Two-way fixed effects and difference-in-differences: problems and solutions

by Hugo Jales

May 21, 2024

Overview

We will take a panoramic view of the modern difference-in-differences literature.

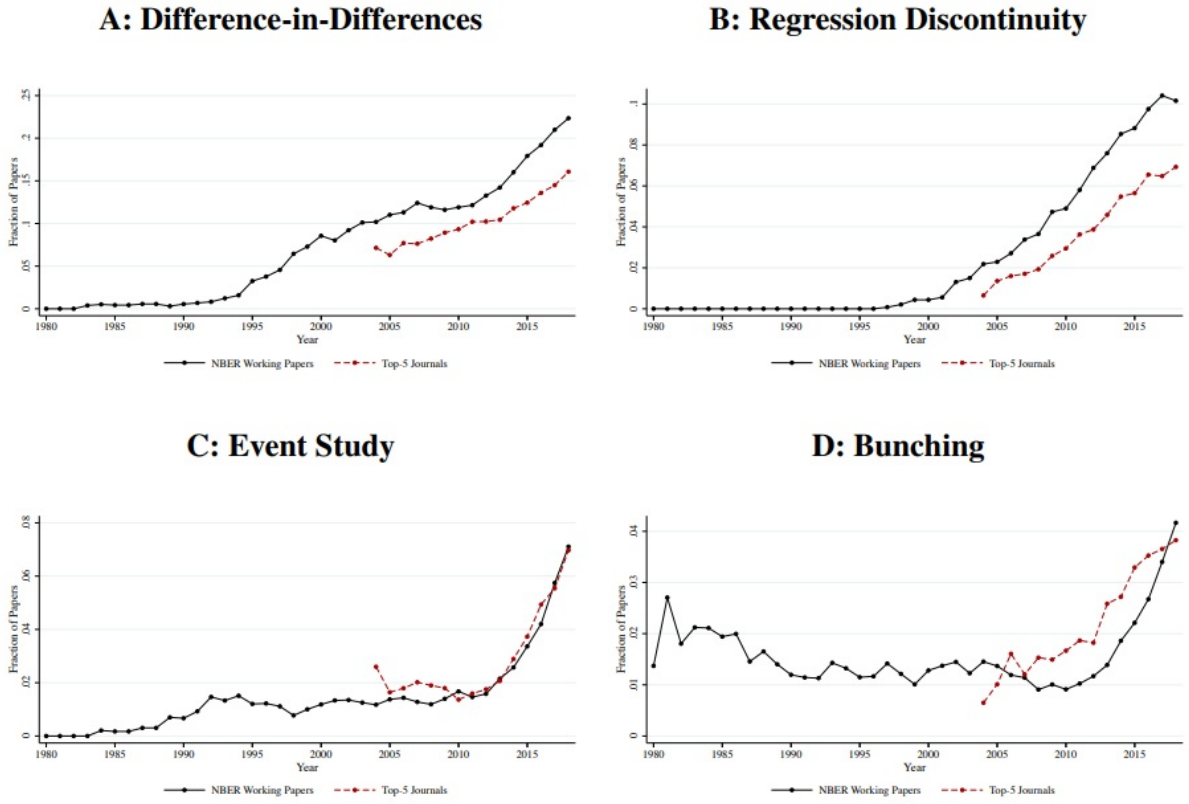
Overview

We will take a panoramic view of the modern difference-in-differences literature.

Our goal is to strengthen our intuition for how these methods work.

DiD

Figure IV: Quasi-Experimental Methods



Notes: This figure shows the fraction of papers referring to each type of quasi-experimental approach. See Table A.I for a list of terms. The series show 5-year moving averages.

Source: Kleven (Annual Review of Economics)

Standard DiD

Suppose we are interested on estimating the magnitude of a particular causal effect.

Standard DiD

Suppose we are interested on estimating the magnitude of a particular causal effect.

We don't have a randomized trial. Our control group is different to the treatment group in important ways.

Standard DiD

Suppose we are interested on estimating the magnitude of a particular causal effect.

We don't have a randomized trial. Our control group is different to the treatment group in important ways.

However, luckily for us, these differences are *constant* over time. In the literature, this is called the parallel paths assumption.

Standard DiD

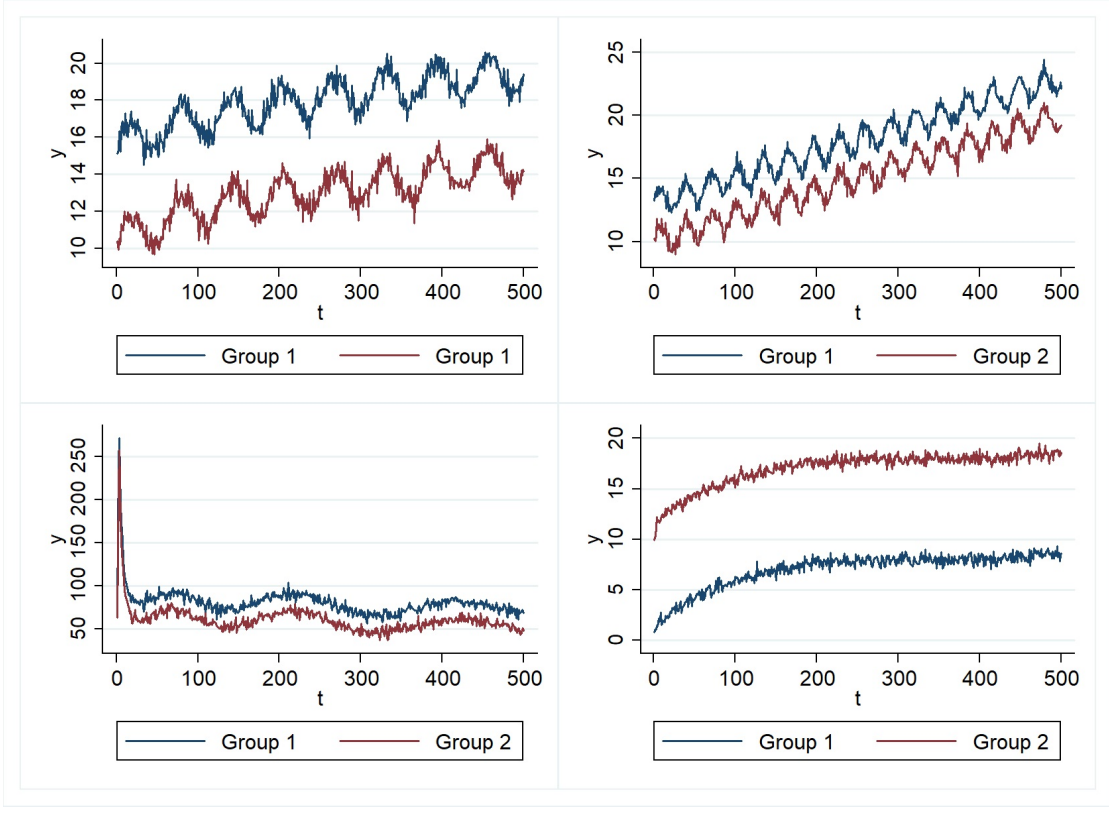
Suppose we are interested on estimating the magnitude of a particular causal effect.

We don't have a randomized trial. Our control group is different to the treatment group in important ways.

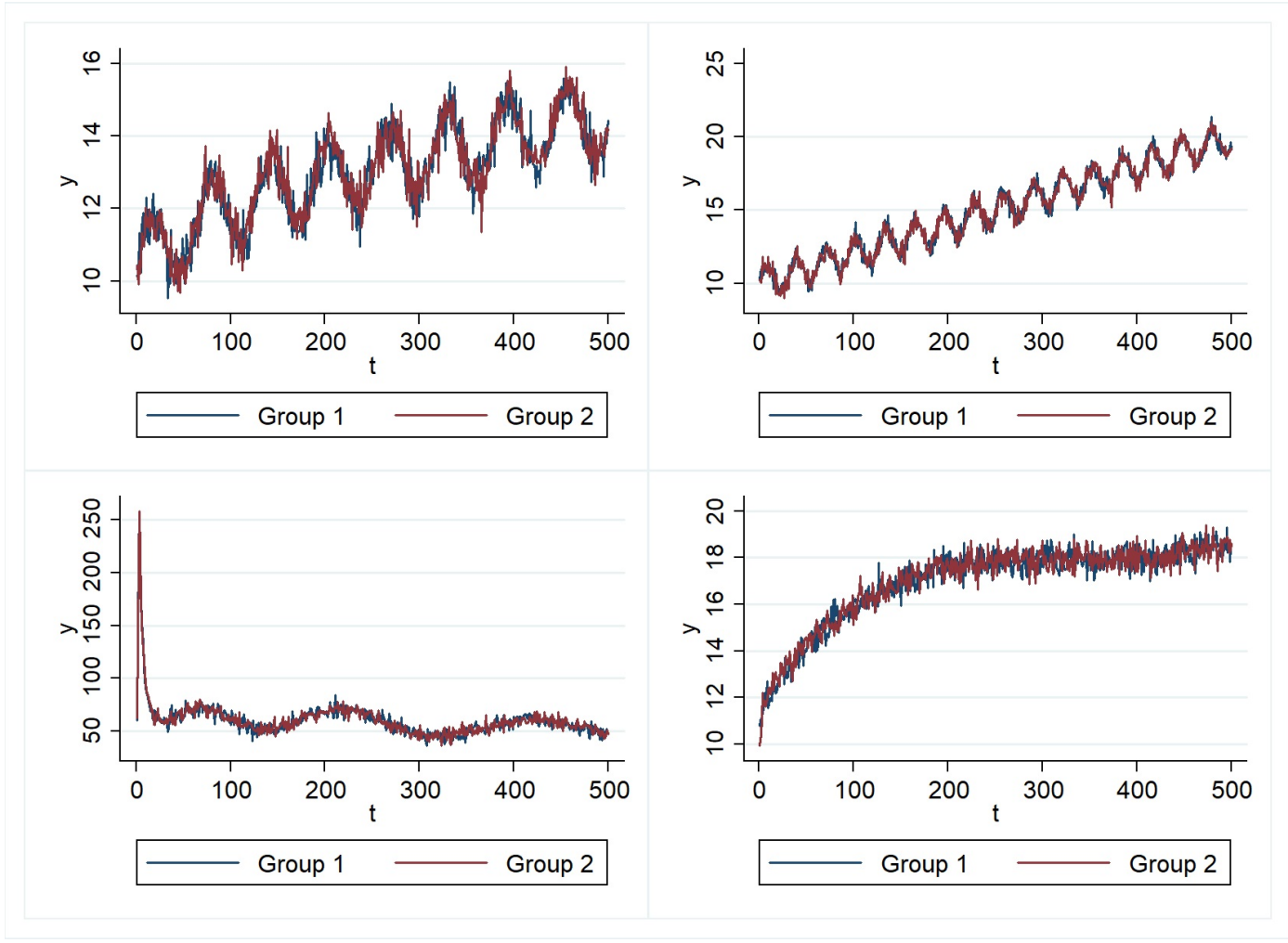
However, luckily for us, these differences are *constant* over time. In the literature, this is called the parallel paths assumption.

Thus, if we observe their outcomes both before and after the treatment, we might be able to recover the causal effect.

Parallel Paths



RCT Paths


y
10
12
14
16
0
100
200
300
400
500
t
Group 1
Group 2
y
10
15
20
25
0
100
200
300
400
500
t
Group 1
Group 2
y
50
100
150
200
250
0
100
200
300
400
500
t
Group 1
Group 2
y
10
12
14
16
18
20
0
100
200
300
400
500
t
Group 1
Group 2

Standard DiD - Strategy

Suppose we just compare the outcomes of the treatment and control groups after the treatment was applied to the former group.

Standard DiD - Strategy

Suppose we just compare the outcomes of the treatment and control groups after the treatment was applied to the former group.

This should not work unless we are in the world of an RCT.

Standard DiD - Strategy

Suppose we just compare the outcomes of the treatment and control groups after the treatment was applied to the former group.

This should not work unless we are in the world of an RCT. It recovers a combination (sum) of the actual causal effect of the policy (ATT) and the underlying differences between these groups (selection bias).

Standard DiD - Strategy

Suppose we just compare the outcomes of the treatment and control groups after the treatment was applied to the former group.

This should not work unless we are in the world of an RCT. It recovers a combination (sum) of the actual causal effect of the policy (ATT) and the underlying differences between these groups (selection bias).

Now, suppose that we instead just compare the outcomes of the treatment and control groups *before* the treatment.

Standard DiD - Strategy

Suppose we just compare the outcomes of the treatment and control groups after the treatment was applied to the former group.

This should not work unless we are in the world of an RCT. It recovers a combination (sum) of the actual causal effect of the policy (ATT) and the underlying differences between these groups (selection bias).

Now, suppose that we instead just compare the outcomes of the treatment and control groups *before* the treatment. Now, there shouldn't be any effect of the treatment, so all that we can recover is a measure of these underlying differences between the groups.

Standard DiD - Strategy

Suppose we just compare the outcomes of the treatment and control groups after the treatment was applied to the former group.

This should not work unless we are in the world of an RCT. It recovers a combination (sum) of the actual causal effect of the policy (ATT) and the underlying differences between these groups (selection bias).

Now, suppose that we instead just compare the outcomes of the treatment and control groups *before* the treatment. Now, there shouldn't be any effect of the treatment, so all that we can recover is a measure of these underlying differences between the groups.

If that measure is stable over time (parallel paths assumption), then you can use it to correct the errors of your first attempt above.

Standard DiD

Notation:

- Y: Observed outcome
- D: Binary Treatment Group Indicator
- T: Binary, before and after the treatment was implemented.
- $Y(1)$: Potential Outcome if exposed to the treatment
- $Y(0)$: Potential Outcome if not exposed to the treatment
- $Y_{it} = Y_{it}(1)D_{it} + Y_{it}(0)(1 - D_{it})$
  where D_{it} is an indicator that individual i at time t is treated ($D_{it} = DT$).

Standard DiD

Notation:

- Y: Observed outcome
- D: Binary Treatment Group Indicator
- T: Binary, before and after the treatment was implemented.
- $Y(1)$: Potential Outcome if exposed to the treatment
- $Y(0)$: Potential Outcome if not exposed to the treatment
- $Y_{it} = Y_{it}(1)D_{it} + Y_{it}(0)(1 - D_{it})$
  where D_{it} is an indicator that individual i at time t is treated ($D_{it} = DT$).

You only observe (Y, D, T), that is, individuals' outcomes, their treatment status, and the data in which you record both the outcome and the treatment status.

Standard DiD

Notation:

- Y: Observed outcome
- D: Binary Treatment Group Indicator
- T: Binary, before and after the treatment was implemented.
- $Y(1)$: Potential Outcome if exposed to the treatment
- $Y(0)$: Potential Outcome if not exposed to the treatment
- $Y_{it} = Y_{it}(1)D_{it} + Y_{it}(0)(1 - D_{it})$
  where D_{it} is an indicator that individual i at time t is treated ($D_{it} = DT$).

You only observe (Y, D, T), that is, individuals' outcomes, their treatment status, and the data in which you record both the outcome and the treatment status.

That is, you observe $Y(1)$ when the individual is treated and $Y(0)$ when the individual is not treated.

Standard DiD

Notation:

- Y: Observed outcome
- D: Binary Treatment Group Indicator
- T: Binary, before and after the treatment was implemented.
- $Y(1)$: Potential Outcome if exposed to the treatment
- $Y(0)$: Potential Outcome if not exposed to the treatment
- $Y_{it} = Y_{it}(1)D_{it} + Y_{it}(0)(1 - D_{it})$
  where D_{it} is an indicator that individual i at time t is treated ($D_{it} = DT$).

You only observe (Y, D, T), that is, individuals' outcomes, their treatment status, and the data in which you record both the outcome and the treatment status.

That is, you observe $Y(1)$ when the individual is treated and $Y(0)$ when the individual is not treated.

Causal Effect: $Y(1) - Y(0)$

DiD

You can easily get:

$$E[Y|D,T] - E[Y|D',T] = E[Y(1)|D,T] - E[Y(0)|D',T]$$

What you want is:

$$E[Y(1)|D,T] - E[Y(0)|D,T]$$

You can look at their difference and you realize that it is:

$$E[Y(0)|D,T] - E[Y(0)|D',T]$$

In other words, your comparison of outcomes has everything that the treatment has caused on the treated group, plus a measure of how appropriate is your control group.

$$E[Y|D,T] - E[Y|D',T] = ATT + \textit{Selection Bias}$$

DiD

Now, what if you do the same thing, but use data before the treatment:

$$E[Y|D, T'] - E[Y|D', T'] = E[Y(0)|D, T'] - E[Y(0)|D', T']$$

Since we are before the treatment, this comparison has no causal effect of the treatment on it. It can only contain underlying differences between these populations.

DiD

Now, what if you do the same thing, but use data before the treatment:

$$E[Y|D, T'] - E[Y|D', T'] = E[Y(0)|D, T'] - E[Y(0)|D', T']$$

Since we are before the treatment, this comparison has no causal effect of the treatment on it. It can only contain underlying differences between these populations.

Parallel paths assumption: This bias is the same in every period. That is:

$$E[Y(0)|D, T'] - E[Y(0)|D', T'] = E[Y(0)|D, T] - E[Y(0)|D', T]$$

DiD

Thus, you can define the DiD estimand as:

$$ATT_{DiD} = \left(E[Y|D,T] - E[Y|D',T]\right) - \left(E[Y|D,T'] - E[Y|D',T']\right)$$

DiD

Thus, you can define the DiD estimand as:

$$ATT_{DiD} = \left(E[Y|D,T] - E[Y|D',T]\right) - \left(E[Y|D,T'] - E[Y|D',T']\right)$$

Hence the name "difference-in-differences".

DiD

We sometimes make things look more complicated than they are.
Let's move some things around:

$$ATT_{DiD} = (E[Y|D,T] - E[Y|D',T]) - (E[Y|D,T'] - E[Y|D',T'])$$

DiD

We sometimes make things look more complicated than they are. Let's move some things around:

$$ATT_{DiD} = (E[Y|D,T] - E[Y|D',T]) - (E[Y|D,T'] - E[Y|D',T'])$$

$$ATT_{DiD} = (E[Y|D,T] - E[Y|D,T']) - (E[Y|D',T] - E[Y|D',T'])$$

Define the change in Y over time as $\Delta y = Y_t - Y_{t-1}$

$$ATT_{DiD} = E[\Delta y|D] - E[\Delta y|D']$$

Difference-in-differences is really a comparison of means between treated and control groups. It is just not in the levels of their outcomes but in changes/growth.

$ATT_{DiD} = E[\Delta y|D] - E[\Delta y|D']$

DiD is a comparison of means between treated and control groups. But, instead of the levels of their outcomes but in changes/growth.

$$ATT_{DiD} = E[\Delta y|D] - E[\Delta y|D']$$

DiD is a comparison of means between treated and control groups. But, instead of the levels of their outcomes but in changes/growth.

In an RCT, the treatment status is independent of any feature of the process of potential outcomes (including levels).

$ATT_{DiD} = E[\Delta y|D] - E[\Delta y|D']$

DiD is a comparison of means between treated and control groups. But, instead of the levels of their outcomes but in changes/growth.

In an RCT, the treatment status is independent of any feature of the process of potential outcomes (including levels).
$(Cov(Y(0), D) = 0 \rightarrow E[Y(0)|D] = E[Y(0)|D'])$)

$ATT_{DiD} = E[\Delta y|D] - E[\Delta y|D']$

DiD is a comparison of means between treated and control groups. But, instead of the levels of their outcomes but in changes/growth.

In an RCT, the treatment status is independent of any feature of the process of potential outcomes (including levels).
$(Cov(Y(0), D) = 0 \rightarrow E[Y(0)|D] = E[Y(0)|D'])$)

In a DiD, the treatment status is *not* independent of the determinant of the levels of potential outcomes. But the treatment status is (mean) independent of *changes* in potential outcomes.

$ATT_{DiD} = E[\Delta y|D] - E[\Delta y|D']$

DiD is a comparison of means between treated and control groups. But, instead of the levels of their outcomes but in changes/growth.

In an RCT, the treatment status is independent of any feature of the process of potential outcomes (including levels).
$(Cov(Y(0), D) = 0 \rightarrow E[Y(0)|D] = E[Y(0)|D'])$)

In a DiD, the treatment status is *not* independent of the determinant of the levels of potential outcomes. But the treatment status is (mean) independent of *changes* in potential outcomes.
$(Cov(\Delta Y(0), D) = 0$

$ATT_{DiD} = E[\Delta y|D] - E[\Delta y|D']$

DiD is a comparison of means between treated and control groups. But, instead of the levels of their outcomes but in changes/growth.

In an RCT, the treatment status is independent of any feature of the process of potential outcomes (including levels).
$(Cov(Y(0), D) = 0 \ \rightarrow E[Y(0)|D] = E[Y(0)|D'])$)

In a DiD, the treatment status is *not* independent of the determinant of the levels of potential outcomes. But the treatment status is (mean) independent of *changes* in potential outcomes.
$(Cov(\Delta Y(0), D) = 0 \ \rightarrow E[\Delta Y(0)|D] = E[\Delta Y(0)|D'])$.

$ATT_{DiD} = E[\Delta y|D] - E[\Delta y|D']$

DiD is a comparison of means between treated and control groups. But, instead of the levels of their outcomes but in changes/growth.

In an RCT, the treatment status is independent of any feature of the process of potential outcomes (including levels).
$(Cov(Y(0), D) = 0 \ \rightarrow E[Y(0)|D] = E[Y(0)|D'])$)

In a DiD, the treatment status is *not* independent of the determinant of the levels of potential outcomes. But the treatment status is (mean) independent of *changes* in potential outcomes.
$(Cov(\Delta Y(0), D) = 0 \ \rightarrow E[\Delta Y(0)|D] = E[\Delta Y(0)|D'])$.

Units that would grow more over time have the same likelihood of being assigned to the treatment and control groups. So any difference in the change in the outcomes of treated and control groups can only be attributed to the treatment itself.

DiD

In other words, the values of outcomes of controls are not valid counterfactuals for the outcomes of the treated.

DiD

In other words, the values of outcomes of controls are not valid counterfactuals for the outcomes of the treated.

But changes in the outcomes of controls are valid counterfactuals for the outcomes of the treated.

DiD

In other words, the values of outcomes of controls are not valid counterfactuals for the outcomes of the treated.

But changes in the outcomes of controls are valid counterfactuals for the outcomes of the treated.

But if you have the counterfactual changes, and baseline levels, you can directly construct counterfactual outcomes of the treated.

DiD

In other words, the values of outcomes of controls are not valid counterfactuals for the outcomes of the treated.

But changes in the outcomes of controls are valid counterfactuals for the outcomes of the treated.

But if you have the counterfactual changes, and baseline levels, you can directly construct counterfactual outcomes of the treated.

You pick their levels before the treatment and apply your identified (from the controls) growth.
That is:

$$E[Y(0)|D,T] = E[Y(0)|D,T'] + E[\Delta y|D']$$

Counterfactual values of the outcome of the treated (LHS) are equal to baseline + growth/change in controls.

DiD

This should help you to think about potential justifications for a DiD strategy in your research.

DiD

This should help you to think about potential justifications for a DiD strategy in your research.

The key is to think about what drives the process of changes over time in outcomes (Solow models for GDP, growth charts for height and weight of kids, future value formulas for balances in investments and retirement accounts, SIR models for pandemics, etc.).

DiD

This should help you to think about potential justifications for a DiD strategy in your research.

The key is to think about what drives the process of changes over time in outcomes (Solow models for GDP, growth charts for height and weight of kids, future value formulas for balances in investments and retirement accounts, SIR models for pandemics, etc.). Theory can be helpful here.

DiD

This should help you to think about potential justifications for a DiD strategy in your research.

The key is to think about what drives the process of changes over time in outcomes (Solow models for GDP, growth charts for height and weight of kids, future value formulas for balances in investments and retirement accounts, SIR models for pandemics, etc.). Theory can be helpful here.

See Marx, Temer, and Tang (2022) for a discussion of when/how parallel paths fit into the models of dynamic choices that we are familiar with (search models, optimal stopping, learning models, etc.).

DiD

This should help you to think about potential justifications for a DiD strategy in your research.

The key is to think about what drives the process of changes over time in outcomes (Solow models for GDP, growth charts for height and weight of kids, future value formulas for balances in investments and retirement accounts, SIR models for pandemics, etc.). Theory can be helpful here.

See Marx, Temer, and Tang (2022) for a discussion of when/how parallel paths fit into the models of dynamic choices that we are familiar with (search models, optimal stopping, learning models, etc.). See Ghanem et al. (2022) for which kinds of selection patterns parallel paths allow.

Estimation

$$ATT_{DiD} = (E[Y|D,T] - E[Y|D,T']) - (E[Y|D',T] - E[Y|D',T'])$$

Note that one could construct an estimator by using the empirical analog of the equation above.

Estimation

$$ATT_{DiD} = (E[Y|D,T] - E[Y|D,T']) - (E[Y|D',T] - E[Y|D',T'])$$

Note that one could construct an estimator by using the empirical analog of the equation above.

But one quick and easy way to do exactly the same thing *in this setting* is to run a regression.

Estimation

$$ATT_{DiD} = (E[Y|D,T] - E[Y|D,T']) - (E[Y|D',T] - E[Y|D',T'])$$

Note that one could construct an estimator by using the empirical analog of the equation above.

But one quick and easy way to do exactly the same thing *in this setting* is to run a regression.

We can run the so-called two-way fixed effects model.

Estimation

$$ATT_{DiD} = (E[Y|D,T] - E[Y|D,T']) - (E[Y|D',T] - E[Y|D',T'])$$

Note that one could construct an estimator by using the empirical analog of the equation above.

But one quick and easy way to do exactly the same thing *in this setting* is to run a regression.

We can run the so-called two-way fixed effects model.

Why would anyone want to do that?

Estimation

$$ATT_{DiD} = (E[Y|D,T] - E[Y|D,T']) - (E[Y|D',T] - E[Y|D',T'])$$

Note that one could construct an estimator by using the empirical analog of the equation above.

But one quick and easy way to do exactly the same thing *in this setting* is to run a regression.

We can run the so-called two-way fixed effects model.

Why would anyone want to do that?

My guess is that it is because it saves us the trouble of computing confidence intervals by hand.

Estimation

Consider the following regression specification:

$$y_{it} = \beta_0 + \beta_1 D_i + \beta_2 T_t + \beta_3 (DT)_{it} + \epsilon_{it}$$

The interaction term in this regression is *numerically* identical to the empirical analog of the DiD estimand.

Intuition

Intuition

OLS attempts to fit the conditional mean function as best as it can.

Intuition

OLS attempts to fit the conditional mean function as best as it can.

There are only 2 covariates here. They are both binary. Thus, there are exactly four possible values for the conditional mean of the outcome given the regressors.
We have 4 free parameters (the betas). Choosing them wisely allows us to recover exactly the conditional mean.

Intuition

OLS attempts to fit the conditional mean function as best as it can.

There are only 2 covariates here. They are both binary. Thus, there are exactly four possible values for the conditional mean of the outcome given the regressors.
We have 4 free parameters (the betas). Choosing them wisely allows us to recover exactly the conditional mean.

Intuition

OLS attempts to fit the conditional mean function as best as it can.

There are only 2 covariates here. They are both binary. Thus, there are exactly four possible values for the conditional mean of the outcome given the regressors.
We have 4 free parameters (the betas). Choosing them wisely allows us to recover exactly the conditional mean.

Conditional Mean	Within the model
$E[Y\|D', T']$	β_0
$E[Y\|D, T']$	$\beta_0 + \beta_1$
$E[Y\|D', T]$	$\beta_0 + \beta_2$
$E[Y\|D, T]$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$

Intuition

We can make sure we get all of those right:

Model Parameter	Interpretation
β_0	$E[Y\|D', T']$
β_1	$E[Y\|D', T] - E[Y\|D', T']$
β_2	$E[Y\|D', T] - E[Y\|D', T']$

Intuition

We can make sure we get all of those right:

Model Parameter	Interpretation
β_0	$E[Y\|D', T']$
β_1	$E[Y\|D', T] - E[Y\|D', T']$
β_2	$E[Y\|D', T] - E[Y\|D', T']$

$$\beta_3 = \big(E[Y|D, T] - E[Y|D', T]\big) - \big(E[Y|D, T'] - E[Y|D', T']\big)$$

Intuition

We can make sure we get all of those right:

Model Parameter	Interpretation
β_0	$E[Y\|D', T']$
β_1	$E[Y\|D', T] - E[Y\|D', T']$
β_2	$E[Y\|D', T] - E[Y\|D', T']$

$$\beta_3 = \big(E[Y|D, T] - E[Y|D', T]\big) - \big(E[Y|D, T'] - E[Y|D', T']\big) = ATT_{DiD}$$

Takeaway: In a 2x2 DiD setup, one can obtain the DiD estimand by setting up a regression model with a dummy for the treatment group, a dummy for the post-treatment period, and an interaction between these. The coefficient on the interaction term will be numerically identical to the difference-in-differences estimator.

Interpretation

This two-way fixed effects specification is a clever way to obtain the DiD estimator as an output of a regression model.

Interpretation

This two-way fixed effects specification is a clever way to obtain the DiD estimator as an output of a regression model.

That means that the coefficient β_3 – *because* it is equal to ATT_{DiD} – can be interpreted as a causal effect (ATT) under the key assumption of parallel paths.

Interpretation

This two-way fixed effects specification is a clever way to obtain the DiD estimator as an output of a regression model.

That means that the coefficient β_3 – *because* it is equal to ATT_{DiD} – can be interpreted as a causal effect (ATT) under the key assumption of parallel paths.

Remark: Do we need the other usual OLS assumptions to view β_3 as causal?

Interpretation

This two-way fixed effects specification is a clever way to obtain the DiD estimator as an output of a regression model.

That means that the coefficient β_3 – *because* it is equal to ATT_{DiD} – can be interpreted as a causal effect (ATT) under the key assumption of parallel paths.

Remark: Do we need the other usual OLS assumptions to view β_3 as causal?

Does this regression assume that the effect is constant across individuals?

Interpretation

This two-way fixed effects specification is a clever way to obtain the DiD estimator as an output of a regression model.

That means that the coefficient β_3 – *because* it is equal to ATT_{DiD} – can be interpreted as a causal effect (ATT) under the key assumption of parallel paths.

Remark: Do we need the other usual OLS assumptions to view β_3 as causal?

Does this regression assume that the effect is constant across individuals? It does not.

New DiD Gospel

Often, the actual empirical application of DiD has a few departures from this simple, 2x2 case.

New DiD Gospel

Often, the actual empirical application of DiD has a few departures from this simple, 2x2 case. Here are a few:

- Multiple time periods/Staggered treatment (early adopters, late adopters, never treated).
- Adjustment for other potential confounders/covariates
- Non-binary treatments.

This is the point of departure for the new DiD literature.

New DiD Gospel

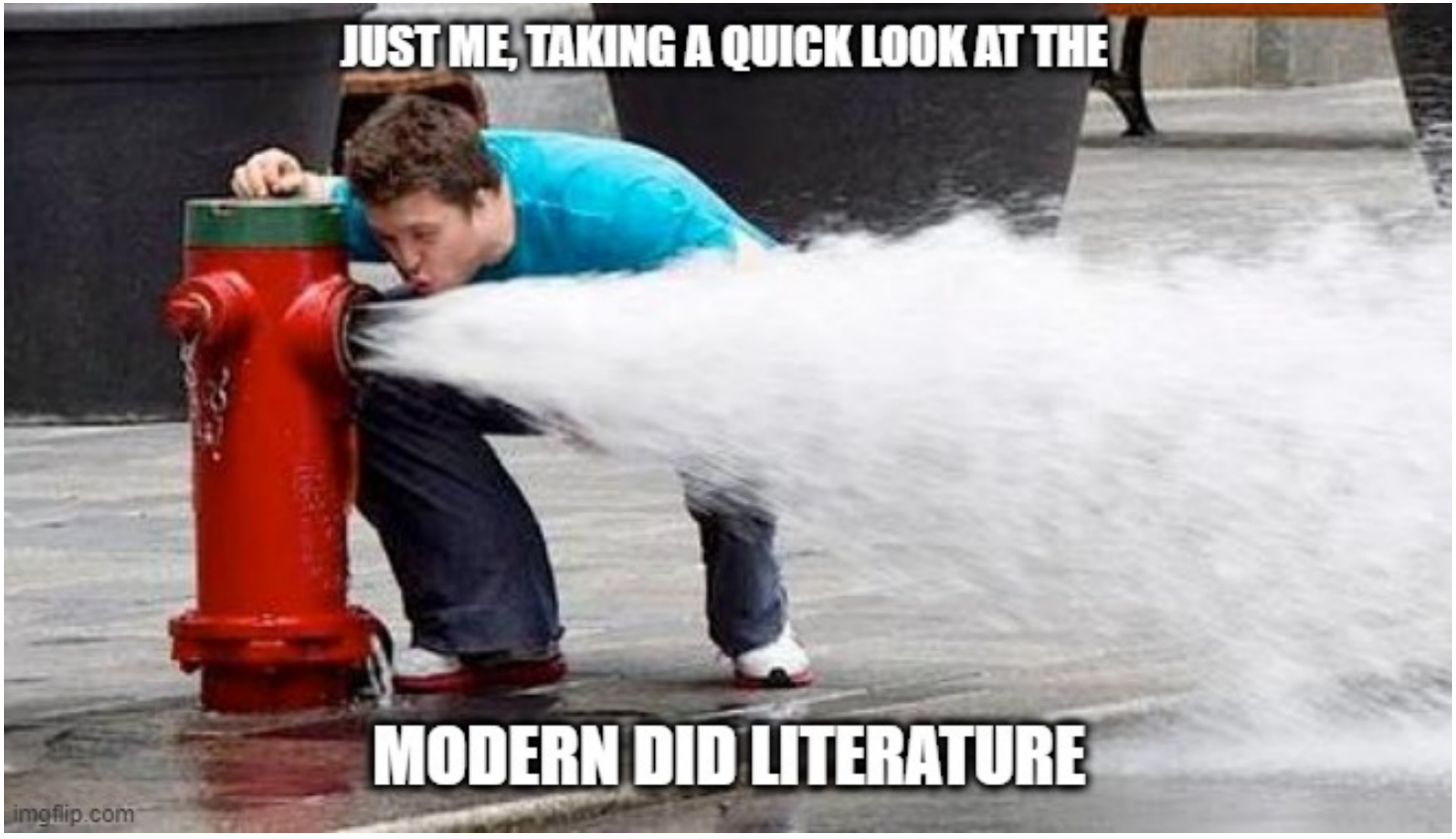
Often, the actual empirical application of DiD has a few departures from this simple, 2x2 case. Here are a few:

- Multiple time periods/Staggered treatment (early adopters, late adopters, never treated).
- Adjustment for other potential confounders/covariates
- Non-binary treatments.

This is the point of departure for the new DiD literature.

So, what is new?

Modern DiD


JUST ME, TAKING A QUICK LOOK AT THE
MODERN DID LITERATURE
imgflip.com

Leveraging Regression Strategies

Another way to write the same regression as we did before is to write it using fixed effects:

$$y_{it} = \alpha_{g_i} + \lambda_t + \beta D_{it} + \epsilon_{it}$$

where:

- α_g: Group fixed-effect
- λ_t: Time fixed-effect.
- β: Two-way fixed effect regression coefficient.
- D_{it}: Dummy that indicates that individual i at time t is treated.
  (Ignoring some irrelevant fine print) this yields yet again another representation for the DiD estimand using regression.

Leveraging Regression Strategies

Another way to write the same regression as we did before is to write it using fixed effects:

$$y_{it} = \alpha_{g_i} + \lambda_t + \beta D_{it} + \epsilon_{it}$$

where:

- α_g: Group fixed-effect
- λ_t: Time fixed-effect.
- β: Two-way fixed effect regression coefficient.
- D_{it}: Dummy that indicates that individual i at time t is treated.
  (Ignoring some irrelevant fine print) this yields yet again another representation for the DiD estimand using regression. But this one *looks* ripe for generalization.

TWFE – Applied practice circa 2015

$$y_{it} = \alpha_{g_i} + \lambda_t + \beta D_{it} + \epsilon_{it}$$

TWFE – Applied practice circa 2015

$$y_{it} = \alpha_{g_i} + \lambda_t + \beta D_{it} + \epsilon_{it}$$

Multiple groups?

TWFE – Applied practice circa 2015

$$y_{it} = \alpha_{g_i} + \lambda_t + \beta D_{it} + \epsilon_{it}$$

Multiple groups? It looks like you just need those fixed-effects.

TWFE – Applied practice circa 2015

$$y_{it} = \alpha_{g_i} + \lambda_t + \beta D_{it} + \epsilon_{it}$$

Multiple groups? It looks like you just need those fixed-effects.

Multiple time periods?

TWFE – Applied practice circa 2015

$$y_{it} = \alpha_{g_i} + \lambda_t + \beta D_{it} + \epsilon_{it}$$

Multiple groups? It looks like you just need those fixed-effects.

Multiple time periods? It looks like you just need those multiple time effects.

TWFE – Applied practice circa 2015

$$y_{it} = \alpha_{g_i} + \lambda_t + \beta D_{it} + \epsilon_{it}$$

Multiple groups? It looks like you just need those fixed-effects.

Multiple time periods? It looks like you just need those multiple time effects.

How about covariates?

$$y_{it} = \alpha_{g_i} + \lambda_t + \delta' \mathbf{X_{it}} + \beta D_{it} + \epsilon_{it}$$

We will discuss the interpretation of the TWFE estimator in each of these cases.

Goodman-Bacon

Let's look at what the TWFE does when you have a case with 3 periods, in a setting without covariates.

Goodman-Bacon

Let's look at what the TWFE does when you have a case with 3 periods, in a setting without covariates.

There is one early adopter, one late adopter, and one never treated.

Goodman-Bacon

Let's look at what the TWFE does when you have a case with 3 periods, in a setting without covariates.

There is one early adopter, one late adopter, and one never treated.

Let's assume that parallel paths hold here.

Goodman-Bacon

Let's look at what the TWFE does when you have a case with 3 periods, in a setting without covariates.

There is one early adopter, one late adopter, and one never treated.

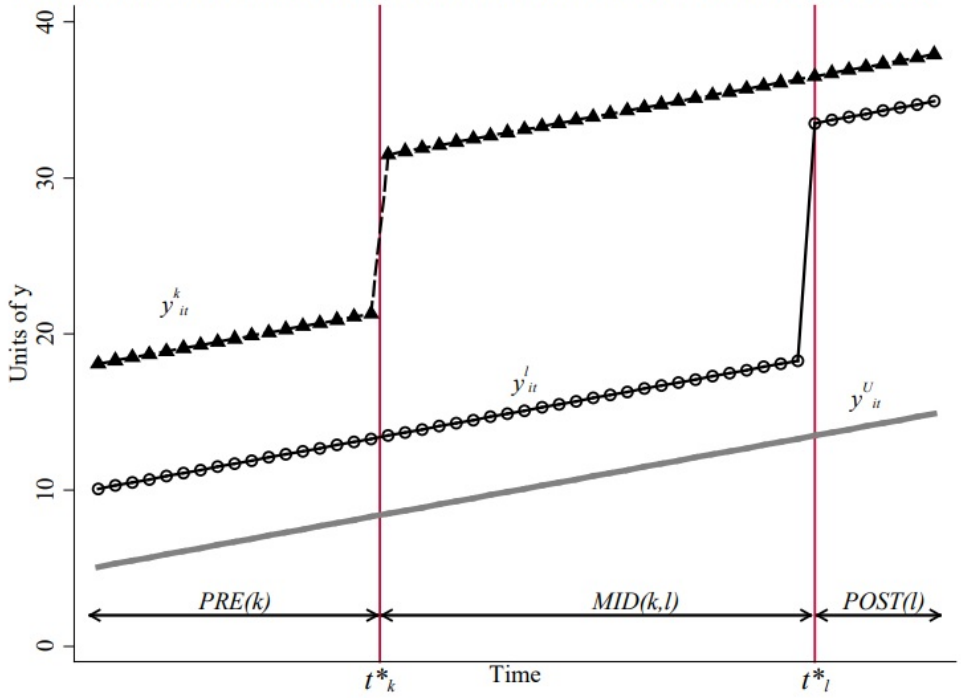
Let's assume that parallel paths hold here.

We will estimate the causal effects using the TWFE specification:

$$y_{it} = \alpha_{g_i} + \lambda_t + \beta D_{it} + \epsilon_{it}$$

Goodman-Bacon

Figure 1. Difference-in-Differences with Variation in Treatment Timing: Three Groups



Notes: The figure plots outcomes in three groups: a control group, U, which is never treated; an early treatment group, E, which receives a binary treatment at $t_k^* = \frac{34}{100}T$; and a late treatment group, ℓ, which receives the binary treatment at $t_\ell^* = \frac{85}{100}T$. The x-axis notes the three sub-periods: the pre-period for group k, $[1, t_k^* - 1]$, denoted by $PRE(k)$; the middle period when group k is treated and group ℓ is not, $[t_k^*, t_\ell^* - 1]$, denoted by $MID(k, \ell)$; and the post-period for group ℓ, $[t_\ell^*, T]$, denoted by $POST(\ell)$. I set the treatment effect to 10 in group k and 15 in group ℓ.

The good, the bad, and the ugly

The good

The good, the bad, and the ugly

The good: The OLS estimator, in a two-way fixed-effects regression, is a weighted average of a bunch two-by-two DiD estimators.

The good, the bad, and the ugly

The good: The OLS estimator, in a two-way fixed-effects regression, is a weighted average of a bunch two-by-two DiD estimators.

The good, the bad, and the ugly

The good: The OLS estimator, in a two-way fixed-effects regression, is a weighted average of a bunch two-by-two DiD estimators.

The bad

The good, the bad, and the ugly

The good: The OLS estimator, in a two-way fixed-effects regression, is a weighted average of a bunch two-by-two DiD estimators.

The bad: These are "variance-weighted" average treatment effects (VwATT) when the effects are different across individuals (but not over time).

The good, the bad, and the ugly

The good: The OLS estimator, in a two-way fixed-effects regression, is a weighted average of a bunch two-by-two DiD estimators.

The bad: These are "variance-weighted" average treatment effects (VwATT) when the effects are different across individuals (but not over time). Not exactly the ATT though (but an interesting parameter nevertheless).

The good, the bad, and the ugly

The good: The OLS estimator, in a two-way fixed-effects regression, is a weighted average of a bunch two-by-two DiD estimators.

The bad: These are "variance-weighted" average treatment effects (VwATT) when the effects are different across individuals (but not over time). Not exactly the ATT though (but an interesting parameter nevertheless).

The ugly

The good, the bad, and the ugly

The good: The OLS estimator, in a two-way fixed-effects regression, is a weighted average of a bunch two-by-two DiD estimators.

The bad: These are "variance-weighted" average treatment effects (VwATT) when the effects are different across individuals (but not over time). Not exactly the ATT though (but an interesting parameter nevertheless).

The ugly: There are typically lots of two-by-two comparisons that you would not do or trust yourself, but OLS is going to use them all.

The good, the bad, and the ugly

The good: The OLS estimator, in a two-way fixed-effects regression, is a weighted average of a bunch two-by-two DiD estimators.

The bad: These are "variance-weighted" average treatment effects (VwATT) when the effects are different across individuals (but not over time). Not exactly the ATT though (but an interesting parameter nevertheless).

The ugly: There are typically lots of two-by-two comparisons that you would not do or trust yourself, but OLS is going to use them all.

The Good

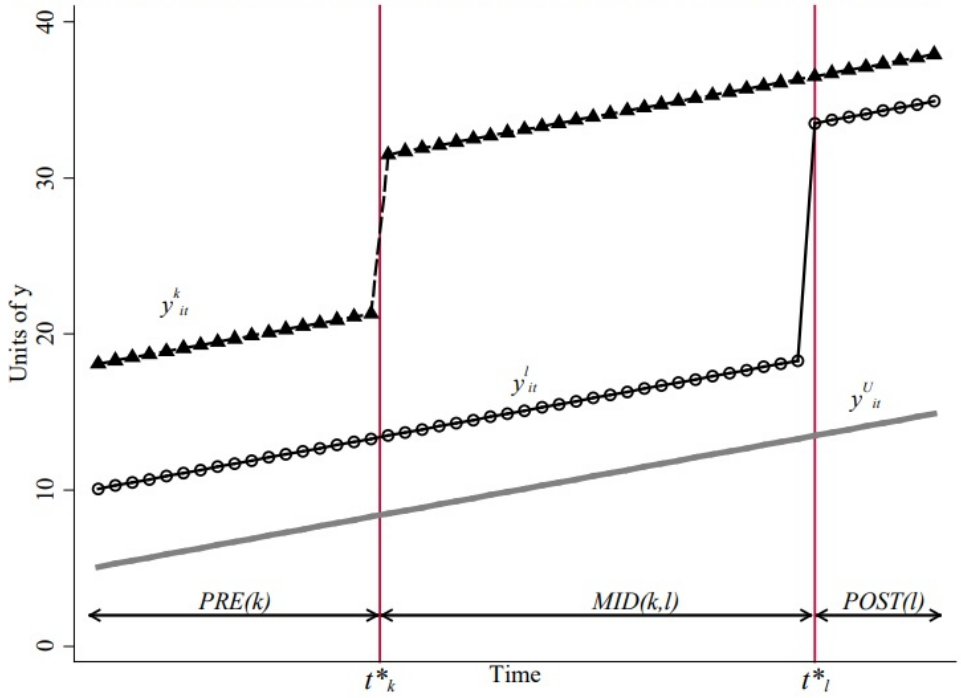
The OLS estimate, in a two-way fixed-effects regression, is a weighted average of all possible two-by-two DID estimators.

The Good

The OLS estimate, in a two-way fixed-effects regression, is a weighted average of all possible two-by-two DID estimators.
Let's see that in an example.

Goodman-Bacon

Figure 1. Difference-in-Differences with Variation in Treatment Timing: Three Groups



Notes: The figure plots outcomes in three groups: a control group, U, which is never treated; an early treatment group, E, which receives a binary treatment at $t_k^* = \frac{34}{100}T$; and a late treatment group, ℓ, which receives the binary treatment at $t_\ell^* = \frac{85}{100}T$. The x-axis notes the three sub-periods: the pre-period for group k, $[1, t_k^* - 1]$, denoted by $PRE(k)$; the middle period when group k is treated and group ℓ is not, $[t_k^*, t_\ell^* - 1]$, denoted by $MID(k, \ell)$; and the post-period for group ℓ, $[t_\ell^*, T]$, denoted by $POST(\ell)$. I set the treatment effect to 10 in group k and 15 in group ℓ.

The bad

These are "variance-weighted" average treatment effects (VwATT) when the effects are different across individuals (but not over time).

The bad

These are "variance-weighted" average treatment effects (VwATT) when the effects are different across individuals (but not over time). Not exactly the ATT though (but an interesting parameter nevertheless).

The bad

These are "variance-weighted" average treatment effects (VwATT) when the effects are different across individuals (but not over time). Not exactly the ATT though (but an interesting parameter nevertheless).

Groups treated in the middle of your (time) sample will have larger weights.

The bad

These are "variance-weighted" average treatment effects (VwATT) when the effects are different across individuals (but not over time). Not exactly the ATT though (but an interesting parameter nevertheless).

Groups treated in the middle of your (time) sample will have larger weights.

Such a weighting scheme is great when you expect low heterogeneity in the treatment effects.

The bad

These are "variance-weighted" average treatment effects (VwATT) when the effects are different across individuals (but not over time). Not exactly the ATT though (but an interesting parameter nevertheless).

Groups treated in the middle of your (time) sample will have larger weights.

Such a weighting scheme is great when you expect low heterogeneity in the treatment effects.

With heterogeneity, however, the parameter will change even if you just add more years of data for everyone before the treatment starts (or ends).

The Ugly

There are typically lots of two-by-two comparisons that you would not do or trust yourself, but OLS is going to use them all.

The Ugly

There are typically lots of two-by-two comparisons that you would not do or trust yourself, but OLS is going to use them all.

Individuals treated earlier in the sample act as controls for those treated later.

The Ugly

There are typically lots of two-by-two comparisons that you would not do or trust yourself, but OLS is going to use them all.

Individuals treated earlier in the sample act as controls for those treated later.

The original parallel paths assumption had no business allowing you to do that.

The Ugly

There are typically lots of two-by-two comparisons that you would not do or trust yourself, but OLS is going to use them all.

Individuals treated earlier in the sample act as controls for those treated later.

The original parallel paths assumption had no business allowing you to do that.

The key was that potential outcomes in the *untreated state* (Y(0)) would evolve in parallel over time.

The Ugly

There are typically lots of two-by-two comparisons that you would not do or trust yourself, but OLS is going to use them all.

Individuals treated earlier in the sample act as controls for those treated later.

The original parallel paths assumption had no business allowing you to do that.

The key was that potential outcomes in the *untreated state* (Y(0)) would evolve in parallel over time.

The Ugly

How did we get here?

The Ugly

How did we get here? Look at the regression:

$$y_{it} = \alpha_{g_i} + \lambda_t + \beta D_{it} + \epsilon_{it}$$

It is clear that if this is the process that generates the data and we take it seriously, we get:

Parallel paths between all units not yet treated:

$$E[y|G = g, T = t, D = 0] - E[y|G = g', T = t, D = 0] = \alpha_g - \alpha_{g'}$$

The Ugly

How did we get here? Look at the regression:

$$y_{it} = \alpha_{g_i} + \lambda_t + \beta D_{it} + \epsilon_{it}$$

It is clear that if this is the process that generates the data and we take it seriously, we get:

Parallel paths between all units not yet treated:

$$E[y|G = g, T = t, D = 0] - E[y|G = g', T = t, D = 0] = \alpha_g - \alpha_{g'}$$

Parallel paths between units already treated and those not yet treated. For $t_g^* < t < t_{g'}^*$:

$$E[y|G = g, T = t, D = 1] - E[y|G = g', T = t, D = 0] = \alpha_g - \alpha_{g'} + \beta$$

The Ugly

OLS will take advantage of all experiments you thought of and some that you did not. Including this one here, for $t_g^* < t < t_{g'}^*$

$$E[y|G = g, T = t, D = 1] - E[y|G = g', T = t, D = 0] = \alpha_g - \alpha_{g'} + \beta$$

The Ugly

OLS will take advantage of all experiments you thought of and some that you did not. Including this one here, for $t_g^* < t < t_{g'}^*$

$$E[y|G = g, T = t, D = 1] - E[y|G = g', T = t, D = 0] = \alpha_g - \alpha_{g'} + \beta$$

If you are only comfortable assuming that the paths were supposed to be parallel *before* treatment takes place, you wouldn't feel comfortable exploiting this source of variation.

The Ugly

OLS will take advantage of all experiments you thought of and some that you did not. Including this one here, for $t_g^* < t < t_{g'}^*$

$$E[y|G = g, T = t, D = 1] - E[y|G = g', T = t, D = 0] = \alpha_g - \alpha_{g'} + \beta$$

If you are only comfortable assuming that the paths were supposed to be parallel *before* treatment takes place, you wouldn't feel comfortable exploiting this source of variation.

Even units that were *always treated* since the beginning of your sample act as control groups in the standard 2WFE specification.

The Ugly

OLS will take advantage of all experiments you thought of and some that you did not. Including this one here, for $t_g^* < t < t_{g'}^*$

$$E[y|G = g, T = t, D = 1] - E[y|G = g', T = t, D = 0] = \alpha_g - \alpha_{g'} + \beta$$

If you are only comfortable assuming that the paths were supposed to be parallel *before* treatment takes place, you wouldn't feel comfortable exploiting this source of variation.

Even units that were *always treated* since the beginning of your sample act as control groups in the standard 2WFE specification.

Essentially, you are assuming parallel paths of Y(0) and also Y(1). OLS will use both.

The Ugly

At at time $t_g^* < t < t_{g'}^*$ in which Group g is treated but group g' is not treated yet, we get:

$$E[y|G = g, T = t, D = 1] - E[y|G = g', T = t, D = 0] = \alpha_g - \alpha_{g'} + \beta$$

At any time t after both of them are treated:

$$E[y|G = g, T = t, D = 1] - E[y|G = g', T = t, D = 1] = \alpha_g - \alpha_{g'} + \beta - \beta$$

The difference between these differences is:

The Ugly

At at time $t_g^* < t < t_{g'}^*$ in which Group g is treated but group g' is not treated yet, we get:

$$E[y|G = g, T = t, D = 1] - E[y|G = g', T = t, D = 0] = \alpha_g - \alpha_{g'} + \beta$$

At any time t after both of them are treated:

$$E[y|G = g, T = t, D = 1] - E[y|G = g', T = t, D = 1] = \alpha_g - \alpha_{g'} + \beta - \beta$$

The difference between these differences is: $-\beta$. Inverting the contrast you get your β.

The Ugly

At at time $t_g^* < t < t_{g'}^*$ in which Group g is treated but group g' is not treated yet, we get:

$$E[y|G = g, T = t, D = 1] - E[y|G = g', T = t, D = 0] = \alpha_g - \alpha_{g'} + \beta$$

At any time t after both of them are treated:

$$E[y|G = g, T = t, D = 1] - E[y|G = g', T = t, D = 1] = \alpha_g - \alpha_{g'} + \beta - \beta$$

The difference between these differences is: $-\beta$. Inverting the contrast you get your β.

This is one of the comparisons that the 2WFE will do, unless you explicitly work to rule it out.

The Ugly

At at time $t_g^* < t < t_{g'}^*$ in which Group g is treated but group g' is not treated yet, we get:

$$E[y|G = g, T = t, D = 1] - E[y|G = g', T = t, D = 0] = \alpha_g - \alpha_{g'} + \beta$$

At any time t after both of them are treated:

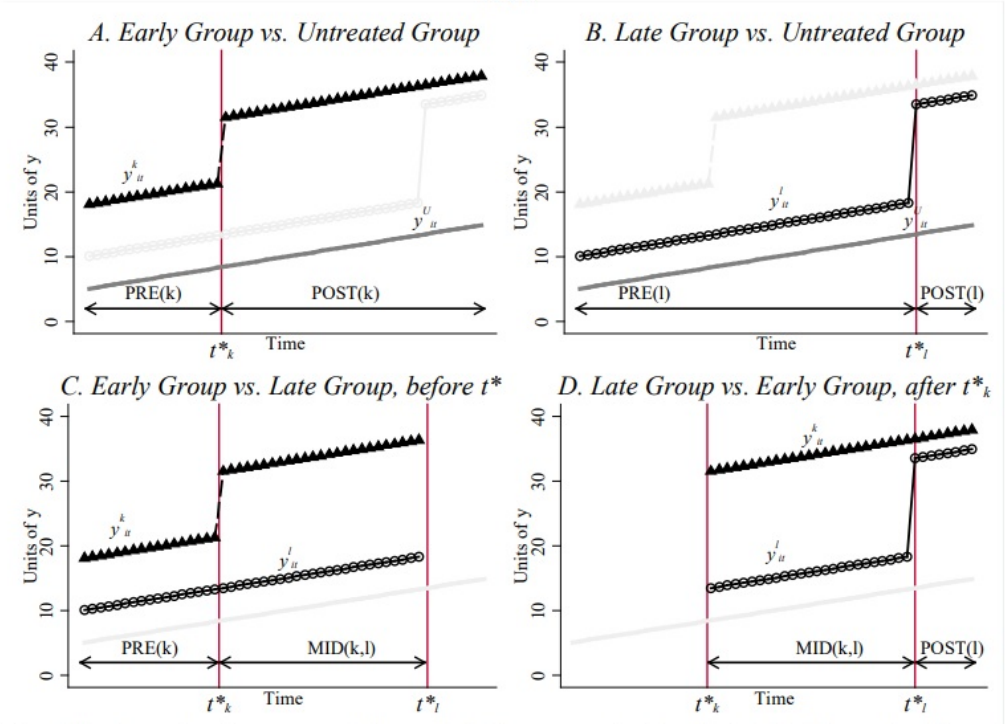
$$E[y|G = g, T = t, D = 1] - E[y|G = g', T = t, D = 1] = \alpha_g - \alpha_{g'} + \beta - \beta$$

The difference between these differences is: $-\beta$. Inverting the contrast you get your β.

This is one of the comparisons that the 2WFE will do, unless you explicitly work to rule it out. It uses parallel paths of $Y(1)$ which only holds if the treatment has no time heterogeneity.

Goodman-Bacon

Figure 2. The Four Simple (2x2) Difference-in-Differences Estimates from the Three Group Case



Notes: The figure plots the groups and time periods that generate the four simple 2x2 difference-in-difference estimates in the case with an early treatment group, a late treatment group, and an untreated group from Figure 1. Each panel plots the data structure for one 2x2 DD. Panel A compares early treated units to untreated units ($\hat{\beta}_{kU}^{DD}$); panel B compares late treated units to untreated units ($\hat{\beta}_{\ell U}^{DD}$); panel C compares early treated units to late treated units during the late group's pre-period ($\hat{\beta}_{k\ell}^{DD,k}$); panel D compares late treated units to early treated units during the early group's post-period ($\hat{\beta}_{k\ell}^{DD,\ell}$). The treatment times mean that $\overline{D}_k = 0.67$ and $\overline{D}_\ell = 0.16$, so with equal group sizes, the decomposition weights on the 2x2 estimate from each panel are 0.365 for panel A, 0.222 for panel B, 0.278 for panel C, and 0.135 for panel D.

How many 2x2 comparisons is OLS using here again?

The Ugly

Individuals treated earlier in the sample act as controls for those treated later.

The Ugly

Individuals treated earlier in the sample act as controls for those treated later.

The Ugly

Individuals treated earlier in the sample act as controls for those treated later.

If the effects are heterogeneous over time (they pick up pace and rise, or they decay), then some treatment effects have a negative weight.

The Ugly

Individuals treated earlier in the sample act as controls for those treated later.

If the effects are heterogeneous over time (they pick up pace and rise, or they decay), then some treatment effects have a negative weight.

That is, the larger the effect is, the smaller your estimate of it will be.

The Ugly

Individuals treated earlier in the sample act as controls for those treated later.

If the effects are heterogeneous over time (they pick up pace and rise, or they decay), then some treatment effects have a negative weight.

That is, the larger the effect is, the smaller your estimate of it will be. You never, ever, *ever* want to be in that spot.

How bad can it be, and how to fix it

How bad can it be? Well, with heterogeneity in the effects over time, enough that even the sign can be wrong.

How bad can it be, and how to fix it

How bad can it be? Well, with heterogeneity in the effects over time, enough that even the sign can be wrong.

How to fix: There are multiple solutions in the literature.

How bad can it be, and how to fix it

How bad can it be? Well, with heterogeneity in the effects over time, enough that even the sign can be wrong.

How to fix: There are multiple solutions in the literature.

They all amount to only using the comparisons that you are willing to use.

Diagnostics

We know the weights. We can check to see what is going on underneath.

Diagnostics

We know the weights. We can check to see what is going on underneath. Use bacondecomp in Stata/R to see if you have negative weights and which comparisons are driving your estimate (they might be the ones you want anyway).

Example: Stevenson and Wolfers (AER, 2006)

Stevenson and Wolfers exploit "the natural variation resulting from the different timing of the adoption of unilateral divorce laws" in 37 states from 1969-1985 (see table 1) using the "remaining fourteen states as controls" to evaluate the effect of these reforms on female suicide rates.

Example: Stevenson and Wolfers (AER, 2006)

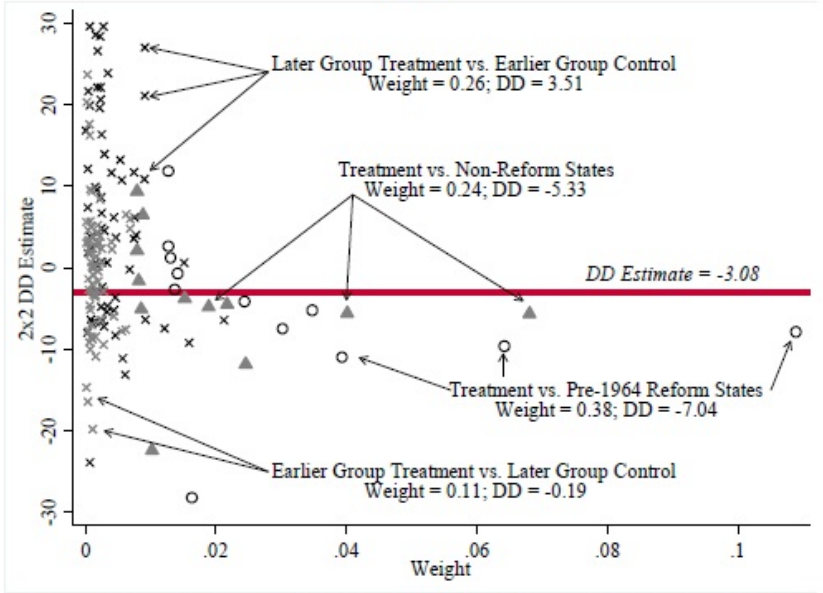
Table 1. The No-Fault Divorce Rollout: Treatment Times, Group Sizes, and Treatment Shares

No-Fault Divorce Year (t_k^*)	Number of States	Share of States (n_k)	Treatment Share ($\bar{D}_k$)
Non-Reform States	5	0.10	.
Pre-1964 Reform States	8	0.16	.
1969	2	0.04	0.85
1970	2	0.04	0.82
1971	7	0.14	0.79
1972	3	0.06	0.76
1973	10	0.20	0.73
1974	3	0.06	0.70
1975	2	0.04	0.67
1976	1	0.02	0.64
1977	3	0.06	0.61
1980	1	0.02	0.52
1984	1	0.02	0.39
1985	1	0.02	0.36

Notes: The table lists the dates of no-fault divorce reforms from Stevenson and Wolfers (2006), the number and share of states that adopt in each year, and the share of periods each treatment timing group spends treated in the estimation sample from 1964-1996.

Example: Stevenson and Wolfers (AER, 2006)

Figure 6. Difference-in-Differences Decomposition for Unilateral Divorce and Female Suicide



Notes: The figure plots each 2x2 DD components from the decomposition theorem against their weight for the unilateral divorce analysis. The open circles are terms in which one timing group acts as the treatment group and the pre-1964 reform states act as the control group. The closed triangles are terms in which one timing group acts as the treatment group and the non-reform states act as the control group. The x's are the timing-only terms. The figure notes the average DD estimate and total weight on each type of comparison. The two-way fixed effects estimate, -3.08, equals the average of the y-axis values weighted by their x-axis value.

Goodman-Bacon

"The bias resulting from time-varying effects is also apparent in figure 6. The average of the post-treatment event-study estimates in figure 5 is -4.92, but the DD estimate is 60 percent as large (-3.08).

Goodman-Bacon

"The bias resulting from time-varying effects is also apparent in figure 6. The average of the post-treatment event-study estimates in figure 5 is -4.92, but the DD estimate is 60 percent as large (-3.08).

The difference stems from the comparisons of later- to earlier-treated groups. The average treated/untreated estimates are negative (-5.33 and -7.04), as are the comparisons of earlier- to later-treated states (although less so: -0.19).

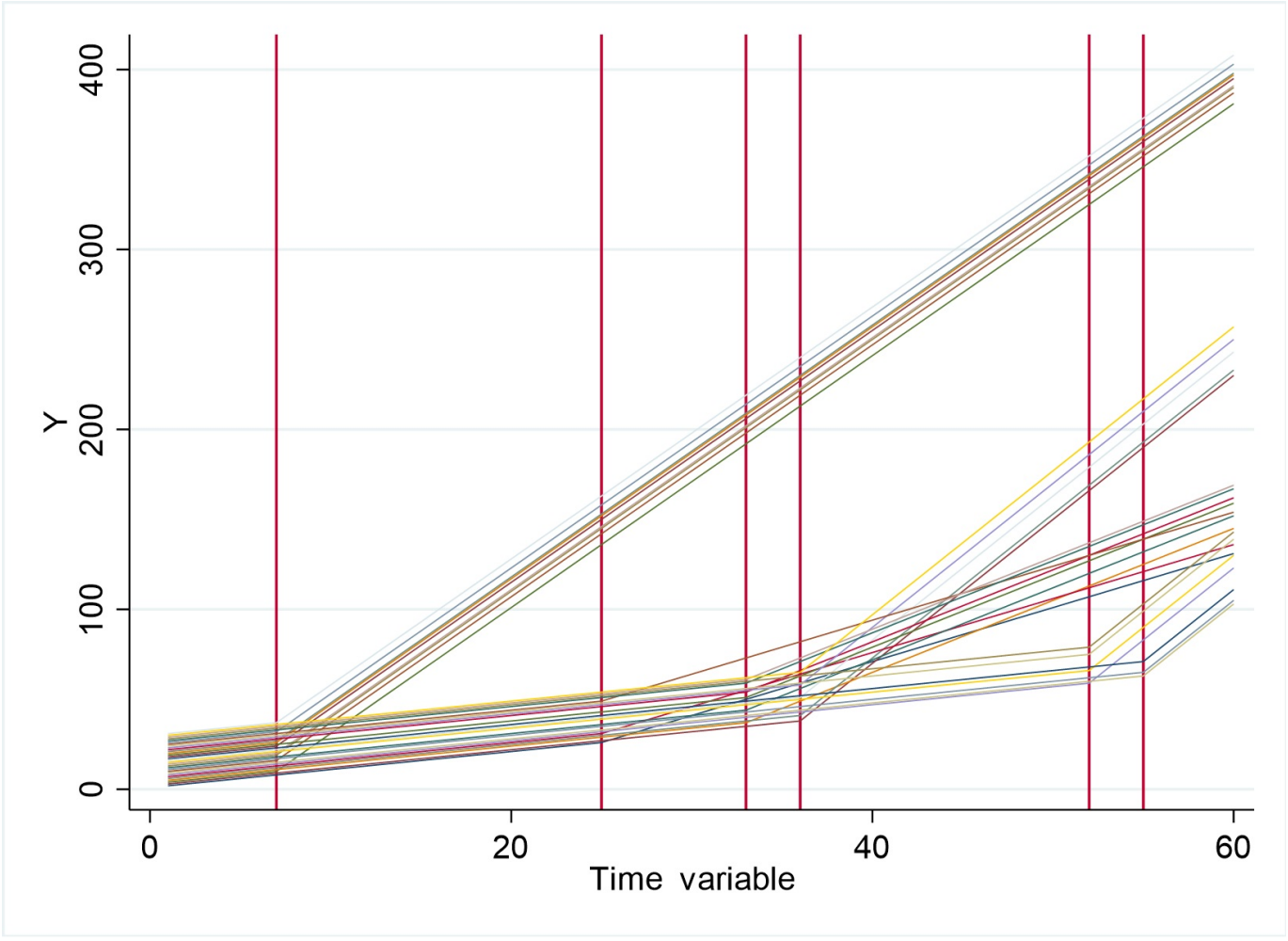
Goodman-Bacon

"The bias resulting from time-varying effects is also apparent in figure 6. The average of the post-treatment event-study estimates in figure 5 is -4.92, but the DD estimate is 60 percent as large (-3.08).

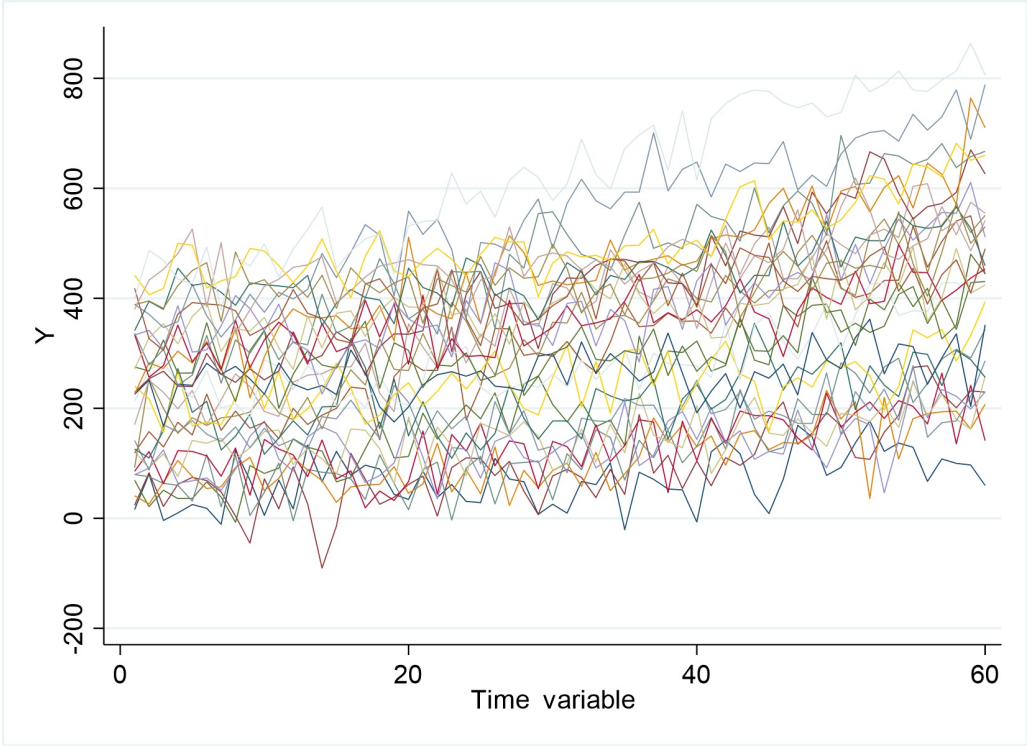
The difference stems from the comparisons of later- to earlier-treated groups. The average treated/untreated estimates are negative (-5.33 and -7.04), as are the comparisons of earlier- to later-treated states (although less so: -0.19).

The comparisons of later- to earlier-treated states, however, are positive on average (3.51) and account for the bias in the overall DD estimate. Using the decomposition theorem to take these terms out of the weighted average yields an effect of -5.44."

Negative Weights

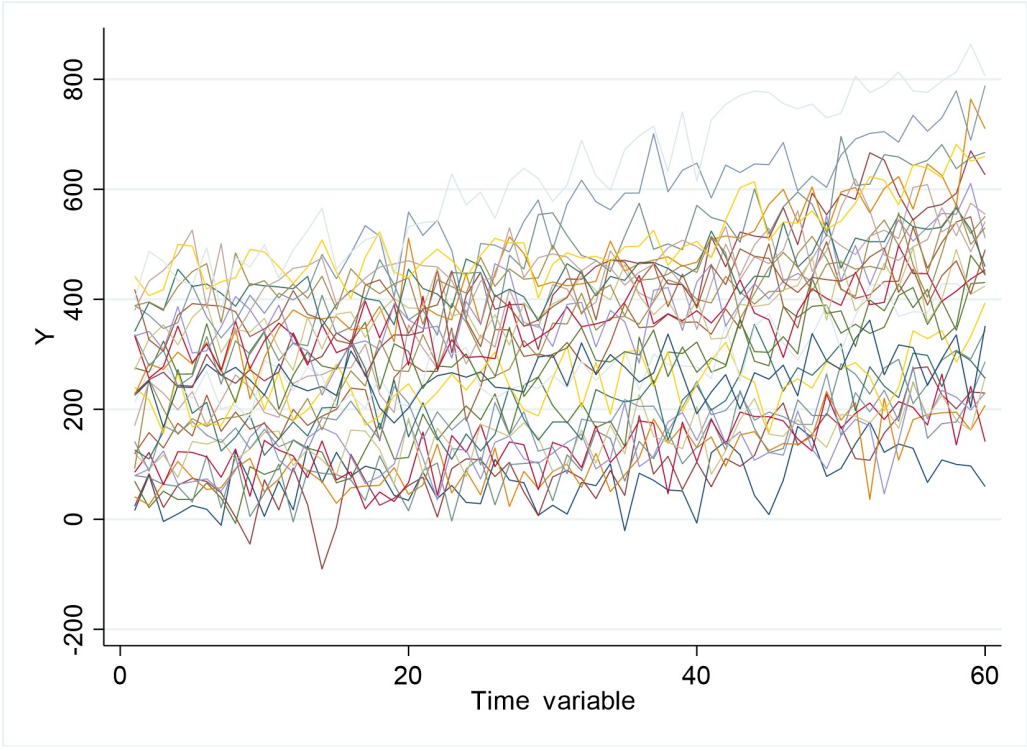


Negative Weights



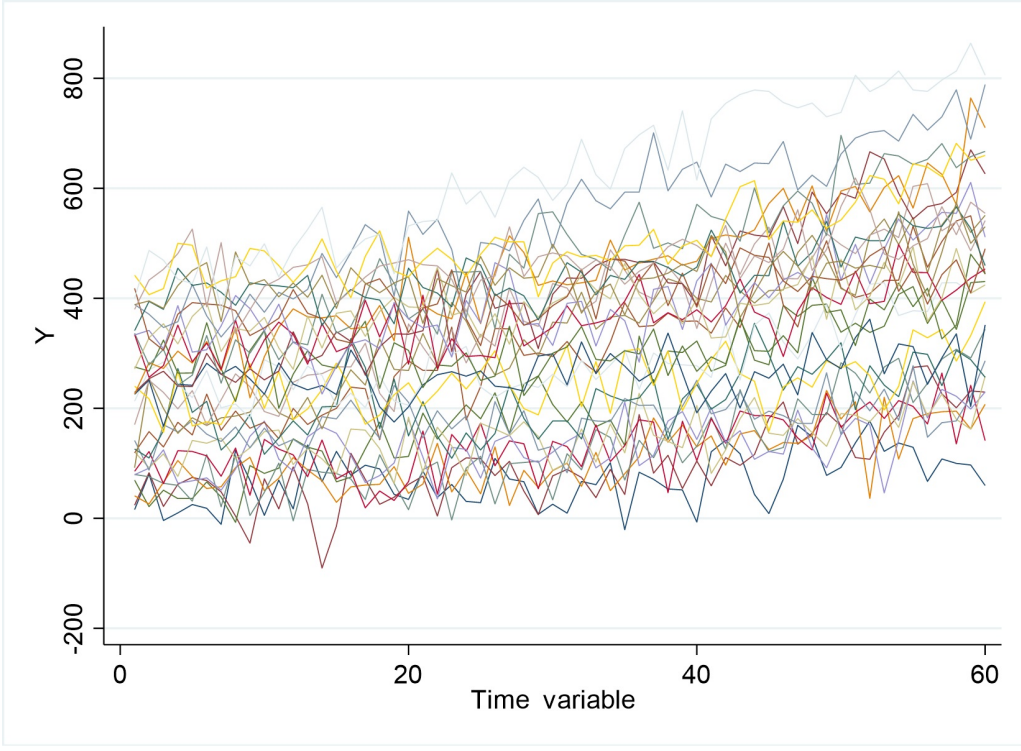
Parallel paths hold here.

Negative Weights



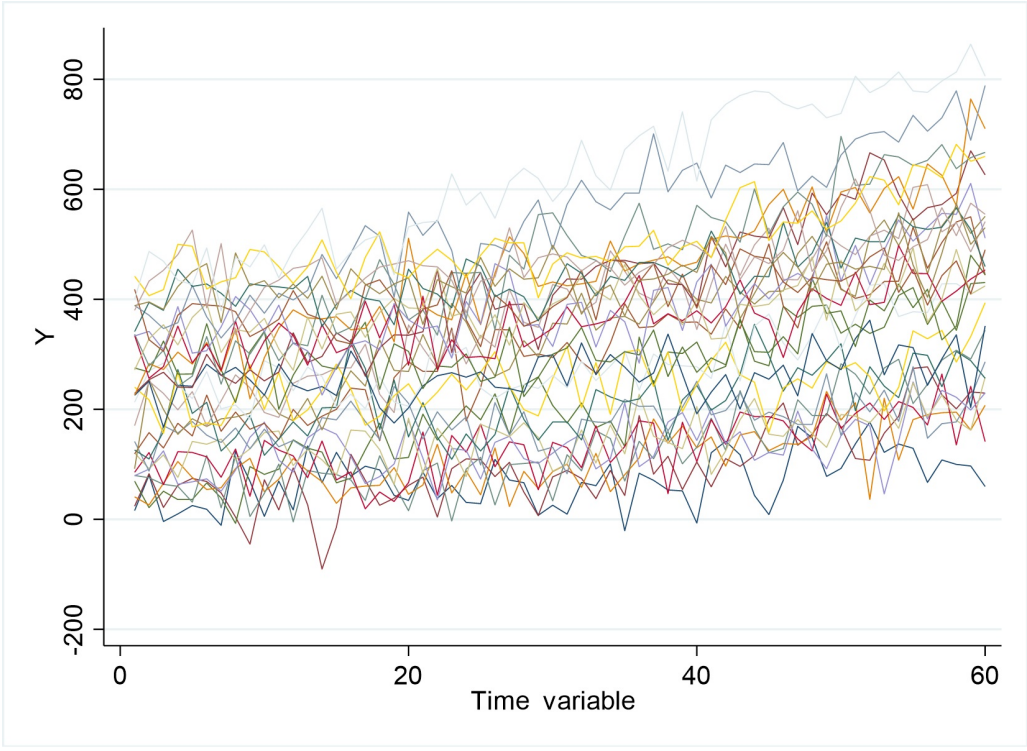
Parallel paths hold here. There are no pre-trends.

Negative Weights



Parallel paths hold here. There are no pre-trends. The DiD design is a valid way to get causal effects here.

Negative Weights



Parallel paths hold here. There are no pre-trends. The DiD design is a valid way to get causal effects here. This is the same DGP as before with more noise, so that it is not entirely obvious that the effects grow over time.

Negative Weights - Red Flags

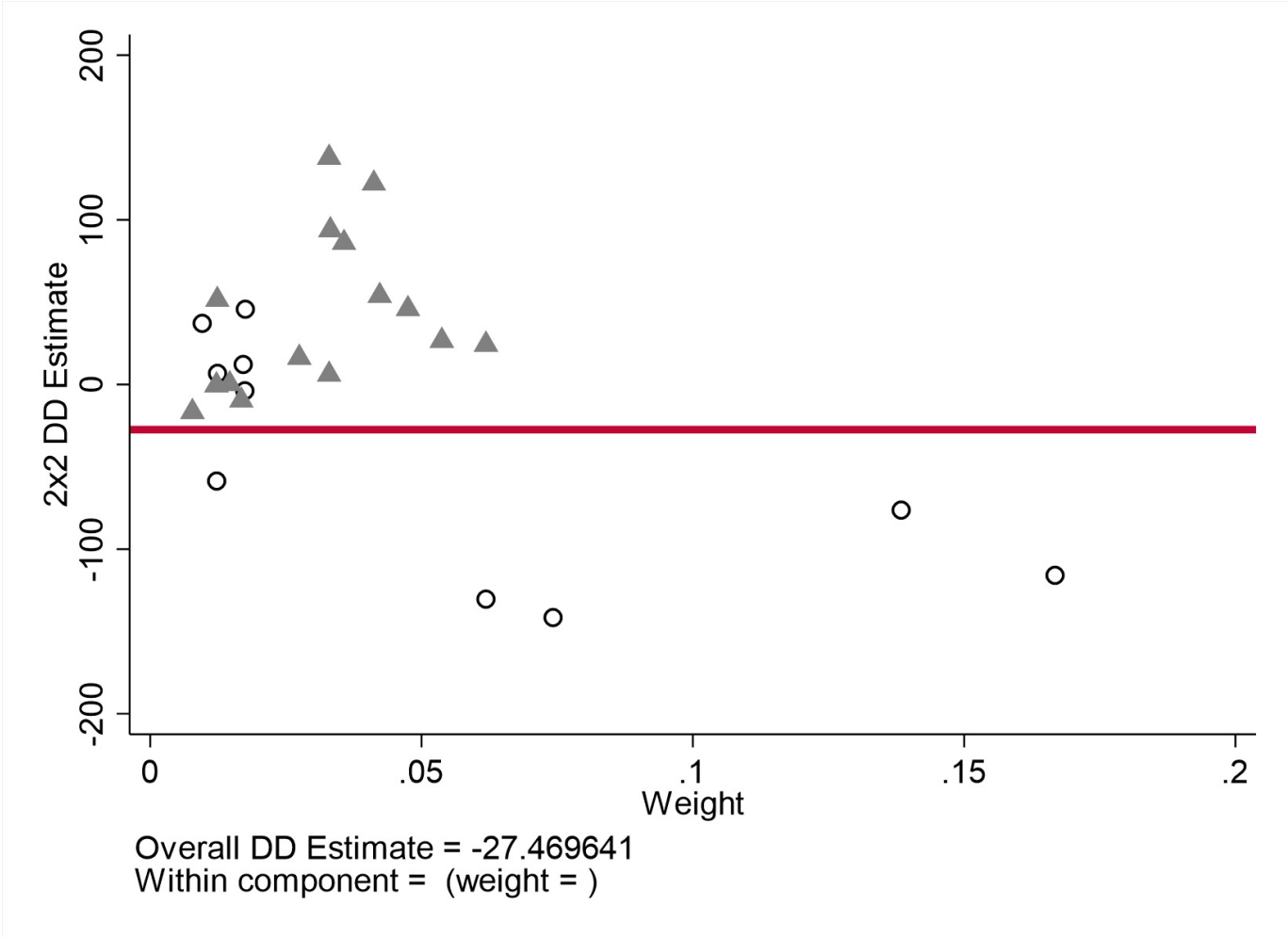
```
. reghdfe Y D, absorb(id t)
(MWFE estimator converged in 2 iterations)

HDFE Linear regression                            Number of obs   =      1,800
Absorbing 2 HDFE groups                           F(   1,   1710) =      33.35
                                                  Prob > F        =     0.0000
                                                  R-squared       =     0.8959
                                                  Adj R-squared   =     0.8905
                                                  Within R-sq.    =     0.0191
                                                  Root MSE        =    56.0058

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           D |  -27.46964   4.756903    -5.77   0.000    -36.7996   -18.13968
       _cons |   331.3546   2.815453   117.69   0.000    325.8325    336.8767
------------------------------------------------------------------------------
```

Negative Weights - Red Flags

Bacon Decomposition

	Beta	TotalWeight
Early_v_Late	51.37289047	.0123657302
Late_v_Early	-141.4384918	.0741943846
Early_v_Late	85.9851532	.0357232214
Late_v_Early	-115.7802353	.1667083601
Early_v_Late	.1877050549	.0146556807
Late_v_Early	12.28703403	.0170982933
Early_v_Late	93.74787903	.0332042752
Late_v_Early	-76.17219543	.1383511554
Early_v_Late	-9.751236916	.0167929674
Late_v_Early	45.83772278	.0174926737
Early_v_Late	-.4106702209	.0122130672
Late_v_Early	37.28028107	.0095414589
Early_v_Late	121.9943924	.0412191018
Late_v_Early	-130.2444458	.0618286496
Early_v_Late	6.010273457	.0329752828
Late_v_Early	6.879799843	.0123657302
Early_v_Late	24.23727417	.0618795396
Late_v_Early	-3.70816493	.0174036209
Early_v_Late	45.85042572	.0474952625
Late_v_Early	-58.51732635	.0122130672
Early_v_Late	52.13389577	.1642784771

Bacon Decomposition



Chaisemartin and D'Haultfœuille (2020)

They show another decomposition of the TWFE coefficient.

Chaisemartin and D'Haultfœuille (2020)

They show another decomposition of the TWFE coefficient. Why is there more than one decomposition of the TWFE coefficient? Didn't Bacon solve this problem already?

Chaisemartin and D'Haultfœuille (2020)

They show another decomposition of the TWFE coefficient. Why is there more than one decomposition of the TWFE coefficient? Didn't Bacon solve this problem already?

This is sort of why the literature is so vast, by the way.

Chaisemartin and D'Haultfœuille (2020)

They show another decomposition of the TWFE coefficient. Why is there more than one decomposition of the TWFE coefficient? Didn't Bacon solve this problem already?

This is sort of why the literature is so vast, by the way.

Answer: There's more than one way to skin a cat.

Chaisemartin and D'Haultfœuille (2020)

They show another decomposition of the TWFE coefficient. Why is there more than one decomposition of the TWFE coefficient? Didn't Bacon solve this problem already?

This is sort of why the literature is so vast, by the way.

Answer: There's more than one way to skin a cat.

Bacon sliced the TWFE in terms of every possible 2X2 DID estimator you could conceivably run with multiple groups and different timing of treatment.

Chaisemartin and D'Haultfœuille (2020)

They show another decomposition of the TWFE coefficient. Why is there more than one decomposition of the TWFE coefficient? Didn't Bacon solve this problem already?

This is sort of why the literature is so vast, by the way.

Answer: There's more than one way to skin a cat.

Bacon sliced the TWFE in terms of every possible 2X2 DID estimator you could conceivably run with multiple groups and different timing of treatment. His key contribution is to show that you might be running some weird DIDs inadvertently.

Chaisemartin and D'Haultfœuille (2020)

They show another decomposition of the TWFE coefficient. Why is there more than one decomposition of the TWFE coefficient? Didn't Bacon solve this problem already?

This is sort of why the literature is so vast, by the way.

Answer: There's more than one way to skin a cat.

Bacon sliced the TWFE in terms of every possible 2X2 DID estimator you could conceivably run with multiple groups and different timing of treatment. His key contribution is to show that you might be running some weird DIDs inadvertently.

Chaisemartin and D'Haultfœuille show how to slice the TWFE into weighted averages of treatment effects for each treated group and each period. They show how large the weights of different TEs will be.

Chaisemartin and D'Haultfœuille (2020)

They show another decomposition of the TWFE coefficient. Why is there more than one decomposition of the TWFE coefficient? Didn't Bacon solve this problem already?

This is sort of why the literature is so vast, by the way.

Answer: There's more than one way to skin a cat.

Bacon sliced the TWFE in terms of every possible 2X2 DID estimator you could conceivably run with multiple groups and different timing of treatment. His key contribution is to show that you might be running some weird DIDs inadvertently.

Chaisemartin and D'Haultfœuille show how to slice the TWFE into weighted averages of treatment effects for each treated group and each period. They show how large the weights of different TEs will be. They show when you might have a TE with a negative weight.

Chaisemartin and D'Haultfœuille (2020)

Key result: Let $W_{g,t}$ be the weight associated with the treatment effects of a group g that at time t was treated will have in the TWFE regression specification. The weights sum to one and are proportional to:

$$W_{g,t} \propto D_{g,t} - E_g[D_{g,t}|T = t] - E_t[D_{g,t}|G = g] + E_{g,t}[D_{g,t}]$$

Chaisemartin and D'Haultfœuille (2020)

Key result: Let $W_{g,t}$ be the weight associated with the treatment effects of a group g that at time t was treated will have in the TWFE regression specification. The weights sum to one and are proportional to:

$$W_{g,t} \propto D_{g,t} - E_g[D_{g,t}|T = t] - E_t[D_{g,t}|G = g] + E_{g,t}[D_{g,t}]$$

- $D_{g,t}$ is the dummy indicating whether or not group g is treated at time t.
- $E_g[D_{g,t}|T = t]$ is the average of the treatment indicator when taken across all groups but holding constant the time t – the between groups at time t average value of D.

Chaisemartin and D'Haultfœuille (2020)

Key result: Let $W_{g,t}$ be the weight associated with the treatment effects of a group g that at time t was treated will have in the TWFE regression specification. The weights sum to one and are proportional to:

$$W_{g,t} \propto D_{g,t} - E_g[D_{g,t}|T = t] - E_t[D_{g,t}|G = g] + E_{g,t}[D_{g,t}]$$

- $D_{g,t}$ is the dummy indicating whether or not group g is treated at time t.
- $E_g[D_{g,t}|T = t]$ is the average of the treatment indicator when taken across all groups but holding constant the time t – the between groups at time t average value of D.
- $E_t[D_{g,t}|G = g]$ is the average of the treatment indicator when taken across all time periods but holding constant the group g – the across time average value of D for group g.
- $E_{g,t}[D_{g,t}]$ is the across groups and time average of the treatment indicator D

Chaisemartin and D'Haultfœuille (2020)

A special case in which the weights are equal to one over the number of (g,t) treated cells (that is, they are ATT weights):

- The design is staggered (treatment either increases or stays the same over time).
- Treatment is binary.
- There is no variation in treatment timing.

Chaisemartin and D'Haultfœuille (2020)

A special case in which the weights are equal to one over the number of (g,t) treated cells (that is, they are ATT weights):

- The design is staggered (treatment either increases or stays the same over time).
- Treatment is binary.
- There is no variation in treatment timing.

Note that this is all true in the 2x2 DID case since the treatment is indeed binary, and everyone who gets treated gets treated in period two and stays treated until the end of the observation period, which is period 2.

Chaisemartin and D'Haultfœuille (2020)

These weights are wonderfully easy to compute by hand.

Chaisemartin and D'Haultfœuille (2020)

These weights are wonderfully easy to compute by hand.

I will say that again. It takes you one second to program them by hand in your application.

Compute the mean of the treatment dummy.

Compute the by group across time average of $D_{g,t}$ (there will be one for each group).

Compute the by time, across groups, average of $D_{g,t}$ (there will be one for each time).

Chaisemartin and D'Haultfœuille (2020)

These weights are wonderfully easy to compute by hand.

I will say that again. It takes you one second to program them by hand in your application.

Compute the mean of the treatment dummy.

Compute the by group across time average of $D_{g,t}$ (there will be one for each group).

Compute the by time, across groups, average of $D_{g,t}$ (there will be one for each time).

Generate the weight as

Chaisemartin and D'Haultfœuille (2020)

These weights are wonderfully easy to compute by hand.

I will say that again. It takes you one second to program them by hand in your application.

Compute the mean of the treatment dummy.

Compute the by group across time average of $D_{g,t}$ (there will be one for each group).

Compute the by time, across groups, average of $D_{g,t}$ (there will be one for each time).

Generate the weight as

$$W_{g,t} \propto D_{g,t} - E_g[D_{g,t}|T = t] - E_t[D_{g,t}|G = g] + E_{g,t}[D_{g,t}]$$

Proof

The only command you need to know to do that is the one you already typically use to compute some descriptive statistics. In Stata, it will be the egen.

- sum D

Proof

The only command you need to know to do that is the one you already typically use to compute some descriptive statistics. In Stata, it will be the egen.

- sum D
- gen OverallAverage = r(mean)

Proof

The only command you need to know to do that is the one you already typically use to compute some descriptive statistics. In Stata, it will be the egen.

- sum D
- gen OverallAverage = r(mean)
- bysort groupid: egen timeAverage = mean(D)

Proof

The only command you need to know to do that is the one you already typically use to compute some descriptive statistics. In Stata, it will be the egen.

- sum D
- gen OverallAverage = r(mean)
- bysort groupid: egen timeAverage = mean(D)
- bysort year: egen GroupAverage = mean(D)

Proof

The only command you need to know to do that is the one you already typically use to compute some descriptive statistics. In Stata, it will be the egen.

- sum D
- gen OverallAverage = r(mean)
- bysort groupid: egen timeAverage = mean(D)
- bysort year: egen GroupAverage = mean(D)
- gen weight = D - timeAverage - GroupAverage + OverallAverage

Proof

The only command you need to know to do that is the one you already typically use to compute some descriptive statistics. In Stata, it will be the egen.

- sum D
- gen OverallAverage = r(mean)
- bysort groupid: egen timeAverage = mean(D)
- bysort year: egen GroupAverage = mean(D)
- gen weight = D - timeAverage - GroupAverage + OverallAverage
- sum weight if D==1

Proof

The only command you need to know to do that is the one you already typically use to compute some descriptive statistics. In Stata, it will be the egen.

- sum D
- gen OverallAverage = r(mean)
- bysort groupid: egen timeAverage = mean(D)
- bysort year: egen GroupAverage = mean(D)
- gen weight = D - timeAverage - GroupAverage + OverallAverage
- sum weight if D==1
  This last line shows the summary statistics of the weights;

Proof

The only command you need to know to do that is the one you already typically use to compute some descriptive statistics. In Stata, it will be the egen.

- sum D
- gen OverallAverage = r(mean)
- bysort groupid: egen timeAverage = mean(D)
- bysort year: egen GroupAverage = mean(D)
- gen weight = D - timeAverage - GroupAverage + OverallAverage
- sum weight if D==1
  This last line shows the summary statistics of the weights;it will show if you have negative weights.

Proof

The only command you need to know to do that is the one you already typically use to compute some descriptive statistics. In Stata, it will be the egen.

- sum D
- gen OverallAverage = r(mean)
- bysort groupid: egen timeAverage = mean(D)
- bysort year: egen GroupAverage = mean(D)
- gen weight = D - timeAverage - GroupAverage + OverallAverage
- sum weight if D==1
  This last line shows the summary statistics of the weights;it will show if you have negative weights.

  In Stata, type twowayfeweights.

Solutions

Only allow comparisons with the never-treated (or the not-yet-treated) group.

Solutions

Only allow comparisons with the never-treated (or the not-yet-treated) group.

No comparison between "timing groups", in which already treated acts as the control for the late treated group.

Solutions

Only allow comparisons with the never-treated (or the not-yet-treated) group.

No comparison between "timing groups", in which already treated acts as the control for the late treated group.

As Pedro Sant'anna likes to say: Regression is variation-hungry, and causal inference is variation-cautious.

Solutions

Only allow comparisons with the never-treated (or the not-yet-treated) group.

No comparison between "timing groups", in which already treated acts as the control for the late treated group.

As Pedro Sant'anna likes to say: Regression is variation-hungry, and causal inference is variation-cautious.

Tailor your procedure to only use the comparisons you trust.

Solutions

Only allow comparisons with the never-treated (or the not-yet-treated) group.

No comparison between "timing groups", in which already treated acts as the control for the late treated group.

As Pedro Sant'anna likes to say: Regression is variation-hungry, and causal inference is variation-cautious.

Tailor your procedure to only use the comparisons you trust.

You can do that by cleverly specifying your regression (see Wooldridge, 2021) by using lots of interaction terms.

Gardner's 2SDiD

Parallel paths specify a particular process for $y(0)$, not for y.

$$y(0) = \alpha_{g_i} + \lambda_t + \epsilon_{it}$$

This is essentially what parallel paths imply, not more, not less.

Gardner's 2SDiD

Parallel paths specify a particular process for $y(0)$, not for y.

$$y(0) = \alpha_{g_i} + \lambda_t + \epsilon_{it}$$

This is essentially what parallel paths imply, not more, not less.

Estimate the group (α) and the time (λ) fixed-effects using data for which $D_{it} = 0$ (before the treatment).

Gardner's 2SDiD

Parallel paths specify a particular process for $y(0)$, not for y.

$$y(0) = \alpha_{g_i} + \lambda_t + \epsilon_{it}$$

This is essentially what parallel paths imply, not more, not less.

Estimate the group (α) and the time (λ) fixed-effects using data for which $D_{it} = 0$ (before the treatment).

Generate $\tilde{y}_i = y_i - \widehat{\alpha}_{g_i} + \widehat{\lambda}_t$:

Gardner's 2SDiD

Parallel paths specify a particular process for $y(0)$, not for y.

$$y(0) = \alpha_{g_i} + \lambda_t + \epsilon_{it}$$

This is essentially what parallel paths imply, not more, not less.

Estimate the group (α) and the time (λ) fixed-effects using data for which $D_{it} = 0$ (before the treatment).

Generate $\tilde{y}_i = y_i - \widehat{\alpha}_{g_i} + \widehat{\lambda}_t$:

$$\tilde{y}_i = \beta_0 + \beta D_{it} + \epsilon_{it}$$

Under parallel path assumptions, the error is allowed to be correlated with the level of y through α_i.

Gardner's 2SDiD

Parallel paths specify a particular process for $y(0)$, not for y.

$$y(0) = \alpha_{g_i} + \lambda_t + \epsilon_{it}$$

This is essentially what parallel paths imply, not more, not less.

Estimate the group (α) and the time (λ) fixed-effects using data for which $D_{it} = 0$ (before the treatment).

Generate $\tilde{y}_i = y_i - \widehat{\alpha}_{g_i} + \widehat{\lambda}_t$:

$$\tilde{y}_i = \beta_0 + \beta D_{it} + \epsilon_{it}$$

Under parallel path assumptions, the error is allowed to be correlated with the level of y through α_i. There is no OVB in here anymore. We also know how to think of OLS with a binary regressor.

2SDID

Back to that last example:

Source	SS	df	MS	Number of obs	=	1,800
				F(1, 1798)	=	854.16
Model	5878808.86	1	5878808.86	Prob > F	=	0.0000
Residual	12374893.6	1,798	6882.5882	R-squared	=	0.3221
				Adj R-squared	=	0.3217
Total	18253702.4	1,799	10146.5828	Root MSE	=	82.961

dy	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
D	114.4167	3.914901	29.23	0.000	106.7385	122.095
_cons	-5.44e-08	2.830605	-0.00	1.000	-5.551622	5.551621

2SDID

Back to that last example:

Source	SS	df	MS	Number of obs	=	1,800
				F(1, 1798)	=	854.16
Model	5878808.86	1	5878808.86	Prob > F	=	0.0000
Residual	12374893.6	1,798	6882.5882	R-squared	=	0.3221
				Adj R-squared	=	0.3217
Total	18253702.4	1,799	10146.5828	Root MSE	=	82.961

dy	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
D	114.4167	3.914901	29.23	0.000	106.7385	122.095
_cons	-5.44e-08	2.830605	-0.00	1.000	-5.551622	5.551621

The ATE in this example is around 105.

Covariates

First thing to ask:

Covariates

First thing to ask: Why add covariates?

Covariates

First thing to ask: Why add covariates?

There are good reasons to add and good reasons to avoid doing so.

Covariates

First thing to ask: Why add covariates?

There are good reasons to add and good reasons to avoid doing so.

The good reasons:

- They might increase precision.

Covariates

First thing to ask: Why add covariates?

There are good reasons to add and good reasons to avoid doing so.

The good reasons:

- They might increase precision.
- You only believe in parallel paths *conditional* on their levels.

Covariates

First thing to ask: Why add covariates?

There are good reasons to add and good reasons to avoid doing so.

The good reasons:

- They might increase precision.
- You only believe in parallel paths *conditional* on their levels.

The counterpoint:

- My understanding is that parallel paths of $Y(0)$ versus $Y(0)|X$ are non-nested.
- They certainly will add another layer of comparisons that you did not think through. Even the outcomes of individuals with the same level of the treatment will be compared and affect your estimate of β, as long as they have different levels of expected treatment (given X).

Covariates

What is the interpretation of the TWFE when you add covariates?

$$y_{it} = \alpha_{g_i} + \lambda_t + \delta' \mathbf{X_{it}} + \beta D_{it} + \epsilon_{it}$$

Well, it is complicated.

Covariates

What is the interpretation of the TWFE when you add covariates?

$$y_{it} = \alpha_{g_i} + \lambda_t + \delta' \mathbf{X_{it}} + \beta D_{it} + \epsilon_{it}$$

Well, it is complicated. It gets easier if you use propensity scores or add a bunch of interactions to deal with different effects at different levels of $\mathbf{X}$. (See Callaway and Sant'Anna).

Covariates

"Adding controls introduces a new source of identifying variation – within-group changes in "predicted treatment" – that was not there in the unadjusted version."

Covariates

"Adding controls introduces a new source of identifying variation – within-group changes in "predicted treatment" – that was not there in the unadjusted version."

Goodman Bacon showed that a DiD with covariates is a weighted average of the 2x2 DiD with covariate-adjusted levels of Y (which is what you hope that it would do) plus another term that comes entirely from within-cell variation (no actual variation on D).

Covariates

Let's ask a simpler question.
Suppose that you have a cross-section and that conditional on the vector of covariates, the treatment is randomly assigned.

$$E[Y(1)|X, D] = E[Y(1)|X]$$

$$E[Y(0)|X, D] = E[Y(0)|X]$$

This is the standard setting in which we discuss regression, matching, etc.

Covariates

What is the right way to interpret the coefficient of a regression of y on D controlling for X?

Let's assume that the conditional mean function of $Y(1)$ and $Y(0)$ are linear in $E[D|X] = p(x)$, the propensity score (otherwise things will get even worse):

$$E[Y(0)|X, D] = E[Y(0)|X] = a_0 + b_0 p(X)$$

$$E[Y(1)|X, D] = E[Y(0)|X] = a_1 + b_1 p(X)$$

Covariates – Słoczyński (2022)

If it is a causal effect – or some approximation for it – which one is it?

Covariates – Słoczyński (2022)

If it is a causal effect – or some approximation for it – which one is it? The ATT, ATE, ATU?

Covariates – Słoczyński (2022)

If it is a causal effect – or some approximation for it – which one is it? The ATT, ATE, ATU? Some weighted average of the causal effects for everyone?

$$\tau_{ols} = \omega_1 \tau_{ATT} + \omega_0 \tau_{ATU}$$

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

Covariates – Słoczyński (2022)

If it is a causal effect – or some approximation for it – which one is it? The ATT, ATE, ATU? Some weighted average of the causal effects for everyone?

$$\tau_{ols} = \omega_1 \tau_{ATT} + \omega_0 \tau_{ATU}$$

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

The Good: So, it is indeed some weighted average of the effects of the treatment on the treated and the average effect of the treatment on those not treated.

The Bad: The weights are weird.

Covariates – Słoczyński (2022)

If it is a causal effect – or some approximation for it – which one is it? The ATT, ATE, ATU? Some weighted average of the causal effects for everyone?

$$\tau_{ols} = \omega_1 \tau_{ATT} + \omega_0 \tau_{ATU}$$

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

The Good: So, it is indeed some weighted average of the effects of the treatment on the treated and the average effect of the treatment on those not treated.
The Bad: The weights are weird. Very weird.

Covariates

Assume that $Var[p(x)|D = 1] \approx Var[p(x)|D = 0]$. Then:

$$\tau_{ols} = Pr[D = 0]\tau_{ATT} + Pr[D = 1]\tau_{ATU}$$

We are weighting the causal effects on the treated by the proportion of controls, and the causal effects on the controls by the proportion of treated.

Covariates

Assume that $Var[p(x)|D = 1] \approx Var[p(x)|D = 0]$. Then:

$$\tau_{ols} = Pr[D = 0]\tau_{ATT} + Pr[D = 1]\tau_{ATU}$$

We are weighting the causal effects on the treated by the proportion of controls, and the causal effects on the controls by the proportion of treated.

If you had a dataset in which $Pr[D = 1] = 0.5$, you are safe.

Covariates

Assume that $Var[p(x)|D = 1] \approx Var[p(x)|D = 0]$. Then:

$$\tau_{ols} = Pr[D = 0]\tau_{ATT} + Pr[D = 1]\tau_{ATU}$$

We are weighting the causal effects on the treated by the proportion of controls, and the causal effects on the controls by the proportion of treated.

If you had a dataset in which $Pr[D = 1] = 0.5$, you are safe. Also whenever your causal effects are homogeneous.

Covariates

Takeaway: You are not getting the ATT, or ATE, except in specific cases.

Covariates

Takeaway: You are not getting the ATT, or ATE, except in specific cases.

For small programs, $Pr[D = 1]$ is close to zero, (lots of controls observations, few units treated), your estimate is close to:

$$\tau_{ols} = 1\tau_{ATT}$$

In this case, you are probably safe and OLS is near the ATT, although it might be far form the ATE.

Covariates

Takeaway: You are not getting the ATT, or ATE, except in specific cases.

Covariates

Takeaway: You are not getting the ATT, or ATE, except in specific cases.

For larger programs $Pr[D = 1]$ is much larger than $Pr[D = 0]$, (lots of treated individuals, not as many controls), your estimate is close to:

$$\tau_{ols} = 1\tau_{ATU}$$

Your estimate is representative for those not yet treated, but not for those who actually took it.

Example 1 - Słoczyński (2022)

TABLE 1.—THE EFFECTS OF A TRAINING PROGRAM ON EARNINGS

	(1)	(2)	(3)	(4)
	Original estimates			
OLS	−3,437***	−78	623	794
	(612)	(596)	(610)	(619)
	Diagnostics			
$\hat{w}_0$	0.019	0.001	0.017	0.017
$\hat{w}_0^* = \hat{\rho}$	0.011	0.011	0.011	0.011
$\hat{\delta}$	−0.970	−0.987	−0.971	−0.971
$\hat{\delta}^* = 2\hat{\rho} - 1$	−0.977	−0.977	−0.977	−0.977
	Decomposition			
$\widehat{\text{ATT}}$	−3,373***	−69	754	928
	(620)	(595)	(619)	(630)
$\hat{w}_1$	0.981	0.999	0.983	0.983
$\widehat{\text{ATU}}$	−6,753***	−6,289**	−6,841***	−6,840***
	(1,219)	(2,807)	(1,294)	(1,319)
$\hat{w}_0$	0.019	0.001	0.017	0.017
$\widehat{\text{ATE}}$	−6,714***	−6,218**	−6,754***	−6,751***
	(1,206)	(2,777)	(1,281)	(1,305)
Demographic controls	√		√	√
Earnings in 1974				√
Earnings in 1975		√	√	√
$\hat{\rho} = \hat{\text{P}}(d = 1)$	0.011	0.011	0.011	0.011
Observations	16,177	16,177	16,177	16,177

The estimates in the top panel correspond to column 2 in table 3.3.3 in Angrist and Pischke (2009, p. 89). The dependent variable is earnings in 1978. Demographic controls include age, age squared, years of schooling, and indicators for married, high school dropout, Black, and Hispanic. For treated individuals, earnings in 1974 correspond to real earnings in months 13 to 24 prior to randomization, which overlaps with calendar year 1974 for a number of individuals. Formulas for w_0, w_1, and δ are given in theorem 1 and corollary 2. Following these results, $\text{OLS} = \hat{w}_1 \times \widehat{\text{ATT}} + \hat{w}_0 \times \widehat{\text{ATU}}$. Estimates of ATE, ATT, and ATU are sample analogs of τ_{APLE}, $\tau_{APLE,1}$, and $\tau_{APLE,0}$, respectively. Also, $\widehat{\text{ATE}} = \hat{\rho} \times \widehat{\text{ATT}} + (1 - \hat{\rho}) \times \widehat{\text{ATU}}$. Huber–White standard errors (OLS) and bootstrap standard errors ($\widehat{\text{ATE}}$, $\widehat{\text{ATT}}$, and $\widehat{\text{ATU}}$) are in parentheses. Statistically significant at *10%, **5%, and ***1%.

Example 1 - Słoczyński (2022)

TABLE 2.—THE EFFECTS OF CASH TRANSFERS ON LONGEVITY

	(1)	(2)	(3)	(4)
	Original estimates			
OLS	0.0157***	0.0158***	0.0182***	0.0167***
	(0.0058)	(0.0059)	(0.0062)	(0.0061)
	Diagnostics			
$\hat{w}_0$	0.861	0.870	0.784	0.784
$\hat{w}_0^* = \hat{\rho}$	0.875	0.875	0.875	0.875
$\hat{\delta}$	0.736	0.745	0.659	0.659
$\hat{\delta}^* = 2\hat{\rho} - 1$	0.750	0.750	0.750	0.750
	Decomposition			
$\widehat{\text{ATT}}$	0.0129**	0.0149**	0.0097	0.0089
	(0.0064)	(0.0071)	(0.0078)	(0.0079)
$\hat{w}_1$	0.139	0.130	0.216	0.216
$\widehat{\text{ATU}}$	0.0162***	0.0160***	0.0206***	0.0188***
	(0.0057)	(0.0059)	(0.0063)	(0.0064)
$\hat{w}_0$	0.861	0.870	0.784	0.784
$\widehat{\text{ATE}}$	0.0133**	0.0150**	0.0110	0.0102
	(0.0063)	(0.0068)	(0.0073)	(0.0074)
State fixed effects	√			
County fixed effects			√	√
Cohort fixed effects	√	√	√	√
State characteristics		√	√	√
County characteristics		√		
Individual characteristics		√	√	√
$\hat{\rho} = \hat{\text{P}}(d = 1)$	0.875	0.875	0.875	0.875
Observations	7,860	7,859	7,859	7,857

The estimates in the top panel correspond to columns 1 to 4 in panel A of table 4 in Aizer et al. (2016, p. 952). The dependent variable is log age at death, as reported in the MP records (columns 1 to 3) or on the death certificate (column 4). State, county, and individual characteristics are listed in table E2.1 in online appendix E2. Formulas for w_0, w_1, and δ are given in theorem 1 and corollary 2. Following these results, OLS $= \hat{w}_1 \times \widehat{\text{ATT}} + \hat{w}_0 \times \widehat{\text{ATU}}$. Estimates of ATE, ATT, and ATU are sample analogs of τ_{APLE}, $\tau_{APLE,1}$, and $\tau_{APLE,0}$, respectively. Also, $\widehat{\text{ATE}} = \hat{\rho} \times \widehat{\text{ATT}} + (1 - \hat{\rho}) \times \widehat{\text{ATU}}$. Huber–White standard errors (OLS) and bootstrap standard errors ($\widehat{\text{ATE}}$, $\widehat{\text{ATT}}$, and $\widehat{\text{ATU}}$) are in parentheses. Statistically significant at *10%, **5%, and ***1%.

OLS implicit weights

Suppose that you see the output of this regression:

```
reg y x d
```

Source	SS	df	MS	Number of obs	=	10,000
				F(2, 9997)	>	99999.00
Model	225906355	2	112953178	Prob > F	=	0.0000
Residual	2801636.73	9,997	280.247748	R-squared	=	0.9878
				Adj R-squared	=	0.9877
Total	228707992	9,999	22873.0865	Root MSE	=	16.741

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	9.978818	.0111145	897.82	0.000	9.957031	10.0006
d	.144409	.3348259	0.43	0.666	-.5119172	.8007351
_cons	-.0881683	.2376766	-0.37	0.711	-.5540624	.3777257

OLS implicit weights – Aronow and Samii

Does any of this matters in practice?

OLS implicit weights – Aronow and Samii

Does any of this matters in practice?

TABLE 1 Summary Statistics for the Nominal and Effective Samples Relevant to Model (4) in Table 4A of Gerber and Huber Gerber and Huber (2010)

	Nominal Sample		Effective Sample	
Variable	**Mean**	**S.D.**	**Mean**	**S.D.**
Party ID (Scale: −2 = Strong Republican to +2 = Strong Democrat)	0.05	1.35	−0.10	1.74
Age (years)	47.58	15.11	48.06	15.78
Female (1 = yes)	0.52	0.50	0.52	0.50
Hispanic (1 = yes)	0.04	0.21	0.06	0.23
Black (1 = yes)	0.04	0.19	0.03	0.16
Union member (1 = yes)	0.08	0.27	0.07	0.26
Income (Scale: 0 to 1)	0.59	0.26	0.58	0.25
Income Refused/Don't Know	0.11	0.31	0.09	0.29
Education (Scale: 0 to 5)	2.51	1.32	2.41	1.30
Pre-election household income forecast	0.46	0.96	0.54	0.94
Post-election household income forecast	0.42	0.98	0.48	0.95
Pre-election national economy forecast	−0.11	0.97	0.04	1.01
Post-election national economy forecast	−0.14	0.89	−0.05	0.90
Log change in holiday spending	−0.04	1.19	−0.02	1.09
Log change in vacation spending	0.05	2.18	0.05	2.21
Pre-election happiness	1.93	0.83	2.01	0.84
Post-election happiness	1.95	0.76	2.02	0.75
Pre-election state economy forecast	0.04	0.84	0.11	0.85
Post-election state economy forecast	−0.03	0.88	0.00	0.88

Aronow and Samii

Jensen (2003) studies the effects of regime type on FDI using cross-section, time-series cross-sectional, and other regression analyses of a set of 114 countries observed over the years 1970 to 1997. The results suggest that "democratic political institutions are associated with higher levels of FDI inflows".

Aronow and Samii

The specification incorporates country and decade-fixed effects lagged FDI, and a set of control variables, including lagged values of market size, development level, growth, trade, budget deficit, government consumption, and democracy. The resulting estimate implies that a one-unit increase in the Polity score corresponds to a 0.020 increase in net FDI inflows as a percentage of gross domestic product ($p < 0.001$).

Aronow and Samii

The specification incorporates country and decade-fixed effects lagged FDI, and a set of control variables, including lagged values of market size, development level, growth, trade, budget deficit, government consumption, and democracy. The resulting estimate implies that a one-unit increase in the Polity score corresponds to a 0.020 increase in net FDI inflows as a percentage of gross domestic product ($p < 0.001$).

For which population is this effect representative?

Aronow and Samii

From the sample of 114 countries, 12 contribute over half (51%) of the weight used to construct the estimate of the effect of regime type on FDI, and 32 contribute 90% of the weight.

Aronow and Samii

From the sample of 114 countries, 12 contribute over half (51%) of the weight used to construct the estimate of the effect of regime type on FDI, and 32 contribute 90% of the weight.
The top 12 contributing countries, in descending order of their weights are Uruguay, Hungary, Niger, Philippines, Argentina, Madagascar, Pakistan, Zimbabwe, Poland, Peru, Lesotho, and Belarus.

Importantly, the lowest-contributing 14 countries combine to contribute less than 0.05% of the weight used to construct the effect estimate. This lowest-contributing countries, in ascending order of their weights, are Central African Republic, Albania, Germany, Russia, Haiti, Benin, South Africa, Yemen, Honduras, DR Congo, Czech Republic, Lithuania, China, and Latvia.

Aronow and Samii

We see that the findings are driven primarily by the experiences of Latin American, Eastern European, and African cases, the exceptions being the Philippines and Pakistan. A substantively important finding is that none of the fast-growing economies of East Asia (the "Asian Tiger" economies of Hong Kong, Singapore, South Korea, and Taiwan, along with Indonesia, Malaysia, or Thailand) are major contributors to this estimate, even though they may be countries of interest for this study.

Aronow and Samii

In addition, large and theoretically interesting countries such as Russia, China, and Germany are essentially not represented at all in the estimate of the causal effect. Second, the results help in determining which cases one may want to investigate further to check the accuracy of one's interpretation of the quantitative results. This provides a refinement to the nested mixed-methods strategies proposed by Lieberman (2005) for using quantitative results to select cases for qualitative investigation.

Aronow and Samii

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.

Mundlack Device

Remember that we started by recognizing that if we have parallel paths of untreated potential outcomes $Y(0)$, then a difference-in-differences strategy could be used to obtain estimates of the ATT.

Then we realized that a regression of the outcome with two-ways (group and time) fixed-effects had to yield numerically the same estimate, so it also could be used to obtain the ATT.

Well, there was another way we could specify the regression.

$$y_{it} = \theta_0 + \theta_1 D_{i,t} + \theta_2 \widehat{E}_g[D_{g,t}|T = t] + \theta_3 \widehat{E}_t[D_{g,t}|G = g] + u_{i,t}$$

That is, we regress the outcome on the treatment status D and two controls: one is the time-average value of the treatment, and the other is the group average value of the treatment.

Turns out, this will give you the same estimate for the effect of the treatment $\widehat{\theta}_1$ as the TWFE and the DiD.

This trick is called the (two-way) Mundlack device.

Mundlack Device

Mundlack used it to show a different way to run standard fixed-effects models.

Mundlack Device

Mundlack used it to show a different way to run standard fixed-effects models. Instead of adding a dummy for each group, you could instead add as a control for the average value of all covariates in each group to your regression. The coefficients on the X will be numerically identical to the fixed-effect regression coefficient.

Take-away:

Mundlack Device

Mundlack used it to show a different way to run standard fixed-effects models. Instead of adding a dummy for each group, you could instead add as a control for the average value of all covariates in each group to your regression. The coefficients on the X will be numerically identical to the fixed-effect regression coefficient.

Take-away:An entirely analogous way to add, or to control for fixed-effects, is to add as a covariate the average value of the regressors.

Mundlack Device

Mundlack used it to show a different way to run standard fixed-effects models. Instead of adding a dummy for each group, you could instead add as a control for the average value of all covariates in each group to your regression. The coefficients on the X will be numerically identical to the fixed-effect regression coefficient.

Take-away:An entirely analogous way to add, or to control for fixed-effects, is to add as a covariate the average value of the regressors. This will force OLS to rely only on the residualized version of the variation of the key covariates after purging its correlation with the group averages. It implicitly gives you the within transformation that the fixed-effects regression relies upon.

Mundlack Device

This generalizes to the TWFE: You just need the time and the group averages as controls.

Mundlack Device

This generalizes to the TWFE: You just need the time and the group averages as controls.

If you have a panel of 50 states and 20 years, the two-way fixed effects will require around 70 fixed-effects (one for each year and one for each group).

Mundlack Device

This generalizes to the TWFE: You just need the time and the group averages as controls.

If you have a panel of 50 states and 20 years, the two-way fixed effects will require around 70 fixed-effects (one for each year and one for each group).

You will get exactly the same coefficient on your treatment variable (the policy) if you add two regressors (the group average and the time average).

Mundlack Device

This generalizes to the TWFE: You just need the time and the group averages as controls.

If you have a panel of 50 states and 20 years, the two-way fixed effects will require around 70 fixed-effects (one for each year and one for each group).

You will get exactly the same coefficient on your treatment variable (the policy) if you add two regressors (the group average and the time average).
Isn't that cute?

This turns out to be quite helpful when dealing with continuous treatments.

Multivalued and continuous treatments

Often, the policy/variable of interest takes more than just two values.

Multivalued and continuous treatments

Often, the policy/variable of interest takes more than just two values. Sometimes, the treatment has different dosages (no treatment, low dosage, high dosage).
The applied practice until a few years ago was to run the following regression:

$$y_{igt} = \beta D_{igt} + \alpha_{g_i} + \lambda_t + \epsilon_{igt}$$

That is, one would run an OLS regression of the treatment intensity controlling for group and time fixed-effects.

Multivalued and continuous treatments

Often, the policy/variable of interest takes more than just two values. Sometimes, the treatment has different dosages (no treatment, low dosage, high dosage).
The applied practice until a few years ago was to run the following regression:

$$y_{igt} = \beta D_{igt} + \alpha_{g_i} + \lambda_t + \epsilon_{igt}$$

That is, one would run an OLS regression of the treatment intensity controlling for group and time fixed-effects.

As you can imagine, this will be subject to similar issues.

Multivalued and continuous treatments

Often, the policy/variable of interest takes more than just two values. Sometimes, the treatment has different dosages (no treatment, low dosage, high dosage).
The applied practice until a few years ago was to run the following regression:

$$y_{igt} = \beta D_{igt} + \alpha_{g_i} + \lambda_t + \epsilon_{igt}$$

That is, one would run an OLS regression of the treatment intensity controlling for group and time fixed-effects.

As you can imagine, this will be subject to similar issues.

Multivalued and continuous treatments

Often, the policy/variable of interest takes more than just two values. Sometimes, the treatment has different dosages (no treatment, low dosage, high dosage).
The applied practice until a few years ago was to run the following regression:

$$y_{igt} = \beta D_{igt} + \alpha_{g_i} + \lambda_t + \epsilon_{igt}$$

That is, one would run an OLS regression of the treatment intensity controlling for group and time fixed-effects.

As you can imagine, this will be subject to similar issues.

Defining the parameter of interest

When the treatment can take more than two values, we need to slow down to even define what we are after.

Defining the parameter of interest

When the treatment can take more than two values, we need to slow down to even define what we are after.There is no such a thing as "the effect" of the treatment.

Defining the parameter of interest

When the treatment can take more than two values, we need to slow down to even define what we are after.There is no such a thing as "the effect" of the treatment. Maybe we should say "an effect" of the treatment.

Defining the parameter of interest

When the treatment can take more than two values, we need to slow down to even define what we are after.There is no such a thing as "the effect" of the treatment. Maybe we should say "an effect" of the treatment.

There is the effect of getting a check of 1000 bucks,

Defining the parameter of interest

When the treatment can take more than two values, we need to slow down to even define what we are after.There is no such a thing as "the effect" of the treatment. Maybe we should say "an effect" of the treatment.

There is the effect of getting a check of 1000 bucks,there is another effect of getting a check of 5 cents.

Defining the parameter of interest

When the treatment can take more than two values, we need to slow down to even define what we are after.There is no such a thing as "the effect" of the treatment. Maybe we should say "an effect" of the treatment.

There is the effect of getting a check of 1000 bucks,there is another effect of getting a check of 5 cents.

There is an effect of taking an aspirin; there is another effect of taking 20.

Defining the parameter of interest

When the treatment can take more than two values, we need to slow down to even define what we are after.There is no such a thing as "the effect" of the treatment. Maybe we should say "an effect" of the treatment.

There is the effect of getting a check of 1000 bucks,there is another effect of getting a check of 5 cents.

There is an effect of taking an aspirin; there is another effect of taking 20.

If the groups are heterogeneous, there is the effect of taking a small dosage (of the treatment) relative to not taking it for those who took it.

Defining the parameter of interest

When the treatment can take more than two values, we need to slow down to even define what we are after.There is no such a thing as "the effect" of the treatment. Maybe we should say "an effect" of the treatment.

There is the effect of getting a check of 1000 bucks,there is another effect of getting a check of 5 cents.

There is an effect of taking an aspirin; there is another effect of taking 20.

If the groups are heterogeneous, there is the effect of taking a small dosage (of the treatment) relative to not taking it for those who took it.

There is the effect of taking a large dosage relative to not taking it for those who took the large dosage.

Defining the parameter of interest

When the treatment can take more than two values, we need to slow down to even define what we are after.There is no such a thing as "the effect" of the treatment. Maybe we should say "an effect" of the treatment.

There is the effect of getting a check of 1000 bucks,there is another effect of getting a check of 5 cents.

There is an effect of taking an aspirin; there is another effect of taking 20.

If the groups are heterogeneous, there is the effect of taking a small dosage (of the treatment) relative to not taking it for those who took it.

There is the effect of taking a large dosage relative to not taking it for those who took the large dosage.

There is also the effect of taking a large dosage, relative to not taking it, for those who took the small dosage.

Defining the parameter of interest

Defining the parameter of interest

There is an effect for every pair of group (defined by the dosage taken) and treatment intensity (defined by the dosage into consideration).

Defining the parameter of interest

There is an effect for every pair of group (defined by the dosage taken) and treatment intensity (defined by the dosage into consideration). When the treatment is binary, there is only one possible dosage and two groups, so you get only two "effects": The ATT (the effect on those who took it) and the ATU (the effect on those not treated yet).

Defining the parameter of interest

There is an effect for every pair of group (defined by the dosage taken) and treatment intensity (defined by the dosage into consideration). When the treatment is binary, there is only one possible dosage and two groups, so you get only two "effects": The ATT (the effect on those who took it) and the ATU (the effect on those not treated yet).

With no restriction on how heterogeneous the effects can be, the more values the treatment can take, the more parameters there will be.

Defining the parameter of interest

To see this, suppose that the treatment takes three values: Zero, small (1), and large (2).
Effect of the small dosage (relative to no treatment), on those who took the small dosage (D=1):

$$E[Y(1) - Y(0)|D = 1]$$

Defining the parameter of interest

To see this, suppose that the treatment takes three values: Zero, small (1), and large (2).
Effect of the small dosage (relative to no treatment), on those who took the small dosage (D=1):

$$E[Y(1) - Y(0)|D = 1]$$

Effect of the large dosage (relative to no treatment) on those who took the large dosage (D=2):

$$E[Y(2) - Y(0)|D = 2]$$

Defining the parameter of interest

To see this, suppose that the treatment takes three values: Zero, small (1), and large (2).
Effect of the small dosage (relative to no treatment), on those who took the small dosage (D=1):

$$E[Y(1) - Y(0)|D = 1]$$

Effect of the large dosage (relative to no treatment) on those who took the large dosage (D=2):

$$E[Y(2) - Y(0)|D = 2]$$

Defining the parameter of interest

Effect of the large dosage (relative to no treatment) on those who took the small dosage (D=2):

$$E[Y(2) - Y(0)|D = 1]$$

Defining the parameter of interest

Effect of the large dosage (relative to no treatment) on those who took the small dosage (D=2):

$$E[Y(2) - Y(0)|D = 1]$$

Effect of the large dosage (relative to no treatment) on those who got no treatment (D=0):

$$E[Y(2) - Y(0)|D = 0]$$

Key Result

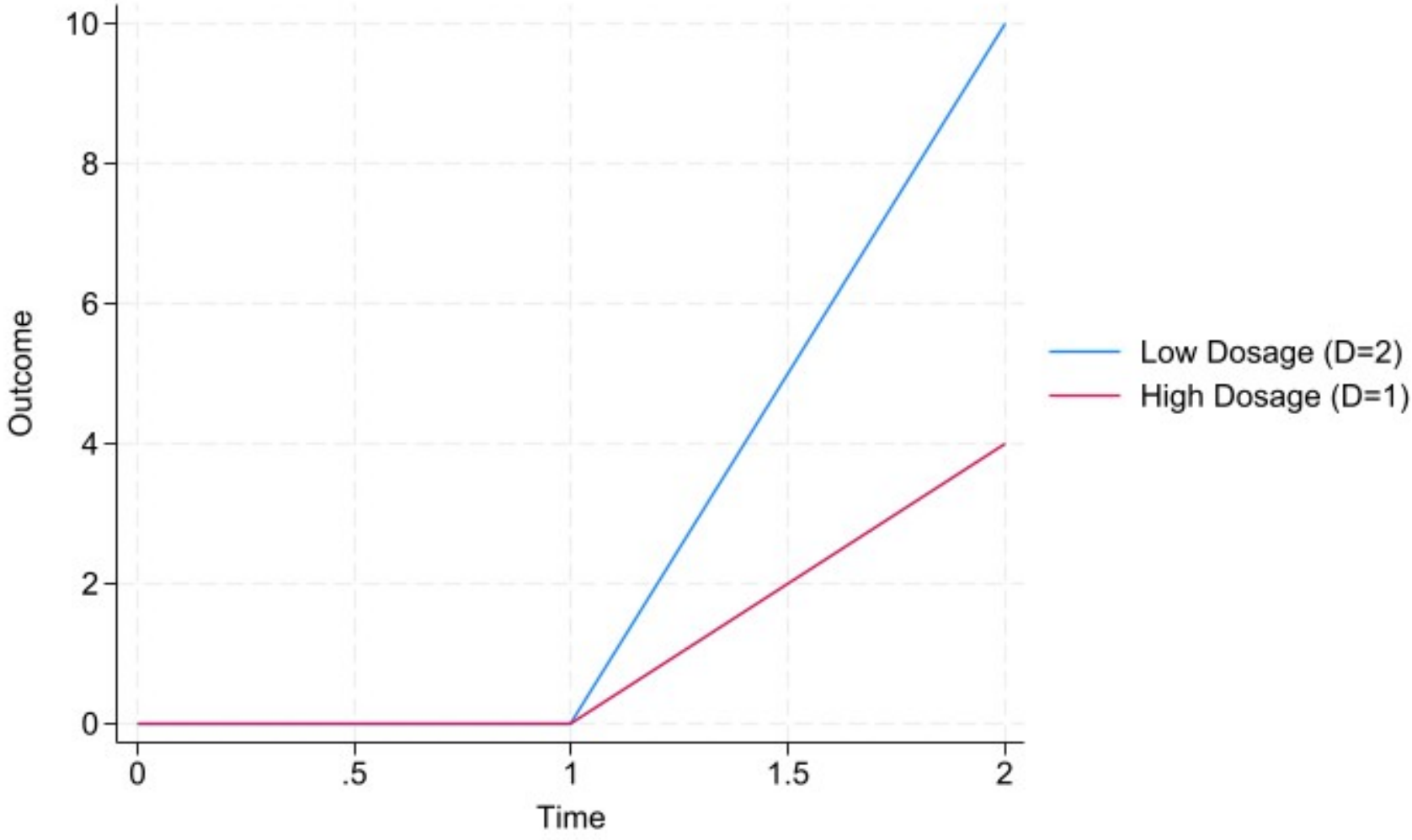
The standard TWFE constraints the marginal effect of every dosage to be the same for every group and dosage.

Key Result

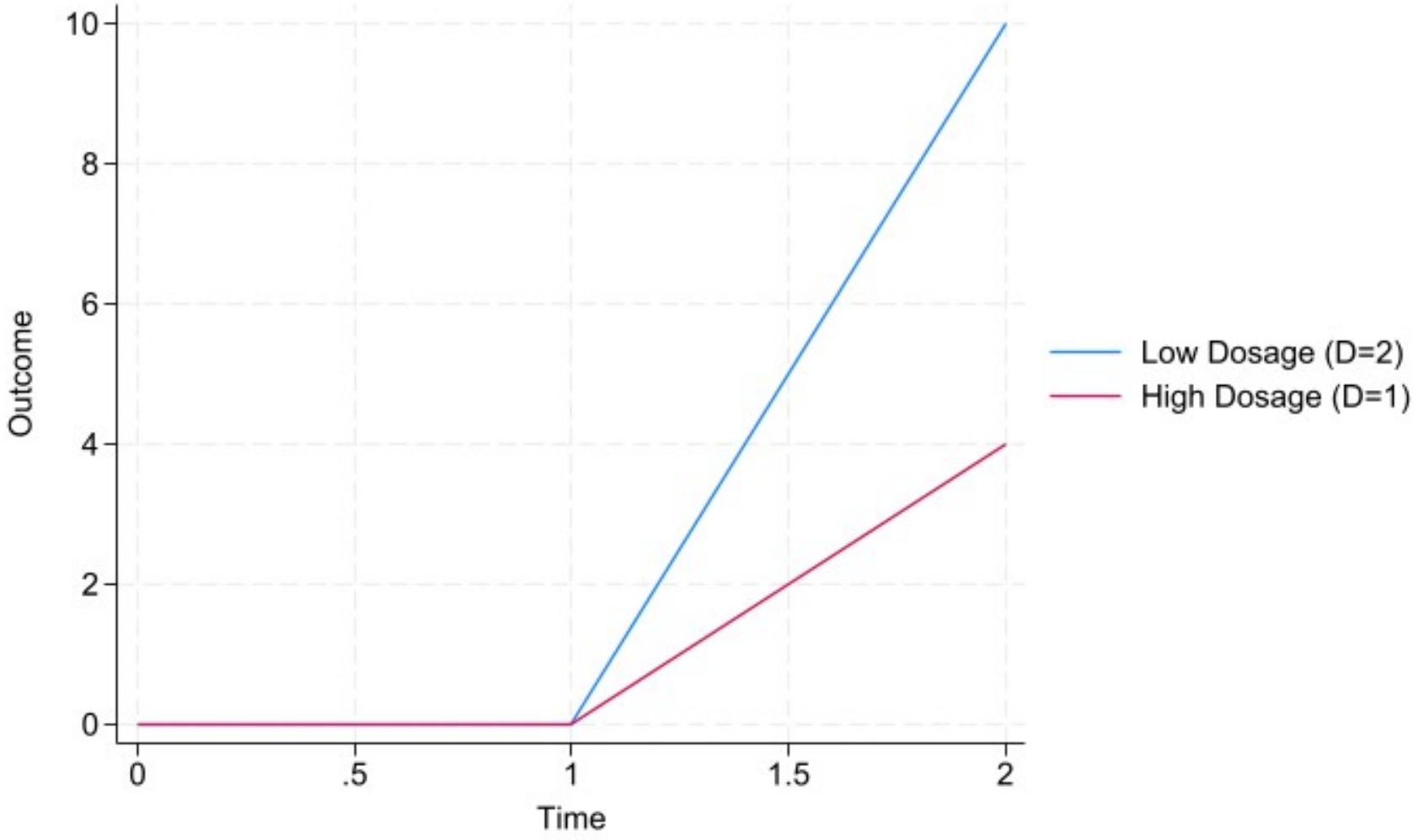
The standard TWFE constraints the marginal effect of every dosage to be the same for every group and dosage.

When there is heterogeneity, it might induce unexpected comparisons and negative weights.

One Example

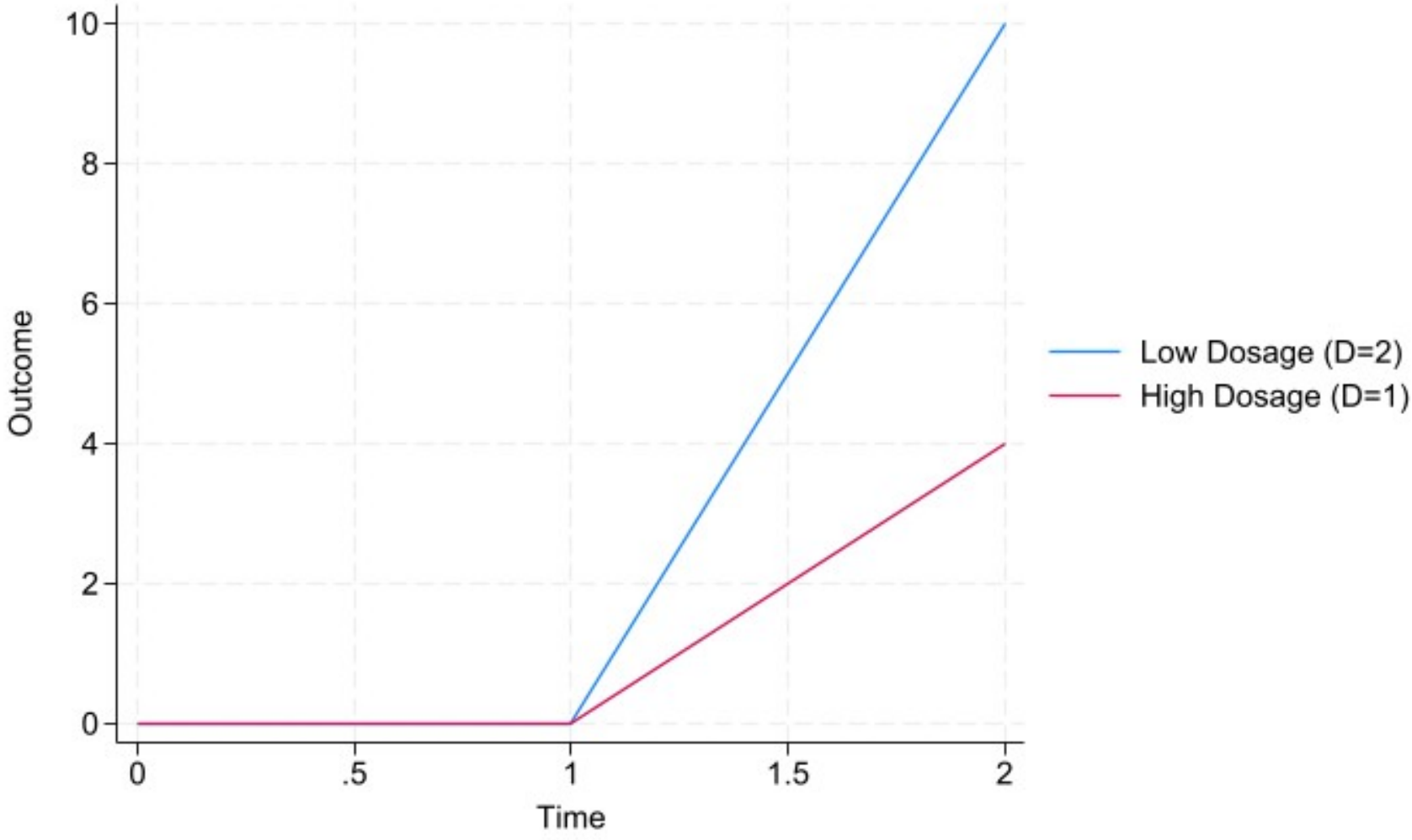


One Example



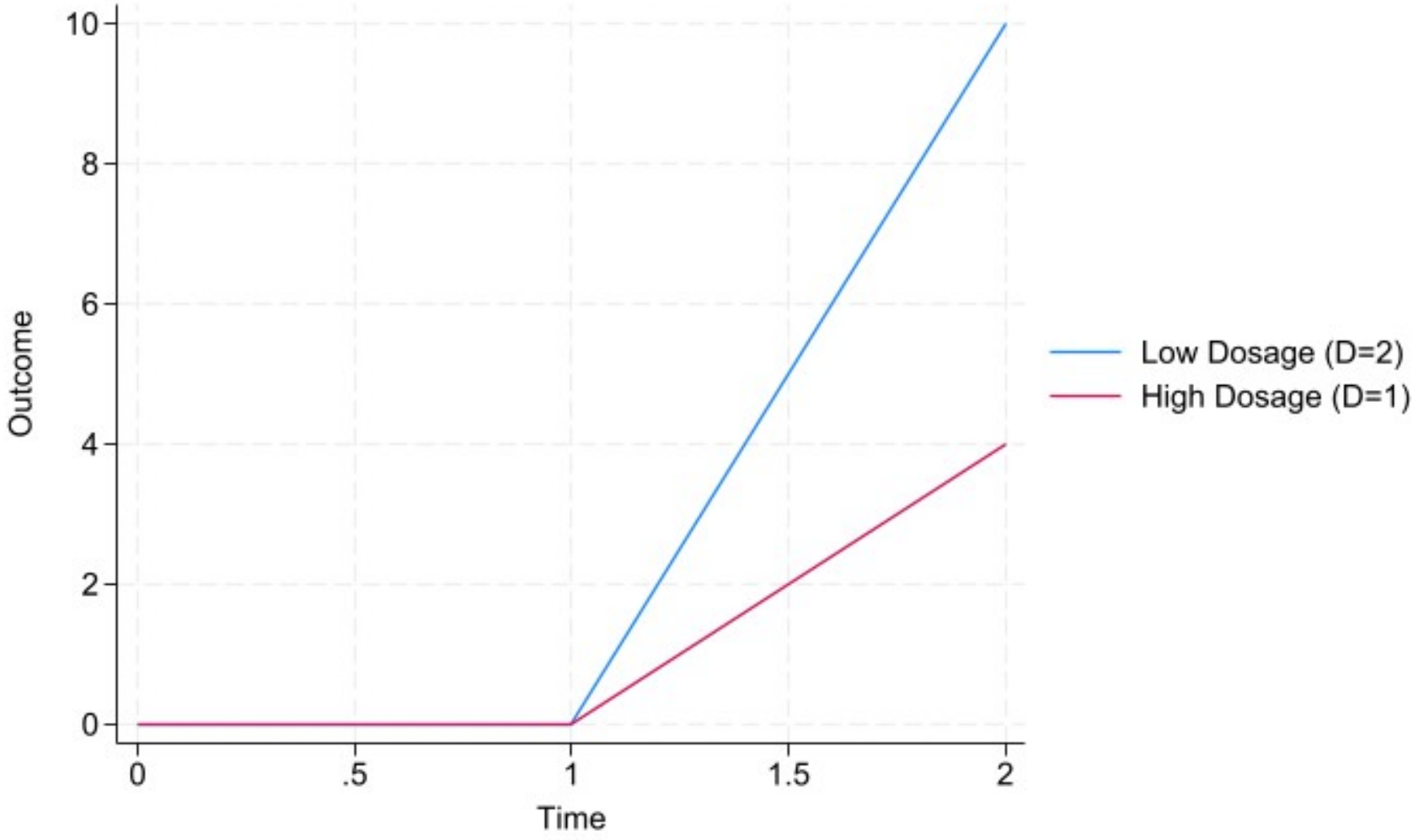
Parallel paths hold here.

One Example



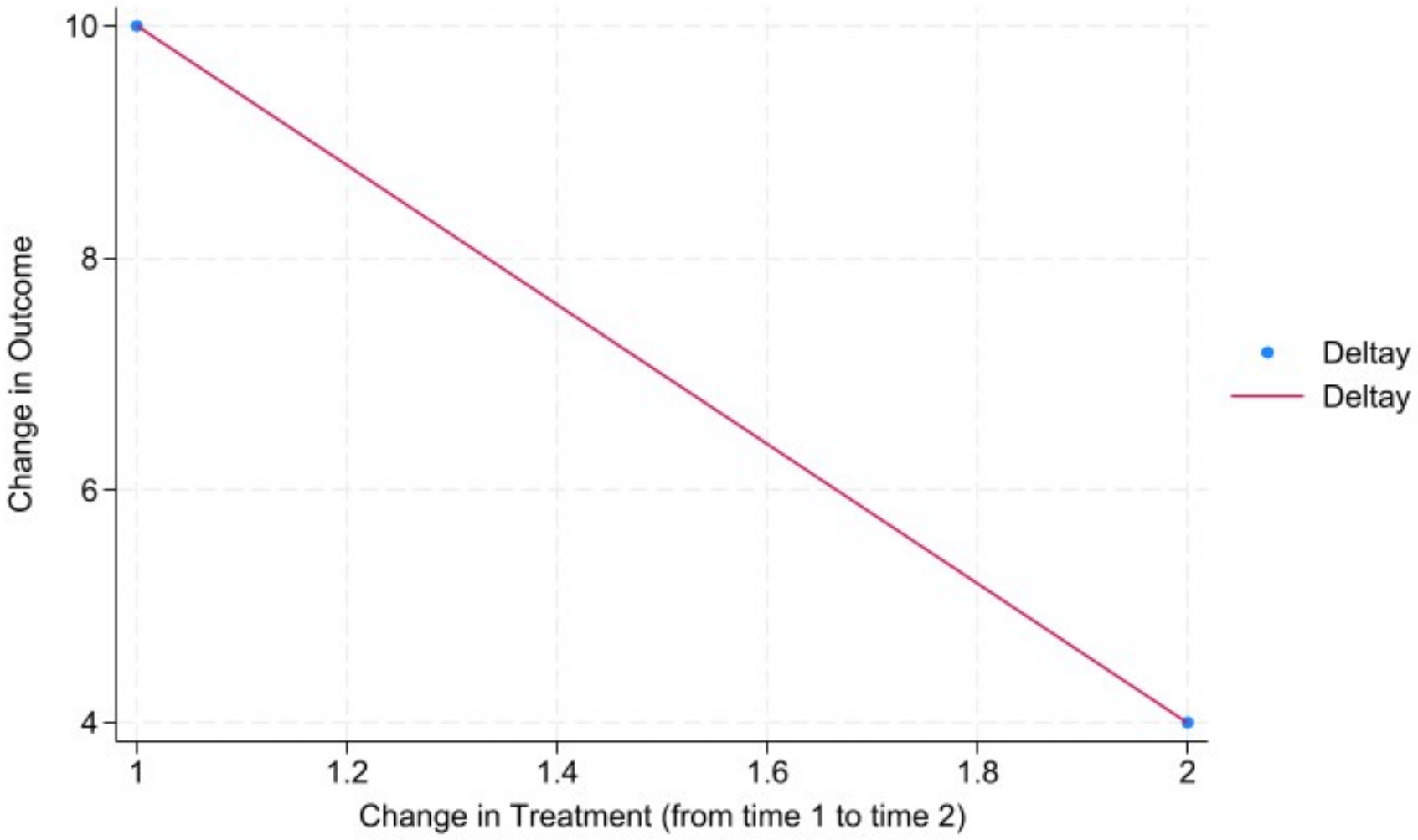
Parallel paths hold here. The effect is positive for both groups ($+4$ for the large dosage and $+10$ for the small dosage).

One Example



Parallel paths hold here. The effect is positive for both groups (+4 for the large dosage and +10 for the small dosage).What does the TWFE regression pick up?

TWFE



The TWFE coefficient on D = -6.

Why does the TWFE exhibit this behavior?

Why does the TWFE exhibit this behavior?

The group that had the largest change in the treatment had the *smallest change* in the outcome. If the effects are not heterogeneous, then the larger D is, the smaller y has to be.

Why does the TWFE exhibit this behavior?

The group that had the largest change in the treatment had the *smallest change* in the outcome. If the effects are not heterogeneous, then the larger D is, the smaller y has to be.

Note that the constant treatment effect assumption is implicitly enforced/assumed/used in the procedure.

Why does the TWFE exhibit this behavior?

The group that had the largest change in the treatment had the *smallest change* in the outcome. If the effects are not heterogeneous, then the larger D is, the smaller y has to be.

Note that the constant treatment effect assumption is implicitly enforced/assumed/used in the procedure.

This is actually the correct result – for a different parameter, the ATT of going from dosage 1 to 2 for the group that got the high dosage 2 – under *another assumption*, that is not parallel paths of $y(0)$ (untreated outcome).

Why does the TWFE exhibit this behavior?

The group that had the largest change in the treatment had the *smallest change* in the outcome. If the effects are not heterogeneous, then the larger D is, the smaller y has to be.

Note that the constant treatment effect assumption is implicitly enforced/assumed/used in the procedure.

This is actually the correct result – for a different parameter, the ATT of going from dosage 1 to 2 for the group that got the high dosage 2 – under *another assumption*, that is not parallel paths of $y(0)$ (untreated outcome). The assumption is that the large dosage group would have a change in outcome of $+10$ if it were treated with a dosage of 1 instead of 2.

Some concluding thoughts

The design (identification coming from paths that are supposed to be parallel) might be valid.

Some concluding thoughts

The design (identification coming from paths that are supposed to be parallel) might be valid. The key takeaway from this literature is to be careful in the implementation stage.

Some concluding thoughts

The design (identification coming from paths that are supposed to be parallel) might be valid. The key takeaway from this literature is to be careful in the implementation stage.

The more your setting departs from the textbook, two groups, two periods case, the more the implicit optimization from OLS might drive your estimand away from the target parameter.

Some concluding thoughts

The design (identification coming from paths that are supposed to be parallel) might be valid. The key takeaway from this literature is to be careful in the implementation stage.

The more your setting departs from the textbook, two groups, two periods case, the more the implicit optimization from OLS might drive your estimand away from the target parameter.

Proceed with caution, particularly if the treatment is: (i) staggered, (ii) multi-valued/continuous (iii) reversible, (iv) if you are adding covariates, and (v) if you expect some heterogeneity of effects.

Some concluding thoughts

The design (identification coming from paths that are supposed to be parallel) might be valid. The key takeaway from this literature is to be careful in the implementation stage.

The more your setting departs from the textbook, two groups, two periods case, the more the implicit optimization from OLS might drive your estimand away from the target parameter.

Proceed with caution, particularly if the treatment is: (i) staggered, (ii) multi-valued/continuous (iii) reversible, (iv) if you are adding covariates, and (v) if you expect some heterogeneity of effects.

You can still use OLS, but you will have to adjust your specification (adding interactions (Wooldridge), estimating the group and time fixed-effects using only untreated observations (Gardner's 2sDiD) to address these issues.

Taking stock

There is a debate on how important these issues are in practice.

Taking stock

There is a debate on how important these issues are in practice.

As with many other issues, there is a lot of heterogeneity.

Taking stock

There is a debate on how important these issues are in practice.

As with many other issues, there is a lot of heterogeneity.

Whenever treatment effects are close to homogeneous, this shouldn't matter much.

Taking stock

There is a debate on how important these issues are in practice.

As with many other issues, there is a lot of heterogeneity.

Whenever treatment effects are close to homogeneous, this shouldn't matter much.

Whenever treatment effect heterogeneity is independent of group indicators and time, it should not matter much again.

Taking Stock

However, at the very least, this literature pointed out a discrepancy between the variation that is supposed to be identifying the parameter of interest, as it is discussed in plain English in the body of (many) DiD papers, and the variation that is actually driving the parameter estimates.

Taking Stock

However, at the very least, this literature pointed out a discrepancy between the variation that is supposed to be identifying the parameter of interest, as it is discussed in plain English in the body of (many) DiD papers, and the variation that is actually driving the parameter estimates.

The TWFE exploits many natural experiments, sometimes more than we are aware of.

Taking Stock

However, at the very least, this literature pointed out a discrepancy between the variation that is supposed to be identifying the parameter of interest, as it is discussed in plain English in the body of (many) DiD papers, and the variation that is actually driving the parameter estimates.

The TWFE exploits many natural experiments, sometimes more than we are aware of.

Sometimes this is fine.

Taking Stock

However, at the very least, this literature pointed out a discrepancy between the variation that is supposed to be identifying the parameter of interest, as it is discussed in plain English in the body of (many) DiD papers, and the variation that is actually driving the parameter estimates.

The TWFE exploits many natural experiments, sometimes more than we are aware of.

Sometimes this is fine. Smart people knew how to avoid these problems all along (see Wooldridge, Angrist, Abadie).

Taking Stock

However, at the very least, this literature pointed out a discrepancy between the variation that is supposed to be identifying the parameter of interest, as it is discussed in plain English in the body of (many) DiD papers, and the variation that is actually driving the parameter estimates.

The TWFE exploits many natural experiments, sometimes more than we are aware of.

Sometimes this is fine. Smart people knew how to avoid these problems all along (see Wooldridge, Angrist, Abadie). But the way that they did things was more sophisticated than the applied practice.

Useful resources

On the process of working on this, I learned that there are plenty of good resources to help you in your DiD project.

Useful resources

On the process of working on this, I learned that there are plenty of good resources to help you in your DiD project.

Goodman-Bacon has a FAQ on his website about his Decomposition paper. It is called "So you have been told to do my Difference-in-Differences' thing: A guide".

Useful resources

On the process of working on this, I learned that there are plenty of good resources to help you in your DiD project.

Goodman-Bacon has a FAQ on his website about his Decomposition paper. It is called "So you have been told to do my Difference-in-Differences' thing: A guide".

Better than that, Asjad Naqvi has a GitHub page with R, Stata, and Julia code, walking you through everything relevant in the new DiD literature (estimators, decompositions, getting the weights).

Useful resources

On the process of working on this, I learned that there are plenty of good resources to help you in your DiD project.

Goodman-Bacon has a FAQ on his website about his Decomposition paper. It is called "So you have been told to do my Difference-in-Differences' thing: A guide".

Better than that, Asjad Naqvi has a GitHub page with R, Stata, and Julia code, walking you through everything relevant in the new DiD literature (estimators, decompositions, getting the weights). It has code to draw some nice simulated data so you can see how things work and when they don't. Here is the link: https://asjadnaqvi.github.io/DiD/

Useful resources

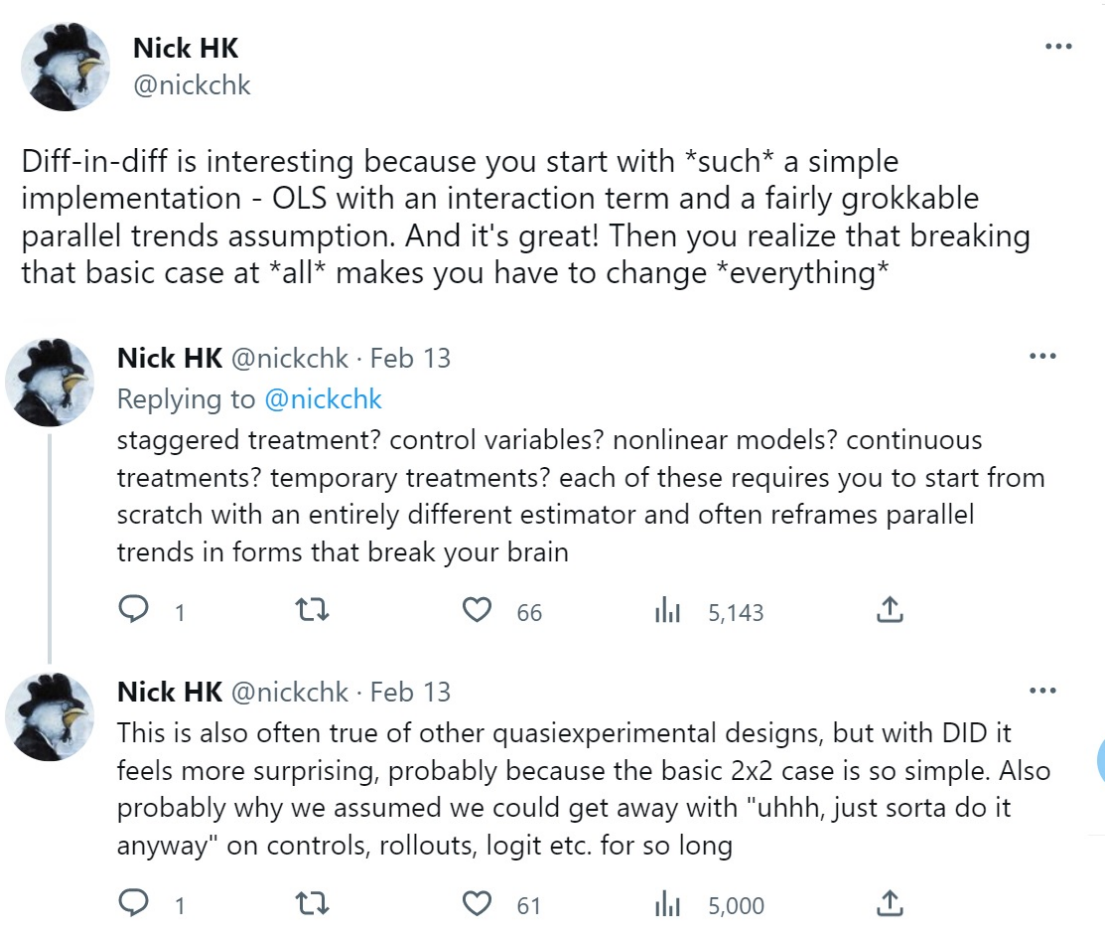
On the process of working on this, I learned that there are plenty of good resources to help you in your DiD project.

Goodman-Bacon has a FAQ on his website about his Decomposition paper. It is called "So you have been told to do my Difference-in-Differences' thing: A guide".

Better than that, Asjad Naqvi has a GitHub page with R, Stata, and Julia code, walking you through everything relevant in the new DiD literature (estimators, decompositions, getting the weights). It has code to draw some nice simulated data so you can see how things work and when they don't. Here is the link: https://asjadnaqvi.github.io/DiD/

I wish I had known all of that before I started writing code to generate my graphs.

Some concluding thoughts



Nick HK
@nickchk

Diff-in-diff is interesting because you start with *such* a simple implementation - OLS with an interaction term and a fairly grokkable parallel trends assumption. And it's great! Then you realize that breaking that basic case at *all* makes you have to change *everything*

Nick HK @nickchk · Feb 13
Replying to @nickchk
staggered treatment? control variables? nonlinear models? continuous treatments? temporary treatments? each of these requires you to start from scratch with an entirely different estimator and often reframes parallel trends in forms that break your brain

1 66 5,143

Nick HK @nickchk · Feb 13
This is also often true of other quasiexperimental designs, but with DID it feels more surprising, probably because the basic 2x2 case is so simple. Also probably why we assumed we could get away with "uhhh, just sorta do it anyway" on controls, rollouts, logit etc. for so long

1 61 5,000

References

Goodman-Bacon, Andrew. "Difference-in-differences with variation in treatment timing." Journal of Econometrics 225.2 (2021): 254-277.

Gardner, John. "Two-stage differences in differences." arXiv preprint arXiv:2207.05943 (2022).

Aronow, Peter M., and Cyrus Samii. "Does regression produce representative estimates of causal effects?." American Journal of Political Science 60.1 (2016): 250-267.

Słoczyński, Tymon. "Interpreting OLS estimands when treatment effects are heterogeneous: Smaller groups get larger weights." The review of economics and statistics 104.3 (2022): 501-509.

Słoczyński, Tymon. "When should we (not) interpret linear iv estimands as late?." arXiv preprint arXiv:2011.06695 (2020).

Callaway, Brantly, and Pedro HC Sant'Anna. "Difference-in-differences with multiple time periods." Journal of Econometrics 225.2 (2021): 200-230.

Roth, Jonathan, et al. "What's trending in difference-in-differences? A synthesis of the recent econometrics literature." arXiv preprint arXiv:2201.01194 (2022).

Wooldridge, Jeffrey M. "Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators." Available at SSRN 3906345 (2021).

Athey, Susan, and Guido W. Imbens. "Design-based analysis in difference-in-differences settings with staggered adoption." Journal of Econometrics 226.1 (2022): 62-79.

Arkhangelsky, Dmitry, et al. Synthetic difference in differences. No. w25532. National Bureau of Economic Research, 2019.