Applied Microeconometrics

Lecture 6

Manuel Bagues

August 2022
Lecture Slides

Previous lecture:

- Canonical model
- More than two periods and variation in treatment timing
    - When treatments are heterogeneous across units or time $\rightarrow$ TWFE estimate does not have a meaningful interpretation
- Two problems:
    - Arbitrary weights (can be even negative)
    - 'Forbidden' comparisons (using already treated units)
- Dynamic TWFE: including lags and leads
    - Works if only heterogeneity in time since treatment...
    - ...but fails in the presence of heterogeneous treatment effects across adoption cohorts.
    - Note: this problem affects also evaluation of pre-trends

- Diagnostic approaches: assess how relevant is the problem!
    - Chaisemartin and D'Haultfoeuille (2020)
        - heterogeneity in treatment effects that would reverse the sign of the estimate
    - Goodman-Bacon (2021)
        - report weights for each group of comparisons: how much weight on 'forbidden' comparisons?
    - Jakiela (2021)
        - negative weights?
        - test the constant treatment assumption
- Several new estimators for staggered DID, with some common features:
    - Use only 'clean' comparisons between treated and non-treated groups
    - Aggregate them using some type of user-specified weights
- Let us see one of them in practice: Callaway and Sant'Anna (2021)

E.g. Callaway and Sant'Anna (2021)

- Let us assume parallel trends, no anticipation, and SUTVA
- Two possible control groups:
  - never-treated units
  - all not-yet treated units
- Several options for weights available
- When the number of periods and groups is small you may report all relevant ATT(g,t)
- Example from problem set 3:
  - Impact of minimum wage on teen employment
  - Sample of US counties, years 2003-2007, N=2,500
  - Some states treated in 2004, 2006 and 2007

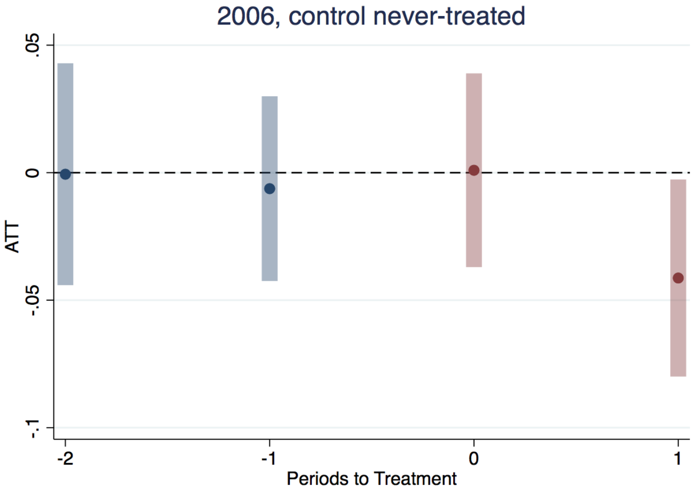
```
. csdid lemp lpop, i(countyreal) t(year) g(first_treat)
............
Difference-in-difference with Multiple Time Periods

                                                Number of obs = 2,500
Outcome model  : weighted least squares
Treatment model: inverse probability tilting
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
g2004						
t_2003_2004	-.0145329	.0221264	-0.66	0.511	-.0579	.0288341
t_2003_2005	-.0764267	.0286661	-2.67	0.008	-.1326113	-.0202422
t_2003_2006	-.1404536	.035373	-3.97	0.000	-.2097834	-.0711239
t_2003_2007	-.1069093	.0328863	-3.25	0.001	-.1713652	-.0424533
g2006						
t_2003_2004	-.0006112	.022198	-0.03	0.978	-.0441185	.0428961
t_2004_2005	-.006267	.018481	-0.34	0.735	-.042489	.0299551
t_2005_2006	.0009473	.0193812	0.05	0.961	-.0370391	.0389337
t_2005_2007	-.0413123	.0197171	-2.10	0.036	-.0799571	-.0026674
g2007						
t_2003_2004	.0266993	.0140628	1.90	0.058	-.0008633	.0542619
t_2004_2005	-.0045906	.0157101	-0.29	0.770	-.0353818	.0262007
t_2005_2006	-.0284515	.0181775	-1.57	0.118	-.0640787	.0071758
t_2006_2007	-.0287821	.0162333	-1.77	0.076	-.0605988	.0030347

Control: Never Treated

Plot results for the 2006 group:

csdid_plot, group(2006)



- Aggregate result for ATE

```
. // Aggregate result for ATE
. estat simple
Average Treatment Effect on Treated
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
ATT	-.0417577	.0115008	-3.63	0.000	-.0642989	-.0192165

- Test for pretrend

```
. // Test for pretrend
. estat pretrend
Pretrend Test. H0 All Pre-treatment are equal to 0
chi2(5) =     6.8436
p-value  =     0.2325
```

Other estimators:

- Chaisemartin and D'Haultfoeuille (2020)
  - Can be applied when the treatment switches on and off
  - Stata command: twowayfeweights
- Sun and Abraham (2021)
  - last-to-be-treated as control group
- Alternative solution: stacked regression

Stacked differences-in-differences: Steps

Cengiz, Dube, Lindner, and Zipperer (2019)

1. Create separate datasets for each treatment-cohort g.
2. Keep all units treated in that cohort, and all units that are not treated by year $g+k$ where g is the cohort-treatment year and k is the outermost relative year that you want to test (e.g. if you do an event study plot from -5 to 5 , would equal 5).
3. Keep only observations within years $g-k$ and $g+k$ for each cohort-specific dataset, and then stack them in relative time.
4. Append all cohort-specific datasets together.
5. Run the same TWFE estimates as in standard DiD but include interactions for the cohort-specific dataset with all of the fixed effects, controls, and clusters.

Shortcoming: Prevents negative weighting but shorter-run estimates and less statistical power (smaller sample)

Stacked differences-in-differences: Application

Cengiz, Dube, Lindner, and Zipperer (2019)

- Impact of minimum wage changes in US on low-wage jobs across a series of 138 state-level minimum wage changes between 1979-2016.
- 138 event h-specific datasets including the outcome variable and controls for the treated state h and all other 'clean controls states' in timeframe (-3 to +4)
- For each event, run a 'single treatment' diff-in-diff:
- Comparing only switchers to not (yet) treated units (drop already treated states).

Summary

- Different heterogeneity-robust DID methods available (see Table 2 in Roth, Sant'Anna, Bilinski and Poe 2022)
- Which one? Trade-off between efficiency and required assumptions
- Typically very similar results

Other issues: non-parallel trends

- Causal interpretation of DD valid only under "parallel trends" assumption
- Untestable: parallel trends in the past provide only supportive evidence, they are neither a necessary nor a sufficient condition
  - Example by Kahn-Lang and Lang (2020): boys' and girls' height follow parallel trends until age 13, but this does not imply that bar-mitzvahs (for boys at age 13) affect height
- If groups differ in levels, why should we expect similar trends?
  - Example: liberal states tend to adopt certain policies and they may be exposed to different shocks

  $\rightarrow$ Similarity in levels, not only trends, makes common trends assumption more plausible: *why* do levels differ, and can the same mechanism affect trends?
- Parallel trends assumption sensitive to the functional form assumption
  - If levels (or distribution) differs, functional form matters, and implies a different counterfactual - should be theoretically justified.
  - Example: levels vs. log.
- Parallel trends conditional on covariates

Other issues: non-parallel trends

Three problems:

1. Absence of a statistically significant pre-trend does not necessarily imply that parallel trends hold
2. Conditioning on the result of a pre-test can introduce pre-test bias
3. Even if a significant pre-trend is observed, we may want to learn something from the data.

Solutions:

1. Increase the power of pre-tests
    - Power calculations
    - Reverse the role of the null and the alternative hypotheses: e.g. test the null of pre-existing trends sufficient to eliminate effect
2. Bounds approach: post-treatment violation of parallel trends assumed to be no larger than maximal pre-treatment violation
3. Adjust your S.E. for pre-testing as in Roth (2019)

Roth (2019)

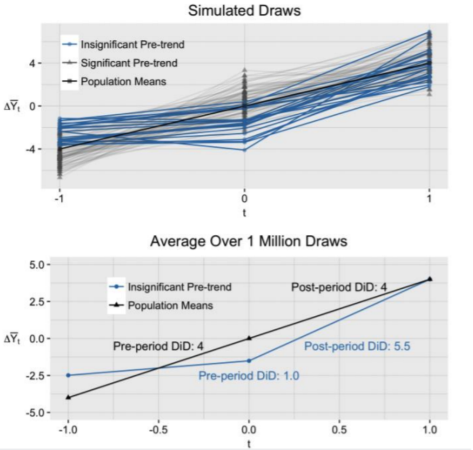
- Pretrend tests are often underpowered
- Reporting DID effects conditional on surviving a pre-trend test of introduces a pre-testing problem, which can exacerbate the bias from an underlying trend, and lead to wrong CI
- If a pre-trend truly exists, then with sample noise, cases leading to non-rejection of parallel trends in the pre-period would also have stronger difference in the post, resulting in an overstatement of the TE ('mean-reversion')

Pre-trends: Power issues - Roth 2019

True causal effect is 0 ($y_{it}(1) = y_{it}(0)$), and true model is:

$$y_{it}(0) = \alpha_i + \phi_t + D_i \times g(t) + \epsilon_{it} \tag{1}$$

With underlying upward trend $g(t) = \gamma t$



Pre-trends: Power issues, take-aways from simulation

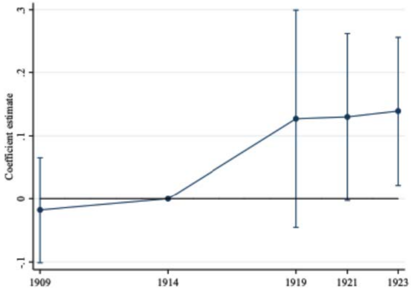
Roth 2019

- When there is an underlying trend, pre-trends testing exacerbates bias.
- Statistical noise in finite sample may prevent detecting trend
- Blue draws would not detect a pre-trend
- True slope between -1 and 0 would be $-\beta_{-1}$, and β between 0 and 1, but in the blue ones $\beta = 0$
- If we get these draws (the cases where we fail to detect the underlying trend), we will produce large treatment estimates because of this failure.
- $\rightarrow$ "Passing" the pre-trends test, paradoxically leads to more biased estimates.

How to proceed?

- Parametric approaches: impose a structure for differential trends (e.g. linear), control parametrically for it without pre-testing
- Alternative relaxations of parallel trends assumptions: Rambarachan Roth (2019), Freyaldenhoven et al. (2019) - but provide valid inference only from an ex ante sampling perspective, not conditional on passing a pre-test.

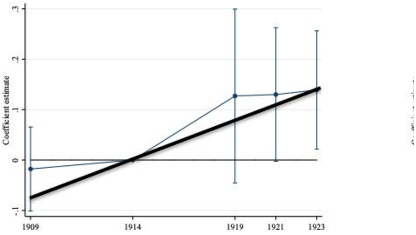
Example: Did non-pharmaceutical interventions (lockdowns) increased growth post1918 Flu? (CLV 2020)



(a) NPI intensity and log manufacturing employ-

Example: Did non-pharmaceutical interventions (lockdowns) increased growth post1918 Flu? (CLV 2020)

Could a linear difference in trends explain the evidence?



(a) NPI intensity and log manufacturing employment.

(b)

Example: Did non-pharmaceutical interventions (lockdowns) increased growth post1918 Flu? (CLV 2020)