0] - \mathbb{E}[Y(0)|D=0,T=0]$$

mean the same thing.

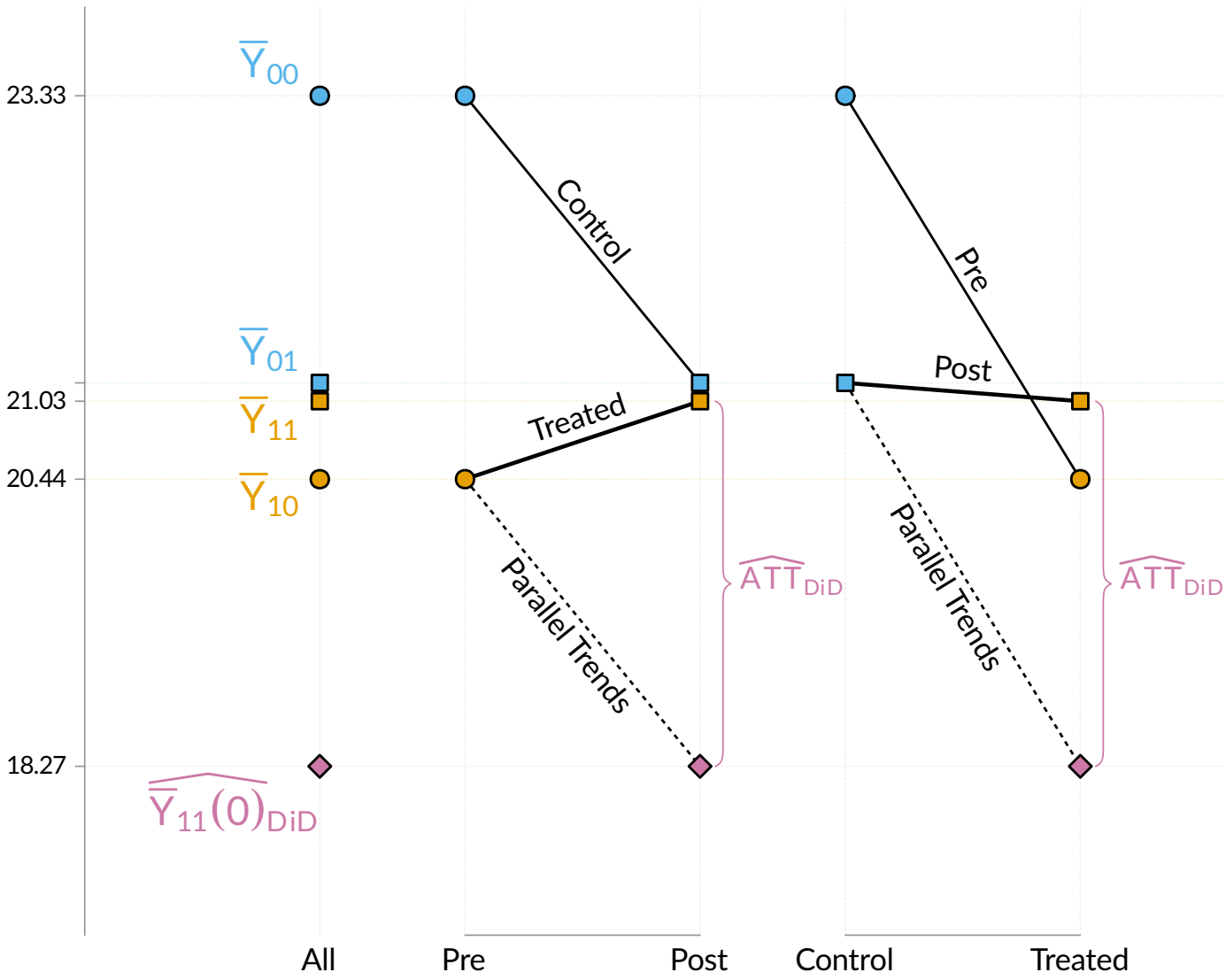
Diff-in-Diff: Parallel Trends



Diff-in-Diff: Parallel Trends



Diff-in-Diff: Parallel Trends



When to use Diff-in-Diff?

You

- Want to learn the effects of a treatment on some outcome.
- Have data on both treated and untreated units.
- Have the data for at least 2 periods: before and after the treatment.

But

- Treatment assignment is not randomized.
- Don't know or see all potential confounders.
- Other things may have changed between before and after.

Then you need to assume

- Parallel trends.

And to check for them, it helps to have

- At least 3 time periods, of which at least 2 are in the before period.

Diff-in-Diff: Diagnostics and Forensics

Check

- For parallel trends in the before period.
- DiD estimates with placebo outcomes, i.e. known to be unaffected by the treatment.
- With different control groups.
- For things that happened simultaneously with the treatment affecting only one group.
- Relax the linearity assumption with Changes-in-Changes.

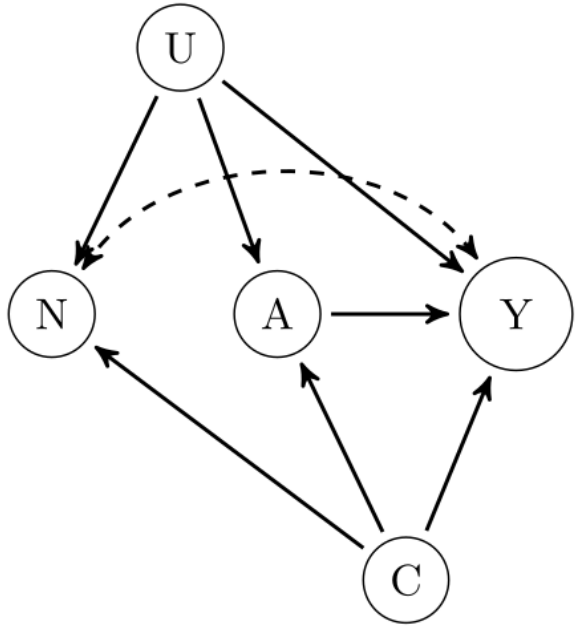
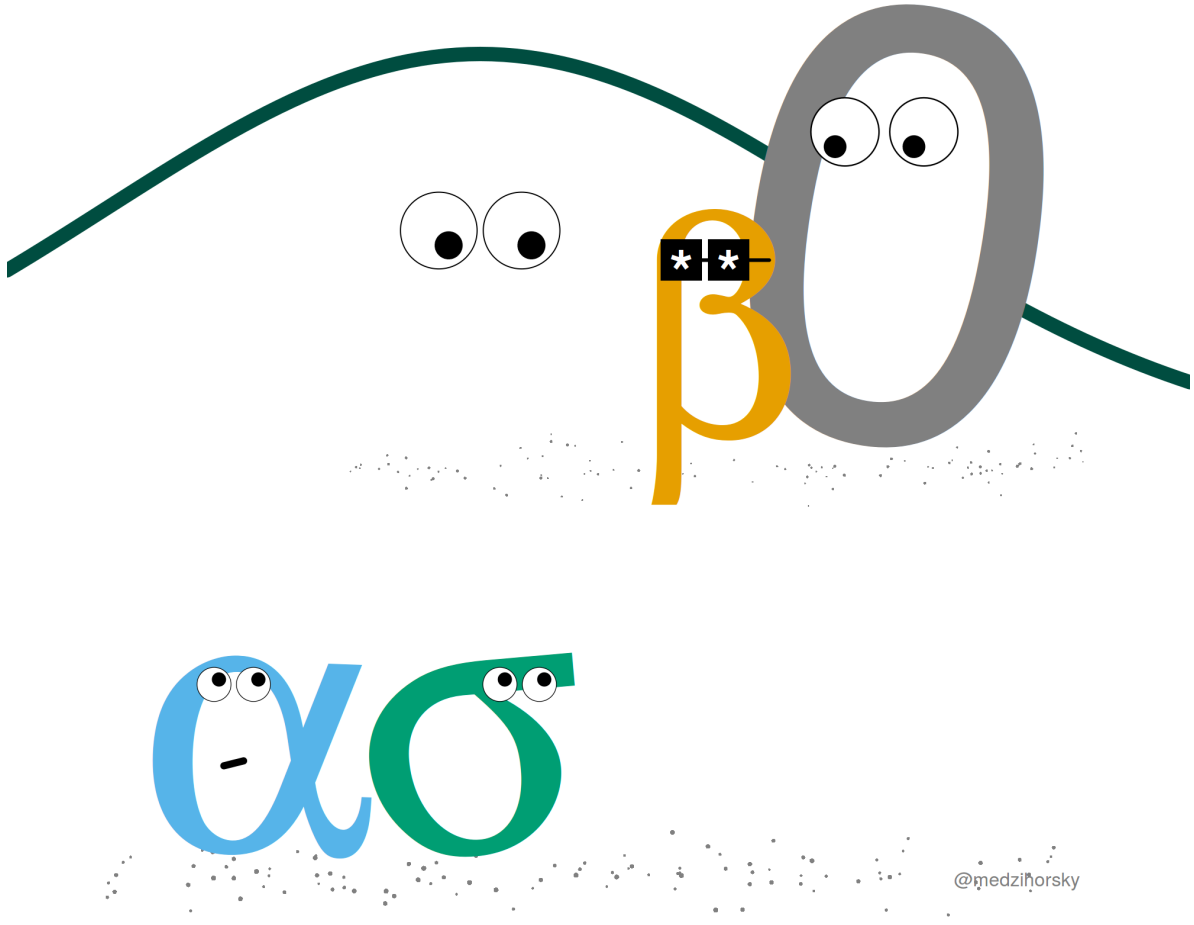


FIG. 1. *Directed acyclic graph depicting the causal association between the treatment A, primary outcome Y, negative control outcome N, measured pre-exposure confounders C and unmeasured confounders U.*

Source: Sofer et al. (2016)

Effects



"He thinks he's an effect."

Thank you!

APPENDIX

Changes-in-Changes

Identification and inference in nonlinear difference-in-differences models

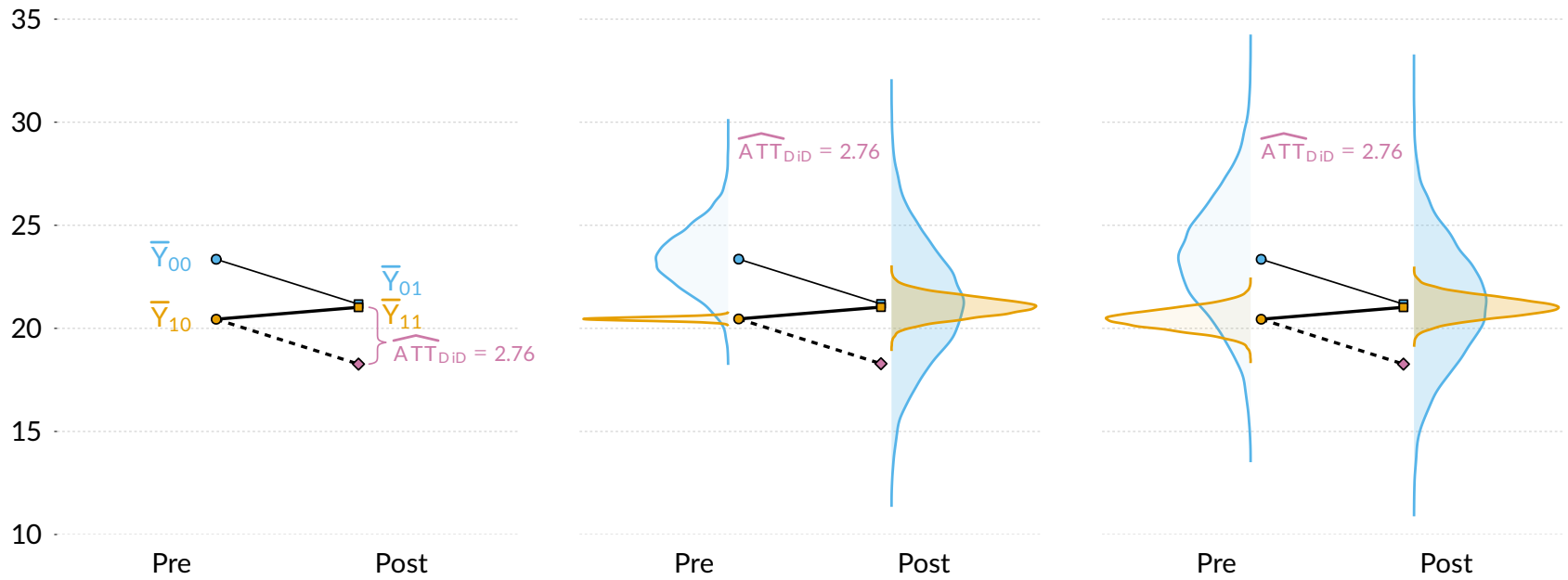
S **Athey**, GW **Imbens** - Econometrica, 2006 - Wiley Online Library

… Finally, in the supplementary material to this article, available on the Econometrica website (**Athey** and **Imbens** (2006)), we apply the methods developed in this paper to study the effects …

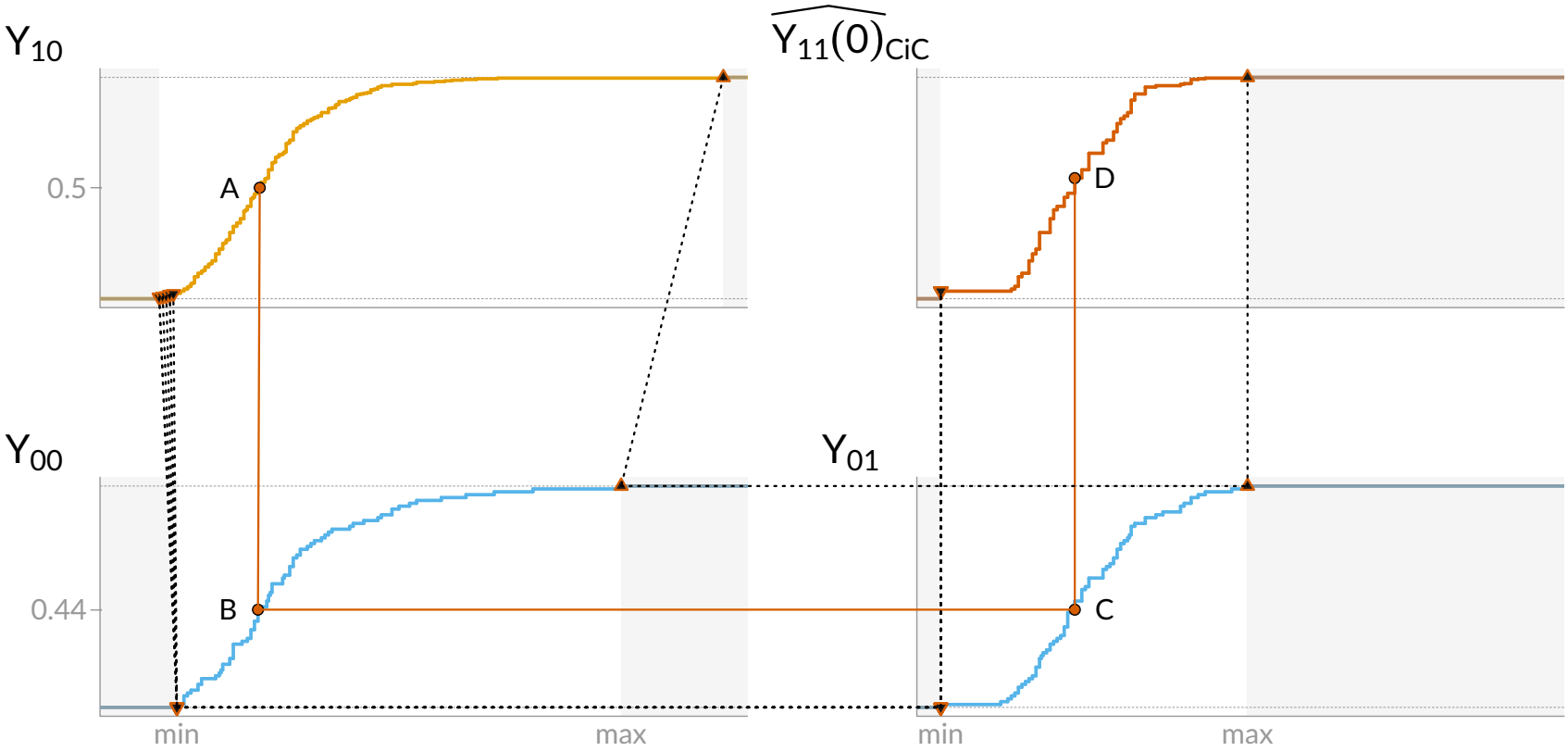
☆ Save Cite Cited by 1423 Related articles All 26 versions »

Changes-in-Changes: Motivation

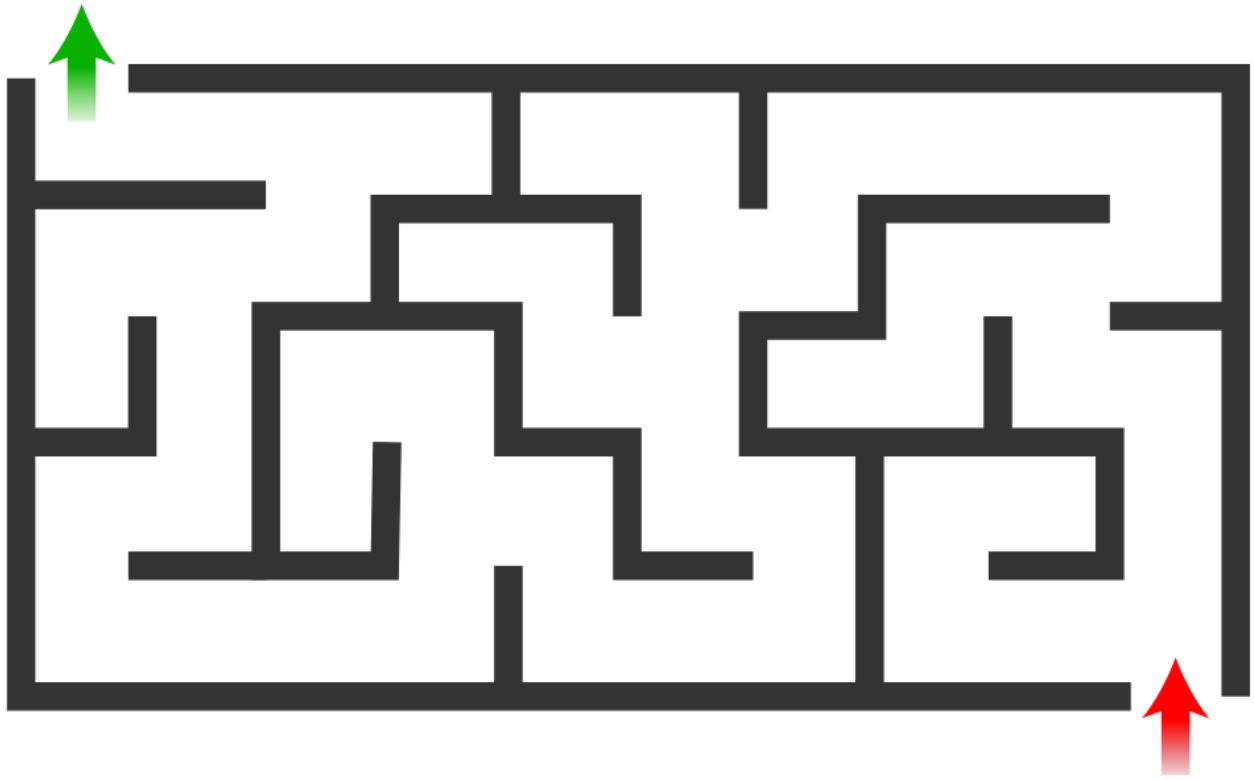
$$\widehat{ATT}_{DiD} = \Big(\widehat{\mathbb{E}}[Y_{11}] - \widehat{\mathbb{E}}[Y_{10}]\Big) - \Big(\widehat{\mathbb{E}}[Y_{01}] - \widehat{\mathbb{E}}[Y_{00}]\Big)$$



Athey & Imbens (2006) generalizes DiD using cdfs and inverse cdfs.

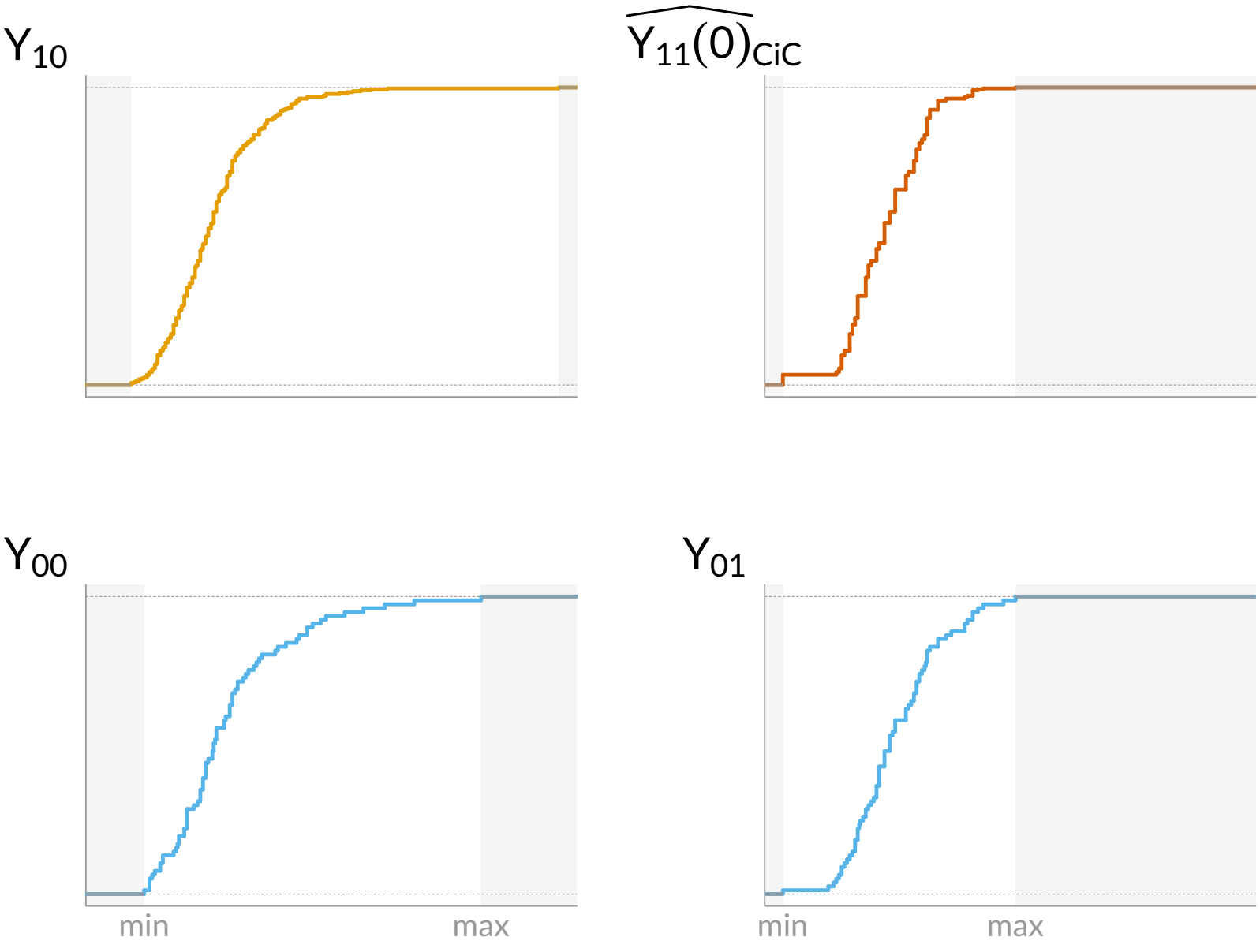


Changes-in-Changes: Walkthrough

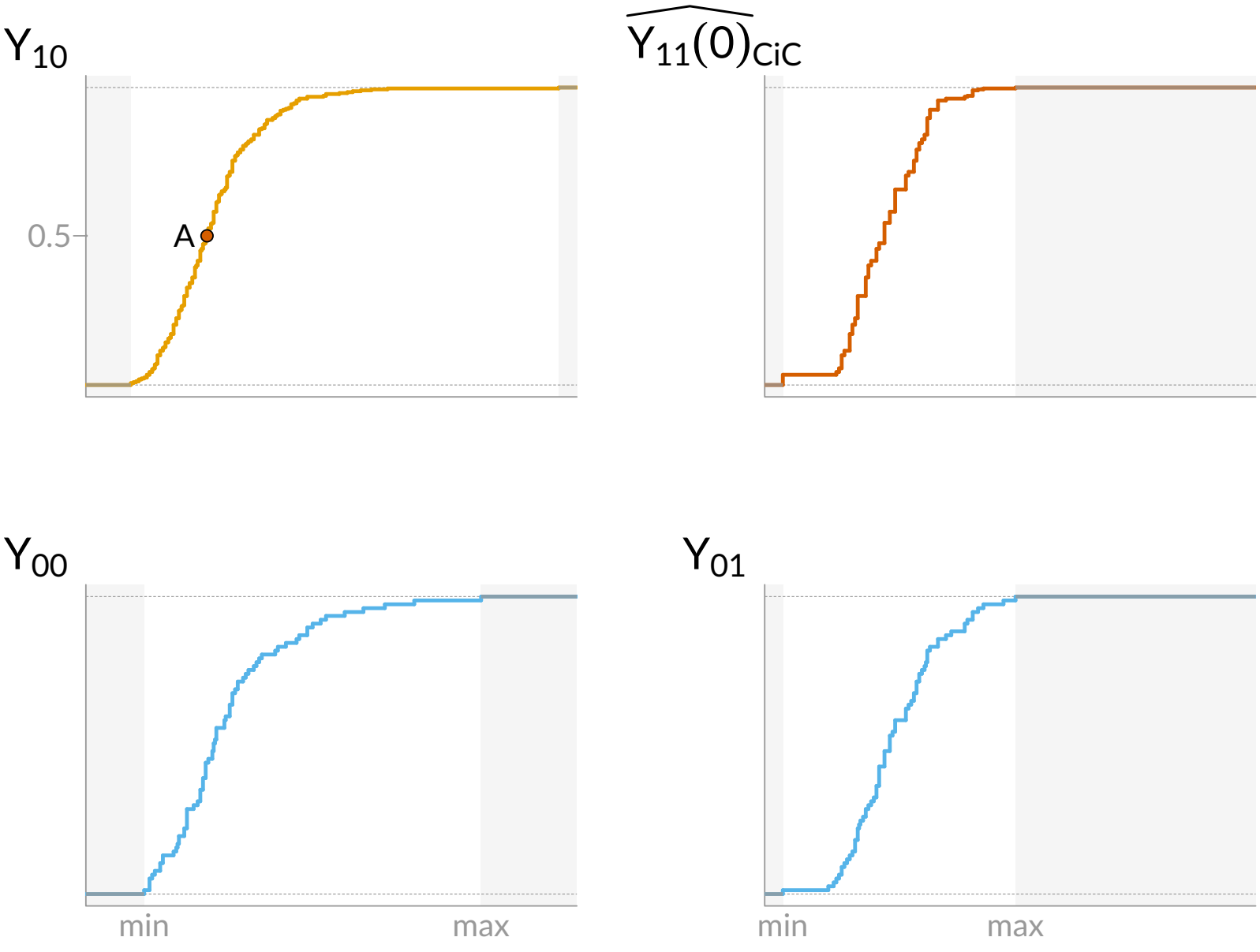


Source: *Wikipedia*

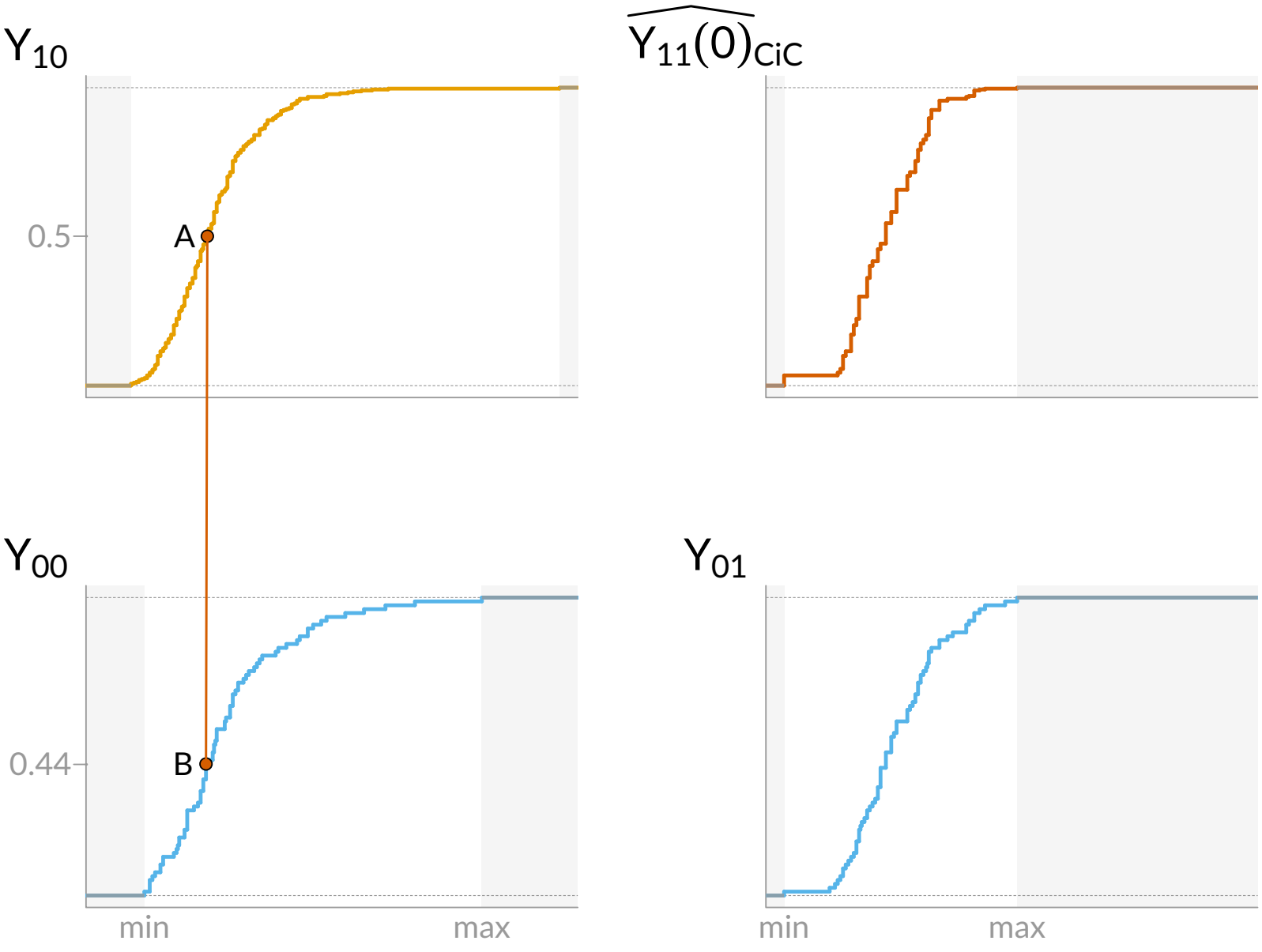
Changes-in-Changes: Walkthrough



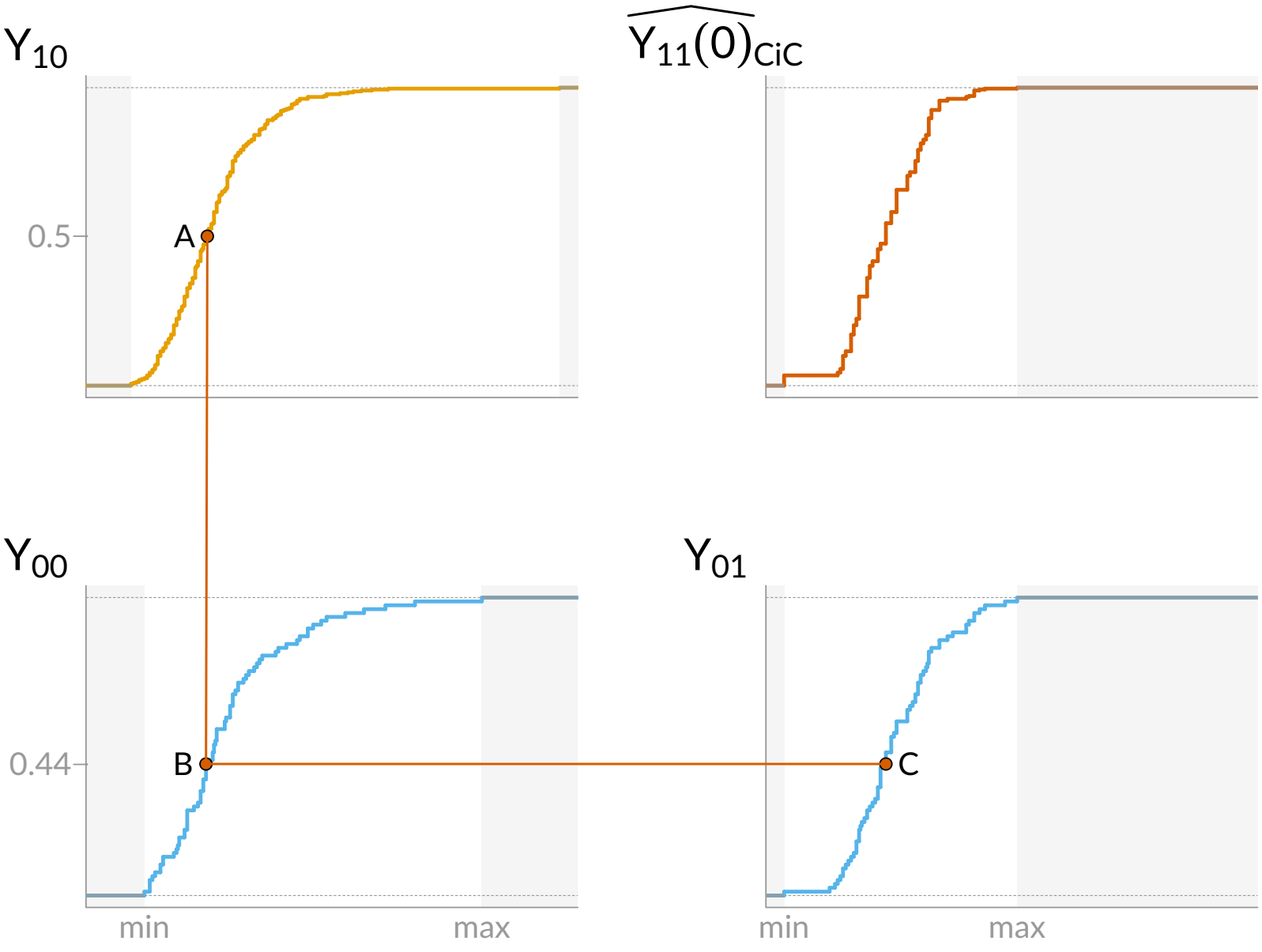
Changes-in-Changes: Walkthrough



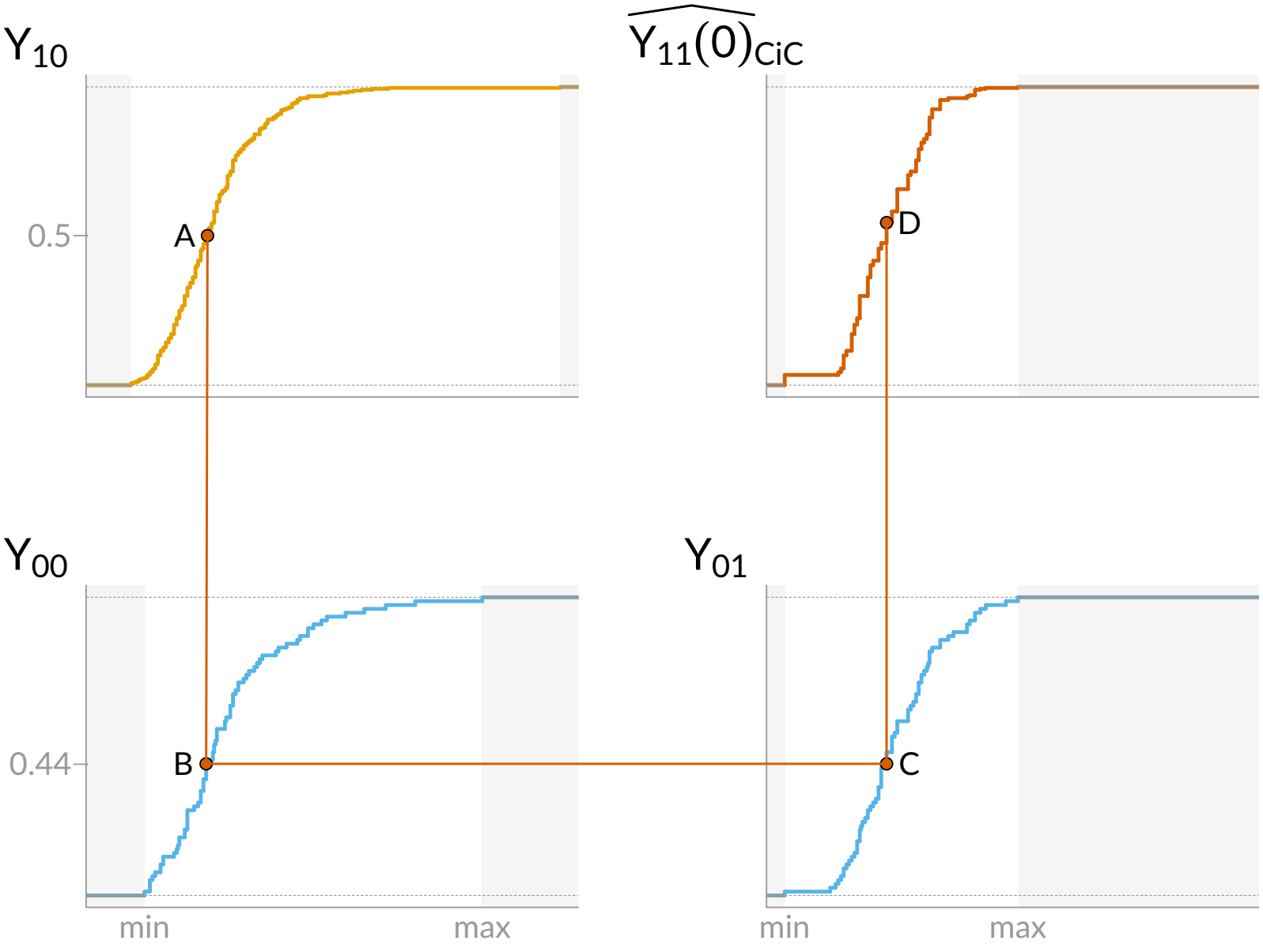
Changes-in-Changes: Walkthrough



Changes-in-Changes: Walkthrough

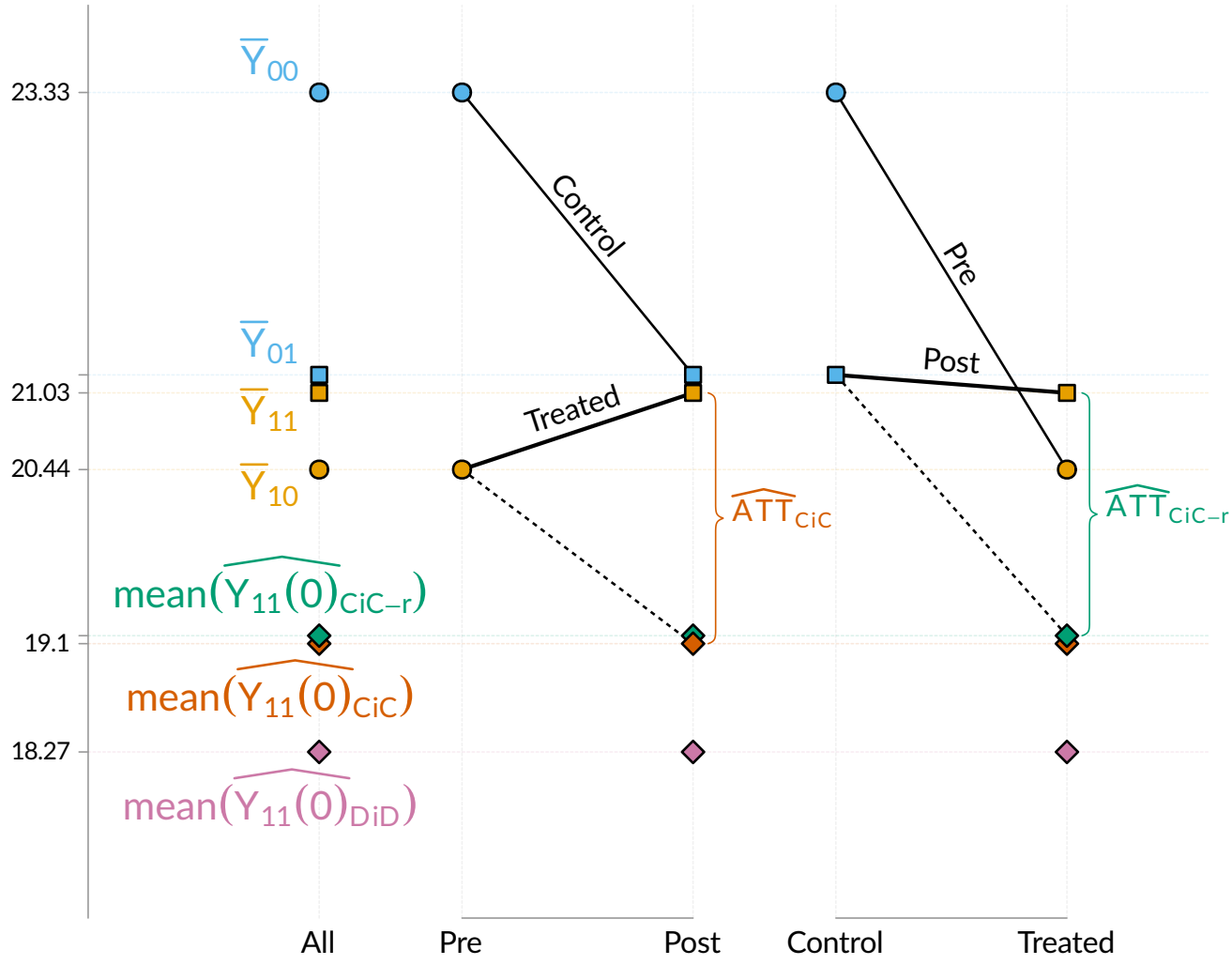


Changes-in-Changes: Walkthrough



CiC vs. DiD in an example

Card & Kruger (1994)



Negative Outcome Control

[HTML] On **negative outcome** control of unobserved confounding as a generalization of difference-in-differences

T **Sofer**, DB Richardson, E Colicino… - Statistical science: a …, 2016 - ncbi.nlm.nih.gov

… In Section 3 we show how **negative outcomes** potentially can be used in broader settings than the classical DID, and develop a general NOC approach to indirectly account for …

☆ Save Cite Cited by 49 Related articles All 11 versions

NOC procedure

Sofer et al. (2016) presents a DiD-like procedure using placebo (negative) outcomes (N) that are known to not be affected by the treatment but may be affected by the same confounding factors.

- pre-treatment outcomes are a special case of placebo outcomes
- NOC average estimate:

$$\frac{1}{n_{11}}\sum_{i=1}^{n_{11}} Y_{11,i} - \frac{1}{n_{11}}\sum_{i=1}^{n_{11}} \hat{F}^{-1}_{Y_{01}}\left(\hat{F}_{N_{01}}\left(N_{11,i}\right)\right)$$

- CiC average estimate:

$$\frac{1}{n_{11}}\sum_{i=1}^{n_{11}} Y_{11,i} - \frac{1}{n_{10}}\sum_{i=1}^{n_{10}} \hat{F}^{-1}_{Y_{01}}\left(\hat{F}_{Y_{00}}\left(Y_{10,i}\right)\right)$$

NOC example I.

Sofer et al. (2016): individuals' health and air pollution

- Treatment: black carbon (air pollution) exposure
- Primary Outcome (Y): fibrinogen, a blood inflamation marker
- Negative/Placebo Outcome (N): Body Mass Index

NOC example II.

Glynn & Ichino (2019): individuals' sense of political efficacy and attending candidate debates

- Treatment: debate attendance
- Primary Outcome (Y): political efficacy
- Negative/Placebo Outcome (N): knowledge of government (not mentioned in debates)