

First Annual CodeChella: Difference-in-differences

Presenter: Scott Cunningham

CODE-CHELLA



CodeChella Outline

- Introduction to basics of the simple design without and with covariates
 - Potential outcomes review
 - Regression and sample averages equivalence
 - Covariates and multiple periods
- Diagnosing the problems created by differential timing
 - Strict exogeneity and heterogenous treatment effects
 - Static specification
 - Dynamic specification (i.e., event study)

CodeChella Outline

- Differential timing and the Bacon decomposition
- Two solutions to differential timing
 - Manual aggregation
 - Explicit Imputation
- Fuzzy difference-in-differences

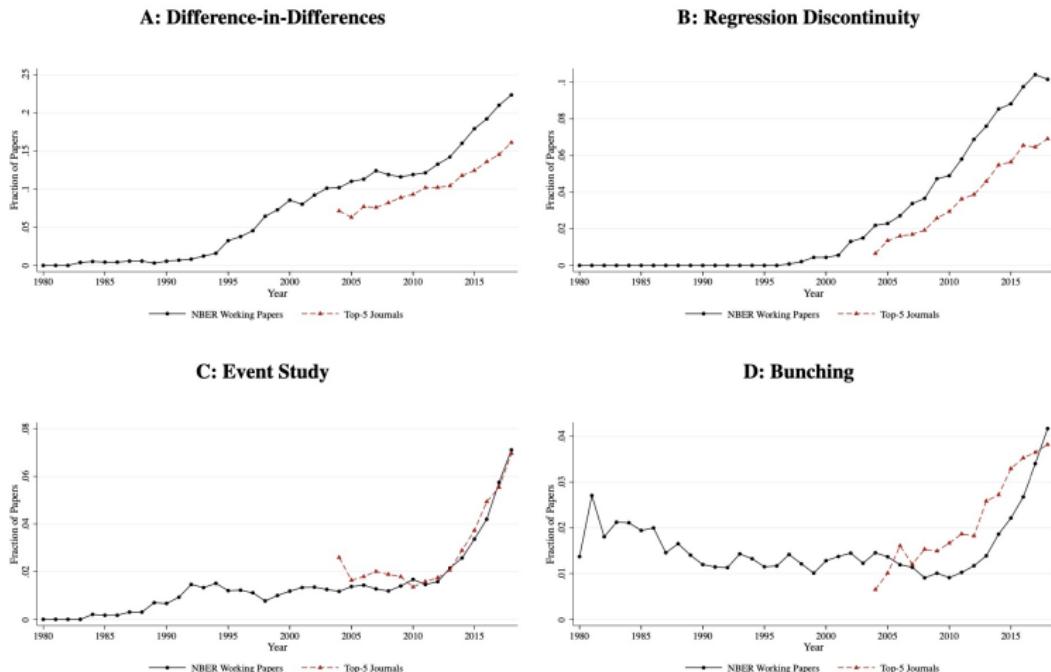
What is Difference-in-differences?

- A non-random treatment is applied to one or more groups
- A group of units do not receive the units at the same time (either never, or not yet, for comparison purposes)
- Observations are taken before and after for each group
- Researcher differences before and after, then differences the difference – hence the name

Difference-in-differences (DiD)

- Very old, conceptually intuitive research design
- Early attempts at using date back to several health policy debates in the 19th century
- Brought into labor economics with Orley Ashenfelter (1978), LaLonde (1985), Card and Krueger (1994)
- Has since become the most popular quasi-experimental method, even more than RDD

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.

Why is there entire seminar devoted to DiD?

- Good news: DiD is very popular because of economists interest in large potentially impactful policies
- Bad news: Lots of new papers suggesting standard methods are biased
- Good news: Lots of new papers suggesting solutions
- Bottom line: This is a chance for us to learn more applied econometrics – always a good thing

Pedagogy of the seminar

- Emphasize assumptions
- Emphasize papers stand-alone
- Weakness of the seminar: I can't really advise you on how these are connected to one another, as that level of depth I'm still working on myself
- Weakness of the seminar: It can feel like drinking from a firehose, but I mainly want you to learn the papers

Potential outcomes review

- DiD really can't be understood without committing to some common causality language
- Standard language is the potential outcomes model, sometimes called the Rubin-Neyman model
- Don't go over potential outcomes too fast or you'll miss all the fun
- Potential outcomes are thought experiments about worlds that never existed, but which *could have*
- Peter Hull and Pedro Sant'Anna insist that I use the wrong notation, but even smart people are wrong occasionally

Introduction to Counterfactuals and Causality

- Aliens come and orbit earth, see sick people in hospitals and conclude “these ‘hospitals’ are hurting people”
- Motivated by anger and compassion, they kill the doctors to save the patients
- Sounds stupid, but earthlings do this too - all the time
- Let’s look at the challenges of making causality synonymous with correlations

#1: Correlation and causality are very different concepts

- Causal question:

"If a doctor puts a patient on a ventilator (D), will her covid symptoms (Y) improve?"

- Correlation question:

$$\frac{1}{n} \frac{\text{Cov}(D, Y)}{\sqrt{\text{Var}_D} \sqrt{\text{Var}_Y}}$$

- These are not the same thing

#2: Coming first may not mean causality!

- Every morning the rooster crows and then the sun rises
- Did the rooster cause the sun to rise? Or did the sun cause the rooster to crow?
- *Post hoc ergo propter hoc*: “after this, therefore, because of this”



#3: No correlation does not mean no causality!

- A sailor sails her sailboat across a lake
- Wind blows, and she perfectly counters by turning the rudder
- The same aliens observe from space and say “Look at the way she’s moving that rudder back and forth but going in a straight line. That rudder is broken.” So they send her a new rudder
- They’re wrong but why are they wrong? There is, after all, no correlation
- Question: What if she had been moving the rudder by flipping coins?

Potential outcomes notation

- Let the treatment be a binary variable:

$$D_{i,t} = \begin{cases} 1 & \text{if hospitalized at time } t \\ 0 & \text{if not hospitalized at time } t \end{cases}$$

where i indexes an individual observation, such as a person

- Potential outcomes:

$$Y_{i,t}^j = \begin{cases} 1 & \text{health if hospitalized at time } t \\ 0 & \text{health if not hospitalized at time } t \end{cases}$$

where j indexes a counterfactual state of the world

Moving between worlds

- I'll drop t subscript, but note – these are potential outcomes for the same person at the exact same moment in time
- A potential outcome Y^1 is not the historical outcome Y either conceptually or notationally
- Potential outcomes are hypothetical states of the world but historical outcomes are ex post realizations
- Major philosophical move here: go from the potential worlds to the actual (historical) world based on your treatment assignment

Important definitions

Definition 1: Individual treatment effect

The individual treatment effect, δ_i , equals $Y_i^1 - Y_i^0$

Definition 3: Switching equation

An individual's observed health outcomes, Y , is determined by treatment assignment, D_i , and corresponding potential outcomes:

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

$$Y_i = \begin{cases} Y_i^1 & \text{if } D_i = 1 \\ Y_i^0 & \text{if } D_i = 0 \end{cases}$$

Definition 2: Average treatment effect (ATE)

The average treatment effect is the population average of all i individual treatment effects

$$\begin{aligned} E[\delta_i] &= E[Y_i^1 - Y_i^0] \\ &= E[Y_i^1] - E[Y_i^0] \end{aligned}$$

So what's the problem?

Definition 4: Fundamental problem of causal inference

If you need both potential outcomes to know causality with certainty, then since it is impossible to observe both Y_i^1 and Y_i^0 for the same individual, δ_i , is *unknowable*.

Conditional Average Treatment Effects

Definition 5: Average Treatment Effect on the Treated (ATT)

The average treatment effect on the treatment group is equal to the average treatment effect conditional on being a treatment group member:

$$\begin{aligned} E[\delta|D = 1] &= E[Y^1 - Y^0|D = 1] \\ &= E[Y^1|D = 1] - E[Y^0|D = 1] \end{aligned}$$

Definition 6: Average Treatment Effect on the Untreated (ATU)

The average treatment effect on the untreated group is equal to the average treatment effect conditional on being untreated:

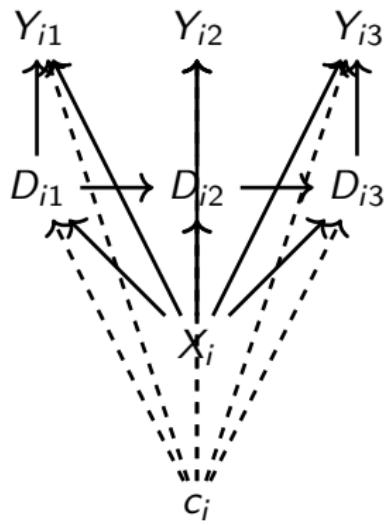
$$\begin{aligned} E[\delta|D = 0] &= E[Y^1 - Y^0|D = 0] \\ &= E[Y^1|D = 0] - E[Y^0|D = 0] \end{aligned}$$

Twoway fixed effects

- When working with panel data, the so-called “twoway fixed effects” (TWFE) estimator is the workhorse estimator
- It's easy to run, a version of OLS, and many people are just interested in mean effects anyway
- It's been the most common model for estimating treatment effects in a difference-in-differences, and so we need to spend some time understanding what it is

Panel Data

- Panel data: we observe the same units (individuals, firms, countries, schools, etc.) over several time periods
- Often our outcome variable depends on unobserved factors which are also correlated with our explanatory variable of interest
- If these omitted variables are constant over time, we can use panel data estimators to consistently estimate the effect of our explanatory variable



Sorry - drawing the DAG for a simple panel model is somewhat messy!

When to use this

- Traditionally, this was used for estimating constant treatment effects with unobserved time-invariant heterogeneity – recall the c_i was constant across all time periods
- It's a linear model, so you'll be estimating conditional mean treatment effects – if you want the median, you can't use this
- Once you enter into a world with dynamic treatment effects and differential timing, this loses all value

Problems that fixed effects cannot solve

- Reverse causality: Becker predicted police reduce crime, but when you regress crime onto police, it's usually positive
 - $\hat{\beta}_{FE}$ inconsistent unless strict exogeneity conditional on c_i holds
 - $E[\varepsilon_{it}|x_{i1}, x_{i2}, \dots, x_{iT}, c_i] = 0; t = 1, 2, \dots, T$
 - implies ε_{it} uncorrelated with past, current and future regressors
- Time-varying unobserved heterogeneity
 - It's the time-varying unobservables you have to worry about in fixed effects
 - Can include time-varying controls, but as always, don't condition on a collider
- Treatment effect heterogeneity under differential timing (Gardner 2021)

Formal panel notation

- Let y and $x \equiv (x_1, x_2, \dots, x_k)$ be observable random variables and c be an unobservable random variable
- We are interested in the partial effects of variable x_j in the population regression function

$$E[y|x_1, x_2, \dots, x_k, c]$$

Formal panel notation cont.

- We observe a sample of $i = 1, 2, \dots, N$ cross-sectional units for $t = 1, 2, \dots, T$ time periods (a balanced panel)
 - For each unit i , we denote the observable variables for all time periods as $\{(y_{it}, x_{it}) : t = 1, 2, \dots, T\}$
 - $x_{it} \equiv (x_{it1}, x_{it2}, \dots, x_{itk})$ is a $1 \times K$ vector
- Typically assume that cross-sectional units are i.i.d. draws from the population: $\{y_i, x_i, c_i\}_{i=1}^N \sim i.i.d.$ (cross-sectional independence)
 - $y_i \equiv (y_{i1}, y_{i2}, \dots, y_{iT})'$ and $x_i \equiv (x_{i1}, x_{i2}, \dots, x_{iT})$
 - Consider asymptotic properties with T fixed and $N \rightarrow \infty$

Formal panel notation

Single unit:

$$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{it} \\ \vdots \\ y_{iT} \end{pmatrix}_{T \times 1} \quad X_i = \begin{pmatrix} X_{i,1,1} & X_{i,1,2} & X_{i,1,j} & \dots & X_{i,1,K} \\ \vdots & \vdots & \vdots & & \vdots \\ X_{i,t,1} & X_{i,t,2} & X_{i,t,j} & \dots & X_{i,t,K} \\ \vdots & \vdots & \vdots & & \vdots \\ X_{i,T,1} & X_{i,T,2} & X_{i,T,j} & \dots & X_{i,T,K} \end{pmatrix}_{T \times K}$$

Panel with all units:

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_N \end{pmatrix}_{NT \times 1} \quad X = \begin{pmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_N \end{pmatrix}_{NT \times K}$$

Unobserved heterogeneity

- For a randomly drawn cross-sectional unit i , the model is given by

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}, \quad t = 1, 2, \dots, T$$

- y_{it} : log wages i in year t
- x_{it} : $1 \times K$ vector of variable events for person i in year t , such as education, marriage, etc. plus an intercept
- β : $K \times 1$ vector of marginal effects of events
- c_i : sum of all time-invariant inputs known to people i (but unobserved for the researcher), e.g., ability, beauty, grit, etc., often called unobserved heterogeneity or fixed effect
- ε_{it} : time-varying unobserved factors, such as a recession, unknown to the farmer at the time the decision on the events x_{it} are made, sometimes called idiosyncratic error

Fixed effect regression

- Our unobserved effects model is:

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}; t = 1, 2, \dots, T$$

- If we have data on multiple time periods, we can think of c_i as **fixed effects** to be estimated
- OLS estimation with fixed effects yields

$$(\hat{\beta}, \hat{c}_1, \dots, \hat{c}_N) = \underset{b, m_1, \dots, m_N}{\operatorname{argmin}} \sum_{i=1}^N \sum_{t=1}^T (y_{it} - x_{it}b - m_i)^2$$

this amounts to including N individual dummies in regression of y_{it} on x_{it}

Derivation: fixed effects regression

$$(\hat{\beta}, \hat{c}_1, \dots, \hat{c}_N) = \underset{b, m_1, \dots, m_N}{\operatorname{argmin}} \sum_{i=1}^N \sum_{t=1}^T (y_{it} - x_{it} b - m_i)^2$$

The first-order conditions (FOC) for this minimization problem are:

$$\sum_{i=1}^N \sum_{t=1}^T x'_{it} (y_{it} - x_{it} \hat{\beta} - \hat{c}_i) = 0$$

and

$$\sum_{t=1}^T (y_{it} - x_{it} \hat{\beta} - \hat{c}_i) = 0$$

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

Derivation: fixed effects regression

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

$$\hat{c}_i = \frac{1}{T} \sum_{t=1}^T (y_{it} - x_{it}\hat{\beta}) = \bar{y}_i - \bar{x}_i\hat{\beta},$$

where

$$\bar{x}_i \equiv \frac{1}{T} \sum_{t=1}^T x_{it}; \bar{y}_i \equiv \frac{1}{T} \sum_{t=1}^T y_{it}$$

Plug this result into the first FOC to obtain:

$$\hat{\beta} = \left(\sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)'(x_{it} - \bar{x}_i) \right)^{-1} \left(\sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)'(y_{it} - \bar{y}) \right)$$

$$\hat{\beta} = \left(\sum_{i=1}^N \sum_{t=1}^T \ddot{x}_{it}' \ddot{x}_{it} \right)^{-1} \left(\sum_{i=1}^N \sum_{t=1}^T \ddot{x}_{it}' \ddot{y}_{it} \right)$$

with time-demeaned variables $\ddot{x}_{it} \equiv x_{it} - \bar{x}$, $\ddot{y}_{it} \equiv y_{it} - \bar{y}$

Fixed effects regression

Running a regression with the time-demeaned variables
 $\ddot{y}_{it} \equiv y_{it} - \bar{y}_i$ and $\ddot{x}_{it} \equiv x_{it} - \bar{x}$ is numerically equivalent to a regression of y_{it} on x_{it} and unit specific dummy variables.

Even better, the regression with the time demeaned variables is consistent for β even when $Cov[x_{it}, c_i] \neq 0$ because time-demeaning eliminates the unobserved effects

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}$$

$$\bar{y}_i = \bar{x}_i\beta + c_i + \bar{\varepsilon}_i$$

$$(y_{it} - \bar{y}_i) = (x_{it} - \bar{x})\beta + (c_i - \bar{c}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

$$\ddot{y}_{it} = \ddot{x}_{it}\beta + \ddot{\varepsilon}_{it}$$

Fixed effects regression: main results

- Identification assumptions:

- ① $E[\varepsilon_{it}|x_{i1}, x_{i2}, \dots, x_{iT}, c_i] = 0; t = 1, 2, \dots, T$
 - regressors are strictly exogenous conditional on the unobserved effect
 - allows x_{it} to be arbitrarily related to c_i
- ② $\text{rank}\left(\sum_{t=1}^T E[\ddot{x}'_{it} \ddot{x}_{it}]\right) = K$
 - regressors vary over time for at least some i and not collinear

- Fixed effects estimator

- ① Demean and regress \ddot{y}_{it} on \ddot{x}_{it} (need to correct degrees of freedom)
- ② Regress y_{it} on x_{it} and unit dummies (dummy variable regression)
- ③ Regress y_{it} on x_{it} with canned fixed effects routine
 - Stata: `xtreg y x, fe i(PanelID)`

FE main results

- Properties (under assumptions 1-2):
 - $\widehat{\beta}_{FE}$ is consistent: $\plim_{N \rightarrow \infty} \widehat{\beta}_{FE,N} = \beta$
 - $\widehat{\beta}_{FE}$ is unbiased conditional on \mathbf{X}

Fixed effects regression: main issues

- Inference:
 - Standard errors have to be “clustered” by panel unit (e.g., farm) to allow correlation in the ε_{it} ’s for the same i .
 - Yields valid inference as long as number of clusters is reasonably large
- Typically we care about β , but unit fixed effects c_i could be of interest
 - \hat{c}_i from dummy variable regression is unbiased but not consistent for c_i (based on fixed T and $N \rightarrow \infty$)

Application: SASP

- From 2008-2009, I fielded a survey of Internet sex workers (685 respondents, 5% response rate)
- I asked two types of questions: static provider-specific information (e.g., age, weight) and dynamic session information over last 5 sessions
- Let's look at the panel aspect of this analysis together

Risk premium equation

$$\begin{aligned}Y_{is} &= \beta_i X_i + \delta D_{is} + \gamma_{is} Z_{is} + u_i + \varepsilon_{is} \\ \ddot{Y}_{is} &= \gamma_{is} \ddot{Z}_{is} + \ddot{\eta}_{is}\end{aligned}$$

where Y is log price, D is unprotected sex with a client in a session, X are client and session characteristics, Z is unobserved heterogeneity, and u_i is both unobserved and correlated with Z_{is} .

Table: POLS, FE and Demeaned OLS Estimates of the Determinants of Log Hourly Price for a Panel of Sex Workers

Depvar:	POLS	FE	Demeaned OLS
Unprotected sex with client of any kind	0.013 (0.028)	0.051* (0.028)	0.051* (0.026)
Ln(Length)	-0.308*** (0.028)	-0.435*** (0.024)	-0.435*** (0.019)
Client was a Regular	-0.047* (0.028)	-0.037** (0.019)	-0.037** (0.017)
Age of Client	-0.001 (0.009)	0.002 (0.007)	0.002 (0.006)
Age of Client Squared	0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
Client Attractiveness (Scale of 1 to 10)	0.020*** (0.007)	0.006 (0.006)	0.006 (0.005)
Second Provider Involved	0.055 (0.067)	0.113* (0.060)	0.113* (0.048)
Asian Client	-0.014 (0.049)	-0.010 (0.034)	-0.010 (0.030)
Black Client	0.092 (0.073)	0.027 (0.042)	0.027 (0.037)
Hispanic Client	0.052 (0.080)	-0.062 (0.052)	-0.062 (0.045)
Other Ethnicity Client	0.156** (0.068)	0.142*** (0.049)	0.142*** (0.045)
Met Client in Hotel	0.133*** (0.029)	0.052* (0.027)	0.052* (0.024)
Gave Client a Massage	-0.134*** (0.029)	-0.001 (0.028)	-0.001 (0.024)
Age of provider	0.003 (0.012)	0.000 (.)	0.000 (.)
Age of provider squared	-0.000 (0.000)	0.000 (.)	0.000 (.)

Table: POLS, FE and Demeaned OLS Estimates of the Determinants of Log Hourly Price for a Panel of Sex Workers

Depvar:	POLS	FE	Demeaned OLS
Body Mass Index	-0.022*** (0.002)	0.000 (.)	0.000 (.)
Hispanic	-0.226*** (0.082)	0.000 (.)	0.000 (.)
Black	0.028 (0.064)	0.000 (.)	0.000 (.)
Other	-0.112 (0.077)	0.000 (.)	0.000 (.)
Asian	0.086 (0.158)	0.000 (.)	0.000 (.)
Imputed Years of Schooling	0.020** (0.010)	0.000 (.)	0.000 (.)
Cohabitating (living with a partner) but unmarried	-0.054 (0.036)	0.000 (.)	0.000 (.)
Currently married and living with your spouse	0.005 (0.043)	0.000 (.)	0.000 (.)
Divorced and not remarried	-0.021 (0.038)	0.000 (.)	0.000 (.)
Married but not currently living with your spouse	-0.056 (0.059)	0.000 (.)	0.000 (.)
N	1,028	1,028	1,028
Mean of dependent variable	5.57	5.57	0.00

Heteroskedastic robust standard errors in parenthesis clustered at the provider level. * p<0.10,

** p<0.05, *** p<0.01

Unit specific time trends often eliminate “results”

Table: Demeaned OLS Estimates of the Determinants of Log Hourly Price for a Panel of Sex Workers with provider specific trends

Depvar:	FE w/provider trends
Unprotected sex with client of any kind	0.004 (0.046)
Ln(Length)	-0.450*** (0.020)
Client was a Regular	-0.071** (0.023)
Age of Client	0.008 (0.005)
Age of Client Squared	-0.000 (0.000)
Client Attractiveness (Scale of 1 to 10)	0.003 (0.003)
Second Provider Involved	0.126* (0.055)
Asian Client	-0.048*** (0.007)
Black Client	0.017 (0.043)
Hispanic Client	-0.015 (0.022)
Other Ethnicity Client	0.135*** (0.031)
Met Client in Hotel	0.073*** (0.019)
Gave Client a Massage	0.022 (0.012)

Concluding remarks

- This is not a review of panel econometrics; for that see Wooldridge and other excellent options
- We reviewed TWFE because it is commonly used with individual level panel data and difference-in-differences
- Their main value is how they control for unobserved heterogeneity through a simple demeaning
- What we will see in this seminar, though, is that strict exogeneity actually imposed not just parallel trends, but also treatment effect homogeneity under differential timing
- Now let's discuss difference-in-differences which will at various times use the TWFE model

John Snow and cholera

- John Snow, epidemiologist in 19th century, usually credited with first use of DiD
- Believed cholera was spread through the Thames water supply which contradicted dominant theory about “dirty air” transmission
- Grand experiment: Lambeth moves its pipe between 1849 and 1854; Southwark and Vauxhall delay
- How can he use this event to test his hypothesis? Three ways: simple comparisons, interrupted time series of the difference in differences (DiD)

Simple cross-sectional design

Table: Lambeth and Southwark and Vauxhall, 1854

Company	Cholera mortality
Lambeth	$Y = L + D$
Southwark and Vauxhall	$Y = SV$

Interrupted time series design

Table: Lambeth, 1849 and 1854

Company	Time	Cholera mortality
Lambeth	1849	$Y = L$
	1854	$Y = L + (T + D)$

Difference-in-differences

Table: Lambeth and Southwark and Vauxhall, 1849 and 1854

Companies	Time	Outcome	D_1	D_2
Lambeth	Before	$Y = L$	$T + D$	D
	After	$Y = L + T + D$		
Southwark and Vauxhall	Before	$Y = SV$	T	
	After	$Y = SV + T$		

Sample averages

$$\hat{\delta}_{kU}^{2 \times 2} = \left(\bar{y}_k^{post(k)} - \bar{y}_k^{pre(k)} \right) - \left(\bar{y}_U^{post(k)} - \bar{y}_U^{pre(k)} \right)$$

Population expectations

$$\widehat{\delta}_{kU}^{2\times 2} = \left(E[Y_k|Post] - E[Y_k|Pre] \right) - \left(E[Y_U|Post] - E[Y_U|Pre] \right)$$

Potential outcomes and the switching equation

$$\widehat{\delta}_{kU}^{2x2} = \underbrace{\left(E[Y_k^1|Post] - E[Y_k^0|Pre] \right) - \left(E[Y_U^0|Post] - E[Y_U^0|Pre] \right)}_{\text{Switching equation}} + \underbrace{E[Y_k^0|Post] - E[Y_k^0|Post]}_{\text{Adding zero}}$$

Parallel trends bias

$$\widehat{\delta}_{kU}^{2\times 2} = \underbrace{E[Y_k^1|Post] - E[Y_k^0|Post]}_{\text{ATT}} + \underbrace{\left[E[Y_k^0|Post] - E[Y_k^0|Pre] \right] - \left[E[Y_U^0|Post] - E[Y_U^0|Pre] \right]}_{\text{Non-parallel trends bias in } 2\times 2 \text{ case}}$$

Bias in our go-to estimators

- OLS will identify the ATT with only two groups and no covariates
- But the more common DD situation is one in which a treatment is adopted by different groups at different times
- OLS with panel and time fixed effects ("twoway fixed effects" or TWFE) is biased we now know
- I'll discuss the bias of TWFE, discuss a new solution, and fingers crossed a simulation if we have time

Regression DD

- There are several good reasons to use TWFE
 - It estimates the ATT under parallel trends
 - It's easy to calculate the standard errors
 - It's easy to include multiple periods
 - We can study treatments with different treatment intensity.
(e.g., varying increases in the minimum wage for different states)
- But there are bad reasons, too, which I'll discuss under differential timing and covariates

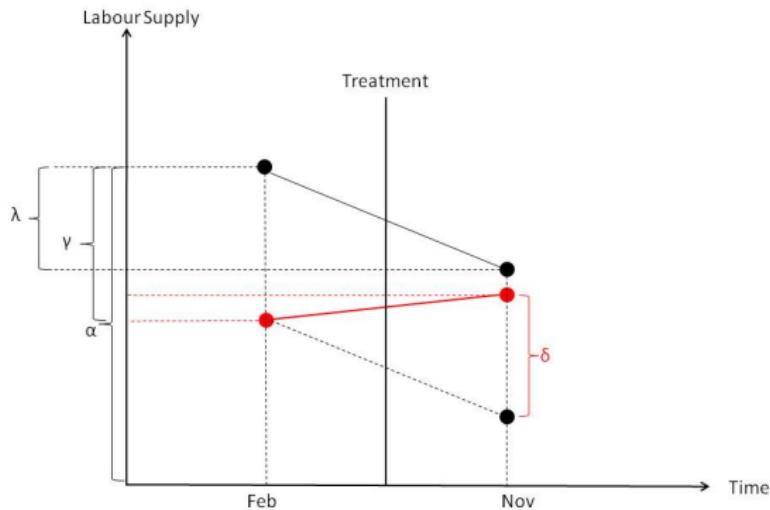
Regression DD - Card and Krueger

- The equivalent regression includes time and group fixed effects:

$$Y_{its} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{its}$$

- NJ is a dummy equal to 1 if the observation is from NJ
- d is a dummy equal to 1 if the observation is from November (the post period)
- This equation takes the following values
 - PA Pre: α
 - PA Post: $\alpha + \lambda$
 - NJ Pre: $\alpha + \gamma$
 - NJ Post: $\alpha + \gamma + \lambda + \delta$
- DD estimate: $(NJ \text{ Post} - NJ \text{ Pre}) - (PA \text{ Post} - PA \text{ Pre}) = \delta$

$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta (NJ \times d)_{st} + \varepsilon_{ist}$$



OLS with twoway fixed effects

Under parallel trends, OLS estimates the ATT for the two group case. Calculating standard errors is easy, multiple time periods is easy. But including covariates and time varying treatment (“differential timing”) will introduce problems.

Now on to some models focused on covariates

- We will discuss two papers now: Abadie (2005) and Sant'Anna and Zhao (2020)
- Abadie (2005) is really used best for longitudinal data or repeated cross sections where treatment occurs at one point in time
- But like other models we'll look into, Abadie (2005) modeled differential selection based on covariates

Overview

- Abadie (2005) proposed an alternative estimator to OLS to estimate the ATT
- The method is a DD type estimator, but isn't using TWFE
- You need treatment and comparison group, before and after treatment
- But you also need conditional parallel trends (based on X)

High level

Short and readable, though when it gets into theorems and proofs,
it's deep

"A good way to do econometrics is to look good for natural experiments and use statistical methods that can tidy up the confounding factors that nature has not controlled for us. – Daniel McFadden

Why do this?

- No randomization. Remember, DD doesn't require randomization – it requires a version of parallel trends
- Treatment is selecting on observable covariates

Why do this anyway?

- In a DD, we may need to control for X because treatment is only conditional on X
- But in TWFE, when you controlling for baseline X , it gets absorbed by the unit fixed effects
- One way around this is to use time-varying controls, but this places restrictions on the DGP as we will see in Sant'Anna and Zhou (2020)
- Abadie (2005) proposes weighting the difference in means using the propensity score estimated with series logit or linear probability models

Three step method

- ① Compute each unit's "after minus before" which is the DD part
- ② Then estimate a propensity score which you'll use to weight each unit
- ③ Finally, compare weighted changes in "after minus before" for treatment versus comparison groups

Inference will take into account step two, which is often the sticky part (see Abadie and Imbens matching paper which shows you can't use the bootstrap for matching, but you can for propensity scores)

You can have heterogeneous treatment effects, but not differential timing

Terms

- t is year of treatment which doesn't vary across units (so no differential timing)
- Y^1 and Y^0 are potential outcomes (counterfactual versus actual)
- D is 1 or 0 based on group and time
- b is the “baseline” which is similar to CS using g as the one year pre-treatment
- X are “baseline” covariates **only** – they do not vary over time, which means propensity scores are estimated off the b period **only**

Assumptions

Kind of common for this propensity score literature to only have two assumptions. But usually the first conditional independence. Now it is parallel trends because this is DD

- ① Conditional parallel trends

$$E[Y_t^0 - Y_b^0 | D = 1, X_b] - E[Y_t^0 - Y_t^0 | D = 0, X_b]$$

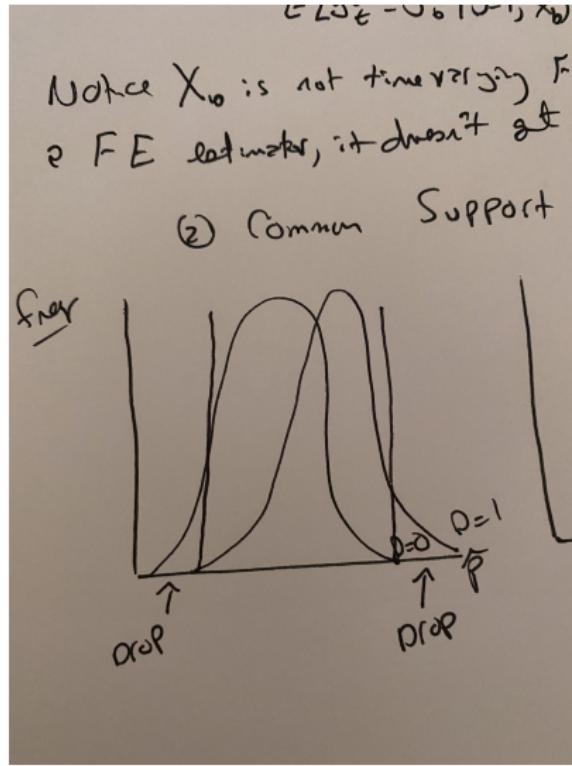
(Notice the b subscript. What is that you think?)

- ② Common support

$$Pr(D = 1) > 0; Pr(D = 1 | X) < 1$$

Let's see a picture of common support that I drew. Apologies it's horrible

Trimming the propensity score to get common support



Definition and estimation

Defining the ATT parameter of interest

$$ATT = E[Y_t^1 - Y_t^0 | D_t = 1] \quad (1)$$

Abadie's estimator

$$E\left[\frac{Y_t - Y_b}{Pr(D_t = 1)} \times \frac{D_t - Pr(D = 1|X_b)}{1 - Pr(D = 1|X_b)}\right] \quad (2)$$

These are also using the “Hajek” (non-normalized) weights from the inverse probability weighting literature

Propensity scores

- Usually there's almost no guidance that I've seen in how to estimate the propensity score except to say use logit or probit
- Dehejia and Wahba (2002) anyway
- Not so here – this is semi-parametric in the sense that you have to use a series of polynomials based on the X controls
- Weirdly, you can use OLS linear probability models (which I've never seen) or something called series logit estimation

Estimating propensity scores

It's common to hear people say that we don't know the propensity score; we can only estimate it. Same here – we approximate it with regressions

$$\widehat{Pr}(X_b) = \widehat{\gamma}_0 + \widehat{\gamma}_1 X + \widehat{\gamma}_2 X^2 + \dots \varepsilon \quad (3)$$

$$\widehat{Pr}(X_b) = F(\widehat{\gamma}_0 + \widehat{\gamma}_1 X + \widehat{\gamma}_2 X^2 + \dots) \quad (4)$$

Stata

Stata command is called `absdid`

You need treatment (varname), X variables (can be a list), the order in which the variables occur (weird, but results change if the order changes), and the exact estimator (LPM or logit)

Why not try it yourselves using the LaLonde NSW job trainings program data?

https://github.com/scunning1975/mixtape/raw/master/nsw_mixtape.dta

Concluding remarks

- LaLonde longitudinal data where you have a baseline and a follow-up
- Repeated cross-sections or panels
- Controls will cause the estimates to vary based on the type of approximation you use (logit for instance vs LPM) and the order in which the polynomials are used

Doubly Robust

- DR literature can be found in the older matching literature (Hirano and Imbens 2001; etc.)
- They combine regression and weighting estimators into one specification and are consistent so long as:
 - The regression specification for the outcome is correctly specified
 - The propensity score specification is correctly specified
- DR is a class of estimators that possess this property
- You're basically controlling for X twice: with a linear regression, with a propensity score, to cover your bases

DR DD

- Sant'Anna and Zhao (2020) incorporate DR into DD
- Think of it as a way of incorporating X into our new DD models (I'll show you why)
- It's in the engine of Callaway and Sant'Anna (2020) that we discuss later so we badly need to understand it
- Dense paper with hairy notation; I'll do my best

Literature

- Pedro is excellent at bridging gaps while simultaneously moving the ball forward – this is a good example
- The outcome regression part of DR goes back to Heckman, et al. (1997) and I use this in section 5.3.2 (“Bias correction”) of the mixtape
- The propensity score part goes back to Abadie (2005) which we’ve discussed
- New work on machine learning fits into this

Organization

- Basic assumptions for DD with covariates
- TWFE assumptions for DD with covariates
- Estimation alternative to TWFE with covariates
- Efficiency and semiparametric bounds

Insurance

- We covered covariates with Abadie (2005); why again?
- Maybe you're unsure whether the propensity score was properly specified
- How about some insurance?
- Two strikes instead of one

ATT

- DD *always* estimates the ATT because it's only the treatment effect for the treatment group in the post-treatment period
- It is not the ATE, or the LATE

$$\delta = E[Y_{it}^1 - Y_{it}^0 | D_i = 1]$$

Basic assumptions of DD

Assumption 1: Assume panel data or repeated cross-sectional data

Handling repeated cross-sectional data is hairy, and so I've chosen to focus on the panel data for this talk, but results are similar for repeated cross sections

Basic assumptions of DD

Assumption 2: Conditional parallel trends

If you were putting covariates into your DD regression, then you were assuming conditional parallel trends

$$E[Y_1^0 - Y_0^0 | X, D = 1] = E[Y_1^0 - Y_0^0 | X, D = 0]$$

Basic assumptions of DD

Assumption 3: Common support or overlap

For some $e > 0$, the probability of being in the treatment group is greater than e and the probability of being in the treatment group conditional on X is $\leq 1 - e$.

Intuition of assumption 3: Called overlap or common support.
Means there is at least a small fraction of the population that is treated and that for every value of the covariates X there is at least a small chance that the unit is not treated. It's called common support when it's a propensity score but it's just about the distribution of treatment and control across values of X .

Estimating DD with Assumptions 1-3

- Assumptions 1-3 gives us a couple of options of estimating the DD
- We can either use the outcome regression (OR) approach of Heckman, et al 1997
- Or we can use the propensity score approach of Abadie (2005)
- What about TWFE? Hold off on that question for a second until we look at the estimators based on Assumptions 1-3

Outcome regression

This is the Heckman, et al. (1997) approach where the outcome evolution is modeled with a regression

$$\hat{\delta}^{OR} = \bar{Y}_{1,1} - \left[\bar{Y}_{1,0} + \frac{1}{n^T} \sum_{i|D_i=1} (\hat{\mu}_{0,1}(X_i) - \hat{\mu}_{0,0}(X_i)) \right]$$

where \bar{Y} is the sample average of Y among units in the treatment group at time t and $\hat{\mu}(X)$ is an estimator of the true, but unknown, $m_{d,t}(X)$ which is by definition equal to $E[Y_t|D = d, X = x]$. See my Section 5.3.2 for more about this.

Inverse probability weighting

This is the Abadie (2005) approach where we use weighting

$$\hat{\delta}^{ipw} = \frac{1}{E_N[D]} E \left[\frac{D - \hat{p}(X)}{1 - \hat{p}(X)} (Y_1 - Y_0) \right]$$

where $\hat{p}(X)$ is an estimator for the true propensity score. Reduces the dimensionality of X into a single scalar.

Caveat

- Outcome regression needs $\hat{\mu}(X)$ to be correctly specified, whereas
- Inverse probability weighting needs $\hat{p}(X)$ to be correctly specified
- It's hard to "rank" these two in practice with regards to model misspecification because each is inconsistent when their own models are misspecified
- Well why don't we just use TWFE? I've never heard anyone complain about including covariates in TWFE and I've been doing it my entire adult life, so we're good right?
- Depends on if you want to assume three more things.
(Mixtape didn't know about this...)

TWFE

Here's the TWFE specification:

$$Y_{it} = \alpha_1 + \alpha_2 T_t + \alpha_3 D_i + \delta(T_i \times D_t) + \varepsilon_{it}$$

Just add in covariates then right?

$$Y_{it} = \alpha_1 + \alpha_2 T_t + \alpha_3 D_i + \delta(T_i \times D_t) + \theta \cdot X_{it} + \varepsilon_{it}$$

Sure! If you're willing to impose the next three assumptions (let's first look at estimators based on .

Decomposing TWFE with covariates

TWFE places restrictions on the DGP. Previous TWFE regression under assumptions 1-3 implies the following:

$$E[Y_1^1 | D = 1, X] = \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X$$

Conditional parallel trends implies

$$E[Y_1^0 - Y_0^0 | D = 1, X] = E[Y_1^0 - Y_0^0 | D = 0, X]$$

$$E[Y_1^0 | D = 1, X] - E[Y_0^0 | D = 1, X] = E[Y_1^0 | D = 0, X] - E[Y_0^0 | D = 0, X]$$

$$E[Y_1^0 | D = 1, X] = E[Y_0^0 | D = 1, X] + E[Y_1^0 | D = 0, X] - E[Y_0^0 | D = 0, X]$$

$$E[Y_1^0 | D = 1, X] = E[Y_0 | D = 1, X] + E[Y_1 | D = 0, X] - E[Y_0 | D = 0, X]$$

Last line from the switching equation. This gives us:

$$E[Y_1^0 | D = 1, X] = \alpha_1 + \alpha_2 + \alpha_3 + \theta X$$

Collecting terms

$$E[Y_1^1|D=1, X] = \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta_1 X$$

$$E[Y_1^0|D=1, X] = \alpha_1 + \alpha_2 + \alpha_3 + \theta_2 X$$

$$E[Y_1^1|D=1, X] - E[Y_1^0|D=1, X]$$

$$= (\alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta_1 X) - (\alpha_1 + \alpha_2 + \alpha_3 + \theta_2 X)$$

$$= \delta + (\theta_1 X - \theta_2 X)$$

By allowing for the possibility that $\theta_1 X \neq \theta_2 X$, we open up the possibility of bias from TWFE which is zero under three additional assumptions.

Assumption 4

The implications of that TWFE regression with assumptions 1-3 gave us those previous expressions which then require placing further restrictions on treatment effects and trends when estimating with TWFE.

TWFE Assumption 4: Homogenous treatment effects in X

$$E[Y_1^1 - Y_1^0 | X, D = 1] = E[Y_1^1 - Y_1^0 | D = 1]$$

This is because when you difference out those previous equations, you need θX to cancel to leave you with δ which implies homogeneity in X .

X-specific trends

TWFE places restrictions on trends for the two groups too. Take conditional expectations of our TWFE equation.

$$E[Y_1|D=1] = \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X_{11}$$

$$E[Y_0|D=1] = \alpha_1 + \alpha_3 + \theta X_{10}$$

$$E[Y_1|D=0] = \alpha_1 + \alpha_2 + \theta X_{01}$$

$$E[Y_0|D=0] = \alpha_1 + \theta X_{00}$$

X-specific trends

Now take DD:

$$\delta^{DD} = \left((\alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X_{11}) - (\alpha_1 + \alpha_3 + \theta X_{10}) \right) - \left((\alpha_1 + \alpha_2 + \theta X_{01}) - (\alpha_1 + \theta X_{00}) \right)$$

Eliminating terms, we get:

$$\delta^{DD} = \delta + (\theta X_{11} - \theta X_{10}) - (\theta X_{01} - \theta X_{00})$$

Second line requires that trends in X for treatment group equal trends in X for control group.

Assumption 5 and 6

For $D = 0, 1$, we need “no X-specific trends in both groups”:

$$E[Y_1 - Y_0 | D = d, X] = E[Y_1 - Y_0 | D = d]$$

Intuition: Sant'Anna and Zhao (2020) say in footnote 4 “[this] follows from analogous arguments” which is the previous slides’ manipulation of terms. Key is to remember these are time-varying covariates so they don’t cancel out within treatment category, so you need the trends in X to cancel out.

Without these six, in general TWFE will not identify ATT. Unclear how off it’ll be, but it will be biased is the point.

Why not both?

- Let's review the problem. What if you claim you need X for conditional parallel trends?
- You have three options:
 - Outcome regression (Heckman, et al. 1997) – needs Assumptions 1-3
 - Inverse probability weighting (Abadie 2005) – needs Assumptions 1-3
 - TWFE (everybody everywhere all the time) – needs Assumptions 1-6
- Problem is 1 and 2 need the models to be correctly specified
- Doubly robust combines them to give us insurance. That's the basic idea. Gives you two chances to be wrong

Next step

- Introduction to the three prior covariate DD models
- Assumptions – check
- Hints about combining OR and IPW
- Now we move into *estimation* phase
- Let's see what doubly robust estimator looks like
- As before, I'm going to only stick to the panel data expressions
bc all repeated cross-section does is add in some terms

Estimation

Some terms

$p(x)$: propensity score model

$\Delta Y = Y_1 - Y_0 = Y_{post} - Y_{pre}$

$\mu_{d,\Delta} = \mu_{d,1}(X) - \mu_{d,0}$, where $\mu(X)$ is a model for

$m_{d,t} = E[Y_t | D = d, X = x]$

So that means $\mu_{1,\Delta}$ is just the treatment group's change in average Y for each $X = x$

We're off to see the (DR) wizard!

Population DR DD model for panel data

$$\delta^{dr} = E \left[\left(\frac{D}{E[D]} - \frac{\frac{p(X)(1-D)}{(1-p(X))}}{E \left[\frac{p(X)(1-D)}{(1-p(X))} \right]} \right) (\Delta Y - \mu_{0,\Delta}(X)) \right]$$

Notice the propensity score modifying the control group (second term inside parentheses) *and* the $\mu(X)$ term modifying the long difference. This is the idea of the doubly robust – you only need one of these models to be correctly specified, not both.

Sidebar: This is also one of the options in the Callaway and Sant'Anna (2020) DD estimator. It lets you pick IPW, regression (OR) or DR. Pedro usually recommends DR because of its advantages.

Efficiency

- Last step is inference
- Authors exploit all the restrictions implied by the assumptions to construct semiparametric bounds
- This is where the influence function comes in, which those who have studied the DID code closely may have noticed
- One of the main results of the paper is that the DR DID estimator is also DR for inference
- Let's skip to Monte Carlos

Monte Carlo details

- Compare DR with TWFE, OR and IPW
- Sample size is 1,000
- 10,000 Monte Carlo experiments
- Propensity score estimated with logit; OR estimated using linear specification

Table: Monte Carlo Simulations, DGP1, Both OR and Propensity score correct

	Bias	RMSE	SE	Coverage	CI length
TWFE	-20.9518	21.1227	2.5271	0.000	9.9061
OR	-0.0012	0.1005	0.1010	0.9500	0.3960
IPW	0.0.257	2.7743	2.6636	0.9518	10.4412
DR	-0.0014	0.1059	0.1052	0.9473	0.4124

Figure 1: Monte Carlo for DID estimators, DGP1: Both pscore and OR are correctly specified

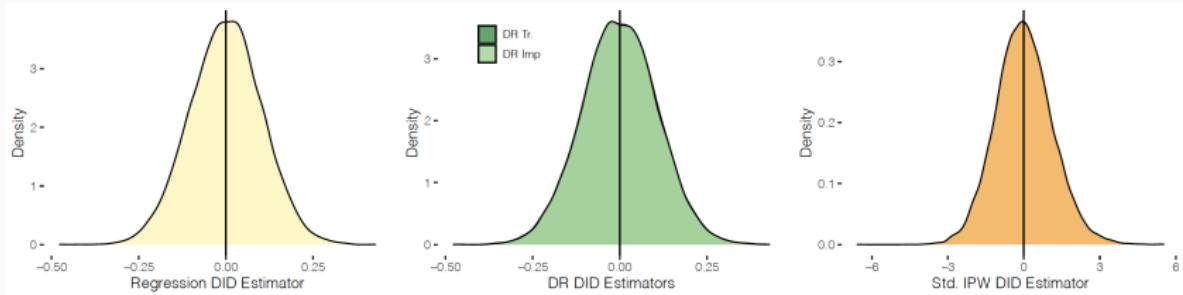
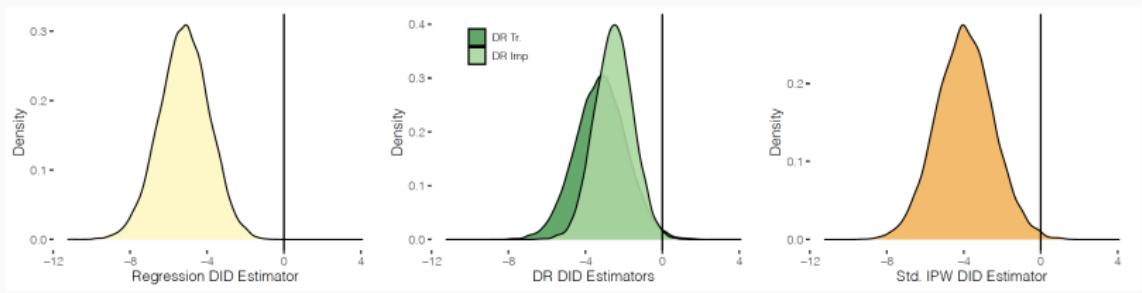


Table: Monte Carlo Simulations, DGP4, Neither OR and Propensity score correct

	Bias	RMSE	SE	Coverage	CI length
TWFE	-16.3846	16.5383	3.6268	0.000	14.2169
OR	-5.2045	5.3641	1.2890	0.0145	5.0531
IPW	-1.0846	2.6557	2.3746	0.9487	9.3084
DR	-3.1878	3.4544	1.2946	0.3076	5.0749

Figure 4: Monte Carlo for DID estimators, DGP4: Both OR and PS are misspecified



To a kid with a hammer, everything is a nail

- Use the right tool (oven) for the job (making lasagna), not the same tool (hammer) regardless of the job (making lasagna)
- One of the main things I learned from this paper was again biases in TWFE with covariates – Mixtape and MHE don't cover this
- This method only needed three assumptions not the six for TWFE
- Like everything Pedro does, there is code for this but it's only in R – DRDID
- But it's one of the main options in Callaway and Sant'anna under differential timing, and therefore it's crucial we understand this
- But you still have to have specified correctly either at least the outcome model or propensity score model



The Journal of Human Resources

[Home Page](#) | [Current Issue](#) | [Archive](#) | [Subscribe](#) | [Alerts](#) | [Customer Service](#) | [Feedback](#)

Institution: Baylor University Libraries

Does Strengthening Self-Defense Law Deter Crime or Escalate Violence? Evidence from Expansions to Castle Doctrine

Cheng Cheng

Mark Hoekstra

Abstract

From 2000 to 2010, more than 20 states passed so-called "Castle Doctrine" or "stand your ground" laws. These laws expand the legal justification for the use of lethal force

Summary

- Cheng and Hoekstra (2013) are interested in whether expansions to “castle doctrine statutes” at the state level increase or decrease gun violence.
- Prior to these expansions, English common law principle required “duty to retreat” before using lethal force against an assailant except when the assailant is an intruder in the home
 - The home is one’s “castle” – hence, “castle doctrine”
 - When intruders threatened the victim in the home, the duty to retreat was waived and lethal force in self-defense was allowed

Castle doctrine law explained

- In 2005, Florida passed a law that expanded self-defense protections beyond the house
 - 2000 to 2010, 21 states explicitly put “castle doctrine” into statute, and (more importantly) extended it to places outside the home
 - In other words, 21 states removed the duty to retreat in specified circumstances
- Other changes:
 - Presumption of reasonable fear is added
 - Civil liability for those acting under the law is removed

Economic theory predicts more lethal homicides

- Workers supply legal or illegal labor and are therefore responsive to costs and benefits
- Castle doctrine expansions lowered the (expected) cost of killing someone in self-defense
- If people are rational, then lowering the price of lethal self-defense should increase lethal homicides

Economic theory also predicts less crime from deterrence

- Although deterrence is a theoretical possibility, note that the goal of the laws was to protect enhance victim rights, not deter crime
- Testable prediction with data and same design

Treatment passage

- Summary:

- 21 states passed laws removing “duty to retreat” in places outside the home
- 17 states removed “duty to retreat” in any place one had a legal right to be
- 13 states include a presumption of reasonable fear
- 18 states remove civil liability when force was justified under law

Cheng and Hoekstra's identification strategy

- Panel fixed effects estimation

$$Y_{it} = \beta_1 D_i + \beta_2 T_t + \beta_3(CDL_{it}) + \alpha_1 X_{it} + c_i + u_t + \varepsilon_{it}$$

- CDL is a fraction between 0 and 1 depending on the percent of the year the state has a castle doctrine law
- Preferred specifications includes “region-by-year fixed effects”

Data

- FBI Uniform Crime Reports Part 1 Offenses (2000-2010)
 - State-level crime rates, or “offenses per 100,000 population”
 - Falsification outcomes: motor vehicle theft and larceny
- Dataset on justifiable homicides by private citizens

Outcomes (in order)

- Deterrence and homicide outcomes:
 - ① Burglary: the unlawful entry of a structure to commit a felony or a theft
 - ② Robbery: the taking or attempting to take anything of value from the care, custody or control of a person or persons by force or threat of force or violence and/or putting the victim in fear
 - ③ Aggravated assault: unlawful attack by one person upon another for the purpose of inflicting severe or aggravated bodily injury
- Homicide categories
 - ① Total homicides – murder plus non-negligent manslaughter (~14,000 per year)
 - ② Justifiable homicides by private citizens (~250/year)

Inference: Clustering

- Statistical inference: cluster standard errors at the state level
 - Are disturbances random draws from individually identical distribution?
 - It's likely that within a state, unobserved determinants of crime are serially correlated
 - They follow Bertrand, Duflo and Mullainathan (2004) and adjust for serial correlation in unobserved disturbances within states at the level of the treatment

Inference: Fisher's sharp null

- How likely is it that we estimate effects of this magnitude when using randomly chosen pre-treatment time periods and randomly assigning placebo treatments?
- Randomizes dates within-state for the pre-treatment period (<2000)
- Randomization inference and exact p-values

Region-by-year fixed effects

- Absent passing castle doctrine laws, outcomes in these 21 states would have changed similar to other states in their same region
 - Recall the “region-by-year fixed effects” in the X term
 - By including “region-by-year fixed effects”, they are arguing that unobserved changes in crime are running “parallel” to the treatment states within region over time
 - Need not hold across regions since the across region variation is not being used in this analysis due to the saturation of the model with “region-by-year fixed effects”

State specific time trends

- Alabama, et al. dummy interacted with TREND which equals 1 in 2000, 2 in 2001, ..., 11 in 2010
- Forces the identification to come from variation in outcomes around the state-specific linear trend
 - Outcomes must be large enough and different enough from a state-specific linear trend otherwise it is collinear with the state-trend
 - Same argument applies to any control though
 - Goodman-Bacon (2019) suggests group-trends are less taxing and satisfying than unit-specific trends

Control variables

- Controls (X matrix in earlier equation)
 - Full-time police employment per 100,000 state residents from the LEKOA data (FBI data)
 - Persons incarcerated in state prison per 100,000 residents
 - Shares of white/black men in 15-24 and 25-44 age groups
 - State per capita spending on public assistance
 - State per capita spending on public welfare

Parallel Leads

- Look at each set of treatment states against never-treated figure by figure (rare)
- Use a one-period lead in the regression model (not as common)
- I'm going to look at event study coefficients (most common)

Step one: Falsification test

- Policy-makers are not just randomly flipping coins when passing laws, but presumably do so because of things they observe on the ground
- Address concerns up front this isn't driven by spurious crime results
- Cheng and Hoekstra (2013) present falsification of larceny and motor vehicle theft first, then results

Step one (cont.)

- Results will be presented separately under six different specifications
 - Each new specification adds more controls
- Pop quiz: What should you expect to find on key variables of interest when conducting a falsification and why?

Answer

- No statistically significant association between the CDL passage and the placebos; preferably precise zeroes
- No association on the one-year lead either
- Basically, you should not find effects where there are no theoretical policy effects; gun laws shouldn't affect non-violent offenses

Step one (cont.)

- How do you interpret coefficients?
 - His model is “log outcomes” regressed onto a dummy variable (level), so these are semi-elasticities and approximate percentage changes – but you should transform them by taking the exponential of each coefficient and then differencing it from one to find the actual percentage change
 - Ex: CDL = -0.0137 (column 12, Table 3, “Log (larceny rate)” outcome.) $\text{Exp}(-0.0137) = 0.986$, and so $1-0.986 = 1.4$. Thus, CDL reduced larceny rates by 1.4 percent, which is not statistically significant.

Results – Falsification Exercise

Table 3: Placebo Tests

	OLS - Unweighted					
	7	8	9	10	11	12
Panel A: Larceny						
Castle Doctrine Law	0.00745 (0.0227)	0.00145 (0.0205)	-0.00188 (0.0210)	-0.00445 (0.0226)	-0.00361 (0.0201)	-0.0137 (0.0228)
One Year Before Adoption of Castle Doctrine Law					-0.0103 (0.0114)	
Observation	550	550	550	550	550	550
Panel B: Motor Vehicle Theft						
			Log (Motor Vehicle Theft Rate)			
Castle Doctrine Law	0.0767* (0.0413)	0.0138 (0.0444)	0.00814 (0.0407)	0.00775 (0.0462)	0.00977 (0.0391)	-0.00373 (0.0361)
One Year Before Adoption of Castle Doctrine Law				-0.00155 (0.0287)		
Observation	550	550	550	550	550	550
State and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Region-by-Year Fixed Effects		Yes	Yes	Yes	Yes	Yes
Time-Varying Controls			Yes	Yes	Yes	Yes
Controls for Larceny or Motor Theft					Yes	
State-Specific Linear Time Trends						Yes

Notes: Each column in each panel represents a separate regression. The unit of observation is state-year. Robust standard errors are clustered at the state level. Time-varying controls include policing and incarceration rates, welfare and public assistance spending, median income, poverty rate, unemployment rate, and demographics.

Step two: testing the deterrence hypothesis

- Having found no effect on their placebos, Cheng and Hoekstra (2013) examine the effect of CDL on three deterrence outcomes: burglary, robbery and aggravated assault
 - They will, again, have six specifications per outcome in the “weighted” regression, and then another five for the “unweighted” regression
- Pop quiz: What does deterrence look like?

Answer

- Negative signs on the CDL variable is consistent with deterrence – these crimes were “deterred”, in other words
- Based on early work by Becker (1968) and 1970s work by his student Isaac Ehrlich; higher probabilities of getting hurt in public may cause offenders to avoid violence in public altogether
- Bounds on the magnitudes from the standard errors are used to provide some confidence about the estimates as well

Results – Deterrence

	OLS - Weighted by State Population						OLS - Unweighted					
	1	2	3	4	5	6	7	8	9	10	11	12
Panel A: Burglary												
Castle Doctrine Law							Log (Burglary Rate)					
	0.0780***	0.0290	0.0223	0.0164	0.0327*	0.0237		0.0572**	0.00961	0.00663	0.00277	0.00683
	(0.0255)	(0.0236)	(0.0223)	(0.0247)	(0.0165)	(0.0207)		(0.0272)	(0.0291)	(0.0268)	(0.0304)	(0.0222)
One Year Before Adoption of												
Castle Doctrine Law							-0.0201					
							(0.0139)					
Panel B: Robbery												
Castle Doctrine Law							Log (Robbery Rate)					
	0.0408	0.0344	0.0262	0.0216	0.0376**	0.0515*		0.0448	0.0320	0.00839	0.00552	0.00874
	(0.0254)	(0.0224)	(0.0229)	(0.0246)	(0.0181)	(0.0274)		(0.0331)	(0.0421)	(0.0387)	(0.0437)	(0.0339)
One Year Before Adoption of												
Castle Doctrine Law							-0.0156					
							(0.0167)					
Panel C: Aggravated Assault												
Castle Doctrine Law							Log (Aggravated Assault Rate)					
	0.0434	0.0397	0.0372	0.0362	0.0424	0.0414		0.0555	0.0698	0.0343	0.0305	0.0341
	(0.0387)	(0.0407)	(0.0319)	(0.0349)	(0.0291)	(0.0285)		(0.0604)	(0.0630)	(0.0433)	(0.0478)	(0.0405)
One Year Before Adoption of												
Castle Doctrine Law							-0.00343					
							(0.0161)					
Observations	550	550	550	550	550	550		550	550	550	550	550
State and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes		Yes	Yes	Yes	Yes	Yes
Region-by-Year Fixed Effects		Yes	Yes	Yes	Yes	Yes			Yes	Yes	Yes	Yes
Time-Varying Controls			Yes	Yes	Yes	Yes			Yes	Yes	Yes	Yes
Contemporaneous Crime Rates					Yes						Yes	
State-Specific Linear Time Trends						Yes						Yes

Conclusion

- “In short, these estimates provide strong evidence against the possibility that castle doctrine laws cause economically meaningful deterrence effects” (p. 17)
 - Translation: They can't find evidence of large deterrence effects
- “Thus, while castle doctrine law may well have benefits to those legally justified in protecting themselves in self-defense, there is no evidence that the law provides positive spillovers by deterring crime more generally” (p. 17)
 - They note in footnote 24 that they cannot measure the benefits to victims whose crimes were deterred, or the benefits from lower legal costs; their focus is limited to whether it deterred the crimes, not whether the net benefits from the laws were positive
 - Obviously, if there is no deterrence, though, then the net benefits are lower from CDL than they would be if they did deter

Step 3: Homicides

- The key finding in this study focuses on CDL and its effect on homicides and non-negligent manslaughter
- Pop quiz: what should the sign on CDL be here?

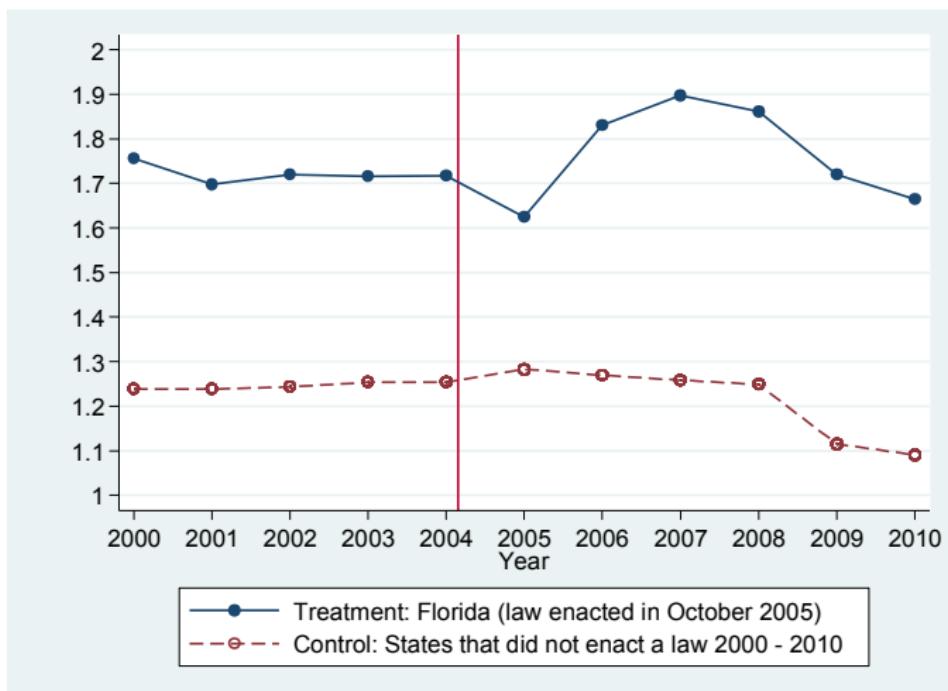
Answer

- Effects should be **positive**
- Cheng and Hoekstra want to show the raw data, but have differential timing
- Differential timing means you can't show pre-treatment raw data for the never-treated groups
- So they show it one by one – which isn't the most aesthetically pleasing way to do it, but which has the benefit of being transparent

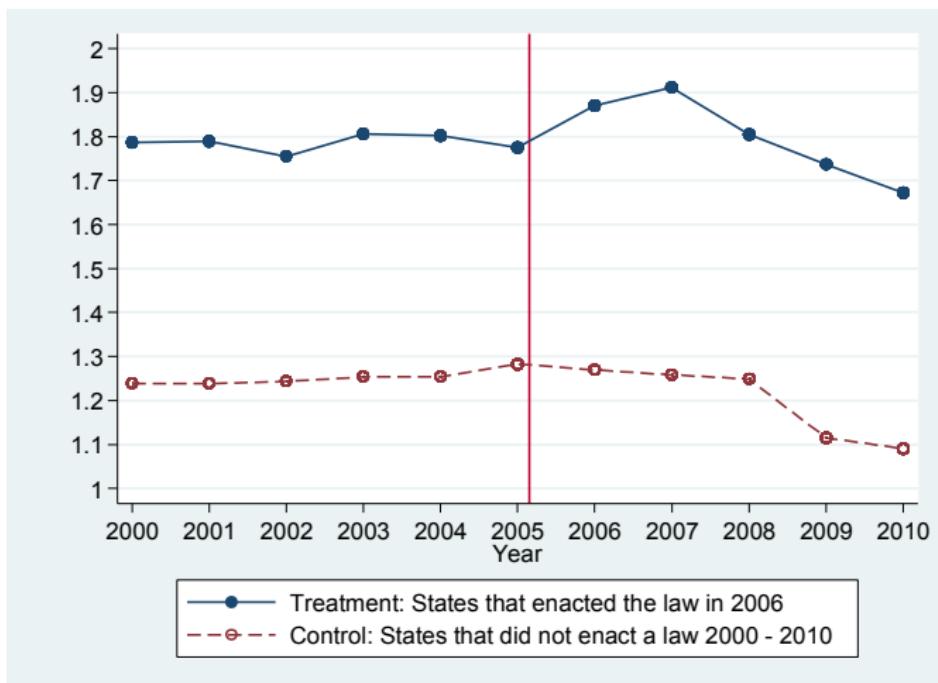
Parallel pre-treatment trends

- Keep your eyes on whether pre-treatment trends are parallel for treatment and control groups
 - Remember, though – he needs parallel trends within-region – these figures don't show that
 - But starting with pictures and raw data has value

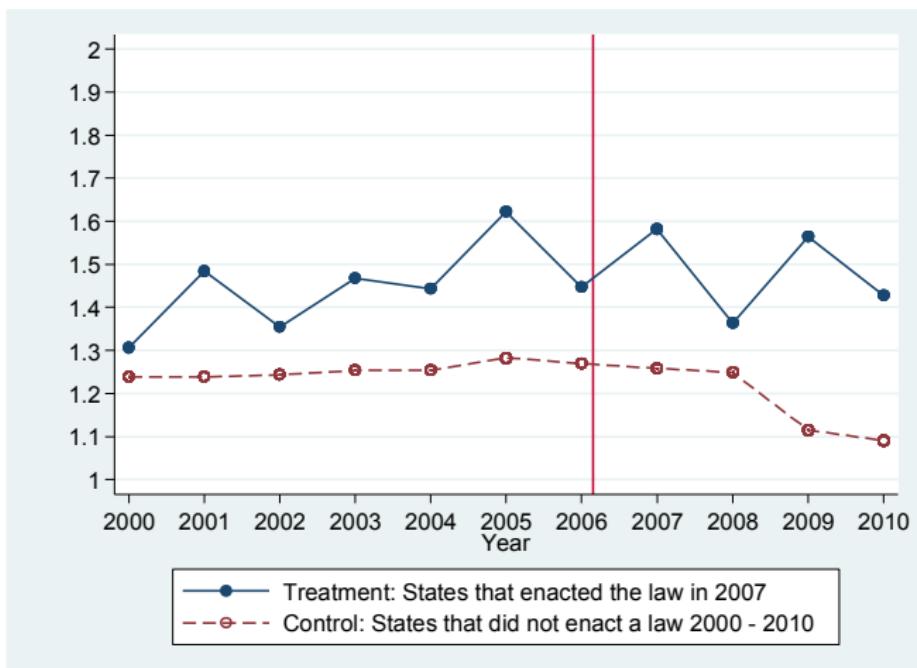
Log Homicide Rates – 2005 Adopter = Florida



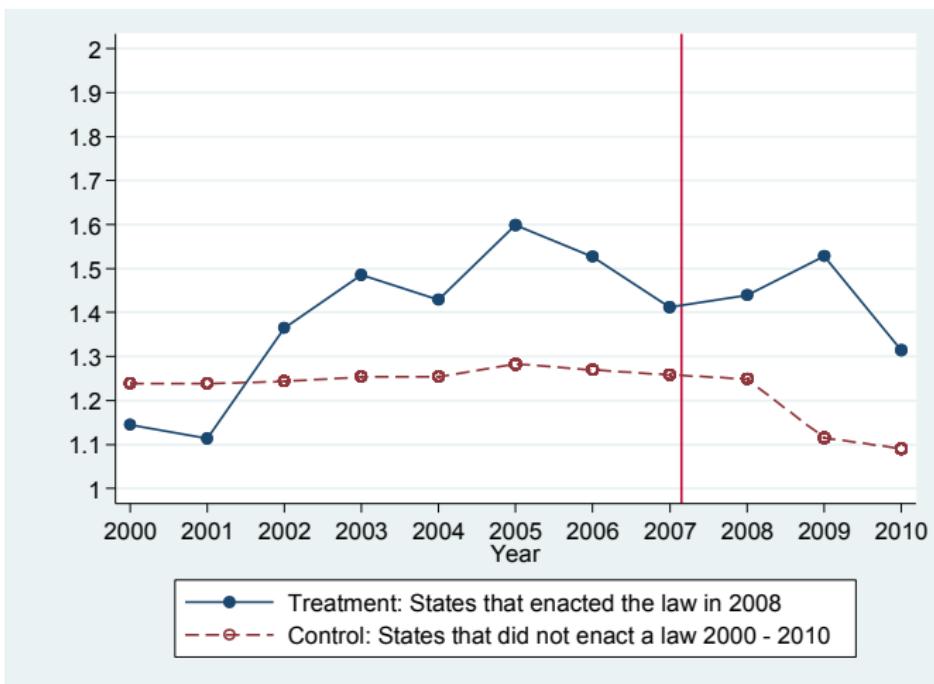
Log Homicide Rates – 2006 Adopter (13 states)



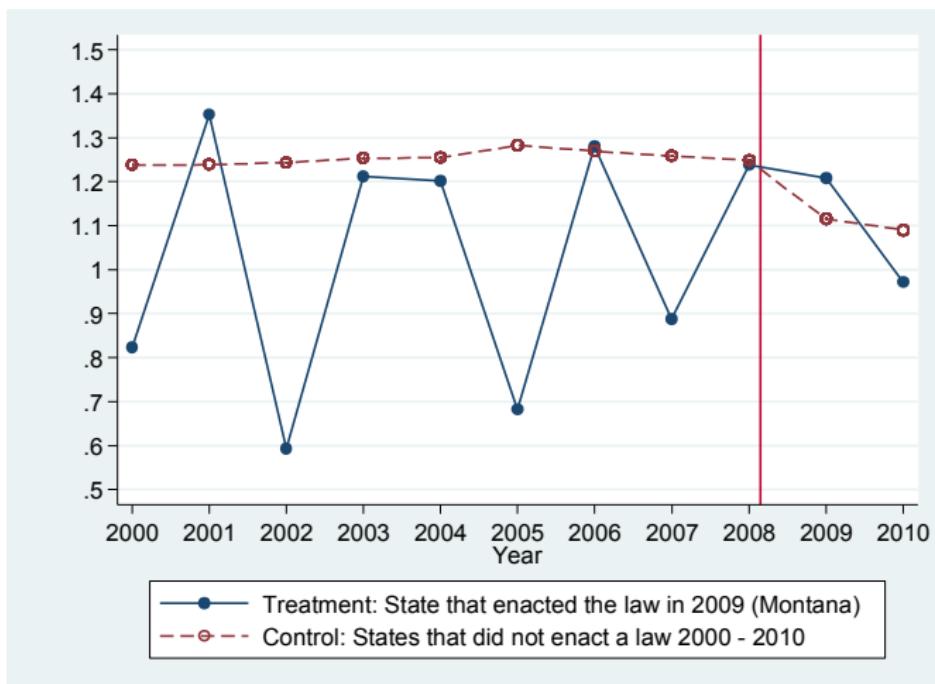
Log Homicide Rates – 2007 Adopter (4 states)



Log Homicide Rates – 2008 Adopter (2 states)



Log Homicide Rates – 2009 Adopter = Montana



Modeling

- He uses a class of estimators more appropriate for “counts” called “count models”, like the negative binomial estimated with maximum likelihood
- Results are robust to least squares and count models

Homicide – Negative Binomial; Murder – OLS

	1	2	3	4	5	6
<u>Panel C: Homicide (Negative Binomial - Unweighted)</u>						
Castle Doctrine Law	0.0565*	0.0734**	0.0879***	0.0783**	0.0937***	0.108***
	(0.0331)	(0.0305)	(0.0313)	(0.0355)	(0.0302)	(0.0346)
One Year Before Adoption of Castle Doctrine Law				-0.0352		
				(0.0260)		
Observations	550	550	550	550	550	550
<u>Panel D: Log Murder Rate (OLS - Weighted)</u>						
Castle Doctrine Law	0.0906**	0.0955**	0.0916**	0.0884**	0.0981**	0.0813
	(0.0424)	(0.0389)	(0.0382)	(0.0404)	(0.0391)	(0.0520)
One Year Before Adoption of Castle Doctrine Law				-0.0110		
				(0.0230)		
Observations	550	550	550	550	550	550
State and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Region-by-Year Fixed Effects		Yes	Yes	Yes	Yes	Yes
Time-Varying Controls			Yes	Yes	Yes	Yes
Contemporaneous Crime Rates					Yes	
State-Specific Linear Time Trends						Yes

Fisher sharp null

Move the 11-year panel back one year at a time (covering 1960-2009) and estimate 40 placebo “effects” of passing CDL 1 to 40 years earlier

Method	Average estimate	Estimates larger than actual estimate
Weighted OLS	-0.003	0/40
Unweighted OLS	0.001	1/40
Negative binomial	0.001	0/40

My replication using event study plots

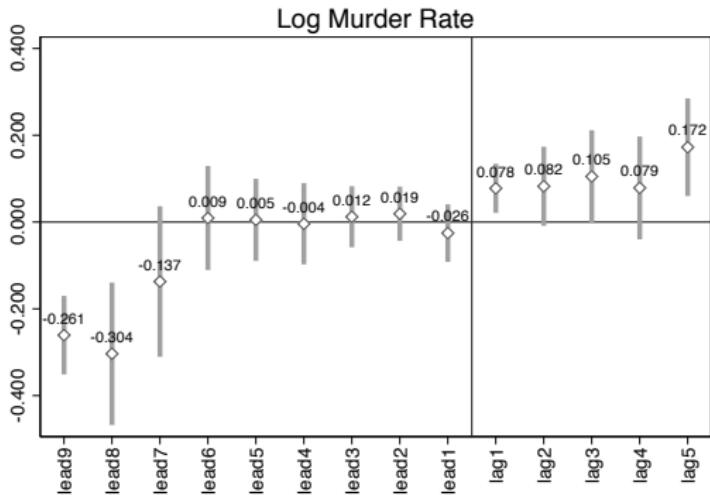


Figure: Homicide event study plots using coefplot

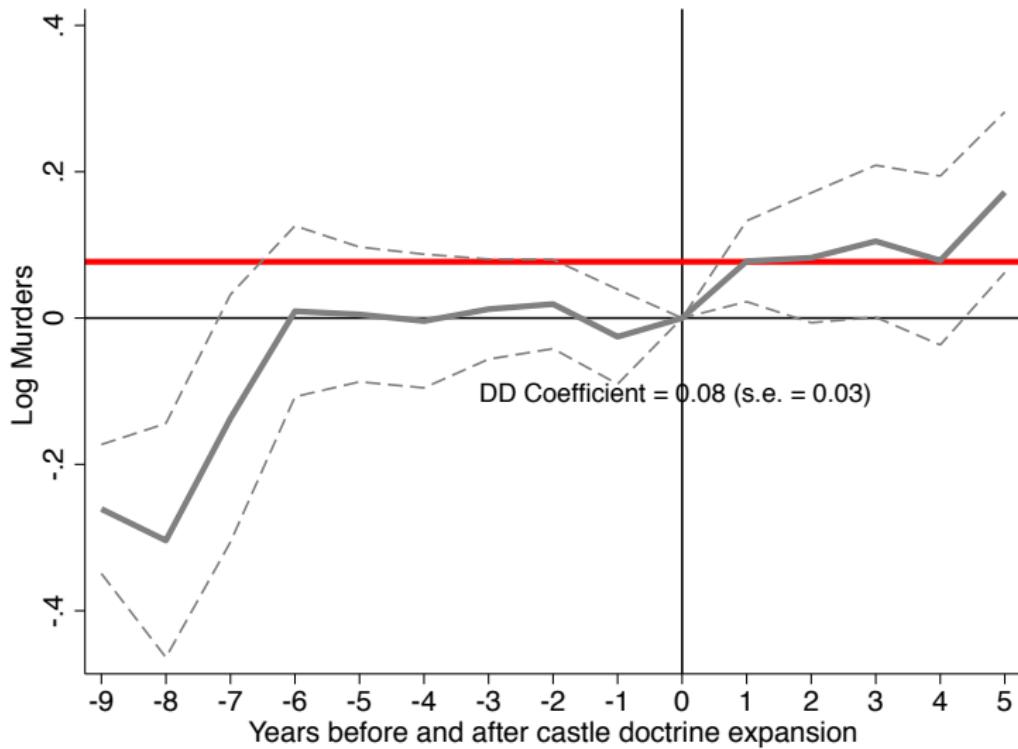


Figure: Homicide event study plots using two way

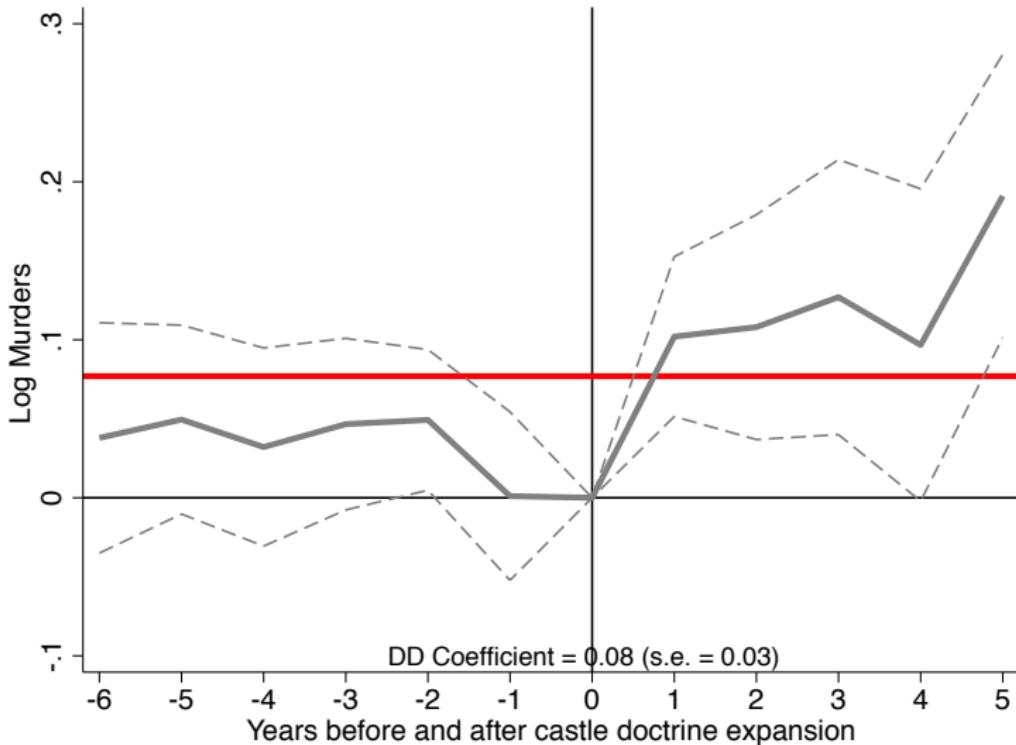


Figure: Homicide event study plots using two way and force early leads into one coefficient

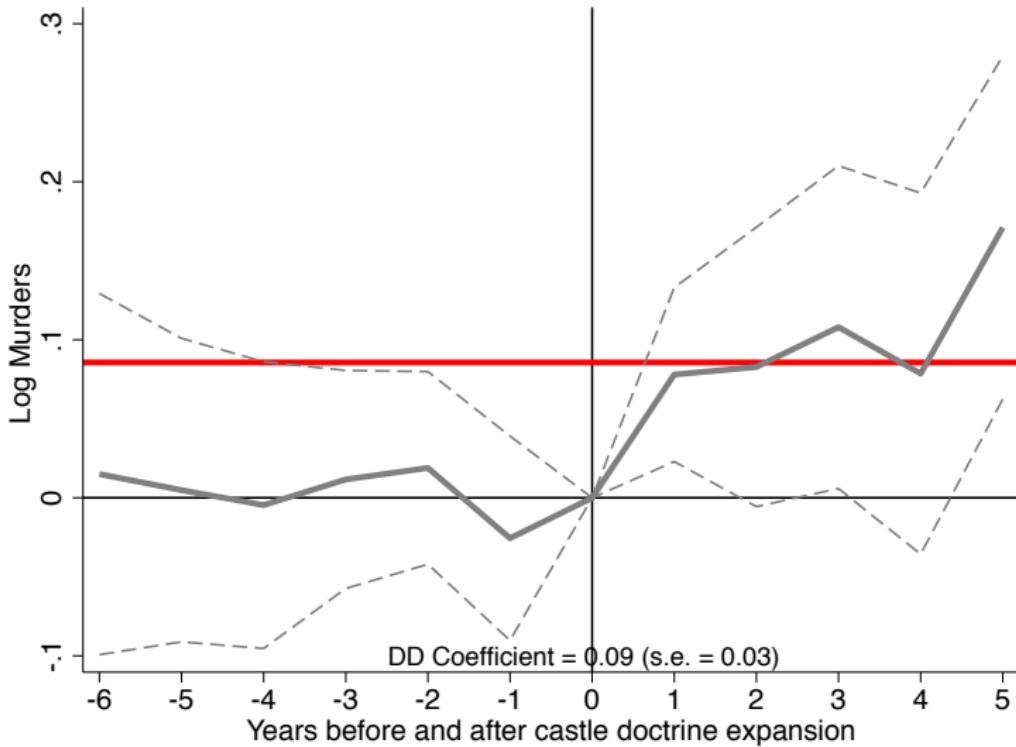


Figure: Homicide event study plots using two-way dropping imbalanced states

Interpretation

- No evidence that Castle Doctrine/Stand Your Ground Laws deter violent crimes such as burglary, robbery and aggravated assault
- These laws do lead to an 8% net increase in homicide rates, translating to around 600 additional homicides *per year* across the 21 adopting states
- Unlikely that all of the additional homicides were legally justified
- Incentives matter in some contexts (lethal force) but not others (deterrence)

Differential timing

- We covered mostly the simple two group case
- In the two group case, we can estimate the ATT under parallel trends using OLS with unit and time fixed effects
- If we have covariates, then we can use TWFE under restrictive assumptions, or we have other options (OR, IPW, DR)
- Now let's move to a more common scenario where we have more than two groups who get treated at various times

2x2 versus differential timing

- For this next part, similar to how we did with Sant'Anna and Zhou (2020), we will decompose TWFE to understand what it needs for unbiasedness under differential timing
- All of this is from Goodman-Bacon (2021, forthcoming) though the expression of the weights is from 2018 for personal preference
- Goodman-Bacon (2021, forthcoming) shows that parallel trends is **not enough** for TWFE to be unbiased when treatment adoption is described by differential timing
- TWFE with differential timing uses treated groups as controls – not all estimators do – and this can introduce bias

Decomposition Preview

- TWFE estimates a parameter that is a weighted average over all 2x2 in your sample
- TWFE assigns weights that are a function of sample sizes of each “group” and the variance of the treatment dummies for those groups

Decomposition (cont.)

- TWFE needs two assumptions: that the variance weighted parallel trends are zero (far more parallel trends iow) and no dynamic treatment effects (not the case with 2x2)
- Under those assumptions, TWFE estimator estimates the variance weighted ATT as a weighted average of all possible ATTs

Terms and notation

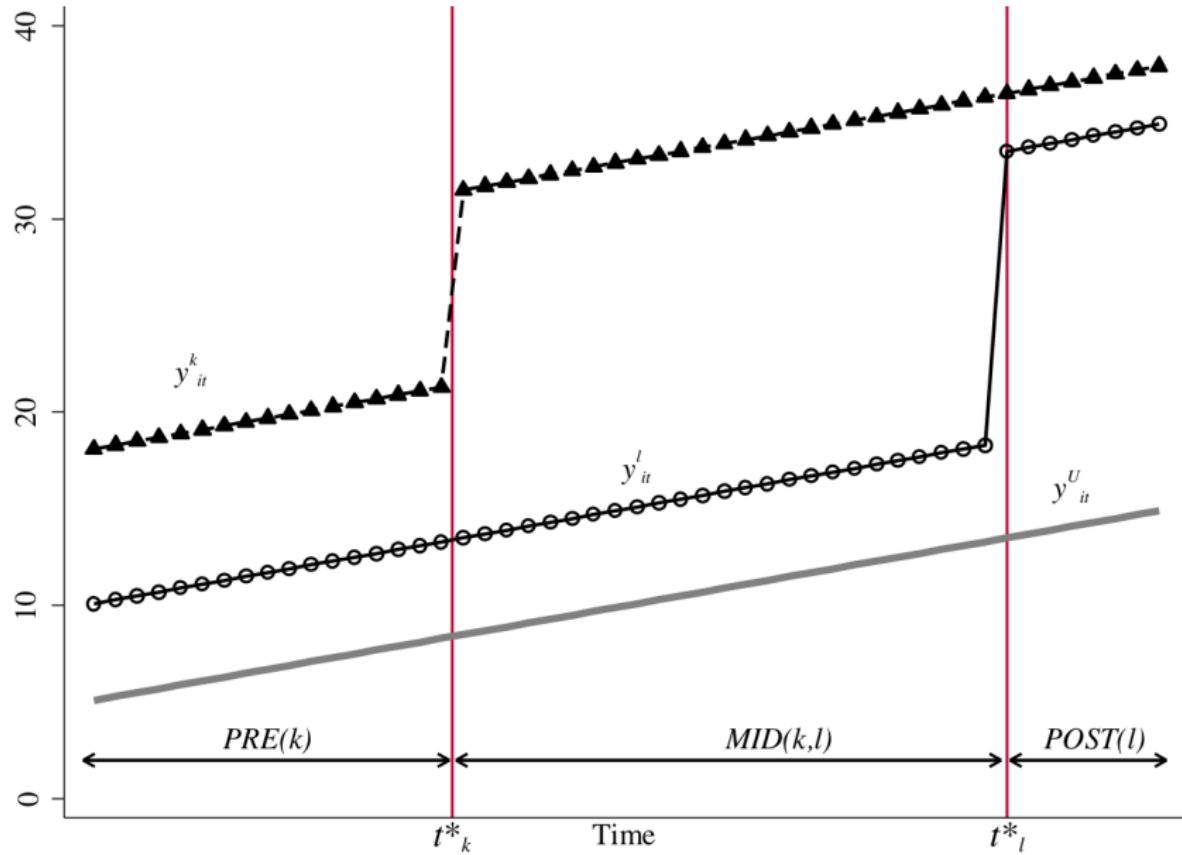
- Let there be two treatment groups (k, l) and one untreated group (U)
- k, l define the groups based on when they receive treatment (differently in time) with k receiving it earlier than l
- Denote \bar{D}_k as the share of time each group spends in treatment status
- Denote $\widehat{\delta}_{jb}^{2 \times 2}$ as the canonical 2×2 DD estimator for groups j and b where j is the treatment group and b is the comparison group

K^2 distinct DDs

Let's look at 3 timing groups (a, b and c) and one untreated group (U). With 3 timing groups, there are 9 2x2 DDs. Here they are:

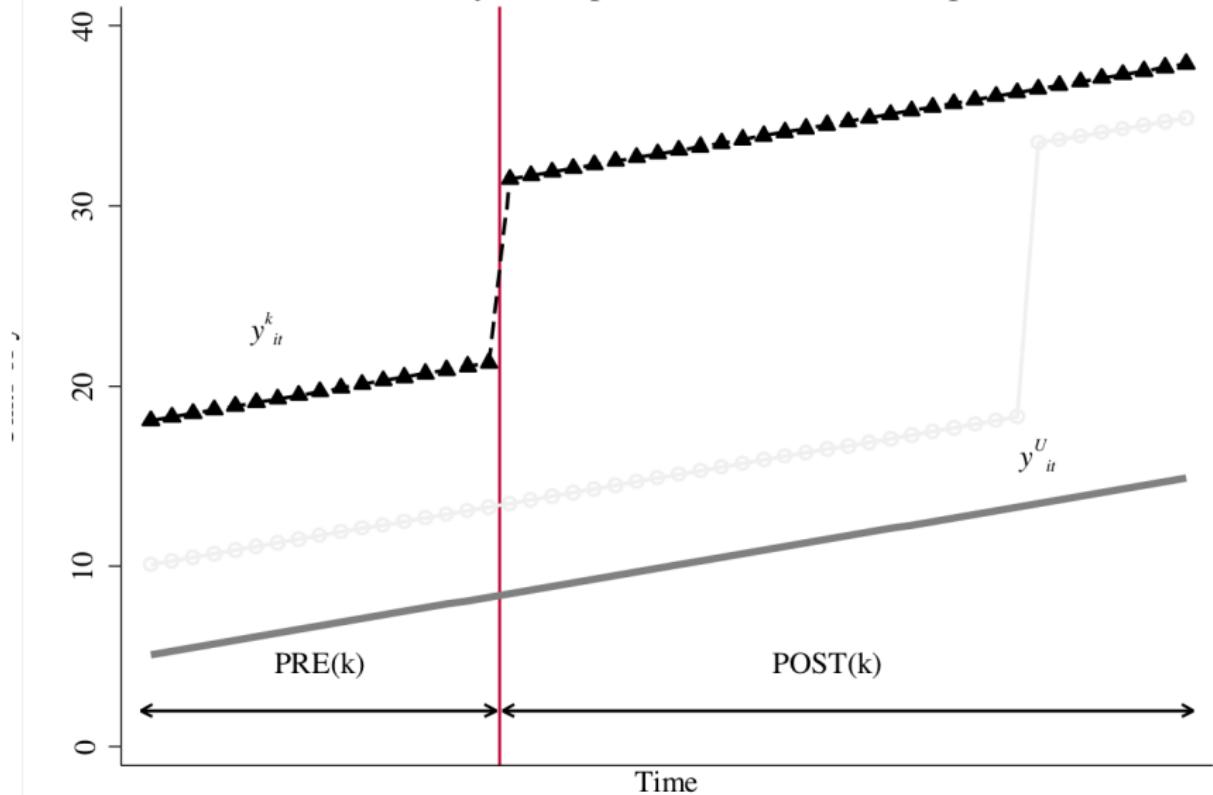
a to b	b to a	c to a
a to c	b to c	c to b
a to U	b to U	c to U

Let's return to our simpler example with k group treated at t_k^* and l treated at t_l^* plus the U untreated group



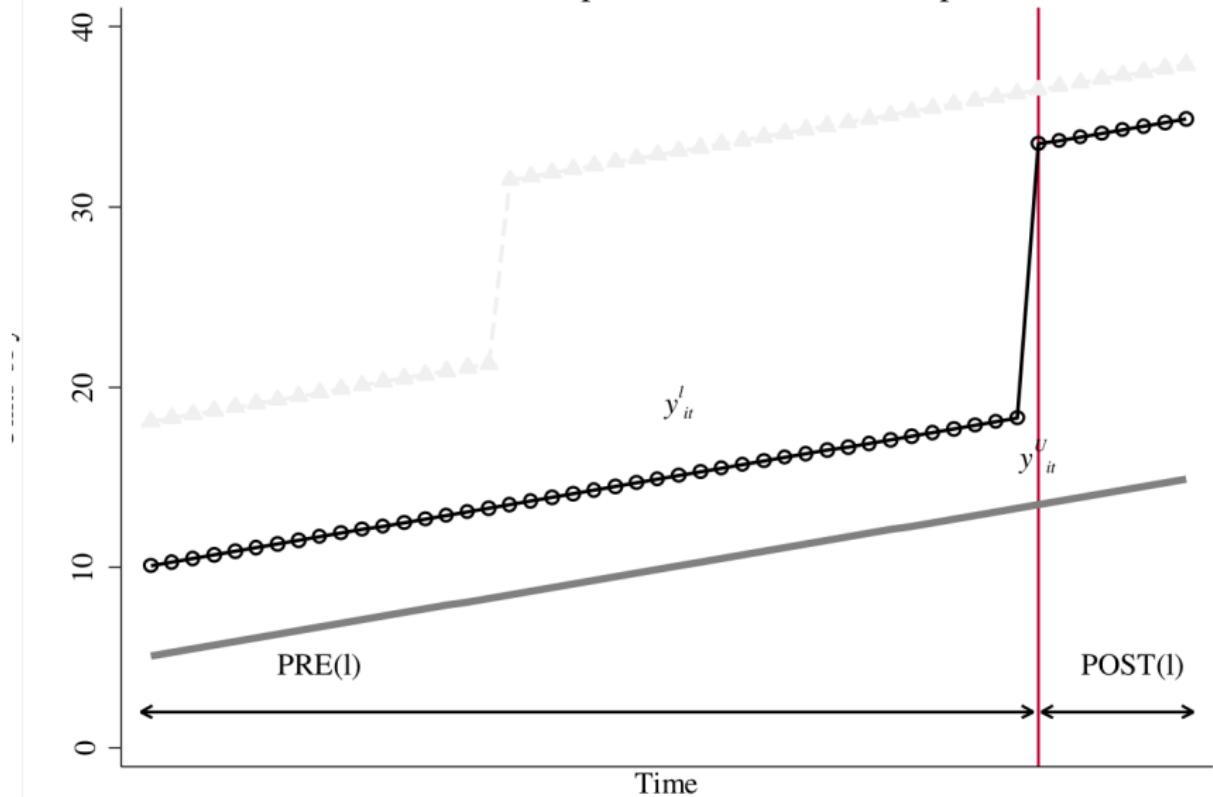
$$\widehat{\delta}_{kU}^{2 \times 2} = \left(\bar{y}_k^{post(k)} - \bar{y}_k^{pre(k)} \right) - \left(\bar{y}_U^{post(k)} - \bar{y}_U^{pre(k)} \right)$$

A. Early Group vs. Untreated Group

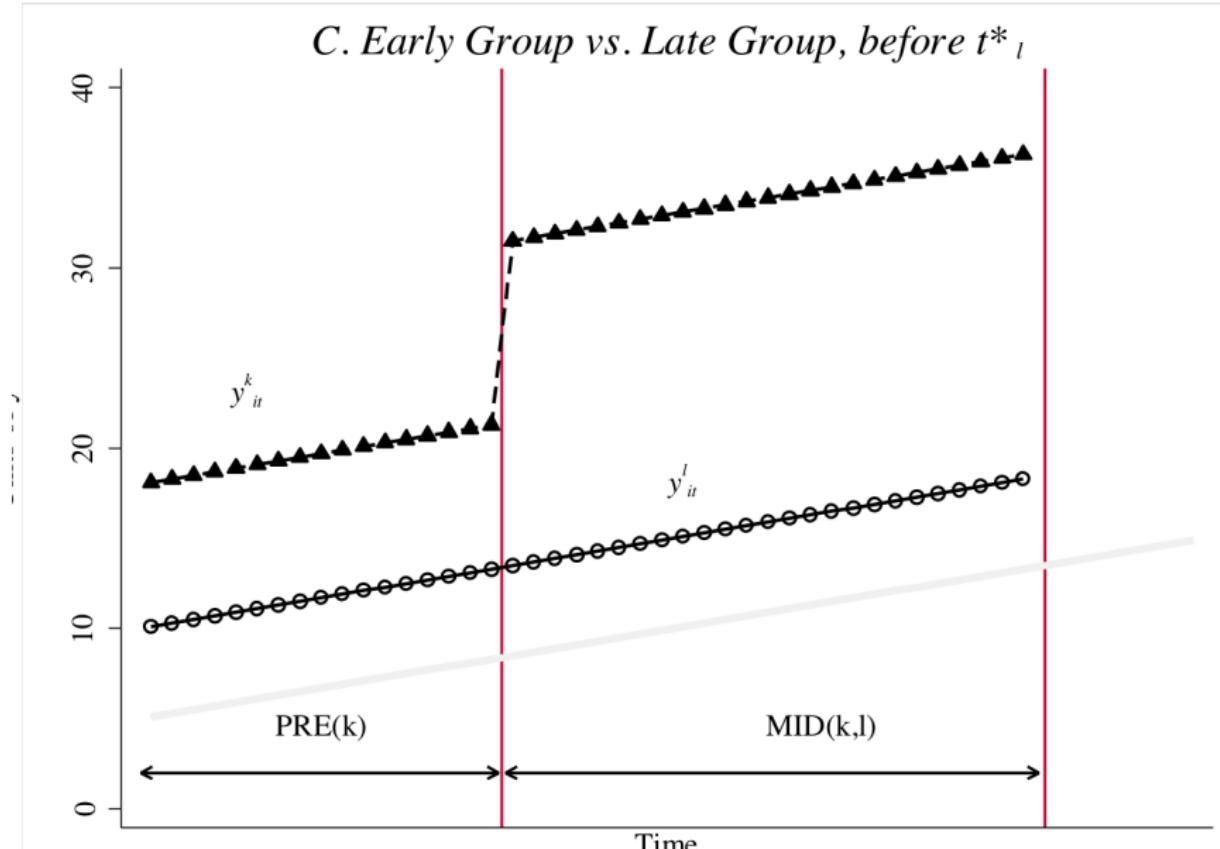


$$\widehat{\delta}_{IU}^{2 \times 2} = \left(\bar{y}_I^{post(I)} - \bar{y}_I^{pre(I)} \right) - \left(\bar{y}_U^{post(I)} - \bar{y}_U^{pre(I)} \right)$$

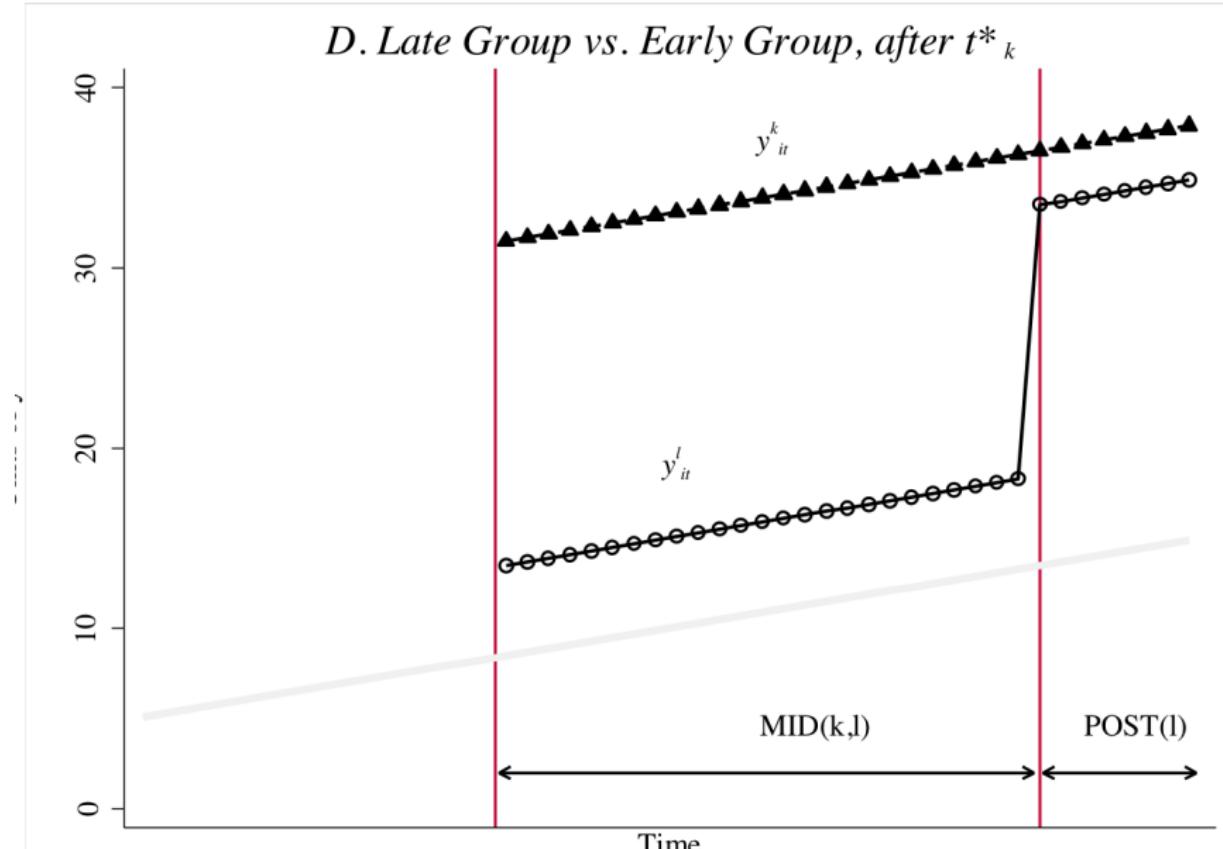
B. Late Group vs. Untreated Group



$$\delta_{kl}^{2 \times 2, k} = \left(\bar{y}_k^{MID(k,l)} - \bar{y}_k^{Pre(k,l)} \right) - \left(\bar{y}_l^{MID(k,l)} - \bar{y}_l^{PRE(k,l)} \right)$$



$$\delta_{lk}^{2 \times 2,I} = \left(\bar{y}_I^{POST(k,l)} - \bar{y}_I^{MID(k,l)} \right) - \left(\bar{y}_k^{POST(k,l)} - \bar{y}_k^{MID(k,l)} \right)$$



Bacon decomposition

TWFE estimate yields a weighted combination of each groups' respective 2x2 (of which there are 4 in this example)

$$\hat{\delta}^{DD} = \sum_{k \neq U} s_{kU} \hat{\delta}_{kU}^{2 \times 2} + \sum_{k \neq U} \sum_{l > k} s_{kl} \left[\mu_{kl} \hat{\delta}_{kl}^{2 \times 2, k} + (1 - \mu_{kl}) \hat{\delta}_{lk}^{2 \times 2, l} \right]$$

where that first 2x2 combines the k compared to U and the l to U (combined to make the equation shorter)

Third, the Weights

$$\begin{aligned}s_{ku} &= \frac{n_k n_u \bar{D}_k (1 - \bar{D}_k)}{\widehat{\text{Var}}(\tilde{D}_{it})} \\ s_{kl} &= \frac{n_k n_l (\bar{D}_k - \bar{D}_l) (1 - (\bar{D}_k - \bar{D}_l))}{\widehat{\text{Var}}(\tilde{D}_{it})} \\ \mu_{kl} &= \frac{1 - \bar{D}_k}{1 - (\bar{D}_k - \bar{D}_l)}\end{aligned}$$

where n refer to sample sizes, $\bar{D}_k(1 - \bar{D}_k)$
 $(\bar{D}_k - \bar{D}_l)(1 - (\bar{D}_k - \bar{D}_l))$ expressions refer to variance of
treatment, and the final equation is the same for two timing groups.

Weights discussion

- Two things to note:
 - More units in a group, the bigger its 2x2 weight is
 - Group treatment variance weights up or down a group's 2x2
- Think about what causes the treatment variance to be as big as possible. Let's think about the s_{ku} weights.
 - $\bar{D} = 0.1$. Then $0.1 \times 0.9 = 0.09$
 - $\bar{D} = 0.4$. Then $0.4 \times 0.6 = 0.24$
 - $\bar{D} = 0.5$. Then $0.5 \times 0.5 = 0.25$
 - $\bar{D} = 0.6$. Then $0.6 \times 0.4 = 0.24$
- This means the weight on treatment variance is maximized for *groups treated in middle of the panel*

More weights discussion

- But what about the “treated on treated” weights (i.e., $\bar{D}_k - \bar{D}_I$)
- Same principle as before - when the difference between treatment variance is close to 0.5, those 2x2s are given the greatest weight
- For instance, say $t_k^* = 0.15$ and $t_I^* = 0.67$. Then $\bar{D}_k - \bar{D}_I = 0.52$. And thus $0.52 \times 0.48 = 0.2496$.

Summarizing TWFE centralities

- Groups in the middle of the panel weight up their respective 2x2s via the variance weighting
- Decomposition highlights the strange role of panel length when using TWFE
- Different choices about panel length change both the 2x2 and the weights based on variance of treatment

Moving from 2x2s to causal effects and bias terms

Let's start breaking down these estimators into their corresponding estimation objects expressed in causal effects and biases

$$\begin{aligned}\widehat{\delta}_{kU}^{2\times 2} &= ATT_k Post + \Delta Y_k^0(Post(k), Pre(k)) - \Delta Y_U^0(Post(k), Pre) \\ \widehat{\delta}_{kl}^{2\times 2} &= ATT_k(MID) + \Delta Y_k^0(MID, Pre) - \Delta Y_l^0(MID, Pre)\end{aligned}$$

These look the same because you're always comparing the treated unit with an untreated unit (though in the second case it's just that they haven't been treated yet).

The dangerous 2x2

But what about the 2x2 that compared the late groups to the already-treated earlier groups? With a lot of substitutions we get:

$$\hat{\delta}_{lk}^{2\times 2} = ATT_{I, Post(I)} + \underbrace{\Delta Y_I^0(Post(I), MID) - \Delta Y_k^0(Post(I), MID)}_{\text{Parallel trends bias}} - \underbrace{(ATT_k(Post) - ATT_k(Mid))}_{\text{Heterogeneity bias!}}$$

Substitute all this stuff into the decomposition formula

$$\widehat{\delta}^{DD} = \sum_{k \neq U} s_{kU} \widehat{\delta}_{kU}^{2 \times 2} + \sum_{k \neq U} \sum_{l > k} s_{kl} \left[\mu_{kl} \widehat{\delta}_{kl}^{2 \times 2, k} + (1 - \mu_{kl}) \widehat{\delta}_{kl}^{2 \times 2, l} \right]$$

where we will make these substitutions

$$\begin{aligned}\widehat{\delta}_{kU}^{2 \times 2} &= ATT_k(Post) + \Delta Y_I^0(Post, Pre) - \Delta Y_U^0(Post, Pre) \\ \widehat{\delta}_{kl}^{2 \times 2, k} &= ATT_k(Mid) + \Delta Y_I^0(Mid, Pre) - \Delta Y_I^0(Mid, Pre) \\ \widehat{\delta}_{lk}^{2 \times 2, l} &= ATT_l(Post(l)) + \Delta Y_I^0(Post(l), MID) - \Delta Y_k^0(Post(l), MID) \\ &\quad - (ATT_k(Post) - ATT_k(Mid))\end{aligned}$$

Notice all those potential sources of biases!

Potential Outcome Notation

$$p \lim \widehat{\delta}_{n \rightarrow \infty}^{TWFE} = VWATT + VWPT - \Delta ATT$$

- Notice the number of assumptions needed even to estimate this very strange weighted ATT (which is a function of how you drew the panel in the first place).
- With dynamics, it attenuates the estimate (bias) and can even reverse sign depending on the magnitudes of what is otherwise effects in the sign in a reinforcing direction!
- Let's look at each of these three parts more closely

Variance weighted ATT

$$\begin{aligned} VWATT &= \sum_{k \neq U} \sigma_{kU} ATT_k(Post(k)) \\ &+ \sum_{k \neq U} \sum_{l > k} \sigma_{kl} \left[\mu_{kl} ATT_k(MID) + (1 - \mu_{kl}) ATT_l(POST(l)) \right] \end{aligned}$$

where σ is like s only population terms not samples.

- Weights sum to one.
- Note, if all the ATT are identical, then the weighting is irrelevant.
- But otherwise, it's basically weighting each of the individual sets of ATT we have been discussing, where weights depend on group size and variance

Variance weighted parallel trends

$$\begin{aligned} VWPT &= \sum_{k \neq U} \sigma_{kU} \left[\Delta Y_k^0(Post(k), Pre) - \Delta Y_U^0(Post(k), Pre) \right] \\ &+ \sum_{k \neq U} \sum_{l > k} \sigma_{kl} \left[\mu_{kl} \{ \Delta Y_k^0(Mid, Pre(k)) - \Delta Y_l^0(Mid, Pre(k)) \} \right. \\ &\quad \left. + (1 - \mu_{kl}) \{ \Delta Y_l^0(Post(l), Mid) - \Delta Y_k^0(Post(l), Mid) \} \right] \end{aligned}$$

There are K^2 parallel trends inside the weights. Their weighted average must equal zero.

Heterogeneity bias

$$\Delta ATT = \sum_{k \neq U} \sum_{l > k} (1 - \mu_{kl}) \left[ATT_k(Post(l) - ATT_k(Mid)) \right]$$

Now, if the ATT is constant over time, then this difference is zero, but what if the ATT is not constant? Then TWFE is biased, and depending on the dynamics and the VWATT, may even flip signs

Callaway and Sant'Anna 2020

- New papers are coming out focused on the issues that we are seeing with TWFE
- I'll discuss one though by Callaway and Sant'anna (2020) due to time constraints (call it CS)
- If we have time, I'll run through a simulation illustrating both the bias of TWFE and the unbiased estimation of this CS estimator
- Interesting ancestry – CS is a descendent of Abadie (2005) from earlier

Preliminary

CS considers identification, aggregation, estimation and inference procedures for ATT in DD designs with

- ① multiple time periods
- ② variation in treatment timing (i.e., differential timing)
- ③ parallel trends only holds after conditioning on observables

When might you use this estimator

Probably in the very situations describing your own study

- ① When treatment effects heterogenous by time of adoption
- ② When treatment effects change over time
- ③ When shortrun effects more pronounced than longrun effects
- ④ When treatment effect dynamics differ if people are first treated in a recession relative to expansion years

Group-time ATT is the parameter of interest in CS

$$ATT(g, t) = E[Y_t^1 - Y_t^0 | G_g = 1]$$

Group-time ATT

Group-time ATT is the ATT for a specific group and time

- Groups are basically cohorts of units treated at the same time
- CS will calculate an ATT per group/time which will be the sum of all $T - t_k$ for all groups (i.e., a lot)
- Group-time ATT estimates are not determined by the estimation method one adopts (first difference or FE) bc they are simple differences in means
- Does not directly restrict heterogeneity with respect to observed covariates, timing or the evolution of treatment effects over time
- Provides a way to aggregate over these to get a single ATT
- Inference is the bootstrap

Notation

- T periods going from $t = 1, \dots, T$
- Units are either treated ($D_t = 1$) or untreated ($D_t = 0$) but once treated cannot revert to untreated state
- G_g signifies a group and is binary. Equals one if individual units are treated at time period t .
- C is also binary and indicates a control group unit equalling one if “never treated” (can be relaxed though to “not yet treated”)
 - Recall the problem with TWFE on using treatment units as controls
- Generalized propensity score enters into the estimator as a weight:

$$\widehat{p(X)} = \Pr(G_g = 1 | X, G_c + C = 1)$$

Assumptions

Assumption 1: Sampling is iid (panel data)

Assumption 2: Conditional parallel trends (for either never treated or not yet treated)

$$E[Y_t^0 - Y_{t-1}^0 | X, G_g = 1] = [Y_t^0 - Y_{t-1}^0 | X, C = 1]$$

Assumption 3: Irreversible treatment

Assumption 4: Common support (propensity score)

Assumption 5: Limited treatment anticipation (i.e., treatment effects are zero pre-treatment)

CS Estimator (the IPW version)

$$ATT(g, t) = E \left[\left(\frac{G_g}{E[G_g]} - \frac{\frac{\hat{p}(X)C}{1-\hat{p}(X)}}{E \left[\frac{\hat{p}(X)C}{1-\hat{p}(X)} \right]} \right) (Y_t - Y_{g-1}) \right]$$

This is the inverse probability weighting estimator. Alternatively, there is an outcome regression approach and a doubly robust. Sant'Anna recommends DR. Notice how CS doesn't use already-treated as controls.

Staggered adoption (i.e., universal coverage)

Proof.

Remark 1: In some applications, eventually all units are treated, implying that C is never equal to one. In such cases one can consider the “not yet treated” ($D_t = 0$) as a control group instead of the “never treated?” ($C = 1$). □

Aggregated vs single year/group ATT

- The method they propose is really just identifying very narrow ATT per group time.
- But we are often interested in more aggregate parameters, like the ATT across all groups and all times
- They present two alternative methods for building “interesting parameters”
- Inference from a bootstrap

Stata simulation

Let's now review a simulation in Stata which can be downloaded from my github repo called `baker.do`.

Pedro Sant'Anna for the win

- Now a word from a good friend – Pedro Sant'Anna. Legend!
- He'll be discussing deChaisemartin and D'Haultfoeiller (2020) because:
 - He's a good guy
 - He's a great presenter
 - I fell behind

Sun and Abraham 2020

- Recall our discussion of event studies estimated with TWFE under differential timing
- Now that we know about the biases of TWFE when estimating aggregate DD parameters, let's revisit event studies under differential timing
- Callaway and Sant'Anna (2020) propose alternative estimators for event studies that estimate group-time ATT in relative event time
- But now we will discuss Sun and Abraham (2020) [SA] which is like a blend of Goodman-Bacon's decomposition and Callaway and Sant'anna alternative estimator to TWFE

Summarizing

- Goodman-Bacon (2021, forthcoming) focused on decomposition of TWFE to show bias under differential timing
- Callaway and Sant'anna (2020) presents alternative estimator that yields unbiased estimates of group-time ATTs which can be aggregated or put into event study plots
- Sun and Abraham (SA) is like a combination of the two papers

Summarizing (cont.)

- ① SA is a decomposition of the population regression coefficient on event study leads and lags with differential timing estimated with TWFE
- ② They show that the population regression coefficient is “contaminated” by information from other leads and lags
- ③ SA presents an alternative estimator that is not so dissimilar to CS

Summarizing (cont.)

- Problems seem to occur with DD when we introduce treatment effect heterogeneity
- Under treatment effect heterogeneity, spurious non-zero positive lead coefficients even when there is no pretrend
- This problem is exacerbated by the TWFE related weights as under some scenarios, the weights sum to zero and “cancel out” the treatment effects from other periods
- They present a 3-step TWFE based alternative estimator which addresses the problems that they find

Summarizing (cont.)

- Only decomposition of TWFE estimating dynamic leads and lags (Goodman-Bacon focused on a “static” specification)
- Contamination of coefficients on leads and lags by treatment effects depends on the magnitude of the weights on the true group-time ATT, or “cohort-specific ATT”
- Weights are a function of cohort composition
- Examining weights lets you gauge how treatment effect heterogeneity would interact with potential non-zero and non-convex weighting in population regression coefficients on the leads and lags

Difficult notation sadly

- When treatment occurs at the same time, we say they are part of the same cohort, e
- If we bin the data, then a lead or lag l will appear in the bin g so sometimes they use g instead of l or $l \in g$
- Building block is the “cohort-specific ATT” or $CATT_{e,l}$ – same thing as CS group-time ATT
- Estimate $CATT_{e,l}$ with population regression coefficient μ_l

Difficult notation (cont.)

- At each time t there are two possible treatment status $D_{i,t} \in \{0, 1\}$ over $T + 1$ time periods
- Path of treatment status scales exponentially with T and can take on 2^{T+1} possible values
- They focus on irreversible treatment where treatment status is non-decreasing sequence of zeroes and ones

Difficult notation (cont.)

- If a group is never treated, the ∞ symbol is used to either describe the group ($E_i = \infty$) or the potential outcome (Y^∞)
- $Y_{i,t}^\infty$ is the potential outcome for unit i if it had never received treatment (versus received it later), also called the baseline outcome
- Other counterfactuals are possible – maybe unit i isn't "never treated" but treated later in counterfactual

More difficult notation (cont.)

- Treatment effects are the difference between the observed outcome relative to the never-treated counterfactual outcome:
$$Y_{i,t} - Y_{i,t}^{\infty}$$
- We can take the average of treatment effects at a given relative time period across units first treated at time $E_i = e$ (same cohort) which is what we mean by $CATT_{e,l}$
- Doesn't use t index time ("calendar time"), rather uses l which is time until or time after treatment date e ("relative time")
- Think of it as $l = \text{year} - \text{treatment date}$

Definition 1

Definition 1: The cohort-specific ATT / periods from initial treatment date e is:

$$CATT_{e,I} = E[Y_{i,e+I} - Y_{i,e+I}^{\infty} | E_i = e]$$

Identifying assumption 1

Assumption 1: Parallel trends in baseline outcomes:

$E[Y_{i,t}^\infty - Y^\infty + i, s | E_i = e]$ is the same for all $e \in \text{supp}(E_i)$ and for all s, t and is equal to $E[Y_{i,t}^\infty - Y_{i,s}^\infty]$

Interesting SA comment: Never-treated units are likely to differ from ever-treated units in many ways; think of a Roy model. What does it imply that they chose not to get treated? It may imply net negative treatment effects and that could mean they may not share the same evolution of baseline outcomes as the treatment groups. If you think they are unlikely to satisfy this assumption, then drop them. Almost like a synthetic control approach.

Assumption 2

Assumption 2: No anticipator behavior in pre-treatment periods: There is a set of pre-treatment periods such that $E[Y_{i,e+I}^e - Y_{i,e+I}^\infty | E_i = e] = 0$ for all possible leads.

Basically means that potential outcomes prior to treatment at baseline are on average the same. This means there is no pre-trends, essentially. This is most plausible if the full treatment paths are not known to the units (e.g., Craigslist opening erotic services without announcement)

Assumption 3

Assumption 3: Treatment effect homogeneity: For each relative time period I , the $CATT_{e,I}$ doesn't depend on the cohort and is equal to $CATT_I$.

Assumption 3 requires each cohort experience the same path of treatment effects. Treatment effects need to be the same across cohorts in every relative period for homogeneity to hold, whereas for heterogeneity to occur, treatment effects just need to differ across cohorts in one relative time period. Doesn't preclude dynamic treatment effects, though. It just imposes that cohorts share the same treatment path.

Treatment effect heterogeneity

- Assumption 3 is violated when different cohorts experience different paths of treatment effects
- Cohorts may differ in their covariates which affect how they respond to treatment (e.g., if treatment effects vary with age, and there is variation in age across units first treated at different times, then there will be heterogeneous treatment effects)
- Doesn't rule out parallel trends

TWFE Regression

$$Y_{i,t} = \alpha_i + \delta_t + \sum_{g \in G} \mu_g 1\{t - E_i \in g\} + \varepsilon_{i,t}$$

They say E_i is the initial time of a binary variable absorbing treatment for unit i . Fixed effects should be obvious. μ_g is the population regression coefficient on the leads and lags that we want to estimate. We estimate this using OLS and get $\widehat{\mu}_g$.

We are interested in the properties of μ_g under differential timing as well as whether there are any never-treated units

Specifying the leads and lags

How will we specify the $1\{t - E_i \in g\}$ term? SA considers a couple:

- ① Static specification:

$$Y_{i,t} = \alpha_i + \delta_t + \mu_g \sum_{l \geq 0} D_{i,t}^l + \varepsilon_{i,t}$$

- ② Dynamic specification:

$$Y_{i,t} = \alpha_i + \delta_t + \sum_{l=-K}^{-2} \mu_l D_{i,t}^l + \sum_{l=0}^L \mu_l D_{i,t}^l + \varepsilon_{i,t}$$

Multicollinearity

Dynamic specification requires deciding which leads to drop. They recommend dropping two: $I = -1$ and some other one (they seem to favor $I = -4$). The reason is twofold. You drop one of them to avoid multicollinearity in the relative time indicators. You drop a second one because of the multicollinearity coming from the linear relationship between TWFE and the relative period indicators.

Trimming and binning

- First some terms: trimming and binning, I do both in the Mixtape when analyzing Cheng and Hoekstra (2013)
- Binning means placing all “distant” relative time indicators into a single one. Done because of the sparseness of units in such distant bins. So if there’s 3 distant leads and lags that aren’t balanced, combine them all into the last lead and lag
- Trimming means excluding any relative period for which you don’t have balance in relative time. This creates a balanced panel “in relative time”, but imbalanced panel length overall.
- They’ll analyze both and how they affect $\widehat{\mu}_g$ estimation using TWFE

Interpreting $\widehat{\mu}_g$ under no to all assumptions

Proposition 1 (no assumptions): The population regression coefficient on relative period bin g is a linear combination of differences in trends from its own relative period $l \in g$, from relative periods $l \in g'$ of other bins $g' \neq g$, and from relative periods excluded from the specification (e.g., trimming).

$$\begin{aligned}\mu_g = & \underbrace{\sum_{l \in g} \sum_e w_{e,l}^g (E[Y_{i,e+l} - Y_{i,0}^\infty | E_i = e] - E[Y_{i,e+l}^\infty - Y_{i,0}^\infty])}_{\text{Good stuff}} \\ & + \underbrace{\sum_{g' \neq g} \sum_{l \in g'} \sum_e w_{e,l}^g (E[Y_{i,e+l} - Y_{i,0}^\infty | E_i = e] - E[Y_{i,e+l}^\infty - Y_{i,0}^\infty])}_{\text{Bleh - Other included relative time}} \\ & + \underbrace{\sum_{l \in g^{excl}} \sum_e w_{e,l}^g (E[Y_{i,e+l} - Y_{i,0}^\infty | E_i = e] - E[Y_{i,e+l}^\infty - Y_{i,0}^\infty])}_{\text{More bleh - Excluded}}\end{aligned}$$

Superscript g associates the weight with coefficient μ_g . The weight associated with cohort e in relative period l is equal to the population regression coefficient on the $1\{t - E_i \in g\}$ from regression $D_{i,t}^l \times 1\{E_i = e\}$ on all bin indicators included in the regression and TWFE. Just the mechanics of double demeaning from TWFE

Weight ($w_{e,I}^g$) summation cheat sheet

- ① For relative periods of μ_g own $I \in g$, $\sum_{I \in g} \sum_e w_{e,I}^g = 1$
- ② For relative periods belonging to some other bin $I \in g'$ and $g' \neq g$, $\sum_{I \in g'} \sum_e w_{e,I}^g = 0$
- ③ For relative periods not included in G , $\sum_{I \in g^{excl}} \sum_e w_{e,I}^g = -1$

Estimating the weights

Regress $D_{i,t}^I \times 1\{E_i = e\}$ on:

- ① all bin indicators included in the main TWFE regression,
- ② $\{1\{t - E_i \in g\}\}_{g \in G}$ (i.e., leads and lags) and
- ③ the unit and time fixed effects

Interpretation of coefficients under parallel trends only

Proposition 2: Under the parallel trends only, the population regression coefficient on the indicator for relative period bin g is a linear combination of $CATT_{e,I \in g}$ as well as $CATT_{d,I'}$ from other relative periods $I' \notin g$ with the same weights stated in Proposition 1:

$$\begin{aligned}\mu_g = & \underbrace{\sum_{I \in g} \sum_e w_{e,I}^g CATT_{e,I}}_{\text{Desirable}} \\ & + \underbrace{\sum_{g' \neq g, g' \in G} \sum_{I' \in g'} \sum_e w_{e,I'}^g CATT_{e,I'}}_{\text{Undesirable - other specified bins}} \\ & + \underbrace{\sum_{I' \in g^{excl}} \sum_e w_{e,I'}^g CATT_{e,I'}}_{\text{Undesirable - excluded relative time indicators}}\end{aligned}$$

Comment on Proposition 2

The coefficient μ_g can be written as an average of $CATT_{e,I}$ from own periods but also $CATT_{e,I'}$ from other periods.

The weights are still functions of cohort comparisons, like in Proposition 1, which means μ_g can be written as non-convex averages of not only $CATT_{e,I}$ from own periods $I \in g$, but also $CATT_{e,I'}$ from other periods.

Means μ_g could in fact be the wrong sign to all $CATT_{e,I \in g}$.

Weights can help us gauge the severity of this problem.

When the weights have larger magnitude, treatment effect heterogeneity matters more as a particular $CATT_{e,I}$ can drive the overall estimates. But when weights are uniform, treatment effect heterogeneity matters less.

Interpretation under parallel trends and no anticipation

Proposition 3: If parallel trends holds and no anticipation holds for all $l < 0$ (i.e., no anticipatory behavior pre-treatment), then the population regression coefficient μ_g for g is a linear combination of post-treatment $CATT_{e,l'}$ for all $l' \geq 0$.

$$\begin{aligned}\mu_g = & \sum_{l' \in g, l' \geq 0} \sum_e w_{e,l'}^g CATT_{e,l'} \\ & + \sum_{g' \neq g, g' \in G} \sum_{l' \in g', l' \geq 0} \sum_e w_{e,l'}^g CATT_{e,l'} \\ & + \sum_{l' \in g^{excl}, l' \geq 0} \sum_e w_{w,l'}^g CATT_{e,l'}\end{aligned}$$

Proposition 3 comment

Notice how once we impose zero pre-treatment treatment effects, those terms are gone (i.e., no $l \in g, l < 0$). But the second term remains unless we impose treatment effect homogeneity (homogeneity causes terms due to weights summing to zero to cancel out). Thus μ_g may be non-zero for pre-treatment periods even though parallel trends hold in the pre period.

Proposition 4

Proposition 4: If parallel trends and treatment effect homogeneity, then $CATT_{e,I} = ATT_I$ is constant across e for a given I , and the population regression coefficient μ_g is equal to a linear combination of $ATT_{I \in g}$, as well as $ATT_{I' \notin g}$ from other relative periods

$$\begin{aligned}\mu_g &= \sum_{I \in g} w_I^g ATT_I \\ &+ \sum_{g' \neq g} \sum_{I' \in g'} w_{I'}^{g'} ATT_{I'} \\ &+ \sum_{I' \in g^{excl}} w_{I'}^{g'} ATT_{I'}\end{aligned}$$

Proposition 4 comment

The weight $w_I^g = \sum_e w_{e,I}^g$ sums over the weights $w_{e,I}^g$ from Proposition 1 and is equal to the population regression coefficient from the following auxiliary regression:

$$D'_{i,t} = \alpha_i + \lambda_t + \sum_{g \in G} w_I^g \cdot 1\{t - E_i \in g\} + u_{i,t}$$

which regresses $D'_{i,t}$ on all bin indicators and TWFE

On binning

- Many propose either binning or trimming to create “balanced” panels (in relative event time)
- But SA notes that binning in simulations creates uninterpretable weights (due to the binned $CATT_{e,I'}$ inclusion in μ_g), whereas trimming creates weights that are more reasonable
- This may be because trimming subtracts the corresponding $CATT_{e,I'}$ from μ regression coefficient

Intuition for contamination

- Stupid notation make Hulk smash!
- Let's do a simple toy example instead

Balanced panel $T = 2$ with cohorts $E_i \in \{1, 2\}$. We drop two relative time periods to avoid multicollinearity, so we will include bins $\{-2, 0\}$ and drop $\{-1, 1\}$.

Toy example

$$\begin{aligned}\mu_{-2} = & \underbrace{CATT_{2,-2}}_{\text{own period}} + \underbrace{\frac{1}{2}CATT_{1,0} - \frac{1}{2}CATT_{2,0}}_{\text{other included bins}} \\ & + \underbrace{\frac{1}{2}CATT_{1,1} - CATT_{1,-1} - \frac{1}{2}CATT_{2,-1}}_{\text{Excluded bins}}\end{aligned}$$

- Parallel trends gets us to all of the $CATT$
- No anticipation makes $CATT = 0$ for all $l < 0$ (all $l < 0$ cancel out)
- Homogeneity cancels second and third terms
- Still leaves $\frac{1}{2}CATT_{1,1}$ – you chose to exclude a group with a treatment effect

Lesson: drop the relative time indicators on the left, not things on the right, bc lagged effects will contaminate through the excluded bins

Interaction-weighted estimator

- They propose an interacted weighted estimator (IW) as a consistent estimator for μ_g
- Estimator uses either never-treated as controls or “last cohort treated” if no never-treated (contra CS which uses “not yet treated”)
- No covariates bc this is a regression with fixed effects and time-varying covariates create own biases, although they note you can plug in CS for the DD calculation and recover *CATT* that way
- The interaction is a TWFE regression specification that interacts relative period indicators with cohort/group indicators, excluding indicators for never-treated cohorts

Interaction-weighted estimator

- **Step one:** Do this DD regression and hold on to $\hat{\delta}_{e,l}$

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{e \notin C} \sum_{l \neq -1} \delta_{e,l} (1\{E_i = e\} \cdot D_{i,t}^l) + \varepsilon_{i,t}$$

Can use never-treated or last-treated cohort. Drop always treated.
The $\delta_{e,l}$ is a DD estimator for $CATT_{e,l}$ with particular choices for pre-period and cohort controls

Interaction-weighted estimator

- **Step two:** Estimate weights using sample shares of each cohort in the relevant periods:

$$Pr(E_i = e | E_i \in [-l, T - l])$$

IW estimator

- **Step three:** Take a weighted average of estimates for $CATT_{e,I}$ from Step 1 with weight estimates from step 2

$$\hat{v}_g = \frac{1}{|g|} \sum_{I \in g} \sum_e \hat{\delta}_{e,I} \widehat{Pr}\{E_i = e | E_i \in [-I, T - I]\}$$

Consistency and Inference

- Under parallel trends and no anticipation, $\hat{\delta}_{e,I}$ is consistent, and sample shares are also consistent estimators for population shares.
- Thus IV estimator is consistent for a weighted average of $CATT_{e,I}$ with weights equal to the share of each cohort in the relevant period(s).
- They show that each IW estimator is asymptotically normal and derive its asymptotic variance. Doesn't rely on bootstrap like CS.

DD Estimator of CATT

Definition 2: DD estimator with pre-period s and control cohorts C estimates $CATT_{e,I}$ as:

$$\widehat{\delta}_{e,I} = \frac{E_N[(Y_{i,e+I} - Y_{i,s}) \times 1\{E_i = e\}]}{E_N[1\{E_i = e\}]} - \frac{E_N[(Y_{i,e+I} \times 1\{E_i \in C\})]}{E_N[1\{E_i \in C\}]}$$

Proposition 5: If parallel trends and no anticipation both hold for all pre-periods, then the DD estimator using any pre-period and non-empty control cohorts (never-treated or not-yet-treated) is an unbiased estimate for $CATT_{e,I}$.

Software

Use staggered from

<https://github.com/jonathandroth/staggered> by Jon Roth
(Brown University). There is also a Stata wraparound using the
rcall package in Stata. See instructions on the URL above.

Conclusion of SA

- Bacon shows the TWFE coefficient on the static parameter is “contaminated” by other periods leads and lags
- Three strong assumptions needed for TWFE to be unbiased: parallel trends, no anticipation, and treatment homogeneity
- Three step interaction-weighted estimator is an alternative
- Doesn’t restrict to treatment profile homogeneity
- Callaway and Sant’Anna (2020) and Sun and Abraham (2020) use different controls, but under certain situations (no covariates, never treated) they are the same (“nested”)
- Software in R and Stata exist

Difference-in-differences

Covariates

Differential timing

Imputation DiD

Alternative estimators

Basic suggestions going forward

2SDiD

Borusyak, et al. 2021

Athey, et al. 2021

Imputation

Some methods are more obviously imputations than others though. I am thinking of imputation as occurring mainly at the unit level, as opposed to manipulations of the data that go directly to estimating the ATT such as DiD.

Explicit imputation methods

Previous discussions had focused on a manual aggregation of weights.

But other recent alternatives have emerged that I will simply term “explicit imputation estimators”

Here we will discuss three: Gardner (20210, Borusyak, Jaravel and Spiess (2021) imputation estimator and Athey, et al. (2018) matrix completion for panel data

2SDiD

- I'd like to go back to a more traditional form of analysis by reviewing Gardner (2021)
- Like a few other papers, Gardner (2021) is both a diagnosis of the illness and a cure, and I'm putting his cure into an explicit imputation framework
- John Gardner is an assistant professor and applied econometrician at University of Mississippi – smart, cool, and former colleague of Brant Callaway of Callaway and Sant'Anna
- The cure will be nicely called two-stage difference-in-differences (2SDiD) – Nice name!

Highlights

- Why does TWFE fail under differential timing? Violates strict exogeneity under heterogeneity
- The logic of the failure suggests an obvious, but previously unknown, solution which is the 2SDiD
- I'll explain 2SDiD, focus on the parallel trends implications, and show we can get a consistent and unbiased estimate of group and relative time fixed effects
- If you can get consistent and unbiased estimates of group and relative time fixed effects, then you can delete them and run normal analysis
- We'll work through some code

Background

- By now, we all agree that TWFE just doesn't handle heterogeneity under differential timing very well
- We've seen in the Goodman-Bacon decomposition why – it's caused by TWFE implicitly calculating late to early 2x2s, which are a source of bias
- But some of you are coming straight from a panel econometrics course that maybe didn't use potential outcomes notation
- Isn't strict exogeneity enough for consistent estimates? What then does strict exogeneity have to do with heterogeneity and differential timing?
- Everything

More background

"It seems natural that TWFE should identify the ATT" – Gardner (2021)

It just seems like TWFE with a DiD will estimate the ATT with weights that we'll find intuitive. Was this just a conjecture and was never true? Why isn't this working?

High level discussion

- TWFE identifies the ATT when the heterogeneous effects are distributed equally across all groups and periods, but since that is a knife-edge situation, it is likely that TWFE will not in our applications meet this special scenario
- In the two group case, that is what happens though which is why TWFE worked fine there
- Metaphorically, the two group case that we always used to pin our intuition of what DiD was doing was the exception not the rule
- Goodman-Bacon (2021) shows the problem is caused by late-to-early comparisons; Gardner (2021) will show that the problem is misspecification
- Think of these as different perspectives on the same problem

Model misspecification

"Misspecified DiD regression models project heterogenous treatment effects onto group and period fixed effects rather than the treatment status itself"

Spoiler: This analysis of the problem suggests solution – why don't we remove those?

2SDiD

“What’s the name of that kid from Mexico?” – Ted Lasso

“Dani Rojas” – Nate the Great

“Great name” – Ted Lasso

- Two stage DiD is a great name because of its connection to that classic IV model 2SLS
- If you can link it to 2SLS in your mind, it may help you because it’ll show you that Gardner’s model is a two stage model
- First stage – estimate the group and relative time fixed effects using only the $D = 0$ observations
- Second stage – using predicted values based off those fixed effect coefficients, run your model off the transformed outcome
- Get the standard errors right just like 2SLS by taking the first stage into account

More high level

- The second step recovers the average difference in outcomes between treated and untreated units after removing group and period fixed effects
- What I like about Gardner's method is its pleasant familiarity, its speed
- But note, it's not going to allow you to do the kind of heterogeneity analysis that CS allows for
- Some of the differences will be due to slightly different PT assumptions, and some will because 2SDID will be using all of the data for analysis, not just the baseline for calculating the DID estimates

Notation

i : panel units

t : calendar time – think of real dates

$g \in \{0, 1, \dots, G\}$ – groups

$p \in \{0, 1, \dots, P\}$ – relative time or “periods”

Periods are successive. Group 0 – never treated. Group 1 – treated in period 1, 2, and on. Group 2 – treated in period 2, etc.

Parameters

$$\beta_{gp} = E \left[Y_{gpit}^1 - Y_{gpit}^0 | g, p \right]$$

It's a group-time ATT but expressed in a more traditional econometric notation that you could easily find in Wooldridge or some such

Modeling basics

Under parallel trends, mean outcomes will satisfy the following equation

$$E\left[Y_{gpit}|g, p, D_{gp}\right] = \lambda_g + \gamma_p + \beta_{gp}D_{gp}$$

In two-group, group and period effects are eliminated with dummies because TWFE uses dummies to demean across multiple dimensions. Then TWFE identifies ATT. But this does not hold when average effects vary across group and period. There are many ways to express a treatment effect's across group and time, but Gardner presented it as a weighted average of the coefficients for only that group-period situation:

$$E\left(\beta_{gp}|D_{gp} = 1\right) = E\left(Y_{gpit}^1 - Y_{gpit}^0|D_{gp} = 1\right)$$

Strict exogeneity violation

Rewriting the above we get:

$$\begin{aligned} E\left[Y_{gpit}|g, p, D_{gp}\right] &= \lambda_g + \gamma_p + E\left[\beta_{gp}|D_{gp} = 1\right]D_{gp} \\ &\quad \left[\beta_{gp} - E(\beta_{gp}|D_{gp} = 1)\right]D_{gp} \end{aligned}$$

The problem is there's this weird new error term and it isn't mean zero under heterogenous treatment effects spread across group and period. Unlike the two group case, the coefficient on D_{gp} from TWFE doesn't identify the average $E(\beta_{gp}|D_{gp} = 1)$

So let's see Gardner's solution, but note – his solution was suggested by the problem itself. Gardner is thoughtful and observant.

DiD regression estimand

- So if TWFE isn't recovering $E(\beta_{gp}|D_{gp} = 1)$, then what is it recovering?
- He shows that under PT, the coefficient on D_{gp} is:

$$\beta^* = \sum_{g=1}^G \sum_{p=g}^P w_{gp} \beta_{gp}$$

- So then – what are the weights w_{gp} ?
- Groan – It's a huge mess, and I hate even showing it to you because I find the weights almost impossible to decipher, but maybe you'll have a better go at it than me

Weights

$$w_{gp} = \frac{\left\{ [1 - P(D_{gp} = 1|g)] - [P(D_{gp} = 1|p) - P(D_{gp} = 1)] \right\} P(g, p)}{\sum_{g=1}^G \sum_{p=g}^P \left\{ [1 - P(D_{gp} = 1|g)] - [P(D_{gp} = 1|p) - P(D_{gp} = 1)] \right\} P(g, p)}$$

Terms:

- $P(D_{gp} = 1|p)$: share of units treated in period p
- $P(D_{gp} = 1|g)$: share of periods in which g is treated
- $P(D_{gp} = 1)$: share of unit \times time treated
- $P(g, p)$: population share of observation corresponding to group g and period p

I thought about changing all those probabilities into means, but honestly, it really didn't help me at all. But Gardner notes that this is from theorem 1 of deChaisemartin and D'Haultfoeuiller (2020) and his Appendix A

Estimation

$$Y_{gpit} = \lambda_g + \gamma_p + \beta D_{gp} + \varepsilon_{gpit}$$

This specification assumes a conditional expectation function that is linear in group, period and treatment status. But when the model is misspecified, it will attribute some of the heterogeneity impacts of the treatment to group and period fixed effects. The longer the treatment, the greater \bar{D} is, the more that group's treatment effects will be absorbed by group fixed effects. When misspecified, TWFE doesn't recover $E[\beta|D = 1]$.

Statistical issues

- Common support: “as long as there are untreated and treated observations for each group and period, λ_g and γ_p are identified from the subpopulation of untreated groups and periods.”
- Identification: “the overall group \times period ATT is identified from a comparison of mean outcomes between treated and untreated groups after removing group and period effects.”

Estimation: First stage

First stage:

$$Y_{gpit} = \lambda_g + \gamma_p + \varepsilon_{gpit}$$

using only $D_{gp} = 0$, retaining the fixed effects. Collect the $\widehat{\lambda}_g$ and $\widehat{\gamma}_p$.

Estimation: Second stage

Second stage:

$$\begin{aligned}\hat{y}_{gpit} &= y_{gpit} - \widehat{\lambda}_g - \widehat{\gamma}_p \\ \widehat{y}_{gpit} &= \alpha + \beta D_{gp} + \psi_{gpit}\end{aligned}$$

Why does this work? Parallel trends assumption implies:

$$E(y_{gpit}|g, p, D_{gp}) - \lambda_g - \gamma_p = E\left[\beta_{gp}|D_{gp} = 1\right]D_{gp} + \left[\beta_{gp} - E(\beta_{gp}|D_{gp} = 1)\right]D_{gp}$$

But because

$$E\left\{ [\beta_{gp} - E(\beta_{gp}|D_{gp} = 1)]D_{gp}|D_{gp} \right\} = 0$$

Estimand

Then this procedure will identify $E(\beta_{gp}|D_{gp} = 1)$. Consistency and unbiasedness proofs.

This is $E(\beta_{gp}|D_{gp} = 1) = \sum^G \sum^P \beta_{gp} P(g, p|D_{gp} = 1)$. It will tend to put more weight, by definition, on groups earlier into their treatment. But this isn't the same as the negative weighting that BJS say occurs oof the long lags. It just means there are more of them.

Event studies are:

$$y_{gpit} = \lambda_g + \gamma_p + \sum_{r=-R}^P \beta_r D_{rgp} + \varepsilon_{gpit}$$

Just change the second stage with the transformed outcome.

Inference

- Standard errors are wrong on the second stage because the dependent variable uses estimates obtained from the first stage.
- The asymptotic distribution of the second stage can be obtained by interpreting the two-stage procedure as a joint GMM

Background

- Arguably the first paper in this genre of “new did” was Borusyak and Jaravel (2017) – they provided analysis of TWFE flaws under heterogeneity as well as event study analysis
- The paper was not published, but instead was recently updated with a new coauthor (Spiess) and modified considerably
- This is a paper that shows the problems with TWFE under heterogeneity, but then writes out a solution that uses imputation

Themes

- Somewhat of an ambitious goal set for applied microeconomists in that they strongly recommend defining ahead of time exactly what target parameter you're wanting to identify
- With that parameter, then build the estimator – a Field of Dreams paradigm
- The TWFE error was caused by users conflating estimation and assumptions as well as poor defining of target parameters
- As a result, the model suffered from misspecification

Themes

Static model

$$y_{it} = \alpha_i + \gamma_t + \delta D_{it} + \varepsilon_{it}$$

Contribution: Define target parameters and assumptions. Proposes a more formal disciplined approach of choosing the weighted average of treatment effects

Basics

- Potential outcomes without treatment will follow a parallel trend (but one with a bit more structure)
- No anticipatory effects
- Treatment effects follow some model that restricts heterogeneity *a priori* for economic reasons

Event study contributions

- ① Can't identify point estimates of leads in event study design
 - Separate out the testable assumptions about pre-trends from dynamic treatment effects under these assumptions
- ② Implicit homogeneity assumption in event study may lead to estimates putting negative weight on long-run lags under differential timing
 - When we have long lags, regression is using extrapolation based on forbidden regressions which negatively weights
 - This is fine with homogenous treatment effects and in fact is an argument for TWFE, but not with heterogeneity
- ③ Spurious identification of longrun effects can happen under heterogeneity with staggered rollouts

Heterogeneity is hard

Again, this paper is telling us that the presence of heterogeneous treatment effects is largely what makes analysis so challenging

Imputation estimator

"The most efficient linear unbiased estimator of any pre-specified weighted average of treatment effects under homoskedasticity"

- Separate assumption from estimation; incorporate the former
- Estimate a flexible high-dimensional regression
- Aggregate the coefficients

Characteristics

- All other unbiased linear estimators are less efficient
- Imputation
- Avoids pre-test problems pointed out by Roth (2018) (just wasn't able to work it in unfortunately)

Setting

- Irreversible treatment, of “absorbing state”
- Never treated units are like SA described with potential outcomes of the infinity symbol
- Definition of the target parameter, which will be the static parameter is a weighted average of underlying treatment effects

$$\delta_w = \sum_{it \in \omega} w_{it} \delta_{it} = w_1' \delta$$

Target parameter

$$\delta_w = \sum_{it \in \omega} w_{it} \delta_{it} = w_1' \delta$$

- The static parameter is essentially the final product of a weighting procedure of individual treatment effects
- Thus it's in the tradition of CS and SA who also build aggregated parameter estimates from smaller building blocks
- Two elements to the estimator then:
 - ① Recovery of these smaller treatment effects
 - ② Weights

Value of target parameters

- With a target parameter in mind, we see the elements that we need
- When we see the elements we need, we have a guide for estimation
- This is the flavor of new applied work when contrasted against the tradition DiD which conflated target parameters with the parameter in a TWFE regression

Assumptions

- ① Parallel trends with a twist

$$Y^0 = \alpha_i + \beta_t + \varepsilon_{it}$$

$$E[\varepsilon_{it}] = 0 \forall it \in \Omega$$

- ② No anticipation (needed for leads)
- ③ Restricted causal effects $\beta_\tau = 0$

First assumption has potential outcome following a linear model; third assumption is new and the authors hit it over and over. If you're willing to impose zero treatment effects based on economic theory, then as they'll show, there are gains you incur from doing so

Conventional event studies

Diagnose problems with TWFE both in static and dynamic specifications

Consider a saturated model:

$$y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \sum_{h=-a; h \neq -1}^{b-1} \delta_h 1[k_{it} = h] + \delta_{bt} 1[k_{it} \geq b] + \tilde{\varepsilon}_{it}$$

Conditional on “passing” a pretrend test (e.g., $\delta_n = 0$ for all event study periods before treatment), lags represent dynamic treatment effects

Discussion of this conventional model

- This specification assumes no anticipation as well as homogeneity
- TWFE coefficients are interpreted as averages of causal effects for some horizon where weights are left to the mechanics of whatever TWFE algebra
- (We just pray the weights make sense because, well, proof by assumption and all)

Under identification of fully dynamic specification

- Problem 1: it does not impose a strong enough version of assumption 2 (no anticipation)
- If there is no never-treated group, then these specifications are under-identified
- Path of leads is not identified under fully dynamic OLS
- Adding a linear trend to this path fits the data just as easily

Unrestricted dynamics

"Formally, the problem arises because a linear time trend t nad a linear term in the cohort E_i (subsumed by unit FE) can perfectly reproduce a linear term in relative time $K_{it} - A - E_i$. Therefore a complete set of treatment leads and lags, which is equivalent to the FE of relative time, is collinear with the unit and period fixed effects."

Resolving this collinearity problem requires *stronger* restrictions on leads. Specifically, we must drop **two leads**, not just one. We drop one -1 and -a, which is why they insist on a priori reasoning about the data. Increases power. Stronger than minimal assumptions allow for more powerful tests.

Assumption 3

- It's assumption 3 – homogeneity – that BJS (and really the first paper, Borusyak and Jaravel) showed was a problem for traditional event studies
- And we saw that earlier with Sun and Abraham
- What happens is that with heterogeneity, the weights on the treatment effects can become negative

Negative weights on the longrun coefficients

- If you have no anticipation and parallel trends, then TWFE will in the static model estimate a weighted average of TE
- Problem is the weights could be negative
- They give a simple illustration

Simple illustration

Table: TWFE dynamics

$E(y_{it})$	$i = A$	$i = B$
t=1	α_A	α_B
t=2	$\alpha_A + \beta_2 + \delta_{A2}$	$\alpha_B + \beta_2$
t=3	$\alpha_A + \beta_3 + \delta_{A3}$	$\alpha_B + \beta_3 + \delta_{B3}$
Event date	$E_i = 2$	$E_i = 3$

Static: $\delta = \delta_{A2} + \frac{1}{2}\delta_{B3} - \frac{1}{2}\delta_{A3}$.

Notice the negative weight on the furthest lag. This is how TWFE rolls.

Main takeaway

The larger the effects in the longrun, the smaller the coefficient will be (hence why the longrun lags will be “downward biased”).

“With a large never-treated group, our setting becomes closer to a classical non-staggered DiD design, and therefore the negative weights disappear”

OLS extrapolation, without homogeneity, will be biased

Roth (2018) shows that OLS pre-trend estimates are correlated with the estimates of TE obtained from the same dynamic specification

Robust imputation-based estimation and testing

- Robust and efficient estimator from first principles
- Asymptotic analysis
- Consistent and asymptotically normal
- Standard errors and pre-tests

Modifications of general model

- Modification of assumption 1 to A1':

$$Y_{it}^0 = A'_{it}\lambda_i + X'_{it}\delta + \varepsilon_{it}$$

- Assumption 4 is introduced (homoskedastic residuals)

Using A1' to A4, we get the “efficient estimator” which is for all linear unbiased estimates of δ_W , the unique efficient estimator $\widehat{\delta}_W^*$ can be obtained with 3 steps

Role of the untreated observations

"At some level, all methods for causal inference can be viewed as imputation methods, although some more explicitly than others." – Imbens and Rubin (2015)

"The idea is to estimate the model of Y_{it}^0 using the untreated observations and extrapolate it to impute Y_{it}^0 for treated observations."

Steps

- ① Estimate expected potential outcomes using OLS and only the untreated observations (this is similar to Gardner 2021)
- ② Then calculate $\hat{\delta}_{it} = Y_{it}^1 - \hat{Y}_{it}^0$
- ③ Then estimate target parameters as weighted sums

$$\hat{\delta}_W = \sum_{it} w_{it} \hat{\delta}_{it}$$

Testing for parallel trends

- Perform pre-trend testing using untreated sample only
- This separation is preferable conceptually because it presents the conflation of using an identification assumption and validating it
- Traditional regression-based tests use the full sample, including the treated observations though
- Therefore it is not a test for A1 and A2; rather it is a joint test that is also sensitive to A3
- BJS test uses the untreated observations for which Y_{it}^0 is ok under A2

Test

- ① Choose an alternative model for Y_{it}^0 richer than A1

$$Y_{it}^0 = A'_{it}\lambda_i + X'_{it}\beta + w'_{it}\delta + \tilde{\varepsilon}_{it}$$

- ② Estimate δ with $\hat{\delta}$ using OLS on untreated units only
- ③ Test $\delta = 0$ using F-test or visually

Comparisons to other estimators

Table 3: Efficiency and Bias of Alternative Estimators

Horizon	Estimator	Baseline simulation		More pre-periods	Heterosk. residuals	AR(1) residuals	Anticipation effects
		Variance (1)	Coverage (2)				
$h = 0$	Imputation	0.0099	0.942	0.0080	0.0347	0.0072	-0.0569
	DCDH	0.0140	0.938	0.0140	0.0526	0.0070	-0.0915
	SA	0.0115	0.938	0.0115	0.0404	0.0066	-0.0753
$h = 1$	Imputation	0.0145	0.936	0.0111	0.0532	0.0143	-0.0719
	DCDH	0.0185	0.948	0.0185	0.0703	0.0151	-0.0972
	SA	0.0177	0.948	0.0177	0.0643	0.0165	-0.0812
$h = 2$	Imputation	0.0222	0.956	0.0161	0.0813	0.0240	-0.0886
	DCDH	0.0262	0.958	0.0262	0.0952	0.0257	-0.1020
	SA	0.0317	0.950	0.0317	0.1108	0.0341	-0.0850
$h = 3$	Imputation	0.0366	0.928	0.0255	0.1379	0.0394	-0.1101
	DCDH	0.0422	0.930	0.0422	0.1488	0.0446	-0.1087
	SA	0.0479	0.952	0.0479	0.1659	0.0543	-0.0932
$h = 4$	Imputation	0.0800	0.942	0.0546	0.3197	0.0773	-0.1487
	DCDH	0.0932	0.950	0.0932	0.3263	0.0903	-0.1265
	SA	0.0932	0.954	0.0932	0.3263	0.0903	-0.1265

Notes: See Section 4.6 for a detailed description of the data-generating processes and reported statistics.

Big idea

"The main part of the article is about the statistical problem of imputing the missing values of Y. Once these are imputed, we can estimate the causal effect of interest, δ ."

"To estimate average causal effect of the treatment on the treated units, we impute the missing potential control outcomes" – Athey, et al. (2021)

Overview

- Athey, et al. (2021) unites two literatures – unconfoundedness and synthetic control
- Combines computer science with statistics to create the matrix completion with nuclear norm (MCNN) estimator
- Nuclear norm regularization is used for the imputation

What is matrix completion

- Completing a matrix means guessing at the correct values that are missing
- Hence the “completion” is just another name for “filling in” the matrix
- In causal inference, if the matrix is a matrix of potential outcomes (e.g., Y^0), then missingness is caused by treatment assignment

Here's a matrix of potential outcomes, Y^0 , representing units at time t that had not been treated.

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & Y_{12}^0 & Y_{13}^0 & \dots & Y_{1t}^0 \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & Y_{2t}^0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & Y_{i3}^0 & \dots & Y_{it}^0 \end{pmatrix}$$

Now imagine a treatment assignment, SUTVA, that flips treatment from 0 to 1 in the last period t :

$$Y = DY^0 + (1 - D)Y^1$$

Ask yourself: why are there question marks in the last column?

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & Y_{12}^0 & Y_{13}^0 & \dots & ? \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & Y_{i3}^0 & \dots & ? \end{pmatrix}$$

Matrix completion seeks to do the following:

Matrix completion with nuclear norm will impute the last column using regularized regression:

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & Y_{12}^0 & Y_{13}^0 & \dots & \widehat{Y_{1t}^0} \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & \widehat{Y_{2t}^0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & Y_{i3}^0 & \dots & \widehat{Y_{it}^0} \end{pmatrix}$$

And once you have those, you can calculate individual level treatment effects that can be used to aggregate to the ATT

History of matrix completion

- Open competition by Netflix in 2006 – winner would get \$1m if they could improve predictive model by ten points on RMSE
- Invited a ton of competition – from MIT teams to regular everyday joes working out of their home office
- Everyone was given a database which was then tested by Netflix on a holdout dataset
- Quick progress was made followed by very slow advances
- Winner was announced in 2009

Netflix prize

- Gigantic sparsely populated matrix (100m users ranking 100k movies)
- I like Silver Linings Playbook and Lars and the Real Girl and you like Silver Linings Playbook
- Probably you'll also like Lars and the Real Girl
- So we are using correlations in the columns to "complete" missing values
- When you think about it, while it seems predictive (and it is), isn't it really a causal design?
- "If I watch Lars and the Real Girl, will I like it?"

Types of imputation

- I didn't always think of causal inference in terms of imputation because often the method was just taking existing values and manipulating them, rather than filling in missing values
- But the fundamental problem of causal inference states that causal inference is a missing data problem, so it makes sense you'd be imputing
- I tend to think therefore in terms of implicit and explicit imputation methods
- Borusyak, et al. (2021) and Athey, et al. (2021) both seem more like "explicit" imputation methods
- Callaway and Sant'Anna (2020) on the other hand is an implicit method, as is did methods more generally

Two literatures

- Lots of moving parts in this interesting paper, so my goal here is purely explainer and mostly high level at that.
- I want you to be competent and conversant in it so we also have some R code
- There's two literatures they want you to have in your mind:
 - ① Unconfoundedness – $(Y^0, Y^1) \perp\!\!\!\perp D|X$ – sometimes explicitly imputes (nearest neighbor), sometimes more implicit (inverse probability weighting)
 - ② Synthetic control – literally calculating a counterfactual as a weighted average over all donor pool units
- Their MCNN method will show that both are “nested” within the general framework they've developed making them actually special cases

Differences

- Conceptually different in the way they exploit patterns for causal inference
- Unconfoundedness assumes that patterns over time are stable *across units*
- Synth assumes patterns across units are stable *over time*
- Regularization (particularly the nuclear norm) nests them both

The Gist

- Factor models and interactive effects model the observed outcome as the sum of a linear function of covariates and a unobserved component that is a low rank matrix plus noise
- Estimates are typically based on minimizing the sum of squared errors given the rank of the matrix of unobserved components with the rank itself estimated

Three contributions

- ① Formal results for non-random missingness when block structure allows for correlation over time. Nuclear norm is important here
- ② Shows unconfoundedness and synth are in fact matrix completion methods
 - they all have the same objective function based on the Frobenius norm for the difference between the latent matrix and the observed matrix
 - Each approach imposes different sets of restrictions on the factors in the matrix factorization
 - MCNN by contrast doesn't impose any restrictions – just regularization to characterize the estimator
- ③ Applies the method to two datasets, but I'm going to skip that bc I find that stuff tedious once I've muscled my way through a paper like this

Block structure

- Lots of jargon in this article – unconfoundedness, vertical and horizontal regression, fat and thin matrices.
- Unfortunately, you need to learn it all so let me try and organize it
- We define the matrix first in terms of its block structure which is describing where and when the missingness is occurring in the matrix

Unconfoundedness

- Much of the unconfoundedness literature estimates an ATE under unconfoundedness
- But it tends to focus only on a simple setup where the missingness is the last period
- Think about LaLonde (1986) – NSW treats the workers, and then you don't observe Y^0 for the treated group in the *last period*
- This is the “single-treated-period block structure” because only one *period* is missing

Single-treated-period block structure

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & Y_{12}^0 & Y_{13}^0 & \dots & ? \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & Y_{i3}^0 & \dots & ? \end{pmatrix}$$

Single-treated-period block structure

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & Y_{12}^0 & Y_{13}^0 & \dots & Y_{1t}^0 \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & Y_{2t}^0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & ? & \dots & ? \end{pmatrix}$$

Notice, this is the synthetic control design because a single unit (unit i) is missing Y^0 for the 3rd and t th periods.

Staggered adoption

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & ? & ? & \dots & ? \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & ? & \dots & ? \end{pmatrix}$$

So all of these so-called designs can be expressed in terms of missingness in the block structure, and our job therefore is to find an estimator that is general enough to manage all of them. Their MCNN will be that.

Thin and Fat matrices

- We also have to consider the relative number of panel units N and time periods T because this also shapes which regression style will be used for imputation
- Thin matrices are basically where $N \gg T$, but fat matrices are ones where $T \gg N$
- Approximately square ones are where T is approximately equal to N
- My conditionally accepted JHR (fingers crossed) had around 400 panel units and 180 months, so it was $N \gg T$ which is a thin matrix

Vertical and horizontal regression

- Two special combinations of missing data patterns and matrix shape need special attention because they are the focus of large but separate literatures
- Unconfoundedness has that single-treated period block structure with a thin matrix ($N \gg T$).
- You use a large number of units and impute missing potential outcomes in the last period using controls with similar lagged outcomes
- This is the horizontal regression – imagine just running OLS on the lags and taking predicted values
- The horizontal regression holds under unconfoundedness

Vertical regression

Doudchenko and Imbens (2016) and Pinto and Furman (2019) show that Abadie, Diamond and Hainmueller (2011) can be interpreted as regressing the outcomes for the treated prior to treatment on the outcomes for controls in the same period

Fixed effects and factor models

- Both horizontal and vertical regressions exploit other patterns
- An alternative to each of them though is to consider an approach that allows for the exploitation of both stable patterns over time and stable patterns across units
- This is where their matrix completion with nearest neighbor model comes in – it does that very thing

Matrix completion with nuclear norm

- Model the $N \times T$ matrix of complete outcomes data matrix Y as:

$$Y = L^* + e$$

where $E[e|L^*] = 0$

- The error term can be thought of as measurement error if you need a frame to think about it
- So you have this complete matrix, L^* , and zero mean conditional independence holds

Assumption 1

Apart from the unconfoundedness assumption, we have this weird assumption!

Assumption 1

e is independent of L^* and the elements of e are σ -sub-Gaussian and independent of each other

Lots of matrix forms can be defined this way. But let's not get lost in the weeds – we are still just trying to estimate L^* ! That's the main storyline, not the side quest, to use Red Dead Redemption words I understand

All imputations are wrong but some are useful

- You can impute something a million different ways.
- $1 + 1 + 1 + 1 = 4$ is an imputation of the fifth unknown element and frankly just looking at it, seems wrong.
- You could minimize the sum of squared differences but if the objective function doesn't depend on L^* , the estimator would just spit back Y and $\delta = 0$.
- They add a penalty term $||\lambda||$ to the objective function, but even then, not all of them do well.
- Turns out, it actually matters whether you regularize the fixed effects or not (just like it matters whether you regularize the constant in LASSO apparently – I decided to take their word for it)

Estimator

$$L^* = \widehat{L} + \widehat{\Gamma} \mathbf{1}_T^T + I_N \widehat{\Delta}^T$$

where the objective function is:

$$= \arg \min_{L, \Gamma, \Delta} \left\{ \frac{1}{O} \| P_0(Y - L - \Gamma \mathbf{1}_T^T - \mathbf{1}_N \Delta^T) \|_F^2 + \Lambda \| L \| \right\}$$

Fixed effects and regularization

- The penalty will likely be the nuclear norm but notice that the fixed effects are outside the penalty term. You could subsume them into L , they say, but they recommend you not doing this.
- Fraction of observations is relatively high and so the fixed effects can actually be estimated separately (apparently that is one difference between MCNN and the rest of the MC literature)
- The penalty will be chosen using cross-validation

Other norms

- One thing I thought was interesting was that the nuclear norm allowed for the construction of a low rank L^* matrix, but other norms actually would have weird properties
- I remember once me asking Imbens (like I had even a clue what I was talking about), "Why not use elastic net? Why are you using the nuclear norm?" He said elastic net would spit out all zeroes. I remember thinking "Why did I think I would understand what he told me?"
- One advantage of NN is its fast and convex optimization programs will do it, whereas some others won't because of the large N or T issues
- There's almost like a cross walk, too, between this and Borusyak, et al. (2021) but I don't quite see it except they both leverage imputation

Conclusion

- Let's just review the R code. It uses gsynth
- We'll look at Cheng and Hoekstra (2013) again, because frankly it's a dataset I know
- Ultimately, this is just another model though that can be used for differential timing but at the moment, no one knows how it performs in simulations alongside Borusyak, et al. (2021), Callaway and Sant'Anna (2020) or any of the others
- So I can't really answer questions about when to use it and not to – it comes down to these very narrow assumptions
- You choose the estimator based on the problem you're studying and the assumptions – you must justify it, no one else can, but you do so by appealing to assumptions

Sharp DiD

- In a “sharp” DiD, a group gets treated in period 1, a control group does not
- Parallel trends allows you to identify ATT
- We discussed several methods
- But sometimes the lines between treatment and control groups get “fuzzy”

Fuzziness

- In a “fuzzy” DiD design, there’s growth in treatment occurring among units for reasons other than the treatment assignment in the control group
 - They discuss an early 2000s Duflo paper where Indonesia pushed for more primary schooling
 - Used earlier cohorts as controls bc they were already past the age
 - But they saw growth in schools too
- In many applications, the “treatment rate” increase more in some groups than in others but there is no group that goes from fully untreated to fully treated
- But there is no group that also remains fully untreated

Fuzzy estimators

- Popular fuzzy estimator (10% of AERs from 2010-2012) divides DiD of the outcome by the DiD of the treatment

$$Wald_{DiD} = \frac{\left(E[Y_k|Post] - E[Y_k|Pre] \right) - \left(E[Y_U|Post] - E[Y_U|Pre] \right)}{\left(E[D_k|Post] - E[D_k|Pre] \right) - \left(E[D_U|Post] - E[D_U|Pre] \right)}$$

- It's Wald IV in that we scale the reduced form by the first stage but they call it Wald DiD
- de Chaisemartin and D'Haultfoeuille (2017) estimates the LATE for groups who go from untreated to treated

Two proposed estimators

Propose two other estimators

- ① Time corrected Wald ratio, $Wald_{TC}$ – relies on PT within subgroup of units sharing the same treatment at the first date
- ② Changes in changes extension, $Wald_{CiC}$ – extension of Athey and Imbens (2006) “changes in changes” paper. Generalizes CiC to fuzzy. CiC is invariant to outcome scaling but puts restrictions on the full distribution of potential outcomes instead of the mean

Personal takeaway

- Two main values of this paper that I found:
 - Situations where the control group is getting treated with unrelated policy shocks
 - Continuous treatments
- Code to do it is simple but in Stata

Most basic notation

For any random variable, R, we interpret as R_{dgt} as treatment status, treatment group, time

$$R_{101} \sim R | D = 1, G = 0, T = 1$$

Individual treatment status (D) is whether a unit is treated regardless of group; Group (G) is treatment or control *groups*; Time (T) is before or after

Sharp: $D = G \times T$; Fuzzy: $D \neq G \times T$

Cases under consideration

Case 1: Share of treated units in control don't change between periods

$$E[D_{01}] = E[D_{00}]$$

Wald_{DiD} identifies the LATE parameter for “switchers” (i.e., people whose treatment status changed between 0 and 1) if parallel trends holds and if the ATE of treated units at both dates is stable over time; proposes new estimators that don't depend on this

Stable ATE isn't required in a typical “sharp” DiD

Cases under consideration

Case 2: Share of treated units changes over time in control

$$E[D_{01}] > E[D_{00}]$$

Wald_{DiD} identifies the LATE of switchers under PT and stable ATE assumption and LATE of treatment and control group switchers are the same

Under certain assumptions, their alternative estimator will only be partially identified, and it depends on the size of the change of treated units in the control.

Fuzzy design assumptions

A1: Dominating growth of treated units in the treatment group

The treatment group is the one experiencing the larger increase in its treatment rate.

This rules out the case where the two groups experience the same evolution of their treatment rates. Let $R_{gt} \sim R|G = g, T = t$; Assumption 1 implies the following conditions:

$$E(D_{11}) > E(D_{10})$$

$$E(D_{11}) - E(D_{10}) > E(D_{01}) - E(D_{00})$$

Fuzzy design assumptions

A2: Stable percent of treated units in the control group

$0 < E(D_{01}) = E(D_{00}) < 1$ means there is stable percent of treatment units in the control group.

This is a special case where number of treatment units in control group is fixed.

Fuzzy design assumptions

A3: Treatment participation equation

In the treatment group, no one switches from treatment to control.
Formally this is

$$D = 1 \text{ if } V \geq v_{gt} \text{ with every } V \perp\!\!\!\perp T|G$$

Where V is the propensity to get treatment, v_{gt} is a threshold specific to each group/time

A little more notation

- We say a unit is treated as $D(t) = 1\{V \geq v_{gt}\}$
- Switchers are units who go from control to treatment between 0 and 1 $S = \{D(0) < D(1), G = 1\}$
- LATE is for switchers: $\Delta = E(Y_{11}(1) - Y_{11}(0)|S)$
- LQTE is also for switchers: $\tau_q = f_{y_{11}(1)|S(q)}^{-1} - F_{y_{11}(0)|S(q)}^{-1}$

Switcher LATE/LQTE

Why only switchers?

- Sometimes only ones affected are switchers; a policy occurs but only eligibility for some. Switchers end up treated
- Identifying more than the LATE places more restrictions and this already has like 8 assumptions

First estimator: Wald_{DiD}

Commonly used strategy in these fuzzy designs is to normalize the DiD on the outcome by the DiD on the treatment status itself (because remember, in the fuzzy design, units are *becoming* treated as well as *being in treatment groups*)

$$\text{Wald}_{DiD} = \frac{\text{DiD}_Y}{\text{DiD}_D}$$

Wald-DiD

Let $S' = \{D(0) \neq D(1), G = 0\}$ be control group switchers. Then we define relevant parameters as:

$$\begin{aligned}\Delta' &= E(Y_{01}(1) - Y_{01}(0)|S') \\ \alpha &= \frac{[P(D_{11} = 1) - P(D_{10} = 1)]}{DiD_D}\end{aligned}$$

Assumptions

A4: Parallel trends

Standard assumption. Not worth repeating for the millionth time.

Assumptions

A5: Stable treatment effect over time

In both groups, the average effect of going from 0 to d units of treatment among units with $D(0) = d$ is stable over time. This is the same as assuming that among these units, the mean of $Y(d)$ and $Y(0)$ follow the same evolution over time

$$E \left[Y(d) - Y(0) | G, T = 1, D(0) = d \right] - \\ E \left[Y(d) - Y(0) | G, T = 0, D(0) = d \right] = 0$$

for units in the switching population

Assumptions

A6: Homogenous treatment effect over time

Switchers have the same LATE in both groups. This isn't necessary in sharp DiD, just fuzzy

Wald DiD theorems

There's a reason we just listed six assumptions. We need them for this traditional scaled DiD method for fuzzy designs called the Wald DiD. We'll go in order.

Theorem 1: Wald DiD

If A1, A3-A5 hold, then Wald DiD equals

$$\alpha\Delta + (1 - \alpha)\Delta'$$

but if A2 or A6, then Wald DiD equals Δ

Interpretation of theorem 1: case 1

Case 1: when treatment grows in the control group, then $\alpha > 1$.
Then if we assume A1, A3-A5, a lot of things cancel out under A1, A3-A5, but the Wald DiD becomes a weighted *difference* of the LATEs of treatment and control group switchers in period 1.

Since it is a difference in LATEs, then even two positive LATEs can flip sign if the first is less than the second.

But if you assume A6, you just get the LATE.

Interpreting theorem 1: case 2

Case 2: When treatment diminishes in controls, then $\alpha < 1$.

Then under A1, A3-A5, Wald DiD will equal a weighted average of LATEs of treatment and control group switchers in period 1.

This quantity will not reverse signs, but won't equal the LATE without A6.

Interpreting theorem 1: case 3

Case 3: Treatment rate is stable in control, then $\alpha = 1$ and Wald DiD will equal LATE under A1, A3-A5.

This requires that the ATE among units treated at $T=0$ remain stable over time – necessary condition.

Under A1, A3-A4, Wald DiD is equal to LATE plus a bias term involving several LATEs, and unless they cancel out exactly, Wald DiD will be different from the LATE

Alternative estimators

- Wald TC – Time Corrected Wald DiD
- Wald CiC – Changes in changes generalization to fuzzy design

Now we review alternative assumptions under which Wald TC or Wald CiC identify the LATE of switchers in the fuzzy. First let's look at Wald TC which won't depend on A4-A5.

Alternative assumptions for the Wald TC

A4': Conditional parallel trends

This requires $Y(0)$ mean average follow the same trends as all the other groups.

Wald TC estimator

Wald TC equals

$$\frac{E(Y_{11}) - E(Y_{10} + \delta_{D_{10}})}{E(D_{11} - E(D_{10})}$$

where

$$\delta_d = E[Y_{d_01}] - E[Y_{d_00}]$$

which is the change in mean outcome between periods 0 and 1 for controls and treatment status d (not groups T and C – individual units d).

Theorem 2

Theorem 2 and the Wald TC

If A1-A3 and A4', then Wald TC equals Δ

Note that: Wald TC equals

$$\frac{E(Y|G=1, T=1) - E(Y + (1-D)\delta_0 + D\delta_1|G=1, T=0)}{E(D|G=1, T=1) - E(D|G=1, T=0)}$$

This is almost the Wald DiD ratio except for that second term with the $Y + (1 - D)\delta_0 + D\delta_1$ instead of just Y .

This arises because time can independently affect the outcome.

When treatment is stable for a group G , then $\delta_0 = 0$.

Comment on Theorem 2

Wald TC equals

$$\frac{E(Y|G=1, T=1) - E(Y + (1-D)\delta_0 + D\delta_1|G=1, T=0)}{E(D|G=1, T=1) - E(D|G=1, T=0)}$$

The numerator of Wald TC compares the mean outcome in the treatment group in the post period 1 to the counterfactual mean we would have had if switchers had remained untreated.

Then normalized by the change in switching, we get the LATE for switchers

Wald CiC

Here we have continuous outcomes and an estimator for quantiles of the LATE called LQTE. New assumption is complicated but is needed for the Wald CiC

Assumptions for changes in changes Wald ratio

A7: Monotonicity and time invariance of unobservables

Potential outcomes are strictly increasing functions of some scalar unobserved heterogeneity term whose distribution is stationary over time. Also imposes the distribution of that unobserved heterogeneity be stationary within subgroups of units sharing the same treatment status at baseline.

Data restrictions

A8: Data restrictions

First, Y must have the same support in each of the eight $D \times G \times T$ cells (common support). Second, the distribution of Y be continuous with positive density in each of the eight cells.

This will allow us to bound treatment effects (Athey and Imbens 2006). Now the ugliest estimator ever.

Wald CiC estimator

Let $Q(y) = F_{Y_{01}}^{-1} \cdot F_{Y_{00}}(Y)$ be the quantile-quantile transform of Y from period 0 to 1 in the control group. Also let:

$$F_{CiC,d(Y)} = \frac{P(D_{11} = d)F_{Y_{d11}} - P(D_{10} = d)F_{Y_{d10}}}{P(D_{11} = d) - P(D_{10} = d)}$$

And our Wald CiC estimator is:

$$W_{CiC} = \frac{E(Y_{11}) - E(Q_{D10}(Y_{10}))}{E(D_{11}) - E(D_{10})}$$

Theorem 3: Wald CiC

Theorem 3: Wald CiC

Under A1-A3 and A7-A8, then W_{CiC} is the LATE and equivalently we get the LQTE

$$W_{CiC} = \frac{E(Y|G=1, T=1) - E((1-D)Q_0(Y) + DQ_1(Y)|G=1, T=0)}{E(D|G=1, T=1) - E(D|G=1, T=0)}$$

Comment on theorem 3

Almost the standard Wald DiD except for that $(1 - D)Q_0(Y) + DQ_1(Y)$ instead of Y in the second term of the numerator. So again, we are simply making adjustments for the fuzziness but under different set of assumptions. This term accounts for the fact that time directly affects the outcome, but in a CiC setup.

Which to use

It's about choosing your poison. Do you want A4' or A7?

When T and C have different outcome distributions conditional on D in the first period, then scaling of the outcome may have large effect on the Wald-TC. Whereas Wald-CiC isn't sensitive to the scaling of Y.

But when the two groups have similar outcome distributions conditional on D in the first period, Wald-TC may be preferable as A4' only restricts the mean of the potential outcomes, whereas Wald-CiC restricts the entire distribution

Extensions to non-binary, ordered treatment

Theorem 6

Under continuous treatments, the estimators we've been considering are equal to the average causal response parameter that Angrist and Imbens (1995) discuss. This parameter is a weighted average over all values of d of the effect of increasing treatment from $d - 1$ to d for any switchers where treatment status goes from strictly below to strictly above d over time.

Theorem 6 extends to a continuous treatment. Under theorem 6, each of the estimators is identifying a weighted average of the derivative of potential outcomes with respect to changing d

Stata code

Only code I know of at the moment is the fuzzydid the authors published in Stata Journal. But it allows you to specify which estimator. Here's sample code for Wald DiD:

```
fuzzydid lngonf g_decr post1 inverse_fee, did breps  
(1000) cluster(county1)
```

where *g_{decr}* is the treatment group dummy, *post1* is the post period dummy, and *inversefee* is our continuous treatment variable. We specify the Wald DiD by noting *did* after the comma.

Concluding remarks

- Paper is hard but worth it. It's possible your controls are getting treated for unrelated reasons, but this is testable
- The Wald DiD is a conventional approach but suffers bias without a layering in of assumptions
- Alternative estimators for when control group stabilization isn't possible or you don't want to impose treatment effect homogeneity are available
- `fuzzydid` can handle continuous treatments as well as dummies.

← Thread



Analisa Packham
@analisapackham

...

referees who keep suggesting Calloway and Sant'Anna (2019) and Goodman-Bacon (2019) when the treatment happens to everyone in the same year:

plz stop it

10:58 AM · Jul 15, 2021 · Twitter Web App

23 Retweets 16 Quote Tweets 540 Likes



- Differential timing with heterogeneity – Bacon, Callaway and Sant'anna, etc.
- Covariates – Abadie, Sant'Anna and Zhao
- Fuzzy - de Chaisemartin and D'Haultfoeuille

Concluding remarks on DD

- You're probably going to write a paper using DiD at least once in your life, but probably more
- Even if you don't, you're going to read a lot of papers using DiD, referee them, or advise students using them
- It's in your best interest to make the fixed cost investment in the new econometrics of DiD because the old methods are mostly harmful
- Good news is we are at the conclusion of this wave of papers, software is now widely available, solutions tend to have common features, and overall presentations (static and dynamic) aren't all that different

Concluding remarks

- Simple 2x2 has its own problems when estimated using TWFE
if you include covariates
- Stronger assumptions needed to include covariates, and bias can be large
- Don't control for covariates that could be affected by the outcome
- Why pay more for the same car?

Concluding remarks

- Main problem in differential timing is heterogeneity and the use of already-treated units as controls
- If you use TWFE for differential timing, report the Bacon decomposition and report the number of never-treated units
- If you are estimating event studies using TWFE, remember to drop *two* leads to address multiple forms of collinearity (SA; BJS)
- If you have differential timing, consider going directly to one of the robust estimators we discussed
- CS has additional benefits like examining heterogeneous responses by timing – this is part of the value of defining target parameters as weighted averages

Concluding remarks

- Causal claims depends on valid assumptions, high quality and appropriate data, and appropriate estimators
- Use this opportunity to remember how much fun econometrics is
- Don't sweat whether you learned everything in this seminar – check out my substack "Causal Inference: the Remix" for simple explainers, go back to the papers, talk to the authors (they are all very smart, but also extremely kind people)
- Have fun! Remember that applied work is exciting, so don't sweat it. Don't forget how great it is to learn something new
- Don't forget that season 2 of Ted Lasso comes out July 23rd