Difference-in-differences Part II

Shuowen Chen

BW Econometrics Reading Group (06/08/2023)

DID in a nutshell

A method to estimate the effect of a policy (treatment) on an outcome by comparing over time groups experiencing different evolutions to their exposure to the policy

Potential outcomes framework for causal inference

Two units, two periods

- $y_{i,t}(0,0)$: unit i's potential outcome at t if remains untreated in both periods
- $y_{i,t}(0,1)$: unit i's potential outcome at t if untreated at $t = 1$ but treated at $t = 2$
- potential outcomes correspond with a path of treatments

Potential outcomes framework for causal inference

Two units, two periods

- $y_{i,t}(0,0)$: unit i's potential outcome at t if remains untreated in both periods
- $y_{i,t}(0,1)$: unit i's potential outcome at t if untreated at $t=1$ but treated at $t=2$
- potential outcomes correspond with a path of treatments

Fundamental problem of causal inference: missing data

- cannot compute unit–level treatment effect
- for each i, only observe $y_{i,t}(0,0)$ or $y_{i,t}(0,1)$
- observed outcome for i:

$$y_{i,t} = D_i y_{i,t}(0,1) + (1 - D_i) y_{i,t}(0,0)$$

Estimand of interest

- average treatment effect on the treated (ATT)

$$\tau := \mathbb{E}[y_{i,2}(0,1) - y_{i,2}(0,0) | D_i = 1]$$

- challenge: $y_{i,2}(0,0)$ is not observable for any treated unit
  - ATT is not identified
- DID point identifies ATT by imposing additional assumptions

DID assumptions and ATT

Parallel trend:
In the absence of the treatment, both the treated and the control groups would have experience the same outcome evolution

$$\mathbb{E}[y_{i,2}(0,0) - y_{i,1}(0,0)|D_i = 1] = \mathbb{E}[y_{i,2}(0,0) - y_{i,1}(0,0)|D_i = 0]$$

No anticipatory effects:
The treatment has no causal effect prior to its implementation

$$y_{i,1}(0,0) = y_{i,1}(0,1) \text{ for all } i \text{ with } D_i = 1$$

DID assumptions and ATT

Parallel trend:
In the absence of the treatment, both the treated and the control groups would have experience the same outcome evolution

$$\mathbb{E}[y_{i,2}(0,0) - y_{i,1}(0,0)|D_i = 1] = \mathbb{E}[y_{i,2}(0,0) - y_{i,1}(0,0)|D_i = 0]$$

No anticipatory effects:
The treatment has no causal effect prior to its implementation

$$y_{i,1}(0,0) = y_{i,1}(0,1) \text{ for all } i \text{ with } D_i = 1$$

Average treatment effects on the treated:

$$\tau = \underbrace{\mathbb{E}[y_{i,2}(0,1) - y_{i,1}(0,1)|D_i = 1]}_{\text{treated group difference}} - \underbrace{\mathbb{E}[y_{i,2}(0,0) - y_{i,1}(0,0)|D_i = 0]}_{\text{control group difference}}$$

▸ Derivations

Implementation: two-way fixed effect estimation

Estimating ATT:

$$\widehat{\tau} = \left(\overline{y}_{t=2,D=1} - \overline{y}_{t=1,D=1}\right) - \left(\overline{y}_{t=2,D=0} - \overline{y}_{t=1,D=0}\right)$$

Linear regression specification

$$y_{i,t} = (\mathbf{1}[t = 2] \cdot D_i) \times \beta + \alpha_i + \delta_t + \varepsilon_{i,t}$$

- $y_{i,t}$: outcome of i at period t
- D_i: treatment status
- $\mathbf{1}[t = 2]$: post–treatment indicator
- α_i: individual-specific characteristics that are unobservable but fixed over time
- δ_t: period-specific characteristics that are unobservable but common among all is

▸ Computing SE

Staggered adoption

A special kind of treatment timing

- units adopt a treatment of interest at different points in time
  - different states expanded Medicaid in different years
- absorbing state: once a unit is treated, it remains treated

Rules out the case in which treatment can be on or off

Notations of treatment status

- $D_{i,t}$: an indicator whether i is treated at t
- $G_i := \min\{t : D_{i,t} = 1\}$
  - the earliest period in which i is treated
  - if i is never treated in the data sample: $G_i = \infty$
- under absorbing state: $D_{i,t} = 1$ for all $t \geq G_i$

Potential outcomes framework

recall potential outcomes correspond to a path of outcomes

- $y_{i,t}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$: unit i's potential outcome at t if it is first treated at time g
- $y_{i,t}(\mathbf{0}_T)$: potential outcome at t if never treated
- under absorbing state: the entire path of potential outcomes can be summarized by g
  - $y_{i,t}(g) \equiv y_{i,t}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$
  - $y_{i,t}(\infty) \equiv y_{i,t}(\mathbf{0}_T)$
- treatment effect at t for i that are treated at g:

$$\tau_{i,t}(g) := y_{i,t}(g) - y_{i,t}(\infty)$$

▸ A parallel trend assumption   ▸ No anticipatory effects

TWFE SPECIFICATIONS

Static:

$$y_{i,t} = \alpha_i + \delta_t + D_{i,t}\beta + \varepsilon_{i,t}$$

- β: overall treatment effects across groups and time periods

Dynamic:

$$y_{i,t} = \alpha_i + \delta_t + \sum_{r \neq 0} \mathbf{1}[R_{i,t} = r]\beta_r + \varepsilon_{it}$$

- $R_{i,t} = t - G_i + 1$: the time relative to treatment
- G_i: the earliest period in which i is treated

Outline of the presentation

- Pitfalls of TWFE estimator in static specifications
  - source of the issue
  - examples for illustration

Static TWFE estimator: DID decomposition

Theorem 1 in Goodman-Bacon (2021)

$$\widehat{\beta} = \sum_{m \neq m', t < t'} v_{m,m',t,t'} DID_{m,m',t,t'}$$

- weighted average of DID comparisons between any pairs (m and m') and time periods (t and t')
- $v_{m,m',t,t'} \geq 0$ and sums up to 1
  - $v_{m,m',t,t'} > 0$ when group m changed its treatment status (treated group) and the other m' did not ("control" group)
  - proportional to group sizes and treatment variations
- two types of "control" group
  - group that is not treated at both t and t'
  - group that is treated at both t and t'

Heterogeneity treatment effects in time

Consider two groups and three periods:

- Group A: untreated at time 1, treated at both time 2 and 3
- Group B: untreated at time 1 and 2, treated at time 3

$$\widehat{\beta} = \frac{1}{2}(DID_{A,B,1,2} + DID_{B,A,2,3}),$$

- $DID_{A,B,1,2} = y_{A,2} - y_{A,1} - (y_{B,2} - y_{B,1})$ ✓
- $DID_{B,A,2,3} = y_{B,3} - y_{B,2} - (y_{A,3} - y_{A,2})$

Forbidden comparison and negative weights

- recall treatment effect at t for i that are treated at g:

$$\tau_{i,t}(g) := y_{i,t}(g) - y_{i,t}(\infty)$$

- $DID_{B,A,2,3} = y_{B,3} - y_{B,2} - (y_{A,3} - y_{A,2}) = \tau_{B,3}(3) - (\tau_{A,3}(2) - \tau_{A,2}(2))$
- $DID_{A,B,1,2} = y_{A,2} - y_{A,1} - (y_{B,2} - y_{B,1}) = \tau_{A,2}(2)$

therefore

$$\begin{aligned}
\widehat{\beta} &= \frac{1}{2}(DID_{A,B,1,2} + DID_{B,A,2,3}) \\
&= \frac{1}{2}[\tau_{A,2}(2) + \tau_{B,3}(3) - \tau_{A,3}(2) + \tau_{A,2}(2)] \\
&= \tau_{A,2}(2) + \frac{1}{2}\tau_{B,3}(3) - \frac{1}{2}\tau_{A,3}(2)
\end{aligned}$$

- early-treated group gets a negative weight in its longer run treatment effect

▸ Generalized theorems

Discussion: homogeneous treatment effects

Suppose for all i and $t \geq g$

$$\tau_{i,t}(g) = \tau$$

- all units have the same treatment effects
- the treatment has the same effect regardless of how long it has been since treated

The population regression coefficient β is τ

- is this a sensible assumption?

Some discussion on the negative weights

By FWL

$$\widehat{\beta} = \frac{\sum_{i,t}(D_{i,t} - \widehat{D}_{i,t})y_{i,t}}{\sum_{i,t}(D_{i,t} - \widehat{D}_{i,t})^2},$$

- $\widehat{D}_{i,t} = \overline{D}_i + \overline{D}_t - \overline{D}$
  - $\overline{D}_i = \sum_t D_{i,t}/T$: time average of treatment status for i
  - $\overline{D}_t = \sum_i D_{i,t}/n$: cross sectional average at time t
  - $\overline{D} = \sum_{i,t} D_{i,t}/nT$: broad average

Negative weights for early-treated group

- early-treated group gets a negative weight in its longer run treatment effect
- $\overline{D}_i \approx 1$: i has been treated most of the periods
- $\overline{D}_t \approx 1$: almost all i's are treated at t
- $\overline{D} < 1$: there is a fraction of non-treated at some time
- $D_{i,t} - \widehat{D}_{i,t} < 0$
- Proposition 1 in de Chaisemartin and D'Haultfœuille (2020, AER): TWFE estimator tends to assign negative weights to periods where a large fraction of groups are treated, and to groups treated for many periods.

Outline of the presentation

- Pitfalls of TWFE estimator in dynamic specifications

Dynamic specification

$$y_{i,t} = \alpha_i + \delta_t + \sum_{r \neq 0} \mathbf{1}[R_{i,t} = r]\beta_r + \varepsilon_{it}$$

- $R_{i,t} = t - G_i + 1$: the time relative to treatment
- G_i: the earliest period in which i is treated
- $CATT_{e,l}$ (cohort ATT): average treatment effect l periods from the initial treatment for cohort g

$$CATT_{g,l} = \mathbb{E}[y_{i,g+l} - y^{\infty}_{i,g+l} | G_i = g]$$

- $y^{\infty}_{i,g+l}$: never-treated counterfactual

Heterogeneous treatment effects across cohorts

Proposition 3 in Sun and Abraham (2021)

$$\widehat{\beta}_r = \sum_g w_{g,l} CATT_{g,l} + \sum_{l' \neq l} \sum_g w_{g,l'} CATT_{g,l'}$$

- $\sum_g w_{g,l} = 1$, $\sum_g w_{g,l'} = 0$

blue term: resembles the issue in static specifications

red term: contamination from CATT at other time periods

- if $CATT_{g,l'}$ is the same across all g (homogeneneity across group)

$$\sum_g w_{g,l'} CATT_{g,l'} = CATT_{g,l'} \sum_g w_{g,l'} = 0$$

- invalidates using $\widehat{\beta}_r = 0$ $(r < 0)$ as pre-testing

Next presentation

- alternative estimators robust to heterogeneity
- diagnostic tools
- computing standard errors

ATT under the two assumptions

Rearrange the parallel trend assumption equation to be

$$\mathbb{E}[y_{i,2}(0,0)|D_i=1]=\mathbb{E}[y_{i,1}(0,0)|D_i=1]+\mathbb{E}[y_{i,2}(0,0)-y_{i,1}(0,0)|D_i=0]$$

Plug into the ATT expression

$$\begin{aligned}
\tau &:= \mathbb{E}[y_{i,2}(0,1)-y_{i,2}(0,0)|D_i=1] \\
&= \mathbb{E}[y_{i,2}(0,1)|D_i=1]-\mathbb{E}[y_{i,1}(0,0)|D_i=1]-\mathbb{E}[y_{i,2}(0,0)-y_{i,1}(0,0)|D_i=0] \\
&= \mathbb{E}[y_{i,2}(0,1)|D_i=1]-\mathbb{E}[y_{i,1}(0,1)|D_i=1]-\mathbb{E}[y_{i,2}(0,0)-y_{i,1}(0,0)|D_i=0] \\
&= \underbrace{\mathbb{E}[y_{i,2}(0,1)-y_{i,1}(0,1)|D_i=1]}_{\text{treated group difference}}-\underbrace{\mathbb{E}[y_{i,2}(0,0)-y_{i,1}(0,0)|D_i=0]}_{\text{control group difference}}
\end{aligned}$$

Hence the name difference-in-differences

◂ Back

Sampling assumption for inference

Let $W_i = (y_{i,2}, y_{i,1}, D_i)'$ denote the vector of outcomes and treatment status for unit i. The data is a sample of n i.i.d. draws $W_i \sim F$ for some distribution F that satisfies parallel trends

- $\widehat{\beta}$ is a consistent estimate of τ and has asymptotically valid confidence intervals (Roth et al., 2022)
- how to estimate (clustered) standard errors is worth a separate presentation (Bertrand et al., 2004; Abadie et al., 2023)

◂ Back

A parallel trend assumption

If the treatment had not occurred, the average outcomes for all adoption groups would have evolved in parallel

- for all $t' \neq t$ and $g \neq g'$

$$\mathbb{E}[y_{i,t}(\infty) - y_{i,t'}(\infty) | G_i = g] = \mathbb{E}[y_{i,t}(\infty) - y_{i,t'}(\infty) | G_i = g']$$

◂ Back

No anticipatory effects assumption

For all i and $t < g$,

$$y_{i,t}(\mathrm{g}) = y_{i,t}(\infty)$$

- unit i does not act on the knowledge of their future treatment date before treatment starts

◂ Back

Generalized results in static TWFE specifications

- Theorem 1 in de Chaisemartin and D'Haultfœuille (2020, AER)
- Proposition 2 in Borusyak, Jaravel and Spiess (2023)

◂ Back