Advanced DID

Violations of Parallel Trends

MIXTAPE SESSIONS

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 parallel trends

- Three substrands of this literature:
  - → **Parallel trends only conditional on covariates**
  - → **Testing for violations of (conditional) parallel trends**
  - → **Sensitivity analysis and bounding exercises**
- I 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
  - $\rightarrow$ E.g. states that pass a minimum wage increase might also change unemployment insurance at the same time
  - $\rightarrow$ 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:
  - $\rightarrow$ In period 0, all control units have outcome 10; all treated units have outcome 5.
  - $\rightarrow$ In period 1, all control units have outcome 15.
  - $\rightarrow$ If treatment hadn’t occurred, would treated units’ outcome have increased by 5 also (PT in levels)?
  - $\rightarrow$ Or would they have increased by 50% ($\sim$ PT in logs)?

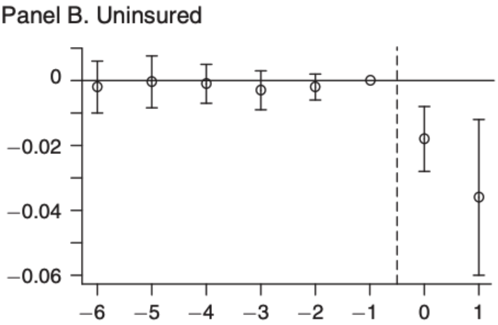
Roth and Sant'Anna (2023) show that PT will depend on functional form unless:

- **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:

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

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 (Freyaldenhoven et al., 2019; Kahn-Lang and Lang, 2020; Bilinski and Hatfield, 2018; Roth, 2022)
- Remainder of the talk today will focus on these issues, as well as some solutions.
- Perhaps selfishly, will focus mainly on two of my papers
  - → 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. min wage and UI 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. min wage and UI 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

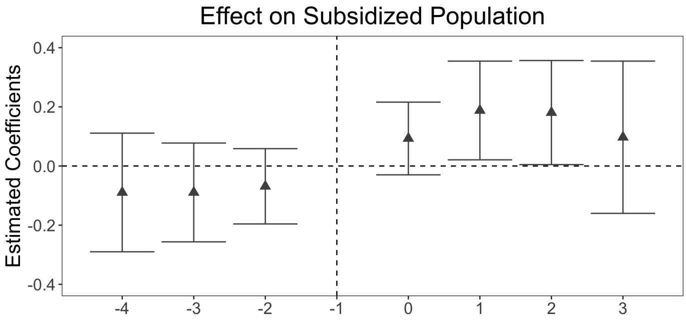
Overview of Limitations

- 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. min wage and UI 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. min wage and UI 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")

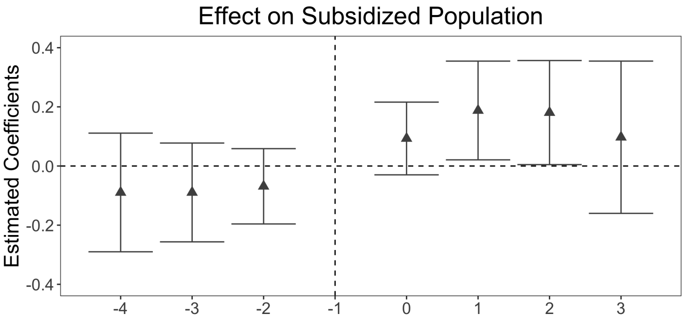
Issue 1 - Low Power



- He & Wang (2017) study impacts of placing college grads as village officials in China
- Use an “event-study” approach comparing treated and untreated villages

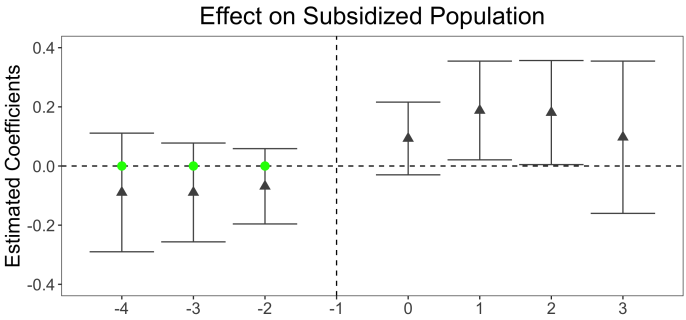
$$Y_{it} = \sum_{k \neq -1} D_{it}^{k} \beta_k + \alpha_i + \phi_t + \epsilon_{it}$$

Issue 1 - Low Power



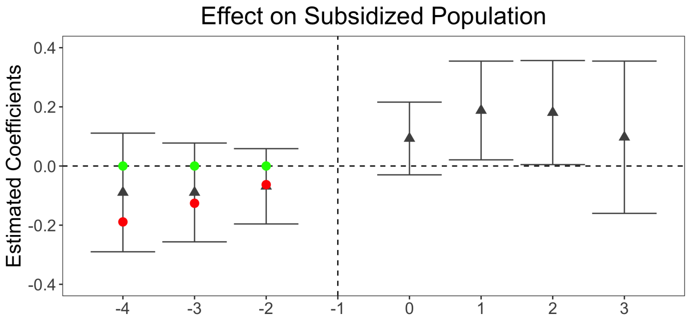
"The estimated coefficients on the leads of treatment ... are statistically indifferent from 0. ... We conclude that the pretreatment trends in the outcomes in both groups of villages are similar, and villages without CGVOs can serve as a suitable control group for villages with CGVOs in the treatment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

Issue 1 - Low Power



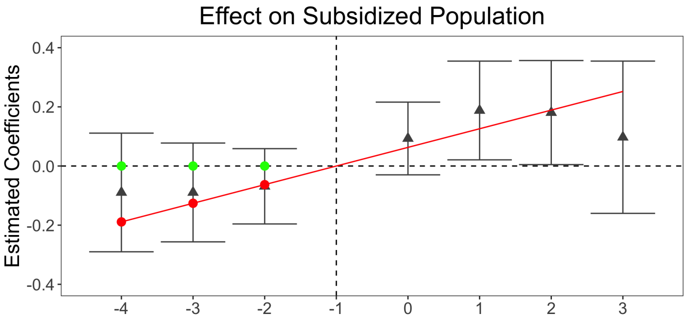
- P-value for $H_0 : \beta_{pre} =$ green dots (no pre-trend): 0.81

Issue 1 - Low Power



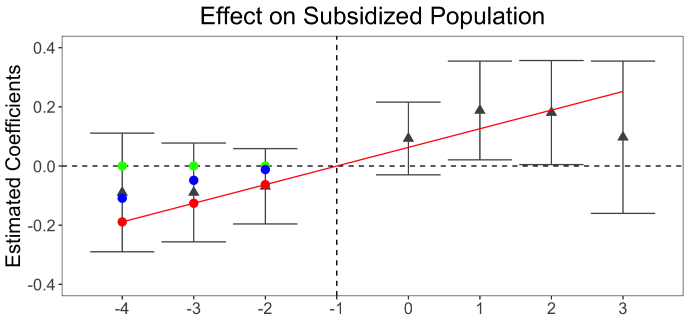
- P-value for $H_0 : \beta_{pre} =$ green dots (no pre-trend): 0.81
- P-value for $H_0 : \beta_{pre} =$ red dots: 0.81

Issue 1 - Low Power



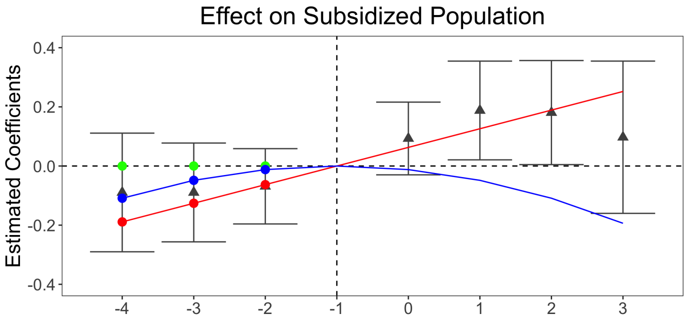
- P-value for $H_0 : \beta_{pre} =$ green dots (no pre-trend): 0.81
- P-value for $H_0 : \beta_{pre} =$ red dots: 0.81

Issue 1 - Low Power



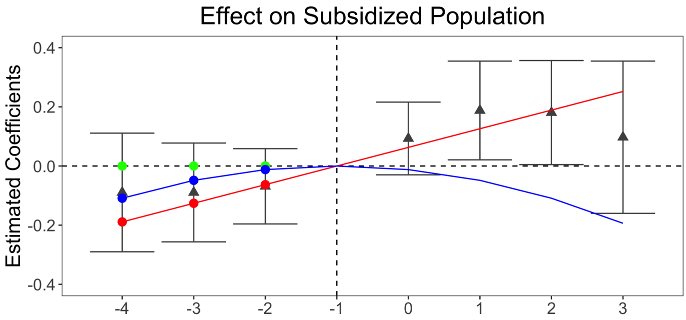
- P-value for $H_0 : \beta_{pre} =$ green dots (no pre-trend): 0.81
- P-value for $H_0 : \beta_{pre} =$ red dots: 0.81
- P-value for $H_0 : \beta_{pre} =$ blue dots: 0.81

Issue 1 - Low Power



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

Issue 1 - Low Power



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

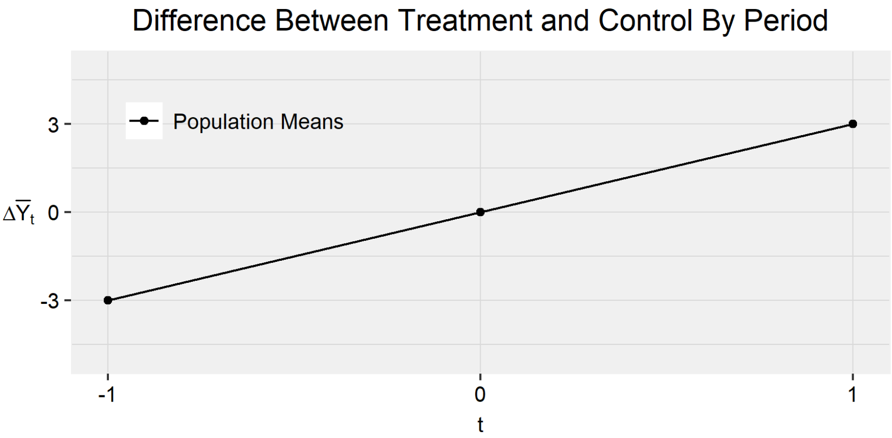
- Roth (2022): simulations calibrated to papers published in *AER*, *AEJ: Applied*, and *AEJ: Policy* between 2014 and mid-2018
  - → 70 total papers contain an event-study plot; focus on 12 w/available data
- Evaluate properties of standard estimates/CIs under linear violations of parallel trends against 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

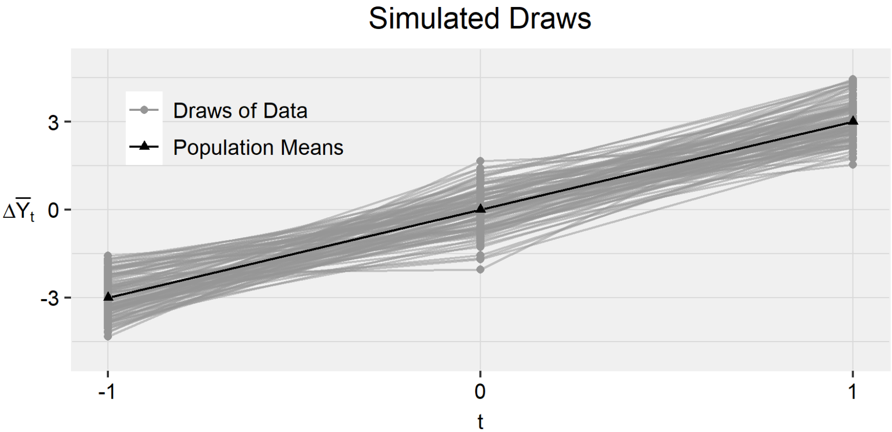
- When parallel trends is violated, we will sometimes fail to find a significant pre-trend
- But the draws of data where this happens are a **selected sample**. This is known as *pre-test bias*.
- Analyzing this selected sample introduces additional statistical issues, and can make things worse!

Stylized Three-Period DiD Example

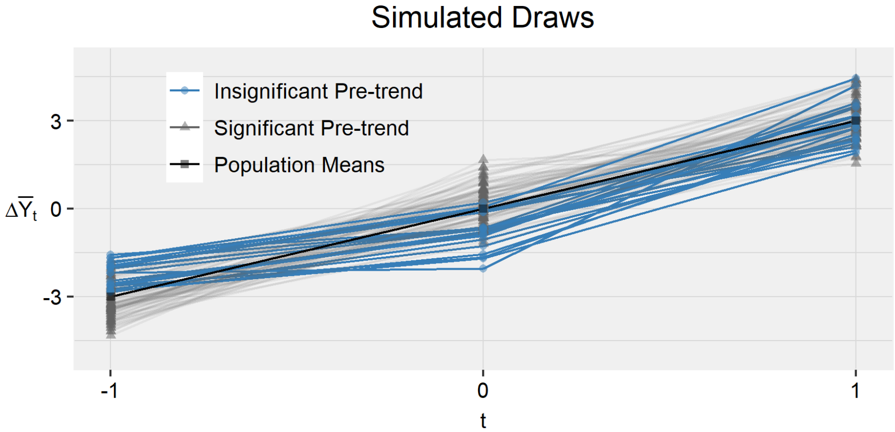
- Consider a 3-period model ($t = -1, 0, 1$) where treatment occurs in last period
- No causal effect of treatment: $Y_{it}(0) = Y_{it}(1)$ in all periods
- In population, treatment group is on a linear trend relative to the control group with slope δ
  - → Control group mean in period t: $E[Y_{it}(0) \mid \text{Control group}] = 0$
  - → 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



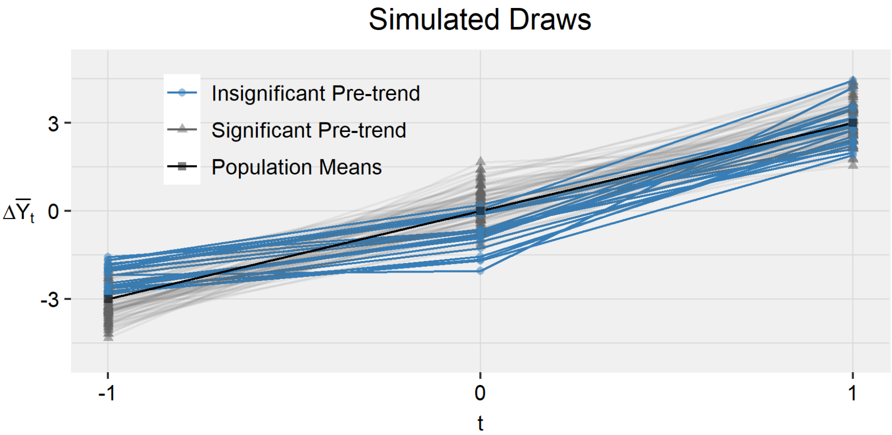
- Example: In population, there is a linear difference in trend with slope 3



- 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

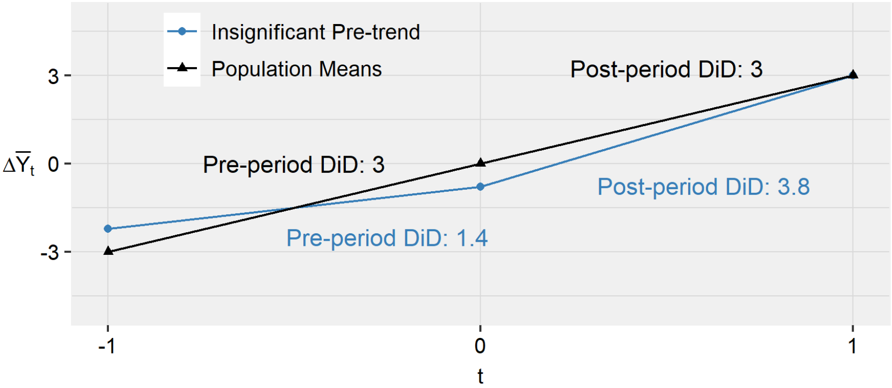


- Example: In population, there is a linear difference in trend with slope 3
- In some of the draws of the data, highlighted in blue, the difference between period -1 and 0 will be insignificant



- 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 $t = 0$

Average Over 1 Million Draws



- 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:

$$\delta_1 = \mathbb{E}\left[Y_{i1}(0) - Y_{i0}(0) \mid D_i = 1\right] - \mathbb{E}\left[Y_{i1}(0) - Y_{i0}(0) \mid D_i = 0\right]$$

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

$$\delta_{-1} = \mathbb{E}\left[Y_{i,-1}(0) - Y_{i0}(0) \mid D_i = 1\right] - \mathbb{E}\left[Y_{i,-1}(0) - Y_{i0}(0) \mid D_i = 0\right]$$

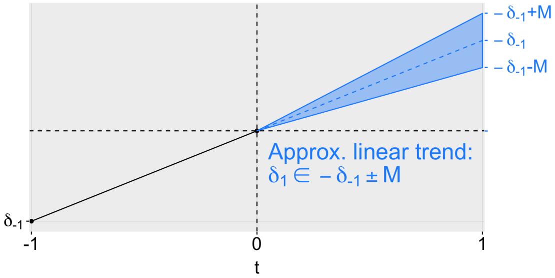
- Key idea: restrict possible values of δ_1 given δ_{-1}
  Intuitively, counterfactual trend can't be too different from pre-trend

Examples of Restrictions on δ

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

Examples of Restrictions on δ

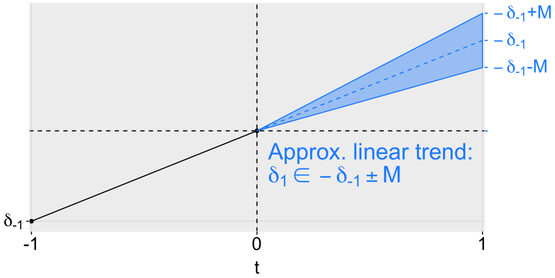
- **Bounds on relative magnitudes:** Require that $|\delta_1| \leq \bar{M}|\delta_{-1}|$
- **Smoothness restriction:** Bound how far δ_1 can deviate from a linear extrapolation of the pre-trend: $\delta_1 \in [-\delta_{-1} - M, -\delta_1 + M]$



- 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
  - $\rightarrow$ Building on tools in partial ID literature (Andrews et al., 2023; Armstrong and Kolesár, 2018)
- The CIs account for the fact that we don't observe the true (population) pre-trend δ_{pre}, only our estimate $\widehat{\beta}_{pre}$.
- 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 $\beta_{pre} = 0$ when the SEs are larger!

Benzarti & Carloni (2019)

- 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.

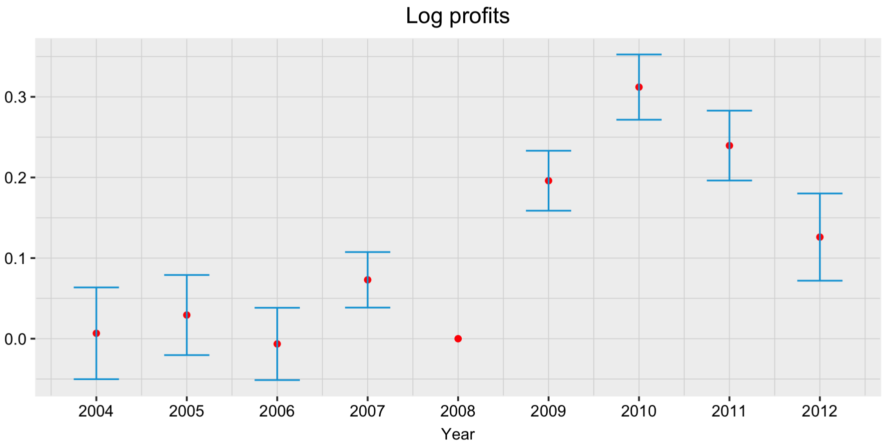
- BC analyze the impact of this change using a difference-in-differences design comparing restaurants to a control group of other market services firms

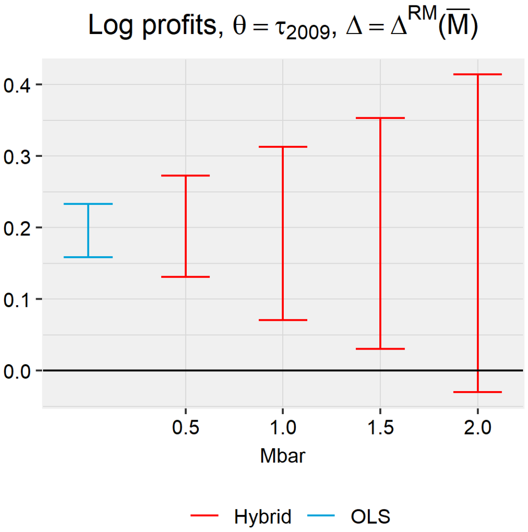
$$Y_{irt} = \sum_{s=2004}^{2012} \beta_s \times 1[t = s] \times D_{ir} + \phi_i + \lambda_t + \epsilon_{irt}, \tag{1}$$

  → 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. We 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 $(\widehat{\beta}_{pre}, \widehat{\beta}_{post})$ and are willing to bound the biases of $\widehat{\beta}_{post}$ using β_{pre}.
- The theory only relies on asymptotic normality of $\widehat{\beta}$ (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.
  - $\rightarrow$ See the HonestDiD package README for examples!

So to summarize!

- 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
- And these tools play nicely with recent estimators developed for heterogeneous treatment effects; see the HonestDiD package README for examples!

Other Related Papers

- Bayesian version of HonestDiD (Kwon and Roth, 2024)
- Other bounding exercises (Manski and Pepper, 2018; Ye et al., 2021) Ye et al
- Non-inferiority approaches to pre-testing (Bilinski and Hatfield, 2018; Dette and Schumann, 2020)
- Impose structure on the confounds (Freyaldenhoven et al., 2019)

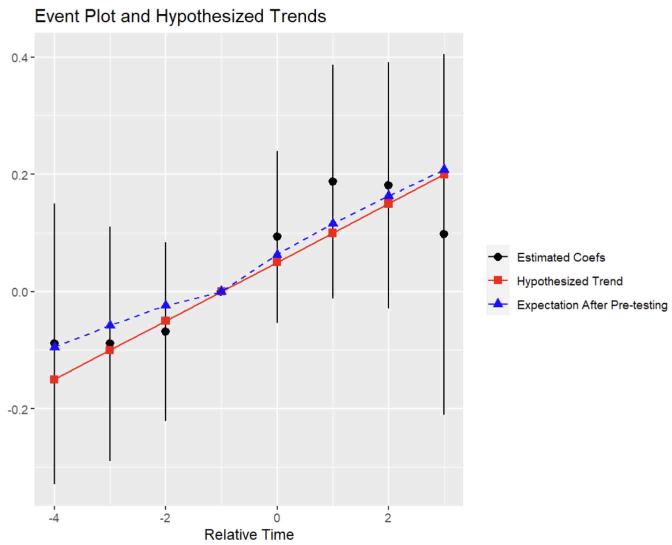
Thank you!

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

Pre-testing Diagnostics

- A "low-touch" intervention is to evaluate the likely power/distortions from pre-testing under *context-relevant* violations of parallel trends
- Enter the `pretrends` package / Shiny app



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.

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.

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 straightforward 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

References I

Andrews, Isaiah, Jonathan Roth, and Ariel Pakes, “Inference for Linear Conditional Moment Inequalities,” *The Review of Economic Studies*, January 2023, p. rdad004.

Armstrong, Timothy B. and Michal Kolesár, “Optimal Inference in a Class of Regression Models,” *Econometrica*, 2018, *86* (2), 655–683.

Bilinski, Alyssa and Laura A. Hatfield, “Seeking evidence of absence: Reconsidering tests of model assumptions,” *arXiv:1805.03273 [stat]*, May 2018.

Dette, Holger and Martin Schumann, “Difference-in-Differences Estimation Under Non-Parallel Trends,” *Working Paper*, 2020.

Freyaldenhoven, Simon, Christian Hansen, and Jesse M. Shapiro, “Pre-event Trends in the Panel Event-Study Design,” *American Economic Review*, 2019, *109* (9), 3307–3338.

References II

Kahn-Lang, Ariella and Kevin Lang, "The Promise and Pitfalls of Differences-in-Differences: Reflections on 16 and Pregnant and Other Applications," *Journal of Business & Economic Statistics*, 2020, *38* (3), 613–620.

Kwon, Soonwoo and Jonathan Roth, "(Empirical) Bayes Approaches to Parallel Trends," *AEA Papers and Proceedings*, May 2024, *114*, 606–609.

Manski, Charles F. and John V. Pepper, "How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions," *The Review of Economics and Statistics*, 2018, *100* (2), 232–244.

Rambachan, Ashesh and Jonathan Roth, "A More Credible Approach to Parallel Trends," *The Review of Economic Studies*, October 2023, *90* (5), 2555–2591.

References III

Roth, Jonathan, "Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends," *American Economic Review: Insights*, 2022, *4* (3), 305–322.

__ **and Pedro H. C. Sant'Anna**, "When Is Parallel Trends Sensitive to Functional Form?," *Econometrica*, 2023.

Ye, Ting, Luke Keele, Raiden Hasegawa, and Dylan S. Small, "A Negative Correlation Strategy for Bracketing in Difference-in-Differences," *arXiv:2006.02423 [econ, stat]*, 2021.