Recitation 11: DiD and IV approaches to Treatment Effects

Seung-hun Lee

Columbia University
Introduction to Econometrics II Recitation

April 18th, 2022

Difference-in-Difference estimation

Simple DiD regression: 2 Groups and 2 periods

- 'Before and after' (t_0 and t_1) and treated vs untreated ($G_i = 1$ and $G_i = 0$)
- No one treated at t_0 but $G_i = 1$ group is treated at t_1, $G_i = 0$ are never treated
- We can define a treatment framework as

$$y_{i,t} = G_i y_i(1, t) + (1 - G_i) y_i(0, t) \ \ (t \in \{t_0, t_1\})$$

- What we observe: $(y_{i,t_1}, y_{i,t_0}, G_i, X_i)$ for every unit i
- What we do not: $y_i(1, t_1)$ for those in $G_i = 0$ and $y_i(0, t_1)$ for those in $G_i = 1$

Parallel trend is a CIA assumption in this setup

- We define it as

  $$(y_i(1,t_1) - y_i(1,t_0), y_i(0,t_1) - y_i(0,t_0)) \perp\!\!\!\perp G_i | X_i$$

  Or we can let $y_i(G_i, t_0) = y_i(t_0)$ for $G_i \in \{0,1\}$ since no one is treated then and write

  $$(y_i(1,t_1) - y_i(t_0), y_i(0,t_1) - y_i(t_0)) \perp\!\!\!\perp G_i | X_i$$

- Changes in treatment outcome for both groups are independent of G_i with X_i controlled for
  - Violated if uncontrolled factors affect the changes in the outcome or if there are differences in the trends of $y_{i,t}$ across treated and control groups before the experiment.

Parallel trend is a CIA assumption in this setup

- Let $D_i = G_i$ and write

$$
\begin{aligned}
y_i = y_{i,t_1} - y_{i,t_0} &= D_i(\underbrace{y_i(1,t_1) - y_i(1,t_0)}_{y_i(1)}) + (1 - D_i)(\underbrace{y_i(0,t_1) - y_i(0,t_0)}_{y_i(0)}) \\
&= D_i y_i(1) + (1 - D_i) y_i(0)
\end{aligned}
$$

  Thus, we can write $(y_i(1), y_i(0)) \perp\!\!\!\perp D_i | X_i$

- Testing for this is technically not feasible
  - As a close approach: Select a time period $\tilde{t} < t_0$ and find out if the difference $y_i(t_0) - y_i(\tilde{t})$ is independent with G_i
  - Do this by plotting pre-treatment outcome (not perfect: does not perfectly take into account the influence of unobserved factors)
  - If nonzero difference/pre-trends observed: Sure sign of nonparallel trends

Not to difficult to implement

- We can write

$$y_{it} = \beta_0(X_i) + \beta_1(X_i) \cdot \mathbb{I}[t = t_1] + \beta_2(X_i) \cdot G_i + \beta_3(X_i) \cdot G_i \cdot \mathbb{I}[t = t_1] + \epsilon_{it}$$

- In this context, the treatment effect for $X_i = x$ would be

$$\begin{aligned} TE(x) &= E[y_i(1) - y_i(0)|X_i = x] \\ &= E[(y_i(1, t_1) - y_i(1, t_0)) - (y_i(0, t_1) - y_i(0, t_0))|X_i = x] \\ &= x \cdot E[\{(\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_2)\} - \{(\beta_0 + \beta_1) - (\beta_0)\}|X_i = x] \\ &= x \cdot E[(\beta_1 + \beta_3) - \beta_1|X_i = x] \\ &= \beta_3 x \end{aligned}$$

So β_3 would be our parameter of interest.

Extending to TWFE

- in 2 $\times$ 2 setup, DiD is equivalent to TWFE

$$y_{it} = a_i + b_t + cD_{it} + e_{it}$$

  where $D_{it} = 1$ if unit i is treated at period t and 0 if otherwise
- Treatments can be implemented in multiple periods: Let $D_{i\tau} = 0$ if $D_{it} = 1$ and $\tau > t$
- New problem: Overall trend in the data is reversed or attenuated relative to a trend that appears in the groups composing the data (Simpson's paradox)
  - In the case where there is a strong heterogeneity of treatment across groups or time, the coefficient of c may be reversed from the true effect

Treatment effect as an weighted average of $i \times t$ cells

- Let G_i be the first period in which i is treated ($G_i = \infty$ if never treated)
- By applying (two-step) within estimation, we can get

$$\hat{c} = \frac{\sum_{i.t} \tilde{y}_{it} \widetilde{D}_{it}}{\sum_{i.t} \widetilde{D}_{it}^2}$$

- By FWL, we get the same estimate by regressing y_{it} onto the residual of

$$D_{it} = a_i + b_i + u_{it}$$

In this way, we can write

$$\hat{c} = \frac{\sum_{i,t} y_{it} \hat{u}_{it}}{\sum_{i,t} \hat{u}_{it}^2} = \sum_{i,t} y_{it} \left(\frac{\hat{u}_{it}}{\sum_{i,t} \hat{u}_{it}^2} \right)$$

Problematic case: Early-adopters with negative weights at large t

- Some observations end up getting *negative* weights.
  - This happens since $\hat{u}_{it}$ is not necessarily binary, and those whose treatment intensity is below the mean gets negative weights.
  - Early adopters in later years (high unit-level treatment mean and time-level treatment mean) end up with negative weights.
- Minimum requirement is to vary the coefficient c across different implementation period, $c(G_i)$ (a simple event-study)
- Not enough: We want to compare the average treatment effect for those treated earlier vs. later. Specifically, for a specific $g \leq t$, and for any $g' > t$, write

$$E[y_{it}(g) - y_{it}(\infty)|G_i = g] = E[y_{it} - y_{i,g-1}|G_i = g] - E[y_{it} - y_{i,g-1}|G_i = g']$$

- It is nicer to have those who are never treated (pure control)

Selection on unobservables: Setup and IV approach

Violation of CIA: Selection on unobservables

- Even with controls, assignment may fail to be random
  - Participants self-select based on expected benefit: Failure to predict expected benefit with observables
  - Participants may be selected, consciously or not, to join: Participants and nonparticipants systematically differ on something that is not usually observed.
  - There may be equilibrium effects: Outcomes may be subject to spillover effects (Think of control groups that are also inadvertently exposed to treatment)
- Mathematically, we end up with

$$\begin{aligned} E[y_i|D_i = 1, x] - E[y_i|D_i = 0.x] &= \mu(x,1) - \mu(x,0) + E[\epsilon_i(1)|D_i = 1, x] - E[\epsilon_i(0)|D_i = 0, x] \\ &= TE + \underbrace{E[\epsilon_i(1)|D_i = 1, x] - E[\epsilon_i(0)|D_i = 0, x]}_{\text{(Positive/Negative) selection bias}} \end{aligned}$$

  - $E[\epsilon(d)|D_i, X_i]$ no longer zero, and $(\epsilon_i(1), \epsilon_i(0))$ and the u_i in $D_i = \mathbb{I}[u_i < p(x_i)]$ can covary

Old approach: Heckman's two-step estimates

- We have a data generating process

$$y_i = \max\{X_i\beta + \sigma\eta_i, 0\}, \eta_i \sim N(0,1)$$

  So we only see y_i if $X_i\beta + \sigma\eta_i > 0$. We are able to observe D_i, specified as

$$D_i = \begin{cases} 1 & \text{if } \eta > -\frac{X_i\beta}{\sigma} \\ 0 & \text{otherwise} \end{cases}$$

- Then, for the observed sample, we are likely to have an η_i that is positively selected
- Adjust by regressing D_i on X_i to get $(\widehat{\beta/\sigma})$ and include estimate of $\frac{\phi(X_i\beta/\sigma)}{\Phi(X_i\beta/\sigma)}$ into the regression ($y_i = X_i\beta + \gamma\widehat{\frac{\phi(X_i\beta/\sigma)}{\Phi(X_i\beta/\sigma)}} + \epsilon_i$)
- Not recommended: If errors are non-normal, incorrect specification

IV approach: Relevance, Exclusion, and Monotonic

- **Relevancy**: It effects the propensity score $p(w, z) = \Pr(D_i = 1 | W_i = w, Z_i = z)$
- **Exclusion**: Distribution of the counterfactual outcomes and u_i does not depend on $Z_i | X_i$. To put it in mathematical notation,

$$(y_i(1), y_i(0), u_i) \perp\!\!\!\perp Z_i | W_i$$

- **Monotonicity**: For a given $W_i = w$, z changes the treatment in the same direction for everyone. This is also called a no two-way movement condition.
  - Only partially testable: For each i, only one of z or z' exists (intuition and proxy based on changes in treatment enrollment after changes in Z_i)

We identify treatment effects on compliers!

- Exclusion: u_i and the outcome can be correlated, but the changes in Z_i not affect u_i and outcome conditional on W_i
  - In that regard, we need to rewrite our $y_i(d)$ notation as

$$y_i(d) = \mu(w, d) + \epsilon_i(d)$$

- Relevance: By moving from $p(w, z)$ to some larger $p(w, z')$, we find compliers, a group of individuals who do not get treated at $p(w, z)$ but do get treated at $p(x, z')$
  - Also: Always-takers and never-takers
- Monotonicity: Prevents the case where $u_i < p(w, z)$ when $Z_i = z$ but $u_i' > p(x, z') > p(w, z)$ when $Z_i = z'$
  - This groups is referred to as defiers
  - u_i moves with Z_i and in an opposite direction: This violate exclusion as well (and monotonicity is also known as no two-way movement)

Local average treatment effect: Treatment localized to compliers!

- Narrow the interest to compliers, and get average treatment effect on this group
- Setup
  - Fix $W_i = w$
  - Z_i will be a binary instrumental variable. Think of this as a variable that affects eligibility but not related to outcome
  - $D_i(w, z)$ can be characterized as $D_i(w, z) = \mathbb{I}[u_i < p(w, z)]$. Note that as z rises, so will $p(w, z)$ due to relevance condition.
  - Z_i itself has no bearing, at least directly, on the outcome. So $y_i(d) = \mu(w, d) + \epsilon_i(d)$. So we still have $(\epsilon_i(1), \epsilon_i(0), u_i) \perp\!\!\!\perp Z_i | X_i$.
- LATE is defined as

$$LATE(w, z, z') = E[y_i(1) - y_i(0) | p(w, z) < u_i < p(w, z'), W_i = w]$$

Obtaining LATE from the data: Other group ruled out

- Breaking down the equation

$$\begin{aligned}
E[y_i|X_i = x'] - E[y_i|X_i = x] &= E[\mathbb{I}[u_i < p(x')]y_i(1) + \mathbb{I}[u_i > p(x')]y_i(0)|x'] \\
&\quad - E[\mathbb{I}[u_i < p(x)]y_i(1) + \mathbb{I}[u_i > p(x)]y_i(0)|x] \\
&= E[\mathbb{I}[u_i < p(x')]y_i(1) + \mathbb{I}[u_i > p(x')]y_i(0)] \\
&\quad - E[\mathbb{I}[u_i < p(x)]y_i(1) + \mathbb{I}[u_i > p(x)]y_i(0)] \ (\because \text{Exclusion}) \\
&= E[(\mathbb{I}[u_i < p(x')] - \mathbb{I}[u_i < p(x)])y_i(1) \\
&\quad - (\mathbb{I}[u_i > p(x')] - \mathbb{I}[u_i > p(x)])y_i(0)] \\
&= E[(\mathbb{I}[u_i < p(x')] - \mathbb{I}[u_i < p(x)])(y_i(1) - y_i(0))] \\
&= \Pr(\mathbb{I}[u_i < p(x')] - \mathbb{I}[u_i < p(x)] = 1) \\
&\quad \times E[y_i(1) - y_i(0)|\mathbb{I}[u_i < p(x')] - \mathbb{I}[u_i < p(x)] = 1]
\end{aligned}$$

- $\mathbb{I}[u_i < p(x')] - \mathbb{I}[u_i < p(x)] = 0$ for never-takers and always-takers
- Defiers ruled out by monotonicity

Obtaining LATE from the data: LATE as Wald

- As a result, we get

$$\begin{aligned}\Pr(\mathbb{I}[u_i < p(x')] - \mathbb{I}[u_i < p(x)] = 1)E[y_i(1) - y_i(0)|\mathbb{I}[u_i < p(x')] - \mathbb{I}[u_i < p(z)] = 1] \\ = \Pr(p(x) < u_i < p(x'))E[y_i(1) - y_i(0)|p(x) < u_i < p(x')]\end{aligned}$$

- Therefore, we are able to back out the definition of the LATE and can identify them as

$$LATE(w,z,z') = \frac{E[y_i|X_i = x'] - E[y_i|X_i = x]}{\Pr(p(x) < u_i < p(x'))} = \frac{E[y_i|X_i = x'] - E[y_i|X_i = x]}{p(x') - p(x)}$$

- We can go further: Estimate propensity scores with $p(w,z) = E[D_i|W_i = w, Z_i = z]$ and get

$$LATE(w,z,z') = \frac{E[y_i|w,z'] - E[y_i|w,z]}{E[D_i|w,z'] - E[D_i|w,z]}$$

Estimating LATE

- To obtain this from regressions, we follow these steps (parametrically or nonparametrically):
  1. Regress D on Z and other covariates W to get $\widehat{D} = \hat{p}(w, z)$
  2. Regress Y on other covariates W and $\widehat{D}$.

  The LATE estimator can then be obtained here is called the Wald estimator.

IV with continuous variation: Marginal treatment effect

- The marginal treatment effect at $p(w,z) = p$ is defined as the treatment effect on individuals whose $u_i = p(w,z)$
- We can write

  $$MTE(p) = E[y_i(1) - y_i(0) | u_i = p]$$

  The conditional expectation above is not directly obtainable from the data. Heckman and Vytlacil show that

  $$MTE(p) = \frac{\partial E[y_i | p(w,z) = p]}{\partial p}$$

- By changing p slightly, we identify ‘marginal compliers’ who change treatment assignment. MTE measures the treatment effect on this group

Deriving MTE, optional

- Let $G(p) = E[y_i \cdot \mathbb{I}[p(w,z) = p]]$, which we can rewrite as

$$\begin{aligned} G(p) &= E[y_i(1) \cdot \mathbb{I}[u_i < p(w,z)] \cdot \mathbb{I}[p(w,z) = p] + y_i(0) \cdot \mathbb{I}[u_i > p(w,z)] \cdot \mathbb{I}[p(w,z) = p]] \\ &= E[y_i(1) \cdot \mathbb{I}[u_i < p] \cdot \mathbb{I}[p(w,z) = p]] + E[y_i(0) \cdot \mathbb{I}[u_i > p] \cdot \mathbb{I}[p(w,z) = p]] \end{aligned}$$

- By the exclusion and since $u_i \sim U[0,1]$ (so $f(u_i) = 1$),

$$\begin{aligned} E[y_i(1) \cdot \mathbb{I}[u_i < p] \cdot \mathbb{I}[p(w,z) = p]] &= E[y_i(1) \cdot \mathbb{I}[u_i < p]] \Pr(p(w,z) = p) \\ &= \int_0^p E[y_i(1)|u = t] dt \Pr(p(w,z) = p) \\ E[y_i(0) \cdot \mathbb{I}[u_i > p] \cdot \mathbb{I}[p(w,z) = p]] &= \int_p^1 E[y_i(0)|u = t] dt \Pr(p(w,z) = p) \end{aligned}$$

Deriving MTE, optional

- Combine the two to get

$$G(p) = E[y_i \cdot \mathbb{I}[p(w,z) = p]] = \left(\int_0^p E[y_i(1)|u = t]dt + \int_p^1 E[y_i(0)|u = t]dt \right) \Pr(p(w,z) = p)$$

- Since $E[y_i \cdot \mathbb{I}[p(w,z) = p]] = E[y_i|p(w,z) = p] \cdot \Pr(p(w,z) = p)$, this implies that

$$E[y_i|p(w,z) = p] = \frac{G(p)}{\Pr(p(w,z) = p)} = \int_0^p E[y_i(1)|u = t]dt + \int_p^1 E[y_i(0)|u = t]dt$$

- Then, by Leibniz's integral rule

$$\frac{\partial E[y_i|p(w,z) = p]}{\partial p} = E[y_i(1)|u = p] - E[y_i(0)|u = p] = MTE(p)$$

Estimating MTE

1. Estimate $p(w, z) = \Pr(D_i = 1 | W_i = w, Z_i = z)$
2. Regress y_i on the estimated $p(w, z)$ and W_i in a flexible setting - preferably not just linearly but with some nonlinearities and interaction between W_i and $p(w, z)$.
3. Take a derivative with respect to p. (or local linear estimator)

4 For treatment effects, evaluate the $E[y_i | p(w, z), x]$ at $p(w, z) = 1$ and $p(w, z) = 0$ and identify the difference. (You can obtain $E[y_i | \cdot]$ by getting the predicted values).

Caveats for LATE and MTE

- Z belongs in the treatment equation (relevancy): $D_i = 1(u_i < p(W_i, Z_i))$
- Z does not belong in the outcome equation (exclusivity): $y_i(d) = \mu(w, d) + \epsilon_i(d)$
- In other words, we get $(\epsilon_i(1), \epsilon_i(0), u_i) \perp\!\!\!\perp Z_i | W_i$
- (MTE): $p(w, z)$ should be continuous with z so that derivatives are defined

Testing for CIA condition

- Recall that conditional independence assumption is satisfied when

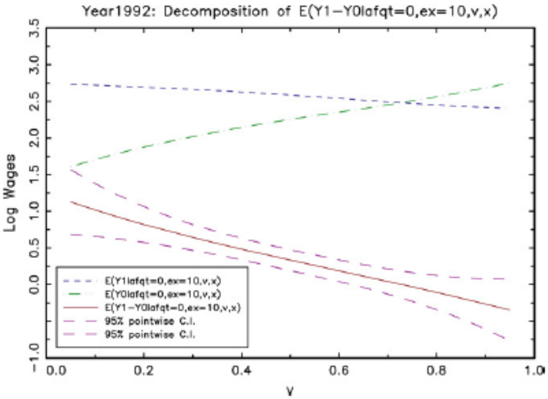
$$(\epsilon_i(1), \epsilon_i(0)) \perp\!\!\!\perp u_i | X_i$$

- In cases where this is true, then the outcomes are independent of u_i conditional on X_i.

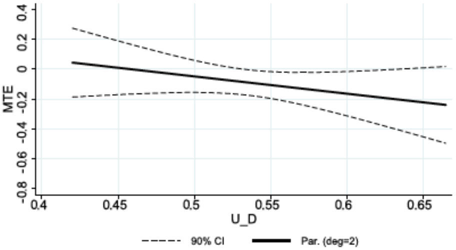
$$MTE(x,p) = E[Y_i(1) - Y_i(0)|X_i = x, u_i = p] = E[Y_i(1) - Y_i(0)|X_i = x]$$

- We know that $MTE(x,p) = \frac{\partial E[Y_i|X_i = x, p(w,z) = p]}{\partial p}$
- This suggests that $E[Y_i|X_i = x, p(x,z) = p]$ is linear in p if CIA holds.
- Thus, it is highly recommended to put polynomial terms of $p^k, k = 1, 2, 3, \ldots$ when you estimate marginal treatment effects. Then, test to find whether the nonlinear terms have coefficient zero. This is feasible if you have 3 or more points of $Z|W$

MTE is presented in graphs like these



(a) Carneiro, Lee (2009)



(b) Johnson, Taylor (2019)