From TWFE to Callaway–Sant'Anna

A Derivation of the Modern Difference-in-Differences Toolkit

Bas Machielsen

2026-05-18

Introduction

What this lecture is about

- The Difference-in-Differences (DiD) **revolution** of the last few years.
- The classical workhorse — the **Two-Way Fixed Effects** (TWFE) regression — can be badly biased under **staggered** treatment adoption.
- We will see exactly *why* it fails, and then build the **Callaway and Sant'Anna (2021)** estimator that fixes it.
- The path: derive everything from potential outcomes, line by line.

Why care?

- Staggered adoption is the norm: minimum wages, RTW laws, abortion policy, Medicaid expansions, ...
- A long literature ran TWFE on staggered designs and reported a single $\hat{\beta}$.
- Goodman-Bacon (2021), de Chaisemartin & D'Haultfœuille (2020), Sun & Abraham (2021), and Callaway & Sant'Anna (2021) showed: that $\hat{\beta}$ can be sign-flipped from the true ATT even when *every* unit has a positive effect.
- The "fix" is conceptual: stop starting from an estimator; start from an **estimand**.

Roadmap

1. The classic 2×2 DiD.
2. The TWFE regression in 2×2.
3. The breakdown of TWFE under staggered timing — *forbidden comparisons*.
4. The Goodman-Bacon decomposition (mechanics).
5. The Frisch-Waugh-Lovell (FWL) theorem, properly proved.
6. The Callaway-Sant'Anna building block: $ATT(g, t)$.
7. **Never-treated** vs **not-yet-treated** controls.
8. Identifying assumptions: no anticipation, random timing.
9. Aggregation: event-study and group-specific.

Notation conventions

- Panel: units $i = 1, \dots, N$, time $t = 1, \dots, T$.
- Y_{it}: observed outcome. $Y_{it}(0), Y_{it}(1)$: potential outcomes.
- D_i: treatment-group indicator. Treat_{it}: treated-now indicator.
- G_i: first treatment period (the *cohort*). $G_i = \infty$ for the never-treated.
- All expectations are taken in the population unless noted; replace with sample averages for the estimator.

The classic 2×2 DiD

Setup

- Two periods $t \in \{1, 2\}$, two groups.
- $D_i \in \{0, 1\}$: 1 if unit i is ever treated, 0 otherwise.
- Nobody is treated in $t = 1$; treated units are treated in $t = 2$.
- Observed outcome:

$$Y_{it} = D_i \cdot \mathbb{I}(t = 2) \cdot Y_{it}(1) + \big(1 - D_i \cdot \mathbb{I}(t = 2)\big) \cdot Y_{it}(0).$$

The target: ATT

The **Average Treatment Effect on the Treated** in period 2:

$$ATT = \mathbb{E}[Y_{i2}(1) - Y_{i2}(0) \,|\, D_i = 1]\,.$$

- $\mathbb{E}[Y_{i2}(1) \mid D_i = 1]$ is identified directly: it is $\mathbb{E}[Y_{i2} \mid D_i = 1]$.
- $\mathbb{E}[Y_{i2}(0) \mid D_i = 1]$ is the **counterfactual** — never observed.
- We need an assumption to pin it down.

Parallel trends

The Parallel Trends Assumption (PT):

$$\mathbb{E}[Y_{i2}(0) - Y_{i1}(0) \,|\, D_i = 1] \;=\; \mathbb{E}[Y_{i2}(0) - Y_{i1}(0) \,|\, D_i = 0]\,.$$

- "Absent treatment, the treated group's outcome would have trended in parallel with the control group's."
- *Not* a statement about levels — only differences.

Identification from observables

Start from $ATT = \mathbb{E}[Y_{i2}(1) \mid D = 1] - \mathbb{E}[Y_{i2}(0) \mid D = 1]$.

Identification from observables

Start from $ATT = \mathbb{E}[Y_{i2}(1) \mid D = 1] - \mathbb{E}[Y_{i2}(0) \mid D = 1]$.

Use PT to express the missing counterfactual:

$$\mathbb{E}[Y_{i2}(0) \mid D = 1] = \mathbb{E}[Y_{i1}(0) \mid D = 1] + \mathbb{E}[Y_{i2}(0) - Y_{i1}(0) \mid D = 0].$$

Identification from observables

Start from $ATT = \mathbb{E}[Y_{i2}(1) \mid D = 1] - \mathbb{E}[Y_{i2}(0) \mid D = 1]$.

Use PT to express the missing counterfactual:

$$\mathbb{E}[Y_{i2}(0) \mid D = 1] = \mathbb{E}[Y_{i1}(0) \mid D = 1] + \mathbb{E}[Y_{i2}(0) - Y_{i1}(0) \mid D = 0].$$

Since nobody is treated in $t = 1$ and the control is never treated, $Y(0)$ equals Y in those cells:

$$ATT = \left(\mathbb{E}[Y_{i2} \mid D = 1] - \mathbb{E}[Y_{i1} \mid D = 1]\right) \; - \; \left(\mathbb{E}[Y_{i2} \mid D = 0] - \mathbb{E}[Y_{i1} \mid D = 0]\right).$$

This is the canonical 2×2 DiD estimand.

The TWFE regression

The standard specification

Let $\text{Treat}_{it} = D_i \cdot \mathbb{I}(t = 2)$. Run:

$$Y_{it} = \alpha_i + \lambda_t + \beta^{TWFE}\text{Treat}_{it} + \varepsilon_{it},$$

with strict exogeneity $\mathbb{E}[\varepsilon_{it} \mid D_i, t] = 0$.

- α_i: unit fixed effects (time-invariant heterogeneity).
- λ_t: time fixed effects (common shocks).
- β^{TWFE}: the coefficient of interest. Target: the ATT from Section 2.

We will show, step by step, that $\beta^{TWFE} = ATT$.

Conditional expectations, cell by cell

Take $\mathbb{E}[\cdot \mid D_i, t]$ of the regression equation. Strict exogeneity zeroes the error, leaving:

Cell	$\mathbb{E}[Y_{it} \mid D_i, t]$
$(D = 1,\, t = 2)$	$\mathbb{E}[\alpha_i \mid D = 1] + \lambda_2 + \beta^{TWFE}$
$(D = 1,\, t = 1)$	$\mathbb{E}[\alpha_i \mid D = 1] + \lambda_1$
$(D = 0,\, t = 2)$	$\mathbb{E}[\alpha_i \mid D = 0] + \lambda_2$
$(D = 0,\, t = 1)$	$\mathbb{E}[\alpha_i \mid D = 0] + \lambda_1$

Within-group **change** from $t = 1$ to $t = 2$ (unit FE drops out, since it is constant within group):

$$\mathbb{E}[Y_{i2} \mid D = 1] - \mathbb{E}[Y_{i1} \mid D = 1] = (\lambda_2 - \lambda_1) + \beta^{TWFE},$$

$$\mathbb{E}[Y_{i2} \mid D = 0] - \mathbb{E}[Y_{i1} \mid D = 0] = (\lambda_2 - \lambda_1).$$

β^{TWFE} equals the 2×2 DiD estimand

Subtract the control-group change from the treated-group change. The common time effect $(\lambda_2 - \lambda_1)$ cancels:

$$\underbrace{\left(\mathbb{E}[Y_{i2} \mid D = 1] - \mathbb{E}[Y_{i1} \mid D = 1]\right) - \left(\mathbb{E}[Y_{i2} \mid D = 0] - \mathbb{E}[Y_{i1} \mid D = 0]\right)}_{2 \times 2 \text{ DiD estimand}}$$

$$= \left((\lambda_2 - \lambda_1) + \beta^{TWFE}\right) - (\lambda_2 - \lambda_1) = \beta^{TWFE}.$$

So under the TWFE specification + strict exogeneity, the regression coefficient *is* the population 2×2 DiD estimand — no approximation, exact identity.

β^{TWFE} identifies the ATT

Chain this with the result from Section 2 (PT $\Rightarrow 2 \times 2$ DiD identifies the ATT):

$$\beta^{TWFE} \overset{\text{regression algebra}}{=} 2 \times 2 \text{ DiD estimand} \overset{\text{PT}}{=} ATT.$$

- The first equality is mechanical: it follows from the regression equation and strict exogeneity.
- The second equality requires **Parallel Trends** to map the DiD onto a causal object.
- Together, in the 2×2 design, TWFE *directly* recovers the ATT in period 2.

This is exactly why TWFE became the workhorse for two-period, two-group DiD. Trouble begins only beyond 2×2 — and that is Section 4.

What TWFE quietly assumes

In the regression equation, every treated cell carries the *same* β:

$$Y_{i2}(1) - Y_{i2}(0) = \beta, \quad \forall i.$$

- The treatment effect is **constant across units**.
- It is also **constant across time** (no dynamics).
- This is the **homogeneous treatment effects** restriction. Innocent here, fatal later.

Staggered timing breaks TWFE

A three-period thought experiment

- $t \in \{1, 2, 3\}$. Treatment is absorbing.
- **Group A** (early): treated from $t = 2$.
- **Group B** (late): treated from $t = 3$.

With staggered timing, the TWFE estimator is a weighted average of *all* possible 2×2 DiD comparisons. One of them is the *forbidden* comparison: using already-treated Group A as a control for newly treated Group B.

The forbidden 2×2

Treat B as "treated" (changes status between $t = 2$ and $t = 3$) and A as "control" (no change in Treat between $t = 2$ and $t = 3$):

$$DiD_{B,A} = (\mathbb{E}[Y_{i3} \mid B] - \mathbb{E}[Y_{i2} \mid B]) - (\mathbb{E}[Y_{i3} \mid A] - \mathbb{E}[Y_{i2} \mid A]).$$

What does this identify? Substitute in potential outcomes carefully.

Potential outcomes in $DiD_{B,A}$

For Group B: treated only at $t = 3$, so $Y_{i3} = Y_{i3}(1)$ but $Y_{i2} = Y_{i2}(0)$.

Potential outcomes in $DiD_{B,A}$

For Group B: treated only at $t = 3$, so $Y_{i3} = Y_{i3}(1)$ but $Y_{i2} = Y_{i2}(0)$.

For Group A: treated at both $t = 2$ and $t = 3$, so $Y_{i3} = Y_{i3}(1)$ and $Y_{i2} = Y_{i2}(1)$.

Potential outcomes in $DiD_{B,A}$

For Group B: treated only at $t = 3$, so $Y_{i3} = Y_{i3}(1)$ but $Y_{i2} = Y_{i2}(0)$.

For Group A: treated at both $t = 2$ and $t = 3$, so $Y_{i3} = Y_{i3}(1)$ and $Y_{i2} = Y_{i2}(1)$.

Therefore:

$$DiD_{B,A} = \big(\mathbb{E}[Y_{i3}(1) \mid B] - \mathbb{E}[Y_{i2}(0) \mid B]\big) - \big(\mathbb{E}[Y_{i3}(1) \mid A] - \mathbb{E}[Y_{i2}(1) \mid A]\big).$$

Apply parallel trends

Assume PT for untreated potential outcomes:

$$\mathbb{E}[Y_{i3}(0) - Y_{i2}(0) \mid B] = \mathbb{E}[Y_{i3}(0) - Y_{i2}(0) \mid A].$$

In the **first** bracket: add and subtract $\mathbb{E}[Y_{i3}(0) \mid B]$:

$$\mathbb{E}[Y_{i3}(1) \mid B] - \mathbb{E}[Y_{i2}(0) \mid B] = ATT_{B,3} + \mathbb{E}[Y_{i3}(0) - Y_{i2}(0) \mid B].$$

In the **second** bracket: add and subtract $\mathbb{E}[Y_{i3}(0) \mid A]$ and $\mathbb{E}[Y_{i2}(0) \mid A]$:

$$\mathbb{E}[Y_{i3}(1) \mid A] - \mathbb{E}[Y_{i2}(1) \mid A] = (ATT_{A,3} - ATT_{A,2}) + \mathbb{E}[Y_{i3}(0) - Y_{i2}(0) \mid A].$$

The forbidden comparison, simplified

Subtract; PT cancels the $Y(0)$-drift terms:

$$\boxed{DiD_{B,A} = ATT_{B,3} - (ATT_{A,3} - ATT_{A,2}).}$$

- The clean piece $ATT_{B,3}$ is what we want.
- The contamination $-(ATT_{A,3} - ATT_{A,2})$ is **the change in Group A's treatment effect over time**.

What goes wrong

- If Group A's treatment effect **grows** ($ATT_{A,3} > ATT_{A,2}$), the forbidden comparison **subtracts** that growth from our estimate of $ATT_{B,3}$.
- TWFE blindly uses A as a control and mistakes the *evolution of A's treatment effect* for a baseline time trend.
- In dramatic cases, $\hat{\beta}^{TWFE}$ can be **negative even when every unit has a positive effect**.

Why does OLS do this?

The ΔTreat illusion

OLS sweeps out unit fixed effects by looking at within-unit *changes*. Between $t = 2$ and $t = 3$:

$$\Delta\text{Treat}_A = 1 - 1 = 0, \qquad \Delta\text{Treat}_B = 1 - 0 = 1.$$

- To OLS, $\Delta\text{Treat}_A = 0$ looks identical to "never treated."
- It happily uses Group A as the control to soak up the $t = 2 \to t = 3$ time effect.
- That is exactly how the forbidden comparison is born.

A two-period mini-proof

Restrict to $t \in \{2, 3\}$. TWFE is equivalent to first differences. For each unit:

$$\Delta Y_i = \Delta\lambda + \beta\,\Delta\text{Treat}_i + \Delta\varepsilon_i.$$

- Group A: $\Delta\text{Treat} = 0$.
- Group B: $\Delta\text{Treat} = 1$.

A regression of ΔY on a single binary regressor (with intercept) returns the difference of means in ΔY:

$$\hat{\beta}^{TWFE} = \mathbb{E}[\Delta Y \mid B] - \mathbb{E}[\Delta Y \mid A].$$

The two-period mini-proof, expanded

Expand ΔY_i back to levels:

$$\hat{\beta}^{TWFE} = (\mathbb{E}[Y_{i3} \mid B] - \mathbb{E}[Y_{i2} \mid B]) - (\mathbb{E}[Y_{i3} \mid A] - \mathbb{E}[Y_{i2} \mid A]).$$

That **is** $DiD_{B,A}$. OLS did not "decide" to use A as a control — the fixed-effects sweep mechanically turned A's constant treatment status into "no change" and made A look like a valid control.

Variance weighting (preview)

In a panel with many windows, OLS combines many 2×2's by **variance-weighting**:

- The variance of a binary Treat with mean p is $p(1-p)$.
- Cohorts treated near the *middle* of the panel maximize $p(1-p)$ and get the most weight.
- Within a window where a cohort's Treat is constant ($p = 0$ or 1), it contributes **zero variance** there, and OLS recruits whichever other cohort *does* vary in that window — even if that means using an already-treated cohort as a control.

The exact weights drop out of the FWL theorem next.

The Goodman-Bacon decomposition

Setup

- Two cohorts: **Early** (E, share n_E) and **Late** (L, share n_L), $n_E + n_L = 1$.
- Three windows, shares $t_1 + t_2 + t_3 = 1$:
  - **Pre** (window 1): nobody treated.
  - **Mid** (window 2): only E treated.
  - **Post** (window 3): both treated.
- Six cells (g, w). Treatment $D_{g,w} = 1$ in $(E, 2), (E, 3), (L, 3)$; zero elsewhere.

FWL applied to TWFE

By Frisch-Waugh-Lovell (proved next section),

$$\hat{\beta}^{TWFE} = \frac{\sum_{i,t} \widetilde{D}_{it} Y_{it}}{\sum_{i,t} \widetilde{D}_{it}^2}, \tag{1}$$

where $\widetilde{D}_{it}$ is the residual from regressing D_{it} on the unit and time fixed effects:

$$\widetilde{D}_{it} = D_{it} - \bar{D}_{g(i)} - \bar{D}_{w(t)} + \bar{D}. \tag{2}$$

Because D_{it}, $\bar{D}_g$, $\bar{D}_w$ are constant inside a (g, w) cell, so is $\widetilde{D}_{g,w}$.

Computing the means

- **Group means**: $\bar{D}_E = t_2 + t_3$, $\bar{D}_L = t_3$.
- **Time means**: $\bar{D}_1 = 0$, $\bar{D}_2 = n_E$, $\bar{D}_3 = 1$.
- **Grand mean**: sum cell mass $\times$ cell treatment over the three treated cells $(E, 2), (E, 3), (L, 3)$:

$$\bar{D} = \sum_{g,w} n_g \, t_w \, D_{g,w} = n_E t_2 \cdot 1 \; + \; n_E t_3 \cdot 1 \; + \; n_L t_3 \cdot 1$$

$$= n_E t_2 + (n_E + n_L) \, t_3 = n_E t_2 + t_3,$$

using $n_E + n_L = 1$. (Cross-check: also equals $n_E \bar{D}_E + n_L \bar{D}_L = n_E(t_2 + t_3) + n_L t_3 = n_E t_2 + t_3$.)

Computing $\widetilde{D}_{g,w}$ in one cell

Take cell $(E, 1)$ (early group, pre window):

$$\widetilde{D}_{E,1} = D_{E,1} - \bar{D}_E - \bar{D}_1 + \bar{D} = 0 - (t_2 + t_3) - 0 + (n_E t_2 + t_3) = t_2(n_E - 1) = -n_L t_2.$$

Doing the same for the other five cells:

Cell	Pre $(w = 1)$	Mid $(w = 2)$	Post $(w = 3)$
Early ($g = E$)	$-n_L t_2$	$+n_L(t_1 + t_3)$	$-n_L t_2$
Late ($g = L$)	$+n_E t_2$	$-n_E(t_1 + t_3)$	$+n_E t_2$

A symmetry to notice

- $\widetilde{D}_{E,1} = \widetilde{D}_{E,3}$ and $\widetilde{D}_{L,1} = \widetilde{D}_{L,3}$.
- In Pre and Post, the two groups have *equal* treatment status (0–0 in Pre, 1–1 in Post).
- The fixed-effects projection cannot separate Pre from Post — they look identical *to the projection*.
- Only the **Mid** window carries asymmetric information about treatment. The Mid window is the fulcrum.

The numerator

The numerator collapses to cells (justified in Section 7):

$$\sum_{i,t} \widetilde{D}_{it}\, Y_{it} \quad \propto \quad \sum_{g,w} (n_g t_w)\, \widetilde{D}_{g,w}\, \bar{Y}_{g,w}. \tag{3}$$

Plug in $\widetilde{D}_{g,w}$:

$$\text{Num} = n_E n_L \Big[-t_1 t_2\, \bar{Y}_{E,1} + t_1 t_2\, \bar{Y}_{L,1} + t_2(t_1+t_3)\bar{Y}_{E,2} - t_2(t_1+t_3)\bar{Y}_{L,2}$$

$$-t_2 t_3\, \bar{Y}_{E,3} + t_2 t_3\, \bar{Y}_{L,3} \Big].$$

Every cell carried a factor of $n_E n_L$ that we pulled out front.

Numerator, regrouped

Expand $t_2(t_1 + t_3) = t_1 t_2 + t_2 t_3$ in the two middle terms, then collect by $t_1 t_2$ and $t_2 t_3$:

$$\text{Num} = n_E n_L \Big[t_1 t_2 \underbrace{\big((\bar{Y}_{E,2} - \bar{Y}_{E,1}) - (\bar{Y}_{L,2} - \bar{Y}_{L,1})\big)}_{= DiD_{\text{Clean}}}$$

$$+ t_2 t_3 \underbrace{\big((\bar{Y}_{L,3} - \bar{Y}_{L,2}) - (\bar{Y}_{E,3} - \bar{Y}_{E,2})\big)}_{= DiD_{\text{Forbidden}}} \Big].$$

Out pops the clean comparison (E treats, L still control) and the forbidden one (L treats, E used as control).

The denominator

The denominator likewise collapses to

$$\sum_{g,w} (n_g t_w)\, \widetilde{D}_{g,w}\, D_{g,w} \tag{4}$$

(note the raw $D_{g,w}$, not $\widetilde{D}_{g,w}$ — justified by orthogonality in Section 7). It has nonzero contributions only where $D_{g,w} = 1$:

$$\text{Denom} = n_E t_2 \cdot n_L(t_1 + t_3) + n_E t_3 \cdot (-n_L t_2) + n_L t_3 \cdot n_E t_2$$

$$= n_E n_L \big[t_2(t_1 + t_3) - t_2 t_3 + t_2 t_3\big] = n_E n_L \cdot t_2(t_1 + t_3).$$

The Goodman-Bacon decomposition

Take the ratio. The $n_E n_L$ and t_2 cancel:

$$\boxed{\hat{\beta}^{TWFE} = \left(\frac{t_1}{t_1 + t_3}\right) DiD_{\text{Clean}} + \left(\frac{t_3}{t_1 + t_3}\right) DiD_{\text{Forbidden}}.}$$

The weights depend only on the **relative durations** of the Pre and Post windows. The Mid window is the fulcrum and washes out of the *weights*.

Insight 1: the long-post-period trap

When the **post window is long relative to the pre window**, OLS leans heavily on the forbidden comparison.

- 20-year panel, treatment adoption between years 2 and 5: $t_1 \approx 0.10$, $t_3 \approx 0.75$.
- Forbidden weight $\approx t_3/(t_1 + t_3) \approx 0.88$.

Most applied DiD papers have exactly this structure. That is why TWFE was so vulnerable.

Insight 2: when TWFE is "safe"

When **all treatment happens at the very end of the panel**, $t_3 \to 0$.

- Forbidden weight $\to 0$. Clean weight $\to 1$.
- TWFE reduces to the clean DiD.

Diagnostic: before reporting $\hat{\beta}^{TWFE}$, compute $t_3/(t_1 + t_3)$ for your panel.

Insight 3: group-size amplification

In the multi-group case, the $n_E n_L$ factor *does not* fully cancel: it modulates each pairwise weight.

- A large early-adopter cohort acts as a big, stable pool of "constant-treatment" units — ideal OLS bait for soaking up time fixed effects.
- A large early cohort therefore **amplifies** the weight of the forbidden comparison across the entire regression.

A proper proof of FWL

The FWL theorem

Consider a partitioned linear regression

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon.$$

Claim. The OLS estimator $\hat{\beta}_1$ equals the OLS estimator from regressing Y (or equivalently $\widetilde{Y}$, the residual of Y on X_2) on $\widetilde{X}_1$, the residual of X_1 on X_2:

$$\hat{\beta}_1 = (\widetilde{X}_1'\widetilde{X}_1)^{-1}\widetilde{X}_1'Y = (\widetilde{X}_1'\widetilde{X}_1)^{-1}\widetilde{X}_1'\widetilde{Y}. \tag{5}$$

We will prove both equalities using projection matrices.

Projection matrix toolkit

For a full-column-rank X_2:

$$P_{X_2} = X_2(X_2'X_2)^{-1}X_2', \qquad M_{X_2} = I - P_{X_2}.$$

Properties used below:

1. **Symmetry**: $P_{X_2}' = P_{X_2}$, $M_{X_2}' = M_{X_2}$.
2. **Idempotency**: $P_{X_2}^2 = P_{X_2}$, $M_{X_2}^2 = M_{X_2}$.
3. **Annihilation**: $M_{X_2}X_2 = 0$ and $P_{X_2}X_2 = X_2$.
4. **Orthogonal decomposition**: for any vector v, $v = P_{X_2}v + M_{X_2}v$ and $(M_{X_2}v)'(P_{X_2}v) = 0$.

By definition $\widetilde{X}_1 = M_{X_2}X_1$ and $\widetilde{Y} = M_{X_2}Y$.

Proof: $\hat{\beta}_1 = (\widetilde{X}_1'\widetilde{X}_1)^{-1}\widetilde{X}_1'Y$

Premultiply the normal equations for $\hat{\beta}_1$ on both sides by M_{X_2}:

$$M_{X_2}Y = M_{X_2}X_1\hat{\beta}_1 + M_{X_2}X_2\hat{\beta}_2 + M_{X_2}\hat{\varepsilon}.$$

By Property 3, $M_{X_2}X_2 = 0$. And OLS residuals are orthogonal to *all* regressors, in particular to X_2, so $M_{X_2}\hat{\varepsilon} = \hat{\varepsilon}$.

So $\widetilde{Y} = \widetilde{X}_1\hat{\beta}_1 + \hat{\varepsilon}$. Regressing $\widetilde{Y}$ on $\widetilde{X}_1$ gives

$$\hat{\beta}_1 = (\widetilde{X}_1'\widetilde{X}_1)^{-1}\widetilde{X}_1'\widetilde{Y}.$$

Proof: why Y may replace $\widetilde{Y}$

We claimed

$$\hat{\beta}_1 = (\widetilde{X}_1'\widetilde{X}_1)^{-1}\widetilde{X}_1'Y.$$

Start from the just-derived $(\widetilde{X}_1'\widetilde{X}_1)^{-1}\widetilde{X}_1'\widetilde{Y}$ and show $\widetilde{X}_1'Y = \widetilde{X}_1'\widetilde{Y}$.

Decompose $Y = P_{X_2}Y + M_{X_2}Y$. Then

$$\widetilde{X}_1'Y = (M_{X_2}X_1)'(P_{X_2}Y) \;+\; (M_{X_2}X_1)'(M_{X_2}Y).$$

The first term:

$$(M_{X_2}X_1)'P_{X_2}Y = X_1'M_{X_2}'P_{X_2}Y = X_1'M_{X_2}P_{X_2}Y.$$

But $M_{X_2}P_{X_2} = (I - P_{X_2})P_{X_2} = P_{X_2} - P_{X_2}^2 = 0.$

So the first term vanishes, and the second is exactly $\widetilde{X}_1'\widetilde{Y}$:

$$\widetilde{X}_1'Y = \widetilde{X}_1'\widetilde{Y}. \qquad \blacksquare$$

What that means in words

- $\widetilde{X}_1$ lives entirely in the orthogonal complement of Col(X_2).
- $P_{X_2}Y$ — the part of Y that X_2 explains — lives in Col(X_2).
- These two pieces are perpendicular: their inner product is zero.
- So whether we put Y or $\widetilde{Y}$ on the right, the part of Y that X_2 explains contributes **nothing** to $\hat{\beta}_1$. Only the orthogonal piece counts.

It does not matter whether we residualize Y as long as we residualize X_1.

FWL specialized to TWFE

Apply the FWL theorem (Equation 5) to the TWFE equation with

$$X_1 = D_{it} \text{ (treatment dummy)}, \qquad X_2 = [\,\text{unit FE} \mid \text{time FE}\,].$$

Two results from the FWL proof are used:

1. **FWL identity** (Equation 5): $\hat{\beta}^{TWFE} = (\widetilde{X}_1'\widetilde{X}_1)^{-1}\widetilde{X}_1'Y$. Written out as sums,

$$\hat{\beta}^{TWFE} = \frac{\sum_{i,t} \widetilde{D}_{it} Y_{it}}{\sum_{i,t} \widetilde{D}_{it}^2}.$$

   This is **exactly** the assertion Equation 1 that we used without proof in Section 6.

2. **"Y may replace $\widetilde{Y}$"** (the second equality in Equation 5): the numerator keeps the *raw* Y_{it} — we never have to compute $\widetilde{Y}_{it}$.

To be justified

What remains for Section 6 to be fully justified:

- (a) the closed form $\widetilde{D}_{it} = D_{it} - \bar{D}_g - \bar{D}_w + \bar{D}$ from Equation 2,
- (b) the cell-level numerator Equation 3,
- (c) the cell-level denominator Equation 4 (and *why it uses raw* $D_{g,w}$).

We tackle them next.

(a) The within-transform closed form

In a **balanced panel** (every i observed in every t), the OLS fit of D_{it} on unit + time dummies admits a closed form:

$$\hat{D}_{it} = \bar{D}_{g(i)} + \bar{D}_{w(t)} - \bar{D}.$$

Why? Unit-demeaning and time-demeaning each project onto a subspace of Col(X_2). In a balanced panel those two projections **commute**; applying both subtracts the unit mean *and* the time mean, but along the way subtracts the grand mean **twice** — so add it back once.

Subtracting from D_{it}:

$$\widetilde{D}_{it} = D_{it} - \hat{D}_{it} \;=\; D_{it} - \bar{D}_{g(i)} - \bar{D}_{w(t)} + \bar{D}.$$

That is exactly Equation 2. The "classic two-way within transform" is no longer a black box.

(b) Numerator: from observations to cells

Within each cell (g, w):

- D_{it} is constant: every unit in cohort g has the same treatment timeline.
- $\bar{D}_g$, $\bar{D}_w$, $\bar{D}$ are constants by construction.
- Hence $\widetilde{D}_{it} = \widetilde{D}_{g,w}$, a single number, for every (i, t) in the cell.

Cell size: $|\text{cell}\,(g, w)| = \underbrace{N n_g}_{\#\text{ units in } g} \cdot \underbrace{T t_w}_{\#\text{ periods in } w} = NT \cdot n_g t_w.$

Reorder the observation-level sum by cell and pull out the cell constant:

$$\sum_{i,t} \widetilde{D}_{it}\, Y_{it} = \sum_{g,w} \widetilde{D}_{g,w} \underbrace{\sum_{(i,t)\in(g,w)} Y_{it}}_{= NT \cdot n_g t_w \cdot \bar{Y}_{g,w}} = NT \sum_{g,w} (n_g t_w)\, \widetilde{D}_{g,w}\, \bar{Y}_{g,w}.$$

That confirms Equation 3 (up to the NT prefactor that will cancel against the denominator).

(c) Denominator: orthogonality trick

The observation-level denominator looks asymmetric: $\sum_{i,t} \widetilde{D}_{it}^2$ uses *residualized* $\widetilde{D}$, while Section 6's Equation 4 uses *raw* $D_{g,w}$. Why is that legitimate?

By OLS first-order conditions, $\hat{D}_{it}$ (the fitted value on X_2) is orthogonal to the residual $\widetilde{D}_{it}$:

$$\sum_{i,t} \widetilde{D}_{it}\, \hat{D}_{it} = 0.$$

Therefore, writing $D = \hat{D} + \widetilde{D}$,

$$\sum_{i,t} \widetilde{D}_{it}^2 = \sum_{i,t} \widetilde{D}_{it}\, (\widetilde{D}_{it} + \hat{D}_{it}) = \sum_{i,t} \widetilde{D}_{it}\, D_{it}.$$

Now apply the same cell-aggregation as for the numerator:

$$\sum_{i,t} \widetilde{D}_{it}\, D_{it} = NT \sum_{g,w} (n_g t_w)\, \widetilde{D}_{g,w}\, D_{g,w}.$$

This is exactly the cell-level denominator in Equation 4 — and it explains the swap $\widetilde{D} \to D$.

Closing the loop

Numerator ÷ denominator. The NT factor cancels:

$$\hat{\beta}^{TWFE} = \frac{\sum_{g,w}(n_g t_w)\,\widetilde{D}_{g,w}\,\bar{Y}_{g,w}}{\sum_{g,w}(n_g t_w)\,\widetilde{D}_{g,w}\,D_{g,w}}.$$

Everything Section 6 used without proof is now justified:

(i) Equation 1 — the FWL representation (projection-matrix proof).
(ii) Equation 2 — the two-way within transform (balanced-panel closed form).
(iii) Equation 3 — the cell-level numerator (constancy of $\widetilde{D}$ in cells).
(iv) Equation 4 — the cell-level denominator *with raw* $D_{g,w}$ (orthogonality identity).

Plugging in the six $\widetilde{D}_{g,w}$ values and the six $\bar{Y}_{g,w}$ from Section 6, the numerator regroups into $t_1 t_2 \cdot DiD_{\text{Clean}} + t_2 t_3 \cdot DiD_{\text{Forbidden}}$ and the t_2 factor cancels against the denominator, leaving

$$\hat{\beta}^{TWFE} = \frac{t_1}{t_1 + t_3}\,DiD_{\text{Clean}} + \frac{t_3}{t_1 + t_3}\,DiD_{\text{Forbidden}}. \qquad \blacksquare$$

Callaway-Sant'Anna: the building block

The philosophical reset

- TWFE starts with an *estimator* (OLS on a fixed equation) and hopes β maps to something meaningful.
- Callaway and Sant'Anna (2021) flip the order: **start with an estimand**, identify it under explicit assumptions, then construct an estimator.
- The estimand is allowed to depend on cohort and time — no homogeneity imposed.

Cohort notation

- $G_i \in \{2, 3, \ldots, T, \infty\}$ is unit i's first treatment period.
- $G_i = \infty$ means **never treated**. Call this set C.
- The building block:

$$ATT(g,t) = \mathbb{E}[Y_{it}(1) - Y_{it}(0) \,|\, G_i = g], \quad t \geq g.$$

That is the average treatment effect for cohort g at calendar time t. A whole (g,t) matrix of these.

PT using the never-treated as control

To identify $ATT(g,t)$ for $t \geq g$, compare cohort g across t and $g-1$ to the never-treated across the same two periods. PT:

$$\mathbb{E}\left[Y_{it}(0) - Y_{i,g-1}(0) \,\middle|\, G_i = g\right] = \mathbb{E}\left[Y_{it}(0) - Y_{i,g-1}(0) \,\middle|\, G_i = \infty\right].$$

- The baseline period is $g-1$, just before treatment.
- The drift in untreated potential outcomes is the same for cohort g as for the never-treated.

Identification

Starting from the definition and using $Y_{it} = Y_{it}(0)$ for the never-treated and $Y_{i,g-1} = Y_{i,g-1}(0)$ for cohort g (pre-treatment, assuming no anticipation):

$$ATT(g,t) = \mathbb{E}[Y_{it} \mid G = g] - \mathbb{E}[Y_{i,g-1} \mid G = g]$$

$$- \left(\mathbb{E}[Y_{it} \mid G = \infty] - \mathbb{E}[Y_{i,g-1} \mid G = \infty]\right).$$

A clean 2×2 DiD, but with cohort-and-time-specific objects on every side.

Why this kills the forbidden comparison

- The control set is *restricted* to $G = \infty$.
- Never-treated units have $Y(0)$ observed in every period — no contamination from any treatment effect.
- The forbidden comparison of Section 4 (using already-treated A as a control for B) is **structurally impossible** in this construction.

Not-yet-treated as control

Why an alternative control set?

Two practical reasons:

1. **No never-treated cohort exists** in many panels (e.g., every state eventually adopts the policy).
2. Even when C is non-empty, the never-treated may differ structurally; the **not-yet-treated** are on the same staggered-adoption trajectory and may have more credible parallel trends.

The not-yet-treated control set at calendar time t is

$$\{i : G_i > t\}.$$

These units are untreated at t and at $g-1$ (since $g-1 < t < G_i$), so their observed Y's equal $Y(0)$ in both periods.

PT with not-yet-treated controls

For each (g, t) with $t \geq g$, the parallel-trends statement becomes:

$$\mathbb{E}\left[Y_{it}(0) - Y_{i,g-1}(0) \,\middle|\, G_i = g\right] = \mathbb{E}\left[Y_{it}(0) - Y_{i,g-1}(0) \,\middle|\, G_i > t\right].$$

- The right-hand conditioning set is "any cohort not yet treated through period t."
- If desired, restrict further to a single later cohort g' with $g' > t$ — the same algebra applies.

Identification

Exactly mirroring the never-treated derivation:

$$ATT(g,t) = \mathbb{E}[Y_{it} \mid G = g] - \mathbb{E}[Y_{i,g-1} \mid G = g]$$
$$- \left(\mathbb{E}[Y_{it} \mid G > t] - \mathbb{E}[Y_{i,g-1} \mid G > t]\right).$$

The control side uses only $Y(0)$:

- At t: control units have $G_i > t$, so untreated $\Rightarrow Y_{it} = Y_{it}(0)$.
- At $g-1 < t$: control units are still untreated $\Rightarrow Y_{i,g-1} = Y_{i,g-1}(0)$.

No forbidden comparisons sneak in.

Choosing between the two control sets

	Never-treated	Not-yet-treated
Sample size	Often small	Larger (pools many cohorts)
PT credibility	Risk of structural differences	Drawn from the same adoption process
Existence	May be empty	Always non-empty (if at least one cohort untreated through t)
Efficiency	Lower	Higher in many designs

CS implementations let you pick; the underlying $ATT(g,t)$ is the same target.

Identifying assumptions, formalized

Enriched potential outcomes

Until now we wrote $Y_{it}(0), Y_{it}(1)$ as if treatment were a single binary state. In staggered designs, *when* you are treated matters. Define:

$$Y_{it}(g) = \text{potential outcome for unit } i \text{ at } t \text{ if first treated in cohort } g,$$

with $Y_{it}(\infty)$ the never-treated counterfactual.

No anticipation

Definition. $Y_{it}(g) = Y_{it}(\infty)$ for all $t < g$.

- Before being treated, your potential outcome is your never-treated one.
- Behavior in $t = g - 1$ does *not* depend on the upcoming treatment.

When it fails: agents know a tax is coming next year and shift consumption today. Then $Y_{i,g-1}(g) \neq Y_{i,g-1}(\infty)$.

Identifying $ATT(g,t)$ with $Y_{it}(g)$

Write the estimand in the enriched notation:

$$ATT(g,t) = \mathbb{E}[Y_{it}(g) \mid G_i = g] \; - \; \mathbb{E}[Y_{it}(\infty) \mid G_i = g], \quad t \geq g.$$

The first term is observed: for $G_i = g$, $Y_{it} = Y_{it}(g)$. The second is the counterfactual; we recover it in two steps.

Step 1 (PT in $Y(\infty)$). Apply PT between $G = g$ and $G = \infty$ groups:

$$\mathbb{E}[Y_{it}(\infty) \mid G = g] - \mathbb{E}[Y_{i,g-1}(\infty) \mid G = g] = \mathbb{E}[Y_{it}(\infty) - Y_{i,g-1}(\infty) \mid G = \infty].$$

Step 2 (never-treated observe $Y(\infty)$). For $G = \infty$, $Y = Y(\infty)$ in every period:

$$\mathbb{E}[Y_{it}(\infty) - Y_{i,g-1}(\infty) \mid G = \infty] = \mathbb{E}[Y_{it} - Y_{i,g-1} \mid G = \infty].$$

...where no anticipation bites

The one remaining unobservable is $\mathbb{E}[Y_{i,g-1}(\infty) \mid G = g]$: the *never-treated* outcome at $g-1$ for a cohort that *will* be treated at g.

Step 3 (no anticipation, $g-1 < g$). $Y_{i,g-1}(g) = Y_{i,g-1}(\infty)$, hence

$$\mathbb{E}[Y_{i,g-1}(\infty) \mid G = g] = \mathbb{E}[Y_{i,g-1}(g) \mid G = g] = \mathbb{E}[Y_{i,g-1} \mid G = g].$$

The observed pre-period outcome for cohort g **is** their never-treated counterfactual — because they have not yet been touched by treatment at $g-1$. Substituting Steps 1–3 into the estimand:

$$\boxed{ATT(g,t) = (\mathbb{E}[Y_{it} \mid G = g] - \mathbb{E}[Y_{i,g-1} \mid G = g]) - (\mathbb{E}[Y_{it} \mid G = \infty] - \mathbb{E}[Y_{i,g-1} \mid G = \infty]).}$$

Role summary. PT identifies the *drift* in untreated outcomes; no anticipation lets the *observed* pre-period outcome serve as the pre-treatment baseline.

What no-anticipation failure does

- Callaway-Sant'Anna uses $t = g - 1$ as the **clean pre-treatment baseline**.
- If anticipation inflates $Y_{i,g-1}$ for the treated, the DiD **subtracts an inflated baseline**:

$$\widehat{ATT}(g,t) = (Y_{it} - Y_{i,g-1}^{\text{inflated}}) - (\text{control change}),$$

  and we **underestimate** the true ATT — we have differenced out part of the treatment effect.
- Mitigation: include placebo leads ($e = -2, -3, \ldots$); shift the baseline earlier.

Random timing (a.k.a. mean independence of G on $Y(\infty)$)

Definition. For every period t and any cohorts g, g':

$$\mathbb{E}[Y_{it}(\infty) \mid G_i = g] = \mathbb{E}[Y_{it}(\infty) \mid G_i = g'] \,.$$

Note: no $g - 1$ here. This is a **statement about levels** of the never-treated potential outcome, period by period.

Levels vs. differences

- **Parallel trends** allows *level* differences across cohorts; it constrains only *differences*.
- **Random timing** rules out level differences entirely. Cohorts are exchangeable in their baseline $Y(\infty)$.
- Random timing is much *stronger*: it amounts to "G is as good as random."

If random timing holds, we do not need any DiD machinery — a cross-section in t suffices:

$$ATT(g,t) = \mathbb{E}[Y_{it} \mid G = g] - \mathbb{E}[Y_{it} \mid G = \infty].$$

Why we still need DiD

- Random timing almost never holds in observational data: early adopters, late adopters, and never-adopters typically differ systematically in baseline outcomes.
- Parallel trends concedes the level differences and asks only that the *gap* between groups would have stayed constant absent treatment.
- That weaker assumption is why DiD has been such a workhorse: it buys identification at a credible price.

Aggregation

Why aggregate?

- With G cohorts and T periods we can have dozens or hundreds of $ATT(g,t)$ estimates.
- Too granular to communicate; readers want summary numbers.
- CS propose principled aggregations that respect heterogeneity instead of averaging it away into a single β.

Event-study aggregation

Let $e = t - g =$ event time (periods since treatment). For each e, average across cohorts observable at that event time:

$$ATT(e) = \sum_g ATT(g, g+e) \cdot \mathbb{P}(G = g \,|\, g + e \leq T,\ G < \infty)\,.$$

- Weights are cohort shares **conditional on being observable** at horizon e.
- For $e = 0$: instantaneous effect. For $e = -1, -2, \ldots$: pre-trend / placebo.

The compositional-change pitfall

Event-time e requires observing $t = g + e$. When e is large, only early-treated cohorts contribute:

- $T = 10$, cohort $g = 3$ observable up to $e = 7$; cohort $g = 8$ observable only to $e = 2$.
- The right tail of the event-study plot is mechanically driven by early adopters.
- If early adopters' effects differ from late adopters', the apparent dynamics may just be **changing composition**, not real dynamics.

This is a real and common bug in published event studies.

Group-specific aggregation

For each cohort g, average post-treatment effects within that cohort first:

$$ATT(g) = \frac{1}{T - g + 1} \sum_{t=g}^{T} ATT(g, t).$$

Then aggregate across cohorts by baseline cohort share:

$$\overline{ATT} = \sum_{g<\infty} ATT(g) \cdot \mathbb{P}(G = g \mid G < \infty).$$

- The inner average is **within-cohort** — no composition change.
- The outer weights are fixed at the cohort sizes; the composition is stable.

Pre-trends and placebos

The same machinery for $e < 0$:

$$ATT(e) = \sum_g ATT(g, g+e) \cdot \mathbb{P}(G = g \mid \cdot), \quad e \in \{-1, -2, \dots\}.$$

- Each $ATT(g, g+e)$ for $e < 0$ should be zero under PT + no anticipation.
- Plot them with confidence intervals as a **pre-trend test**.
- A flat plot does not prove PT, but a sharply non-flat plot rejects it.

Wrap-up

The arc

1. **Estimand first.** ATT is what we want; PT is what we need.
2. **TWFE failure.** OLS with fixed effects mechanically converts already-treated units into "controls" whenever their treatment status is constant in a window.
3. **The Goodman-Bacon decomposition** quantifies that failure: forbidden weight $= t_3/(t_1 + t_3)$ in the two-cohort case.
4. **Callaway-Sant'Anna** $ATT(g, t)$ removes forbidden comparisons by construction. Never-treated or not-yet-treated are valid control sets.
5. **No anticipation** + **PT** are the identifying assumptions; **random timing** is the much stronger sibling we usually do not need.
6. **Aggregation** must respect composition — group-specific aggregation is the safe summary.

Implementation

- **R**: `did` package by Callaway and Sant'Anna — function `att_gt()` for the building block, `aggte()` for aggregations.
- **Stata**: `csdid` by Rios-Avila.
- **Python**: `differences` and `pyfixest` (the latter for TWFE diagnostics and Sun-Abraham).
- Always report a **Bacon decomposition** alongside any TWFE estimate as a diagnostic.

References

- Callaway, B. and Sant'Anna, P. H. C. (2021). "Difference-in-Differences with multiple time periods." *Journal of Econometrics*, 225(2): 200–230.
- Goodman-Bacon, A. (2021). "Difference-in-differences with variation in treatment timing." *Journal of Econometrics*, 225(2): 254–277.
- de Chaisemartin, C. and D'Haultfœuille, X. (2020). "Two-way fixed effects estimators with heterogeneous treatment effects." *American Economic Review*, 110(9): 2964–96.
- Sun, L. and Abraham, S. (2021). "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics*, 225(2): 175–199.
- Borusyak, K., Jaravel, X., and Spiess, J. (2024). "Revisiting event study designs: Robust and efficient estimation." *Review of Economic Studies*, 91(6): 3253–3285.