More on Diff-In-Diffs Methods: Recent Developments

Zhiyuan Chen

Empirical Methods
Renmin Business School

April 2025

Outline

- Propensity Score Matching-DID
- Semi-parametric DID
- Synthetic DID
- Staggered Difference-In-Differences
- *Non-linear DID
- Violations of Parallel Trends Assumption
- A Guide for Practitioners

Propensity Score Matching - DID

Potential Outcome Framework

General Idea of Matching

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

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

We aim to estimate

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

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:

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

Conditional Independence Assumption

General Idea of Matching

- **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):** $Y_i^0 \perp D_i | \mathbf{X}_i$
  - Under CIA, $Bias(\mathbf{X}_i) = 0$
  - CIA essentially states that D_i is *randomly assigned conditioning on observable 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."*

Regression vs. Matching

General Idea of Matching

- Statisticians more often use *matching methods* (Cochrane and Rubin, 1973); the idea is quite similar to conditioning on observeables in regressions

- "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: large $\Pr(D_i = 1|\mathbf{X}_i) \times [1 - \Pr(D_i = 1|\mathbf{X}_i)])$
- **Common Support Requirement:** No weights assigned to covariate cells containing no treated and control units: $0 < \Pr(D_i = 1|\mathbf{X}_i) < 1$

Propensity Score Matching

A Structural Interpretation by Heckman, Ichimura, and Todd (1998)

- 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:

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

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

- 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:

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

Propensity Score Matching based DID

Heckman et al. (1998, ReSTud)

- Recall the OLS estimator of 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 with D_{it}, 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:

$$\delta^{PSM-DID} = E(Y_{i1} - Y_{i0} | \mathbf{X}_i, D_{i1} = 1) - E(Y_{i1} - Y_{i0} | \mathbf{X}_i, D_{i1} = 0)$$

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

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

$$E[Y_{i1}^1 - Y_{i1}^0 | \mathbf{X}_i, D_{i1} = 1] = E[\rho_0 (Y_{i1} - Y_{i0}) | \mathbf{X}_i]$$

$$\text{where } \rho_0 = \frac{D_{i1} - P(D_{i1} = 1 | \mathbf{X}_i)}{P(D_{i1} = 1 | \mathbf{X}_i)[1 - P(D_{i1} = 1 | \mathbf{X}_i)]}$$

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

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 $\sum_{i=1}^{N_0} \hat{\omega}_i^{sc} Y_{it} \approx \frac{1}{N_1} \sum_{i=N_0+1}^{N} Y_{it}$ 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:

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

- 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:

$$Y_{gt}^{I} = g_I(D_{gt}) + U_{gt}^{I}, \text{ for } I \in \{0, 1\}$$

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

$$Y_{gt}^{0} = h(D_{gt}, U_{gt}^{0})$$

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

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

  Linear DID:

  $E(Y_{11}^{0}|D_{11} = 1) = E(Y_{10}^{0}|D_{11} = 1) + [E(Y_{01}^{0}|D_{01} = 0) - E(Y_{00}^{0}|D_{01} = 0)]$

- 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)

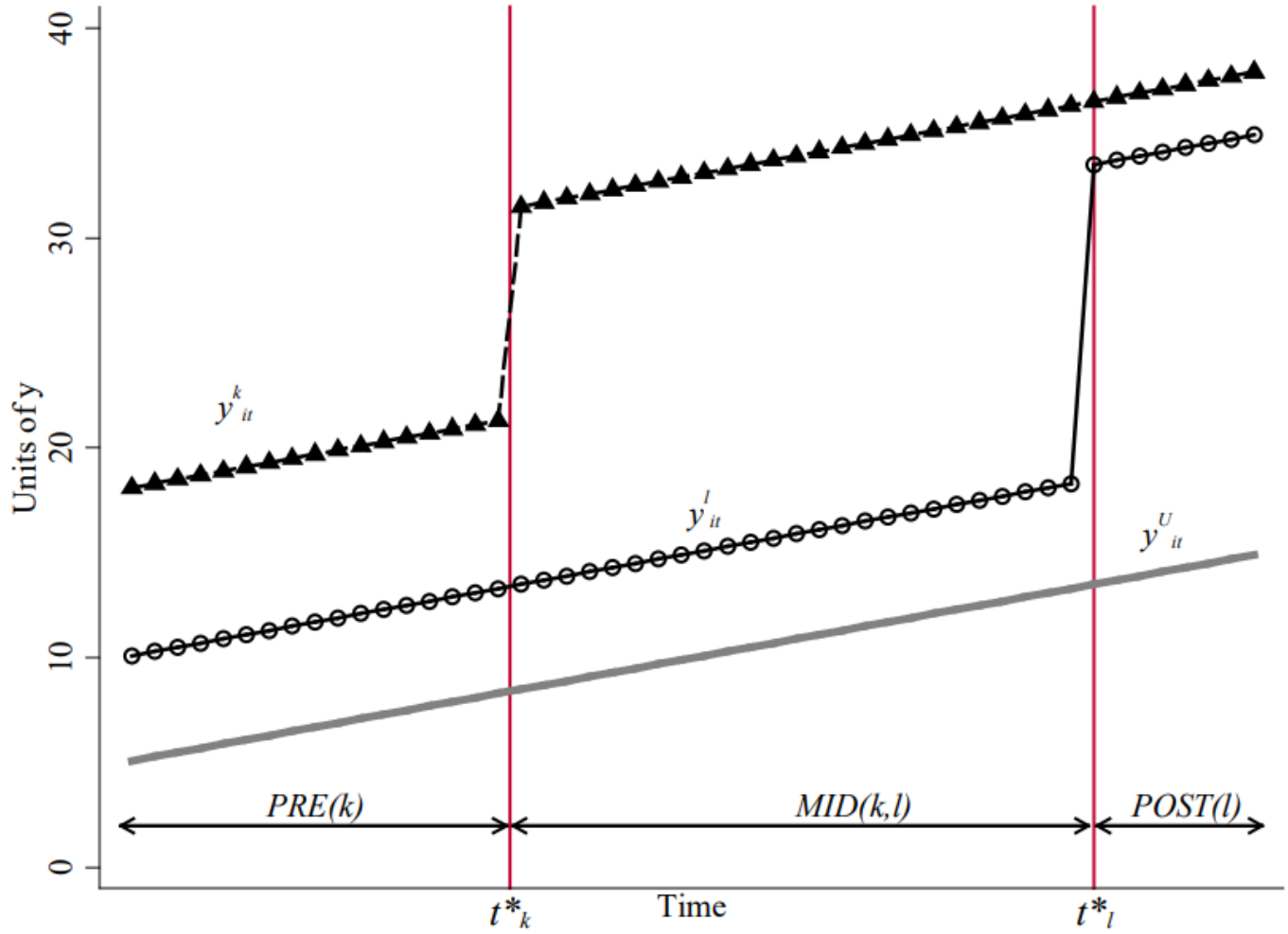


Figure 1: Staggered Difference-in-Differences: 3-by-3 Case

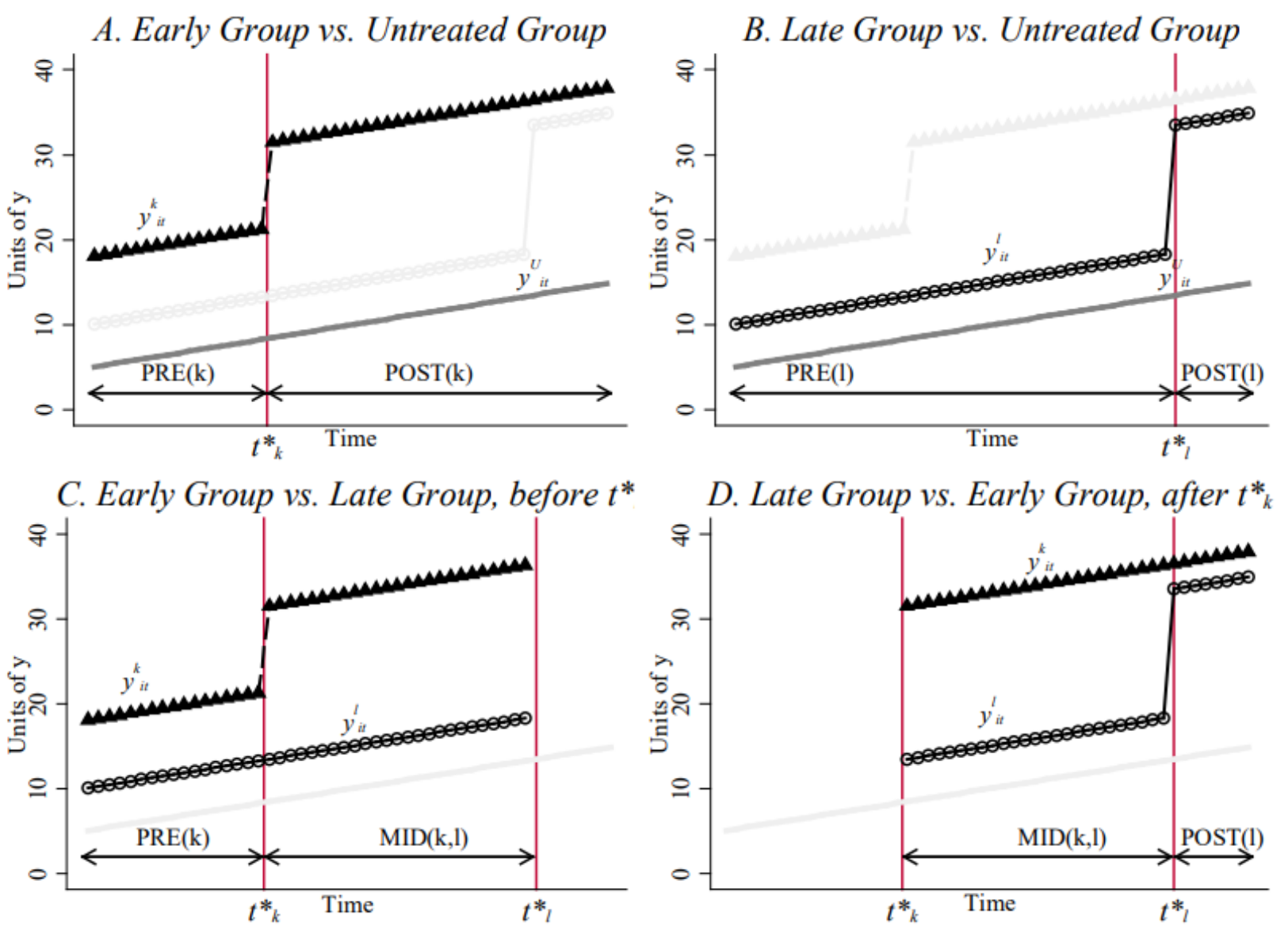


Figure 2: All four simple (2-by-2) DID Estimates

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:*

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

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

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

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

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

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

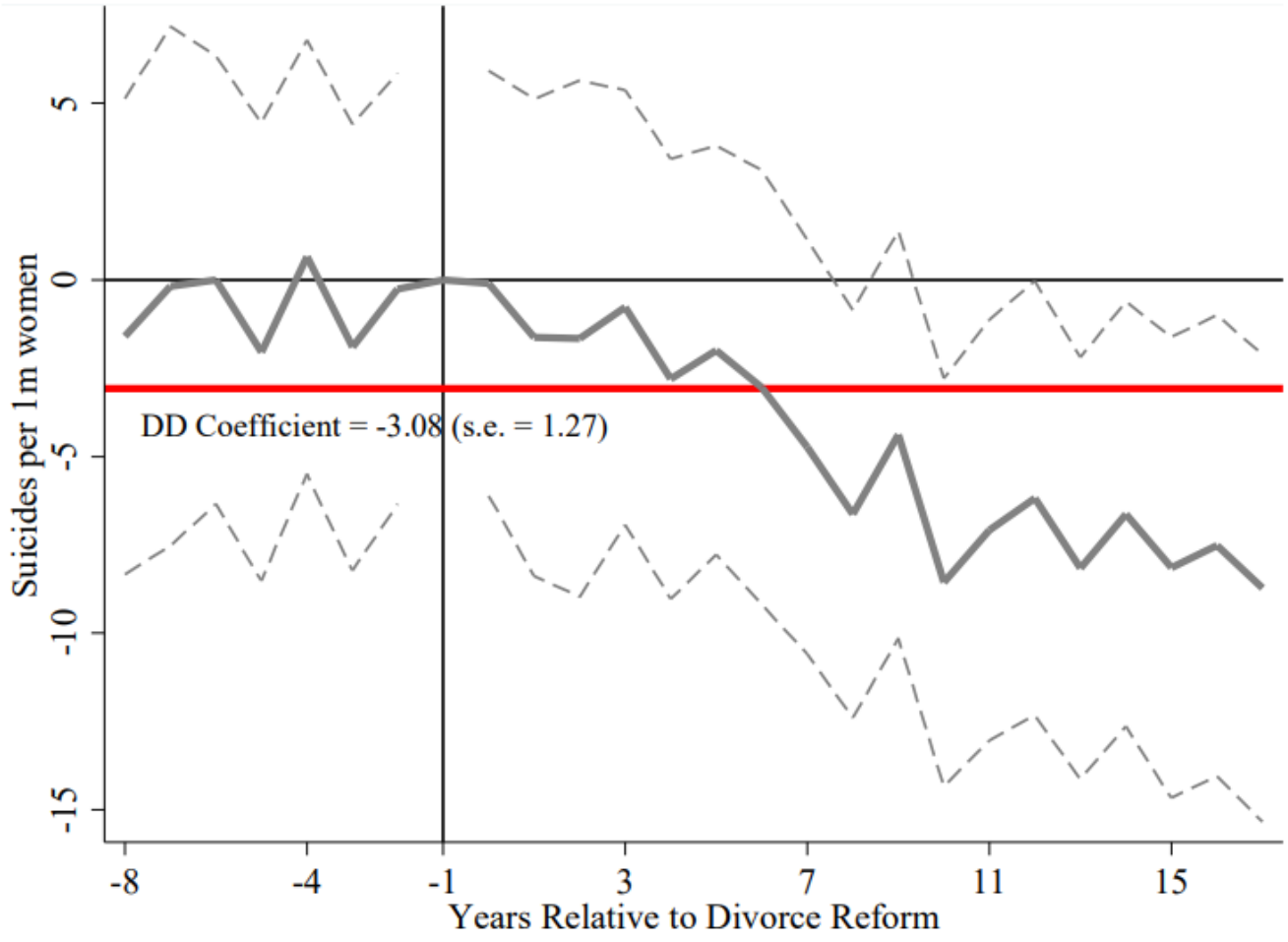


Figure 3: Event-Study and Difference-in-Differences Estimates

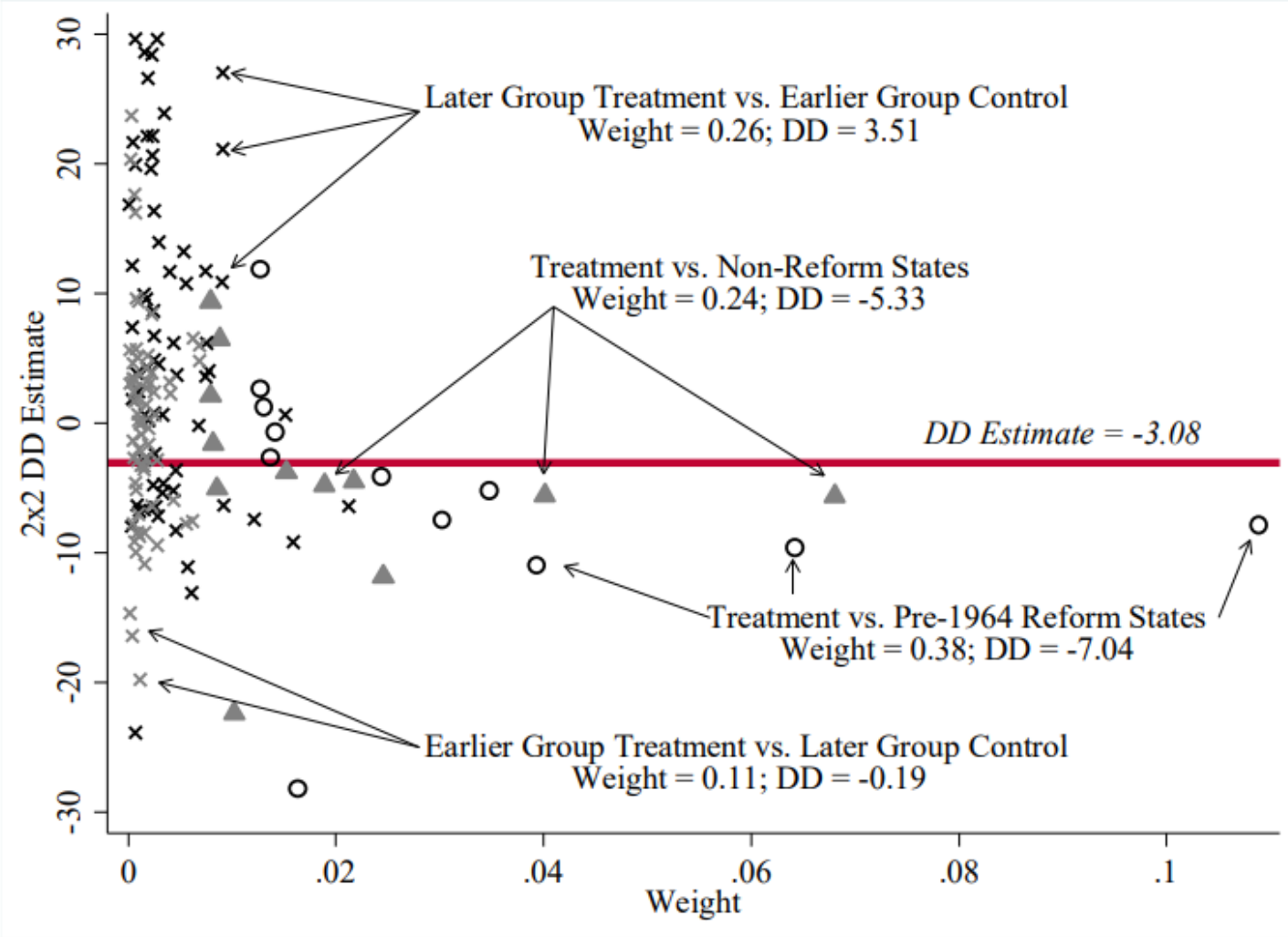


Figure 4: Bacon's Difference-in-Differences Decomposition for Unilateral Divorce and Female Suicide

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 $t \geq g$, $s > t$:

  (1) Never treated units: $E[Y_t^0 - Y_{t-1}^0 | G_g = 1] = E[Y_t^0 - Y_{t-1}^0 | C = 1]$

  (2) Not-yet-treated units: $E[Y_t^0 - Y_{t-1}^0 | G_g = 1] = E[Y_t^0 - Y_{t-1}^0 | D_s = 1]$

- C&S consider two groups of ATT:

$$ATT^{never}(g,t) = E[Y_t - Y_{g-1} | G_g = 1] - E[Y_t - Y_{g-1} | C = 1]$$

$$ATT^{ny}(g,t) = E[Y_t - Y_{g-1} | G_g = 1] - E[Y_t - Y_{g-1} | D_t = 0, G_g = 0]$$

$$\widehat{ATT}^{ny}(g,t) = \frac{\sum_i G_{ig}(Y_{it} - Y_{ig-1})}{\sum_i G_{ig}} - \frac{\sum_i C_i(1 - G_{ig})(Y_{it} - Y_{ig-1})}{\sum_i C_i}$$

- 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

Violations of Parallel Trends Assumption

Based on Roth's lecture notes

Violations of PT

- Remember that in the canonical DiD model we had:
  - Two periods and a common treatment date
  - Identification from parallel trends and no anticipation
  - A large number of clusters for inference
- A second literature has focused on relaxing the second assumption: **what if parallel trends may be violated?**
- The ideas from this literature apply even if there is non-staggered timing, although as we'll see, many of the tools can be applied with staggered timing as well. Large number of clusters is maintained throughout.

Violations of PT

- Three substrands of this literature:
  - Parallel trends only conditional on covariates
  - **Testing for violations of (conditional) parallel trends**
  - **Sensitivity analysis and bounding exercises**
- We will focus on the latter two

Why might we be skeptical of PT?

- Recall PT requires the selection bias to be constant over time. Why might we be skeptical of this?

Why might we be skeptical of PT?

- Recall PT requires the selection bias to be constant over time. Why might we be skeptical of this?
- There might be different confounding factors in period 1 as in period 0
  - E.g. states that pass a minimum wage increase might also change unemployment insurance at the same time
  - Then UI is a confound in period 1 but not in period 0

Why might we be skeptical of PT?

- Recall PT requires the selection bias to be constant over time. Why might we be skeptical of this?
- There might be different confounding factors in period 1 as in period 0
  - E.g. states that pass a minimum wage increase might also change unemployment insurance at the same time
  - Then UI is a confound in period 1 but not in period 0
- The same confounding factors may have different effects on the outcome in different time periods
  - Suppose people who enroll in a job training program are more motivated to find a job
  - Motivation might matter more in a bad economy than in a good economy

Why might we be skeptical of PT? Part 2

- Another reason to be skeptical of parallel trends is "selection bias being constant" depends on the **functional form** chosen for the outcome
- Consider an example:
  - In period 0, all control units have outcome 10; all treated units have outcome 5.
  - In period 1, all control units have outcome 15.
  - If treatment hadn't occurred, would treated units' outcome have increased by 5 also (PT in levels)?
  - Or would they have increased by 50% ($\sim$ PT in logs)?

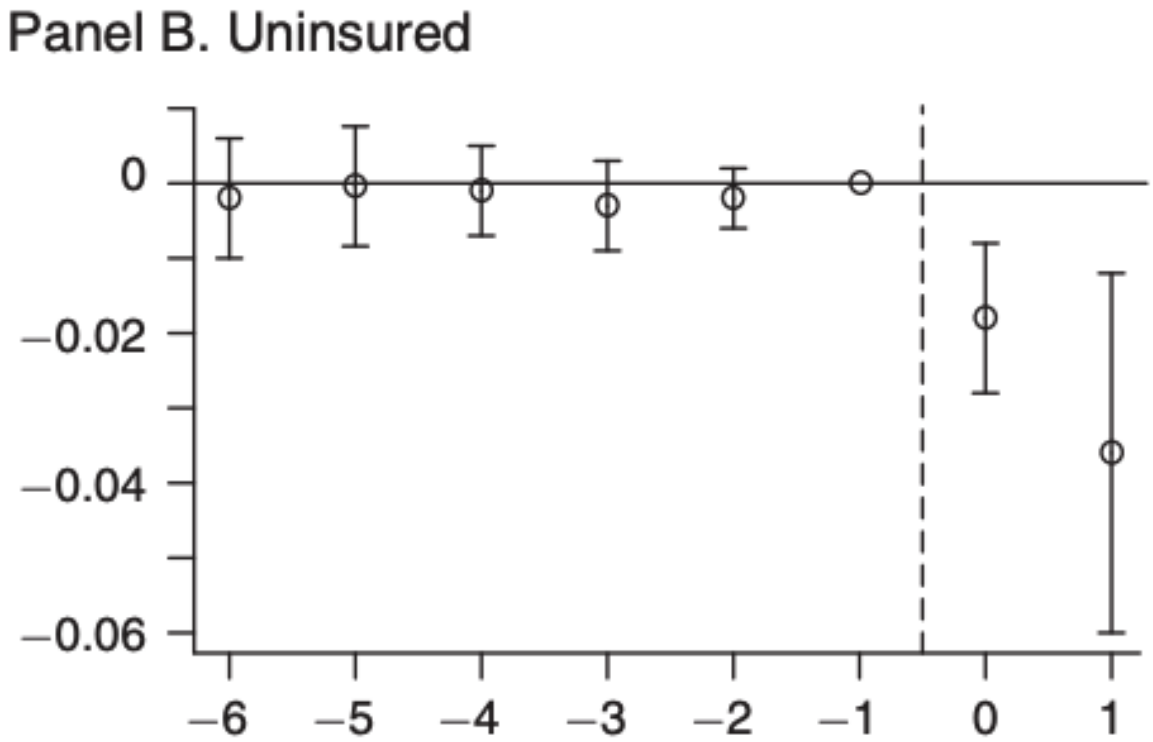
Roth and Sant'Anna (2023) show that PT will depend on **functional form** unless:

- **Randomization:** treated and control group have same dist. of $Y(0)$ in each period
- **No time effects:** distribution of $Y(0)$ doesn't change over time for either group
- **A hybrid**: θ fraction of the population is as good as randomized; the other $1-\theta$ fraction has no time effects.

Absent these conditions, PT will be violated for at least some functional form; often hard to know if we chose the right one!

Pre-trends to the rescue...

- Luckily, in most DiD applications we have several periods before anyone was treated
- We can test whether the groups were **moving in parallel prior to the treatment**
  - If so, *then assumption that confounding factors are stable seems more plausible*
  - If not, *then it's relatively implausible that would have magically started moving in parallel after treatment date*
- Testing for pre-trends provides a natural plausibility check on the parallel trends assumption



- Carey, Miller, and Wherry (2020) do a DiD comparing states who expanded Medicaid in 2014 to states that didn't.
- Report results from "event-study" regression:

$$Y_{its} = \phi_t + \lambda_s + \sum_{r \neq -1} D_i \times 1[t = 2014 + r] \cdot \beta_r + \varepsilon_{it}$$

where Y_{its} is insurance for person i in year t in state s, and $D_i = 1$ if in an expansion state.

- Testing for pre-existing trends is a very natural way to assess the plausibility of the PT assumption
- But it also has several *limitations*, highlighted in recent work grayFreyaldenhoven et al., 2019; Kahn-Lang and Lang, 2020; Bilinski and Hatï¬¬*eld*, 2018; *Roth*, 2022
- Two main possible solutions:
  - Roth (2022 AER:I, "Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends")
  - Rambachan and Roth (2023 RESTUD, "A More Credible Approach to Parallel Trends")

Overview of Limitations

- Parallel pre-trends doesn't necessarily imply parallel (counterfactual) post-treatment trends
  - If **other policies change at the same time as the one of interest**—e.g., mininum wage and unemployment insurance reform together —can produce parallel pre-trends but non-parallel post-trends
  - Likewise, could be that treated/control groups are differentially exposed to recessions, but there is only a recession in the post-treatment period

Overview of Limitations

- Parallel pre-trends doesn't necessarily imply parallel (counterfactual) post-treatment trends
  - If **other policies change at the same time as the one of interest**—e.g., mininum wage and unemployment insurance reform together —can produce parallel pre-trends but non-parallel post-trends
  - Likewise, could be that treated/control groups are differentially exposed to recessions, but there is only a recession in the post-treatment period
- **Low power:** even if pre-trends are non-zero, we may fail to detect it statistically

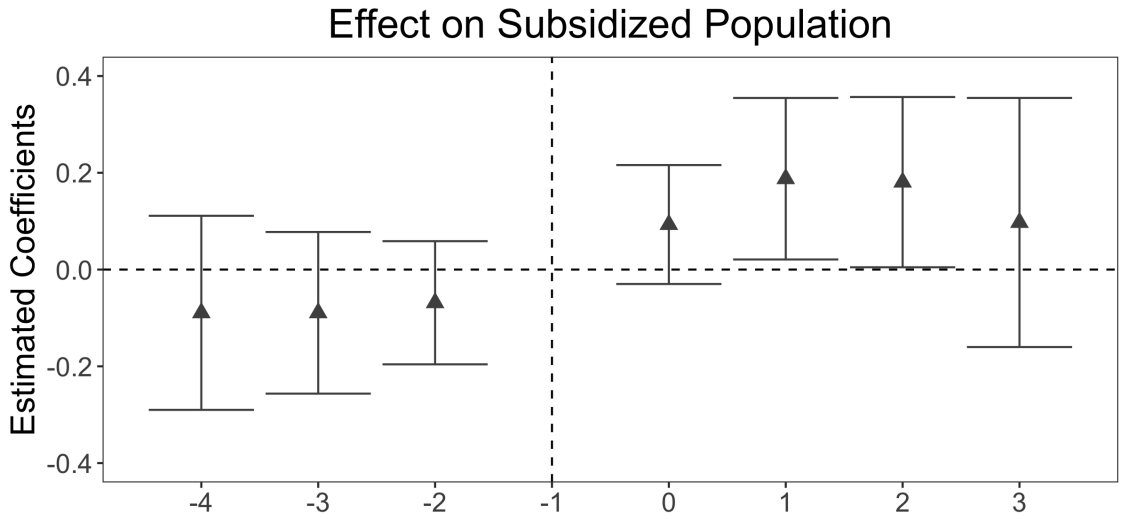
Overview of Limitations

- Parallel pre-trends doesn't necessarily imply parallel (counterfactual) post-treatment trends
  - If **other policies change at the same time as the one of interest**—e.g., mininum wage and unemployment insurance reform together —can produce parallel pre-trends but non-parallel post-trends
  - Likewise, could be that treated/control groups are differentially exposed to recessions, but there is only a recession in the post-treatment period
- **Low power:** even if pre-trends are non-zero, we may fail to detect it statistically
- **Pre-testing issues:** if we only analyze cases without statistically significant pre-trends, this introduces a form of selection bias (which can make things worse)

Overview of Limitations

- Parallel pre-trends doesn't necessarily imply parallel (counterfactual) post-treatment trends
  - If **other policies change at the same time as the one of interest**—e.g., mininum wage and unemployment insurance reform together — can produce parallel pre-trends but non-parallel post-trends
  - Likewise, could be that treated/control groups are differentially exposed to recessions, but there is only a recession in the post-treatment period
- **Low power:** even if pre-trends are non-zero, we may fail to detect it statistically
- **Pre-testing issues:** if we only analyze cases without statistically significant pre-trends, this introduces a form of selection bias (which can make things worse)
- If we fail the pre-test, what next? May still want to write a paper (especially if violation is "small")

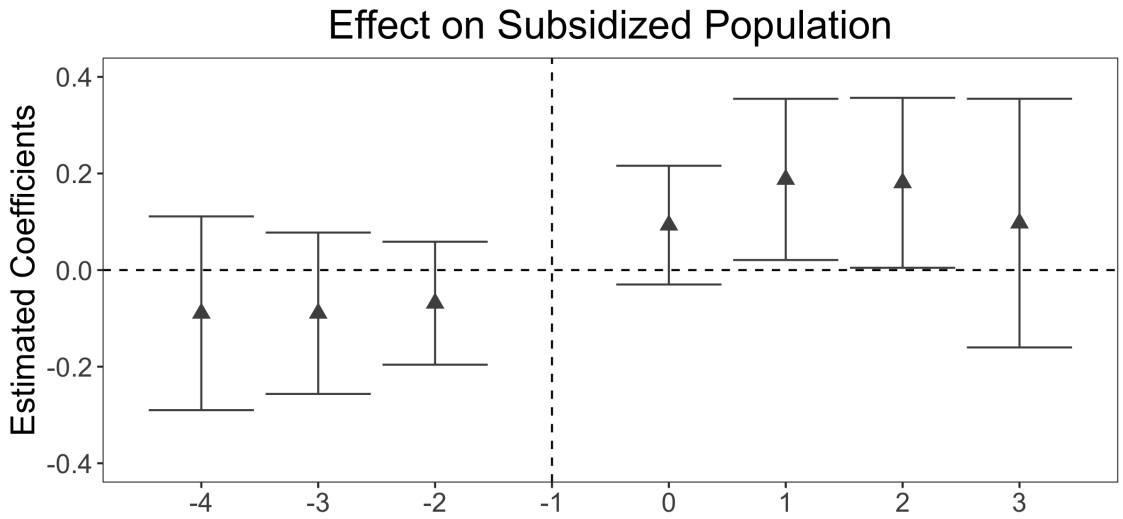
Issue 1 - Low Power



- He & Wang (2017, *AEJ: Applied*) study impacts of placing college grads as village officials in China
- Use an "event-study" approach comparing treated and untreated villages

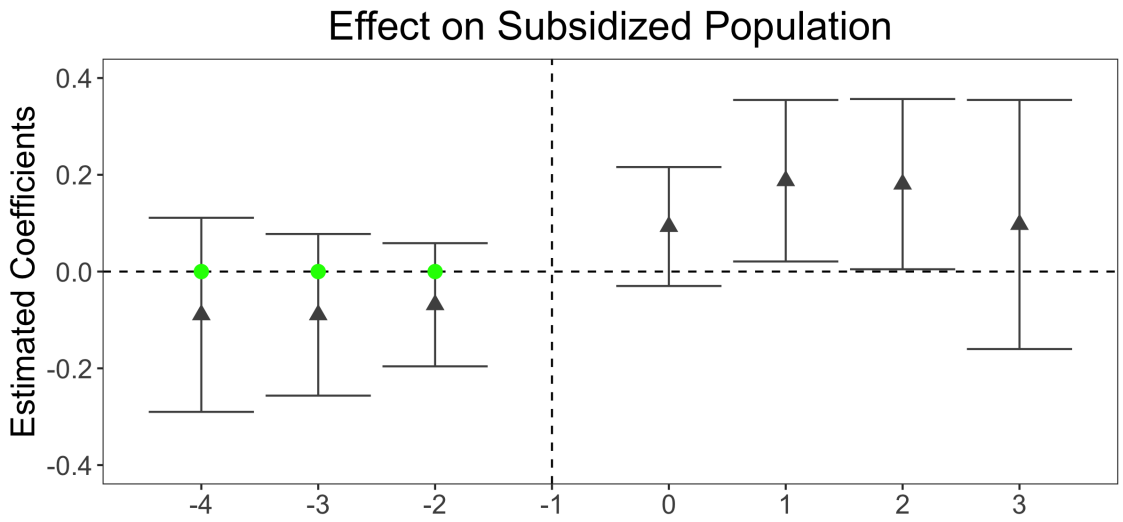
$$Y_{it} = \sum_{k \neq -1} D_{it}^{k} \beta_k + \alpha_i + \phi_t + \varepsilon_{it}$$

Issue 1 - Low Power



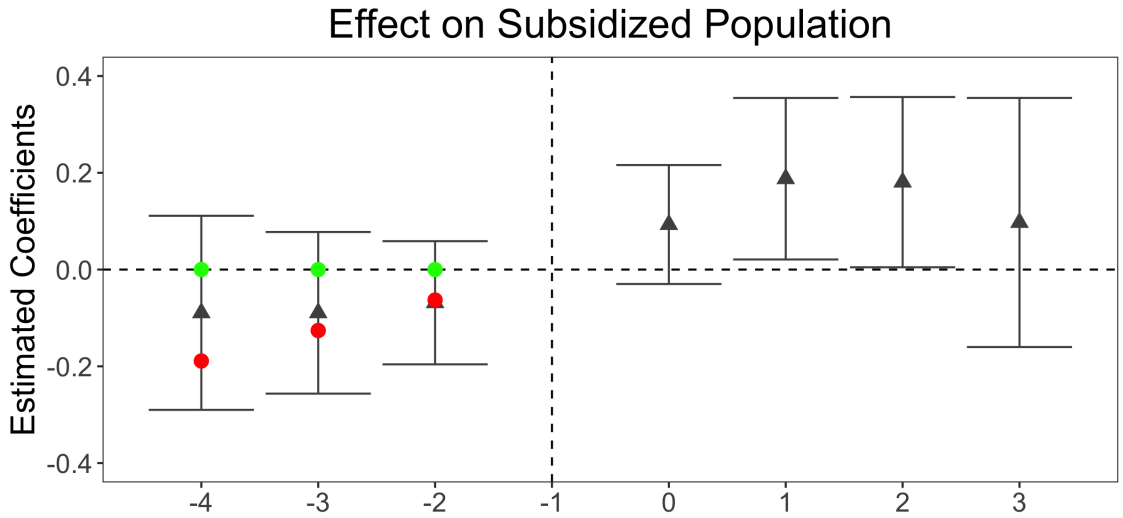
"The estimated coefficients on the leads of treatment ... are statistically indifferent from 0. ... We conclude that the pretreatment trends in the outcomes in both groups of villages are similar, and villages without CGVOs can serve as a suitable control group for villages with CGVOs in the treatment period." (He and Wang, 2017)

Issue 1 - Low Power



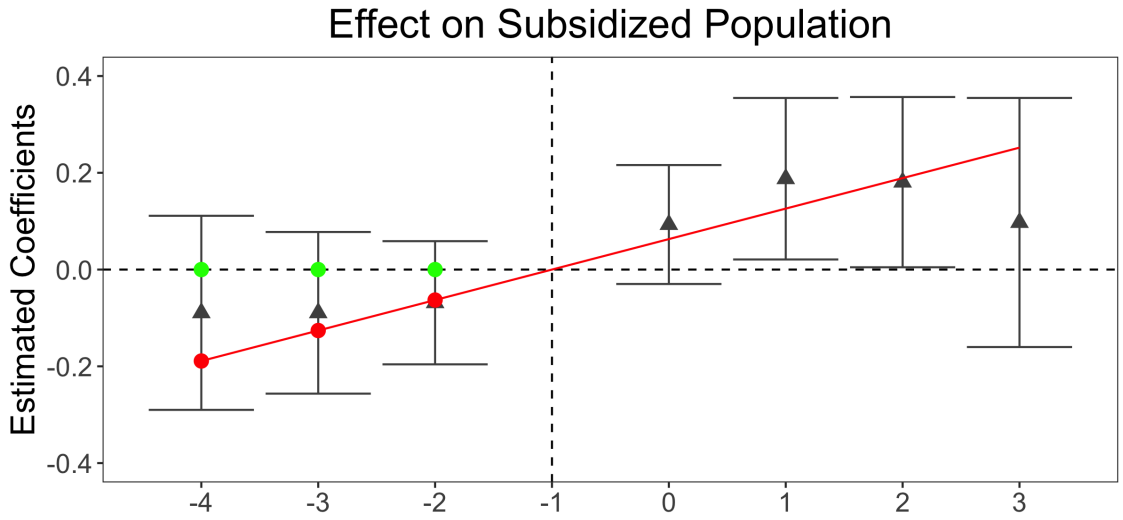
- P-value for H_0 := green dots (no pre-trend): 0.81

Issue 1 - Low Power



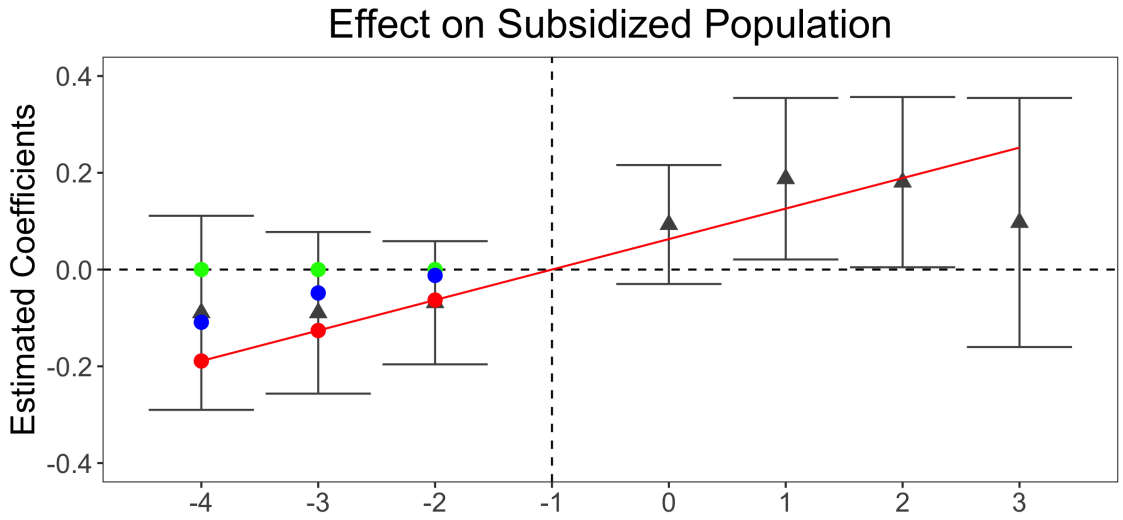
- P-value for $H_0 :=$ green dots (no pre-trend): 0.81
- P-value for $H_0 :=$ red dots: 0.81

Issue 1 - Low Power



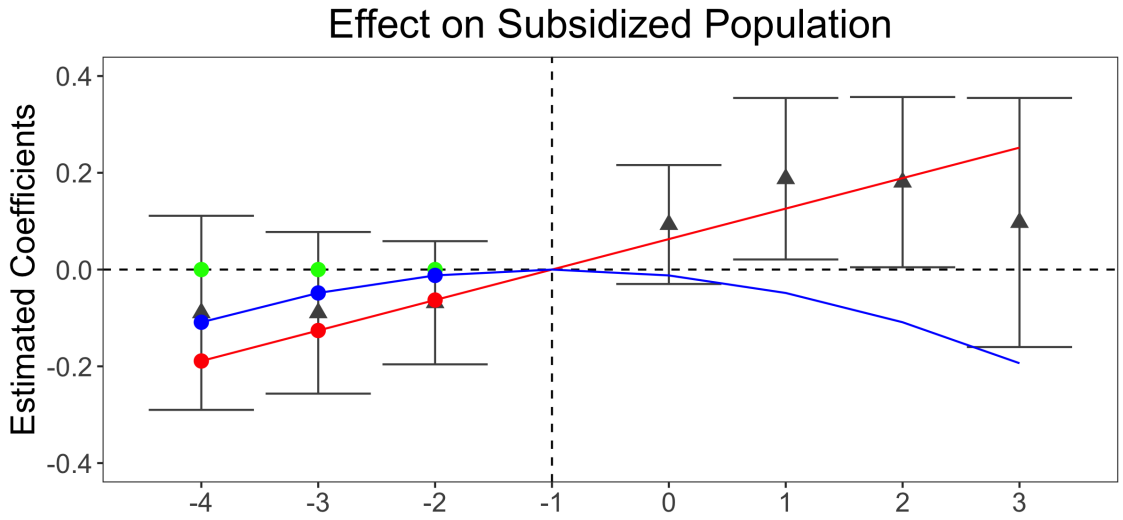
- P-value for $H_0 :=$ green dots (no pre-trend): 0.81
- P-value for $H_0 :=$ red dots: 0.81

Issue 1 - Low Power



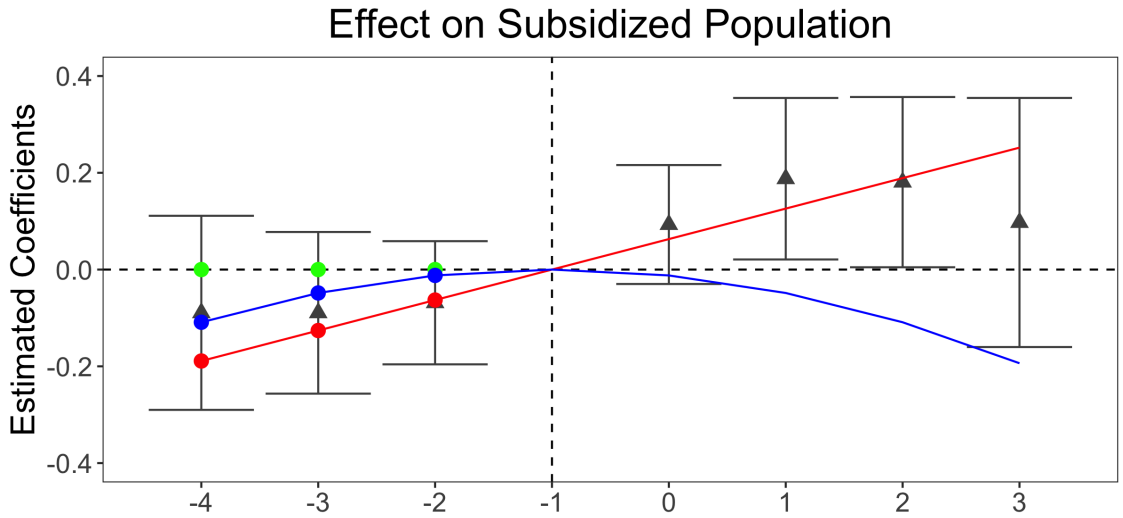
- P-value for $H_0 :=$ green dots (no pre-trend): 0.81
- P-value for $H_0 :=$ red dots: 0.81
- P-value for $H_0 :=$ blue dots: 0.81

Issue 1 - Low Power



- P-value for $H_0 :=$ green dots (no pre-trend): 0.81
- P-value for $H_0 :=$ red dots: 0.81
- P-value for $H_0 :=$ blue dots: 0.81

Issue 1 - Low Power



- P-value for $H_0 :=$ green dots (no pre-trend): 0.81
- P-value for $H_0 :=$ red dots: 0.81
- P-value for $H_0 :=$ blue dots: 0.81
- We can't reject zero pre-trend, but we also can't reject pre-trends that under smooth extrapolations to the post-treatment period would produce substantial bias

More systematic evidence

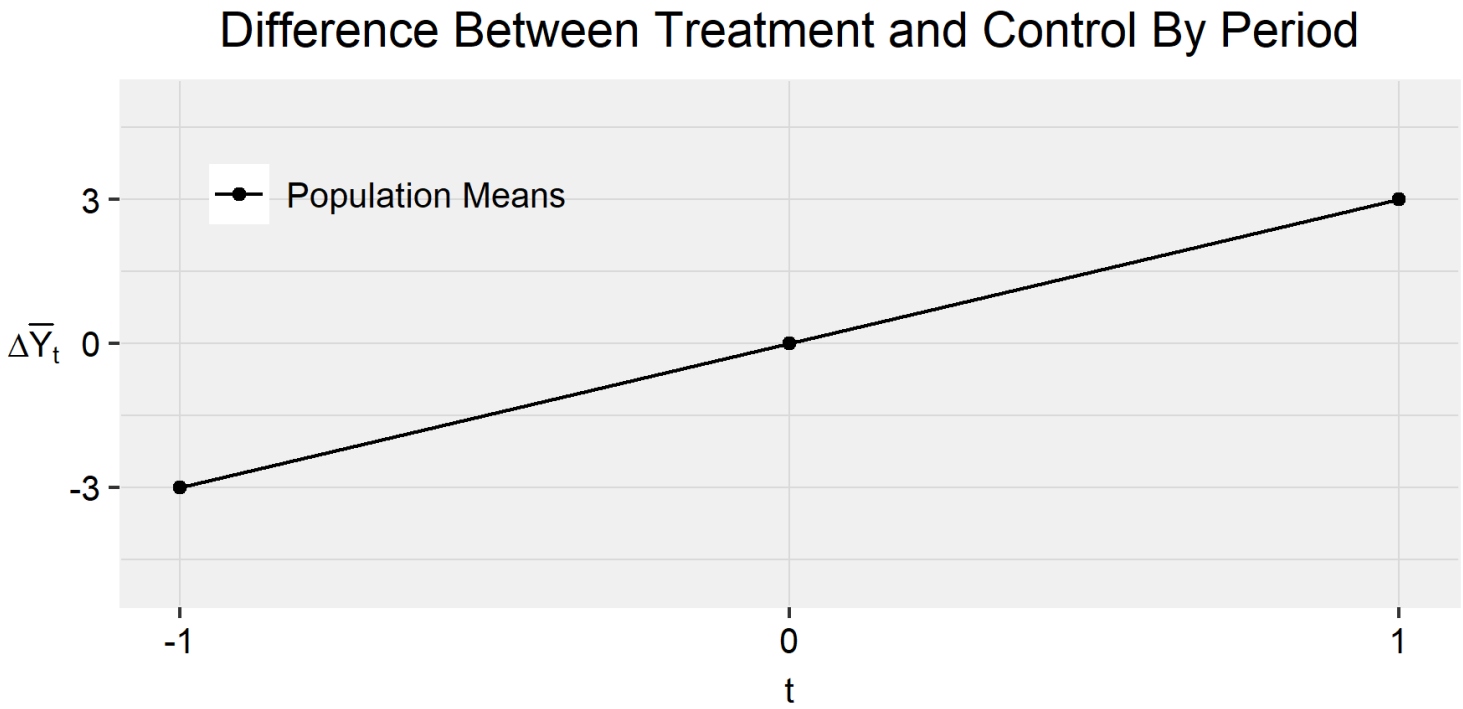
- Roth (2022, ReStud): simulations calibrated to papers published in *AER*, *AEJ: Applied*, and *AEJ: Policy* between 2014 and mid-2018
  - 70 total papers contain an event-study plot; focus on 12 w/available data
- Evaluate properties of standard estimates/CIs under linear violations of parallel trends against which conventional tests have limited power (50 or 80%):
  1. Bias often of magnitude similar to estimated treatment effect
  2. Confidence intervals substantially undercover in many cases
  3. Distortions from pre-testing can further exacerbate these issues

Issue 2 - Distortions from Pre-testing

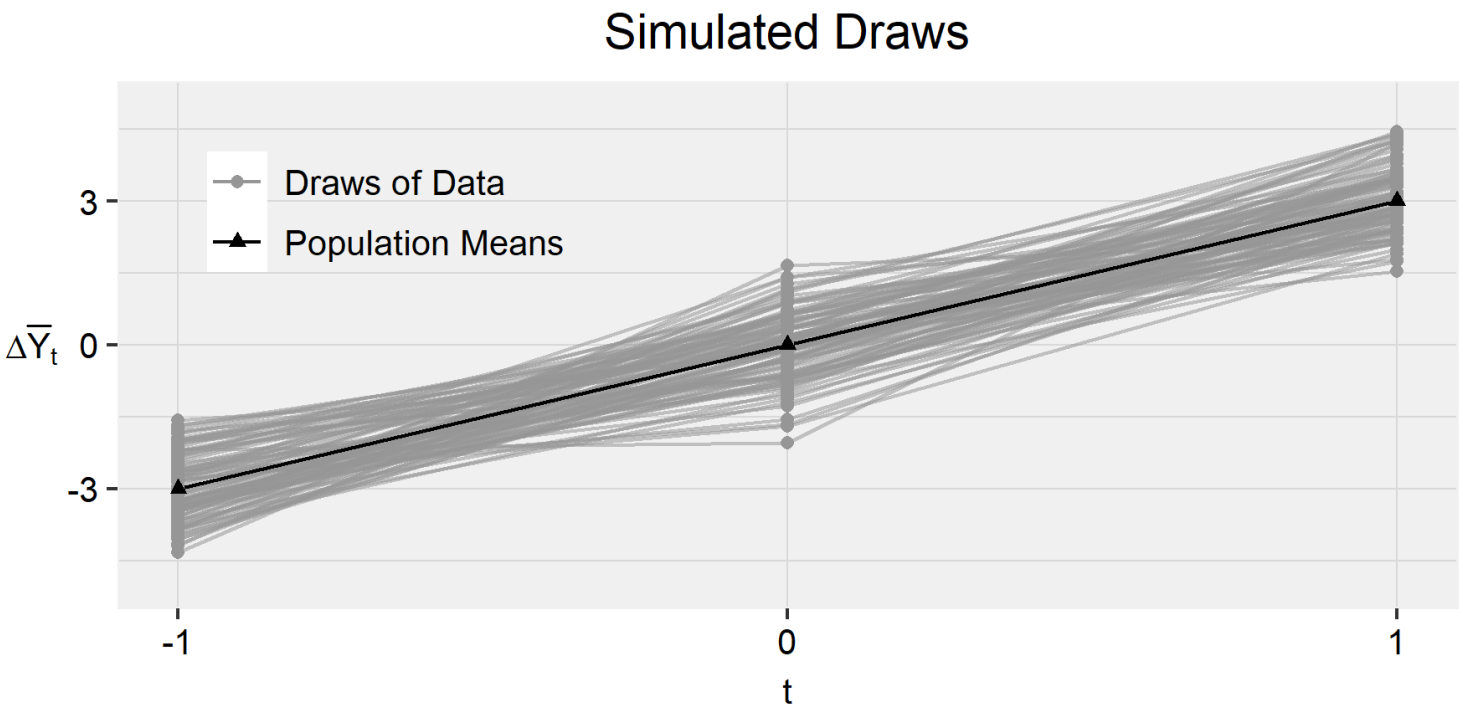
- When parallel trends is violated, we will sometimes fail to find a significant pre-trend
- But the draws of data where this happens are a **selected sample**. This is known as *pre-test bias*.
- Analyzing this selected sample introduces additional statistical issues, and can make things worse!

Stylized Three-Period DiD Example

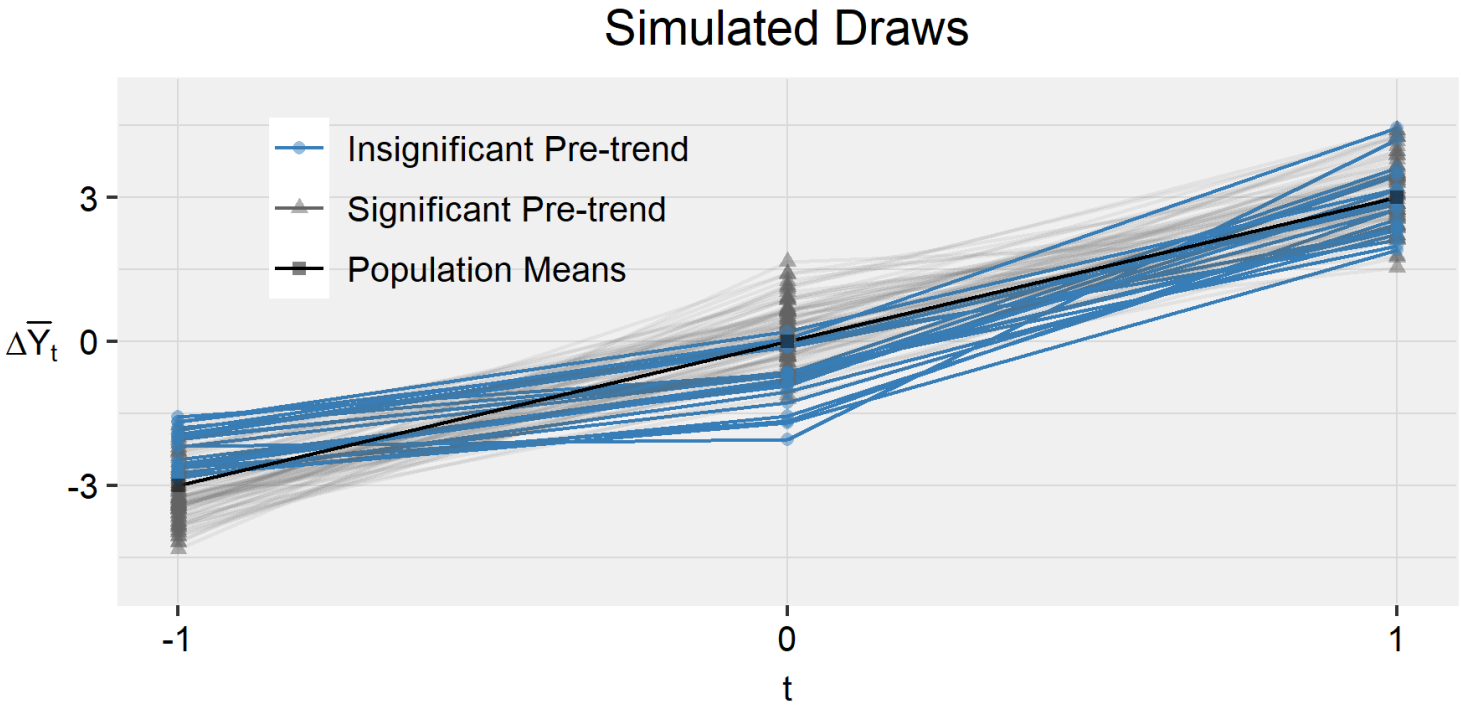
- Consider a 3-period model ($t = -1, 0, 1$) where treatment occurs in last period
- No causal effect of treatment: $Y_{it}^0 = Y_{it}^1$ in all periods
- In population, treatment group is on a `linear trend` relative to the control group with slope δ
  - Control group mean in period t: $E[Y_{it}^0 \mid \text{Control group}] = 0$
  - Treatment group mean in period t: $E[Y_{it}^0 \mid \text{Treated group}] = \delta \cdot t$
- Simulate from this model with Y_{it} equal to the group mean plus independent normal errors



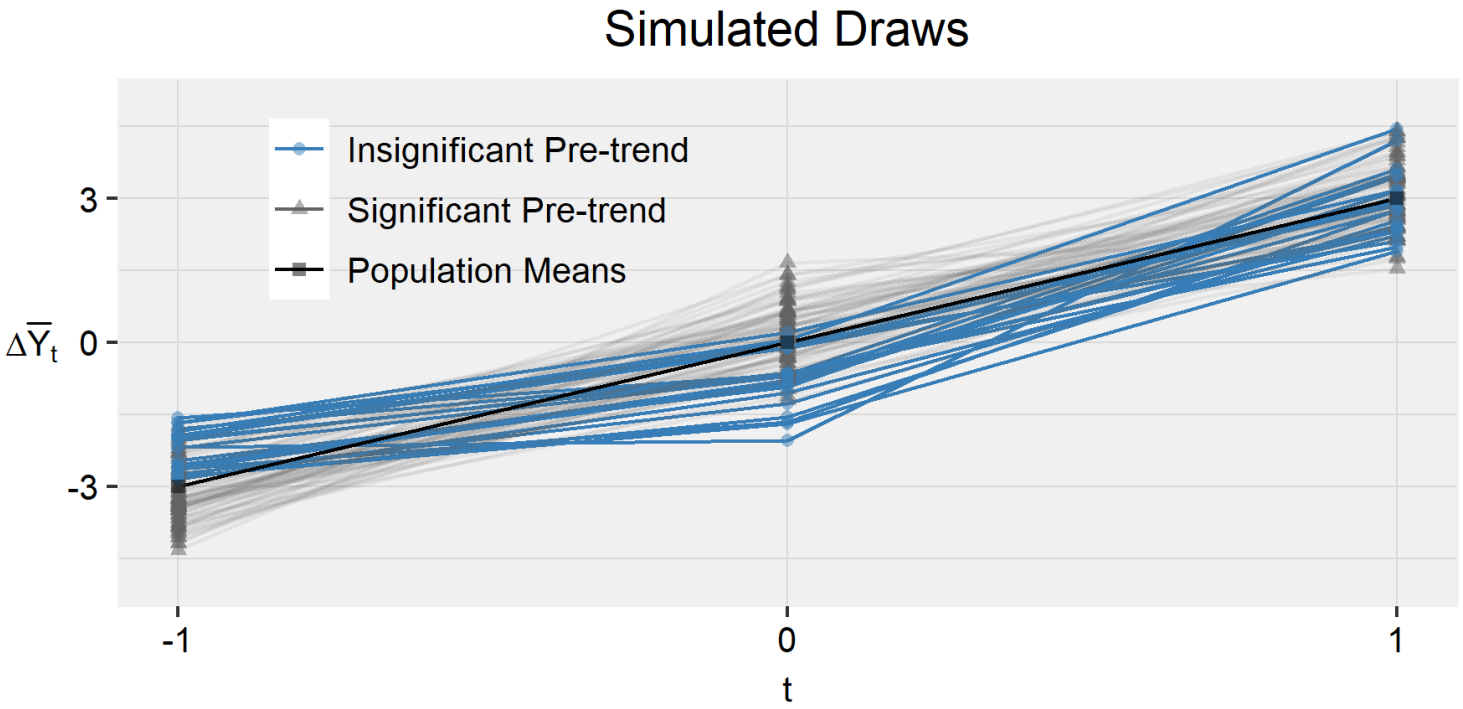
- Example: In population, there is a linear difference in trend with slope 3



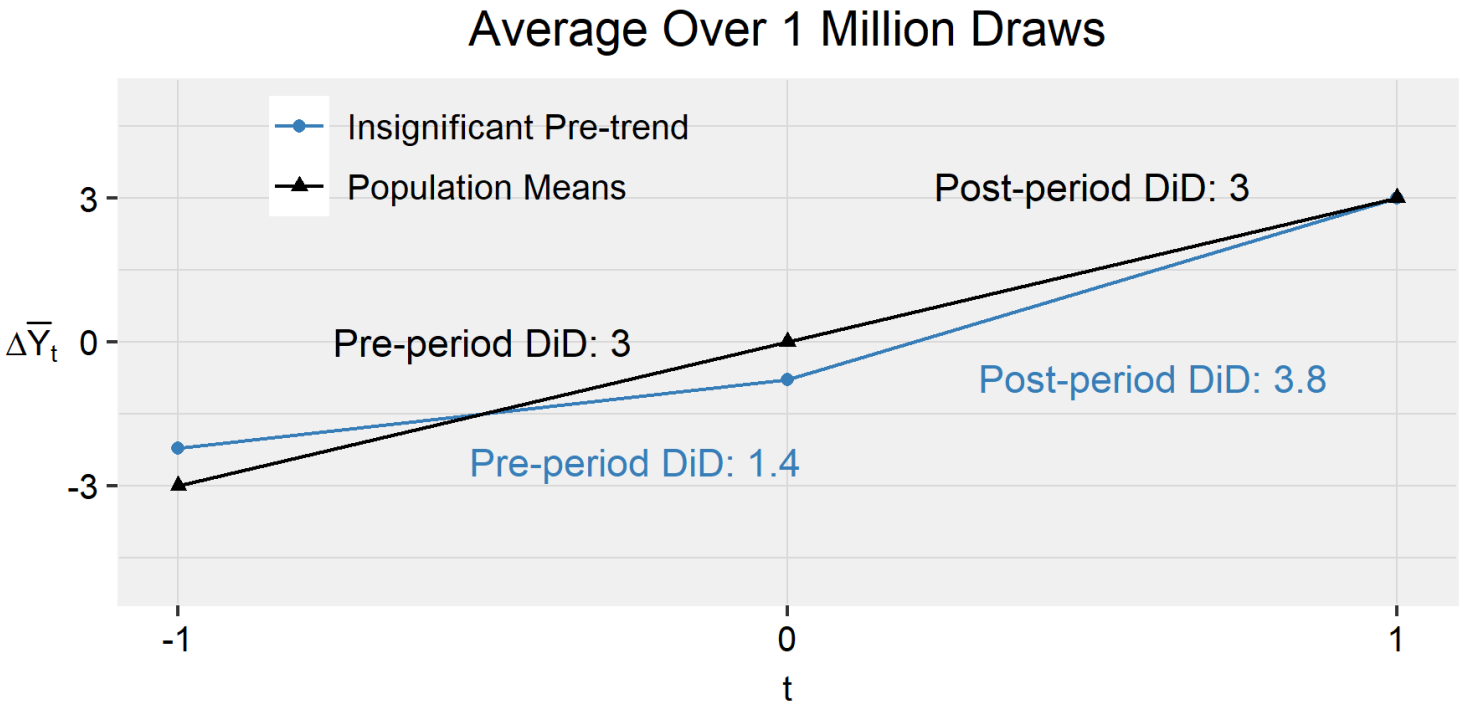
- Example: In population, there is a linear difference in trend with slope 3
- In actual draws of data, there will be noise around this line



- Example: In population, there is a linear difference in trend with slope 3
- In some of the draws of the data, highlighted in blue, the difference between period -1 and 0 will be insignificant



- In some of the draws of the data, highlighted in blue, the difference between period -1 and 0 will be insignificant
- In the insignificant draws, we tend to underestimate the difference between treatment and control at $t = 0$



- In the insignificant draws, we tend to underestimate the difference between treatment and control at $t = 0$
- As a result, the DiD between period 0 and 1 tends to be particularly large when we get an insignificant pre-trend

To Summarize

What are the Limitations of Pre-trends Testing?

1. Parallel pre-trends do not necessarily imply parallel counterfactual post-treatment trends
2. Low Power – May not find significant pre-trend even if PT is violated
3. Pre-testing Issues – Selection bias from only analyzing cases with insignificant pre-trend
4. If reject pre-trends test, what comes next?

To Summarize

What are the Limitations of Pre-trends Testing?

1. Parallel pre-trends do not necessarily imply parallel counterfactual post-treatment trends
2. Low Power – May not find significant pre-trend even if PT is violated
3. Pre-testing Issues – Selection bias from only analyzing cases with insignificant pre-trend
4. If reject pre-trends test, what comes next?

What Can We Do About It?

1. Diagnostics of power and distortions from pre-testing (Roth, 2022, "Pre-Test with Caution..."). See `pretrends` package. Details
2. Formal sensitivity analysis that avoids pre-testing (Rambachan and Roth, 2023, "A More Credible Approach..."). See `HonestDiD` package.

"A More Credible Approach to Parallel Trends"

- The intuition motivating pre-trends testing is that if we knew the true pre-trends, that would be informative about the counterfactual post-treatment diffs in trends
- Formalize this by imposing restrictions that allow us to learn from the pre-trends — intuitively, the counterfactual difference in trends can't be "too different" than the pre-trend
- This allows us to bound the treatment effect and obtain uniformly valid ("honest") confidence sets under the imposed restrictions
- Enables **sensitivity analysis:** How different would the counterfactual trend have to be from the pre-trends to negate a conclusion (e.g. a positive effect)?

Restrictions on Violations of PT

- Consider the 3-period model ($t = -1, 0, 1$) where treatment occurs in last period
- Let δ_1 be the violation of PT:

$$\delta_1 = Y_{i1}^0 - Y_{i0}^0 \,|\, D_i = 1 - Y_{i1}^0 - Y_{i0}^0 \,|\, D_i = 0$$

- We don't directly identify δ_1, but we do identify its pre-treatment analog, δ_{-1}:

$$\delta_{-1} = Y_{i,-1}^0 - Y_{i0}^0 \,|\, D_i = 1 - Y_{i,-1}^0 - Y_{i0}^0 \,|\, D_i = 0$$

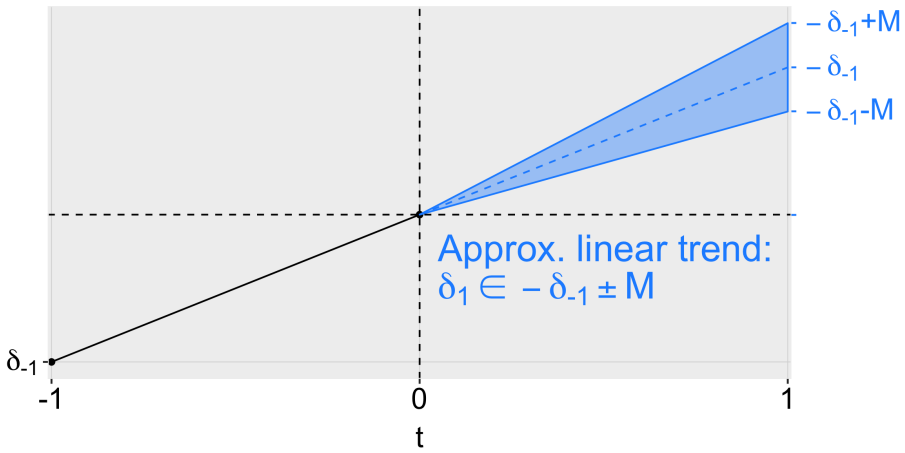
- **Key idea**: restrict possible values of δ_1 given δ_{-1}
  Intuitively, counterfactual trend **can't be too different from pre-trend**

Examples of Restrictions on δ

- **Bounds on relative magnitudes:** Require that $|\delta_1| \leq \bar{M}|\delta_{-1}|$

Examples of Restrictions on δ

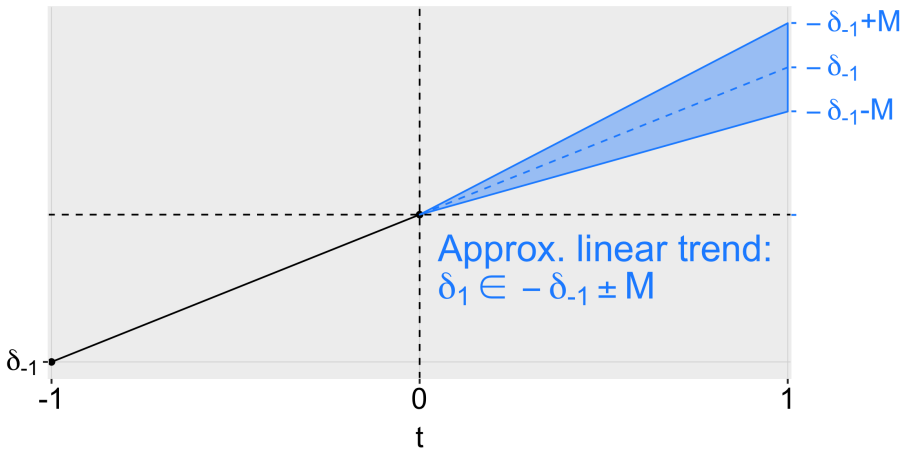
- **Bounds on relative magnitudes:** Require that $|\delta_1| \leq \bar{M}|\delta_{-1}|$
- **Smoothness restriction:** Bound how far δ_1 can deviate from a linear extrapolation of the pre-trend: $\delta_1 \in [-\delta_{-1} - M, -\delta_{-1} + M]$



- Which to choose depends on what types of violations we're worried about
  (e.g., differential shocks vs long-run secular trends)

Examples of Restrictions on δ

- **Bounds on relative magnitudes:** Require that $|\delta_1| \leq \bar{M}|\delta_{-1}|$
- **Smoothness restriction:** Bound how far δ_1 can deviate from a linear extrapolation of the pre-trend: $\delta_1 \in [-\delta_{-1} - M, -\delta_{-1} + M]$



- Which to choose depends on what types of violations we're worried about
  (e.g., differential shocks vs long-run secular trends)

Robust confidence intervals

- In the paper, we develop confidence intervals for the treatment effect of interest under the assumptions on δ discussed above
  - Building on tools in Partial Identification literature Andrews et al., 2023; Armstrong and Kolesar,2018
- The CIs account for the fact that we don't observe the true (population) pre-trend , only our estimate $\hat{\beta}^{pre}$.
- The robust CIs tend to be wider the larger are the confidence intervals on the pre-trends — intuitive, since if we know less about the pre-trends, we should have more uncertainty
- This contrasts with pre-trends tests, where you're less likely to reject the null that $= 0$ when the SEs are larger!

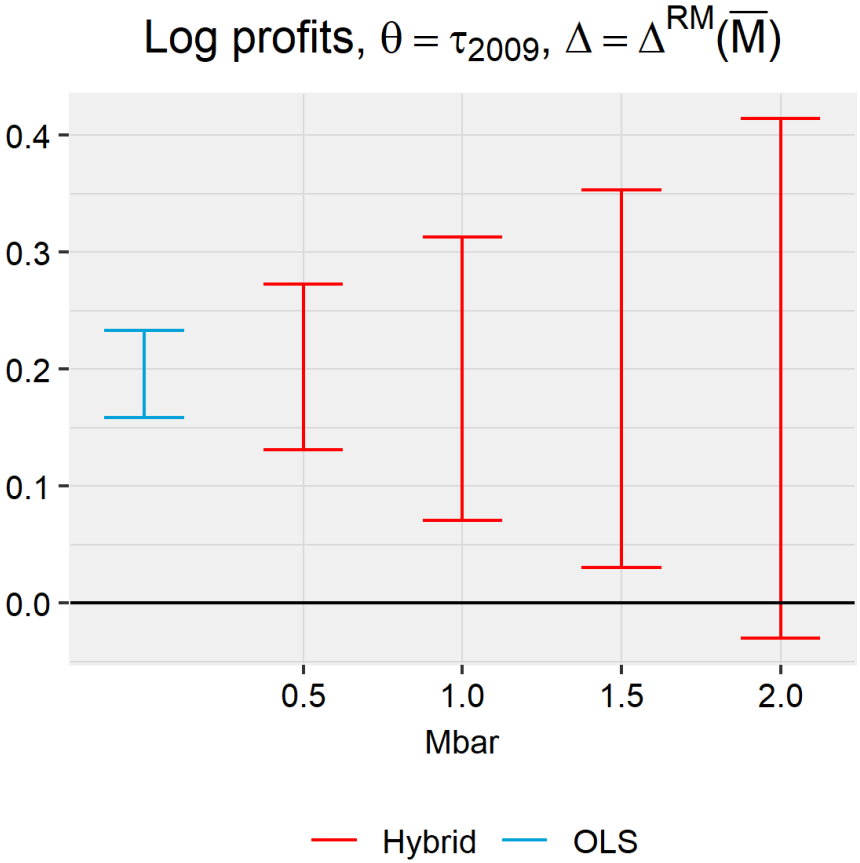
Example: Benzarti & Carloni (2019, AEJ: Economic Policy)

- BC study the incidence of a cut in the value-added tax on sit-down restaurants in France. France reduced the VAT on restaurants from 19.6 to 5.5 percent in July of 2009.
- BC analyze the impact of this change using a difference-in-differences design comparing restaurants to a control group of other market services firms

$$Y_{irt} = \sum_{s=2004}^{2012} \beta_s \times 1[t = s] \times D_{ir} + \phi_i + \lambda_t + \varepsilon_{irt}, \tag{11}$$

  - Y_{irt} = outcome of interest for firm i in region r
  - D_{ir} = indicator if firm i in region r is a restaurant
  - Φ_i, λ_t = firm and year FEs
- Outcomes of interest include firm profits, prices, wage bill & employment.
  – Focus on impact on profits in first year after reform.

Event-study coefficients for log profits



- "Breakdown" $\bar{M}$ for null effect is ~ 2
- Can rule out a null effect unless allow for violations of PT 2x larger than the max in pre-period

More complicated settings

- So far, we have focused on DiD settings with common timing
- But same basic idea works whenever you have "event-study" estimates $(\hat{\beta}^{pre}, \hat{\beta}^{post})$ and are willing to bound the biases of $\hat{\beta}^{post}$ using .
- The theory only relies on asymptotic normality of $\hat{\beta}$ (and consistent estimation of its variance)
- The sensitivity analysis described above can thus be applied to new estimators for staggered treatment timing, IV event-studies, DDD, etc.
  - See the `HonestDiD` package README for examples!

Pros and Cons

Pros

- Very intuitive, easy to visualize.
- Helps identify when pre-testing may be least effective
- Requires minimal changes from standard practice

Cons

- Power will always be < 1, so no guarantee of unbiasedness/correct inference
- Need to specify the hypothesized trend. Will sometimes be difficult to summarize over many of these.
- Still not clear what to do when reject the pre-test.

Summary of Pre-trends Testing

- Tests of pre-trends are intuitive but not a panacea!
- In particular, they may suffer from low power and introduce pre-test bias
- Roth (2022) and Rambachan and Roth (2023) provide tools for diagnostics and sensitivity analysis
- Recent developments:
  - Bayesian version of HonestDiD (Kwon and Roth, 2024)
  - Other bounding exercises (Manski and Pepper, 2018; Ye et al., 2021)
  - Non-inferiority approaches to pre-testing (Bilinski and Hatfield, 2018; Dette and Schumann, 2020)
  - Impose structure on the confounds (Freyaldenhoven et al., 2019)

A Guide for Practitioners

Roth et al.(2023, JOE)

Table 1
A checklist for DiD practitioners.

- Is everyone treated at the same time?

If yes, and panel is balanced, estimation with TWFE specifications such as (5) or (7) yield easily interpretable estimates.

If no, consider using a "heterogeneity-robust" estimator for staggered treatment timing as described in Section 3. The appropriate estimator will depend on whether treatment turns on/off and which parallel trends assumption you're willing to impose. Use TWFE only if you're willing to restrict treatment effect heterogeneity.

- Are you sure about the validity of the parallel trends assumption?

If yes, explain why, including a justification for your choice of functional form. If the justification is (quasi-)random treatment timing, consider using a more efficient estimator as discussed in Section 6.

If no, consider the following steps:

1. If parallel trends would be more plausible conditional on covariates, consider a method that conditions on covariates, as described in Section 4.2.
2. Assess the plausibility of the parallel trends assumption by constructing an event-study plot. If there is a common treatment date and you're using an unconditional parallel trends assumption, plot the coefficients from a specification like (16). If not, then see Section 4.3 for recommendations on event-plot construction.
3. Accompany the event-study plot with diagnostics of the power of the pre-test against relevant alternatives and/or non-inferiority tests, as described in Section 4.4.1.
4. Report formal sensitivity analyses that describe the robustness of the conclusions to potential violations of parallel trends, as described in Section 4.5.

- Do you have a large number of treated and untreated clusters sampled from a super-population?

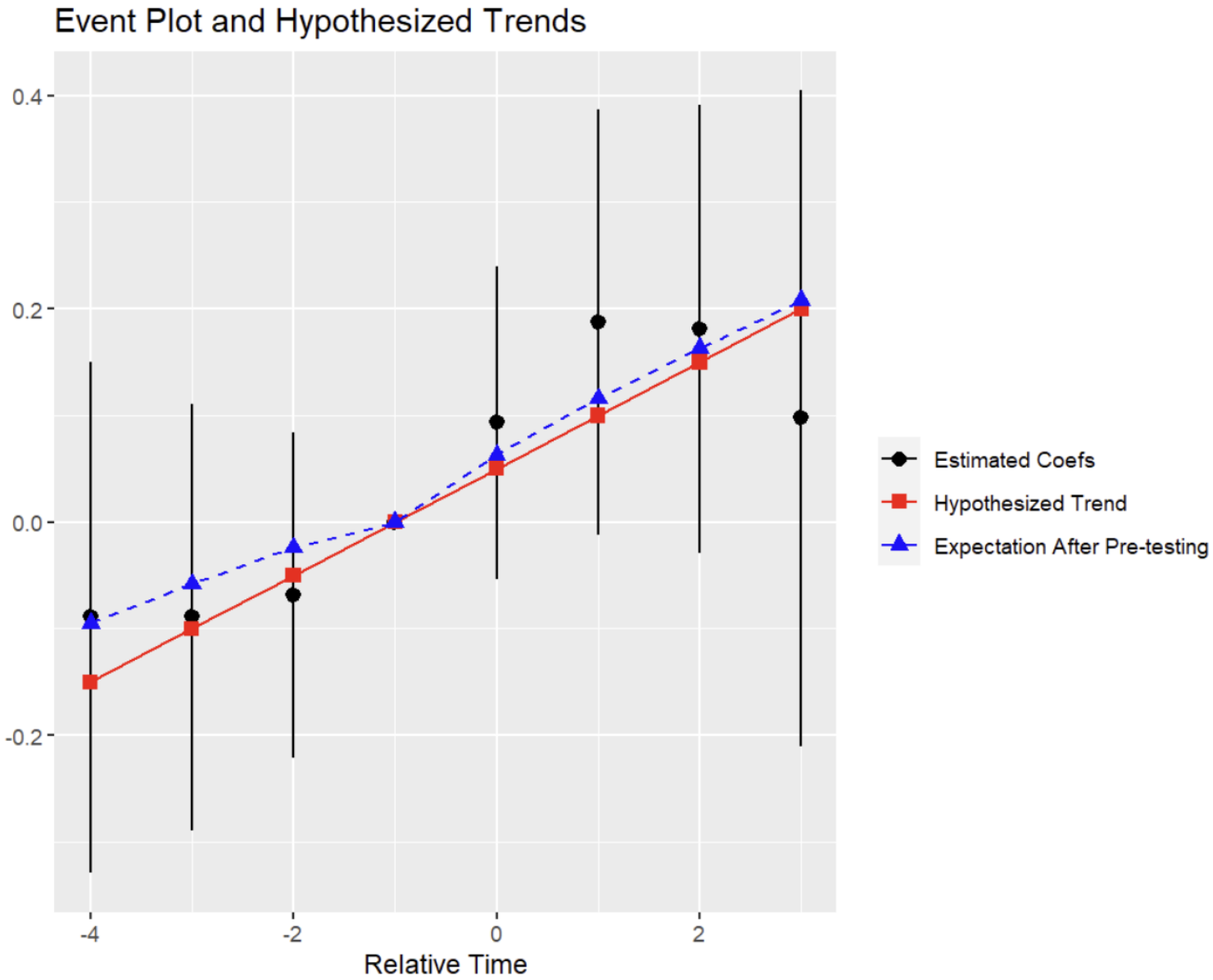
If yes, then use cluster-robust methods at the cluster level. A good rule of thumb is to cluster at the level at which treatment is independently assigned (e.g. at the state level when policy is determined at the state level); see Section 5.2.

If you have a small number of treated clusters, consider using one of the alternative inference methods described in Section 5.1.

If you can't imagine the super-population, consider a design-based justification for inference instead, as discussed in Section 5.2.

Additional Resources

- Roth (2022), "Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends"
  - Paper; `staggered` package ; Shiny app
- Rambachan and Roth (2023), "A More Credible Approach to Parallel Trends"'
  - Paper; `HonestDiD` package ; Vignette


Event Plot and Hypothesized Trends
0.4
0.2
0.0
-0.2
-4
-2
0
2
Relative Time
Estimated Coefs
Hypothesized Trend
Expectation After Pre-testing

Power	Bayes.Factor	Likelihood.Ratio
0.33	0.76	1.23

- **Power.** Chance find significant pre-trend under hypothesized trend.
- **Bayes Factor.** Relative chance you pass the pre-test under hypothesized trend versus under parallel trends.
- **Likelihood Ratio.** Likelihood of observed pre-trend coefs under hypothesized trend versus under parallel trends.

Ye, Keele, Hasegawa, and Small Back

- Consider DiD settings with a **treated group** ($G = trt$) and **two imperfect control groups** $G = a, b$.
- Key assumption: counterfactual trends for the treated group are **bracketed** by counterfactual trends for the control group:

$$\Delta_{trt} \in [\min\{\Delta_a, \Delta_b\}, \max\{\Delta_a, \Delta_b\}]$$

  where $\Delta_g = E[\Delta Y(0) \mid G = g]$ is the average trend in $Y(0)$ for group g.
- Motivating example — expansion of Fair Labor Standards Act (FLSA):
  - Treated group is gov't workers, whose employment is expected to be weakly procyclical
  - Set control groups to be industries whose employment is known to be strongly procyclical (e.g. construction, retail) or countercylical (e.g. agriculture, medical)

Ye, Keele, Hasegawa, and Small Back

- Let DiD_a, DiD_b denote the DiD estimand using groups a, b as the control group
- Then with two periods, it is straighforward to show that

$$\min\{DiD_a, DiD_b\} \leq ATT \leq \max\{DiD_a, DiD_b\}$$

- The paper extends this approach to settings with more than 2 periods, and provides methods for doing inference