

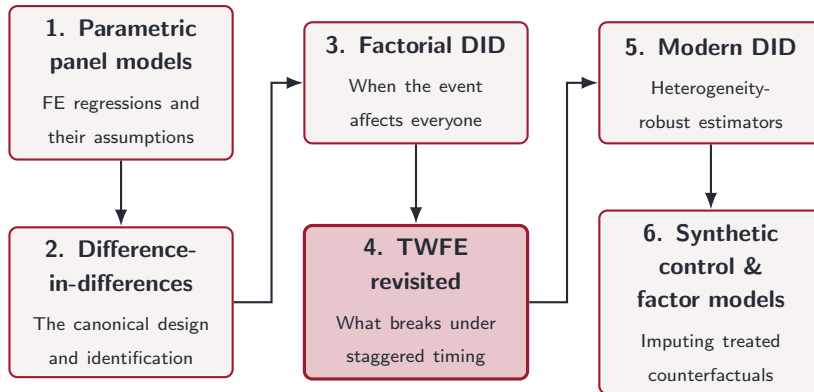
Panel Methods (4): Two-Way Fixed Effects Revisited

Lectures on Causal Panel Analysis

Yiqing Xu
Stanford University

September 2026

Feel free to use or adapt the slides with attribution (CC BY 4.0).



- Lecture 1: parametric panel models, with FE regressions as the workhorse
- Lecture 2: DID as a research design, mostly with two groups and fixed treatment timing
- Lecture 3: factorial DID, extending the design to events that hit every unit
- In the two-period case, the TWFE estimate *is* the DID estimate, so the two traditions seem interchangeable
- Today: what happens when we run TWFE regressions on **many periods**, **staggered timing**, and **heterogeneous effects**, which is what applied work actually looks like

- What assumptions do TWFE models rely upon?
- Three problems
 - Implications for treatment assignment (design)
 - Parallel trends violations
 - The weighting problem

Question

When we estimate $Y_{it} = \tau D_{it} + \alpha_i + \xi_t + \varepsilon_{it}$ on a staggered panel, which comparisons is $\hat{\tau}$ actually making?

- Two groups ($\mathcal{T}, \mathcal{C}$), two periods (t and t'), fixed treatment timing
- Functional form:

$$Y_{it} = \tau_{it}D_{it} + \alpha_i + \xi_t + \varepsilon_{it} \quad \text{or} \quad \begin{cases} Y_{it}(0) = \alpha_i + \xi_t + \varepsilon_{it} \\ Y_{it}(1) = Y_{it}(0) + \tau_{it} \end{cases}$$

where τ_{it} is the treatment effect for unit i at time t

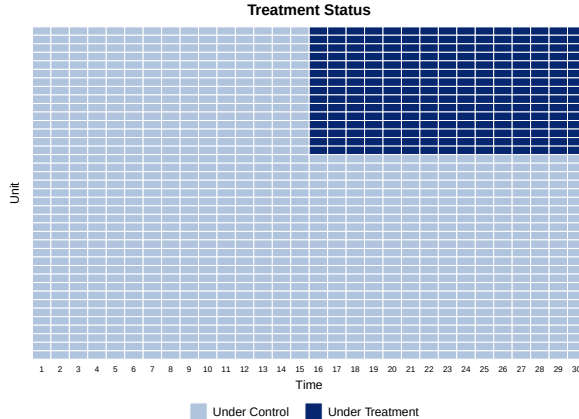
- Parallel trends: $\mathbb{E}[Y_{it'}(0) - Y_{it}(0) \mid i \in \mathcal{T}] = \mathbb{E}[Y_{jt'}(0) - Y_{jt}(0) \mid j \in \mathcal{C}]$
- Or equivalently, $\mathbb{E}[\varepsilon_{it'} - \varepsilon_{it} \mid i \in \mathcal{T}] = \mathbb{E}[\varepsilon_{jt'} - \varepsilon_{jt} \mid j \in \mathcal{C}]$
- $ATT = \mathbb{E}[\tau_{it} \mid D_{it} = 1]$ is nonparametrically identified with two periods (or two treatment histories)

- The treated-group untreated potential outcome after treatment is the one cell we never observe:

	Pre	Post
Treated ($\mathcal{T}$)	$Y_{\mathcal{T}, pre}^0$	??
Comparison ($\mathcal{C}$)	$Y_{\mathcal{C}, pre}^0$	$Y_{\mathcal{C}, post}^0$

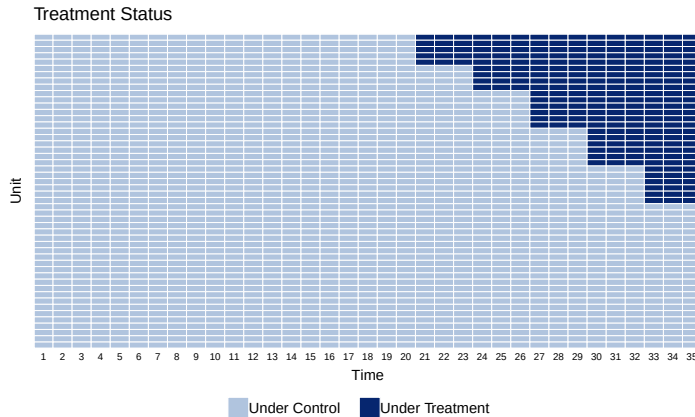
- Parallel trends fills in the missing cell: take the treated pre-period level and add the comparison group's change
- Every estimator in these lectures is a different rule for filling in that cell

Beyond Two Periods: Block Assignment



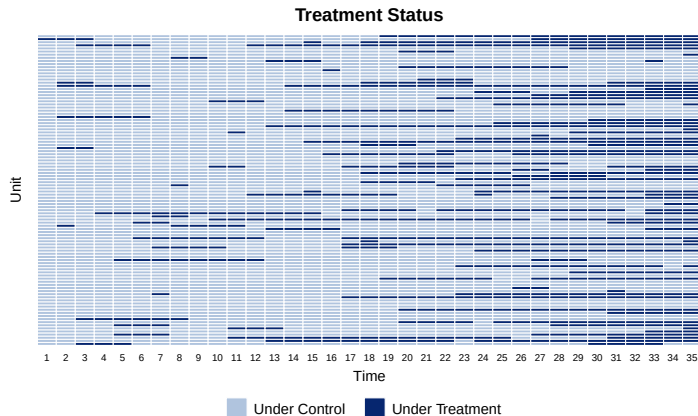
- One adoption date, many periods: still essentially a 2×2 comparison

Beyond Two Periods: Staggered Adoption



- Different units adopt at different times: the classic staggered setting

Beyond Two Periods: The General Case (with Treatment Reversal)



- Treatment can also switch off; the majority of political science applications look like this (we return to it in Lecture 5)

$$Y_{it} = \tau D_{it} + \mathbf{X}_{it}'\boldsymbol{\beta} + \alpha_i + \xi_t + \varepsilon_{it}, \quad D_{it} \in \{0, 1\}$$

- **Functional form**

- Additive fixed effects
- **Constant** and contemporaneous treatment effect
- Linearity in covariates

- Note the difference from the DID setup: DID allowed τ_{it} to vary; the TWFE regression assumes a single τ

⇒ When effects are actually heterogeneous, this assumption fails and induces the **(negative) weighting** problem

- **Strict exogeneity:**

$$\varepsilon_{it} \perp\!\!\!\perp D_{js}, \mathbf{X}_{js}, \alpha_j, \xi_s \quad \forall i, j, t, s$$

- Which implies

$$\{Y_{it}(0), Y_{it}(1)\} \perp\!\!\!\perp D_{js} \mid \mathbf{X}, \alpha, \xi \quad \forall i, j, t, s$$

- With only two groups, this delivers parallel trends:

$$\mathbb{E}[Y_{it}(0) - Y_{it'}(0) \mid \mathbf{X}] = \mathbb{E}[Y_{jt}(0) - Y_{jt'}(0) \mid \mathbf{X}], \quad i \in \mathcal{T}, j \in \mathcal{C}, \forall t, t'$$

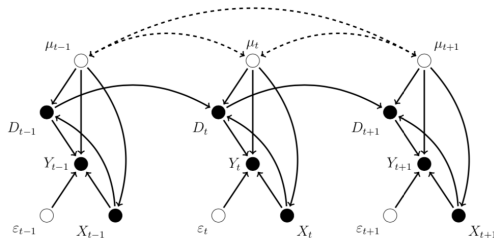
$$Y_{it} = \tau D_{it} + \mathbf{X}_{it}'\boldsymbol{\beta} + \alpha_i + \xi_t + \varepsilon_{it}$$

- Assumptions
 - Functional form
 - Strict exogeneity
- Challenges
 - Strict exogeneity means a lot more than you think
 - Prevalent parallel trends failures
 - Treatment effect heterogeneity leads to bias

We take the three in turn: the design problem, then parallel trends, then weighting.

Strict Exogeneity Means More Than You Think

Reinterpreting strict exogeneity (Imai and Kim 2019):



- 1 No unobserved **time-varying** confounder exists
- 2 No lagged dependent variable $\rightarrow$ adding an LDV is fine if T is large; not adding it is also fine
- 3 No “carryover effect” $\rightarrow$ adding lagged treatment terms is fine
- 4 Past outcomes do not directly affect current treatment (no “feedback”)

Each of these is a substantive claim about the world, not a regression option.

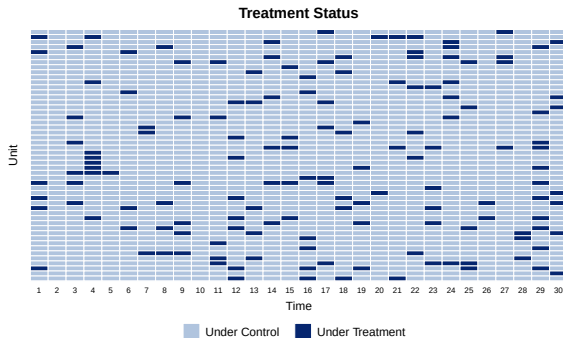
Hypothetical Experiment?

DGPs consistent with strict exogeneity:

$$\alpha_i, X_i \rightarrow D_i \rightarrow Y_i$$

Treatment status is assigned randomly or in one shot, not sequentially!

Examples: **random assignment** within units



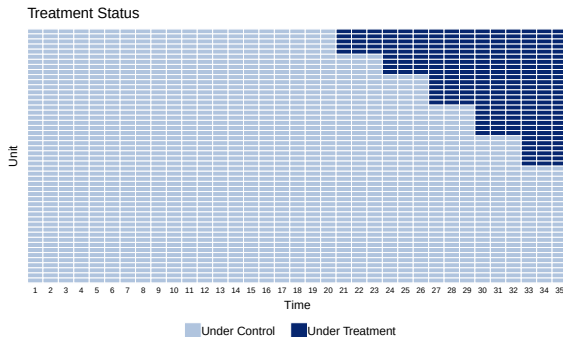
Hypothetical Experiment?

DGPs consistent with strict exogeneity:

$$\alpha_i, \mathbf{X}_i \rightarrow A_i \rightarrow D_i \rightarrow Y_i$$

Treatment status is assigned randomly or in one shot, not sequentially!

Examples: **staggered adoption**, with A_i the adoption time (Athey and Imbens 2022)



The Event-Study (Dynamic TWFE) Specification

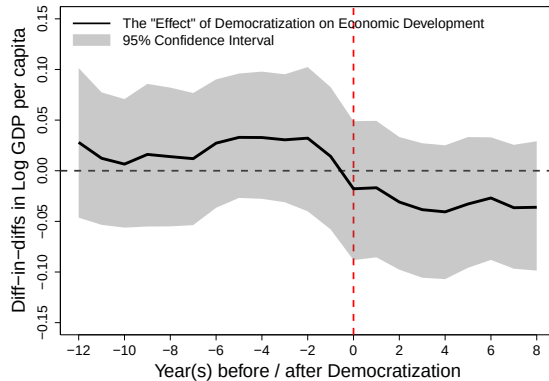
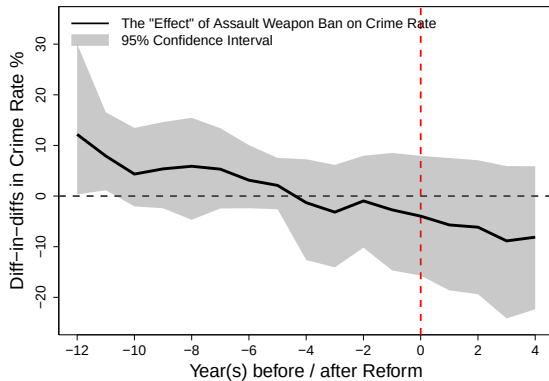
- To probe parallel trends and trace dynamics, applied work estimates

$$Y_{it} = \sum_{s \neq -1} \mu_s \mathbf{1}\{t - T_i = s\} + \alpha_i + \xi_t + \varepsilon_{it}$$

where T_i is unit i 's adoption date and s counts periods relative to adoption

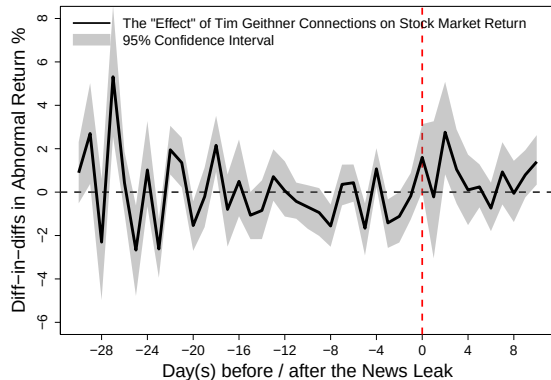
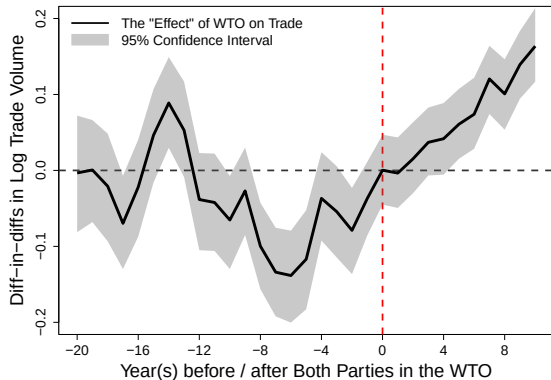
- The period just before adoption ($s = -1$) is the omitted reference category
- **Leads** ($s < -1$) probe pre-treatment differences; **lags** ($s \geq 0$) trace the dynamic effects
- Endpoints are usually binned (e.g. $s \leq -8$, $s \geq 12$) so each coefficient rests on enough observations

Failure of Parallel Trends Is Common (1)



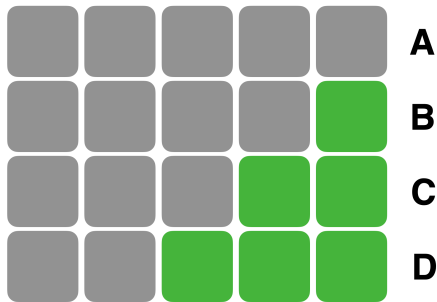
- Pre-treatment differences trend before the treatment ever arrives

Failure of Parallel Trends Is Common (2)



- Lesson: always plot the pre-treatment differences before interpreting the post-treatment ones

What's the Weighting Problem?



- Question: can TWFE recover (some **non-negatively weighted**) ATT when treatment effects are heterogeneous?
- Probably not! e.g. Goodman-Bacon (2021), de Chaisemartin and D'Haultfoeulle (2020), Callaway and Sant'Anna (2021), Borusyak et al. (2024), Strezhnev (2018), Imai and Kim (2021)
- Early adopters (e.g. D) serve as comparisons for late adopters (e.g. B)

⇒ Some **treated** observations are used as comparisons

- Under staggered adoption, the TWFE estimator

$$Y_{it} = \beta^{TWFE} D_{it} + \alpha_i + \xi_t + \varepsilon_{it}$$

is a **weighted average of all possible 2×2 DID estimators** that compare timing groups to each other

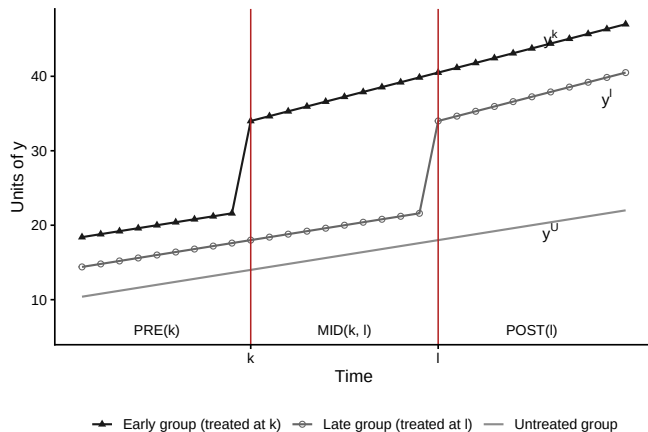
- The weights are proportional to timing-group sizes and the variance of the treatment dummy in each pair, which is highest for units treated in the middle of the panel
- Source of biases:

$$\text{plim}_{N \rightarrow \infty} \hat{\beta}^{TWFE} = VWATT + VWCT - \Delta ATT$$

- $VWATT$: variance-weighted ATT
- $VWCT$: variance-weighted common trends
- ΔATT : change in treatment effects over time

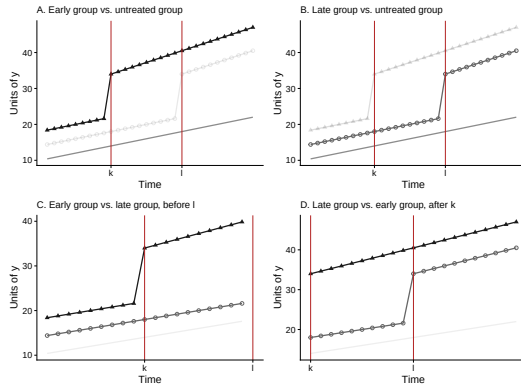
The Three-Group Case

- An early group k , treated at $t_i = k$
- A late group ℓ , treated at $t_i = \ell > k$
- An untreated group U , $t_i = \infty$



Source: Redrawn after Goodman-Bacon (2021), Figure 1 (goodman_bacon_2026.R).

Four Simple (2×2) DID Estimates

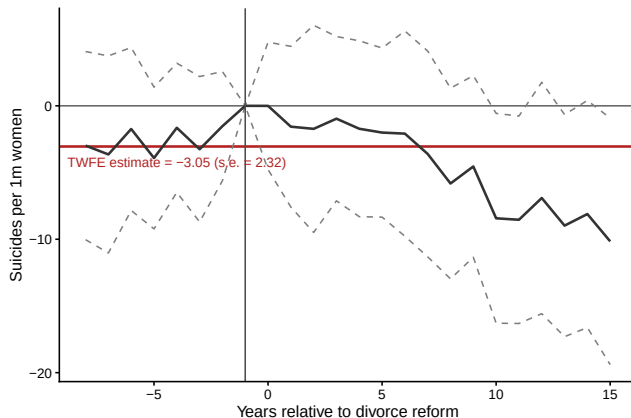


- Panel D is the troublemaker: the **already-treated** early group serves as the comparison, and its evolving treatment effect enters with a **negative sign**

Source: Goodman-Bacon (2021).

Example: No-Fault Divorce Reforms and Female Suicide

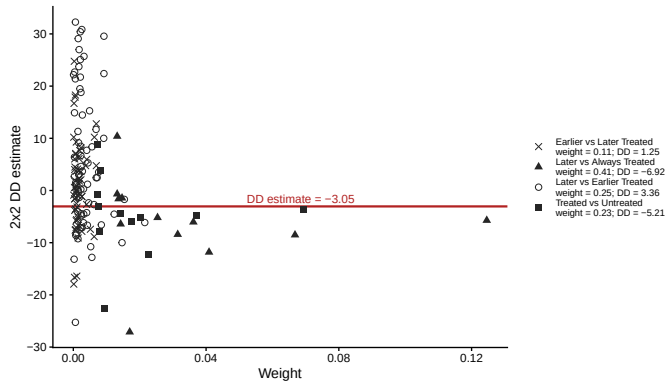
- 37 states adopted unilateral divorce between 1969 and 1985 (9 had reformed before 1964, 5 never did); event-study and TWFE estimates:



Source: Own replication of the divorce example in Goodman-Bacon (2021); Stevenson and Wolfers (2006) data from the `bacondecomp` package

Decomposing the Divorce Estimate

- Plot each 2×2 DID against its weight; average effect and total weight by comparison type:



- The TWFE estimate, -3.05 (Goodman-Bacon reports -3.08), is the average of the y-axis values weighted by the x-axis values

Source: Own replication with the `bacondecomp` package (goodman-bacon 2026.R).

- A complementary result: under heterogeneous effects, the TWFE coefficient satisfies

$$\hat{\beta}^{TWFE} \xrightarrow{p} \mathbb{E} \left[\sum_{(i,t): D_{it}=1} w_{it} \tau_{it} \right]$$

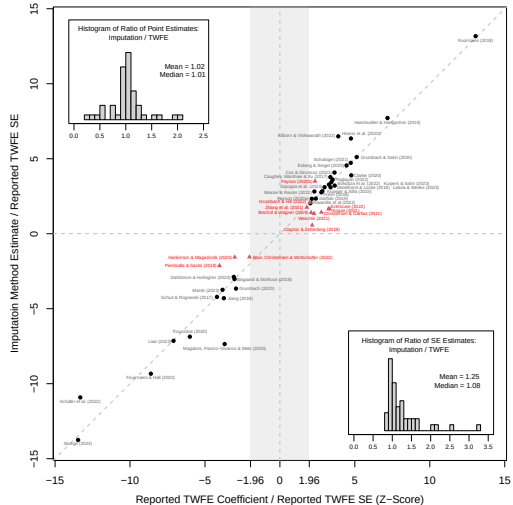
where the weights w_{it} sum to one but **some can be negative**

- Negative weights attach to observations treated for a long time in periods when many units are treated
- With negative weights, $\hat{\beta}^{TWFE}$ can be negative even if every $\tau_{it} > 0$: a **sign reversal**
- Diagnostic: the `twowayfweights` package reports the share and sum of negative weights for your design

- The same logic contaminates the dynamic specification: each lead or lag coefficient $\hat{\mu}_s$ is a weighted average that picks up effects from **other relative periods** and **other cohorts**
- In particular, lead coefficients can appear non-zero even when parallel trends holds, and can appear zero when it fails
- So under staggered adoption with HTE, the standard event-study plot is not a reliable pre-trend diagnostic
- Fix: estimate cohort-specific effects and aggregate them explicitly, e.g. the interaction-weighted estimator (Lecture 5)

How Bad Is the Weighting Problem in Practice? (Chiu et al. 2026)

- Not too bad!
- Reanalyzing published studies from top political science journals, we find:
 - There is almost no sign flipping
 - No systematic overestimation or underestimation of the treatment effects
 - Some marginally significant results become insignificant
 - However, parallel trends violations remain a major concern
- Lecture 5 returns to this study in detail



- TWFE on staggered data mixes clean and contaminated comparisons; with heterogeneous effects, the estimand is no longer any ATT we care about
 - In practice, the first-order threat is usually **parallel trends failure**, not the weighting problem itself
 - Both problems have the same root: the regression does not distinguish *which* observations serve as comparisons for *which* treated observations
 - The fix: make the comparisons explicit, which is what the modern estimators do
- ⇒ The road ahead: Lecture 5 repairs the **weighting problem**; Lecture 6 introduces low-rank methods that can address **some** parallel trends violations.

- **Condition on pre-treatment covariates:** parallel trends holds within strata of X
 - Semiparametric DID, e.g. [Abadie \(2005\)](#); [Callaway and Sant'Anna \(2021\)](#) → Lecture 5
- **Condition on pre-treatment outcomes:** build the comparison from units whose pre-trends match
 - Synthetic control, e.g. [Abadie et al. \(2010, 2015\)](#) → Lecture 6
- **Model the confounding:** impute $Y_{it}(0)$ from a low-rank (latent factor) model fit on untreated data
 - [Gobillon and Magnac \(2016\)](#); [Xu \(2017\)](#); [Athey et al. \(2021\)](#) → Lecture 6
- **Combine weighting and modeling:** doubly robust hybrids
 - Augmented synthetic control ([Ben-Michael et al. 2021](#)); synthetic DID ([Arkhangelsky et al. 2021](#)) → Lecture 6

- TWFE models require stronger assumptions than we normally admit; both treatment effect heterogeneity and parallel trends failures can lead to bias
- Under staggered adoption, TWFE averages all 2×2 comparisons, including ones that use already-treated units as comparisons
- In reanalyses of published work, the weighting bias is usually modest; parallel trends violations are the bigger threat
- Current solutions: matching/reweighting, imputation, and their combinations
- Whichever method you use, run the diagnostics: plot the pre-trend and run placebo tests. They can reject parallel trends, but they cannot prove it.

- `fixest`: fast fixed effects regressions
- `bacondecomp`: Goodman-Bacon decomposition
- `did`: Callaway and Sant'Anna's estimator
- `PanelMatch`: panel matching
- `fect`: imputation estimators (FEct, IFect, MC) with diagnostic tests
- `gsynth`: interactive fixed effects with non-reversible treatments
- `panelView`: panel data visualization

Lectures 5 and 6 demonstrate several of these in worked examples.

References

- Abadie, Alberto. 2005. "Semiparametric Difference-in-Differences Estimators." *Review of Economic Studies* 72 (1): 1–19.
- Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105 (490): 493–505.
- Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2015. "Comparative Politics and the Synthetic Control Method." *American Journal of Political Science* 59 (2): 495–510.
- Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. 2021. "Synthetic Difference-in-Differences." *American Economic Review* 111 (12): 4088–118.
- Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. 2021. "Matrix Completion Methods for Causal Panel Data Models." *Journal of the American Statistical Association* 116 (536): 1716–30.
- Athey, Susan, and Guido W. Imbens. 2022. "Design-Based Analysis in Difference-in-Differences Settings with Staggered Adoption." *Journal of Econometrics* 226 (1): 62–79.
- Ben-Michael, Eli, Avi Feller, and Jesse Rothstein. 2021. "The Augmented Synthetic Control Method." *Journal of the American Statistical Association* 116 (536): 1789–803.
- Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. "Revisiting Event-Study Designs: Robust and Efficient Estimation." *Review of Economic Studies* 91 (6): 3253–85.
- Callaway, Brantly, and Pedro H. C. Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics* 225 (2): 200–230.
- de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review* 110 (9): 2964–96.

References (cont.)

- Chiu, Albert, Xingchen Lan, Ziyi Liu, and Yiqing Xu. 2026. "Causal Panel Analysis under Parallel Trends: Lessons from a Large Reanalysis Study." *American Political Science Review* 120 (1): 245–66.
- Gobillon, Laurent, and Thierry Magnac. 2016. "Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls." *Review of Economics and Statistics* 98 (3): 535–51.
- Goodman-Bacon, Andrew. 2021. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics* 225 (2): 254–77.
- Imai, Kosuke, and In Song Kim. 2019. "When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" *American Journal of Political Science* 63 (2): 467–90.
- Imai, Kosuke, and In Song Kim. 2021. "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data." *Political Analysis* 29 (3): 405–15.
- Strezhnev, Anton. 2018. "Semiparametric Weighting Estimators for Multi-Period Difference-in-Differences Designs." Working paper.
- Sun, Liyang, and Sarah Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics* 225 (2): 175–99.
- Xu, Yiqing. 2017. "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models." *Political Analysis* 25 (1): 57–76.