EC 709: Differences-in-Differences–Estimation

Liang Zhong[1]

Boston University

samzl@bu.edu

September 2023

[1]Thanks to Rubaiyat Alam for kindly sharing his lecture notes

Overview

Table of Contents

Setup for TWFE

$$h(y_{igt}) = \alpha_g + \gamma_t + \beta D_{gt} + \epsilon_{igt}$$
$$E(h(Y_{gt}(D))) = \alpha_g + \gamma_t + \beta D_{gt}$$

- Assume the additive structure between the group FE and time FE
- The regression counterpart to the parallel trends assumption is $E(\epsilon D|t, g) = 0$
- $h(\cdot)$ depends on our assumption of functional form
  - $h(y) = y$ for absolute amounts
  - $h(y) = lny$ for percentage changes
  - $h(y) = Pr(y = 1)$ for binary response models
- For logit and probit, be careful when include group dummies for incidental parameters problem :
  - Fixed effects' dummies cannot be canceled out in non-linear case
  - At least include treatment dummy to replace α_g

Assumptions for TWFE to be DID

- In practice, most papers use two-way fixed effects (TWFE) to implement DID

  $\Leftarrow$ Almost all papers equate DID with TWFE — TWFE is "regression DD" (Angrist & Pischke, 2009)
  - What is TWFE estimating and When would $\beta = DID$?

1. DiD estimate follows the same functional form assumption implied by PT:
   - Assume probabilities move in equal absolute amounts for both groups $\Rightarrow$ Linear model
   - Assume probabilities move by the same standard deviation/logit expression $\Rightarrow$ Probit/Logit
2. Constant treatment effect
   - Often Implausible. E.g., the effect of minimum wage on employment is likely to differ in counties with highly vs. less educated workers

- Both two conditions need to hold!

Functional form and TWFE

$$Pr(y_{igt} = 1) = F(c + \beta D_{gt} + \alpha d_g + \gamma T_t + \epsilon_{igt})$$

- F is logistic; β: coefficient on the interaction term; Assume 2×2:
- If believe the counterfactual is the same change in probability, cannot use interaction term as the estimation:
- $DID = (\frac{e^{c+\beta+\alpha+\gamma}}{1+e^{c+\beta+\alpha+\gamma}} - \frac{e^{c+\alpha}}{1+e^{c+\alpha}}) - (\frac{e^{c+\gamma}}{1+e^{c+\gamma}} - \frac{e^{c}}{1+e^{c}})$

  $\Rightarrow \beta = 0 \not\equiv DiD = 0$

- Use the logit TWFE specification only if PT assumes probabilities move by the same logit expression

$\Rightarrow$ See Wooldridge (2023) for a Non-linear DID framework in general case

Table of Contents

Set up for staggered design

- T time periods: $t = 1, 2, ..., T$.
- Units adopt a binary treatment at different dates $G_i \in \{1, ..., T\} \cup \infty$(where $G_i = \infty$ means "never-treated")
  - If no staggered design, only 2 groups: $G_i = g$ (treated at period g) and $G_i = \infty$ (untreated by period T)
- Potential outcomes $Y_{i,t}(g)$ – depend on time and the time you were first-treated
- For units treated at time g, the time t specific ATT's: $ATT(g,t) = E[Y_{i,t}(g) - Y_{i,t}(\infty)|G_i = g]$

Assumption 1 (No-Anticipation)

For all units i, $Y_{i,t}(g) = Y_{i,t}(\infty)$ for all groups in their pre-treatment periods, i.e., for all $t < g$

$\Rightarrow$ Treatment has no impact before it is implemented, $ATT(g,t) = 0$ for all pre-treatment periods $t < g$

PT for multiple time periods

Assumption 2 (Parallel Trends Assumption (PT))

$$E[Y_{i,t}(\infty) - Y_{i,t-1}(\infty)|G_i = g] = E[Y_{i,t}(\infty) - Y_{i,t-1}(\infty)|G_i = g'] \text{ for all } g, g', t$$

$\Rightarrow$ For identification, we only need to impose slightly weaker versions:

Assumption 3 (PT in post-treatment periods)

$$E[Y_{i,t}(\infty) - Y_{i,t-1}(\infty)|G_i = g] = E[Y_{i,t}(\infty) - Y_{i,t-1}(\infty)|G_i = \infty] \text{ for all } t \geq g$$

$\Rightarrow$ In the absence of treatment, the evolution of the outcomes among the treated units is, on average, the same as the evolution of the outcomes among the untreated units, in all post-treatment periods

Estimation for multiple time periods

- Following the same steps as last time: For all $t \geq g$
  $ATT(g,t) = E[Y_{i,t} - Y_{i,t=g-1}|G_i = g] - E[Y_{i,t} - Y_{i,t=g-1}|G_i = \infty]$
- Very easy to get this via TWFE regressions
  1. Subset your data to have data only for two groups $G_i \in \{g, \infty\}$ and two time periods t and $g-1$, for $t \geq g$
  2. **In this subset of the data,** run the TWFE regression using the following linear specification:

$$Y_i = c + \alpha 1\times\{G_i = g\} + \gamma 1\times\{T_i = t\} + \beta(1\times\{G_i = g\}\cdot 1\times\{T_i = t\}) + \epsilon_i$$

$\Rightarrow$ Estimated treatment effect in group g at time t: $TE_{g,t} \equiv \beta = ATT(g,t)$

- However, in practice it is tempting to run the pooled regression:

$$Y_{it} = \alpha_i + \gamma_t + D_{it}\beta + \epsilon_{it}, \text{where } D_{it} = 1[t \geq G_i] \text{ is a treatment indicator.}$$

$\Rightarrow$ It is unclear what are we estimating in this specification

Heterogeneous effect and TWFE

- dCDH (2020) shows that under PT:

$$E[\hat{\beta}] = E[\sum_{(g,t):D_{g,t}\neq 0} W_{g,t} TE_{g,t}]$$

- $W_{g,t}$ = weights summing to 1
- $W_{g,t} \neq$ proportional to the population of the cell (g,t), so $\hat{\beta} \neq ATET \neq DID$
- More severally, some weights are negative!

$\Rightarrow$ In extreme case, $E[\hat{\beta}] < 0$ even if $TE_{g,t} > 0$ for all (g, t)

- Can use package twowayfeweights to compute all $W_{g,t}$
- Why can we have a negative weight?

$\Leftarrow$ β may compare switchers to always treated

Example from Goodman-Bacon (2021)

- Assume a staggered design: group e treated at $t = 2$, group l treated at t=3, Then

$$\hat{\beta} = \frac{1}{2} \times DID_{e-l}^{1-2} + \frac{1}{2} \times DID_{l-e}^{2-3}$$

  - First term: comparing group e switching from untreated to treated to group l untreated at both periods
  - Second term: comparing switching group l to group e treated at both periods

$\Rightarrow$ With fundamental DID assumptions, $E[DID_{l-e}^{2-3}]$

$$\begin{aligned}
&= E[Y_{l,3} - Y_{l,2} - (Y_{e,3} - Y_{e,2})] \\
&= E[Y_{l,3}(0) + TE_{l,3} - Y_{l,2}(0) - (Y_{e,3}(0) + TE_{e,3} - (Y_{e,2}(0) + TE_{e,2}))] \\
&= E[TE_{l,3} - TE_{e,3} + TE_{e,2}]
\end{aligned}$$

- $TE_{e,3}$ enters with negative weight
- No negative weights if $TE_{g,t} = TE_{g,t'}$

A application from dCDH (2020)

- Revisit Gentzkow et al. (2011) about the effect of the number of newspapers on electoral turnout
- $\beta_{fd} = 0.0026(se = 0.0009)$ and $\beta_{fe} = -0.0011(se = 0.0011)$

$\Rightarrow$ significantly different (t-stat=2.86), so under common trends, reject constant treatment effect

- 45.7% of weights attached to β_{fd} negative, negative weights sum to -1.43
- 40.1% of weights attached to β_{fe} negative, negative weights sum to -0.53
- Corrected: $DID = 0.0043(se = 0.0015)$

$\Rightarrow$ 66% larger and significantly different from β_{fd} at 10% level (t-stat=1.77), has an opposite sign to β_{fe}

Similar issue in the event study plot

Consider dynamic TWFE specifications:

$$Y_{i,t} = \alpha_i + \gamma_t + \sum_{k \neq 0} \beta_k D_{i,t}^k + \epsilon_{it}$$

where $D_{i,t}^k = 1 \times \{t - G_i = k\}$ are "event-time" dummies

- For $k \geq 0$, β_k supposed to estimate cumulative effect of k periods from initial treatment across all groups; For $k \leq -2$, β_k = placebo
- Sun and Abraham (2021) suggests that interpreting the β_k as estimates of the dynamic effects of treatment may be misleading
- The issues arise if the dynamic path of treatment effects is heterogeneous across adoption cohorts

$\Leftarrow$ Biases may be less severe than for "static" specs if dynamic patterns are similar across cohorts

Dynamic TWFE also not robust to heterogeneous effects

Sun and Abraham (2021) show that:

$$E[\hat{\beta}_k] = E[\sum_g w_{g,k} TE_g(k) + \sum_{k' \neq k} \sum_g w_{g,k'} TE_g(k')]$$

$TE_g(k) = ATT(g, g+k)$ is effect k periods from initial treatment in group g

- 1st sum: weighted sum across groups of the effect of $k+1$ treatment periods, with possibly negative weights $\Rightarrow \hat{\beta}_k$ not robust to heterogeneous effects
- 2nd sum: weighted sum, across $k' \neq k$, of effects of $k'+1$ treatment periods $\Rightarrow \hat{\beta}_k$ contaminated by effects of $k'+1$ treatment periods
- Pre-trends tests of the $\hat{\beta}_k$ for $k \leq -2$ may be misleading

$\Leftarrow$ could be non-zero even if PT holds, since they may be "contaminated" by post-treatment effects!

- **eventstudyweights** STATA package can compute above weights

Table of Contents

Callaway and Sant'Anna (2020)

- Several new (closely related) estimators have been proposed to try to address these negative weighting issues
- The key components of all of these are:
  - Be precise about the target parameter (estimand) – i.e., how do we want to aggregate treatment effects across time/units
  - Estimate the target parameter without using already-treated units as the control group
- As shown previously, there is no problem to subset the dataset and run TWFE to estimate each $ATT(g,t)$

$\Leftarrow$ This is precisely the idea behind Callaway and Sant'Anna (2020)!

1. If have a large number of observations and relatively few groups/periods

   $\Rightarrow$ Report $\hat{ATT}(g,t)$ directly
2. If there are many groups/periods compared to the number of observations

   $\Rightarrow$ $\hat{ATT}(g,t)$ may be very imprecisely estimated and/or too numerous to report concisely

Aggregate Scheme

- In these cases, it is often desirable to report sensible averages of the $\hat{ATT}(g,t)$s
- One of the most useful is to report event-study parameters which aggregate $\hat{ATT}(g,t)$s at a particular lag since treatment:

  E.g., $\hat{\theta}_k = AVE_g(\hat{ATT}(g, g+k))$ aggregates effects for cohorts in the kth period after treatment
  - Can also construct for $k < 0$ to estimate "pre-trends"
- If interested in the ATT parallel to the two-period and two-group case

$\Leftarrow$ Computing the average treatment effect for each group and then averaging across groups

- Callaway and Sant'Anna (2020) discuss other sensible aggregations too

E.g., if interested in whether treatment effects differ across good/bad economies, may want to "calendar averages" that pool the $\hat{ATT}(g,t)$ for the same year

- R package **did** is available to implement the whole procedure

Comparisons of new estimators

- Callaway and Sant'Anna also propose an analogous estimator using not-yet-treated rather than never-treated units.
- Sun and Abraham (2021) propose a similar estimator but with different comparison groups (e.g. using last-to-be treated rather than not-yet-treated)
- Borusyak et al. (2021), Wooldridge (2021), Gardner (2021) propose "imputation" estimators that estimate the counterfactual $\hat{Y}_{it}(0)$ using a TWFE model that is fit using only pre-treatment data
  - Main difference from C&S is that this uses more pre-treatment periods, not just period $g-1$
  - This can sometimes be more efficient (if the outcome is not too serially correlated), but also relies on a stronger PT assumption that may be more susceptible to bias
- Roth and Sant'Anna (2021) show that you can get even more precise estimates if you're willing to assume treatment timing is "as good as random"

Roth's Advice to DID practitioners

- Don't freak out about this new literature!
- In most cases, using the "new" DiD methods will not lead to a big change in your results (empirically, TE heterogeneity is not that large in most cases)
- The exceptions are cases where there are periods where almost all units are treated – this is when "forbidden comparisons" get the most weight
- The most important thing is to be precise about who you want the comparison group to be and to choose a method that only uses these "clean comparisons"
- In his experience, the difference between the new estimators is typically not that large – can report multiple new methods for robustness (to make your referees happy!)

Table of Contents

Conditional PT

- Last time mentioned the conditional PT:

Assumption 4 (Conditional Parallel Trends Assumption)

$$E[Y_{i,t=2}(0)|D_i = 1, X] - E[Y_{i,t=1}(0)|D_i = 1, X] = E[Y_{i,t=2}(0)|D_i = 0, X] - E[Y_{i,t=1}(0)|D_i = 0, X]$$

- How to incorporate this assumption in our estimation?
- Temptation: It is very tempting to "extrapolate" and use the "more general" TWFE regression specification:

$$y_{igt} = \alpha_g + \gamma_t + \beta D_{gt} + X_i'\lambda + \epsilon_{igt}$$

  where $E[\epsilon_{igt}|\alpha_g, \gamma_t, X_i] = 0$ almost surely.
- Is $\beta = ATT$? Let's try some simulation exercise

A simulation exercise By Pedro SantAnna

- Samples sizes $n = 1,000$, consider $10,000$ Monte Carlo experiments
- Available data are $Y_{t=2}, Y_{t=1}, D, X_{i=1}^{n}$:
- $X = (X_1, X_2, X_3, X_4), X_j \sim N(0,1), j = 1,2,3,4$
- $f_{reg}(X) = 210 + 27.4X_1 + 13.7(X_2 + X_3 + X_4)$
- $f_{ps}(X) = 0.75(-X_1 + 0.5X_2 - 0.25X_3 - 0.1X_4)$

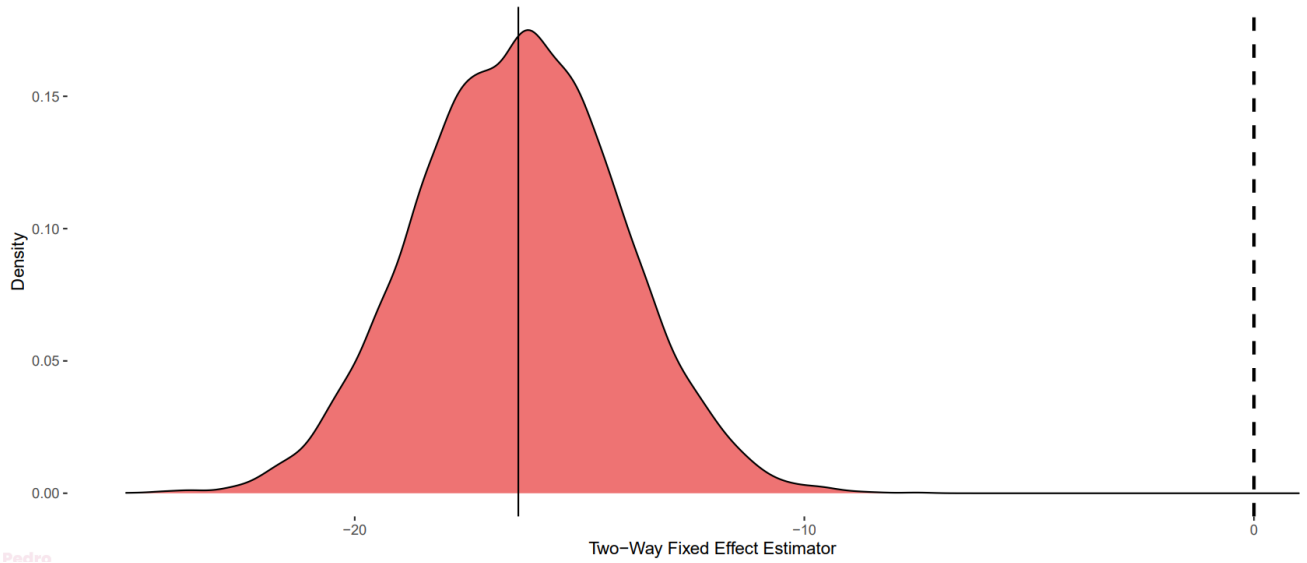
- $Y_{i,t=1}(0) = f_{reg}(X_i) + v_i(X_i, D_i) + \epsilon_{i,t=1}$
- $Y_{i,t=2}(d) = 2f_{reg}(X_i) + v_i(X_i, D_i) + \epsilon_{i,t=2}(d)$ for $d \in \{0,1\}$
  - $v(X, D) \sim N(Df_{reg}(X), 1)$, $\epsilon_{i,t=1}, \epsilon_{i,t=2}(d) \sim N(0,1)$
- $p(X_i) = exp(f_{ps}(X_i))/(1 + exp(f_{ps}(X_i)))$

$\Rightarrow$ $D_i = 1\{p(X_i) \geq U\}, U \sim U(0,1)$

- In this setup, $ATT = 0$

Simulation result from TWFE with covaraites

Figure 1: Monte Carlo for TWFE-based estimators



- Average of $\hat{\beta}$ in the simulations: -16.36 (very biased!)
- Coverage probability of 95% Confidence Interval: 0 (does not control size!)
- Why there is so much bias here?

Implication from TWFE specification

$$y_{igt} = \alpha_g + \gamma_t + \beta D_{gt} + X_i' \lambda + \epsilon_{igt}$$

$\Rightarrow$ $ATT(X)$

$$= (E[Y_{i,t=2} - Y_{i,t=1} | D_i = 1, X_i]) - (E[Y_{i,t=2} - Y_{i,t=1} | D_i = 0, X_i])$$
$$= (\beta + \gamma) - (\gamma) = \beta$$

- It is impossible in practice
  - $\Leftarrow$ Implied Average Treatment effects are homogeneous between covariate subpopulations
  - $\Leftarrow$ Evolution of the outcome among both treated/untreated units does not depend on X
- To capture the heterogeneity across different covariates, we need more econometrics techniques

$\Leftarrow$ Will come back to it when we are talking about "Matching"

Thank You!