Causal Inference II

Mixtape Session

MIXTAPE SESSIONS

Roadmap

Continuous DiD

- Dose causal parameter
- Identification
- Selection bias
- Interpreting TWFE

Fuzzy DiD

Concluding remarks

Continuous DiD

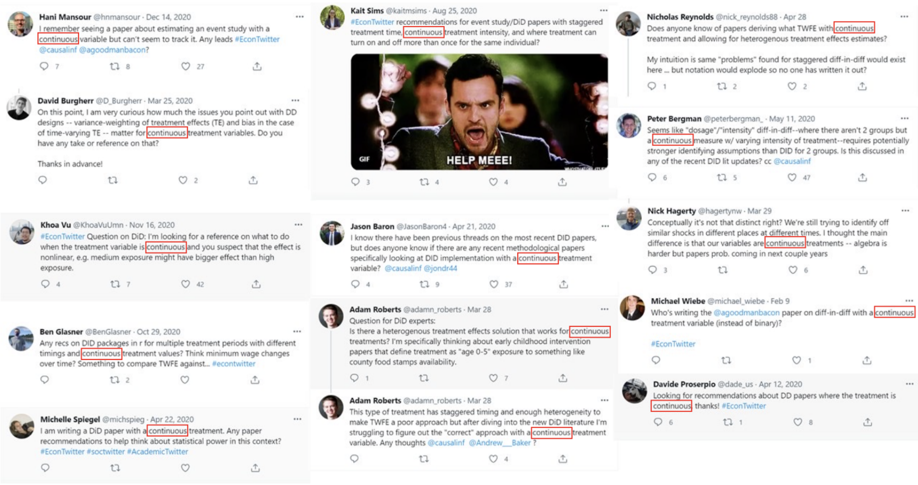
- A very common panel model will use a treatment variable which is continuous, not binary
- Examples include minimum wage papers, my JHR on abortion clinic closures causing increased travel distance, vaccinations, price elasticity of demand etc.
- Variation is in “treatment intensity” and researchers typically use TWFE for estimation, or perhaps count models like Poisson

Quotes

“The two-period regression estimator can be easily modified to allow for continuous, or at least non-binary, treatments.” (Wooldridge 2005)

“A second advantage of regression DiD is that it facilitates the study of policies other than those that can be described by a dummy.” (Angrist and Pischke 2008)

Continuous DiD



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

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

Thanks in advance!
2

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

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

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

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

3 4 4

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

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

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

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

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

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

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

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

#EconTwitter
1

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

Overview

1. What of what we have learned carries forward to the continuous case?
2. Some of the problems with continuous (maybe most) don’t even have to do with differential timing, so I’m not going to cover it

Recommended steps of causal projects

1. Define the parameter we want (“ATT”),
2. Ask what what beliefs do you need (“identification”), and
3. Build cranks that produce the correct numbers (“estimator”)

People often skip 1 and 2 and go straight to 3 and run regressions then go back and assume exogeneity (step 2), and hope that the estimates are weighted averages of individual treatment effects (1), but that is not guaranteed

Dangers of skipping 1 and 2

- TWFE was arguably a case where people skipped 1 and 2 and went straight to 3
- We now know that the “constant treatment effect” static specification does not recover the ATT under parallel trends, but the VWATT and requires no dynamics
- We can see this too in simulations even with matching and regression – defining the parameter ahead of time then clearly indicates what assumptions to make, pushing into specifications
- Let’s look at this now

Data Generating Process

- Covariate imbalance
- Heterogenous treatment effects with respect to covariates
- Linear data generating process
- Question: How do we estimate the ATE vs the ATT?

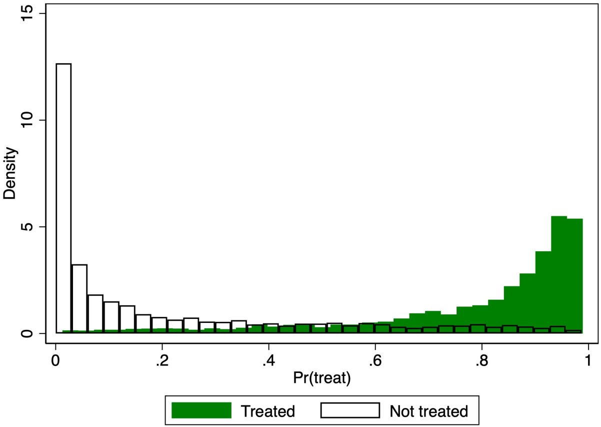
Data Generating Process

- Age is generated from a normal distribution:
  - → Treatment group: mean 25, standard deviation 2.5
  - → Control group: mean 30, standard deviation 3
- GPA is generated from a normal distribution:
  - → Treatment group: mean 1.76, standard deviation 0.5
  - → Control group: mean 2.3, standard deviation 0.75
- Age and GPA are centered around their respective means
- Squared terms and interaction terms are generated:
  - → Age squared (age˙sq), GPA squared (gpa˙sq)
  - → Interaction between age and GPA (interaction)

Data Generating Process

- Outcome variables are generated:
  - → No treatment (y0): $15000 + 10.25 \cdot \text{age} - 10.5 \cdot \text{age\char`\^{}sq} + 1000 \cdot \text{gpa} - 10.5 \cdot \text{gpa\char`\^{}sq} + 500 \cdot \text{interaction} + \epsilon$
  - → Treatment (y1): $y0 + 2500 + 100 \cdot \text{age} + 1000 \cdot \text{gpa}$
  - → Treatment effect (delta): $y1 - y0$
- Average treatment effect (ATE) is estimated at 2500
- Average treatment effect on the treated (ATT) is estimated at 1971

Covariate imbalance (expressed as propensity score)



Regression specifications 1 and 2

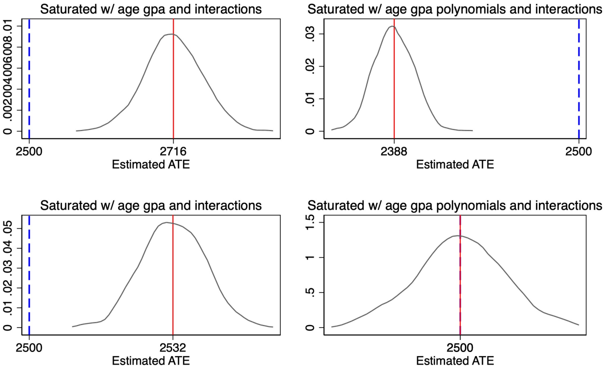
We will look at several different specifications. These first two are standard ways of incorporating covariates. You enter them in linearly as controls.

$$\begin{aligned}
\text{earnings} &= \beta_0 + \beta_1 \text{treat} + \beta_2 \text{age} + \beta_3 \text{gpa} + \epsilon && (1) \\
\text{earnings} &= \beta_0 + \beta_1 \text{treat} + \beta_2 \text{age} + \beta_3 \text{age\_sq} + \beta_4 \text{gpa} \\
&\quad + \beta_5 \text{gpa\_sq} + \beta_6 (\text{gpa} \times \text{age}) + \epsilon && (2)
\end{aligned}$$

Interpretation focuses on $\widehat{\beta_1}$. But what does it mean? Look at top two. Remember ATE is 2500.

ATE estimates across different specifications

OLS Estimates of ATE with heterogenous treatment effects



Four kernel density plots of estimated coefficients from 1000 simulations

ATT Calculation: Regression 3

Saturated regressions: interact treatment dummy with all covariates.

- Regression Specification:

$$\begin{aligned} \text{earnings} &= \beta_0 + \beta_{1t}\text{treat} + \beta_{2a}\text{age} + \beta_{3g}\text{gpa} + \beta_{4ta}(\text{treat} \times \text{age}) + \beta_{5tg}(\text{treat} \times \text{gpa}) + \beta_{6age}(\text{age} \times \text{gpa}) \\ &\quad + \beta_{6tag}(\text{treat} \times \text{age} \times \text{gpa}) + \epsilon \end{aligned}$$

  Estimated ATE is the coefficient, $\widehat{\beta_{1t}}$, but how do we get the estimated ATT?

- Calculating ATT:

$$\text{ATT}_3 = \beta_{1t} + \beta_{4ta} \cdot \bar{\text{age}} + \beta_{5tg} \cdot \bar{\text{gpa}} + \beta_{6tag} \cdot \bar{\text{age}} \cdot \bar{\text{gpa}}$$

  using the means of all covariates

ATT Calculation: Regression 4

Saturated regressions: interact it with covariates, higher order polynomials, and all interactions

- Regression Specification:

$$
\begin{aligned}
\text{earnings} \quad = \quad & \beta_0 + \beta_{1t}\text{treat} + \beta_{2a}\text{age} + \beta_{3a_sq}\text{age\.{}sq} + \beta_{4g}\text{gpa} + \beta_{5g_sq}\text{gpa\.{}sq} + \\
& \beta_{6ta}(\text{treat} \times \text{age}) + \beta_{7ta_sq}(\text{treat} \times \text{age\.{}sq}) + \beta_{8tg}(\text{treat} \times \text{gpa}) + \\
& \beta_{9tg_sq}(\text{treat} \times \text{gpa\.{}sq}) + \beta_{10ag}(\text{age} \times \text{gpa}) + \beta_{11a_sqg}(\text{age\.{}sq} \times \text{gpa}) + \\
& \beta_{12ag_sq}(\text{age} \times \text{gpa\.{}sq}) + \beta_{13a_sqg_sq}(\text{age\.{}sq} \times \text{gpa\.{}sq}) + \epsilon
\end{aligned}
$$

Estimated ATE is the coefficient, $\widehat{\beta_{1t}}$, but how do we get the estimated ATT?

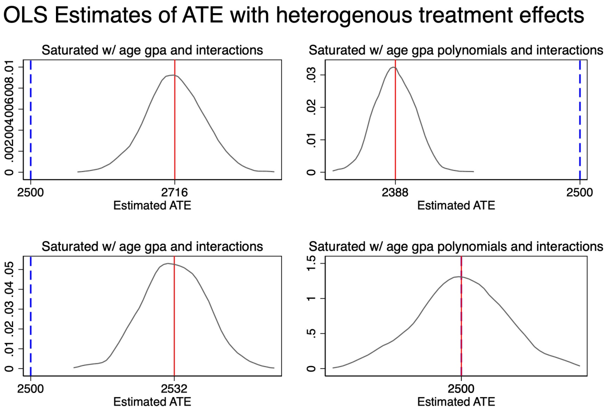
- Calculating ATT:

$$
\begin{aligned}
\text{earnings} \quad = \quad & \beta_0 + \beta_{1t}\text{treat} + \beta_{2a}\text{age} + \beta_{3a\_sq}\text{age\_sq} + \beta_{4g}\text{gpa} + \beta_{5g\_sq}\text{gpa\_sq} + \beta_{6ta}(\text{treat} \times \text{age}) \\
& + \beta_{7ta\_sq}(\text{treat} \times \text{age\_sq}) + \beta_{8tg}(\text{treat} \times \text{gpa}) + \beta_{9tg\_sq}(\text{treat} \times \text{gpa\_sq}) + \beta_{10ag}(\text{age} \times \text{gpa}) \\
& + \beta_{11a\_sqg}(\text{age\_sq} \times \text{gpa}) + \beta_{12ag\_sq}(\text{age} \times \text{gpa\_sq}) + \beta_{13a\_sqg\_sq}(\text{age\_sq} \times \text{gpa\_sq}) + \epsilon
\end{aligned}
$$

Interpretations

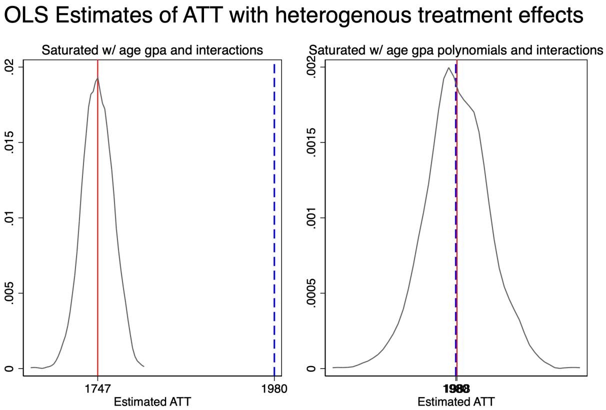
- ATE is 2500
- ATT is 1980
- Key point here: the same regression contains both parameters, but only when done correctly, and only when interpreted correctly

ATE estimates across different specifications



Four kernel density plots of estimated coefficients from 1000 simulations

ATT estimates across different specifications


OLS Estimates of ATT with heterogenous treatment effects

Saturated w/ age gpa and interactions

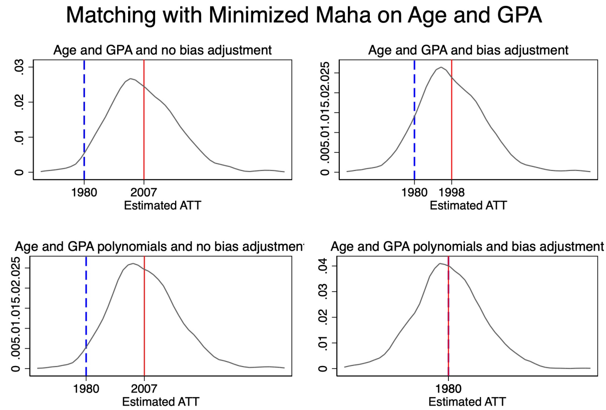
Saturated w/ age gpa polynomials and interactions

Two kernel density plots of estimated coefficients from two regressions and 1000 simulations

Matching by minimizing Maha distance metric

- Next I just estimated the ATT using a simpler model – nonparametric matching using Abadie and Imbens (2006;2011)
- It’s biased when the matches aren’t exact, but you can use regression adjustment to estimate the selection bias
- Similar to what augmented synth does
- Much less difficult syntax

ATT estimates across different matching specifications



Four kernel density plots of estimated ATT from 1000 simulations

Recommended steps of causal projects

1. Define the parameter we want (“ATT”),
2. Ask what what beliefs do you need (“identification”), and
3. Build cranks that produce the correct numbers (“estimator”)

See how when we skep 1 and 2 and go straight to 3, heterogenous treatment effects makes major problems for interpretation? It isn’t that regressions can’t recover parameters, but you have to saturate when you’re attempting to recover ATE or ATT, and even then it’s challenging to interpret – and programming, you often don’t have code that will do it for you.

Introducing a new causal parameter

- **ATT**: Extensive margin causal parameter. Do this versus don’t do this.
- **Dose**: Intensive margin causal parameter. Do this much versus this much.

The dose causal parameter will be based on Angrist and Imbens (1995)

Parameters

Average treated on the treated

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

while the treatment, D, can be any amount, d, that amount is technically a particular dose. We raised the minimum wage, but we raised it to a particular wage.

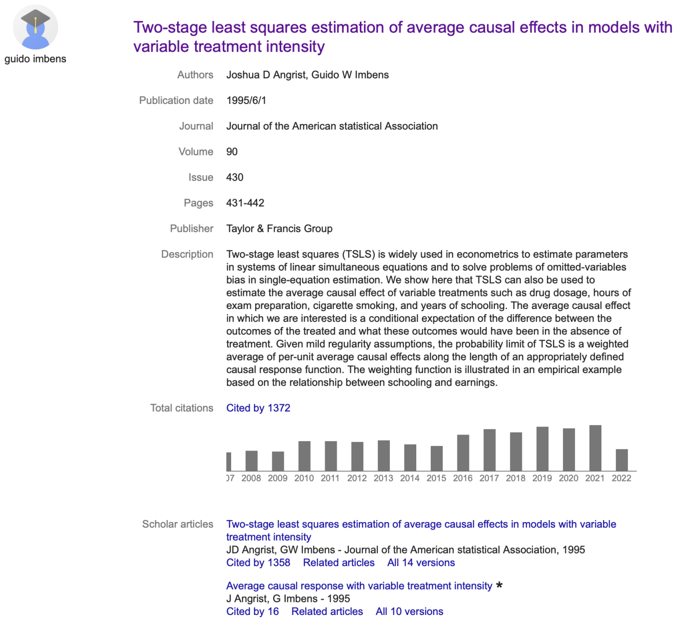
Parameters

Average treated on the treated

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

This is “the ATT of d for the groups that chose d dosage” which uses as its comparison no dose.

Average causal response function



guido imbens

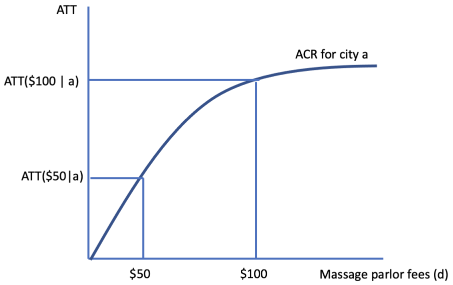
Two-stage least squares estimation of average causal effects in models with variable treatment intensity

Authors	Joshua D Angrist, Guido W Imbens
Publication date	1995/6/1
Journal	Journal of the American statistical Association
Volume	90
Issue	430
Pages	431-442
Publisher	Taylor & Francis Group
Description	Two-stage least squares (TSLS) is widely used in econometrics to estimate parameters in systems of linear simultaneous equations and to solve problems of omitted-variables bias in single-equation estimation. We show here that TSLS can also be used to estimate the average causal effect of variable treatments such as drug dosage, hours of exam preparation, cigarette smoking, and years of schooling. The average causal effect in which we are interested is a conditional expectation of the difference between the outcomes of the treated and what these outcomes would have been in the absence of treatment. Given mild regularity assumptions, the probability limit of TSLS is a weighted average of per-unit average causal effects along the length of an appropriately defined causal response function. The weighting function is illustrated in an empirical example based on the relationship between schooling and earnings.
Total citations	Cited by 1372
Scholar articles	Two-stage least squares estimation of average causal effects in models with variable treatment intensity JD Angrist, GW Imbens - Journal of the American statistical Association, 1995 Cited by 1358 Related articles All 14 versions Average causal response with variable treatment intensity * J Angrist, G Imbens - 1995 Cited by 16 Related articles All 10 versions

Angrist and Imbens 1995

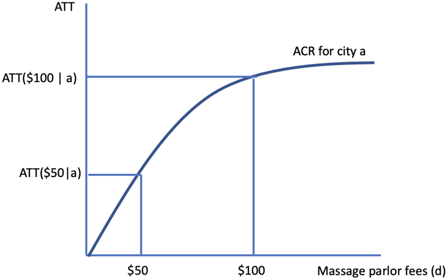
*“We refer to the parameter β as the **average causal response (ACR)**. This parameter captures a weighed average causal responses to a unit change in treatment, for those whose treatment status is affected by the instrument. . . . ”*

ATT for a given dose



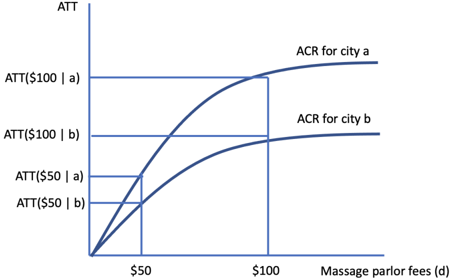
What is the effect of setting fees to $100 versus nothing at all? It's ATT($100—a) for this city.

ATT for a given dose



Assume city a did choose $d = \$100$. Then ATT($50—a) just means that that is its ATT *had* it chosen the lower level. The curve, in other words, is tracing out all average causal response for this city.

ATT for a given dose



What if everyone has different responses? In other words, city a has the higher curve than city b. Then there are several comparisons possible. What is the effect of $50 on outcomes for cities that actually chose $50 versus those than actually chose $100?

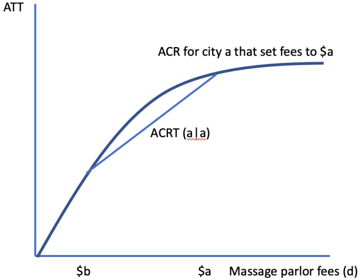
Parameters

Average treated on the treated

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

Notice that you are comparing any dose d to no treatment at all – sort of an extensive margin causal response, but that isn't the only causal concept we have. Elasticities are causal, demand curves are causal, but they aren't based on comparisons to nothing – they are intensive margin comparisons, local comparisons, adjacencies. Zero isn't the only counterfactual in other words.

Average causal response for discrete case vs continuous



What is the ACRT?

- ACRT is the causal effect of dose $D = d_j$ vs a different dose $D = D_{j-1}$ for group d
  - → Easiest example is the demand function: at $p = \$10$, I buy 10 units, but at $p = \$11$, I buy 5 units.
  - → Causal effect of that one dollar increase is -5 units
  - → Demand curves are pairs of potential outcomes and treatments and equilibrium “selects” one of them
- Discrete/multi-valued treatment is linear difference between two ATTs for the same city
- Continuous treatment is the derivative of the function itself

Identification for two period set up

1. Random sampling.
2. No anticipation
3. Parallel trends in Y^0 for units of all doses

Identifying $ATT(d|d)$

We can estimate the $ATT(d|d)$ using the simple DiD equation:

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

No anticipation and parallel trends converts this comparison of before and after into the $ATT(d|d)$

$ATT(d|d)$ is using as its counterfactual the “no treatment”, note. Treatment is a dosage compared to zero iow.

Identifying ACRT

$$\begin{aligned} ATT(b|b) - ATT(a|a) &= (E[\Delta Y_{it}|D_i = a] - E[\Delta Y_{it}|D_i = 0]) \\ &\quad -(E[\Delta Y_{it}|D_i = b] - E[\Delta Y_{it}|D_i = 0]) \\ &= E[\Delta Y_{it}|D_i = a] - E[\Delta Y_{it}|D_i = b] \end{aligned}$$

Comparing high and low dose groups.

Identifying ACRT

$$\begin{aligned} ATT(d_j|d_j) - ATT(d_{j-1}|d_{j-1}) &= \\ (ATT(d_j|d_j) - ATT(d_{j-1}|d_j)) + (ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})) &= \\ (ACRT(d_j|d_j)) + (ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})) &= \end{aligned}$$

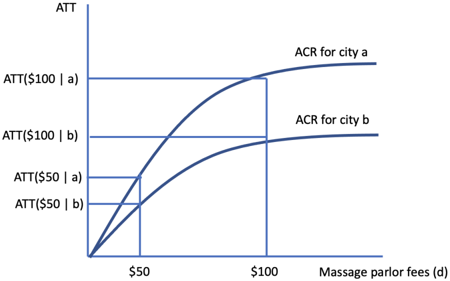
Part in blue is the movement along the average causal response function, the ACRT, and is causal. The part in red is selection bias.

Identifying ACRT

$$\begin{aligned} ATT(d_j|d_j) - ATT(d_{j-1}|d_{j-1}) &= \\ (ATT(d_j|d_j) - ATT(d_{j-1}|d_j)) + (ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})) &= \\ (ACRT(d_j|d_j)) + (ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})) &= \end{aligned}$$

Notice parallel trends allows to identify ATT terms but we need additional assumptions for this red part to vanish. We must assume that the ATT for cities that chose d_j and cities that chose d_{j-1} are the same had they both chose d_{j-1}.

Causality and selection bias



Draw the ACRT for top curve and selection bias.

Interpreting this

- Unrestricted heterogenous treatment effects (across dosage levels and across units with difference dose response functions) is not itself the problem
- If we randomized dosages, then $ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1}) = 0$
- Why? Because then there is no selection on gains from dosages, and average causal response functions are the same for all dosage groups
- So then when is this a problem? Sorting on gains

Interpreting this

- When estimating treatment effects using continuous DiD, you will need to make one of two assumptions
    1. Strong parallel trends: Average change in $E[Y^0]$ for the entire sample is the same as the d group
    2. Parallel trends plus homogenous treatment effect functions
- Roy model like sorting on gains typically lead to violations of the second condition insofar as there is heterogenous returns to dosages across units
- So the question you have to ask yourself is do you think that cities are "optimally setting the minimum wage" around some given minimum wage?

Stronger assumption

- I’m really not so sure I think that when it comes to state legislation that I think a Roy model is likely responsible for the equilibrium
- Solving constrained optimization problems is hard and unlikely is it the case that Florida’s ATT and Georgia’s ATT are terribly different from one another had both chosen the same minimum wage (but that is the bias)
- But notice where this is taking us – we are getting closer and closer back to a place of assuming away the problems!

Interpretation of estimate

- So we are back to interpretation – what is the interpretation, then, of $\widehat{\delta}^{TWFE}$ with continuous DiD?
- Callaway, Goodman-Bacon and Sant'anna (2022) have a decomposition which I encourage you to read

TWFE as weighted average of ACRT parameters

$$\widehat{\delta}^{TWFE} = \int_{d=L}^{d=U} w_1(l) ACRT(l), dl + w_0 \frac{ATE(d_L)}{d_L}$$

- They assume strong parallel trends and find that TWFE coefficient is weighted average of ACRT function
- All weights are non-negative and sum to one
- But weights are strange in that they are maximized at dosage equal to the mean dosage of the entire dataset

New estimator

- There is a new estimator in their new version, but I haven’t read it yet
- Today I just wanted to emphasize the basic principles though – define the parameter, figure out the assumptions, then build the estimators
- Just running regressions doesn’t work once we have heterogenous treatment effects

Conclusion

- Very interesting area, a bit challenging, my suggestion is make clear your assumptions
- Most intuitive for me is the parallel trends plus homogenous treatment effects across units around the dosages
- But interpretations are hard because you're having to think about a new parameter, and you could have nonlinearities in the ACRT that since TWFE is a weighted average of them, could flip sign
- Not a problem but could be a challenge for you as you try to make sense of it (particularly given how the weights are)

Roadmap

Continuous DiD

- Dose causal parameter
- Identification
- Selection bias
- Interpreting TWFE

Fuzzy DiD

Concluding remarks

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

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

- 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 hols 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:

$$\begin{aligned} E(D_{11}) &> E(D_{10}) \\ E(D_{11}) - E(D_{10}) &> E(D_{01}) - E(D_{00}) \end{aligned}$$

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: $\tau_q = f^{-1}_{y_{11}(1)|S^{(}q)} - F^{-1}_{y_{11}(0)|S^{(}q)}$

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*

$$Wald_{DiD} = \frac{DiD_Y}{DiD_D}$$

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

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

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 $Q(y) = F^{-1}_{Y_{01}} \cdot F_{Y_{00}}(Y)$ 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 $inverse_fee$ 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.

Roadmap

Continuous DiD

- Dose causal parameter
- Identification
- Selection bias
- Interpreting TWFE

Fuzzy DiD

Concluding remarks

Empirical micro model

- In 2014, David Card gave a speech at Michigan in which he told the history of the "empirical micro model", particularly labor
- He notes the different role that economic models have played throughout the history of modern applied (empirical) work
    1. Simple approximations of the theoretical model (regress earnings on education)
    2. Structural modeling (often highly parametrized models)
    3. Princeton (essentially this entire workshop) where focus is on physical (manipulated) treatment assignment
- Models are valuable even under #3, but largely to suggest question, interpret results – not to solve the identification problem

Unrestricted heterogeneity and selection on gains

- Over time, applied economists (at least the modal one) is far less likely to believe they know enough about treatment assignment or the underlying treatment effects that they are willing to make any theoretical restrictions
- Treatment effects come from unknown production functions, and whether we have heterogeneity across treatment groups involves selection, most likely from a Roy model type of situation
- By and large, this is my conjecture, economists are either unwilling to restrict heterogeneity ex ante, and the only models many of us are by and large committed to are Roy like rational sorting models
- Without restrictions on heterogeneity ex ante, Roy models imply sorting on gains which create the very problems we find

Consequences of being “deeply a-theoretical”

- Insofar as people sort on the gains from treatment, then given unknown and unrestricted heterogeneous treatment effects, we do not know enough to rule out situations like those in our “baker” dataset
- Baker dataset was extreme, but before this literature, it’s not clear anyone would’ve known it was extreme
- Linear dynamics that never stabilized, large treatment effects in the ATT, heterogenous treatment profiles
- TWFE in the canonical specification performs poorly and in our case flipped sign

Decision making under uncertainty

- If you cannot restrict heterogeneity, and Roy like models are driving selection into treatment, then there's risks
- TWFE will be efficient if model is correctly specified, which requires no dynamic treatment effects, and most of us when pushed cannot credibly claim we have enough prior knowledge to say that as we are often working on something for which underly production technology around treatment effects are not well known or external validity is unknown
- You will need to take this seriously, and adjust practices *away* from naive TWFE modeling
- You are, though, the first wave of researchers to have to, so you are unfortunately caught in a bit of limbo moment

Future work

- We are going to cover more differential timing modeling, but I think some of this is a bit redundant
- We start to see similar fixes to the underlying problem performed in different, though internally valid, ways
- Much of this is a bit cloudy to even the most up-to-date practitioner – differences between estimators are based on which control groups are used, whether PT holds for those control groups, and how best to think about the parallel trends assumption (some of which have stronger additive forms imposing pre-treatment parallel trends and others which do not)
- Cherry picking diff-in-diff could be on the horizon, and this workshop is meant to prepare you for this situation as worst case scenario you will be reading referee reports of people using methods which you may yourself think are inferior but which you still need to understand
- You will need to continue to invest in good judgment, as always; most likely going forward returns to econometric skill has shifted up

Suggestions

- These papers have made the careers of young econometricians to be honest – the growth in citations in short period of time is very weird for econometrics papers (even the IV papers by Angrist and Imbens don't show this pattern)
- Think of a monopolistic competition model – entry stops when zero profits at the margin
- **Profits are shrinking but not zero at the margin**
- There's a lot of activity, it will continue, and you'll need to allocate your time accordingly
- Good luck with your research and pursuit of meaning and happiness – don't lose sight of what's important