More on Diff-In-Diffs Methods

Zhiyuan Chen

RMEB
Remin Business School

May 2022

Outline

- Propensity Score Matching-DID
- Semi-parametric DID
- Synthetic DID
- Non-linear DID
- Staggered Difference-In-Differences

Propensity Score Matching - DID

Think of whether taking Chen's RMEB class as a binary decision, D_i; and income in the future Y_i is the interested outcome:

$$\textit{Potential Outcome} = \begin{cases} Y_i^1, & \textit{if } D_i = 1 \\ Y_i^0, & \textit{if } D_i = 0 \end{cases}$$

We aim to estimate

$$ATT = \mathbf{E}({\color{blue}Y_i^1} - {\color{red}Y_i^0} | D_i = 1, \mathbf{X}_i), \quad D_i \in \{0, 1\}$$

The problem is only $Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$ is observed. In lack of ideal experiment, we use nonparticipants (control group) to approximate participants:

$$\underbrace{E(Y_i|D_i = 1, \mathbf{X}_i) - E(Y_i|D_i = 0, \mathbf{X}_i)}_{\textit{observed income difference}} = \underbrace{E(Y_i^1 - Y_i^0|D_i = 1, \mathbf{X}_i)}_{\delta(\mathbf{X}_i) = ATT|\mathbf{X}_i}$$

$$+ \underbrace{E(Y_i^0|D_i = 1, \mathbf{X}_i) - E(Y_i^0|D_i = 0, \mathbf{X}_i)}_{{\color{red}\textit{Selection Bias}}}$$

- **Selection Bias**:

$$Bias(\mathbf{X}_i) = \mathbf{E}(Y_i^0 | D_i = 1, \mathbf{X}_i) - \mathbf{E}(Y_i^0 | D_i = 0, \mathbf{X}_i)$$

- **Conditional Independence Assumption (CIA):** $\{Y_i^0, Y_i^1\} \perp D_i | \mathbf{X}_i$
  - Under CIA, $Bias(\mathbf{X}_i) = 0$
  - CIA essentially states that D_i is randomly assigned conditioning on observeable characteristics $\mathbf{X}_i$ (see Angrist (1998) on voluntary military service)
  - CIA is likely to fail when *unobserved characteristics* determine select-into-treatment

"The idea of matching between treated and untreated units assumes that $Bias(\mathbf{X}_i) = 0$ *so that conditioning on* $\mathbf{X}_i$ *eliminates the bias."*

- Statisticians more often use matching methods (Cochrane and Rubin, 1973); the idea is quite similar to conditioning on observeables in the regression analysis
- "Regression can be motivated as a particular sort of weighted matching estimator" (Angrist and Pischke: *Mostly Harmless Econometrics*).
    1. **Matching estimator** puts the most weight on covariate cells containing units who are most likely to be treated: high $\Pr(D_i = 1|\mathbf{X}_i)$
    2. **Regression** puts the most weight on covariate cells with the largest conditional variance of treatment status: when $\Pr(D_i = 1|\mathbf{X}_i) \times [1 - \Pr(D_i = 1|\mathbf{X}_i)])$ is large
    3. **Common Support:** No weights assigned to covariate cells containing no treated and control units: $0 < \Pr(D_i = 1|\mathbf{X}_i) < 1$

- The dimension of $\mathbf{X}_i$ can be high, adding to difficulty of matching
- The well-known PSM is to match over the propensity of selection into the treatment (Rosenbaum and Rubin, 1983)
- **Exclusion Restrictions**: Partition $\mathbf{X}$ into $(\mathbf{T}, \mathbf{Z})$ such that:

$$Y_i^0 = g_0(\mathbf{T}_i) + U_i^0, \tag{1}$$

$$Y_i^1 = g_1(\mathbf{T}_i) + U_i^1, \tag{2}$$

$$\Pr(D_i = 1|\mathbf{X}_i) = \Pr(D_i = 1|\mathbf{Z}_i) = P(\mathbf{Z}_i) \quad (\textit{propensity score}) \tag{3}$$

  $\mathbf{T}$ and $\mathbf{Z}$ are not necessarily mutually exclusive
- To identify the ATT, it's enough to assume

$$U_i^0 \perp D_i | P(\mathbf{Z}_i) \quad (\text{PSM CIA})$$

PSM method can be put in the context of classical econometric selection models:

$$Y_i^0 = g_0(\mathbf{T}_i) + U_i^0 \tag{4}$$

$$D_i = \begin{cases} 1 & if \quad \lambda(\mathbf{Z}_i) - v \geq 0 \\ 0 & otherwise \end{cases} \tag{5}$$

- If $\mathbf{Z}_i$ and v are independent, then $P(\mathbf{Z}_i) = \Pr\{v \leq \lambda(\mathbf{Z}_i)\} = F_v[\lambda(\mathbf{Z}_i)]$; the CIA implies that

$$E\{U_i^0 | D_i = 1, F_v[\lambda(\mathbf{Z}_i)]\} = E\{U_i^0 | D_i = 0, F_v[\lambda(\mathbf{Z}_i)]\}$$

- If also $\lambda(\mathbf{Z}_i) \perp (U_i^0, v)$ and $E(U_i^0) = 0$, then for any s:

$$E(U_i^0 | v = s) = 0 \quad \text{(No Selection on Observables)}$$

Propensity Score Matching based DID

Heckman et al. (1998, ReSTud)

- Recall the OLS estimator the two-by-two DID model

$$Y_{it} = \delta D_{it} + \underbrace{\lambda t + \eta_i + v_{it}}_{U_{it}(D_{it})}$$

is simply:

$$\delta^{OLS} = E[Y_{i1} - Y_{i0} | D_{i1} = 1] - E[Y_{i1} - Y_{i0} | D_{i1} = 0]$$

A sufficient condition for the identification is

$$\Pr(D_{i1} = 1 | v_{it}) = \Pr(D_{i1} = 1)$$

- If v_{it} is correlated in time, the model cannot be identified: What are possible solutions?

- The traditional solution is to add covariates $\mathbf{X}_i$ into the model:

$$Y_{it} = \mu + \tau D_{it} + \mathbf{X}_i' \alpha(t) + \lambda t + \varepsilon_{it}$$

$$\Longrightarrow \quad Y_{i1} - Y_{i0} = \tau(D_{i1} - D_{i0}) + \mathbf{X}_i' \alpha + \lambda + \varepsilon_{i1} - \varepsilon_{i0}$$

  - $\mathbf{X}_i$ usually represents pre-treatment characteristics
  - Ideally, covariates $\mathbf{X}_i$ should be treated non-parametrically as $H(\mathbf{X}_i)$
- The PSM-DID method is a semi-parametric way of obtaining ATT:

$$\begin{aligned} \delta^{PSM-DID} &= E(Y_{i1} - Y_{i0} | \mathbf{X}_i, D_{i1} = 1) - E(Y_{i1} - Y_{i0} | \mathbf{X}_i, D_{i1} = 0) \\ Parallel\ Trend \Rightarrow \ \delta^{PSM-DID} &= E(U_{i1} - U_{i0} | P(\mathbf{Z}_i), D_{i1} = 1) \\ &\quad - E(U_{i1} - U_{i0} | P(\mathbf{Z}_i), D_{i1} = 0) \end{aligned}$$

- The PSM-DID estimator permits:
  1. Selection to be dependent on potential outcomes
  2. Selection on unobservables

How to Implement PSM-DID for Panel Data?

- **Estimation**:
  1. Calculate propensity score $P(\mathbf{Z}_i)$ using `probit` or `logit` models
  2. Matching by cohort-year over the estimated propensity score
  3. Generate differenced outcome variables $\Delta Y_{it} = Y_{it} - Y_{i0}$ and calculate treatment effects for oberservation it; Using appropriate weights to obtain ATT_{gt} at a desired aggregation level

  These steps apply to almost any matching-DID estimators (NN matching...)
- **Inference**: Analytical standard errors proposed by Abadie and Imbens (2006).
- Stata command: `teffects` with a nice intro PDF

Semi-parametric DID

Semi-parametric DID

Abadie (2005, ReSTud)

Under the conditional common trend assumption:

$$E[Y_{i1}^0 - Y_{i0}^0|\mathbf{X}_i, D_{i1} = 1] = E[Y_{i1}^0 - Y_{i0}^0|\mathbf{X}_i, D_{i1} = 0]$$

Abadie (2005) shows that ATT can be estimated using simple weighting schemes:

$$\begin{aligned} E[Y_{i1}^1 - Y_{i1}^0|\mathbf{X}_i, D_{i1} = 1] &= E[\rho_0(Y_{i1} - Y_{i0})|\mathbf{X}_i] \\ \text{where } \rho_0 &= \frac{D_{i1} - P(D_{i1} = 1|\mathbf{X}_i)}{P(D_{i1} = 1|\mathbf{X}_i)[1 - P(D_{i1} = 1|\mathbf{X}_i)]} \end{aligned}$$

$$ATT = E[Y_{i1}^1 - Y_{i1}^0|D_{i1} = 1] = E\left[\frac{Y_{i1} - Y_{i0}}{P(D_{i1} = 1)} \cdot \frac{D_{i1} - P(D_{i1} = 1|\mathbf{X}_i)}{1 - P(D_{i1} = 1|\mathbf{X}_i)}\right] \tag{6}$$

Semi-parametric DID

The dis-aggregated ATT can be obtained by approximating $E[Y_{i1}^1 - Y_{i1}^0 | D_{i1} = 1, \mathbf{X}_i^{sub}]$ as:

$$\delta(\mathbf{X}_i^{sub})^{semi-DID} = argmin_{\theta} E\left\{P(D_{i1} = 1|\mathbf{X}_i) \cdot [\rho_0(Y_{i1} - Y_{i0}) - g(\mathbf{X}_i^{sub};\theta)]^2\right\}$$

$\mathbf{X}_i^{sub}$ is a function of $\mathbf{X}_i$; $\mathbf{X}^{sub}$ may contain a subset of variables in $\mathbf{X}_i$

- **Estimation Strategy**:
  1. Estimate propensity score $P(D_{i1} = 1|\mathbf{X}_i)$
  2. Plug the fitted values into the sample analogue of the above equation
- Stata implementation: `absdid` [Link to Stata Manual]

Synthetic DID

Synthetic DID

Arkhangelsky et al. (2021, AER)

- *DID methods:* A substantial number of treated units; researchers invoke "parallel trend" to control for slection effects
- *Synthetic Control (SC) methods:* A single (or small number) of units exposed, seek to *compensate for the lack of parallel trends* by re-weighting units to match their pre-exposure trends
  - Synthetic control estimator by Abadie et al. (2010, JASA):

$$\hat{\delta}_t^{SC} = Y_t^1 - \sum_{i=2}^{I+1} \omega_i^* Y_{it}^* \tag{7}$$

  where $\omega^* = \left(\omega_2^*, \cdots, \omega_{I+1}^*\right)$is chosen to minimize $\|X_1 - X_0\omega\|$
- Synthetic DID have features of both SC and DID:
  1. Like SC, it re-weights and matches pre-exposure trends to *weaken the reliance on parallel trend type assumptions*
  2. Like DID, it is invariant to additive unit-level shifts, and allows for valid large-panel inference.

Synthetic DID

The basic idea

- A balanced panel with N units and T periods, binary treatment $D_{it} \in \{0,1\}$:
  - first N_0 units untreated, last $N_1 = N - N_0$ units treated
  - Units exposed to treatment after T_0
- Basic procedures:
  1. Like SC, find weights $\{\hat{\omega}_i^{sc}\}$ such that $\sum_{i=1}^{N_0} \hat{\omega}_i^{sc} \approx \frac{1}{N_1}\sum_{i=N_0+1}^{N} Y_{it}$ for all $t = 1, \cdots, T_0$
  2. Then use these weights in a basic panel data DID regression (two-way fixed effects) to estimate the ATT:

$$(\hat{\delta}^{sdid}, \hat{\mu}, \hat{\alpha}_i, \hat{\beta}_t) = \operatorname{argmin}_{\delta,\mu,\alpha_i,\beta_t}\left\{\sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{it} - \mu - \alpha_i - \beta_t - \delta D_{it})^2 \hat{\omega}_i^{sdid}\right\} \tag{8}$$

- DID can be thought of a special case of Synthetic DID without unit weights

Synthetic DID

Benefits of Synthetic DID:

- Using similar units and similar periods makes the estimator more robust [*Intuition*: emphasize units that are more similar to treated units]
  - Example: Effect of anti-smoking legislation on California (Abadie et al., 2010)
- Use of the weights may (Not always) improve the estimator's precision by removing systematic (predictable) parts of the outcome

Software Implementation: [Stata]: `sdid`; [R]: `synthdid`

"The synthetic control DID may gain its popularity in the future, especially in the area of public policy evaluation"

Nonlinear Diff-In-Diffs: Change-In-Changes

Nonlinear DID (Change-In-Changes)

Athey and Imbens (2006, Econometrica)

- The classical DID model assumes potential outcomes are separably additive:

$$Y^I_{gt} = g_I(D_{gt}) + U^I_{gt}, \text{ for } I \in \{0, 1\}$$

- Athey and Imbens (2006) generalizes the function of potential outcomes:

$$Y^0_{gt} = h(D_{gt}, U^0_{gt})$$

  Now the goal is to back out their distribution. Under very similar conditions as in DID, the distribution of Y^0_{11} can be identified as

$$F^0_{11}(y) = F_{10}[F^{-1}_{00}(F_{01}(y))]$$

  Linear DID: $E(Y^0_{11}|D_{11}=1) = E(Y^0_{10}|D_{11}=1) + [E(Y^0_{01}|D_{01}=0) - E(Y^0_{00}|D_{01}=0)]$

- Can be extended to count data models (patents), quantile treatment effects...
- Stata implementation: `cic`

Staggered Diff-In-Diffs

Why Staggered Diff-In-Diffs

- Canonical DID framework assumes that policy happen at one time
- But in many empirical settings, policy interventions enact in a staggered manner, generating variations in treatment timing: China's pilot programs, export/import decision, digitalization...
- The two-way fixed effects (TWFE) model were widely adopted by researchers:

$$Y_{it} = \delta D_{it} + \lambda_i + \lambda_t + \varepsilon_{it}$$

- It turns out our understanding of the TWFE estimator is pretty limited, and the interpretation of δ is unclear
- Several recent methodological papers start to deal with this problem

Understanding TWFE

3 groups and 3 periods (Goodman-Bacon, 2020)

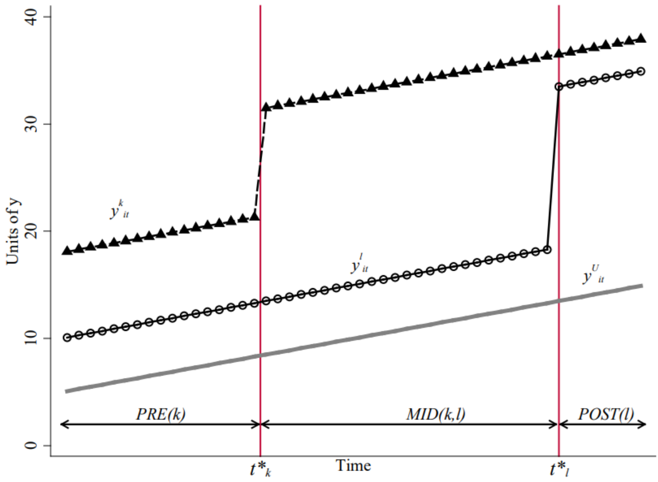

Figure 1: Staggered Difference-in-Differences: 3-by-3 Case

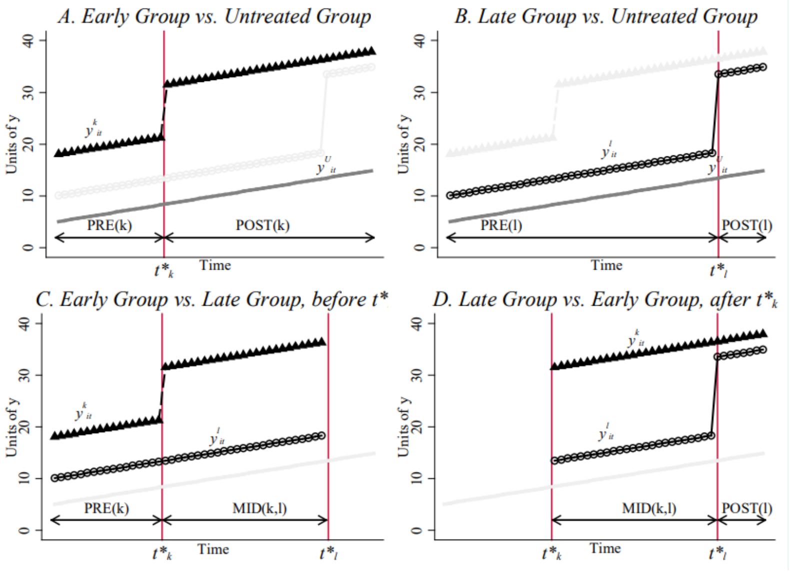

Figure 2: All four simple (2-by-2) DID Estimates

Bacon's Decomposition Theorem (Goodman-Bacon, 2020): *Assume that the data contain $k = 1, \cdots, K$ groups ordered by the treatment timing. The TWFE estimate is a weighted average of all possible two-by-two DID estimators:*

$$\hat{\delta}^{TWFE} = \sum_{k \neq U} \underbrace{w_{kU} \hat{\delta}_{kU}^{2\times 2}}_{\text{treat vs. never treated}} + \sum_{k \neq U} \sum_{\ell > k} w_{k\ell} [\underbrace{\mu_{k\ell} \hat{\delta}_{k\ell}^{2\times 2,k}}_{\text{early vs. late}} + \underbrace{(1-\mu_{k\ell}) \delta_{k\ell}^{2\times 2,\ell}}_{\text{late vs. early}}] \quad (9)$$

where $\sum_{k \neq U} w_{kU} + \sum_{k \neq U} \sum_{\ell > k} w_{k\ell} = 1$ and the weights are

$$w_{kU} = \frac{N_k N_U \bar{D}_k (1 - \bar{D}_k)}{v\hat{a}r(\tilde{D}_{it})}$$

$$w_{k\ell} = \frac{N_k N_\ell (\bar{D}_k - \bar{D}_l)[1 - (\bar{D}_k - \bar{D}_\ell)]}{v\hat{a}r(\tilde{D}_{it})}$$

$$\mu_{k\ell} = \frac{1 - \bar{D}_k}{1 - (\bar{D}_k - \bar{D}_\ell)}$$

where $\tilde{D}_{it} = (D_{it} - \bar{\bar{D}}) - (\bar{D}_i - \bar{\bar{D}}) - (\bar{D}_t - \bar{\bar{D}})$ with $\bar{\bar{D}} = \frac{1}{NT} \sum_i \sum_t D_{it}$, $\bar{D}_i = \sum_t D_{it}$ and $\bar{D}_t = \sum_i D_{it}$

Causal Interpretation of the TWFE Estimator

Goodman-Bacon shows the coefficient δ^{TWFE} can be interpreted as

$$\delta^{TWFE} = \underbrace{VWATT}_{\text{variance-weighted ATT}} + \underbrace{VWCT}_{\text{variance-weighted CT}} + \underbrace{\Delta ATT}_{\text{Bias}} \tag{10}$$

- $VWCT = 0$ if parallel trends assumption hold
- $\Delta ATT = \sum_{k \neq U} \sum_{\ell > k} \sigma_{k\ell}(1 - \mu_{k\ell})[ATT_k^{post(\ell)} - ATT_k^{mid(k,\ell)}]$ is the bias term: **Negative weights** occur when already-treated units act as controls, *changes in their treatment effects over time* get subtracted
- Interpretating the TWFE estimator:
  1. If treatment effects only vary across units: $\Delta ATT = 0$, but the weights may be far away from sample weights
  2. If treatment effects only vary across time: $\Delta ATT \neq 0$, yielding implausible estimates

Application

Impact of unilateral divorce law on female suicide rates (Stevenson and Wolfers, 2006)

- Unilateral (or no-fault) divorce allowed either spouse to end a marriage, redistributing property rights and bargaining power relative to fault-based divorce regimes
- Stevenson and Wolfers exploit *"the natural variation resulting from the different timing of the adoption of unilateral divorce laws"* in 37 states from 1969-1985
- Bacon replicates their study and provide new insights

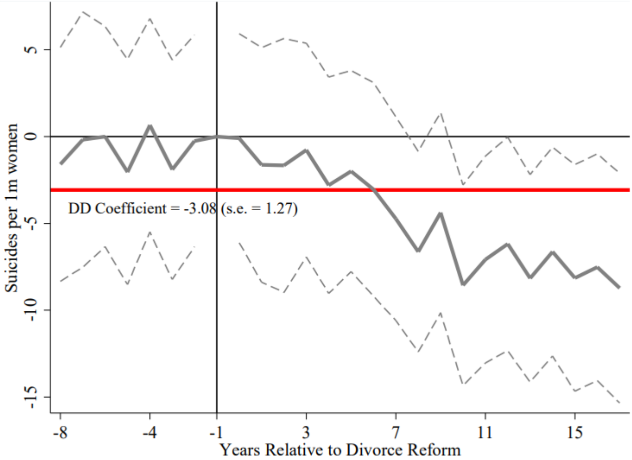

Figure 3: Event-Study and Difference-in-Differences Estimates

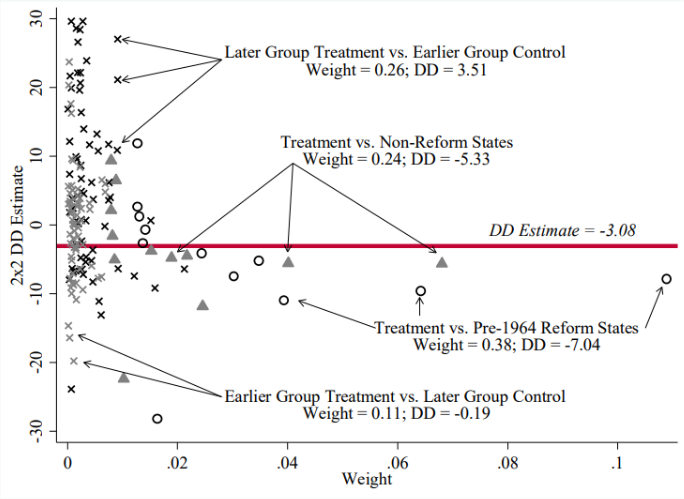

Figure 4: Bacon's Difference-in-Differences Decomposition for Unilateral Divorce and Female Suicide

Sensible Estimators for Staggered DID Designs

Sun and Abraham (2020, JOE)

- Consider group g units start to receive treatment at time $g = 2, \cdots, T$. $G_g \in \{0, 1\}$ and $G_{ig} \in \{0, 1\}$ are group indicator and unit-group indicator
- **Parallel Trend Assumption**: For all $t = 2, \cdots, T$, all $g = 2, \cdots, T$,

$$\begin{aligned} E[Y_t^0 - Y_{t-1}^0 | G_g = 1] &= E[Y_t^0 - Y_{t-1}^0 | C = 1] \\ &= E[Y_t^0 - Y_{t-1}^0] \end{aligned}$$

- S&A proposes the following interaction-weighted estimator for $ATT(g, t)$:

$$Y_{it} = \sum_{g=2}^{T-1} \sum_{e \neq 0} \delta_{ge} G_{ig} \mathbf{I}(t - G_i + 1 = e) + \lambda_i + \lambda_t + \varepsilon_{it}$$

  - $\hat{\delta}_{ge}$ is consistent for $ATT(g, t), t - g + 1 = e$
  - If no never treated group, dropping the last time period is unecessary
- Stata command: `eventstudyinteract`

de Chaisematin and D'Haultfoeuille (2020, AER)

- Parallel Trend Assumption as Sun and Abraham (2020)
- dC&D is interested in estimating instantaneous treatment effect:

$$\delta^{dc\&D} = E\left[\frac{\sum_i^N \sum_{t=2}^T G_{ig}(Y_{it}^1 - Y_{it}^0)}{\sum_i^N \sum_{t=2}^T G_{it}}\right]$$

$$\hat{\delta}^{dc\&D} = \sum_{g=2}^{T} \hat{P}(G_g = 1 | \text{Treated for period} \geq 1) \cdot \widehat{ATT}(g,g)$$

$$\widehat{ATT}^{ny}(g,t) = \underbrace{\frac{\sum_i G_{ig}(Y_{it} - Y_{ig-1})}{\sum_i G_{it}}}_{\textit{treated}} - \underbrace{\frac{\sum_i (1-D_{it})(1-G_{ig})(Y_{it} - Y_{ig-1})}{\sum_i (1-D_{it})(1-G_{ig})}}_{\textit{controls}}$$

- One needs to consider alternative causal parameters if interested in treatment effects dynamics
- Stata command: `did_multiplegt`

Callaway and Sant'Anna (2020, JOE)

- **Parallel Trend Assumption**: For all $g, s, t = 2, \cdots, T$ such that $t \geq g$, $s > t$:

$$\begin{aligned}
\text{(1) Never treated units:} \quad & E[Y_t^0 - Y_{t-1}^0 | G_g = 1] = E[Y_t^0 - Y_{t-1}^0 | C = 1] \\
\text{(2) Not-yet-treated units:} \quad & E[Y_t^0 - Y_{t-1}^0 | G_g = 1] = E[Y_t^0 - Y_{t-1}^0 | D_s = 1]
\end{aligned}$$

- C&S consider two groups of ATT:

$$\begin{aligned}
ATT^{never}(g,t) &= E[Y_t - Y_{g-1} | G_g = 1] - E[Y_t - Y_{g-1} | C = 1] \\
ATT^{ny}(g,t) &= E[Y_t - Y_{g-1} | G_g = 1] - E[Y_t - Y_{g-1} | D_t = 0, G_g = 0] \\
\widehat{ATT}^{ny}(g,t) &= \frac{\sum_i G_{ig}(Y_{it} - Y_{ig-1})}{\sum_i G_{ig}} - \frac{\sum_i C_i(1 - G_{ig})(Y_{it} - Y_{ig-1})}{\sum_i C_i}
\end{aligned}$$

- Stata Command: `csdid`

More on Staggered DID

- Borusyak, Jaravel, and Spiess (2021): Imputation based estimator
    - Stata command: `did_imputation`
- Athey and Imbens (2022, JOE): standard DID estimator still works for randomly assigned adoption date
  and more ......
- Looking forward:
    - Level of treatment heterogeneity is somewhat limited: by cohort-time
    - Non-absorbing treatment: e.g., high-tech firms certification in China
    - Researchers may see benefits by combining different identification methods