CodeChella Madrid 2025

Roadmap

Continuous Treatment

- Discrete multi-valued
- Continuous valued dose

Empirical Example

Setup

Individuals are observed in two periods ($t = 0, 1$)

Treatment turns on in the second period for some units

- The untreated group receives no dose, $D_i = 0$
- The treated group receives a dose $D_i > 0$

How do you do difference-in-differences in this setup?

Initial Example

Let’s work through an example to illustrate what’s novel in this setting...

There are a set of patients who sign up for an experimental medicine. They are either not accepted (receive 0mg) or accepted into the program and allowed a positive dose

- Doctor drugs 0mg, 10mg, 20mg

Estimating effects for each dose group

We have three groups:

1. The ‘untreated’ group who received 0mg
2. The ‘10mg’ group
3. The ‘20mg’ group

The simplest thing to do is to treat the ‘10mg’ and ‘20mg’ group as two treatments and compare both to the ‘untreated’ group

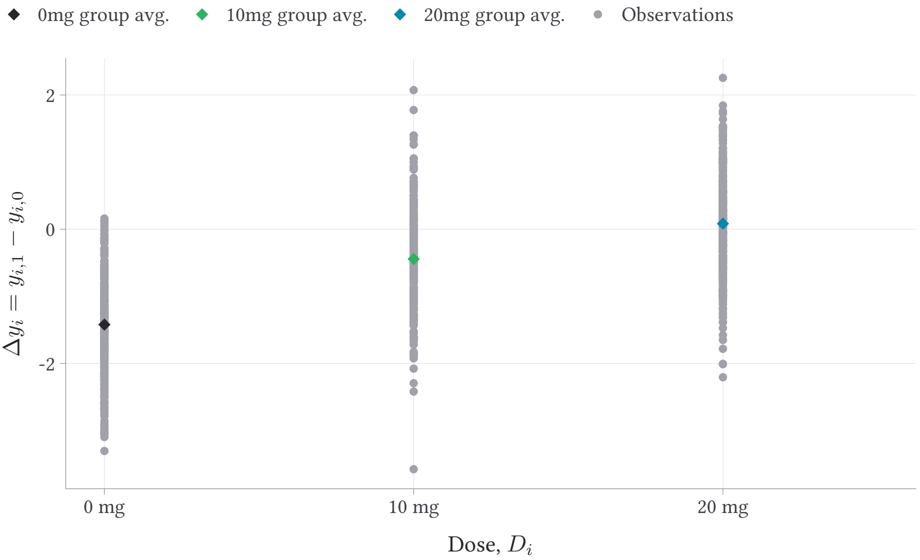
First-differences

The rest of the slides will focus on *first-differences*:

$$\Delta y_i \equiv y_{i1} - y_{i0}$$

Parallel counterfactual trends assumption says that Δy_i would be on average the same for the ‘10mg’ and the untreated group *in the absence of treatment*

- Similarly when looking at the ‘20mg’ group


◆ 0mg group avg.
◆ 10mg group avg.
◆ 20mg group avg.
● Observations
$\Delta y_i = y_{i,1} - y_{i,0}$
2
0
-2
0 mg
10 mg
20 mg
Dose, D_i

Diff-in-Diff

Our diff-in-diff estimate is formed as:

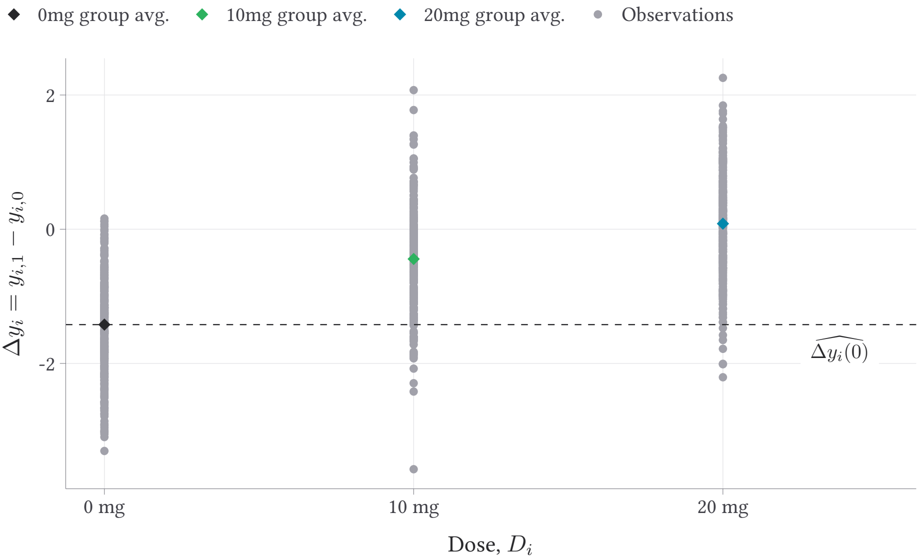
$$\hat{\tau}_{10} = \mathbb{E}[\Delta Y_i \mid D_i = 10] - \mathbb{E}[\Delta Y_i \mid D_i = 0]$$

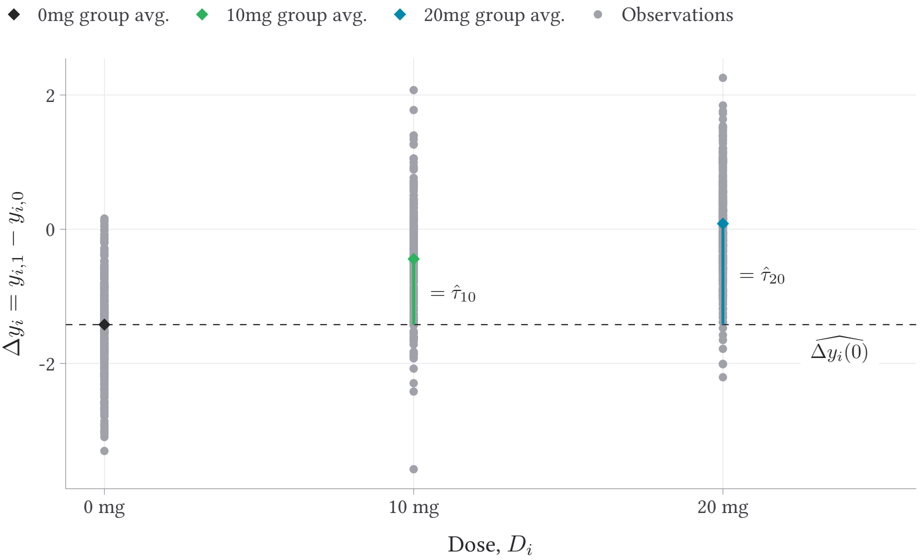
and

$$\hat{\tau}_{20} = \mathbb{E}[\Delta Y_i \mid D_i = 20] - \mathbb{E}[\Delta Y_i \mid D_i = 0]$$

In both cases, we use the untreated group to estimate the counterfactual trend

- $\mathbb{E}[\Delta Y_i \mid D_i = 0]$ is what we think the change in outcomes would have been in the absence of treatment for the treated unit


◆ 0mg group avg.
◆ 10mg group avg.
◆ 20mg group avg.
● Observations
2
0
-2
$\Delta y_i = y_{i,1} - y_{i,0}$
$\widehat{\Delta y_i(0)}$
0 mg
10 mg
20 mg
Dose, D_i



◆ 0mg group avg.
◆ 10mg group avg.
◆ 20mg group avg.
● Observations
2
0
-2
$\Delta y_i = y_{i,1} - y_{i,0}$
$= \hat{\tau}_{10}$
$= \hat{\tau}_{20}$
$\widehat{\Delta y_i(0)}$
0 mg
10 mg
20 mg
Dose, D_i

Estimates

The previous figure shows two difference estimates of $\hat{\tau}_{10}$ and $\hat{\tau}_{20}$

- The 20mg group has a higher estimated treatment effect than the 10mg group. What can we conclude from that?

ATT Estimates

To understand this, let's remember the discussion of ATT vs. ATE

- In the binary DID case, we estimate the average treatment effect *among those who received treatment*

Another way of writing the ATT estimate that will prove useful is

$$ATT \equiv \mathbb{E}[Y_{i1}(1) - Y_{i1}(0) \mid D_i = 1]$$

- The $\mid D_i = 1$ says we are averaging for the group whose treatment was $D_i = 1$

ATT Estimates

$$ATT \equiv \mathbb{E}[Y_{i1}(1) - Y_{i1}(0) \mid D_i = 1]$$

Why might the ATT be different than the ATE?

ATT Estimates

$$ATT \equiv \mathbb{E}[Y_{i1}(1) - Y_{i1}(0) \mid D_i = 1]$$

Why might the ATT be different than the ATE?

The usual “selection on $Y(1)$” story

- Maybe units that benefit most from treatment select into it

ATT Estimates

$$ATT \equiv \mathbb{E}[Y_{i1}(1) - Y_{i1}(0) \mid D_i = 1]$$

Why might the ATT be different than the ATE?

The usual “selection on $Y(1)$” story

- Maybe units that benefit most from treatment select into it
- But, not parallel trends rule out a lot of the $Y(0)$ stories
  - → E.g. Units can not select into treatment because of negative shocks

“ATT” in continuous DID

The concept of ATT vs. ATE is the same, but now there is no single ‘the treated’. Instead, there are two treated groups!
There are now many different potential outcomes

$$Y_{i1}(d),$$

for $d = 0, 10, 20$

- What the outcome would be had the unit taken 0mg, 10mg, or 20mg
- We only observe on potential outcome $Y_{i1}(D_i)$

Treatment effects

We might care about the average effect of moving from 0 to some level $d > 0$:

$$Y_{i1}(d) - Y_{i1}(0)$$

Or, we might want to know about the effect of increasing someone's dose from $d \to d'$

$$Y_{i1}(d') - Y_{i1}(d)$$

- E.g. doctors wonder about increasing someone's dose to the next level

What did our two diff-in-diff estimates identify?

So what do we call the ‘ATT’? ATT$_{10}$ and ATT$_{20}$ perhaps?

Callaway, Goodman-Bacon, and Sant’Anna call this:

$$\text{ATT}(d \mid d) = \mathbb{E}[Y_{i1}(d) - Y_{i1}(0) \mid D_i = d]$$

- Our diff-in-diff’s can only estimate the effect of a certain dose *among the units that received that dose*.

ATT$(d \mid d)$

These dose-specific ATTs are identified from the same parallel trends assumption we are used to:

- Had the units receiving dose $D_i = d$ not received any dose $D_i = 0$, their trend in Y would have evolved the same as the untreated units.

The good news. We know how to think about the parallel trends assumption

- can potentially assess this assumption using pre-trends placebo checks

Comparing estimates

We observe a set of ATT$(d \mid d)$ for a set of doses (in our example 10 and 20)

The initial temptation is to compare these two estimates:

- E.g. finding the ‘correct’ dose of a medicine by picking the highest ATT
- E.g. finding diminishing returns on an treatment investment
- E.g. what is the ‘per-unit’ effect of the treatment?

Thinking about counterfactual treatment effects

We want to think about what would happen if one group of units received a *different* dose?

One example is what Callaway, Goodman-Bacon, and Sant'Anna call the Average Causal Response, ACRT$(d' \mid d)$:

Effect of increasing dose by a little

$$\text{ACRT}(d' \mid d) = \mathbb{E}\Big[Y_{i1}(d' + \varepsilon) - Y_{i1}(d') \mid D_i = d \Big]$$

For the group that received d

Average Causal Response

$$\text{ACRT}(d' \mid d) = \mathbb{E}\Big[Y_{i1}(d'+1) - Y_{i1}(d') \mid D_i = d \Big]$$

- This measures the 'per-unit' effect of the treatment (at some starting point d')
- The effect is averaged over the group of units that received treatment $D_i = d$

Average Causal Response

$$\text{ACRT}(d' \mid d') = \mathbb{E}\Big[Y_{i1}(d'+1) - Y_{i1}(d') \mid D_i = d' \Big]$$

Could imagine estimating this using our two $\text{ATT}(d \mid d)$:

$$\text{ACRT}(d_1 \mid d_1) \approx \frac{\text{ATT}(d_2 \mid d_2) - \text{ATT}(d_1 \mid d_1)}{d_2 - d_1}$$

Thinking about counterfactual treatment effects

Note that two things change whe we move from d_1 to d_2

- The dose amount they receive
- And the group of units we are estimating ATTs for

The latter is what creates difficulty:

- If the group of units receiving treatment at d_1 is different than the group receiving treatment at d_2, then the ATTs are not directly comparable

Thinking about counterfactual treatment effects

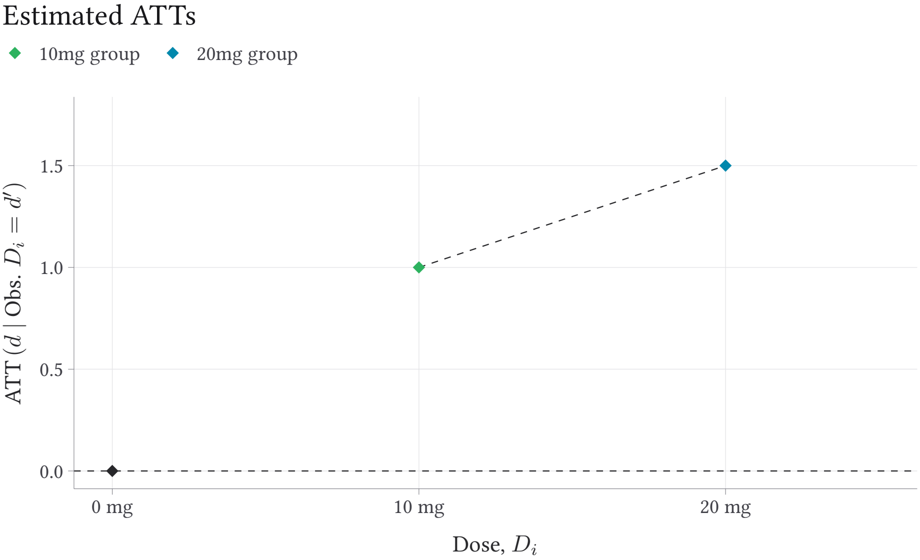
Note that two things change whe we move from d_1 to d_2

- The dose amount they receive
- And the group of units we are estimating ATTs for

The latter is what creates difficulty:

- If the group of units receiving treatment at d_1 is different than the group receiving treatment at d_2, then the ATTs are not directly comparable

Let's dig into this out with our example...


Estimated ATTs
10mg group
20mg group
ATT $(d \mid \text{Obs. } D_i = d')$
1.5
1.0
0.5
0.0
0 mg
10 mg
20 mg
Dose, D_i

Homogeneous dose-ATTs

On the one extreme, we have homogeneity of treatment effects:

$$\text{ATT}(d' \mid d) = \text{ATT}(d')$$

- Treatment effects can depend on dose amount d', but not on which treatment dose-group you are looking at

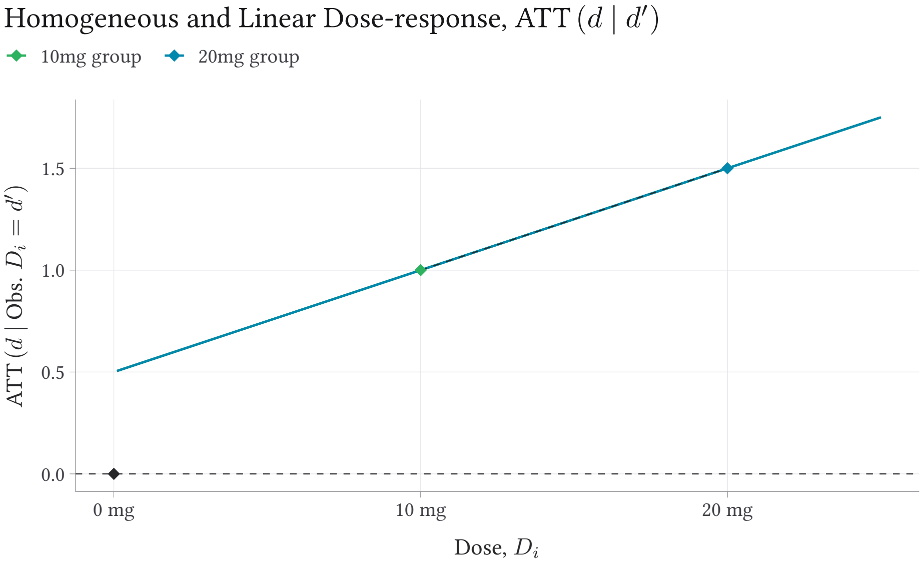
Homogeneous dose-ATTs

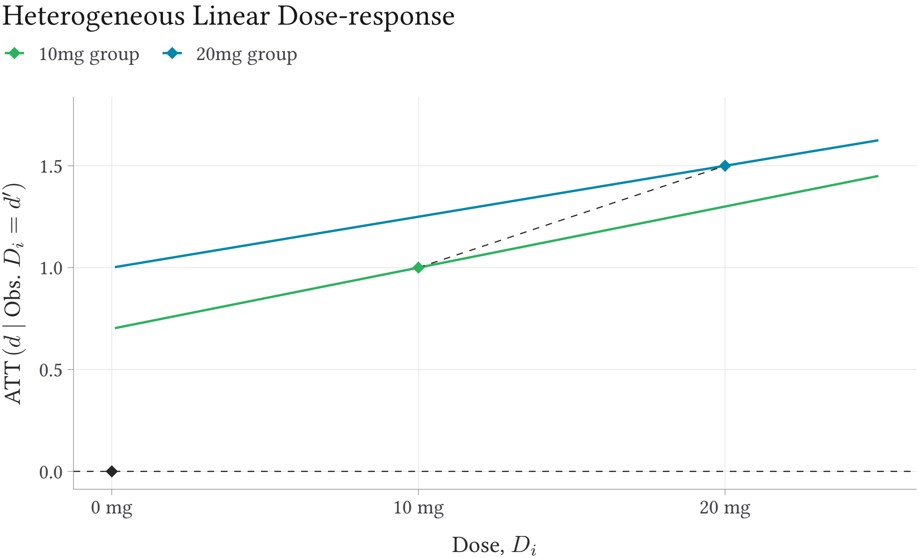
On the one extreme, we have homogeneity of treatment effects:

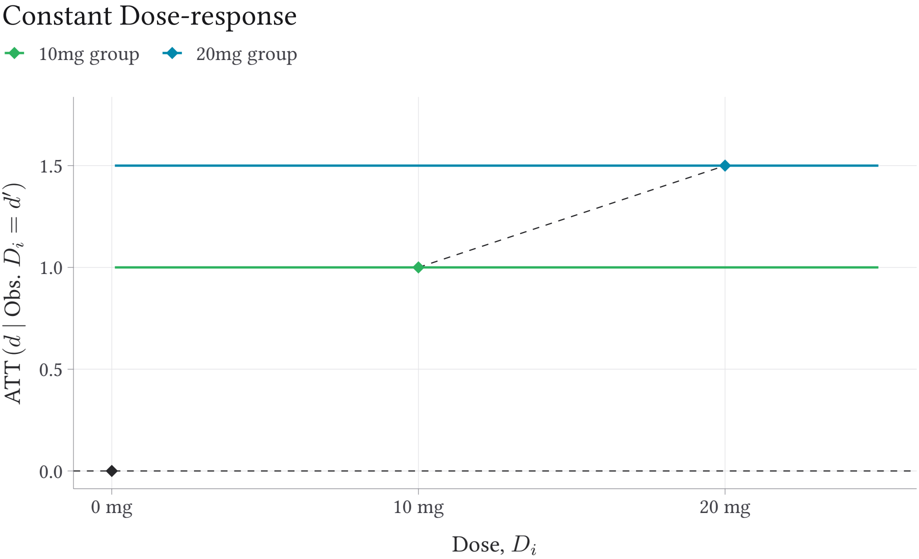
$$\text{ATT}(d' \mid d) = \text{ATT}(d')$$

- Treatment effects can depend on dose amount d', but not on which treatment dose-group you are looking at

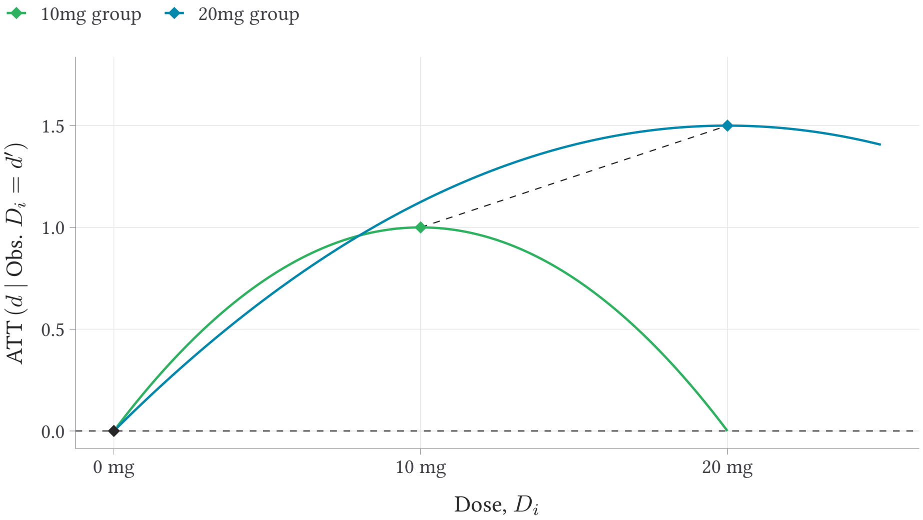
This let's you directly compare $\text{ATT}(d' \mid d')$ and $\text{ATT}(d \mid d)$ to estimate $\text{ACRT}(d' \mid d) = \text{ACRT}(d')$


Homogeneous and Linear Dose-response, ATT $(d \mid d')$
10mg group
20mg group
1.5
1.0
0.5
0.0
0 mg
10 mg
20 mg
ATT $(d \mid \text{Obs. } D_i = d')$
Dose, D_i


Heterogeneous Linear Dose-response
10mg group
20mg group
$\text{ATT}(d \mid \text{Obs. } D_i = d')$
1.5
1.0
0.5
0.0
0 mg
10 mg
20 mg
Dose, D_i


Constant Dose-response
10mg group
20mg group
1.5
1.0
0.5
0.0
0 mg
10 mg
20 mg
$\text{ATT}\,(d \mid \text{Obs. } D_i = d')$
Dose, D_i

People select dose that maximizes their outcome



Recap

So as a recap:

1. Dose-group specific DID estimates identify the ATT$(d \mid d)$ effects

2. Comparing across different doses does not inform us of the marginal effect of changing a dose, the ACR$(d \mid d)$

“Strong Parallel Trends”

In Callaway, Goodman-Bacon, and Sant’Anna, they refer to an assumption called “Strong Parallel Trends” which combines two assumptions:

1. The parallel counterfactual trends assumption: each dose group has the same counterfactual trends as the untreated group
2. Treatment effect homogeneity across groups: $\text{ATT}(d \mid d) = \text{ATT}(d)$
   - → $\implies \text{ACR}(d \mid d) = \text{ACR}(d)$

“Strong Parallel Trends”

In Callaway, Goodman-Bacon, and Sant’Anna, they refer to an assumption called “Strong Parallel Trends” which combines two assumptions:

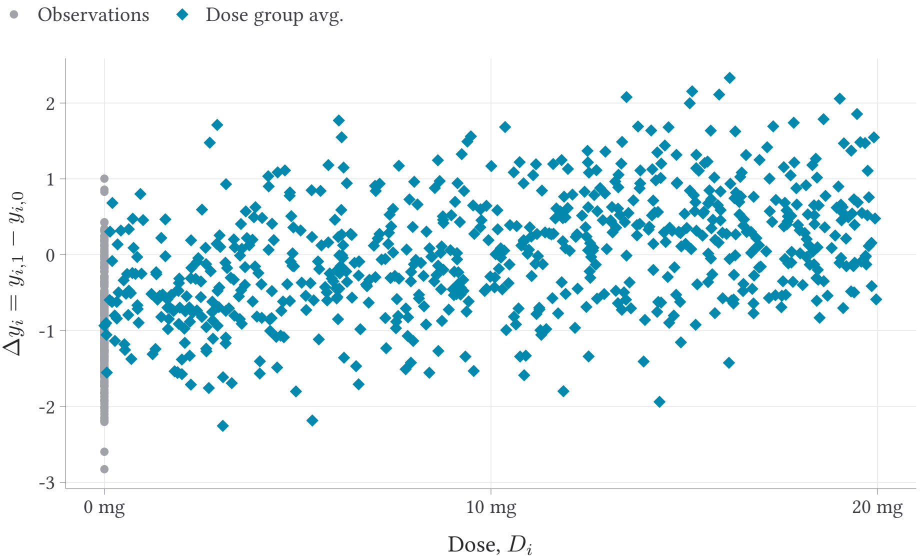
1. The parallel counterfactual trends assumption: each dose group has the same counterfactual trends as the untreated group
2. Treatment effect homogeneity across groups: $\text{ATT}(d \mid d) = \text{ATT}(d)$
   - → $\implies \text{ACR}(d \mid d) = \text{ACR}(d)$

The latter is only needed to identify the ACR curve using your $\text{ATT}(d \mid d)$ estimates

Doctor example extended

Now, let’s say we have a continuous distribution of doses

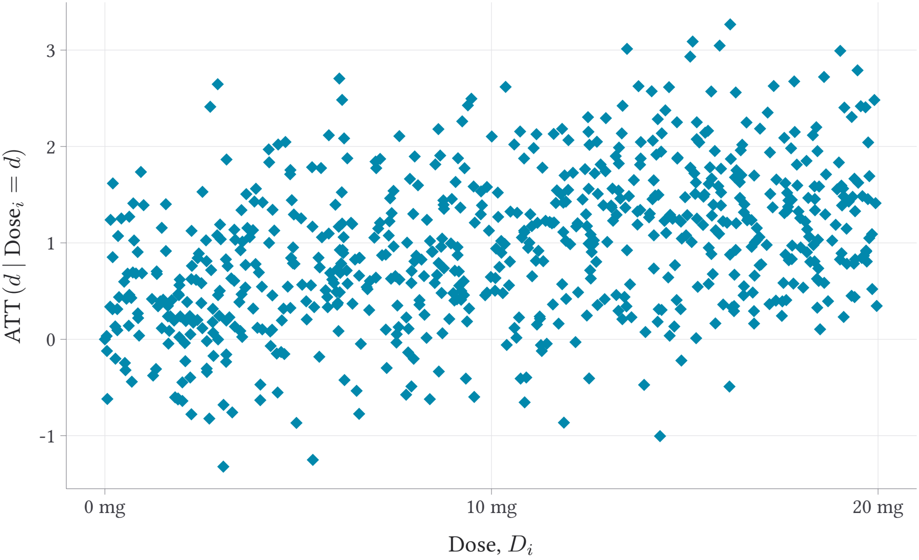
- Some units receive 0mg
- Others receive some amount between 0 and 20mg


● Observations
◆ Dose group avg.
$\Delta y_i = y_{i,1} - y_{i,0}$
2
1
0
-1
-2
-3
0 mg
10 mg
20 mg
Dose, D_i

Dose-specific DID estimates

Our previous method of having dose-specific DID estimates will be quite noisy in this setting

- Very few individuals with the *same* dose
- Throwing out information of individuals with very similar dose



ATT $(d \mid \text{Dose}_i = d)$
3
2
1
0
-1
0 mg
10 mg
20 mg
Dose, D_i

Alternative strategies

Our previous strategy clearly will not suffice

An alternative strategy is to *pool* observations with similar doses to estimate an average $\mathbb{E}[\Delta Y_i \mid D_i = d]$

- E.g. say you break doses into 2mg bins (0–2, 2–4, etc.) and take averages in those bins
- Or any other of your favorite non-parametric method for estimating

ATT($d \mid d$) estimators

In general, the ATT($d \mid d$) estimators will take the form of

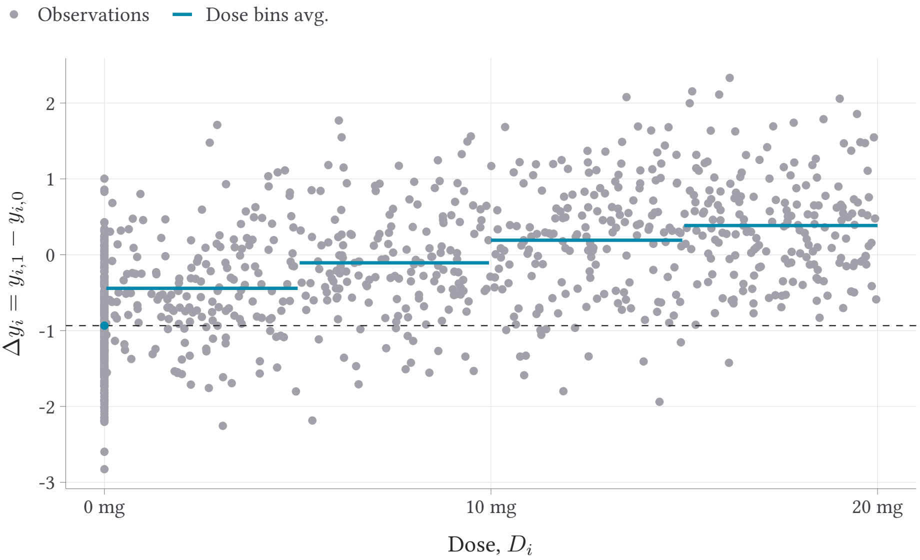
$$\hat{\mathbb{E}}[\Delta Y_i \mid D_i = d] - \hat{\mathbb{E}}[\Delta Y_i \mid D_i = 0]$$

- The second term is just the average change in Y from the control group
- The first-term is where we have some options

Option 1: Bins

The simplest procedure is to ‘discretize’ doses into a set of bins

Then, form a DID estimate for each bin


● Observations
▬ Dose bins avg.
$\Delta y_i = y_{i,1} - y_{i,0}$
2
1
0
-1
-2
-3
0 mg
10 mg
20 mg
Dose, D_i

Pros and Cons of Bins

Advantges:

- Creates a discrete set of groups that are easy to discuss in your paper
- Event-studies are easy to estimate (one for each bin)

Disadvantages:

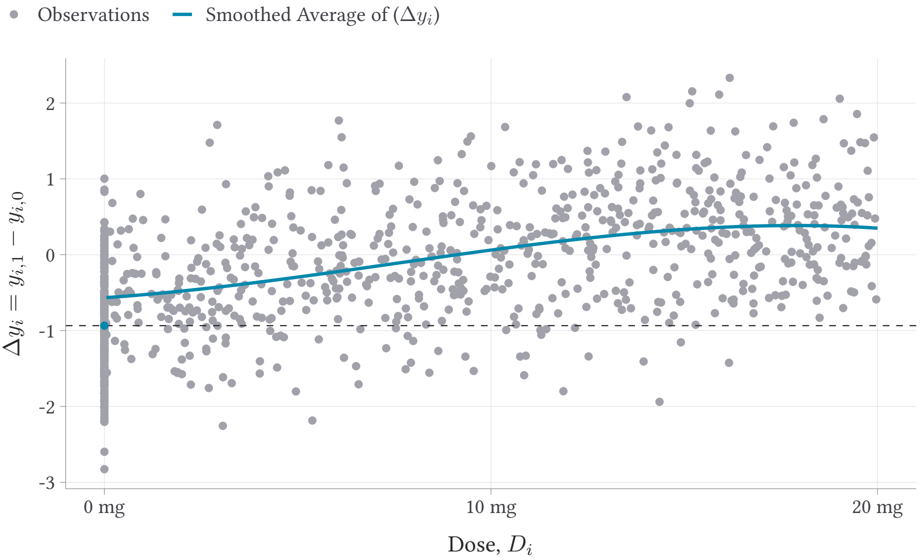
- Bins create a flat surface that is likely not perfectly specified (but gets close as the number of bins grow)

Option 2: Non-parametric smoothing

Alternatively, $\hat{\mathbb{E}}[\Delta Y_i \mid D_i = d]$ can be estimated with non-parametric methods

The simplest way is to use non-parametric method on $\Delta Y_i - \hat{\mathbb{E}}[\Delta Y_i \mid D_i = 0]$ with only the $D_i > 0$ observations.

- The result of this will be the DID estimates
- Standard errors produced should be correct (can always bootstrap to be safe)



● Observations
Smoothed Average of (Δy_i)
$\Delta y_i = y_{i,1} - y_{i,0}$
2
1
0
-1
-2
-3
0 mg
10 mg
20 mg
Dose, D_i

Pros and Cons of Bins

Advantges:

- Better estimates the smooth evolution of $\hat{\mathbb{E}}[\Delta Y_i \mid D_i = d]$

Disadvantages:

- Event-studies are hard to display (one plot for each event-time?)

Roadmap

Continuous Treatment

- Discrete multi-valued
- Continuous valued dose

Empirical Example

Lu and Yu, 2015

This example will use data from “Trade Liberalization and Markup Dispersion: Evidence from China’s WTO Accession” (Lu and Yu, American Economic Journal: Applied Economics 2015)

China’s WTO Ascension

The paper uses the entrance of China into the World Trade Organization (WTO) in 2002

When entering the WTO, tariffs at the industry level were expected to drop to at most 10%

Defining continuous treatment variable

The research strategy is to compare industries (at the 3-digit SIC level) that had higher or lower pre-existing tariff rates

- After the entrance in 2002, industries with higher tariff had *larger* decreases in tariffs $\implies$ a larger shock in import competition

Can define treatment as $D_i = \max(0, \text{Tariff}_{2001} - 0.10)$

- How much tariffs are expected to drop

Outcome variable: Markup Dispersion

The dataset is an industry-by-year panel dataset

- Each row is an industry in China in a given year

The outcome variable is the industry-level *Theil Markup Index*, which measures how much spread in markups at the firm level

- Higher markup index $\implies$ less competitive market

Outcome variable: Markup Dispersion

The dataset is an industry-by-year panel dataset

- Each row is an industry in China in a given year

The outcome variable is the industry-level *Theil Markup Index*, which measures how much spread in markups at the firm level

- Higher markup index $\implies$ less competitive market

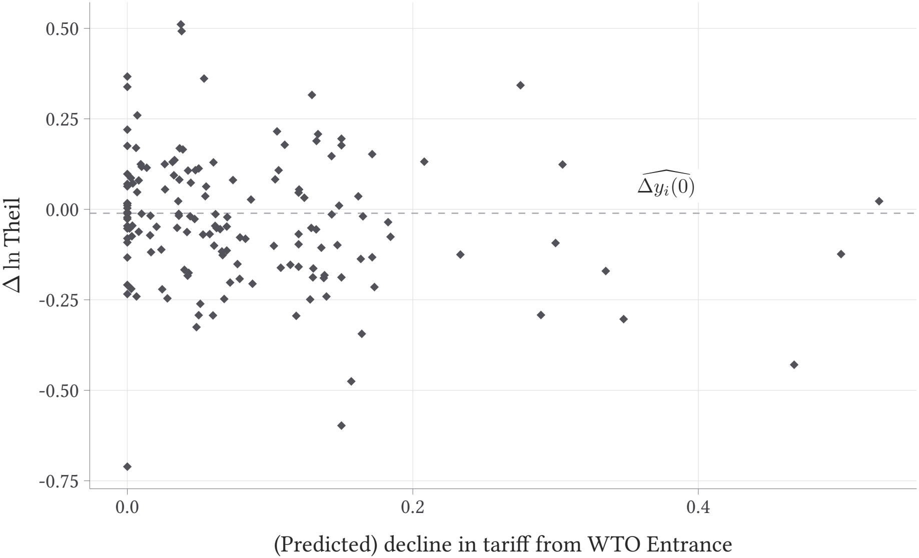
Hypothesis: Larger decrease in tariffs $\implies$ more competition $\implies$ decrease in markup dispersion

2×2 setup

We will use 2000 ($t = 0$) and 2004 ($t = 1$) as our initial years

Let's look at the raw data

- x-axis is our treatment vairable $D_i = \max(0, \text{Tariff}_{2001} - 0.10)$
- y-axis is the change in markup dispersion $\Delta \ln \text{Theil}_i = \ln \text{Theil}_{i,2004} - \ln \text{Theil}_{i,2000}$



Δ ln Theil
0.50
0.25
0.00
-0.25
-0.50
-0.75
0.0
0.2
0.4
$\widehat{\Delta y_i(0)}$
(Predicted) decline in tariff from WTO Entrance

Parallel Trends assumption

Our identification assumption is that if China did not enter the WTO, counterfactual changes in markup dispersion is the same for the unaffected group and the affected doses

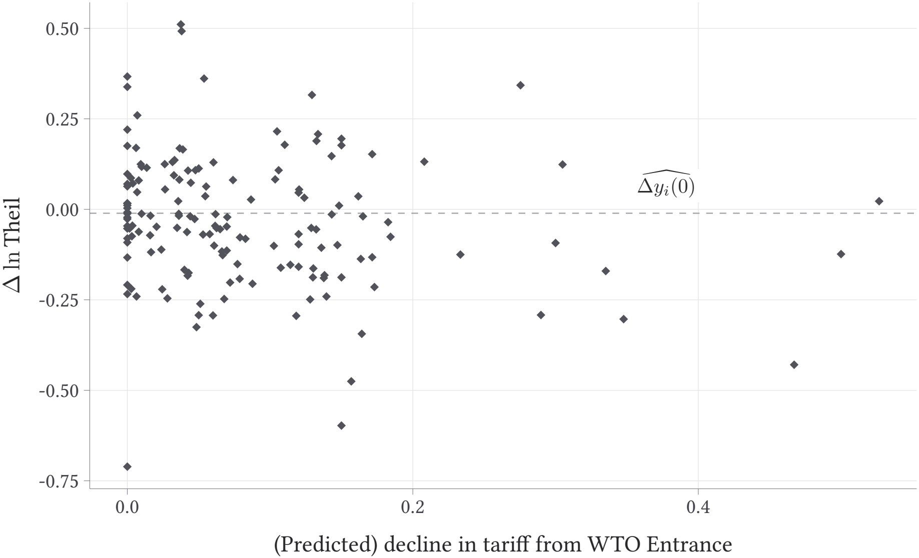
- We can assess the plausability using pre-periods (later)

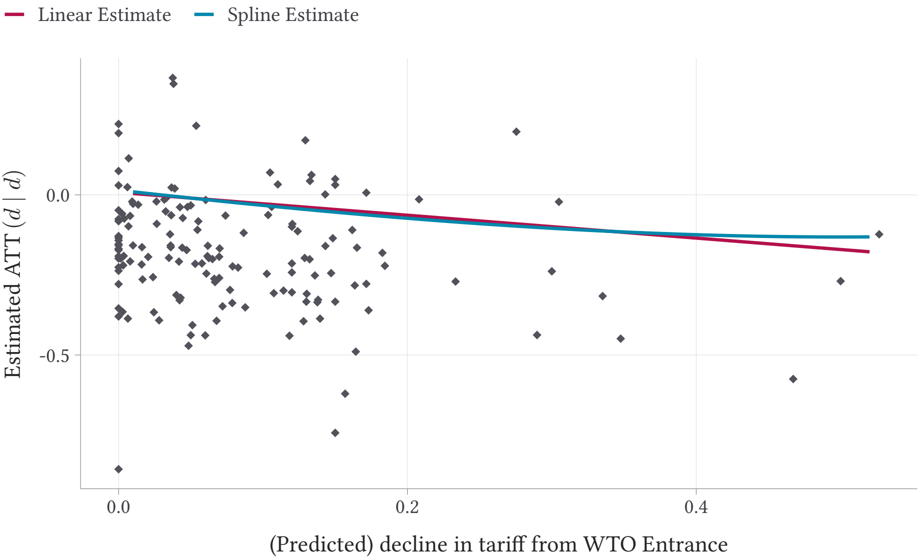
Parallel Trends assumption

Our identification assumption is that if China did not enter the WTO, counterfactual changes in markup dispersion is the same for the unaffected group and the affected doses

- We can assess the plausability using pre-periods (later)

To estimate counterfactual trend, we will take average $\Delta \ln \text{Theil}_i$ for the $D_i = 0$ group


0.50
0.25
0.00
-0.25
-0.50
-0.75
Δ ln Theil
$\widehat{\Delta y_i(0)}$
0.0
0.2
0.4
(Predicted) decline in tariff from WTO Entrance

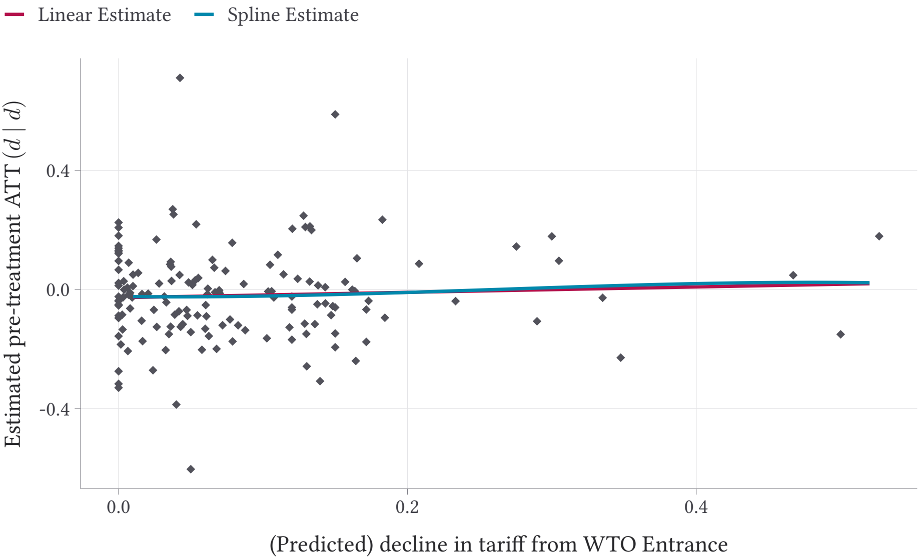

Linear Estimate
Spline Estimate
Estimated ATT $(d \mid d)$
0.0
-0.5
0.0
0.2
0.4
(Predicted) decline in tariff from WTO Entrance

Pre-trends ‘test’

Redo the previous estimates using

$$\Delta \ln \text{Theil}_i = \ln \text{Theil}_{i,1998} - \ln \text{Theil}_{i,2000}$$

We are hoping that the pre-trend change in markup dispersion is uncorrelated with the dose received, D_i



Linear Estimate
Spline Estimate
Estimated pre-treatment ATT $(d \mid d)$
0.4
0.0
-0.4
0.0
0.2
0.4
(Predicted) decline in tariff from WTO Entrance