# Difference-in-Differences

Mixtape Session

MIXTAPE SESSIONS

# Roadmap

Non-binary treatments
    Fuzzy DiD

# 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

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

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

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 $\Delta$

# 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

If A1-A3 and A4', then Wald TC equals $\Delta$

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

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 $inverse_{f}ee$ 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.