

# Panel Data II: Difference-in-Differences

Francisco Villamil

Applied Quantitative Methods II  
MA in Social Sciences, Spring 2026

1/36

## Today's goals

- Understand Difference-in-Differences as an extension of TWFE
- Learn the  $2 \times 2$  DiD setup and the parallel trends assumption
- Interpret the DiD regression coefficient
- Use event studies to visualize dynamic effects and test pre-trends
- Understand the staggered DiD problem and modern solutions

2/36

This lecture is the natural continuation of Panel Data I. Last session, two-way fixed effects used within-unit variation net of common time trends. Today we add the causal design layer: instead of a general panel, we have an intervention that affected some units (the treated group) but not others (the control group). DiD combines the FE logic with an explicit treatment-control comparison. The methods we cover — especially the modern staggered DiD estimators — are at the frontier of applied research in political science and economics.

# Roadmap

From Fixed Effects to DiD

The DiD Estimator

Parallel Trends and Its Threats

Event Studies

Staggered DiD and Recent Advances

Wrap-up

3/36

## Recap: what fixed effects does

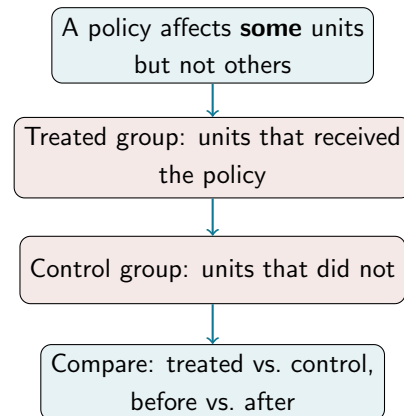
$$y_{it} = \alpha_i + \gamma_t + \beta x_{it} + \varepsilon_{it}$$

- $\alpha_i$ : absorbs all **time-invariant** unit differences
- $\gamma_t$ : absorbs all **unit-invariant** time shocks
- $\hat{\beta}$ : identified from within-unit variation, net of common trends
- But TWFE treats  $x_{it}$  as a continuous variable that varies continuously over time
- What if the variation comes from a **specific, discrete intervention**?

4/36

Briefly ground today's lecture in the previous session. TWFE with a continuously varying  $x_{it}$  is one use case of panel data. But a very important and common use case is a policy intervention: something happens at a specific point in time to a specific set of units. DiD is the framework for this. Stress that DiD does not replace TWFE — it is TWFE with an explicit causal design layered on top. The difference is in how we think about identification and what assumption we are willing to defend.

## A new question: policy interventions



5/36

This frame sets up the conceptual scaffolding for DiD. The scenario is everywhere in social science: a minimum wage increase in one state but not another; electoral reform in some countries but not others; a public health intervention rolled out in some regions. The key is that we have a comparison group that did not receive the treatment and can serve as a counterfactual. DiD exploits this variation: we compare the change in outcomes for the treated group to the change in outcomes for the control group.

## Running example: Card & Krueger (1994)

- **Question:** Does raising the minimum wage reduce employment?
- **Policy:** NJ raised the minimum wage on April 1, 1992 — from \$4.25 to \$5.05 per hour
- **Design:**
  - **Treatment:** fast-food restaurants in New Jersey (NJ)
  - **Control:** fast-food restaurants in Pennsylvania (PA)
  - **Before:** Feb–Mar 1992; **After:** Nov–Dec 1992
- Pennsylvania did not change its minimum wage

6/36

Card and Krueger (1994) is the classic DiD paper and remains one of the most cited papers in labor economics. The conventional wisdom at the time was that minimum wage increases reduce employment (basic supply-demand). Card and Krueger challenged this using a natural experiment: NJ happened to raise its minimum wage while neighboring PA did not. This makes PA a natural control group. The idea is that NJ and PA fast-food restaurants face similar economic conditions, so PA's employment trajectory tells us what NJ's would have been without the policy. Walk through the timing carefully: "before" is measured before the NJ increase, "after" is measured several months after.

## The 2×2 DiD table

	Before	After
Control (PA)	$\bar{y}_{C,pre}$	$\bar{y}_{C,post}$
Treated (NJ)	$\bar{y}_{T,pre}$	$\bar{y}_{T,post}$

$$\hat{\delta}_{DiD} = (\bar{y}_{T,post} - \bar{y}_{T,pre}) - (\bar{y}_{C,post} - \bar{y}_{C,pre})$$

- Control group change: what would have happened **without** treatment
- DiD subtracts this “counterfactual trend” from the treated group change

7/36

This 2×2 table is the most important thing to get right. Walk through each cell: top-left is average employment in PA before the NJ wage hike; top-right is average employment in PA after; bottom-left is NJ before; bottom-right is NJ after. The DiD formula subtracts the control group’s change (PA’s before-to-after difference) from the treated group’s change (NJ’s before-to-after difference). The PA change is our estimate of what would have happened to NJ employment in the absence of the policy. The residual — the DiD — is the causal effect. In Card and Krueger, the DiD was positive: NJ employment actually went up relative to PA, challenging the conventional view.

## Roadmap

From Fixed Effects to DiD

The DiD Estimator

Parallel Trends and Its Threats

Event Studies

Staggered DiD and Recent Advances

Wrap-up

8/36

## The DiD formula

$$\hat{\delta} = \underbrace{(\bar{y}_{T,post} - \bar{y}_{T,pre})}_{\text{treated group change}} - \underbrace{(\bar{y}_{C,post} - \bar{y}_{C,pre})}_{\text{control group change}}$$

- **Treated change:** NJ employment went from 20.4 to 21.0 FTEs per restaurant (+0.59)
- **Control change:** PA employment went from 23.3 to 21.2 FTEs per restaurant (−2.17)
- $\hat{\delta} = 0.59 - (-2.17) = +2.76$
- Interpretation: the minimum wage **raised** NJ employment by  $\approx 2.76$  FTEs relative to PA

9/36

Work through the Card-Krueger numbers concretely. The raw NJ change is small and positive; the raw PA change is negative. Because PA employment fell over the same period (probably due to the broader recession), the counterfactual for NJ is also a decline. DiD shows that NJ outperformed this counterfactual by about 2.76 full-time equivalent employees per restaurant. This is a striking result that upended conventional wisdom. The numbers here are rounded from the original paper — encourage students to look up the original Table 3 in Card and Krueger (1994). Stress that the sign and magnitude of DiD depend critically on what the control group did: if we had not subtracted the PA trend, we would have dramatically underestimated the effect.

## The regression formulation

$$y_{it} = \alpha + \beta_1 \underbrace{Post_t}_{\text{time}} + \beta_2 \underbrace{Treat_i}_{\text{group}} + \underbrace{\delta (Post_t \times Treat_i)}_{\text{interaction = DiD}} + \varepsilon_{it}$$

- $Post_t = 1$  if observation is in the post-treatment period
- $Treat_i = 1$  if unit is in the treatment group
- $\delta =$  **DiD coefficient of interest:** causal effect of treatment
- $\alpha$ : control group mean, pre-period
- $\beta_1$ : time trend (common to both groups)
- $\beta_2$ : pre-period difference between groups

10/36

The regression formulation is algebraically equivalent to the 2×2 table but extends easily to more units and more time periods. Walk through each coefficient:  $\alpha$  is the intercept, i.e., the expected value for the control group before treatment.  $\beta_1$  captures how much outcomes changed from pre to post for the control group.  $\beta_2$  captures the pre-period level difference between treated and control. The interaction  $\delta$  is the DiD: how much the treated group's post-period outcome deviates from what we would predict based on the common time trend plus the pre-period group difference. Emphasize that  $\delta$  is the only parameter of causal interest; the others are nuisance parameters.

## Regression: recovering the 2×2 cells

	Before ( $Post = 0$ )	After ( $Post = 1$ )
Control ( $Treat = 0$ )	$\alpha$	$\alpha + \beta_1$
Treated ( $Treat = 1$ )	$\alpha + \beta_2$	$\alpha + \beta_1 + \beta_2 + \delta$

- DiD from table:  $((\alpha + \beta_1 + \beta_2 + \delta) - (\alpha + \beta_2)) - ((\alpha + \beta_1) - \alpha)$
- $= (\beta_1 + \delta) - \beta_1 = \delta \checkmark$
- The regression recovers  $\hat{\delta}$  automatically via OLS

11/36

This algebraic derivation shows that the regression is doing exactly what the 2×2 table does, just packaged as OLS. Plugging  $Post = 0, Treat = 0$  gives  $\alpha$  (control pre-period). Plugging  $Post = 1, Treat = 0$  gives  $\alpha + \beta_1$  (control post-period). Plugging  $Post = 0, Treat = 1$  gives  $\alpha + \beta_2$  (treated pre-period). Plugging  $Post = 1, Treat = 1$  gives  $\alpha + \beta_1 + \beta_2 + \delta$  (treated post-period). Take the differences as in the DiD formula and everything cancels except  $\delta$ . This is reassuring: the regression is a convenient way to run the same calculation with standard errors.

## DiD in R with feols

- Basic 2×2 DiD (two groups, two periods):  

```
feols(employment ~ post * treat,  
      data = df, cluster = ~state)
```
- With multiple units and periods, use TWFE:  

```
feols(employment ~ treated | unit + time,  
      data = df, cluster = ~unit)
```
- $treated = Post_t \times Treat_i$ : a single binary indicator
- Unit and time FE replace the  $Treat_i$  and  $Post_t$  dummies
- Always cluster standard errors at the **treatment level**

12/36

In practice, with more than two units, the preferred formulation absorbs the  $Treat_i$  and  $Post_t$  terms into unit and time fixed effects and includes only the interaction (the treated dummy) as the regressor of interest. This is the standard panel DiD regression. Note that in `feols`, the `|` separates the regressor from the FEs, so `feols(y ~ treated | unit + time)` is equivalent to running OLS with unit dummies, year dummies, and the treatment dummy. Clustering at the unit level is standard: treatment is assigned at the unit level, and errors within a unit are correlated over time. The two syntaxes (interaction vs. FE absorbing group/time) give identical estimates in the simple 2×2 case.

# Roadmap

From Fixed Effects to DiD

The DiD Estimator

Parallel Trends and Its Threats

Event Studies

Staggered DiD and Recent Advances

Wrap-up

13/36

## The key assumption: parallel trends

### Parallel Trends Assumption

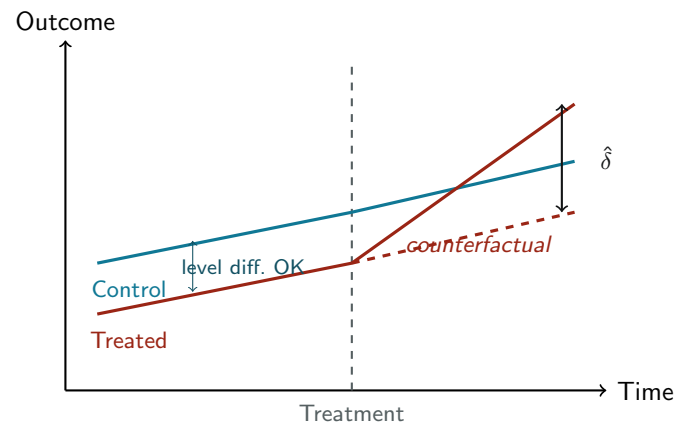
In the absence of treatment, the treated and control groups would have followed the **same time trend**.

- What this allows: use the control group's post-period change as the **counterfactual trend** for the treated group
- What this does **not** require:
  - The two groups need not have the same baseline level
  - Pre-period outcomes can differ — just not the trends
- This is an **assumption**, not a testable fact (we cannot observe the counterfactual)

14/36

Parallel trends is the linchpin of DiD. Make sure students understand what it says and what it does not say. It does NOT require NJ and PA to have the same average employment level. They can differ by any amount (that difference is absorbed by  $\beta_2$  in the regression). What it requires is that absent the minimum wage increase, the change in NJ employment would have been the same as the change in PA employment. If PA had an economic boom in 1992 that NJ missed, or if NJ had structural changes underway independently, parallel trends would be violated. The key phrase “in the absence of treatment” makes this counterfactual and thus untestable in the post-period — we can only observe the actual trajectory of NJ after the wage hike.

## Parallel trends: visualization



15/36

Walk through this diagram carefully. Pre-period: both lines have the same slope (parallel trends holds). There is a level difference between them, but that is fine. At the treatment date, the treated group diverges upward. The dashed line shows the counterfactual: what treated units would have done absent treatment, following the same slope as control. The DiD estimate  $\hat{\delta}$  is the vertical gap between the actual treated trajectory and the counterfactual at the post-period. The diagram also shows that a level difference in the pre-period is completely fine — DiD subtracts it out. What would break the identification is if the two lines had different slopes in the pre-period.

## What can violate parallel trends?

- **Selection into treatment correlated with trends**

- Units that adopt the policy were already on different trajectories
- Example: states with rising employment pre-trend more likely to adopt minimum wage

- **Simultaneous confounders**

- Another event coincides with the treatment
- Example: NJ also changed another labor market policy in 1992

- **Anticipation effects**

- Treated units change behavior before the official treatment date
- Example: firms start hiring/firing when the wage increase is announced

16/36

These three threats are the main reasons DiD can fail. Selection into treatment is the hardest to rule out: if units choose to adopt a policy, they may do so precisely because they were on an upward trajectory (or downward). Simultaneous confounders are like the classic OVB story but applied to time: something else changes at the same time as the treatment. Anticipation effects shift the true treatment date earlier than the nominal date, which means the “pre-period” data already reflects the policy. In Card and Krueger, the anticipation concern is minor (the minimum wage increase was legislated some time before April 1992, but it is unclear how much restaurants adjusted in advance). The simultaneous confounders concern is addressed by using adjacent PA as the control: any statewide NJ confounders are unlikely to affect PA.



The parallel trends assumption says treatment would not have changed the outcome trajectory.

When is this plausible?

What makes NJ and PA a good comparison?

17/36

Discussion prompt. Expected answers: (1) NJ and PA are geographically adjacent with similar labor markets, similar demographics, similar consumer patterns. (2) The policy change was relatively small (a \$0.80 increase). (3) The fast-food sector is similar in both states. (4) One concern: if NJ and PA have different underlying economic trends (NJ's economy was recovering faster), parallel trends might fail. (5) Another concern: if firms had already started adjusting in anticipation of the wage increase, the “pre-period” data from Feb-Mar 1992 might already reflect the shock. Allow 1-2 minutes of student input.

## Roadmap

From Fixed Effects to DiD

The DiD Estimator

Parallel Trends and Its Threats

Event Studies

Staggered DiD and Recent Advances

Wrap-up

18/36

## Beyond the single post-period

- The 2×2 DiD uses one pre and one post period
- With panel data spanning many periods, we can do **more**:
  - Estimate treatment effects at each time point relative to treatment
  - Trace out the **dynamic path** of the effect
- Two purposes:
  - **Pre-trend test**: were pre-treatment trends actually parallel?
  - **Dynamic effects**: does the effect grow, fade, or remain constant?
- This is the **event study** design

19/36

The event study is the modern standard for DiD analysis with multiple periods. It extends the simple post-dummy to a full set of leads and lags around the treatment date. Pre-treatment coefficients (leads) should be near zero if parallel trends holds: they test whether the treated and control groups were already diverging before the intervention. Post-treatment coefficients (lags) show how the effect evolves: does it appear immediately, build over time, or dissipate? Seeing the full dynamic profile is much more informative than a single DiD number. Most good applied papers now include an event study plot alongside the main DiD estimate.

## Event study regression

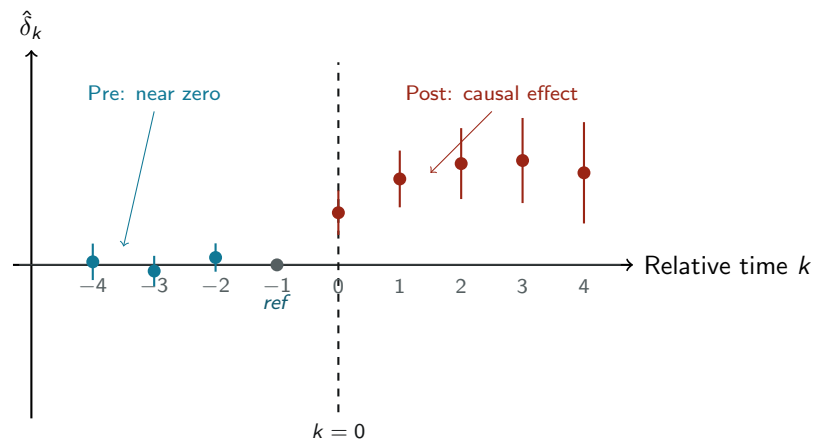
$$y_{it} = \alpha_i + \gamma_t + \sum_{k \neq -1} \delta_k \cdot \mathbf{1}[t - T_i^* = k] + \varepsilon_{it}$$

- $T_i^*$ : treatment date for unit  $i$
- $k = t - T_i^*$ : time relative to treatment ( $k < 0$ : pre,  $k \geq 0$ : post)
- $k = -1$ : **omitted reference period** (normalization)
- **Pre-treatment** ( $k < 0$ ):  $\delta_k \approx 0$  if parallel trends holds
- **Post-treatment** ( $k \geq 0$ ): trace out dynamic causal effects
- Unit and time FE absorb levels;  $\delta_k$  captures relative changes

20/36

Walk through the notation carefully.  $T_i^*$  is the period in which unit  $i$  receives treatment. For unit  $i$  in period  $t$ ,  $k = t - T_i^*$  is the distance from treatment. If  $k = -3$ , we are three periods before treatment; if  $k = 2$ , we are two periods after. The omitted category  $k = -1$  normalizes everything relative to the period just before treatment. So all  $\delta_k$  coefficients measure “how much did outcomes change relative to the period just before treatment?” Pre-treatment  $\delta_k$  should be near zero (no pre-trend); post-treatment  $\delta_k$  should reflect the causal effect. Choosing  $k = -1$  as the reference is conventional and sets the comparison point cleanly just before the intervention.

## Event study: coefficient plot



21/36

This coefficient plot is the standard visualization for event studies. Walk through the four quadrants of interpretation: (1) Pre-period coefficients near zero — good, this supports parallel trends. (2) Post-period coefficients rising after treatment — the policy had a positive and growing effect. (3) Confidence intervals widening over time — normal, since we have fewer observations far from the treatment date. (4) Reference period at zero by construction. Students sometimes confuse the reference period (which is zero by definition) with a pre-trend test result. Emphasize: the pre-trend test comes from the *other* pre-period coefficients ( $k = -4, -3, -2$ ), not from the reference period itself. In the plot shown, the pre-trend test passes: coefficients are small and statistically indistinguishable from zero.

## Event study in R

- Using `feols()` with `i()` for interaction syntax:
 

```
feols(y ~ i(rel_time, treated, ref = -1)
      | unit + year, data = df, cluster = ~unit)
```
- `rel_time`: variable with  $k = t - T_i^*$  (relative time to treatment)
- `treated`: binary indicator for treated units
- `ref = -1`: omit  $k = -1$  as reference period
- Plot with `iplot()` (coefficient plot with CIs):
 

```
iplot(m_es, main = "Event Study")
```
- `fixest` makes event studies very easy to run and plot

22/36

Walk through the `feols()` syntax for event studies. The `i()` function in `fixest` is designed for exactly this kind of interaction: it creates indicators for each value of `rel_time`, interacted with `treated`, and omits the reference value. The output is a set of coefficients  $\hat{\delta}_k$  for each  $k \neq -1$ . The `iplot()` function then plots these automatically with 95% confidence intervals. The full workflow — compute relative time variable, run `feols` with `i()`, plot with `iplot()` — takes about 10 lines of R code. Students should practice this in the assignment. Note: `rel_time` must be created manually as `mutate(rel_time = year - treatment_year)` before running the model.

## Reading an event study plot

- **Pre-trend test:** are pre-treatment  $\hat{\delta}_k$  jointly near zero?
  - If yes: supports parallel trends (not proof, but reassuring)
  - If no: parallel trends likely violated — DiD is not credible
- **Dynamic effects:** what do post-treatment  $\hat{\delta}_k$  look like?
  - Immediate jump at  $k = 0$ : effect appears right away
  - Gradual build-up: effect grows over time (lagged adjustment)
  - Decay: initial effect fades (people adapt, spillovers, attrition)
- **Symmetry:** sometimes useful to check behavior far pre-treatment

23/36

This frame provides the interpretive toolkit for reading event study plots. The pre-trend test is the most important: if the lines are bouncing around before treatment, it signals the two groups had different trends even before the policy, which undermines the entire DiD logic. The dynamic profile is also informative: labor market effects might take quarters to materialize (as firms adjust their workforce slowly); political effects might appear immediately (as people update beliefs); health effects might grow over time (as diseases progress or health habits change). Understanding the expected dynamic profile from theory helps assess whether the event study pattern is plausible.

## Roadmap

From Fixed Effects to DiD

The DiD Estimator

Parallel Trends and Its Threats

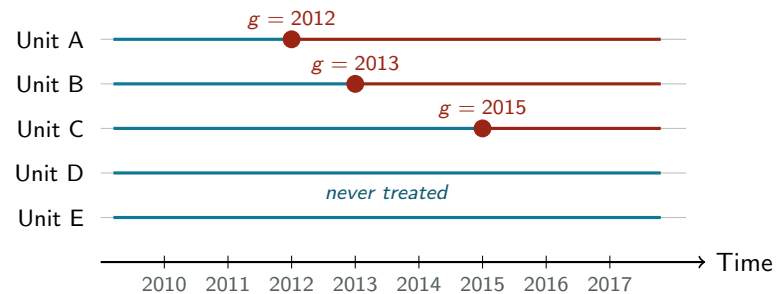
Event Studies

Staggered DiD and Recent Advances

Wrap-up

24/36

## Staggered treatment adoption



25/36

Staggered treatment is the most common pattern in observational policy research. Think of US states adopting right-to-carry laws in different years, countries implementing family leave policies at different times, or municipalities rolling out broadband internet in different waves. Units A, B, C adopt the policy at different times (their treatment cohort  $g$  varies). Units D and E never adopt. The key insight is that in this setting we have multiple  $2 \times 2$  DiDs: early vs. late adopters, early vs. never treated, late vs. never treated. The question is how to aggregate these into a single treatment effect estimate, and whether naive TWFE does so correctly.

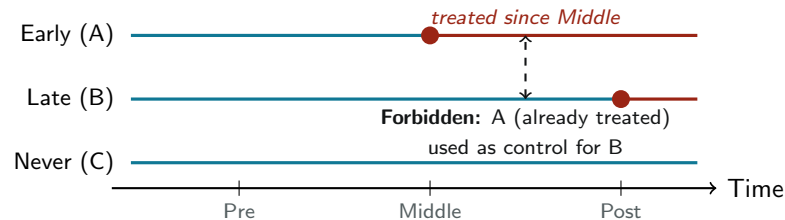
## The problem with naive TWFE in staggered settings

- With staggered timing, TWFE uses **all available  $2 \times 2$  comparisons**
- Including: already-treated units as the **control group** for later-treated units
  - Unit adopted in 2012 becomes "control" for unit adopted in 2015
- This is the "**forbidden comparison**" problem
  - Already-treated units are not a clean control
  - If treatment effects vary over time, their included effect contaminates the estimate
- Result: TWFE can produce a **weighted average with negative weights**
  - Estimate may not correspond to any valid ATT

26/36

This is the key conceptual advance from recent econometrics work (Goodman-Bacon 2021, Callaway-Sant'Anna 2021, de Chaisemartin-D'Haultfoeuille 2020, Sun-Abraham 2021, Borusyak et al. 2024). The problem: in staggered settings, TWFE implicitly treats already-treated units as part of the control group for later-treated units. If treatment effects are heterogeneous (different cohorts have different effects, or effects grow over time), then the already-treated units' changing treatment effects contaminate the estimate. In the extreme, if early-treated units have large effects, TWFE can subtract those effects from the comparison and produce negative weights on some cohorts — making the overall estimate meaningless or even reversed in sign. Goodman-Bacon (2021) showed that the TWFE estimate is a weighted average of all  $2 \times 2$  DiDs, with weights that can be negative.

## Negative weights: the intuition



27/36

This diagram makes the forbidden comparison concrete. Consider the Middle-to-Post window: unit B is being treated in the Post period, so the natural DiD comparison is Middle-to-Post for B vs. C (never treated). But TWFE also uses A vs. B in this window, treating A as the control — even though A has already been treated since the Middle period and its outcome is changing due to the ongoing treatment effect. If A's treatment effect is growing from Middle to Post, then using A as the control produces a negative contribution to the B estimate. The fix is to restrict comparisons to “clean” control groups: never-treated or not-yet-treated units.

## Modern alternatives: Callaway-Sant'Anna (2021)

- Key idea: estimate **group-time ATTs**  $ATT(g, t)$  separately
  - $g$ : treatment cohort (when the unit was first treated)
  - $t$ : calendar time
  - Compare cohort  $g$  to clean controls (never-treated or not-yet-treated)
- Then **aggregate**  $ATT(g, t)$  as desired:
  - Overall average ATT
  - ATT by event-time (dynamic/event study)
  - ATT by calendar time
- Avoids forbidden comparisons entirely
- Works even with **heterogeneous treatment effects**

28/36

Callaway and Sant'Anna (2021) is now the standard reference for staggered DiD. The key innovation is to go back to basics: estimate a separate DiD for every (treatment cohort, calendar time) pair, using only clean controls. These group-time average treatment effects  $ATT(g, t)$  are then aggregated using researcher-specified weights. This modular approach is more transparent and honest about what comparisons are being made. The aggregation step is flexible: you can get an overall average, a dynamic (event-time) profile, or a period-by-period estimate. The cost is slightly less precision (fewer observations in each cell), but the benefit is that the estimates are interpretable under much weaker assumptions.

## Modern alternatives: ETWFE (Wooldridge 2021)

- **Extended TWFE**: stays within the regression framework
- Key idea: interact treatment with cohort dummies and time dummies
  - Each cohort gets its own treatment effect profile
  - Avoids the averaging-over-cohorts problem of naive TWFE
- Estimating a richer model, not a simpler one
- Same identifying assumptions as Callaway-Sant'Anna
- Easier to implement in standard regression software
- In R: `etwfe` package by Grant McDermott

29/36

ETWFE (Wooldridge 2021, implemented in R by McDermott) is a practical alternative to Callaway-Sant'Anna that stays closer to the regression-based TWFE framework that researchers are used to. Instead of running a single treated dummy in TWFE, ETWFE interacts the treatment indicator with cohort fixed effects and time fixed effects, allowing each cohort to have a different treatment effect at each time period. The estimates are then marginal effects (average treatment effects on the treated), aggregated appropriately. The advantage is simplicity: it runs as a single regression with standard post-estimation commands. Students who are comfortable with `feols` will find `etwfe` intuitive.

## In R: did and etwfe packages

- **Callaway-Sant'Anna** via `did` package:

```
library(did)
cs = att_gt(yname = "y", gname = "g", idname = "id",
           tname = "year", data = df)
aggte(cs, type = "simple") # overall ATT
aggte(cs, type = "dynamic") # event-study ATT
```

- **ETWFE** via `etwfe` package:

```
library(etwfe)
m = etwfe(y ~ 1, tvar = year, gvar = g,
          data = df, vcov = ~id)
emfx(m) # marginal effects = ATT
```

30/36

Walk through the syntax of both packages. For `att_gt()`: `gname` is the variable containing the treatment cohort (year of first treatment, 0 for never treated); `tname` is the calendar time variable; `idname` is the unit identifier. The output is a large set of  $ATT(g, t)$  estimates which are then passed to `aggte()` for aggregation. `type = "simple"` gives a single overall ATT; `type = "dynamic"` gives an event-study plot. For `etwfe()`: `tvar` is calendar time, `gvar` is the cohort variable, and `vcov = ~id` clusters SEs by unit. The formula `y ~ 1` indicates no additional controls (add them as `y ~ x1 + x2`). After running `etwfe()`, use `emfx()` to compute marginal effects.

## Naive TWFE vs. modern estimators

	Naive TWFE	Callaway-Sant'Anna	ETWFE
Clean control groups	No	Yes	Yes
Heterogeneous effects	Biased	Handles	Handles
Negative weights	Possible	None	None
Pre-trend test	Via event study	Built-in	Via <code>emfx</code>
Implementation	<code>feols</code>	<code>did</code>	<code>etwfe</code>

- With simultaneous treatment (all treated at same time): TWFE is fine
- With staggered adoption and homogeneous effects: TWFE is approximately fine
- With staggered adoption and **heterogeneous effects**: use CS or ETWFE

31/36

This summary table helps students decide which estimator to use. The key question is whether treatment timing is staggered (different units treated at different times) and whether treatment effects are likely heterogeneous across cohorts or over time. If treatment is simultaneous (everyone treated in the same period), naive TWFE is fine. If treatment is staggered but effects are homogeneous across cohorts, TWFE gives approximately correct results. The problem only arises when both (a) treatment is staggered and (b) effects are heterogeneous. In practice, heterogeneity is the norm: early adopters often differ from late adopters in ways that affect how the policy operates. As a rule of thumb: in staggered settings, report both the naive TWFE and one of the robust estimators; if they agree, TWFE is fine; if they disagree, use the robust estimate.

## Roadmap

From Fixed Effects to DiD

The DiD Estimator

Parallel Trends and Its Threats

Event Studies

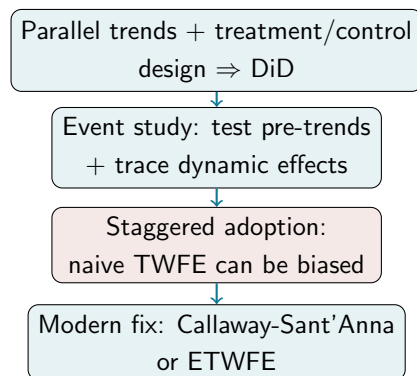
Staggered DiD and Recent Advances

Wrap-up

32/36



## DiD: putting it all together



33/36

Use this summary diagram to consolidate the narrative arc of the lecture. DiD adds a causal design layer on top of TWFE: treated vs. control groups, before vs. after the intervention. The parallel trends assumption is the identification engine. Event studies extend DiD to multiple periods, allowing both a pre-trend test and estimation of dynamic effects. The staggered DiD problem arises when units adopt treatment at different times — naive TWFE can be biased due to forbidden comparisons and negative weights. Modern estimators (Callaway-Sant'Anna, ETWFE) fix this by restricting comparisons to clean controls. Connect back to the broader course theme: good causal inference requires transparent assumptions and careful design, not just powerful methods.

## Key takeaways

- **DiD**: compare treated vs. control, before vs. after
  - Subtract the control trend as the counterfactual
- **Parallel trends**: the key identification assumption
  - Untestable in the post-period; probe with pre-trend tests
- **Event studies**: leads and lags around treatment
  - Pre-treatment  $\hat{\delta}_k \approx 0$ : supports parallel trends
- **Staggered DiD**: naive TWFE can fail with heterogeneous effects
  - Use Callaway-Sant'Anna (did) or ETWFE (etwfe)
- Always **cluster SEs** at the treatment-assignment level

34/36

This is the slide students should be able to reconstruct from memory. Five key points: (1) DiD logic: difference out the control trend. (2) Parallel trends: the core assumption. (3) Event studies: pre-trend test plus dynamic effects. (4) Staggered DiD: the modern problem and its solutions. (5) Clustering. Emphasize that these methods are now standard in top journals in political science, economics, sociology, and public health. Reading a recent applied paper in their field, students should be able to identify which of these elements are being used and evaluate their credibility.

## For next session

- Complete Assignment 6 (DiD application)
- Read the assigned paper (DiD design)
- Next session: Spatial Data (I)
  - Spatial data structures and visualization
  - Spatial autocorrelation
  - Spatial regression models

35/36

Preview the next session. Regression discontinuity is another quasi-experimental design that uses a cutoff rule to assign treatment. Like DiD, it relies on a specific identification assumption (no manipulation around the cutoff, continuity of potential outcomes). It is widely used in political science (electoral threshold studies, policy eligibility cutoffs). The key parallel with DiD: both designs exploit a specific feature of the data-generating process to construct a valid counterfactual. RDD uses spatial/score discontinuities; DiD uses temporal discontinuities combined with a control group.

Questions?

36/36