# Advanced Econometrics

## Lecture 11: Difference-in-Differences

Eduard Brüll

Fall 2025

Lecture 11

## Advanced Econometrics

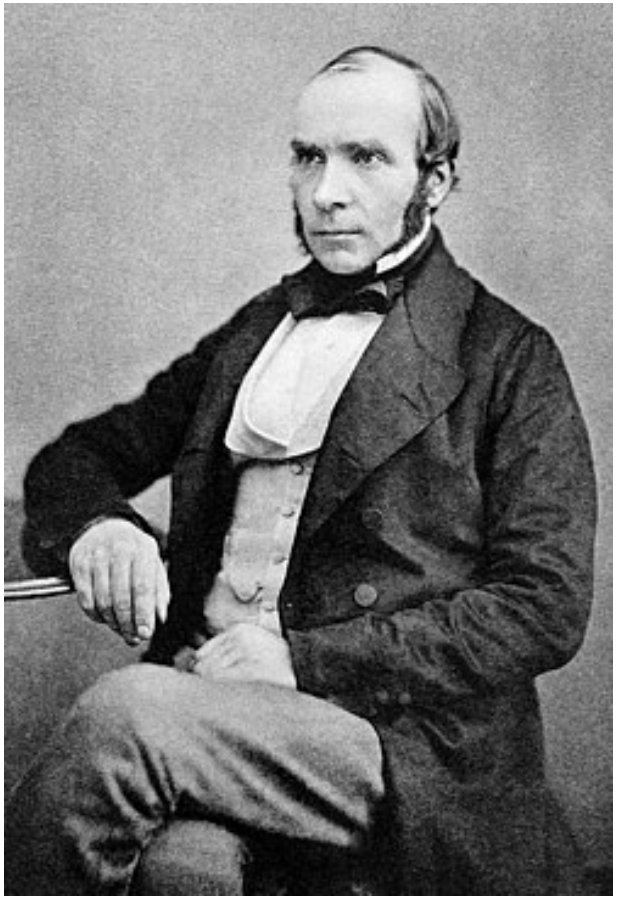


**Literature:** Cunningham (2021, Causal Inference: The Mixtape); Callaway & Sant'Anna (2021); Sun & Abraham (2021); de Chaisemartin & D'Haultfoeuille (2020)

# 11.1: John Snow and the Logic of Causal Inference

# John Snow and the Broad Street Pump

- **John Snow (1813–1858):** A London physician investigating cholera epidemics
- Challenged the prevailing miasma theory (disease from “bad air”)
- Observed stark differences in cholera rates across similar neighborhoods
- Linked outbreaks to **drinking water sources**, not air quality
- The Lambeth Water Company’s upstream intake created a **natural experiment**:
  Cleaner water reduced cholera cases



John Snow (1813–1858)

# Snow's Natural Experiment as a Proto-Difference-in-Differences

- In the early 1850s, two water companies served similar London neighborhoods:
  - **Southwark & Vauxhall:** drew water downstream: contaminated by sewage
  - **Lambeth:** moved intake pipes upstream; had cleaner, uncontaminated water after
- Snow compared cholera mortality rates before and after Lambeth's relocation:

| **Company** | **1849 (Pre)** | **1854 (Post)** |
|---|---|---|
| Southwark & Vauxhall (contaminated) | 135 | 147 |
| Lambeth (clean water) | 85 | 19 |

- Both groups were comparable in poverty, crowding, and sanitation: Only water quality differed

# Snow's Natural Experiment as a Proto-Difference-in-Differences

| Company | 1849 (Pre) | 1854 (Post) |
|---|---|---|
| Southwark & Vauxhall (contaminated) | 135 | 147 |
| Lambeth (clean water) | 85 | 19 |

- Simple DiD estimate:

$$(147 - 135) - (19 - 85) = 78$$

$\Rightarrow$ **78 fewer deaths per 10,000 households** among those receiving clean water.

- A century before randomized trials, Snow had executed one of history's first natural experiments.

John Snow's natural experiment compared mortality changes over time between two otherwise similar groups:

- **Treatment group:** Households supplied by Southwark & Vauxhall Water Company.
- **Control group:** Households supplied by Lambeth Water Company.
- **Intervention:** Lambeth moved its intake upstream (clean water) between 1849 and 1854.

## Snow's implicit causal logic

He compared how cholera mortality **changed over time** in the two groups:

$$(\text{Change in Lambeth}) - (\text{Change in Southwark \& Vauxhall})$$
$$\Rightarrow \text{Causal effect of clean water}$$

# 11.2: Classical Two-Way Fixed Effects DiD

# The Canonical 2×2 Difference-in-Differences Setup

**Groups and periods:**

- Two time-periods $t \in \{1, 2\}$
- A binary treatment $d \in \{0, 1\}$
- Two groups:
  - **Treatment group** $s$ (switchers from untreated to treated in $t = 2$)
  - **Control group** $n$ (never-treated)
- Treatment $d$ turns on only for group $s$ at $t = 2$

**Potential outcomes:**

$Y_{g,t}(d) =$ Potential Outcome of group $g$ at time $t$ given treatment $d \in \{0, 1\}$

**Observed outcomes:**

$$Y_{s,1} = Y_{s,1}(0), \quad Y_{s,2} = Y_{s,2}(1),$$
$$Y_{n,1} = Y_{n,1}(0), \quad Y_{n,2} = Y_{n,2}(0).$$

# Parallel Trends Assumpation

**Parallel Trends Assumpation:**
Absent treatment, treated and control would have followed the same average change

**For the 2×2 case:**

$$\mathbb{E}\big[Y_{s,2}(0) - Y_{s,1}(0)\big] = \mathbb{E}\big[Y_{n,2}(0) - Y_{n,1}(0)\big]$$

**Intuition:** We can use the control group's observed change to stand in for the treated group's **untreated** change.

# Why Counterfactual Parallel Trends Are Not Fully Testable

The **Parallel Trends Assumption (PTA)** involves **unobserved potential outcomes**:

$$\mathbb{E}[\boldsymbol{Y}_{s,2}(0) - \boldsymbol{Y}_{s,1}(0)] = \mathbb{E}[\boldsymbol{Y}_{n,2}(0) - \boldsymbol{Y}_{n,1}(0)]$$

But $\boldsymbol{Y}_{s,2}(0)$ is never observed for treated units after treatment!

- Hence, PTA can only be **partially assessed** (e.g., via pre-trends), not tested directly
- **SUTVA violations** (e.g., spillovers, interference) create the **same problem**:
  $\rightarrow$ They alter the unobserved counterfactual $\boldsymbol{Y}_{s,2}(0)$ or $\boldsymbol{Y}_{n,2}(0)$
- If control outcomes are affected by treatment exposure nearby, parallel pre-trends are not enough

# Stable Unit Treatment Value Assumption (SUTVA)

**What SUTVA means:**

1. **No interference:**
   Unit *i*'s outcome depends only on its own treatment:

$$Y_i(d) \text{ unaffected by } D_{-i}$$

2. **No hidden versions (consistency):**
   Treatment $d \in \{0, 1\}$ is well-defined and uniquely maps to $Y_i(d)$ (no multiple variants or intensities)

**Why it matters for DiD:**
Parallel trends concerns **counterfactual** changes:

$$\mathbb{E}[\boldsymbol{Y}_{s,2}(0) - \boldsymbol{Y}_{s,1}(0)] = \mathbb{E}[\boldsymbol{Y}_{n,2}(0) - \boldsymbol{Y}_{n,1}(0)]$$

If SUTVA fails (spillovers, General Equilibrium effects, contamination), $\boldsymbol{Y}_{s,2}(0)$ or $\boldsymbol{Y}_{n,2}(0)$ are distorted by others' treatment, so the control group no longer mimics the treated group's **untreated** change.

# Why 2×2 DiD recovers ATT

**DiD estimator (2×2):**

$$\text{DID} = \big(Y_{s,2} - Y_{s,1}\big) - \big(Y_{n,2} - Y_{n,1}\big).$$

**Rewrite observed outcomes as potential outcomes:**

$$\text{DID} = \big(Y_{s,2}(1) - Y_{s,1}(0)\big) - \big(Y_{n,2}(0) - Y_{n,1}(0)\big).$$

**Add & subtract the missing counterfactual $Y_{s,2}(0)$ for the treated at $t = 2$:**

$$\text{DID} = \underbrace{\big(Y_{s,2}(1) - Y_{s,2}(0)\big)}_{\text{treatment effect in } s \text{ at } t=2} + \Big[\big(Y_{s,2}(0) - Y_{s,1}(0)\big) - \big(Y_{n,2}(0) - Y_{n,1}(0)\big)\Big].$$

**Take expectations and impose Parallel Trends Assumption:**

$$\mathbb{E}[\text{DID}] = \mathbb{E}\big[Y_{s,2}(1) - Y_{s,2}(0)\big] + \underbrace{\mathbb{E}\big[Y_{s,2}(0) - Y_{s,1}(0)\big] - \mathbb{E}\big[Y_{n,2}(0) - Y_{n,1}(0)\big]}_{= 0 \text{ by Parallel Trends Assumption}}$$

$$= \mathbb{E}\big[Y_{s,2}(1) - Y_{s,2}(0)\big] = \text{ATT}_{s,2}.$$

## Interpretation

In a clean $2 \times 2$ switcher design, the DiD equals the average treatment effect on the treated for the switching group in the post period, provided the untreated trends are parallel across *s* and *n*.

# DiD as two-way fixed effects (TWFE)

To estimate effects of a treatment/policy on an outcome, researchers often run two-way fixed effects (TWFE) regressions:

$$Y_{g,t} = \alpha_g + \gamma_t + \beta_{\text{fe}} D_{g,t} + \varepsilon_{g,t}.$$

where:

- $\alpha_g$: group (e.g. county) fixed effect
- $\gamma_t$: time (e.g. year) fixed effect
- $D_{g,t}$: realized treatment indicator ($= 1$ if treated in group $g$ at time $t$, $0$ otherwise)

## Most DiDs were estimated via TWFE

**26 of the 100 most cited AER papers (2015–2019)** estimate TWFE (de Chaisemartin & D'Haultfoeuille, 2021).

Also widely used in political science, sociology, and environmental sciences.

# 11.3: Event Studies and Dynamic Effects

# Motivation: Beyond a Single Treatment Effect

The basic DiD gives a single average post-treatment effect. But often we want to know:

- **When** effects appear, do they build up or fade out? Are they constant?
- Whether there are **differential trends** before treatment.
- How long effects persist after exposure.

## Idea: Unpack the DiD by event time

Define for each unit $i$:

$$G_i = \text{first period when } D_{i,t} = 1, \qquad \text{Event time: } k = t - G_i.$$

We then interact the treatment indicator with dummies for event time:

$$D_{i,t} \times \mathbf{1}\{t - G_i = k\}.$$

Each interaction measures the treatment effect $k$ periods after (or before) treatment.

# The Event-Study Regression

A conventional specification is:

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{k \neq -1} \beta_k \big( D_i \times \mathbf{1}\{t - G_i = k\} \big) + \varepsilon_{i,t}.$$

- $\alpha_i$: unit fixed effects remove permanent differences.
- $\lambda_t$: time fixed effects absorb aggregate shocks.
- $D_i \times \mathbf{1}\{t - G_i = k\}$: interaction capturing the treatment status at event time $k$.
- The omitted dummy ($k = -1$) is the **reference period**.

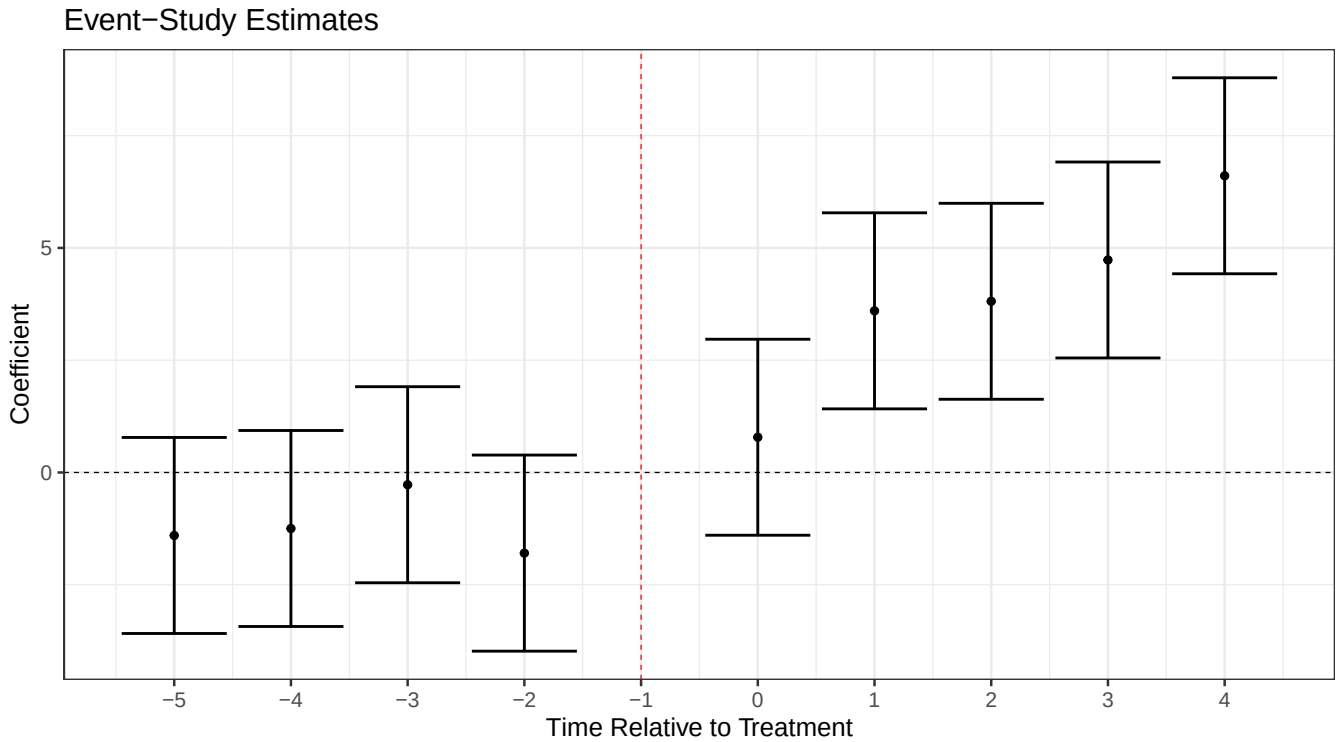
## Interpretation

$$\beta_k = \begin{cases} \text{Effect } k \text{ periods after treatment}, & k \geq 0, \\ \text{Difference } k \text{ periods before treatment}, & k < 0. \end{cases}$$

$\Rightarrow$ Pre-treatment coefficients ($k < 0$) test the **observable part of the parallel-trends assumption.**

# Visualizing Dynamic Effects

Plotting $\widehat{\beta}_k$ with confidence intervals shows how outcomes evolve relative to the last pre-treatment period ($k = -1$):




- **Leads ($k < 0$):** should hover around zero if pre-treatment trends are parallel.
- **Lags ($k \geq 0$):** show the treatment dynamics,do effects grow or fade?

# When Event-Study Plots Mislead

Even though event-studies are intuitive, they can be misleading:

- **Endogenous timing:** units may adopt treatment because of pre-existing outcome trends. $\Rightarrow$ apparent "effects" before treatment may reflect selection.
- **Selective composition:** for large $k$, only early adopters remain observed, so late effects mix dynamics with who is still in sample.
- **Staggered adoption under TWFE:** already-treated units serve as controls for later-treated ones, so $\widehat{\beta}_k$ combines different causal horizons with possibly negative weights.

# 11.4: The Problems with TWFE

# The Problem with TWFE

## Core intuition

TWFE assumes one common treatment effect for everyone, forever. When effects differ by group or timing, this average can get distorted.

**What goes wrong:**

- With staggered adoption, already-treated groups act as *"controls"*.
  $\Rightarrow$ Some comparisons can get **negative or weird weights.**
  $\Rightarrow$ The overall estimate can even flip sign.

**Next:** We will see this in a decomposition of $\widehat{\beta}_{FE}$ into all possible $2\times2$ comparisons based on de Chaisemartin & D'Haultfoeuille (2021)

# Sketch of the Decomposition of $\widehat{\beta}_{FE}$

Start from the TWFE model:

$$Y_{g,t} = \alpha_g + \gamma_t + \beta_{FE} D_{g,t} + u_{g,t}$$

By the Frisch–Waugh–Lovell theorem:

$$\widehat{\beta}_{FE} = \frac{\sum_{g,t} \tilde{D}_{g,t} \tilde{Y}_{g,t}}{\sum_{g,t} \tilde{D}_{g,t}^2}, \qquad \tilde{D}_{g,t} = D_{g,t} - \widehat{D}_{g,t}$$

**Step 1:** Express residualized treatment $\tilde{D}_{g,t}$ as deviations from group and time means:

$$\tilde{D}_{g,t} = D_{g,t} - \bar{D}_{g\cdot} - \bar{D}_{\cdot t} + \bar{D}$$

**Step 2:** Insert this into the covariance term $\sum \tilde{D}_{g,t} \tilde{Y}_{g,t}$ and rearrange by pairs of treatment changes across $(g, t)$

## Main Idea:

Each nonzero $\tilde{D}_{g,t}$ comes from a group that **changes treatment status** relative to other groups or periods
$\Rightarrow$ The estimator can be written as a weighted sum of all possible $2 \times 2$ DiD contrasts.

# How FWL builds the residual $\tilde{D}_{g,t}$

**Goal:** Remove what $D_{g,t}$ shares with group and time patterns

**Step 1: Start from the full treatment indicator**
Each cell starts with its observed treatment $D_{g,t}$

**Step 2: Subtract the group mean $\bar{D}_{g\cdot}$**
Removes how "treated on average" this group is over time
$\Rightarrow$ Controls for permanent differences between groups

**Step 3: Subtract the time mean $\bar{D}_{\cdot t}$**
Removes how many groups are treated in this period overall
$\Rightarrow$ Controls for common time shocks or trends

**Step 4: Add back the grand mean $\bar{D}$**
We subtracted that global average twice, once in each step, so we put one copy back to keep the overall mean at zero

$$\boxed{\tilde{D}_{g,t} = D_{g,t} - \bar{D}_{g\cdot} - \bar{D}_{\cdot t} + \bar{D}}$$

# Potential Outcomes in the Decomposition

Each cell $(g, t)$ has potential outcomes:

$$Y_{g,t}(1), \quad Y_{g,t}(0)$$

and a treatment effect

$$TE_{g,t} = Y_{g,t}(1) - Y_{g,t}(0)$$

Observed outcomes:

$$Y_{g,t} = Y_{g,t}(0) + D_{g,t}\, TE_{g,t}$$

Residualized version:

$$\tilde{Y}_{g,t} = \tilde{Y}_{g,t}(0) + \tilde{D}_{g,t}\, TE_{g,t}$$

Plug into FWL:

$$\sum_{g,t} \tilde{D}_{g,t}\tilde{Y}_{g,t} = \sum_{g,t} \tilde{D}_{g,t}\tilde{Y}_{g,t}(0) + \sum_{g,t} \tilde{D}^2_{g,t}\, TE_{g,t}$$

# Why the first term drops out

Under the parallel trends assumption:

$$\mathbb{E}[Y_{g,t}(0) \mid g,t] = \alpha_g + \gamma_t \quad \Rightarrow \quad \tilde{Y}_{g,t}(0) \text{ has zero covariance with } \tilde{D}_{g,t}$$

$$\mathbb{E}[\tilde{D}_{g,t}\tilde{Y}_{g,t}(0)] = 0$$

**Intuition:**

- The untreated potential outcome only reflects group and time patterns
- Those same patterns were already "partialed out" of $D_{g,t}$ by the FWL residualization
- $\Rightarrow$ Once group and time means are removed, there's no systematic relation left between $\tilde{D}_{g,t}$ and $\tilde{Y}_{g,t}(0)$

**Hence:**

$$\mathbb{E}[\widehat{\beta}_{FE}] = \mathbb{E}\left[\frac{\sum_{g,t}\tilde{D}^2_{g,t}\,TE_{g,t}}{\sum_{g,t}\tilde{D}^2_{g,t}}\right]$$

# How Residualized Treatment Creates Contrasts

$$\tilde{D}_{g,t} = D_{g,t} - \bar{D}_{g\cdot} - \bar{D}_{\cdot t} + \bar{D}$$

- Each $\tilde{D}_{g,t}$ is a **deviation** from both its group's and the period's average treatment.
- A large positive $\tilde{D}_{g,t}$ means "newly treated" relative to others.
- A large negative $\tilde{D}_{g,t}$ means "already treated" when others are not.

## Intuition

Residualization turns $D_{g,t}$ into a measure of how much that cell's treatment status **changes relative to the comparison groups and periods**.

$$\mathbb{E}[\widehat{\beta}_{FE}] = \mathbb{E}\left[\frac{\sum_{g,t} \tilde{D}^2_{g,t} \, TE_{g,t}}{\sum_{g,t} \tilde{D}^2_{g,t}}\right]$$

is therefore an average of **treatment effects weighted by these relative changes**.

# Each Residual Implies a 2×2 DiD Comparison

**For any two groups and periods** $(g, g', t, t')$:

$$\Delta D_{g,t;t'} = (D_{g,t} - D_{g,t'}) - (D_{g',t} - D_{g',t'})$$

- If one group changes treatment while the other does not, $\Delta D_{g,t;t'} = \pm 1$.
- The corresponding outcome contrast

  $$\Delta Y_{g,t;t'} = (Y_{g,t} - Y_{g,t'}) - (Y_{g',t} - Y_{g',t'})$$

  is a standard $2 \times 2$ Difference-in-Differences.
- Summing all such comparisons with $\tilde{D}_{g,t}$ weights reproduces the TWFE estimator:

  $$\sum_{g,t} \tilde{D}_{g,t} \tilde{Y}_{g,t} = \sum_{(g,t)} \omega_{g,t} \, \Delta Y_{g,t;t'} \, .$$

**Thus:** TWFE aggregates all possible $2 \times 2$ DiD contrasts in the data.

# The Weighted-Average Representation

Collecting all $2 \times 2$ comparisons yields:

$$\mathbb{E}[\widehat{\beta}_{FE}] = \mathbb{E}\left[\sum_{(g,t):D_{g,t}\neq 0} W_{g,t}\, TE_{g,t}\right], \quad W_{g,t} = \frac{\tilde{D}_{g,t}^2}{\sum_{g',t'} \tilde{D}_{g',t'}^2}, \quad \sum W_{g,t} = 1.$$

- Each $W_{g,t}$ reflects how strongly that cell's treatment pattern deviates from group and period averages.
- If some $\tilde{D}_{g,t}$ are negative (already-treated groups), their cells receive **negative weights**.
- These negative weights can distort the overall estimate when effects differ by timing.

## Takeaway

$\widehat{\beta}_{FE}$ is a weighted average of all cell-level treatment effects. Weights depend on treatment-timing heterogeneity and can even be negative.

# What TWFE actually estimates

$$\mathbb{E}[\widehat{\beta}_{FE}] = \mathbb{E}\left[\sum_{(g,t):D_{g,t}\neq 0} W_{g,t}\, TE_{g,t}\right], \quad W_{g,t} = \frac{\tilde{D}^2_{g,t}}{\sum_{g',t'} \tilde{D}^2_{g',t'}}, \quad \sum W_{g,t} = 1.$$

- **Combines all** $2 \times 2$ **comparisons** where treatment changes for one group vs. another.
- **Comparison types:**
  1. Switchers vs. never/not-yet treated: What we want
  2. Early vs. late switchers: Mixes timing effects
  3. Switchers vs. already-treated $\Rightarrow$ possible **negative weights**.
- Negatives arise when already-treated units act as (implicit) controls.

## Why it matters

- With heterogeneous effects, TWFE's weighted mean can be **biased or even sign-reversed**.
- $\widehat{\beta}_{FE}$ rarely reflects a single "true" effect under staggered adoption.

# What Does TWFE Weight?

We decomposed $\beta_{\text{TWFE}}$ into ATTs. Now we introduce another useful decomposition:

**Goodman-Bacon (GB):**

- **Unit:** Weighted avg. of **2×2 DiD** designs.
- **Weights:** Non-negative, sum to 1 $\Rightarrow$ which timing contrasts drive TWFE.
- **Pitfall:** Some designs (e.g., later-as-control) can flip sign with dynamics.

**de Chaisemartin & D'Haultfoeuille (dCdH):**

- **Unit:** Weighted sum of $\mathbf{ATT_{g,t}}$ (cohort $\times$ time).
- **Weights:** Can be negative $\Rightarrow$ $\hat{\beta}_{\text{TWFE}}$ not a convex avg.
- **Use:** Shows which $g, t$ cells TWFE (de-)emphasizes.

## Key takeaway

**GB:** Which designs drive TWFE?
**dCdH:** Which $ATT_{g,t}$ does TWFE (even negatively) weight?

# $K^2$ distinct DiDs

Let there be $K$ different **treatment timing groups** and one never-treated group $U$.

- For each ordered pair of timing groups $(j, b)$ we can form a canonical $2 \times 2$ DiD where $j$ is treated and $b$ is the comparison.
- This gives $K^2$ **distinct** $2 \times 2$ **DiDs** (including comparisons to the never-treated group).

**Example:** Three timing groups $a, b, c$ and one untreated group $U$.

| | | |
|---|---|---|
| $a$ to $b$ | $b$ to $a$ | $c$ to $a$ |
| $a$ to $c$ | $b$ to $c$ | $c$ to $b$ |
| $a$ to $U$ | $b$ to $U$ | $c$ to $U$ |

# Example DiD with two timing groups

To keep intuition simple, we now focus on only two timing groups:

- an **early** group $k$ treated at time $t_k^*$,
- a **late** group $l$ treated at time $t_l^*$,
- and a never-treated group $U$.
- $k, l$ are defined by the **time of treatment adoption** $(t_k^*, t_l^*)$, with $k$ treated earlier than $l$.
- $\bar{D}_k$: share of periods in which group $k$ is treated.
- $\widehat{\delta}_{jb}^{2\times 2}$: canonical $2 \times 2$ DiD estimator comparing treatment group $j$ to comparison group $b$.
- In the Bacon decomposition, the TWFE estimate is a weighted average of these $\widehat{\delta}_{jb}^{2\times 2}$ terms, with weights depending on variation in treatment timing and group sizes.

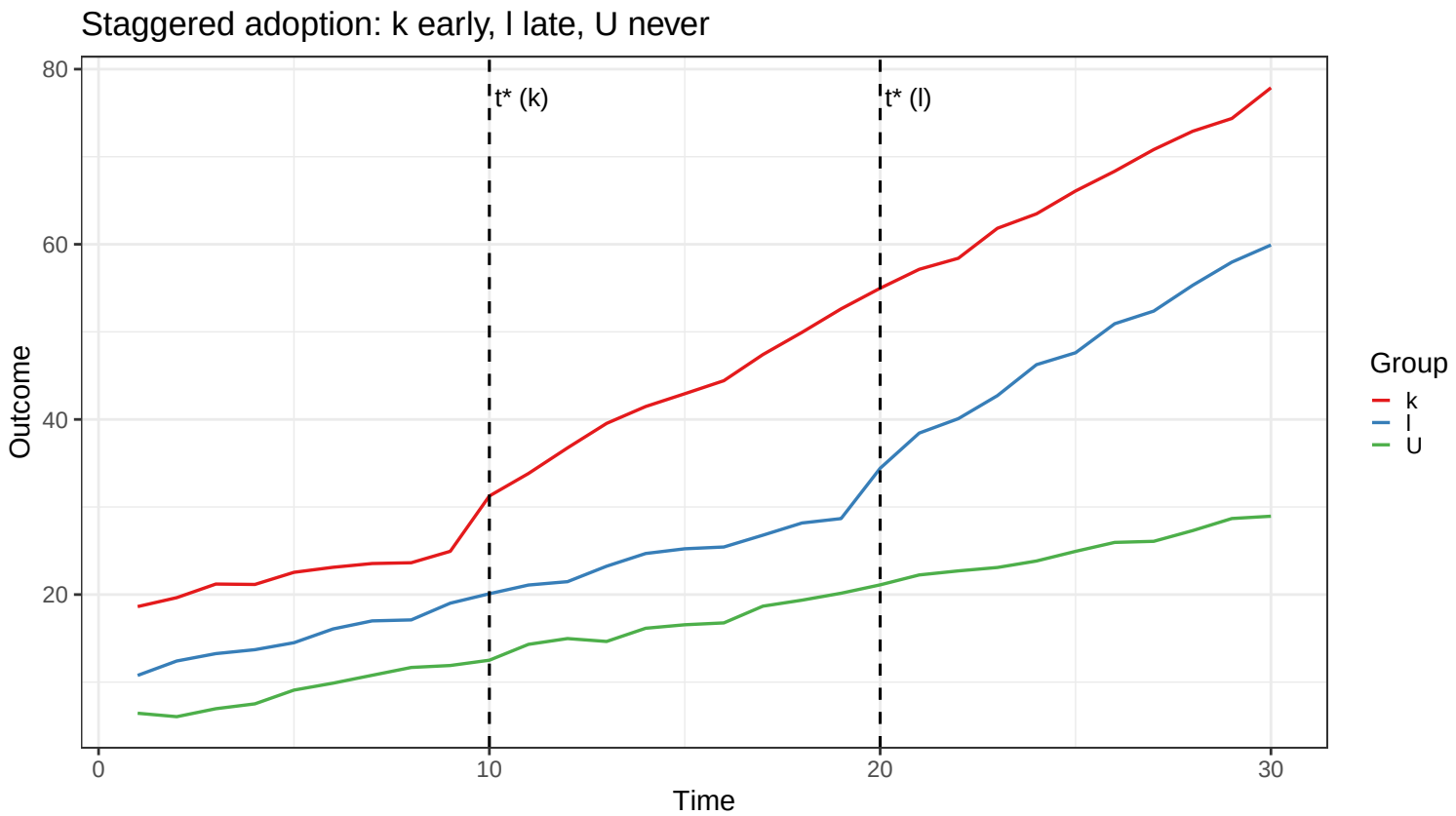
# $2 \times 2$ DiDs with two timing groups

Each treated group can be compared to the others to form a canonical $2 \times 2$ DiD:

| Treated | Comparison | Interpretation |
|---|---|---|
| *k* | *l* | Early vs. late group |
| *l* | *k* | Late vs. early group |
| *k* | *U* | Early vs. never-treated |
| *l* | *U* | Late vs. never-treated |

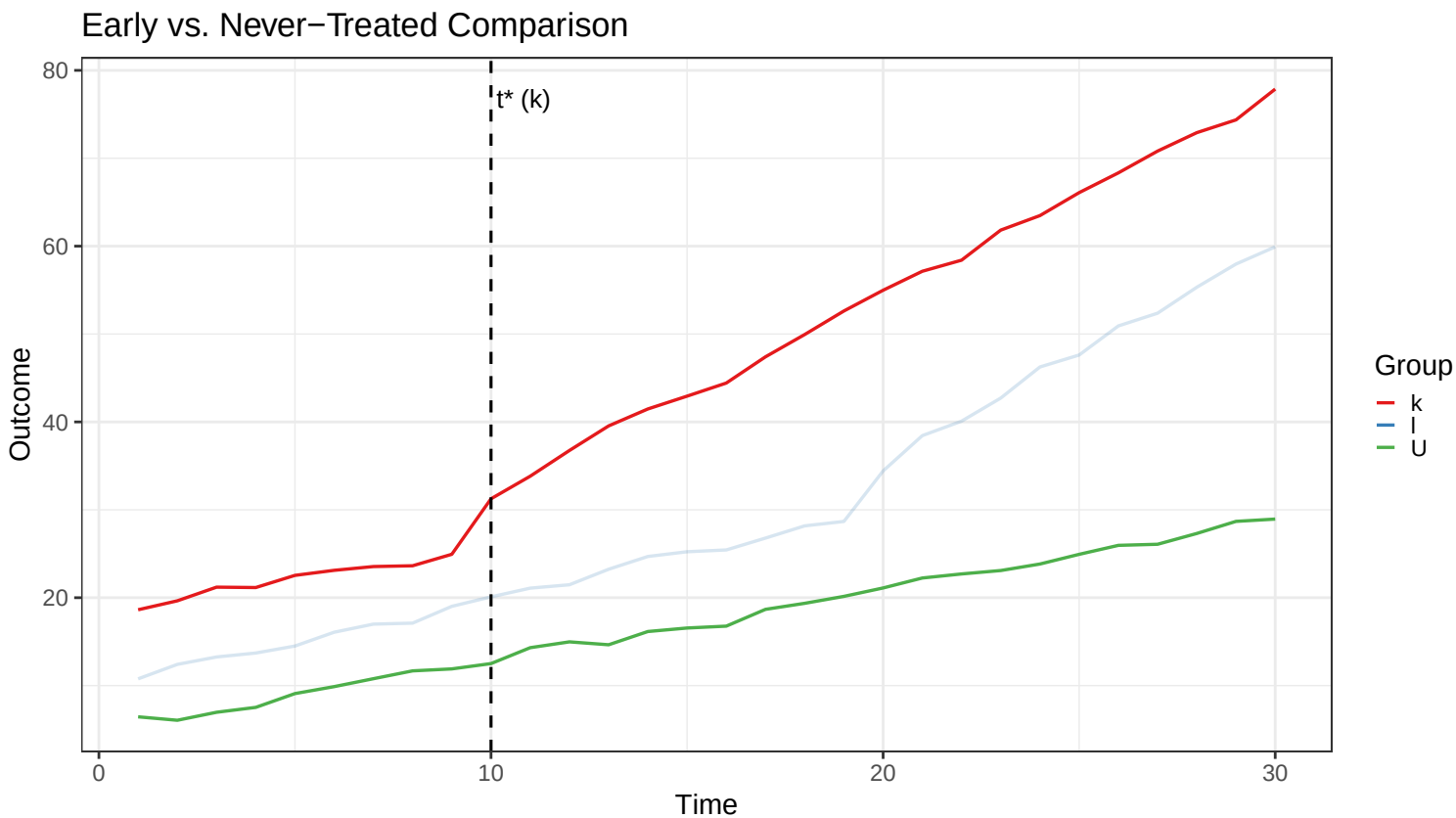
This yields **four distinct ordered $2 \times 2$ DiD comparisons**

# Illustration: Our Example



# Early Group vs. Never-Treated Group

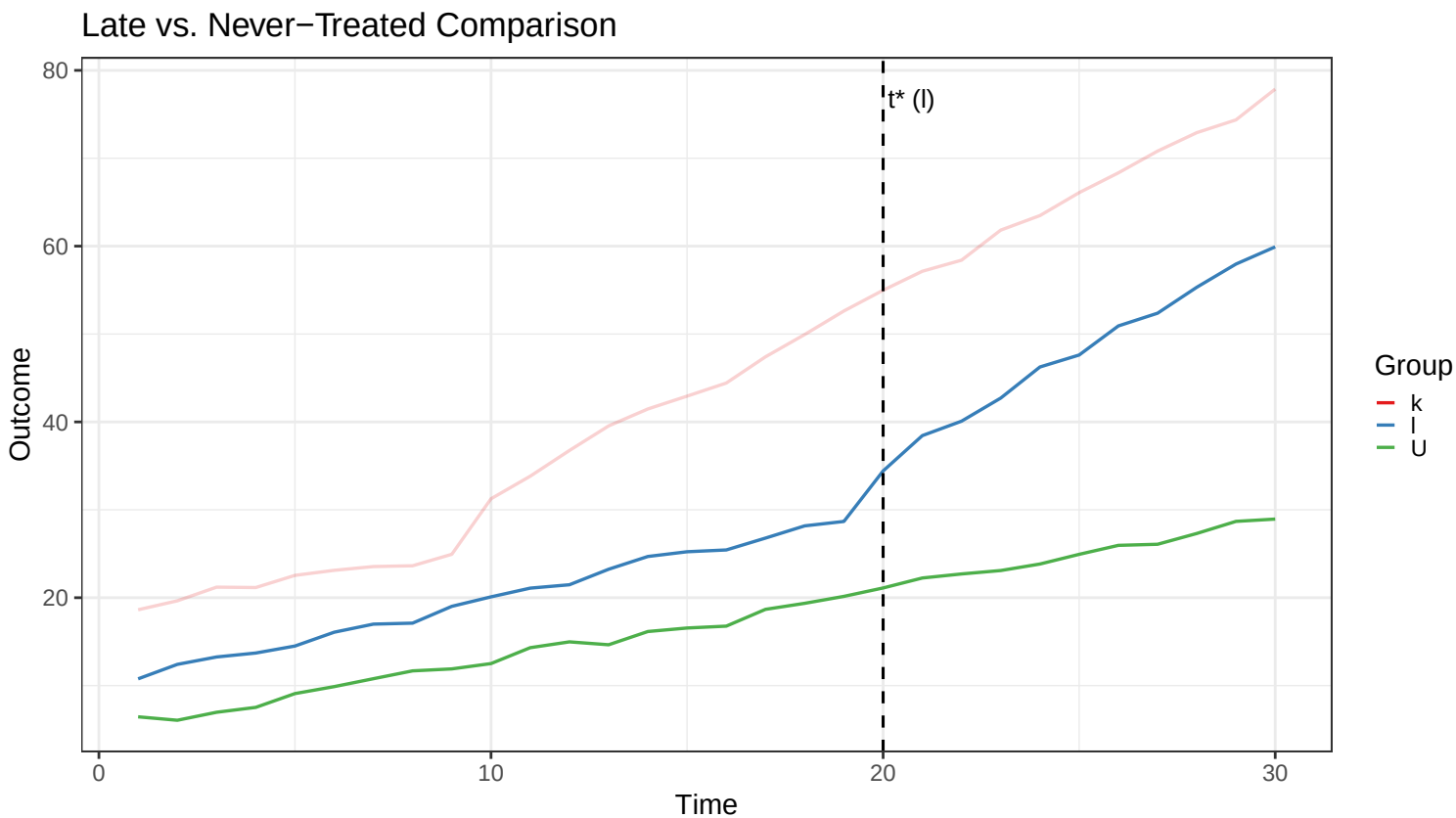
$$\widehat{\delta}_{kU}^{2\times 2} = \left(\bar{y}_k^{\text{post}(k)} - \bar{y}_k^{\text{pre}(k)}\right) - \left(\bar{y}_U^{\text{post}(k)} - \bar{y}_U^{\text{pre}(k)}\right)$$




A. Early group $k$ compared to never-treated $U$ around $t_k^*$.

# Late Group vs. Never-Treated Group

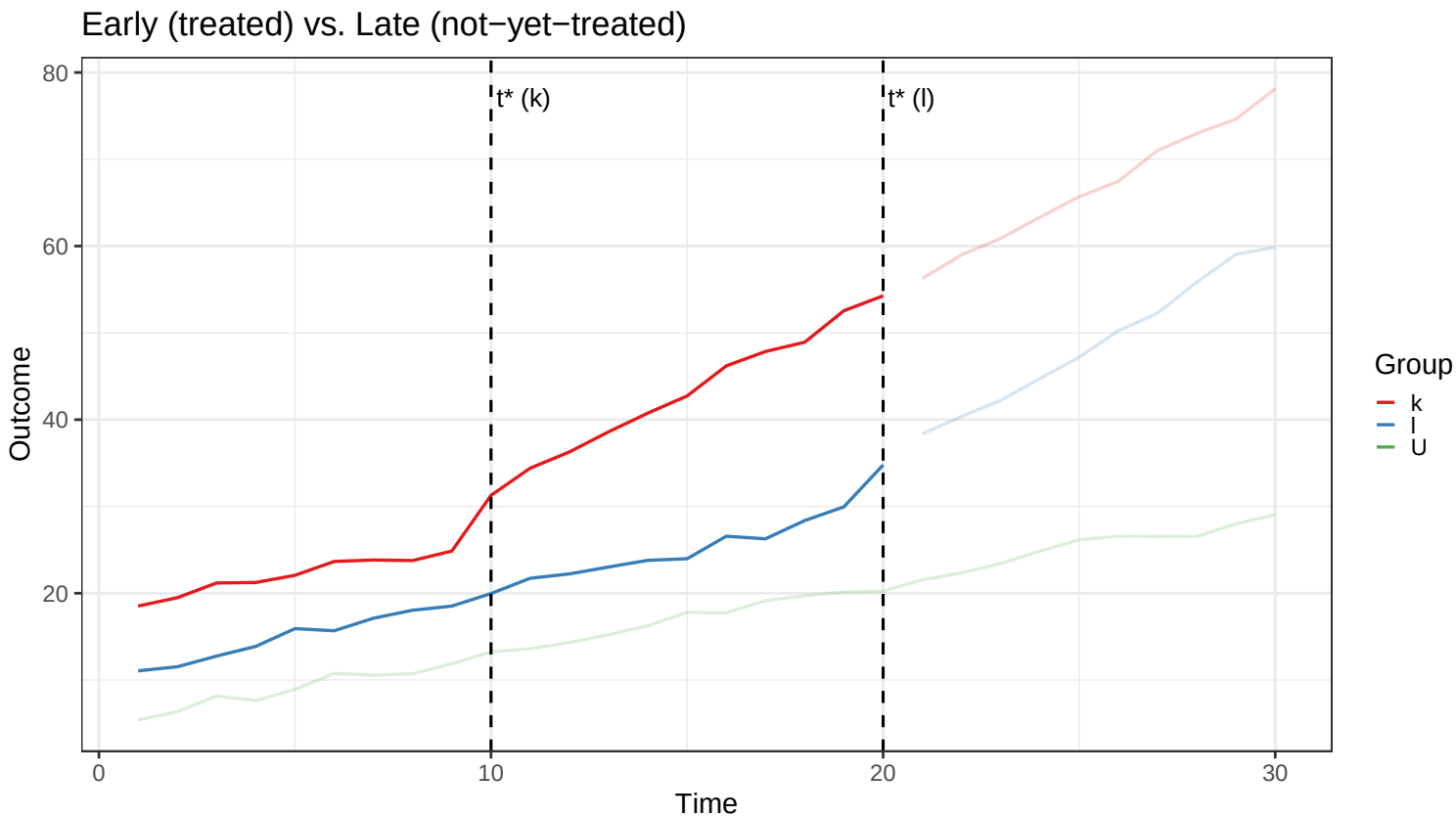
$$\widehat{\delta}_{lU}^{2\times 2} = \left(\bar{y}_l^{\text{post}(l)} - \bar{y}_l^{\text{pre}(l)}\right) - \left(\bar{y}_U^{\text{post}(l)} - \bar{y}_U^{\text{pre}(l)}\right)$$




B. Late group *l* compared to never-treated *U* around $t_l^*$.

# Early Group vs. Late Group, before $t_l^*$

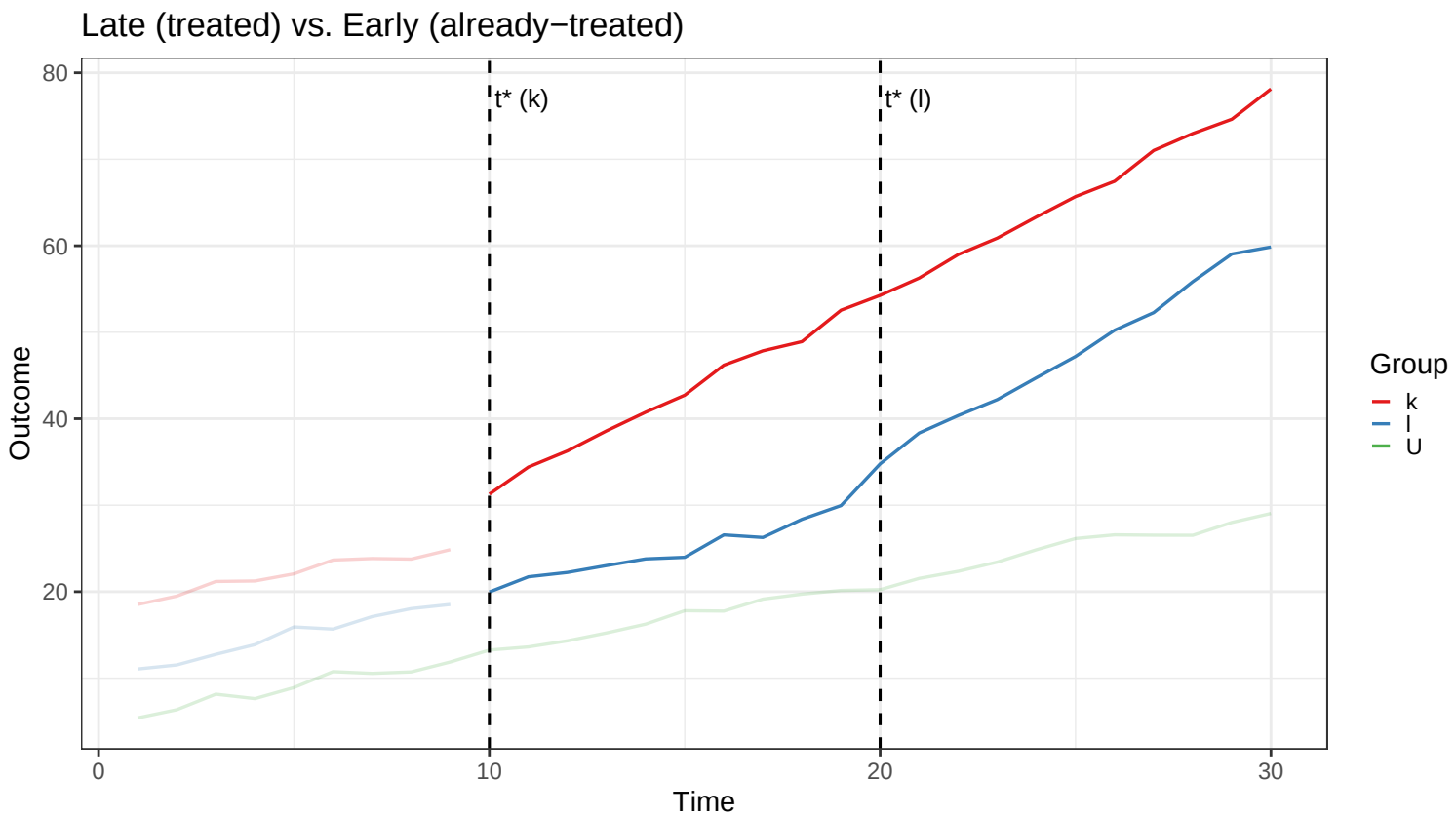
$$\delta_{kl}^{2\times 2,k} = \left(\bar{y}_k^{MID(k,l)} - \bar{y}_k^{PRE(k)}\right) - \left(\bar{y}_l^{MID(k,l)} - \bar{y}_l^{PRE(k)}\right)$$




C. Early group *k* treated, late group *l* not-yet treated; comparison uses only periods before $t_l^*$.

# Late Group vs. Early Group, after $t_k^*$

$$\delta_{lk}^{2\times 2,l} = \left(\bar{y}_l^{POST(k,l)} - \bar{y}_l^{MID(k,l)}\right) - \left(\bar{y}_k^{POST(k,l)} - \bar{y}_k^{MID(k,l)}\right)$$




D. Late group *l* treated, early group *k* already treated; comparison uses only periods after $t_k^*$.

# Bacon decomposition

Consider the TWFE regression estimated at the group level:

$$Y_{ist} = \beta_0 + \delta D_{ist} + \tau_t + \sigma_s + \varepsilon_{ist}$$

With two treated groups ($k, l$) and one never-treated group ($U$), the TWFE estimator of $\widehat{\delta}$ can be written as a weighted average of four distinct $2 \times 2$ DiDs:

$$\widehat{\delta}^{TWFE} = \sum_{k \neq U} s_{kU} \widehat{\delta}_{kU}^{2\times 2} + \sum_{k \neq U} \sum_{l > k} s_{kl} \left[ \mu_{kl} \, \widehat{\delta}_{kl}^{2\times 2,k} + (1 - \mu_{kl}) \, \widehat{\delta}_{lk}^{2\times 2,l} \right].$$

- $s_{kU}, s_{kl}$: non-negative weights that sum to 1.
- $\widehat{\delta}_{kU}^{2\times 2}$ and $\widehat{\delta}_{lU}^{2\times 2}$: DiDs of early/late groups vs. never-treated.
- $\widehat{\delta}_{kl}^{2\times 2,k}$: early vs. late using only *pre* $t_l^*$ periods.
- $\widehat{\delta}_{lk}^{2\times 2,l}$: late vs. early using only *post* $t_k^*$ periods.

# How the Weights Are Determined

Weights in the Goodman–Bacon decomposition depend on two ingredients:

- **Group size:** larger groups $\Rightarrow$ more influence on $\widehat{\delta}^{TWFE}$.
- **Treatment variance:** groups treated for about half of the sample period get the biggest weight.

$$s_{kU} \propto n_k n_U \overline{D}_k(1-\overline{D}_k), \quad s_{kl} \propto n_k n_l (\overline{D}_k - \overline{D}_l)[1-(\overline{D}_k - \overline{D}_l)]$$

## Intuition

Weights are largest where there is most variation in treatment timing, that is, when groups switch from untreated to treated in the middle of the panel.

# Why Some Comparisons Are Dangerous

Not all $2 \times 2$ comparisons identify the same causal contrast:

- Comparisons with never-treated groups are typically valid (if parallel trends hold).
- Comparisons where **later-treated groups serve as controls** for already-treated groups can pick up treatment dynamics, not counterfactual trends.
- These comparisons can even flip the sign of $\widehat{\delta}^{TWFE}$ if effects grow or fade over time.

## What the weights teach us

Weights reveal **which timing contrasts drive your estimate.** If much weight falls on "treated vs. treated" comparisons, your TWFE estimate likely mixes treatment effects at different horizons.

# Interactive Tools for TWFE Problems

Two excellent interactive Shiny apps from *"Causal Inference: The Mixtape"* allow you to explore the issues we discussed:

- **Goodman–Bacon Decomposition App**

  ▸ Bacon Decomposition

  Visualizes how TWFE mixes different 2×2 DiDs and how these affect the overall TWFE parameter

- **Event-Study / TWFE vs. Modern Estimators App**

  ▸ Event-Study Simulation

  Compares TWFE event-study estimates to modern alternatives in the same setup as above

## Recommended Use

Experiment with treatment timing, heterogeneity, and dynamic effects to see how TWFE behaves—and how modern estimators fix these issues.

# From Decomposition to Better Practice

- Use the Bacon decomposition to **diagnose** what drives your estimate:
  - Are most weights on valid (treated vs. untreated) DiDs?
  - Or on questionable (treated vs. treated) ones?
- If the latter, move to estimators that respect staggered timing:
  - Stacking or **Sun & Abraham (2020)**, **Callaway & Sant'Anna (2021)**, **de Chaisemartin and D'Haultfoeuille (2020)**.

## Takeaway

The Goodman–Bacon decomposition helps us understand *why* TWFE can fail and points the way to designs that avoid those pitfalls.

# 11.5: Modern Solutions to TWFE Bias

# Core Idea: Group-Time Average Treatment Effects

**Problem:** TWFE pools all treated units together and implicitly assumes a **common treatment effect**. Under staggered adoption, this mixes effects across:

- different **cohorts** $g$ (time of first treatment)
- different **relative times** $k = t - G_i$

## Solution: Estimate "clean" effects for each cohort and period

$$\text{ATT}(g,t) = \mathbb{E}\big[Y_{g,t}(1) - Y_{g,t}(0) \,\big|\, G_i = g\big]$$

- Compare cohort $g$ only to units that are **untreated** at time $t$.
- Never uses already-treated cohorts as controls.

- **All modern estimators** recover $\text{ATT}(g,t)$ in some way.
- Differences arise in how they aggregate $\text{ATT}(g,t)$ or handle event-time dummies.

# Callaway & Sant'Anna (2021): Core Idea

**Goal:** Estimate treatment effects that vary by **cohort** and by **time** since treatment.

## Core Idea

For each treatment cohort $g$ and period $t$:

$$\text{ATT}(g,t) = (\text{change in } g \text{ around } g) - (\text{change in controls not yet treated at } t)$$

Controls: only units that are **never-treated or not-yet-treated**.

- **Design-based**: builds treatment effects from simple 2×2 DID comparisons.
- **Cohort-specific parallel trends**: only requires PT for each cohort vs. its valid controls.
- **Modular**: $\text{ATT}(g,t)$ can be estimated via IPW, regression, or doubly robust methods.

# Callaway & Sant'Anna (2021): Details

**Group-time ATT:**

$$\text{ATT}(g,t) = \left(\bar{Y}_{g,t} - \bar{Y}_{g,g-1}\right) - \left(\bar{Y}_{C(g,t),t} - \bar{Y}_{C(g,t),g-1}\right)$$

where $C(g,t)$ are units untreated at period $t$.

**Identification:**
Parallel trends need to hold **within cohort**:

$$\mathbb{E}[Y_{g,t}(0) - Y_{g,g-1}(0)] = \mathbb{E}[Y_{C(g,t),t}(0) - Y_{C(g,t),g-1}(0)]$$

**Aggregation into policy parameters**
Overall effect:

$$\text{ATT}^{CS} = \sum_{g,t} w_{g,t} \cdot \text{ATT}(g,t)$$

Weights:

$$w_{g,t} \propto N_g \cdot \mathbf{1}\{t \geq g\}$$

$\Rightarrow$ More weight to larger cohorts and feasible post-treatment periods

# Callaway & Sant'Anna (2021): Pros and Cons

**Advantages**

- Cohort-specific effects $\longrightarrow$ respects staggered adoption.
- No treated-vs-treated comparisons.
- Handles covariates via inverse-probability weighting.
- Produces estimable objects aligned with economic questions.

**Disadvantages**

- **Efficiency loss:** Each ATT$(g, t)$ uses only units that are still untreated at time $t$.
- Precision varies: some cohorts have few units left untreated at certain times.

# Sun & Abraham (2021): Core Idea

**Problem solved:** TWFE event studies are contaminated because already-treated units act as controls.

## Core Idea

Interacting cohort $g$ with **relative time** $k$:

$$\beta_{g,k} = \text{ATT}(g, g+k)$$

Event-study coefficient is a **clean aggregation**:

$$\beta_k^{SA} = \sum_g \omega_{g,k}\, \beta_{g,k}$$

- **Residualized TWFE**: orthogonalizes cohort × time patterns to remove treated-as-control problems
- **Convex aggregation**: final dynamic effects are weighted averages with **non-negative weights**

# Sun & Abraham (2021): Details

**Cohort-specific relative-time effects:**

$$\beta_{g,k} = \text{ATT}(g, g+k)$$

## Aggregation across cohorts

For each event time $k$:

$$\beta_k^{SA} = \sum_g \omega_{g,k} \cdot \text{ATT}(g, g+k)$$

Weights:

$$\omega_{g,k} = \frac{N_g \cdot \mathbf{1}\{g+k \leq T\}}{\sum_{g'} N_{g'} \cdot \mathbf{1}\{g'+k \leq T\}}$$

**Interpretation**

- Only cohorts for which event time $k$ is observed contribute.
- Larger cohorts receive more weight.
- No treated-vs-treated comparisons $\rightarrow$ no contamination.

# Sun & Abraham (2021): Pros and Cons

**Advantages**

- Pre-trends ($k < 0$) are unbiased: only uses not-yet-treated controls.
- No negative weights or sign reversals.
- Provides interpretable dynamic effects over event time.

**Disadvantages**

- Only an **event-study method**: does not give one overall ATT.
- For some event times, only a subset of cohorts provide information $\Rightarrow$ estimates near the edges can be noisy.
- Interpretation becomes harder when few groups contribute at a given relative time.
- Requires enough pre-treatment periods for each cohort to check pre-trends.

# Borusyak, Jaravel & Spiess (2024): Core Idea

**Goal:** Recover unbiased treatment effects in staggered DiD **without** assuming homogeneous effects and **without** using treated units as controls.

$$Y_{it}(0) = \alpha_i + \beta_t$$

1. Estimate untreated potential outcomes using **only never-treated or not-yet-treated units**.
2. Impute counterfactual untreated outcomes:

$$\widehat{Y}_{it}(0) = \hat{\alpha}_i + \hat{\beta}_t$$

3. Estimate cell-level treatment effects:

$$\widehat{\tau}_{it} = Y_{it} - \widehat{Y}_{it}(0)$$

4. Aggregate with weights $w_{it}$ to get any ATT estimand:

$$\widehat{\text{ATT}}_w = \sum_{it \in \Omega_1} w_{it}\, \widehat{\tau}_{it}$$

**Key insight:** All unbiased DiD estimators can be written in this form; BJS derive the **unique efficient** one.

# Borusyak, Jaravel & Spiess (2024): Details

**Model for untreated outcomes**

$$Y_{it}(0) = \alpha_i + \beta_t + \varepsilon_{it}$$

This is estimated **only on untreated observations**:
Never-treated or not-yet-treated.

**Efficient estimator under unrestricted heterogeneity:**

$$\widehat{\tau}_{it} = Y_{it} - \hat{\alpha}_i - \hat{\beta}_t$$

Aggregate via:

$$\widehat{\tau}^*_w = \sum_{it \in \Omega_1} w_{it}\, \widehat{\tau}_{it}$$

**Identification requirements:**

- Parallel trends in $Y_{it}(0)$ (unit FE + time FE structure)
- No anticipation for untreated *it*
- Enough untreated support to estimate FE structure
- Estimand $ATT_w$ must be identified (no ATT beyond comparable horizons)

# BJS with Repeated Cross-Sections (I): Why It Matters

**Many empirical DiD applications use repeated cross-sections:**
CPS, ACS, LFS, Eurostat LFS, DHS, household surveys, etc.

**Problem:** In repeated cross-sections, we **cannot** estimate unit fixed effects $\alpha_i$
because the same units are not followed over time.

**BJS solution:** Replace unit FE with a **flexible function of covariates**:

$$\mathbb{E}[Y_{it}(0) \mid X_i, t] = m(X_i) + \beta_t.$$

- Estimate $m(X_i)$ using untreated observations
- Impute counterfactuals:

$$\widehat{Y}_{it}(0) = \widehat{m}(X_i) + \hat{\beta}_t$$

- Proceed exactly as in the panel case:

$$\widehat{\tau}_{it} = Y_{it} - \widehat{Y}_{it}(0)$$

# BJS with Repeated Cross-Sections (II): Interpretation

**Assumption becomes:**

$$\mathbb{E}[Y_{it}(0) \mid X_i, t] = m(X_i) + \beta_t$$

rather than unit-specific $\alpha_i$.

- This is the standard **conditional parallel trends** assumption used in repeated cross-sections.
- No treated units are ever used as controls for others.
- Flexible $m(X)$ allows rich composition adjustment across years.

**Why this is useful:**

- Works when panel data are unavailable or impossible (e.g. rotating surveys).
- Avoids limitations of TWFE and event-study estimators in repeated XS.
- Often more efficient than IPW-based repeated-XS DiD methods.

# Borusyak, Jaravel & Spiess (2024): Pros and Cons

**Advantages**

- Unbiased under **arbitrary** treatment-effect heterogeneity.
- Efficient among all linear unbiased estimators under spherical errors.
- Works with covariates, triple-differences, repeated cross-sections.
- Clean separation of pre-trend testing and effect estimation.

**Disadvantages**

- Needs many untreated observations to estimate FE precisely.
- Efficiency gains shrink with strong serial correlation.
- Cannot identify long-run effects beyond comparison window.
- Requires two-step implementation (though computationally fast).

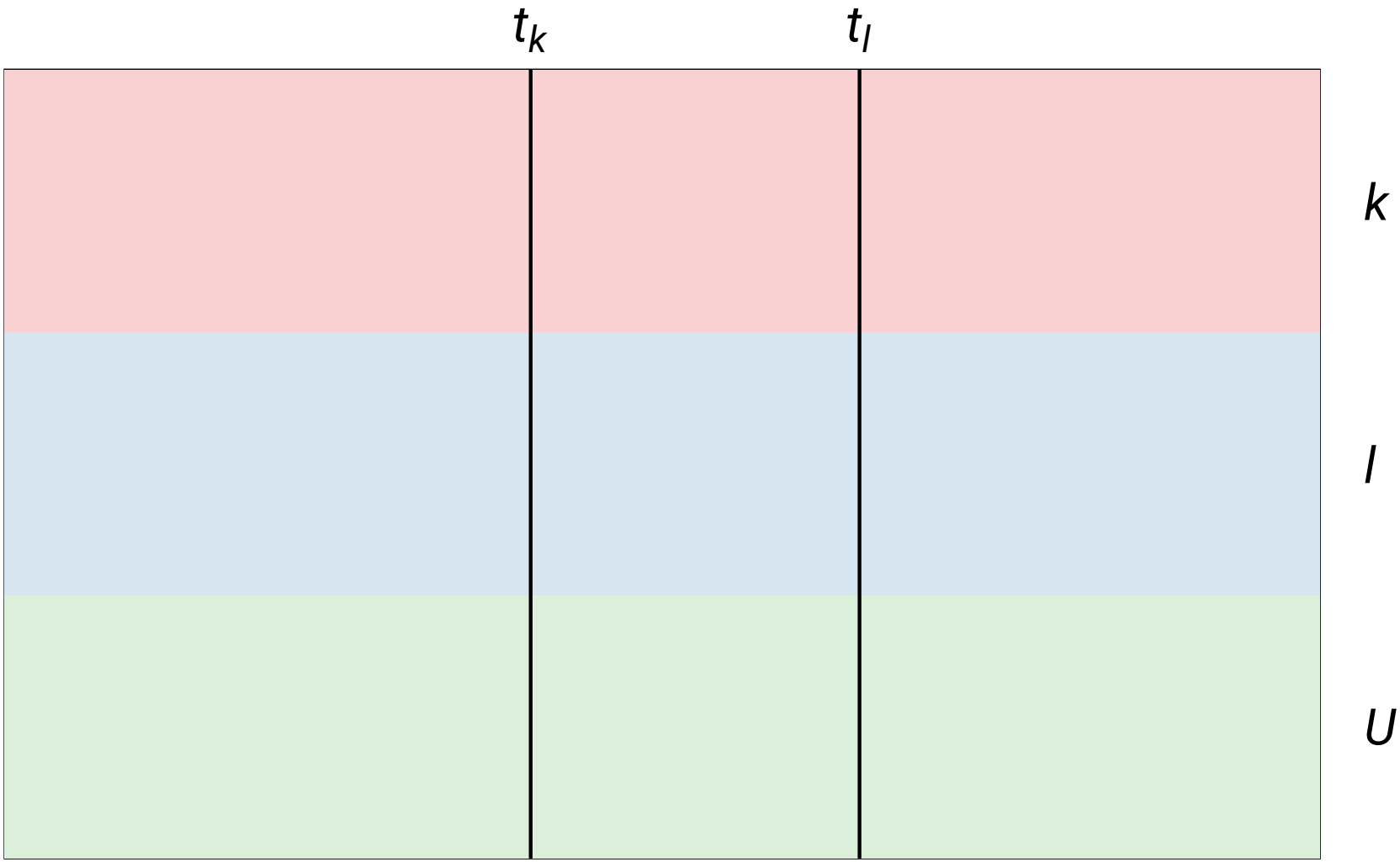
# Core Idea: Stacked Difference-in-Differences

**Idea:** Build a separate "clean" DID for each adoption event (sub-experiment) and **stack** them vertically into one dataset.

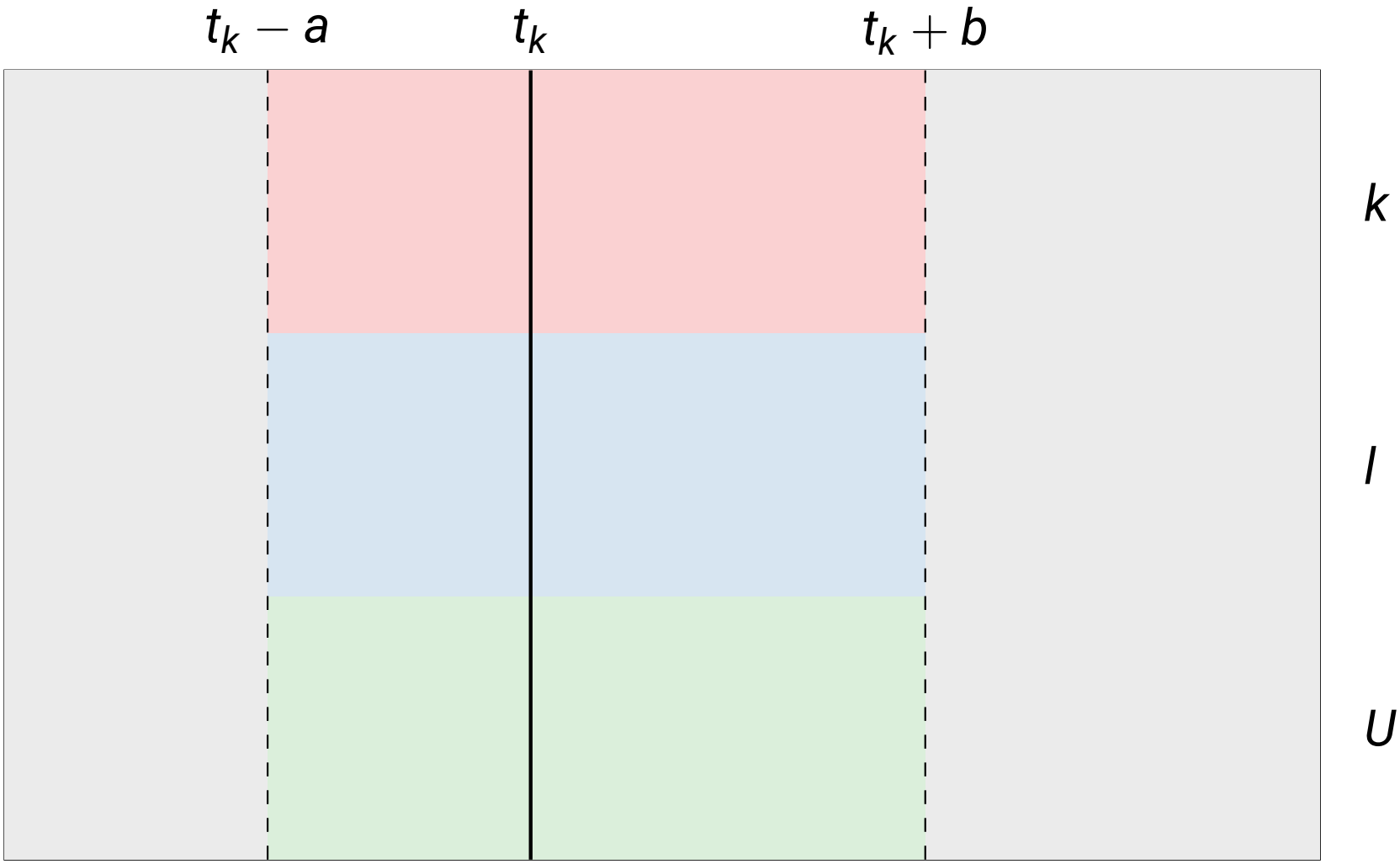
- Removes contaminated late-vs-early comparisons from TWFE.
- Each sub-experiment includes only:
  - Units first treated at *a*
  - "Clean controls" not yet treated at event time
- Run a single TWFE DID or event-study regression on the stacked data.

**Goal:** Recover an average causal effect by pooling multiple valid 2x2 DIDs.

# Stacking: The data with an imbalanced treatment times

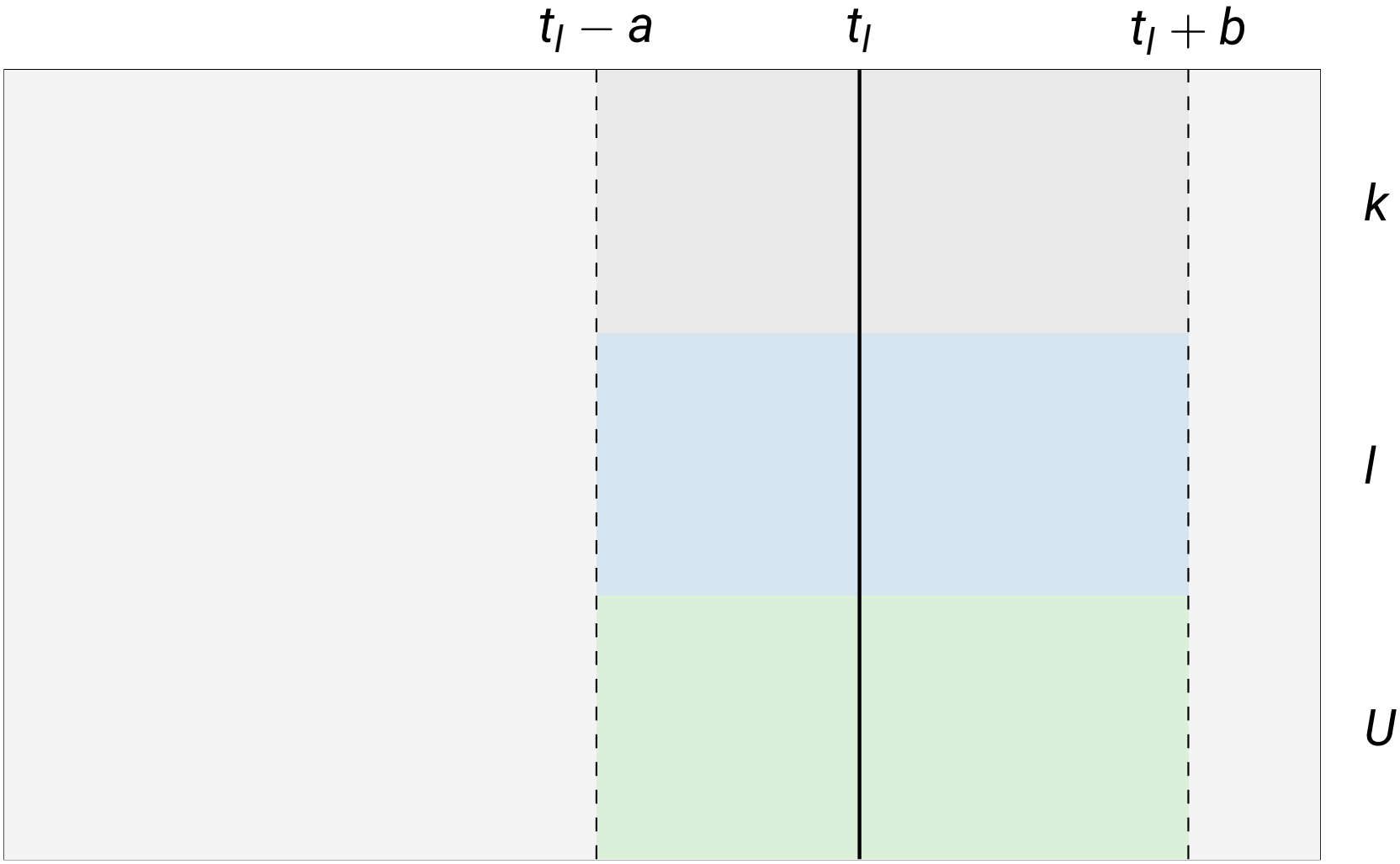


# Stacking: Creating the $k$-dataset




Choose max pre ($-a$) and post ($+b$) periods

# Stacking: Creating the *l*-dataset




Choose max pre ($-a$) and post ($+b$) periods

Already-treated $k$ observations are dropped as controls for $l$

# Stacking: Estimation after creating the datasets

**Goal:** Estimate dynamic treatment effects using the two clean treatment-control datasets (for groups $k$ and $l$).
**Classical steps for stacking:**

1. **Stack the two trimmed datasets**
   Each row corresponds to a triple:

   $$(s, g, e) \quad \text{unit, cohort } g \in \{k, l\},\ e = t - t_g$$

   Only observations inside the window $[t_g - a,\ t_g + b]$ appear.

2. **Estimate an event-study regression on the stack**

   $$Y_{sge} = \alpha_g + \lambda_e + \sum_{e \neq -1} \beta_e \cdot 1\{e\} + \varepsilon_{sge}$$

   with:

   - cohort fixed effects $\alpha_g$,
   - relative-time indicators $1\{e\}$,
   - $e = -1$ as the omitted category.

# Stacking: Identification Problem

**Problem (Wing, Freedman & Hollingsworth, 2024):** The **unweighted** stacked DID does not identify

- the ATT,
- any causal aggregate of ATTs,
- nor any convex combination of treatment effects.

## Reason: Implicit differential weighting (core failure)

$$DID_e^{stack} = \sum_j \frac{N_j^D}{N^D} \Delta Y_{j,e}^D - \sum_j \frac{N_j^C}{N^C} \Delta Y_{j,e}^C$$

- Treated trends weighted by $N_j^D / N^D$
- Control trends weighted by $N_j^C / N^C$

$\Rightarrow$ **Weights differ across treated and control arms.**

Even if parallel trends hold **within every sub-experiment**, these untreated trends **do not cancel**, so:

$$DID_e^{stack} \neq \text{ATT}(a, a+e) \quad \text{(not any meaningful causal parameter).}$$

# Stacking: Pros and Cons

**Advantages**

- Conceptually simple: regression-based estimator familiar to applied researchers.
- Makes research designs explicit (each sub-experiment is a clean 2x2 DID).
- Easy to implement dynamic effects/event studies.

**Disadvantages (critical)**

- **Bias:** Classic stacked regressions apply different implicit weights to treated and control groups across sub-experiments $\Rightarrow$ **not a causal estimand**.
- Not a convex combination of ATTs.
- Compositional change across event time unless data are trimmed.
- Requires corrective weights to identify the trimmed aggregate ATT.

# Where We Go Next

- So far: modern DiD methods correct TWFE's weighting and timing problems.
- But all still rely on a (conditional) parallel trends assumption.
- **Synthetic DiD** relaxes this by blending synthetic controls with DiD.
- It allows for data-driven construction of untreated counterfactuals.

# 11.6: Synthetic Difference-in-Differences

# Why Talk About Synthetic Control in a DiD Lecture?

We will introduce synthetic control first. It was developed for settings, where **parallel trends are doubtful**, especially for a **single treated unit**

- SC solves exactly the core DiD problem: How do we build a credible counterfactual trend?
- Many real-world applications combine:
- It naturally leads to **Synthetic Difference-in-Differences** (Arkhangelsky et al., 2021), a hybrid approach

**Roadmap:** SC $\rightarrow$ California smoking ban $\rightarrow$ placebo inference $\rightarrow$ multi-unit generalization $\rightarrow$ Synthetic DiD

# Synthetic Control: Core Idea

**Abadie, Diamond & Hainmueller (2010, 2015):** Construct a "synthetic" untreated unit that recreates the treated unit's **pre-treatment path**.

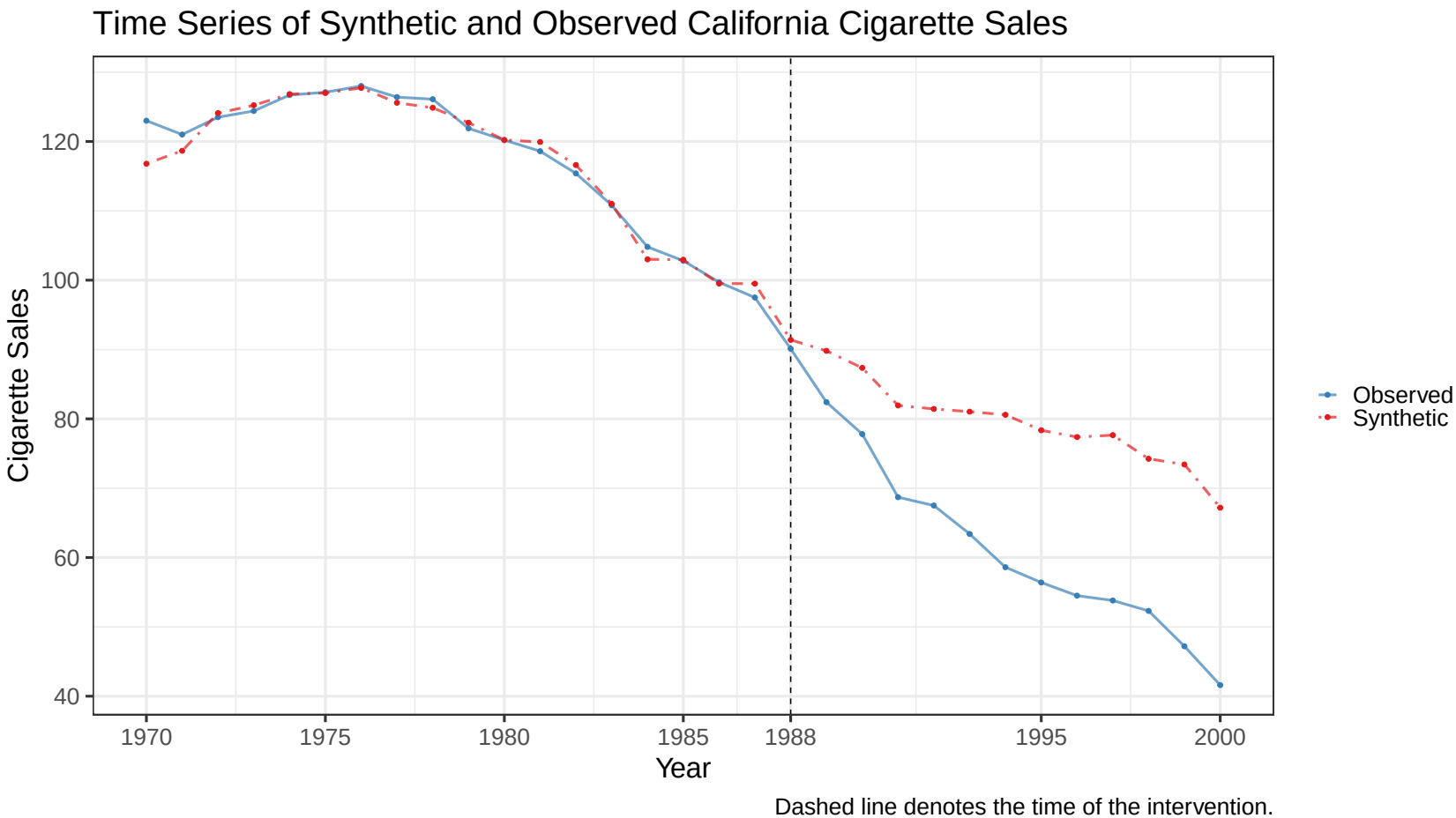
- **Example Treated unit:** e.g. California (Prop 99, 1988 tobacco control program)
- **Donor pool:** other US states never adopting the reform in that period
- Find weights $w_j \geq 0$, $\sum_j w_j = 1$ s.t.

$$Y_{\text{CA, pre}} \approx \sum_j w_j Y_{j,\text{pre}}$$

- **Synthetic CA** = weighted average of control states that best replicate CA's pre-1988 cigarette sales.

**Motivation:** If synthetic CA matches actual CA well before treatment, then deviations after treatment represent the treatment effect.

# Example: Did California's 1988 Smoking Ban Reduce Cigarette Sales?



# Pre-treatment Fit: The Heart of Synthetic Control

**Pre-treatment match is the credibility test**

- If synthetic CA reproduces CA's **trend** and **level**, the donor pool is a valid counterfactual.
- Large pre-treatment Error $\rightarrow$ poor match $\rightarrow$ synthetic control unreliable
- A good synthetic control should:
  - track treated unit year-by-year pre-treatment
  - replicate all main covariates (e.g., income, demographics)

**Summary:** Synthetic control is fundamentally about building a credible counterfactual trend. The better the pre-treatment match, the more trustworthy the post-treatment estimates

# Multiple Predictors in Synthetic Control

Synthetic control matches the treated unit to a weighted average of donors using a **vector of predictors**, not just the outcome:

$$X_{\text{treated}} \approx X_w = \sum_{j \in \mathcal{C}} w_j X_j.$$

**Predictors often include:**

- lagged outcomes (in the example: cigsales 1975, 1980, 1988)
- averages over windows (here for income, cigarette prices, demographics)
- additional behavioral variables (here: beer consumption)

**Motivation:** If donors match California on these predictors, they form a credible synthetic counterfactual.

# How Synthetic Control Chooses Unit Weights

Weights $w_j$ create the synthetic unit:

$$\text{Synthetic} = \sum_j w_j Y_{j,t}, \qquad w_j \geq 0, \ \sum_j w_j = 1.$$

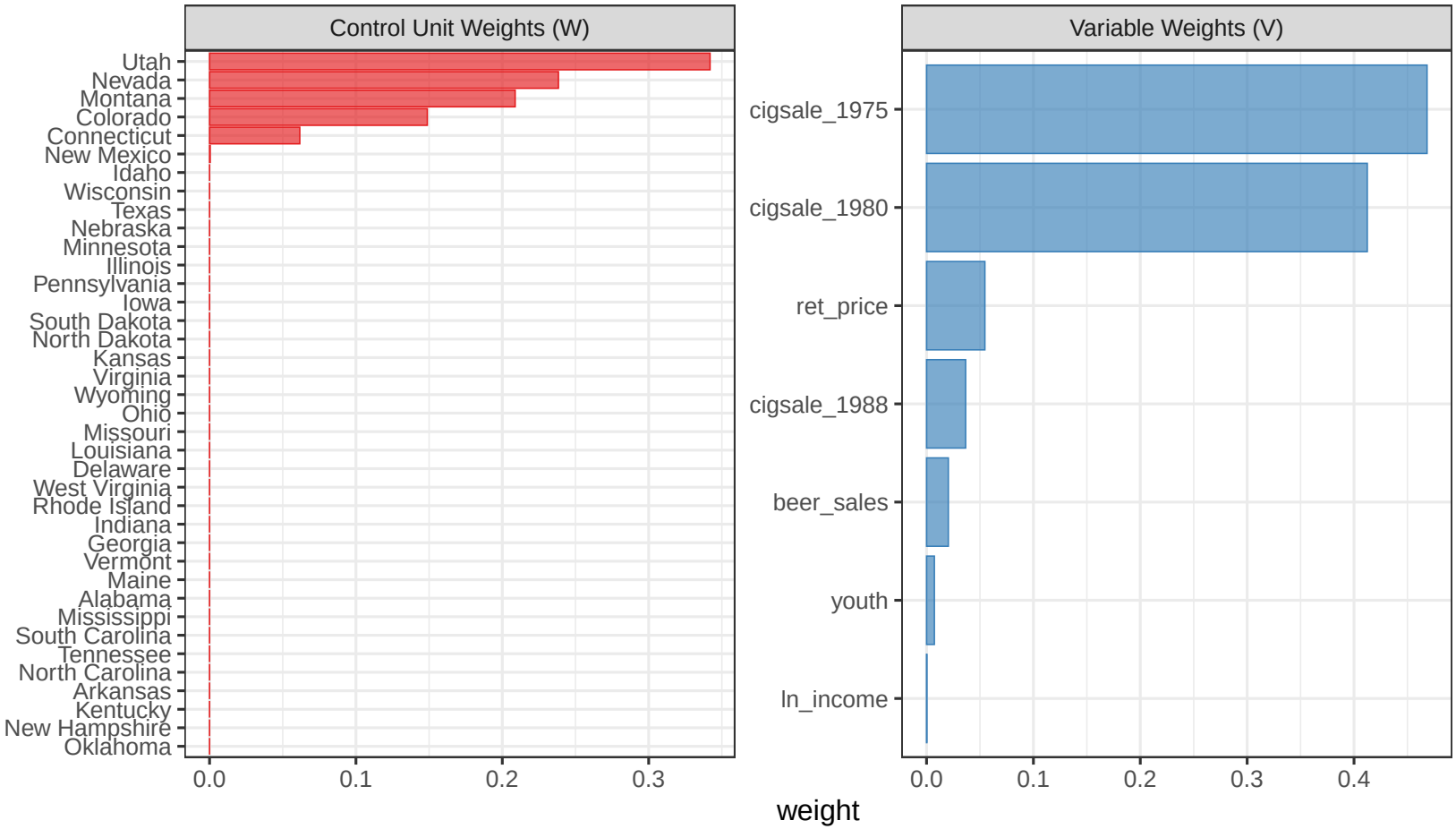
They are chosen to minimize the discrepancy in predictors:

$$\min_w (X_{\text{CA}} - X_w)' V (X_{\text{CA}} - X_w).$$

**Interpretation:**

- Units that closely match California across all predictors receive high weights.
- Poorly matching states receive weights near zero.
- The predictor importance matrix $V$ emphasizes variables most predictive of pre-trends.
- We can see how the synthetic unit is composed

# Unit Weights in our Example



# Placebo Tests: Is California “Special”?

To judge whether CA’s gap is meaningful:
**In-space placebo test:**

- Pretend each donor state received the 1988 ban.
- Construct synthetic controls for all donor states.
- Plot all placebo gaps together with CA’s.

**Interpretation:**

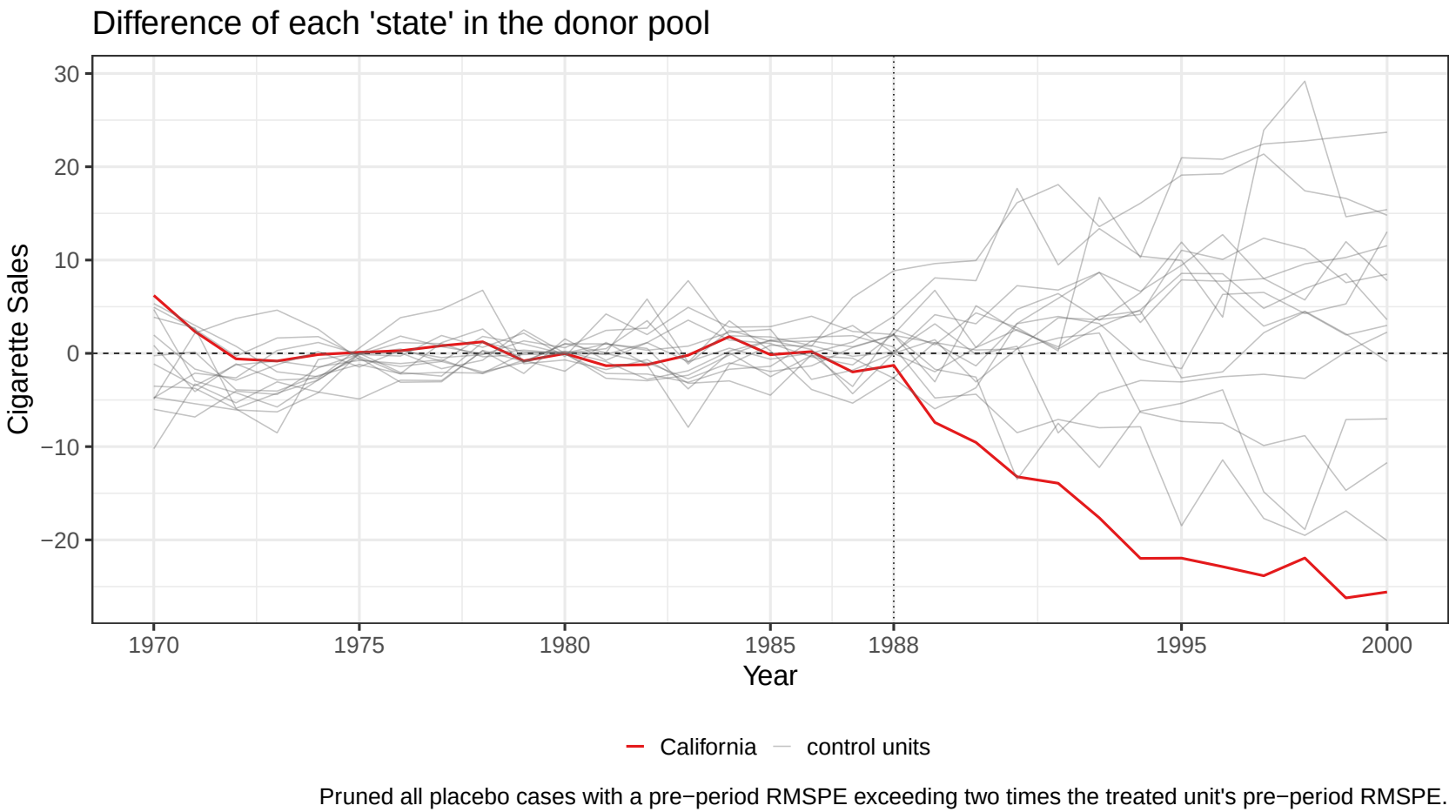
- If CA’s post-treatment gap is among the largest, the effect is unlikely due to chance.
- If many placebo states show equal/larger gaps, results are weak.

**Root Mean Squared Prediction Error ratio test (Abadie et al. 2010):**

$$\text{Ratio}_j = \frac{\text{post-RMSPE}_j}{\text{pre-RMSPE}_j}$$

California should be an extreme outlier if Proposition 99 had real effects.

# Placebos for our Example




Pruned all placebo cases with a pre−period RMSPE exceeding two times the treated unit's pre−period RMSPE.

# Extending Synthetic Control to Multiple Treated Units

Synthetic control was designed for **one treated unit**. But reforms often affect many units at once (*our classic Diff-in-Diff case!*)
**Two main approaches:**

1. Build a separate synthetic control for each treated unit and average the estimated effects.
2. Pool treated units and estimate a single synthetic comparison group.

**Challenges:**

- More treated units $\rightarrow$ harder to find one donor pool that fits all.
- Noise and imbalance across many pre-treatment fits.
- SC becomes computationally heavy and conceptually brittle.

**Need:** A method that combines SC's weighting with the scalability of DiD.

# Synthetic Difference-in-Differences (SDID)

**Arkhangelsky, Athey, Hirshberg, Imbens & Wager (AER, 2021)** develop a hybrid method that combines:

- **Synthetic Control:** Unit weights to match pre-trends
- **Difference-in-Differences:** Time FE, unit FE, large-sample inference

**Idea:**

- Estimate unit weights $\omega_i$ so controls mimic treated units' pre-trends
- Estimate time weights $\lambda_t$ to align pre/post periods
- Run a weighted TWFE regression:

$$\hat{\tau}_{SDID} = \arg\min_{\tau} \sum_{i,t} \omega_i \lambda_t (Y_{it} - \alpha_i - \beta_t - \tau W_{it})^2.$$

**Result:** A robust estimator usable for multiple treated units

# Synthetic DiD: Key Ingredients

**Goal:** Improve DiD when **parallel trends are doubtful** by using data-driven pre-treatment matching, in the spirit of synthetic control.

**Two sets of weights:**

- **Unit weights** $\omega_i$: Make the weighted average of control units mimic the **treated units' pre-treatment path**.

$$\sum_{i \in C} \omega_i Y_{i,t} \approx \frac{1}{N_T} \sum_{i \in T} Y_{i,t} \quad \text{for all pre-treatment } t$$

- **Time weights** $\lambda_t$: Make the weighted average of pre-treatment periods match the **post-treatment level/trend**.

$$\sum_{t < T_0} \lambda_t Y_{\cdot,t} \approx \sum_{t \geq T_0} \lambda_t Y_{\cdot,t}$$

**Idea:** Use SC-style weights to balance pre-trends, then estimate the treatment effect using a weighted TWFE regression for efficiency.

# How SDID Chooses the Weights

SDID chooses weights by matching pre-treatment outcomes, but it also adds a small **ridge penalty** so the weights do not become extreme.

**1. Unit weights: Make controls look like treated units (pre-period)**

$$\omega = \arg \min_{\omega_i \geq 0,\ \sum_i \omega_i = 1} \underbrace{\sum_{t<T_0} \left( \bar{Y}_t^T - \sum_{i \in C} \omega_i Y_{i,t} \right)^2}_{\text{match pre-treatment paths}} + \underbrace{\lambda_\omega \|\omega\|_2^2}_{\text{ridge penalty}}$$

- The first term enforces good pre-period fit
- The ridge penalty pushes weights toward being smoother (shrinks size of the weights)

# How SDID Chooses the Weights

**2. Time weights: Make pre-period look like post-period**

$$\lambda = \arg \min_{\lambda_t \geq 0,\ \sum_t \lambda_t = 1} \underbrace{\sum_{i \in C} \left( \bar{Y}_i^C - \sum_t \lambda_t Y_{i,t} \right)^2}_{\text{match pre- and post-period averages}} + \underbrace{\lambda_\lambda \|\lambda\|_2^2}_{\text{ridge penalty}}$$

- SDID compares post to pre just like DiD, but not all pre-periods are equally informative
- Time weights reweight the pre-period so it better represents the post-period the treated units would have faced
- This avoids letting noisy or unrepresentative early pre-periods distort the counterfactual
- Result: a more credible *"post - pre"* difference and a better-aligned DiD comparison.

# The SDID Estimator and Interpretation

Given unit weights $\omega_i$ and time weights $\lambda_t$, SDID estimates:

$$\hat{\tau}_{SDID} = \sum_{i,t} \omega_i \lambda_t \left( Y_{i,t} - \hat{\alpha}_i - \hat{\beta}_t \right) W_{i,t}$$

where:

- $\hat{\alpha}_i$ = unit FE estimated using weighted controls
- $\hat{\beta}_t$ = time FE estimated using weighted controls
- $W_{i,t}$ = treatment indicator

## Interpretation

- Pre-treatment weighted means of treated and weighted controls are **balanced** by design.
- Post-treatment effect is then identified using a DiD-style contrast of

  (treated units) − (synthetic controls).

- Combines SC's robustness to trend misspecification with DiD's statistical power.

# 11.7: Summary and Guidelines

# Modern DiD: Big Picture

- DiD is a **research design**, not just a regression:
  - Target parameter: typically ATT for some group $\times$ time
  - Identification: (variants of) parallel trends + no anticipation
- Even complicated settings (staggered timing, covariates, weights) can be viewed as **aggregations of 2$\times$2 "building blocks"**
  - Each building block: one group where treatment changes vs. one where it does not
  - Identification comes from simple 2$\times$2 parallel trends
- **Forward-engineering vs. reverse-engineering**
  - Forward: start from causal question $\rightarrow$ parameter $\rightarrow$ assumptions $\rightarrow$ estimator
  - Reverse: start from TWFE and ask "when does this have a causal interpretation?"
- Baker, Callaway, Goodman-Bacon, Sant'Anna (2025): DiD practice should be organized around **clear causal targets**, **transparent assumptions**, and **heterogeneity**.

# A Practical DiD Checklist (Baker et al. + this lecture)

1. **Define the causal object**
   - Which ATT(s)? $2\times2$, event-time ATTs, group-time ATTs, distributional effects?
   - Unit of analysis, timing group, weights (units vs. population)?

2. **State identification assumptions**
   - Which parallel trends? (never-treated, not-yet-treated, all-groups)
   - No anticipation, SUTVA, overlap

3. **Assess plausibility**
   - Pre-trend / event-study diagnostics (Section 11.3)
   - Balance in covariates and outcomes; theory for selection into treatment

# A Practical DiD Checklist (Baker et al. + this lecture)

3. **Choose an estimator consistent with 1–3**
   - Avoid plain TWFE when there is staggered timing and heterogeneity (Section 11.4)
   - Use a modern estimator fitting your setting

4. **Do inference and assess robustness**
   - Cluster appropriately; be explicit about what is “random”
   - Sensitivity to alternative comparison groups, functional forms, weights, estimators

# Selected Advanced Extensions (Not Covered)

Here is an overview of important topics in the DiD literature that we do not discuss in detail that might be interesting for your research:

- **Robustness to Parallel-Trends Violations**
  Rambachan & Roth, 2023: *"A More Credible Approach to Parallel Trends"*
  Instead of assuming exact parallel trends, this approach imposes bounds on how far post-treatment trends may deviate from pre-treatment trends.
- **Continuous Treatments in DiD**
  Callaway et al. 2024: *"Difference-in-Differences with a Continuous Treatment"*
- **Doubly-Robust & Conditional Parallel Trends**
  Sant'Anna & Zhao, 2018: *"Doubly Robust Difference-in-Differences Estimators"*
  Handling conditioning on covariates and more flexible parallel-trends assumptions efficiently.

# Where to Follow Current Developments in DiD

Modern DiD is an active research area. Excellent sources for staying up to date:

- **Mixtape Sessions (Cunningham et al.)** Comprehensive lecture series with slides, code, and applications. `github.com/Mixtape-Sessions`
- **Difference-in-Differences Reading Group (YouTube)** Presentations by leading DiD researchers (Roth, Goodman-Bacon, Callaway, Sant'Anna, etc.). `YouTube: DiD Reading Group`
- **Authors' GitHubs** Cutting-edge methods and replication code: jonathandroth, Callaway, Sant'Anna