Recent advances in Difference-in-Differences

Alejandro Molnar
The World Bank

January 11, 2026

Introduction: updates to methods for estimating DiD

- Difference-in-Differences (DiD) is a workhorse “empirical strategy” for causal inference in applied economics.
- **Basic setup:** Two groups (Treated vs. Control), and two periods (Pre vs. Post).
- Empirical settings can be more complex:
  - *More periods*.
  - *Staggered adoption:* groups of units get treated at different times.
  - *Heterogeneous effects:* treatment effects vary over time and across units.
  - *And more,* e.g., multiple treatment levels.
- A **recent literature** (2018+) has shown that the standard Two-Way Fixed Effects (TWFE) estimator may be biased in these settings.
- **Today:** Describe the issue, diagnose, and overview of solution approaches.

Recap. The canonical 2x2 Difference-in-Differences (DiD) setup

Approach 1. Sample analogue to the DiD of population means
Consider two groups ($g \in \{T, C\}$) and two periods ($t \in \{0, 1\}$). Let $\bar{y}_{g,t}$ denote the sample mean of the outcome for group g in period t.

	Pre-Period ($t = 0$)	**Post-Period** ($t = 1$)	**Time Difference**
Treated (T)	$\bar{y}_{T,0}$	$\bar{y}_{T,1}$	$\Delta\bar{y}_T = \bar{y}_{T,1} - \bar{y}_{T,0}$
Control (C)	$\bar{y}_{C,0}$	$\bar{y}_{C,1}$	$\Delta\bar{y}_C = \bar{y}_{C,1} - \bar{y}_{C,0}$
Diff-in-Diff			$\hat{\beta}^{DiD} = \Delta\bar{y}_T - \Delta\bar{y}_C$

Approach 2. Two-Way Fixed Effects (TWFE) OLS Formulation
The estimator $\hat{\beta}^{DiD}$ is numerically equivalent to $\hat{\beta}$ in the regression:

$$y_{it} = \alpha + \gamma \text{Treat}_i + \lambda \text{Post}_t + \beta(\text{Treat}_i \times \text{Post}_t) + \epsilon_{it}$$

Recap. Numerical example: Cordon pricing in NYC

Scenario: NYC introduces congestion pricing in the CBD (Treated) but it is not introduced "somewhere else" (Control). **Outcome (y):** Average traffic speed (mph).

	Pre (2020)	**Post (2022)**	**Δ Time**
NYC CBD (Treated)	6.0	8.0	$+2.0$
"Elsewhere" (Control)	20.0	18.0	-2.0
Difference			$\hat{\beta}^{DiD} = +$**4.0**

Calculation of the treatment effect:

- **Treated path:** Traffic speed increases by 2 mph ($\Delta \bar{y}_T = +2$).
- **Counterfactual trend:** We assume that absent the policy, CBD would have followed the trend "elsewhere" of decreased by 2 mph ($\Delta \bar{y}_C = -2$).
- **Result:**

$$\hat{\beta}^{DiD} = (2) - (-2) = 4 \text{ mph}$$

The Standard Two-Way Fixed Effects (TWFE) Model

Consider a panel dataset with units i and time periods t. The standard regression is:

$$y_{it} = \alpha_i + \lambda_t + \beta^{TWFE} D_{it} + \varepsilon_{it}$$

- α_i: Unit fixed effects (controls for time-invariant unobservables).
- λ_t: Time fixed effects (controls for common shocks).
- D_{it}: Binary treatment indicator ($= 1$ if treated, 0 otherwise).

Assumptions

- Parallel trends.
- No-anticipation effects

β^{TWFE} is interpreted as the Average Treatment Effect on the Treated (ATT). Note: this still allows that there may be *selection* on who got the treatment (e.g., units where it is more likely to have a positive effect).

Issue: TWFE identifies the ATT only under strictly *homogeneous* treatment effects.

Motivating example: staggered subway expansions

Imagine we are estimating the effect of **subway station openings** on **air pollution** (Gendron-Carrier et al., 2022).

- **City A (Early Adopter):** Opens a new line in 2006.
- **City B (Late Adopter):** Opens a new line in 2016.
- **City C (Never Treated):** No new lines.

Heterogeneity is likely:

1. *Dynamic effects:* Air pollution in City A might decline in 2006, but keep declining in 2007, 2008, ..., as mode switching occurs.
2. *Cohort effects:* City A might be a denser, wealthier city than City B, leading to a different magnitude of effect.

Issues with TWFE

β^{TWFE} is a weighted average of all 2×2 DiDs:

$$\beta^{TWFE} = \sum_k w_k \hat{\beta}_k^{2\times2}$$

- Comparisons between **Treated vs. Never Treated** (OK).
- Comparisons between **Early Treated vs. Late Treated** (OK).
- Comparisons between **Late Treated vs. Early Treated** (may not be valid).

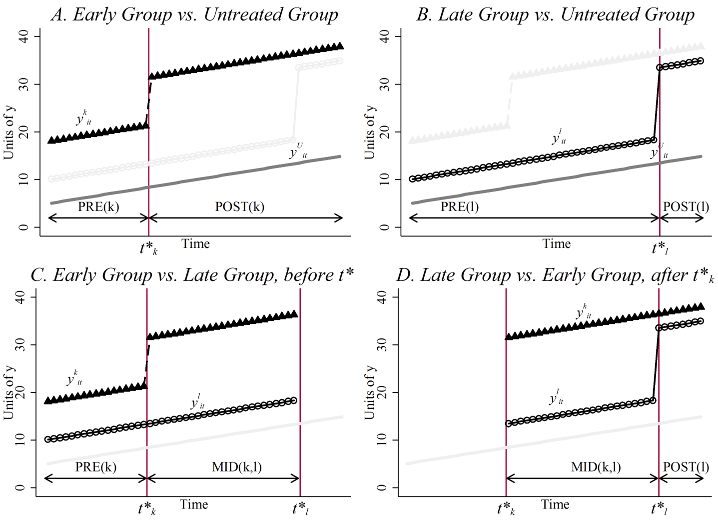
Goodman-Bacon (2021): i) Possibly invalid comparisons are used. ii) Comparisons are weighted by sample size and variance in treatment timing.

de Chaisemartin & D'Haultfoeuille (2020): Some weights $W_{g,t}$ can be **negative**.

- Even with a positive treatment effect everywhere, but some weights are negative, the estimated β could be **negative**.
- Happens primarily when already-treated units are used as controls for newly-treated units.

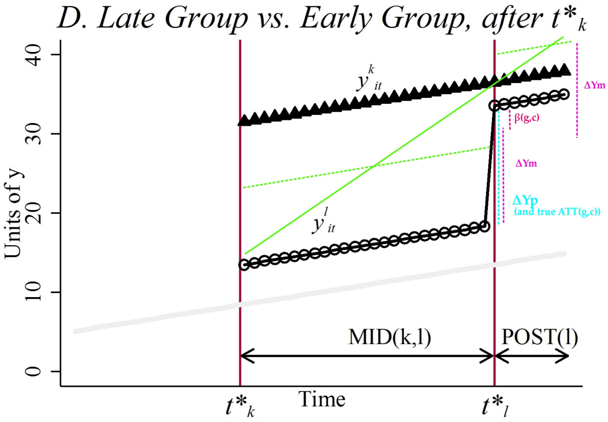
"Forbidden comparisons" issue (Goodman-Bacon, 2021)

In a staggered design, TWFE computes the average of *all* possible 2×2 DiD comparisons.



“Forbidden comparisons” issue (Goodman-Bacon, 2021)

Heterogeneity: If treatment effects grow over time (is dynamic), the already-treated unit is on a different trend path than a never-treated counterfactual - should not serve as counterfactual.



Issue: The estimated effect for the Late Adopter is calculated relative to the steep trend of the Early Adopter, resulting in attenuation bias (or sign flips).

The dynamic "event study" design (TWFE)

1. Regression specification

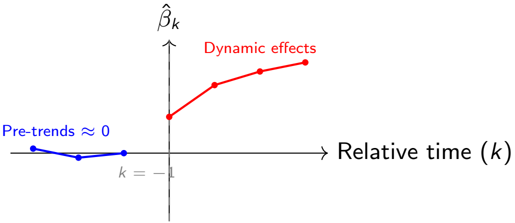
To allow for dynamic effects and "assess" parallel trends, with multiple periods it is standard to decompose the treatment indicator into *relative* time dummies (leads and lags):

$$y_{it} = \underbrace{\mu_i + \lambda_t}_{\text{Fixed Effects}} + \sum_{k=-M, k\neq -1}^{L} \beta_k \cdot \mathbb{1}(t - G_i = k) + \epsilon_{it}$$

- $k = t - G_i$: Time relative to treatment start year G_i.
- $\mathbb{1}(\dots)$: Dummies for being k periods away from treatment.
- $\beta_{-1} = 0$: A normalization (baseline period).

2. Interpretation of coefficients (β_k)

- **Leads ($k < 0$):** Assess *parallel trends*. Ideally $\beta_{pre} \approx 0$ (no anticipation).
- **Lags ($k \geq 0$):** Estimate the dynamic path of the treatment effect.



Event study designs

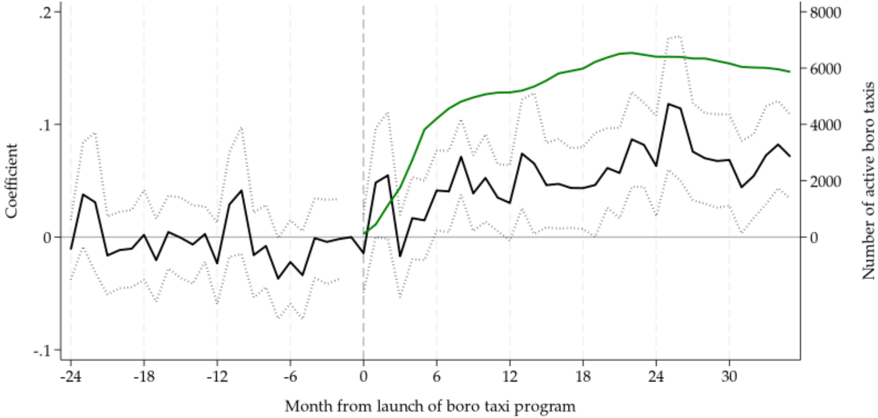


Figure: Effect of NYC boro taxi program on travel times, Mangrum & Molnar (2020)

Economic intuition I: Dynamic effects

Why might effects change over time?
Example: transit infrastructure investments in a sample of cities.

- **Year 0 (Opening):** Possible disruption costs, existing car stock remains in use.
- **Year 3:** Habits change, cars sold or not renewed, transit use increases.
- **Year 5:** Transit revenue funds improved frequency and land use, more transit use.

If we compare a city getting transit in Year 5 to a city that got it in Year 0, we are subtracting the larger "Year 5 growth" of the early city from the "Year 0 start" of the new city. This fits the previous pattern of mechanically biasing down the estimate.

Economic intuition II: cohort selection

Early adopters of the treatment might differ from late adopters.

Example: Uber entry.

- **Cohort 1 (2011):** San Francisco, NYC. (Population density, higher income, existing taxi substitutes).
- **Cohort 5 (2015):** Small-town rural markets. (suburban land use patterns, no taxis).

We cannot assume the treatment effect for SF (Cohort 1) is the same as the effect for a small town (Cohort 5). Standard TWFE averages these into a single β, which is hard to interpret.

Recent methods estimate group-time specific effects ($ATT(g, t)$).

Economic intuition III: anticipation effects

The Issue: Standard DiD assumes the treatment effect is zero prior to the implementation date ($D_{it} = 1$).

Example: infrastructure announcements. Major transport projects have large time-to-build.

- *Example:* A new metro line is announced in 2018, construction begins 2019, opens 2022.
- Investors buy land near future stations in 2018.
- If we set $t = 2022$ as the "treatment date" and use 2018-2021 as the "pre-period," we are comparing the *post-treatment* period against a *partially treated* period.

Consequence: This attenuates the estimated effect (bias toward zero) because the "baseline" was already elevated.

Economic intuition IV: spillovers (SUTVA violations)

Issue: DiD requires the *Stable Unit Treatment Value Assumption* (SUTVA): The treatment of unit i should not affect the outcome of unit j.

Important in a transportation context. Transport networks are inherently connected, resulting in spillover & network effects. Treating one link often affects the "control" links. (Borusyak & Hull, 2023)

Example: congestion pricing.

- **Treated:** Zone A introduces a toll. Traffic drops (outcome improves).
- **Control:** Zone B (adjacent, no toll). Drivers divert from A to B. Traffic worsens.

$$\hat{\beta}_{DiD} = (\Delta Y_A) - (\Delta Y_B)$$

If, say, $\Delta Y_A = -10$ (True Effect) and $\Delta Y_B = +5$ (Spillover damage), the estimated effect is -15. We may overestimate the benefit because the control group was contaminated.

Current estimation approaches for DiD

The consensus is to move away from static TWFE toward methods that strictly control which units are compared.

1. **Callaway & Sant'Anna (2021):** Estimate group-time ATTs using clean controls (never treated or not-yet treated).
2. **de Chaisemartin & D'Haultfoeuille (2020):** Estimators robust to negative weights.
3. **Sun & Abraham (2021):** Interaction-weighted estimator. Fixes the "contamination" between cohorts and trends in dynamic TWFE specifications.
4. **Stacked Regression (Cengiz et al., 2019):** Stack copies of each event.

Method: Callaway & Sant'Anna (2021)

Key idea: Calculate $ATT(g,t)$ for each group g and time t.

$$ATT(g,t) = E[Y_t - Y_{g-1}|G = g] - E[Y_t - Y_{g-1}|G = Never/NotYet]$$

- **Step 1:** Compare treated group g against a "clean" control group (Never Treated or Not-Yet Treated).
- **Step 2:** Aggregate these parameters into summaries (e.g., "Event Study", "Average Effect").

- **Benefit:** Explicit weighting of individual DiD terms, allowing researcher to take a stand on how to average group effects.
- **Also:** links to literature on propensity score matching (modeling selection into treatment; doubly-robust methods).

Method: Sun & Abraham (2021)

Focuses and fixes issues in the standard "Event Study" regression:

$$y_{it} = \alpha_i + \lambda_t + \sum_k \mu_k D_{it}^k + \varepsilon_{it}$$

- **Problem:** In standard regressions, the coefficient for "lead 1" (μ_{-1}) is contaminated by effects from other periods/groups.
- **Solution:** Interaction Weighted Estimation.
- Estimate a fully interacted model (interact time dummies with cohort dummies).
- Aggregate these estimates using sample-share weights.

Alternative: stacked regression

A simpler, intuitive fix sometimes used in applied microeconomics:

Dataset 1	Dataset 2	
Cohort 2012 vs Control	Cohort 2014 vs Control	$\longrightarrow$ Stack & Run FE

- Create a "mini-experiment" for each treatment wave.
- Include unit $\times$ dataset and time $\times$ dataset fixed effects.
- Prevents the "forbidden comparisons."
- Con: data intensive.

Recommendations for empirical practice

More concern about TWFE estimates if:

1. **Staggered treatment:** The policy rolls out over time (common in transport/infrastructure).
2. **Heterogeneity:** You suspect the effect changes over time (dynamic) or depends on when the treatment started (cohort).
3. **Balance:** The "Early Treated" group is large relative to the "Never Treated" group.

Recommendations for empirical practice. Workflow

1. **Diagnostic:** Run the Goodman-Bacon decomposition (Stata: `bacondecomp`, R: `bacondecomp`).
   - Check how much weight is placed on "Late vs. Early" comparisons.
2. **Estimation:** Do not rely solely on static TWFE.
3. **Robustness:** Report results using CS (Callaway & Sant'Anna), SA (Sun & Abraham), and/or dCdH (de Chaisemartin & D'Haultfoeuille).
4. **Plotting:** Plot the event-study coefficients ($ATT(t)$) rather than a single summary number. Visual inspection of pre-trends.

A note on Pre-Trends

The "pre-test" problem (Roth, 2022)

- An event study that "passes" the parallel pre-trends "assessment" (coeffs are zero before $t = 0$) is not guaranteed to be unbiased.
- Low statistical power might hide non-parallel trends.
- And recall that the actual parallel trend is unobservable.

Advice:

- Domain knowledge and discussion of economic issues above is required to argue for parallel trends (Why did the subway open in 2012 and not 2013? Was it random construction delays or strategic economic targeting?).
- "Honest DiD" (Rambachan & Roth, 2023) sensitivity analysis. How large would a violation of parallel trends need to be to "kill" the result?

Conclusion

- Approach to estimating DiD has changed, TWFE usually not sufficient.
- Heterogeneity in treatment effects (over time and cohorts) breaks the standard TWFE estimator in staggered designs.
- In transportation economics (infrastructure, policy rollouts), these issues are common.

References

- **Borusyak, K., Jaravel, X. and Spiess, J.** Revisiting Event-Study Designs: Robust and Efficient Estimation. *The Review of Economic Studies*
- **Borusyak, K. and Hull, P. (2023).** Nonrandom Exposure to Exogenous Shocks. *Econometrica*
- **Callaway, B., & Sant'Anna, P. H. (2021).** Difference-in-differences with multiple time periods. *Journal of Econometrics*.
- **de Chaisemartin, C., & D'Haultfoeuille, X. (2020).** Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*.
- **Goodman-Bacon, A. (2021).** Difference-in-differences with variation in treatment timing. *Journal of Econometrics*.
- **Sun, L., & Abraham, S. (2021).** Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*.

Further reading:

- **Cunningham, Scott (2025).** Causal Inference: The Mixtape *Free online textbook, other resources*.
- **de Chaisemartin, C., & D'Haultfoeuille, X. (2025).** Credible Answers to Hard Questions: Differences-in-Differences for Natural Experiments. *Texbook* (PDF, still freely available).
- **Roth, J., et al. (2023).** What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature. *Journal of Econometrics*.