Difference–in–differences Part I

Shuowen Chen

BW Econometrics Reading Group (04/24/2023)

A personal fun fact

Fernandez-Val, Ivan
To: Chen, Shuowen
Fri 5/3/2019 8:08 PM

Hi Shuowen,

This sounds like an interesting and relevant question. However, since there are already several related papers on the topic, is not clear to me if it can be made into a job market paper. It would help if you can find an interesting empirical application to motivate the problem.

Best,

Ivan

...

On Apr 13, 2019, at 3:24 PM, Chen, Shuowen <swchen@bu.edu> wrote:

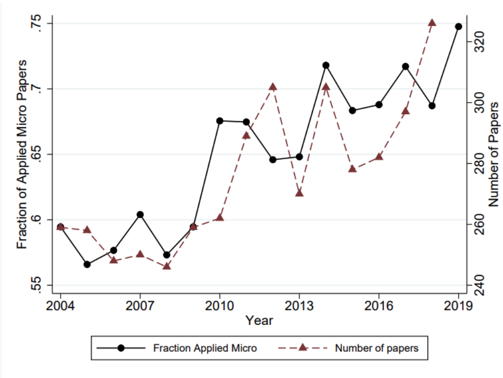
Hi professor,

The attachment is my research proposal. Please let me know your thought on this.

Thank you.

Growing importance of empirical research

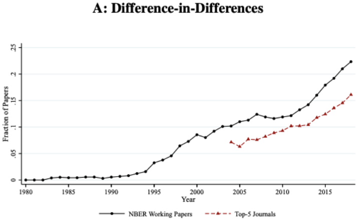
Figure I: Fraction of Applied Microeconomics Articles in Top-5 Journals

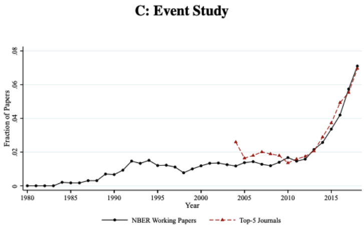


Notes: This figure shows the fraction of papers in top-5 journals that report an applied microeconomics JEL code (left axis) and the total number of papers in the top-5 journals (right axis).

- Currie et al. (2020, AER P&P)

Growing popularity of DID methods





- Five-year moving average time series of the faction of papers referring to each type of quasi–experimental approach

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

A wide applicability of DID

Ashenfelter (1978): estimate the effect of training programs on earnings

Card and Krueger (1994): effect of minimum wage on employment

Gentzkow et al. (2011): effect of newspaper entry and exit on political participation, voter turnout and electoral competitiveness

Wollmann (2019): effect of an abrupt increase in the US merger review threshold on antitrust enforcement and merger activities

BW litigation: price fixing damages

Outline of the presentation

- DID basics: two groups and two periods and a binary treatment
  - estimand of interest
  - assumptions
  - implementation (linear models with fixed effects)
- discussion of DID assumptions
  - parallel trends and no anticipatory effects
- extension to multiple periods and groups
  - staggered adoption
  - extended potential outcomes framework and assumptions

A 2 BY 2 EXAMPLE

- Two periods: $t = 1, t = 2$
- Unit i drawn from one of the two populations
  - $D_i = 1$: receives a treatment between $t = 1$ and $t = 2$
  - $D_i = 0$: remains untreated in both periods
- Observed data
  - $y_{i,t}$: outcome variable
  - D_i: treatment status

Potential outcomes framework

- $y_{i,t}(0,0)$: 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

- $y_{i,t}(0,0)$: 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

Assumption 1: parallel trend

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

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

Assumption 2 : no anticipatory effects

The treatment has no causal effect prior to its implementation

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

ATT estimation 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

Implementation: 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 *i*s

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)

Outline of the presentation

- DID basics: two groups and two periods and a binary treatment
  - estimand of interest
  - assumptions
  - implementation (linear models with fixed effects)
- **discussion of DID assumptions**
  - parallel trends and no anticipatory effects
- extension to multiple periods and groups
  - staggered adoption
  - extended potential outcomes framework and assumptions

Parallel trend allows for selection bias

- let $y_{i,t} = \alpha_i + \delta_t + \varepsilon_{i,t}$ and suppose group one self selects in the treatment based on α_1
- parallel trend assumption

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

  still holds because α_i's are differenced out
- partially why DID is very popular because treatments are usually endogeoenous
- Mark Israel (Rail freight MDL1 rebuttal report) missed this

Is the parallel trend assumption testable/falsifiable

- Suppose the data has 3 time periods: $-1, 0, 1$

$$y_{i,t} = \sum_{s \neq 0} (\mathbf{1}[t = s]) \cdot D_i) \times \beta_s + \alpha_i + \delta_t + \varepsilon_{i,t}$$

- $\widehat{\beta}_1$: DID estimate
- $\widehat{\beta}_{-1}$: pre–period event–study coefficient
- test for pretreatment differences in trends ("pre–trends") by testing the significance of $\widehat{\beta}_{-1}$

Pre-trend test is widely applied

Table 1—Summary of Pre-period Event-Study Coefficients

Paper	# pre-periods	# significant	Max $|t|$	Joint p-value	$|t|$ for slope
Bailey and Goodman-Bacon (2015)	5	0	1.674	0.540	0.381
Bosch and Campos-Vazquez (2014)	11	2	2.357	0.137	0.446
Deryugina (2017)	4	0	1.090	0.451	1.559
Deschenes et al. (2017)	5	1	2.238	0.014	0.239
Fitzpatrick and Lovenheim (2014)	3	0	0.785	0.705	0.977
Gallagher (2014)	10	0	1.542	0.166	0.855
He and Wang (2017)	3	0	0.884	0.808	0.720
Kuziemko et al. (2018)	2	0	0.474	0.825	0.474
Lafortune et al. (2017)	5	0	1.382	0.522	1.390
Markevich and Zhuravskaya (2018)	3	0	0.850	0.591	0.676
Tewari (2014)	10	0	1.061	0.948	0.198
Ujhelyi (2014)	4	1	2.371	0.003	1.954

Notes: This table provides information about the pre-period event-study coefficients in the papers reviewed. The table shows the number of pre-period coefficients in the event study, the number of the pre-period coefficients that are significant at the 95 percent level, the maximum t-stat among those coefficients, the p-value for a chi-squared test of joint significance, and the t-stat for the slope of the linear trend through the pre-period coefficients. See Section I for more detail on the sample of papers reviewed.

- Roth (2022, AER Insights)

The issue with pre–testing: low power

Suppose a researcher tests for K periods of pre–trends

- the test checks if $\widehat{\beta}_{pre} \in B_{NIS}(\Sigma) \equiv \{\beta \in \mathbb{R}^K : |\beta_t| \leq 1.96\sigma_t \text{ for all } t\}$ and σ_t is the standard error of $\widehat{\beta}_{pre,t}$
- an alternative hypothesis:

$$y_{i,t}(0) = \alpha_i + \delta_t + D_i \times g(t) + \varepsilon_{i,t}$$

  linear violation as a special case: $g(t) = \gamma \cdot t$

- power of a test: given that an alternative hypothesis is true, the probability that a test can reject the null
- Roth (2022): the test cannot detect violation of pre–trend unless γ is very large, thus incurs bias and low CI coverage of $\widehat{\beta}$

Is the parallel trend assumption sensitive to functional form?

- a researcher may be interested ATT in levels for a particular policy
- does state-level variation in the policy generate parallel trends in levels or in logs?
- Roth and Sant'Anna (2023, ECMA): when do parallel trends hold regardless of the units in which one measures the outcome

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

for all strictly monotonic function g (Definition 1)

Insensitivity to monotonic transformation

Iff the population can be divided into two groups

- Group one: as good as randomly assigned between the treatment and control
- Group two: has the same potential outcome distribution (CDF) in both groups

A testable implication:

$$F_{y_{i,2}(0,0)|D_i=1}(y) = F_{y_{i,1}(0,0)|D_i=1}(y) + F_{y_{i,2}(0,0)|D_i=0}(y) - F_{y_{i,1}(0,0)|D_i=0}(y)$$

- test if the RHS is weakly increasing in y
- if y has finite support $\mathcal{Y}$, equivalent to testing

$$f_{y_{i,2}(0,0)|D_i=1}(y) = f_{y_{i,1}(0,0)|D_i=1}(y)+f_{y_{i,2}(0,0)|D_i=0}(y)-f_{y_{i,1}(0,0)|D_i=0}(y) \geq 0$$

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

Outline of the presentation

- DID basics: two groups and two periods and a binary treatment
  - estimand of interest
  - assumptions
  - implementation (linear models with fixed effects)
- discussion of DID assumptions
  - parallel trends and no anticipatory effects
- **extension to multiple periods and groups**
  - staggered adoption
  - extended potential outcomes framework and assumptions

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
  - $y_{i,t}(g) \equiv y_{i,t}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$
  - $y_{i,t}(\infty) \equiv y_{i,t}(\mathbf{0}_T)$

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 $g \neq g'$

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

Next presentation

- pitfalls of fixed effect estimates when treatment effects are heterogeneous
- alternative estimates
- dynamic treatment effects