“Difference-in-Differences with multiple time periods” Callaway & Sant’Anna (JoE, 2021)

Jonas Lieber

Diff-in-Diff Reading Group, University of Chicago

April 26, 2022

Big Picture

- extensions to canonical diff-in-diff model
  - multiple time periods
  - variation in treatment timing
  - treatment effect heterogeneity
  - "parallel trendsassumption" only conditional on covariates
- standpoint
  - TWFE can be weird/hard to interpret in these cases
  - What else could/should we done instead?
- plan
  1. identify disaggregated "group-time ATEs"
  2. aggregation of these to causal effects of interest
  3. estimation
     - outcome regression
     - inverse probabilityweighting
     - doubly-robust
  4. inference
     - bootstrap
     - joint inference

Model and Notation

- time periods $t \in \{1, \ldots, \bar{t}\}$
- binary treatment $D_{i,t} \in \{0, 1\}$
- all units are untreated at the start, i.e. $D_{i,1} = 0$
- treatment is irreversible, i.e. $D_{i,t+1} \geq D_{i,t}$ for all $t = 1, \ldots, \bar{t} - 1$
- $C_i \in \{0, 1\}$ is 1 iff i is never treated
- G_i is the time period in which i is first treated ($G_i = \infty$ if $C_i = 1$)
- hence the entire treatment path is characterized by G_i
- Let $G_{i,g} \in \{0, 1\}$ be 1 iff i is first treated in period g
- potential outcomes

$$Y_{i,t} = Y_{i,t}(0) + \sum_{g=2}^{\bar{t}} \left(Y_{i,t}(g) - Y_{i,t}(0) \right) G_{i,g}$$

Group-Time Average Treatment Effects

First objective is to identify

$$ATT(g,t) := \mathbb{E}[Y_{i,t}(g) - Y_{i,t}(0) | G_{i,g} = 1]$$

for all $g, t \in \{2, \ldots, \bar{t}\}$.

Assumptions I

1. treatment
   - all units start untreated
   - treatment is irreversible
2. panel random sampling
   - $(Y_{i,1}, \dots, Y_{i,\bar{t}}, X_i, D_{i,1}, \dots, D_{i,\bar{t}})$ is i.i.d. over i
3. limited treatment anticipation: there is a known $\delta > 0$ such that for all $g < \max\{G\}$, $t \in \{1, \dots, \bar{t}\}$ for which $t < g - \delta$

$$\mathbb{E}[Y_{i,t}(g)|X_i, G_{i,g} = 1] = \mathbb{E}[Y_{i,t}(0)|X_i, G_{i,g} = 1]$$

Assumptions II

4. conditional parallel trends based on a 'never-treated' group
   Let δ be the anticipation horizon from assumption 3. For each $g < \max\{G\}$, $t \in \{2, \dots, \bar{t}\}$ such that $t \geq g - \delta$,
   $$\mathbb{E}[Y_{i,t}(0) - Y_{i,t-1}(0)|X_i, G_{i,g} = 1] = \mathbb{E}[Y_{i,t}(0) - Y_{i,t-1}(0)|X_i, C_i = 1].$$

5. conditional parallel trends based on a 'not-yet' group
   Let δ be the anticipation horizon from assumption 3. For each $g < \max\{G\}$, $t \in \{2, \dots, \bar{t}\}$ and each $s, t \in (2, \dots, \bar{t})$ such that $t \geq g - \delta$ and $t + \delta \leq s < \max\{G\}$
   $$\mathbb{E}[Y_{i,t}(0) - Y_{i,t-1}(0)|X_i, G_{i,g} = 1] = \\ \mathbb{E}[Y_{i,t}(0) - Y_{i,t-1}(0)|X_i, D_{i,s} = 0, G_{i,g} = 0].$$

6. For each $t \geq 2$, $g < \max\{G\}$, there exists some $\varepsilon > 0$ such that $\mathbb{P}[G_g = 1] > \varepsilon$ and
   $$p_{g,t}(X) := \mathbb{P}[G_g = 1|X, G_g + (1 - D_t)(1 - G_g) = 1] < 1 - \varepsilon.$$

Identification Result

Theorem 1

Suppose Assumption 1-4 and 6 hold. Then for all g and t such that $2 + \delta \leq g < \max\{G\}$ and $t \geq g - \delta$

$$ATT(g,t) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{\mathbb{E}\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)\left(Y_t - Y_{g-\delta-1} - m_{g,t,\delta}(X)\right)\right],$$

where

$$m_{g,t,\delta}(X) = \mathbb{E}\left[Y_t - Y_{g-\delta-1} | X, C = 1\right].$$

Choosing the "Right" Comparison Group

- A4 ($\exists$ never-treated gr.) can be replaced w/ A5 ($\exists$ "not-yet" gr.)
- equation in the thm changes slightly, see paper
- both condition on X, see eg [HIT97] for motivation
- If never-treated *and* not-yet treated gr exist, which one to use?
  - Callaway & Sant'Anna favor A4 in applications if
    - never-treated group is large enough
    - never-treated group is "similar enough" to treated groups
  - using a not-yet treated units with A5
    - allows larger comparison groups
    - restricts pre-(anticipation)treatment trends of "not-yet" treated groups
    - can be problematic, see eg [MS21]
  - pretesting to select A4/A5 is problematic, see [Rot19]

Target Parameter

Callaway and Sant'Anna consider linear aggregators of the form

$$\theta = \sum_{g} \sum_{t=2}^{\bar{t}} w(g,t) ATT(g,t)$$

for some "carefully-chosen (known or estimable) weighting functions specified by the researcher such that θ can be used to address a well-posed empirical/policy question."
In particular,

- $w(g,t) \geq 0$ for all g and $t \geq 2$,
- $\sum_{g} \sum_{t=2}^{\bar{t}} w(g,t) = 1$.

Examples

- What are "overall" treatment effects?
- How does the effect of participating in the treatment vary with length of exposure to the treatment?
- Do groups that are treated earlier have, on average, higher/lower average treatment effects than groups that are treated later?
- What is the cumulative average treatment effect of the policy across all groups until some particular point in time?
  - Assume $\delta = 0$.
  - Then Callaway & Sant'Anna propose to consider

$$\theta^{\text{cumulative}} = \sum_{g < maxG} \mathbf{1}_{t \geq g} \mathbb{P}[G = g | G < t] ATT(g, t).$$

Estimation

- recall the identification result

$$ATT(g,t) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{\mathbb{E}\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)(Y_t - Y_{g-\delta-1} - m_{g,t,\delta}(X))\right]$$

- procedure
  1. estimate $ATT(g,t)$
     1. estimate $p_g(X)$ and $m_{g,t,\delta}(X)$
     2. use sample analog of $\mathbb{E}$ above, plugging in $\hat{p}_g(X)$ and $\hat{m}_{g,t,\delta}(X)$
  2. calculate/estimate weights $w(g,t)$
  3. calculate θ

Estimation

- recall the identification result

$$ATT(g,t) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{\mathbb{E}\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)(Y_t - Y_{g-\delta-1} - m_{g,t,\delta}(X))\right]$$

- the identification result in the paper is for three estimators
  - outcome regression
    - leaves out blue term
    - requires *correct* model for $m_{g,t,\delta}(X)$ (outcome evol of comp group)
  - inverse probability weighting
    - leaves out red term
    - requires *correct* model for $p_g(X)$ (propensity score)
  - doubly robust
    - includes both terms
    - robust to misspecification of either $m_{g,t,\delta}(X)$ or $p_g(X)$

Inference: Setting and Assumptions

- asymptotic regime
  - $\bar{t}$ fixed
  - n goes to ∞
- assumptions
  7. parametric model for $m_{g,t,\delta}(X)$ or $p_g(X)$
     - sufficiently smooth
     - $\exists$ $\sqrt{n}$ strongly consistent estimator for parameters
  8. param. model for at least one of $m_{g,t,\delta}(X)$ or $p_g(X)$ is correct
  9. integrability assumptions
- allows different estimators ((non-)least squares, mle,...)

Inference: Theorem

Theorem 2

Suppose Assumption 1-4 and 6-9 hold. Then for all g and t such that $2+\delta \leq g < \max G$ and $g-\delta \leq t \leq \bar{t}-\delta$

$$\sqrt{n}\left(A\hat{T}T_{t\geq g-\delta} - ATT_{t\geq g-\delta}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi_i + o_p(1)$$

where an analytic expression for Ψ_i is provided in the paper. Furthermore

$$\sqrt{n}\left(A\hat{T}T_{t\geq g-\delta} - ATT_{t\geq g-\delta}\right) \xrightarrow{d} \mathcal{N}(0,\Sigma)$$

where an analytic expression for Σ is provided in the paper.

Inference: Implementation

- could use analytic expression to estimate variance, Slutsky etc
- Callaway & Sant'Anna prefer the multiplier bootstrap
  - does not require re-estimation of propensity score in each bootstrap iteration
  - simplifies simultaneous inference
  - for each bootstrap iteration $s = 1, \ldots, S$
    1. draw a sample of N iid random variables with zero mean, unit variance, and finite third moment, independent of the original data, eg

$$V = \begin{cases} 1-\kappa & \text{w.p. } \frac{\kappa}{\sqrt{5}}, \\ \kappa & \text{w.p. } 1-\frac{\kappa}{\sqrt{5}} \end{cases}$$

       for $\kappa = (\sqrt{5}+1)/2$.
    2. $\hat{ATT}^s(t,g) = \hat{ATT} + \overline{V\hat{\Psi}}$.
  - this yields valid point-wise inference
  - see paper for formulas for simultaneous adjustments
  - can draw cluster-specific V's for cluster inference (if clusters are large)

Lemma for Proof of Theorem 1

$$
\begin{aligned}
&\mathbb{E}[Y_t(g) - Y_t(0)|X, G_g = 1] \\
=&\mathbb{E}[Y_t(g) - Y_{g-\delta-1}(0)|X, G_g = 1] - \mathbb{E}[Y_t(0) - Y_{g-\delta-1}(0)|X, G_g = 1] \\
=&\mathbb{E}[Y_t(g) - Y_{g-\delta-1}(0)|X, G_g = 1] - \sum_{l=0}^{t-g-\delta} \mathbb{E}[Y_{t-l}(0) - Y_{t-l-l}(0)|X, G_g = 1 \\
\overset{4}{=}&\mathbb{E}[Y_t(g) - Y_{g-\delta-1}(0)|X, G_g = 1] - \sum_{l=0}^{t-g-\delta} \mathbb{E}[Y_{t-l}(0) - Y_{t-l-l}(0)|X, C = 1] \\
=&\mathbb{E}[Y_t(g) - Y_{g-\delta-1}(0)|X, G_g = 1] - \mathbb{E}[Y_t(0) - Y_{g-\delta-1}(0)|X, C = 1] \\
\overset{3}{=}&\mathbb{E}[Y_t - Y_{g-\delta-1}|X, G_g = 1] - \mathbb{E}[Y_t - Y_{g-\delta-1}|X, C = 1].
\end{aligned}
$$

Note that assumption 1 is used whenever we write the potential outcomes as function of g.

Proof of Theorem 1 for the Outcome Regression

$$
\begin{aligned}
& ATT(g,t) \\
=& \mathbb{E}[\mathbb{E}[Y_t(g) - Y_t(0) | X, G_g = 1] | G_g = 1] \\
=& \mathbb{E}[\mathbb{E}[Y_t - Y_{g-\delta-1} | X, G_g = 1] - \mathbb{E}[Y_t - Y_{g-\delta-1} | X, C = 1] | G_g = 1] \\
=& \mathbb{E}[Y_t - Y_{g-\delta-1} | G_g = 1] - \mathbb{E}[\underbrace{\mathbb{E}[Y_t - Y_{g-\delta-1} | X, C = 1]}_{=: m_{g,t,\delta}(X)} | G_g = 1] \\
=& \mathbb{E}[Y_t - Y_{g-\delta-1} - m_{g,t,\delta}(X) | G_g = 1] \\
=& \mathbb{E}\left[\frac{G_g}{\mathbb{P}[G_g = 1]} (Y_t - Y_{g-\delta-1} - m_{g,t,\delta}(X))\right] \\
=& \mathbb{E}\left[\frac{G_g}{\mathbb{E}[G_g]} (Y_t - Y_{g-\delta-1} - m_{g,t,\delta}(X))\right].
\end{aligned}
$$

References I

- James J Heckman, Hidehiko Ichimura, and Petra E Todd, *Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme*, The review of economic studies **64** (1997), no. 4, 605–654.
- Michelle Marcus and Pedro HC Sant'Anna, *The role of parallel trends in event study settings: An application to environmental economics*, Journal of the Association of Environmental and Resource Economists **8** (2021), no. 2, 235–275.
- Jonathan Roth, *Pre-test with caution: Event-study estimates after testing for parallel trends*, Department of Economics, Harvard University, Unpublished manuscript (2019).