Natural Experiments - Part 2

Professor Ji-Woong Chung
Korea University

Outline

Difference-in-difference continued

- When additional controls are appropriate

How to handle multiple events

- Why they are useful
- Simple estimation approach & its problems
- Better ways to handle multiple events

Falsification tests

Triple difference

- How to estimate & interpret it
- Subsample approach

Outline

Difference-in-difference continued
- When additional controls are appropriate

How to handle multiple events
- Why they are useful
- Simple estimation approach & its problems
- Better ways to handle multiple events

Falsification tests

Triple difference
- How to estimate & interpret it
- Subsample approach

Why the Regression is Helpful

	Post (1)	Pre (2)	Diff (1) − (2)
Treatment (a)	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_0 + \beta_2$	$\beta_1 + \beta_3$
Control (b)	$\beta_0 + \beta_1$	β_0	β_1
Diff. $(a) - (b)$	$\beta_2 + \beta_3$	β_2	β_3

- Some papers report this simple two-by-two table as their estimate.
- Advantages of using regression:
  - Can modify to test timing of treatment. (will discuss later)
  - Allows adding additional controls, X.

Outline

Difference-in-difference continued
- When additional controls are appropriate

How to handle multiple events
- Why they are useful
- Simple estimation approach & its problems
- Better ways to handle multiple events

Falsification tests

Triple difference
- How to estimate & interpret it
- Subsample approach

Adding Controls to Difference-in-Difference

- Easy to add controls in regression:

$$y_{it} = \beta_0 + \beta_1 p_t + \beta_2 d_i + \beta_3 (d_i \times p_t) + X\Gamma + u_{it}$$

- Important to avoid adding controls affected by treatment (Angrist-Pischke term this "bad control").
  - You won't be able to get a consistent estimate of β_3 from estimating the equation

Example of a Bad Control

A policy is introduced at time $t = 1$ for a "treated" group; another group remains untreated. We observe the outcome y_{it} in periods $t = 0$ (pre) and $t = 1$ (post). We also observe a covariate X_{it} measured at post-treatment (so it can be affected by treatment).

Unit	Outcome y_{it} $t = 0$	$t = 1$	Covariate $X_{i,\,t=1}$ (Post)
Treated (i=1)	10	15	20
Control (i=2)	12	14	18

True treatment effect: $(15 - 10) - (14 - 12) = 3$.

Suppose we control for post-treatment covariate $X_{i,1}$ in the regression:

$$y_{it} = \beta_0 + \beta_1 p_t + \beta_2 d_i + \beta_3 (d_i \times p_t) \; + \; \gamma X_{i1} + u_{it}.$$

Because X_{i1} is affected by the treatment (for treated: X=20 vs control X=18), controlling for it will "soak up" some of the treatment's effect and bias β_3.

Result: Estimated β_3 will be less than the true 3 (say ≈ 2) because including X_{i1} removes part of the effect channel from treatment $\rightarrow X \rightarrow y$.

When Controls are Appropriate

- Two main reasons to add controls:
  - Improve precision by reducing standard errors.
  - Restore "random" assignment of treatment.

#1 – To Improve Precision via Controls

- When treatment is randomized, adding baseline covariates that predict the outcome can reduce residual variation ("noise"). That means you get more precise estimates (smaller standard errors).
- **Should adding controls change the treatment effect estimate?**
  - If treatment is truly random and the only difference between treatment and control is the treatment itself, then in expectation the coefficient on the treatment should remain the same regardless of additional controls.
  - The added controls should reduce variance, not introduce bias — they help precision, not identification.
- **When might the estimate change when you add controls?**
  - If treatment isn't fully random, the controls may absorb part of the bias.
  - Or you might have included a "bad control" — a variable that itself is influenced by the treatment. Controlling for such a variable can distort the treatment effect estimate.
- **Practical advice:**
  - Present results both without and with control.
  - Check that added controls are pre-treatment (ex-ante) and not affected by the treatment.
  - If the treatment effect changes substantially when adding controls, investigate why: might be imbalance, omitted confounding, or bad controls.

Example – Improving Precision

- Suppose you have firm-level panel data.
- Some natural experiment "treats" some firms but not others.
  - Could estimate standard difference-in-differences.

$$y_{it} = \beta_0 + \beta_1 p_t + \beta_2 d_i + \beta_3 (d_i \times p_t) + u_{it}$$

  - Could add fixed effects (like firm and year FE) for more precise estimate.
    - p_t is collinear with year FE (doesn't vary across firms).
    - d_i is collinear with firm FE (doesn't vary across time for each firm).
- So, you should estimate:

$$y_{it} = \beta_0 + \beta_3 (d_i \times p_t) + \alpha_i + \delta_t + u_{it}$$

  - α_i: control for treatment
  - δ_t: control for post-treatment

Generalized Difference-in-Differences

- Advantage: can improve precision and provide better model fit.
- Instead of assuming all firms in a group share the same baseline y, firm fixed effects let each firm have its own starting level.
- Instead of assuming the same pre/post change for everyone, time fixed effects allow each year to have its own mean outcome.

#2 – Restoring Randomness of Treatment

- In observational settings, treatment is often **not randomly assigned**.
  - Example: Firms with high x (e.g., size, leverage, productivity) may be *more likely* to receive the treatment.
- If the same x also predicts trends in y (e.g., outcomes grow faster for large firms), then treatment and outcome trends are **confounded**.

  $\Rightarrow$ Differences in y may reflect differences in x, not causal impact of treatment.
- Including x as a control can **restore conditional randomness:**

$$D_i \perp u_i \mid x_i$$

  That is, once we control for x, treatment assignment behaves "as if random."
- Then, comparing treated and untreated firms **with similar** x yields a valid estimate of the treatment effect.

Example – Restoring Randomness

- A new regulation affects some firms but not others.
  - The regulation is meant to be random, but in practice, **larger firms** are more likely to be affected.
  - Firm size (x) also influences y (e.g., profitability or compliance cost trends).
  - Firm size is determined *before* the regulation and is not affected by it (a **pre-treatment variable**).
- If these statements hold:
  - treatment depends on firm size,
  - firm size affects y, and
  - firm size is exogenous to the treatment,

  then adding firm size as a control:

$$y_{it} = \alpha + \beta D_i + \gamma \text{Size}_i + u_i$$

  helps "**restore randomness**" — treatment is random *conditional on size*.
- This correction isolates the **within-size variation** in treatment, removing confounding bias.

Controls Continued...

- But suppose firm size *itself* changes as a result of the new regulation:
  - Then current size is a **post-treatment variable**.
  - Controlling for it would remove part of the treatment's effect on y — a classic **bad control problem**.
- **Alternative:** Use only **pre-treatment size** and allow its effect to differ after treatment:

$$y_{it} = \alpha_i + \lambda_t + \beta(D_i \times Post_t) + \gamma \text{Size}_{i,\,\text{pre}} + \delta(\text{Size}_{i,\,\text{pre}} \times Post_t) + u_{it}$$

- This does two things:
  - Controls for **selection** — treatment may depend on pre-treatment size.
  - Controls for **differential trends** — large vs. small firms may follow different post-treatment paths even without treatment.
- Because $\text{Size}_{i,\,\text{pre}}$ is measured *before treatment*, it's not affected by treatment, avoiding endogeneity.

Restoring Randomness – Caution!

- **Key idea:** Adding controls can, in principle, make treatment "as-if random" once we condition on them.
- **In practice, this is rarely convincing.** Because:
  - We usually cannot verify that all sources of non-random assignment are captured by observables.
  - If treatment depends on both observable and unobservable factors (e.g., firm quality, management skill), controlling for observables alone won't restore true randomness.
  - The identifying assumption becomes:

$$E[u_{it} \mid D_i, x_i] = E[u_{it} \mid x_i],$$

which is strong and often implausible without very specific institutional knowledge.

Restoring Randomness – Caution!

- **When might it be credible?**
  - When the source of non-randomness is narrow and well-understood — for example, treatment depends only on one known variable (like firm size or a cutoff rule).
  - **Regression Discontinuity (RD)** is the classic example:
    - Assignment is non-random overall (higher $x \Rightarrow$ more likely treated),
    - But *around a threshold*, treatment is "as if random" once you control finely for x.
    - Nearby units differ only by whether they fall just above or below the cutoff.
- "Restoring randomness" by adding controls is theoretically neat but empirically fragile—plausible only when the non-randomness mechanism is simple and observable.

Be Careful About Standard Errors

- When you have **multiple pre- and post-treatment periods** (panel data), standard OLS formulas assume that all residuals are independent.
- But in DiD settings, that assumption is usually **violated**:
  - Errors for the same firm (or region, or unit) are often serially correlated over time.
  - If you treat each time–unit observation as independent, your standard errors will be far too small $\Rightarrow$ you'll overstate statistical significance.

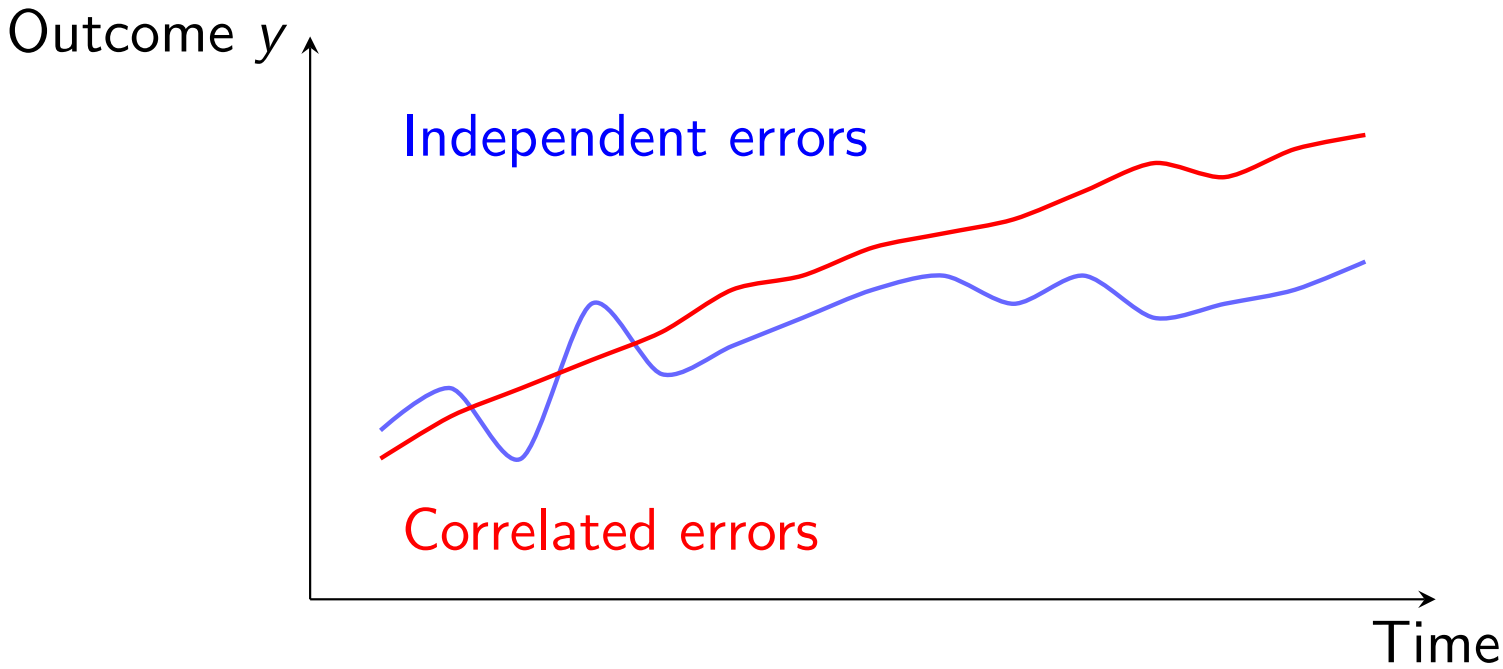
Be Careful About Standard Errors

- **Two standard fixes:**
  1. **Cluster standard errors at the unit level:** Allows arbitrary correlation of residuals over time within each firm (but assumes independence across firms).

$$\widehat{V}(\hat{\beta}) = (X'X)^{-1}\left(\sum_i X_i'\hat{u}_i\hat{u}_i'X_i\right)(X'X)^{-1}$$

  2. **Collapse to one observation per unit:** Compute pre- vs. post-treatment means for each unit, then estimate DiD on these means. This makes the data cross-sectional and avoids serial correlation altogether.
- Both methods give you correct inference for the DiD coefficient $\hat{\beta}_3$, but the clustered SE approach retains all data and is usually preferred.

Serial Correlation Bias – Visual Intuition



- Without clustering, we treat the red series as if each point were independent—even though they move together over time—leading to under-estimated SEs and inflated t-statistics.

Outline

Difference-in-difference continued
- When additional controls are appropriate

How to handle multiple events
- Why they are useful
- Simple estimation approach & its problems
- Better ways to handle multiple events

Falsification tests

Triple difference
- How to estimate & interpret it
- Subsample approach

Outline

Difference-in-difference continued
- When additional controls are appropriate

How to handle multiple events
- Why they are useful
- Simple estimation approach & its problems
- Better ways to handle multiple events

Falsification tests

Triple difference
- How to estimate & interpret it
- Subsample approach

Multiple Treatment Events

- In many real-world settings, a policy or regulation isn't adopted everywhere at once:
  - Instead, it's implemented in **different places at different times**.
- These are called **staggered adoption** or **multiple-event DiD** designs.
- Conceptually, each adoption acts like a small "natural experiment":

$$\text{Group}_g \text{ treated at time } T_g$$

  and other groups that haven't yet been treated can serve as controls for that group at that time.
- Having many such events can make the DiD design more robust and empirically credible:
  - Provides repeated tests of the same hypothesis across groups and times.
  - Helps average out idiosyncratic shocks that might bias a single-event study.

How Multiple Events Strengthen Identification

1. **Replication across time and groups:**
   - If similar treatment effects are observed across many adoption dates, it's less likely that results are driven by a single shock or coincidence.
2. **More demanding test of parallel trends:**
   - Each treated group can be compared to not-yet-treated groups in the same year.
   - For parallel trends to be violated, the bias would need to occur *every time* a group is treated — much less plausible.
3. **More efficient estimation:**
   - Each adoption contributes to the estimation of the treatment effect.
   - Increases statistical power and precision relative to a single event.
4. **Built-in falsification checks:**
   - You can examine pre-trends separately for each event cohort.
   - If pre-trends are flat for all cohorts, that's strong evidence of validity.

Multiple staggered events turn one DiD test into a series of "mini natural experiments" that reinforce each other.

Example

- Suppose 50 U.S. states adopt a new disclosure rule between 1995 and 2010. If you find that:
  - Firms' liquidity rises after adoption,
  - The pattern is consistent across states and adoption years,
  - And no pre-trends appear before any state's adoption,
- Then it's hard to believe that 50 unrelated shocks all just happened to make liquidity rise in exactly those years when each state adopted the rule.
- Caveat: Having many events doesn't automatically guarantee validity:
  - If adoption timing is systematically related to unobserved trends (e.g., states adopt after prior growth spurts), bias can still exist.
  - But with multiple, dispersed adoption dates, such coincidences must repeat many times, which makes the violation less plausible and easier to detect with pre-trend checks.

Outline

Difference-in-difference continued
- When additional controls are appropriate

How to handle multiple events
- Why they are useful
- Simple estimation approach & its problems
- Better ways to handle multiple events

Falsification tests

Triple difference
- How to estimate & interpret it
- Subsample approach

Estimation with Multiple Events

- When a treatment (e.g., new regulation) occurs in different years for different groups, we can extend DiD easily.
- The simplest implementation follows **Bertrand and Mullainathan (2003)**:
  - Stack all groups and periods together.
  - Include **time fixed effects** to absorb common shocks.
  - Include **cohort (group) fixed effects** to absorb permanent differences across groups.
  - Estimate one common treatment indicator that turns on when each group becomes treated.
- This yields a single $\hat{\beta}$ — an *average treatment effect across all events*.

Multiple Events [Part 1]

- The regression model:

$$y_{ict} = \beta d_{ict} + p_t + m_c + u_{ict}$$

  where
  - y_{ict} = outcome for unit i, cohort c, period t
  - d_{ict} = indicator for whether cohort c is treated by time t (i.e., treatment $\times$ post)
  - p_t = **time fixed effects**, capturing economy-wide shocks or macro trends
  - m_c = **cohort fixed effects**, capturing permanent differences across cohorts
- The index c identifies the **cohort** — a set of units treated in the same event.
  - Example: if California adopts a policy in 1999 and Texas in 2003, all California firms belong to the 1999 cohort and all Texas firms to the 2003 cohort.

Multiple Events [Part 2]

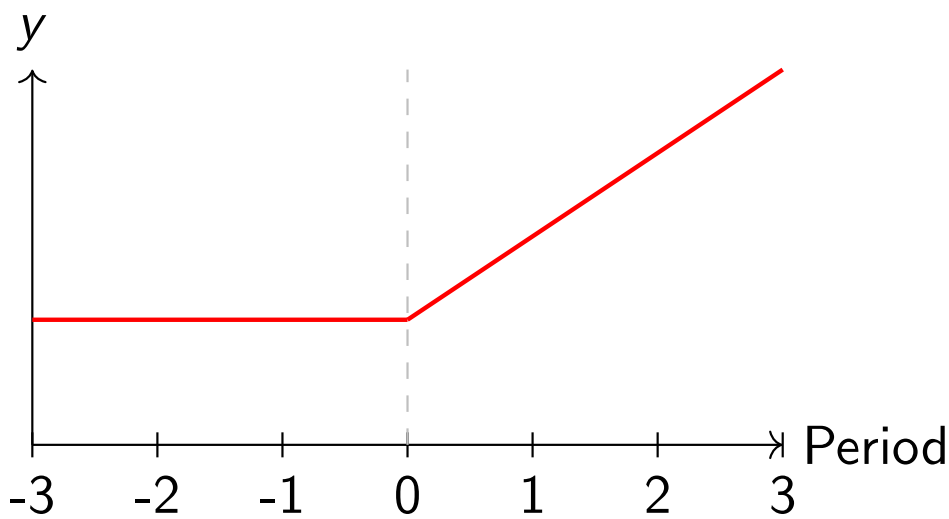
- **Intuition:**
  - In each year, the units that are not yet treated serve as controls for those that have just been treated.
  - Example: A firm treated in 1999 acts as a control for a firm treated in 2004 up until 1999, when the 1999 firm itself becomes treated.
- This approach effectively performs many DiD comparisons simultaneously:
  - "Treated–post" minus "control–post" within each period,
  - then averages those differences across all groups and times.
- The coefficient β represents the average treatment effect across all staggered events, assuming parallel trends hold for each adoption cohort.

A Big Potential Problem [Part 1]

- In the Bertrand–Mullainathan (two-way fixed-effects) setup, **earlier-treated groups are reused as controls** for later-treated groups.
  - Once a state, firm, or cohort is treated, it remains treated forever.
  - So in later periods, that group no longer provides a valid untreated counterfactual.
- Example: California (treated in 1999) is used as a control for Texas (treated in 2003) during 1995–2002, even though California has already begun responding to treatment after 1999.

A Big Potential Problem [Part 2]

- If treatment effects evolve over time — e.g., they grow or fade after treatment — this creates **dynamic treatment effects**.
- When already-treated units act as "controls" for newly-treated ones, their outcomes keep moving because of their own treatment, violating parallel trends.
- Example dynamic pattern:



- In this example, y keeps rising after treatment.
- When we later use these "treated" units as controls, their continued increase biases the estimated effect for later cohorts.

A Big Potential Problem [Part 3]

- Because already-treated groups contaminate the control pool, both the estimated effects *and* the pre-trend tests become biased.
- Even if you add event-study dummies, the "control" observations are not truly untreated, so the estimated pre-period coefficients need not be zero even when parallel trends holds in theory.
- The modern literature (e.g., Callaway–Sant'Anna 2021; Sun–Abraham 2021) proposes alternative estimators that handle these dynamic and heterogeneous effects correctly. See Baker, Cunningham, Goodman-Bacon, and Sant'Anna (2025) and Baker, Larcker, and Wang (JFE 2022) for a review.

Outline

Difference-in-difference continued
- When additional controls are appropriate

How to handle multiple events
- Why they are useful
- Simple estimation approach & its problems
- Better ways to handle multiple events

Falsification tests

Triple difference
- How to estimate & interpret it
- Subsample approach

The Goodman–Bacon Decomposition

- In a staggered-adoption setting, different groups are treated at different times.
- The traditional two-way fixed effects (TWFE) model:

$$y_{it} = \alpha_i + \lambda_t + \beta D_{it} + u_{it}$$

  implicitly averages many 2×2 DiDs:

  - Early vs. never-treated groups
  - Late vs. never-treated groups
  - Early vs. late groups (before and after the late group's adoption)

- Goodman–Bacon (2021, *J. Econometrics*) shows:

$$\hat{\beta}_{TWFE} = \sum_{(g,g')} w_{gg'} \, \widehat{DID}_{gg'},$$

  where each $\widehat{DID}_{gg'}$ is a 2×2 DiD between cohorts (g, g') and the weights $w_{gg'}$ sum to 1. Depend on treatment timing and group sizes.

What "Negative Weights" Mean

- Some comparisons receive **negative weights**, especially when **already-treated** units act as controls for later-treated units.
- The regression enforces orthogonality with fixed effects, effectively **subtracting** parts of those comparisons.
- Intuition:
  - Suppose early adopters' outcomes keep rising after treatment.
  - When they serve as controls for later adopters, this ongoing rise is misattributed to the later treatment — creating a negative contribution.
  - In extreme cases, $\hat{\beta}$ can even flip sign.

Why the Regression "Subtracts" Some Comparisons

- In TWFE:

$$y_{it} = \alpha_i + \lambda_t + \beta D_{it} + u_{it},$$

  we first remove each unit's mean (α_i) and each period's mean (λ_t) before estimating β.

- With staggered adoption:
  - Early adopters have higher average D_{it} (treated longer),
  - Late adopters have lower averages,
  - Demeaning causes some treated observations to have negative residuals $\tilde{D}_{it}$.
- To maintain orthogonality, OLS "balances" these deviations — giving certain group–time cells negative weight.

Intuition: After removing fixed effects, the regression no longer compares treated vs. untreated means—it compares deviations from each group's and year's averages, which can flip signs.

Double Demeaning Makes Some $\tilde{D}_{it}$ Negative

Two units (A=early, B=late), three periods ($t = 0, 1, 2$).
Treatment paths: A: $(0, 1, 1)$; B: $(0, 0, 1)$.

	t=0	t=1	t=2
A (early)	0	1	1
B (late)	0	0	1

Unit means: $\bar{D}_A = \frac{2}{3}$, $\bar{D}_B = \frac{1}{3}$; Time means: $\bar{D}_0 = 0$, $\bar{D}_1 = \frac{1}{2}$, $\bar{D}_2 = 1$; Grand mean: $\bar{D} = \frac{1}{2}$.

$$\tilde{D}_{it} = D_{it} - \bar{D}_i - \bar{D}_t + \bar{D} \quad \Rightarrow$$

	t=0	t=1	t=2
A	$-\frac{1}{6}$	$\frac{1}{3}$	$-\frac{1}{6}$
B	$\frac{1}{6}$	$-\frac{1}{3}$	$\frac{1}{6}$

Note: A is treated at t=2, yet $\tilde{D}_{A,2} < 0$. After removing unit/time means, some treated cells have negative regressor values—the algebraic source of "subtraction."

How Subtraction Appears in OLS: $\sum \tilde{D}_{it}\tilde{y}_{it}$

Let outcomes show dynamic effects:

$$y_{A,t} = (0, 1, 2), \qquad y_{B,t} = (0, 0, 1).$$

After double demeaning, $\tilde{y}_{it} = y_{it} - \bar{y}_i - \bar{y}_t + \bar{y}$ yields:

	t=0	t=1	t=2
A	$-\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$
B	$\frac{1}{3}$	$-\frac{1}{6}$	$-\frac{1}{6}$

Contribution to numerator:

$\sum \tilde{D}_{it}\tilde{y}_{it}$:

	t=0	t=1	t=2
A: $\tilde{D}\tilde{y}$	$+\frac{1}{18}$	$+\frac{1}{18}$	$-\frac{1}{36}$
B: $\tilde{D}\tilde{y}$	$+\frac{1}{18}$	$+\frac{1}{18}$	$-\frac{1}{36}$

Cells like $(A, t{=}2)$ are treated yet contribute negatively—*subtracting* from $\hat{\beta}$. This is the finite-sample manifestation of negative weights in staggered TWFE.

Callaway & Sant'Anna (2021): Setup

- Problem: Standard DiD or TWFE fails when you have multiple time periods and staggered treatment timing.
- Define:
  - $G_i = g$: the time when unit i first receives treatment (if ever).
  - $Y_{it}(0)$, $Y_{it}(1)$: potential outcomes without/with treatment for unit i at time t.
- Key parameter: the *cohort–time average treatment effect*

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

- Compared to TWFE: we estimate separate effects for each cohort g and time t, and then aggregate.

Callaway & Sant'Anna (2021): Identification

- Key assumptions:
  - No anticipation: $Y_{it}(0)$ unchanged just before treatment at $t = g$.
  - Parallel trends (for each cohort g):

$$E\left[Y_t(0) - Y_{t-1}(0) \mid G_i = g\right] = E\left[Y_t(0) - Y_{t-1}(0) \mid \text{controls still untreated at } t\right]$$

- With these assumptions, for each (g, t), one can estimate $ATT(g, t)$.
- Aggregation: Once all $ATT(g, t)$ are computed, you can build summary parameters like

$$\theta_{\text{overall}} = \sum_{g,t} w_{g,t}\, ATT(g, t),$$

where weights $w_{g,t}$ might reflect cohort sizes, exposure times, etc.

Callaway & Sant'Anna (2021): Simpled Example

Three units A, B, C. Two time periods $t = 1, 2$.

Unit	$t = 1$	$t = 2$
$A\,(g = 2)$	10	**15**
$B\,(g = 2)$	8	**13**
$C\,(never)$	12	12

Bold = treated outcome for units treated at $t = 2$. Control group = C (never treated).

Compute for cohort $g = 2$, time $t = 2$:

$$ATT(2,2) = \big((15 - 10) - (12 - 12)\big) = \boxed{5}.$$

Interpretation: The estimated effect for cohort treated at time 2 is +5.
If there were other cohorts/time combinations, we'd compute each $ATT(g, t)$ and then aggregate.

Callaway & Sant'Anna (2021): Why Use It?

- Avoids contamination from previously treated units acting as controls (a major issue in TWFE with staggered adoption).
- Allows treatment effect heterogeneity across cohorts and over time (different g and t values).
- Implemented in R package `did` and Stata command `csdid`.
- When properly applied, yields more credible causal estimates than naïve TWFE in many staggered-treatment settings.

Stacked Regression Approach – Alternative

- For each event cohort c:
  - Restrict to a symmetric pre- and post-window (e.g., 5 years before and after).
  - Drop any unit that receives another treatment within that window (keeps controls clean).
  - Controls are units not yet treated by the event's date.
- Stack the sub-samples (one per event) into one dataset and label each observation by cohort c.
  - Example: A firm may serve as a control in cohort 1999 but later as a treated unit in cohort 2005.
- Estimate the pooled regression:

$$y_{i,c,t} = \beta\, d_{i,c,t} + \alpha_{i,c} + \delta_{t,c} + u_{i,c,t},$$

  where:
  - $\alpha_{i,c}$: unit×cohort fixed effects;
  - $\delta_{t,c}$: time×cohort fixed effects;
  - Optionally include γ_c: cohort FE capturing across-event differences.
- Cluster SEs at the unit (or unit×cohort) level.
- β represents the average treatment effect across events.

Stacked Regression Approach – Not a Panacea

- **Advantages:**
  - Avoids using already-treated units as controls $\rightarrow$ mitigates bias from dynamic effects.
  - Enables focused, interpretable event-window analysis.
  - Flexible: extendable to triple-difference or event-study designs.
- **Caveats:**
  - Requires careful and consistent window choice (pre/post lengths affect weights).
  - Drops later-treated units $\rightarrow$ smaller samples and reduced power.
  - Assumes parallel trends *within each event window*, not across all cohorts.
  - The pooled β still averages heterogeneous effects.
  - Even stacked regressions can mis-weight cohorts when window lengths or variances differ (Wing et al., 2024).

Outline

Difference-in-difference continued
- When additional controls are appropriate

How to handle multiple events
- Why they are useful
- Simple estimation approach & its problems
- Better ways to handle multiple events

Falsification tests

Triple difference
- How to estimate & interpret it
- Subsample approach

Falsification Tests for Difference-in-Differences

Because the key identification assumption (parallel trends) cannot be directly tested, we rely on supporting evidence via falsification tests.

1. Compare pre-treatment observables — do treated and control groups look similar before policy?
2. Check timing of change / pre-trends — did outcomes diverge before treatment?
3. Treatment reversal — if treatment is undone, does effect reverse?
4. Placebo outcomes — test variables that should not be affected by treatment.
5. Triple-difference — use a third dimension where effect should vary, providing extra cross-check.

#1 Pre-treatment Comparison (Part 1)

- Idea: if treatment is as good as random, then treated and control groups should be similar in their characteristics prior to treatment.
- Show tables or graphs of observable covariates (e.g., pre-period means): helps assess balance.
- If large differences exist, it raises concern: why might treated units be different, and could that drive different trends in y?

#1 Pre-treatment Comparison (Part 2)

- If you find a difference in some variable z, does DiD fail?
  - Not necessarily — the question is whether z predicts divergent trends in y independent of treatment.
  - You may control for z (and its interaction with time) to mitigate concern.[1]
- But a key lingering concern: **unobservables**. Differences in observed covariates suggest possible differences in unobservables that affect trends in y.

[1] See Callaway & Sant'Anna (2021) and Caetano et al. (2022) for time-varying covariate extensions.

If you find a difference in some variable z . . .

- Suppose you observe that treated units differ from controls in a covariate z_i (e.g., firm size, prior sales) in the pre-treatment period.
- You can adjust your model:

$$y_{it} = \alpha_i + \lambda_t + \beta\, D_{it} + \gamma\, z_i + \delta\,(z_i \times \text{Post}_t) + u_{it}$$

where $\text{Post}_t = 1$ in the post-treatment period.

- **Example:**

Unit	z_i	$y_{i,\,t=-1} \rightarrow y_{i,\,t=0}$
A (treated)	1,000	$50 \rightarrow 55$
B (treated)	1,100	$52 \rightarrow 57$
C (control)	300	$20 \rightarrow 24$
D (control)	320	$19 \rightarrow 23$

Although z differs (1000 vs 300), the change in y from $t = -1$ to $t = 0$ is similar ($+5$) across groups $\rightarrow$ less concern that z drives differing pre-trends.

- If instead control units changed by $+2$ while treated changed by $+5$ pre-treatment, you would worry that z_i (or other omitted factors) are driving divergent trends $\rightarrow$ DiD may be biased.

#2 Check for Pre-trend (Part 1)

- One of the strongest diagnostics: allow the treatment effect to vary by time (leads and lags) and inspect the coefficients for pre-periods.
- Under parallel trends, one expects no systematic difference in trends for the treated group before the event.

$$y_{it} = \beta_0 + \beta_1 d_i + \beta_2 p_t + \sum_t \gamma_t \left(d_i \times \lambda_t\right) + u_{it}$$

- Here: d_i = indicator for treated unit, p_t = post-treatment period indicator, λ_t = dummy for each period relative to the event (e.g., $t = -2, -1, 0, +1, \ldots$).
- The coefficients γ_t measure the difference in outcome for treated vs controls at each time t.

#2 Check for Pre-trend (Part 2)

- Plot γ_t estimates with confidence intervals across lead and lag periods.
- If pre-treatment γ_t (for $t < 0$) are close to zero and not trending systematically, this supports parallel trends.
- But note: failing to reject pre-trend $\neq$ proof that parallel trends hold. (Roth 2022)

Why "No significant pre-trend" $\neq$ Proof of Parallel Trends

- Researchers often test for pre-treatment differences in trends (leads) to check the Parallel Trends Assumption:

$$H_0 : \gamma_t = 0 \ \ (\text{for } t < 0)$$

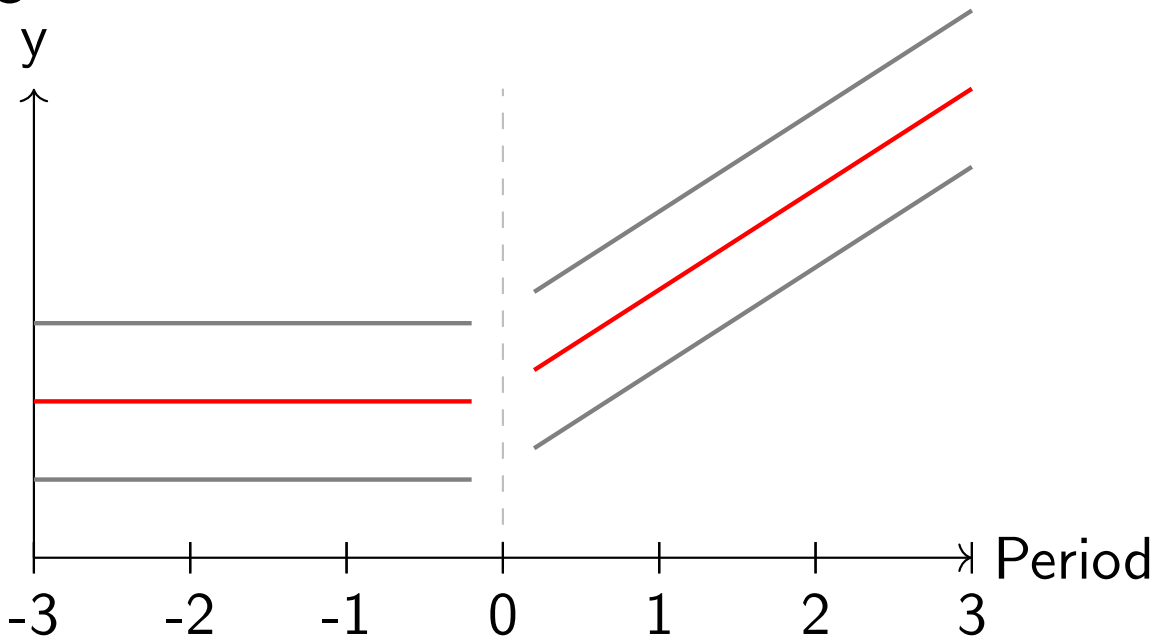
- Roth (2022) identifies two major limitations of this approach:
  1. **Low statistical power:** Conventional pre-trend tests may not detect meaningful violations of parallel trends (i.e., differences in trends that matter for bias) because the test lacks power.
  2. **Selection/conditioning bias:** If investigators proceed only when the pre-trend test "passes" (i.e., no significant difference), they may inadvertently bias their treatment effect estimates (confidence intervals may under-cover) because this conditioning distorts the sampling distribution.
- Failing to reject the null of no pre-trend is necessary for credible DiD design but not sufficient to guarantee the validity of the parallel trends assumption.

Practical Implications for Your DiD Design

- When graphing or estimating pre-treatment coefficients (γ_t) and they appear "flat" (no divergence), this is good—but it does not eliminate the possibility of undetected trend differences.
- **Recommended practices:**
  - Report the magnitude of the pre-trend coefficients (not just p-values), so readers can judge whether any drift is economically meaningful.
  - Compute power or minimal detectable effect (MDE) of your pre-trend test: ask "If there were a trend difference of size X, would we likely detect it?" The R package `pretrends` (Roth 2022) is designed for this.
  - Avoid proceeding conditionally on "pre-test passed" without adjustment, since this may worsen bias.
  - If you observe even small pre-trend drift (or are unsure about power), consider sensitivity methods / bounding approaches (e.g., HonestDiD) rather than relying solely on the standard DiD estimator.
- Use pre-trend tests as a diagnostic tool—helpful for assessing design credibility—but not as definitive proof that the identifying assumption holds.

#2 Check for pre-trend [Part 4]

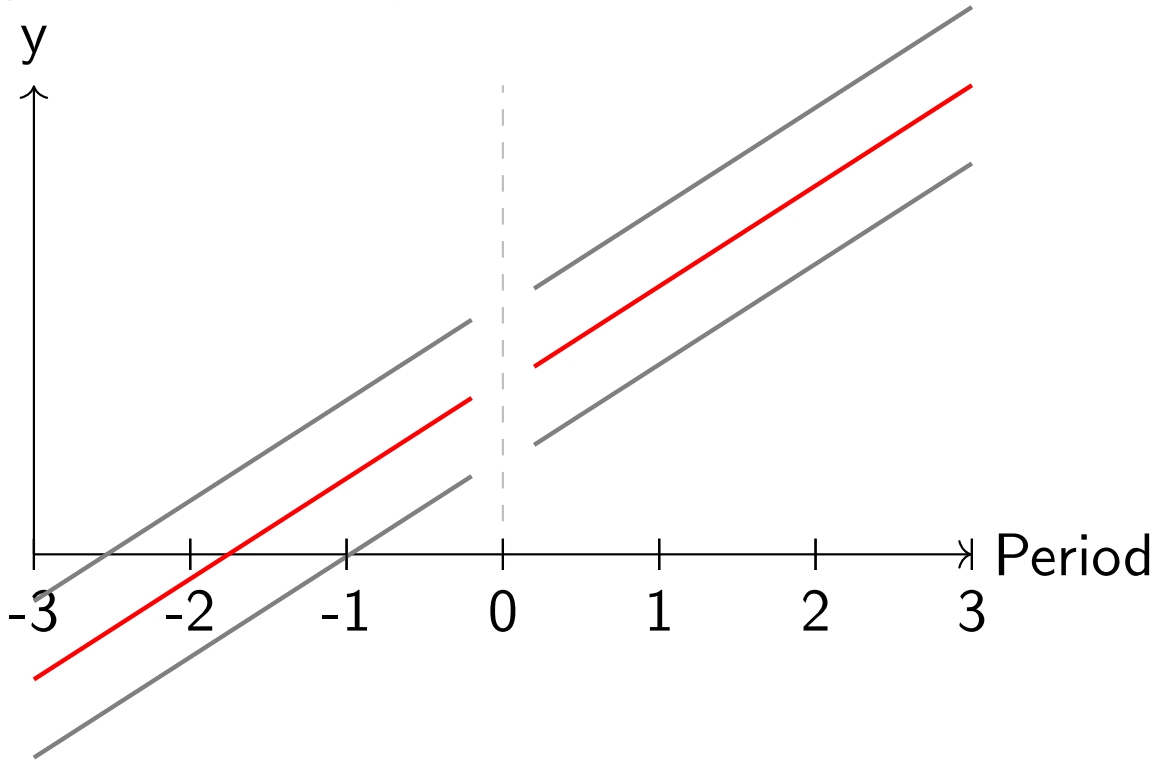
- Something like this is ideal



  - No differential pre-trend
  - Tight confidence intervals

#2 Check for pre-trend [Part 5]

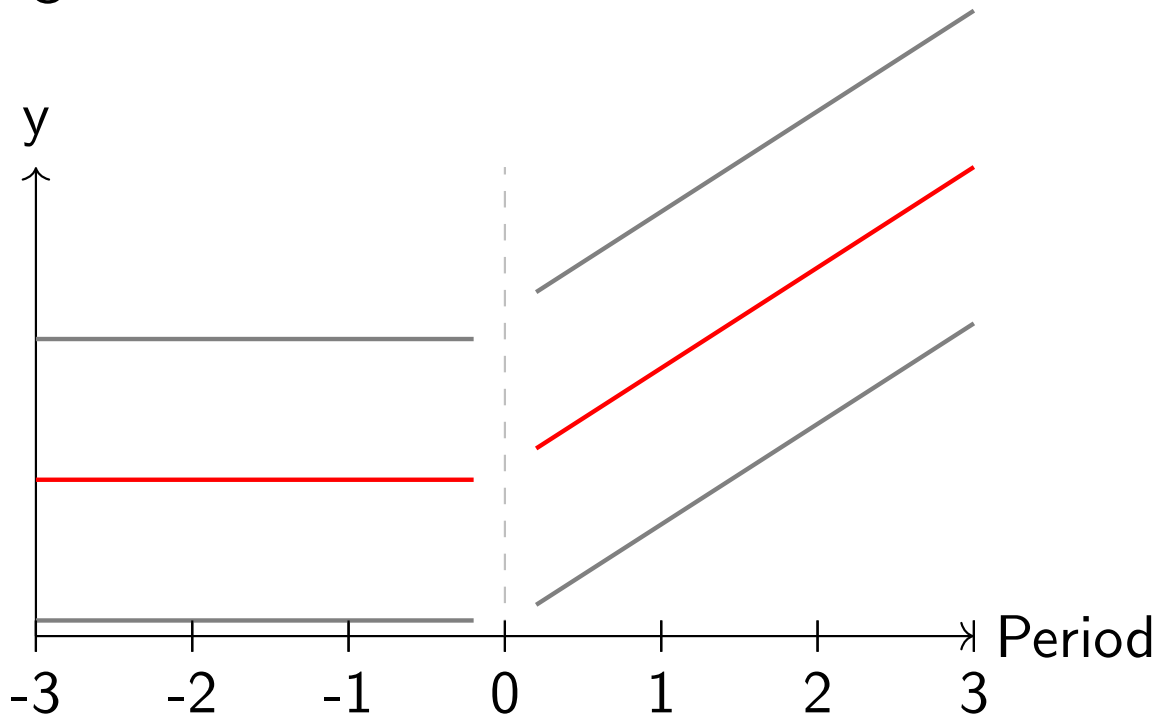
- Something like this is very bad



  - y for treated firms was already going up at faster rate prior to event!

#2 Check for pre-trend [Part 6]

- Should we make much of **wide** confidence intervals in these graphs? E.g.



- Answer: Not too much... Each period point estimate might be noisy; diff-in-diffs will tell us whether post-average y is significantly different then pre-average y

#2 Check for pre-trend [Part 7]

- Another type of pre-trend check done is to do the diff-in-diffs in some "random" pre-treatment to show no effect
- Caveats
  - It is subject to gaming; researcher might choose a particular pre-period to look at that works
  - Prior approach allows us to see what the timing was and determine whether it is plausible

#3 Treatment Reversal

- If treatment is reversed (e.g., regulation is removed), then we can test whether the outcome moves back toward control levels.
- This strengthens the causal story by demonstrating the treatment's temporal link to the outcome.
- Requires the reversal to be exogenous and clean — otherwise inference is weak.

#4 Placebo Outcomes / Unaffected Variables

- Identify variables that theory predicts should *not* be affected by the treatment.
- Run the DiD on these placebo outcomes — if you find large "effects," it raises concerns of residual confounding or misspecification.
- Example: If a labor regulation should only affect wages, check an outcome like "firm color change" that shouldn't be affected.

#5 Triple Difference

- Use when theory suggests treatment effect should differ by a third dimension (e.g., high-vs-low exposure subgroup).
- Model:

$$y_{it} = \cdots + \delta \left(d_i \times p_t \times s_i \right) + \ldots$$

  where s_i is subgroup indicator.
- The triple difference adds a further layer of variation:
  (post vs pre) $\times$ (treated vs control) $\times$ (sensitive vs less-sensitive).
- Helps isolate the effect and test consistency of theory.

Outline

Difference-in-difference continued
- When additional controls are appropriate

How to handle multiple events
- Why they are useful
- Simple estimation approach & its problems
- Better ways to handle multiple events

Falsification tests

Triple difference
- How to estimate & interpret it
- Subsample approach

Outline

Difference-in-difference continued
- When additional controls are appropriate

How to handle multiple events
- Why they are useful
- Simple estimation approach & its problems
- Better ways to handle multiple events

Falsification tests

Triple difference
- How to estimate & interpret it
- Subsample approach

Triple Difference (DDD) – Specification

$$y_{it} = \beta_0 + \beta_1 p_t + \beta_2 d_i + \beta_3 h_i + \beta_4 (p_t \times h_i) + \beta_5 (d_i \times h_i) + \beta_6 (p_t \times d_i) + \beta_7 (p_t \times d_i \times h_i) + u_{it}$$

- p_t: indicator for post-treatment period.
- d_i: indicator for treated group.
- h_i: indicator for high-sensitivity subgroup (effect modifier).
- β_6: DiD effect for low-sensitivity subgroup ($h_i = 0$).
- β_7: Additional effect for high-sensitivity subgroup ($h_i = 1$); total for high $= \beta_6 + \beta_7$.

Triple Difference – Combinations Table

		h_i	
		0 (low-sens)	1 (high-sens)
$d_i = 0$	Controls	pre/post	pre/post
$d_i = 1$	Treated	pre/post	pre/post

- There are $2 \times 2 \times 2 = 8$ cells (pre/post $\times$ treated/control $\times$ low/high).
- The full model with the eight coefficients (including constant) allows estimation of each cell's mean.
- The triple interaction ($p_t \times d_i \times h_i$) isolates how much more (or less) the treatment effect is for the high-sensitivity group compared to the low-sensitivity group.

Triple Difference – Example

You sponsor a job-training program at $t = 0$. You believe larger firms benefit more.

$$h_i = 1 \text{ if firm assets above median before } t = 0.$$

Estimated coefficients:

$$\beta_6 = 2, \quad \beta_7 = 3.$$

- For low-sensitivity firms ($h_i = 0$): effect $= \beta_6 = 2$.
- For high-sensitivity firms ($h_i = 1$): effect $= \beta_6 + \beta_7 = 2 + 3 = 5$.
- Interpretation: training yields $+2$ units for smaller firms, $+5$ units for larger firms (difference of 3 units tied to size).

Triple Difference – Key Assumption & Reference

- The DDD estimator hinges on a single parallel-trends-on-differences assumption: the difference between high vs low sensitivity groups' trends is the same in treated and control groups, in absence of treatment.
- According to Olden & Møen (2019): "The difference between two biased DiD estimators will be unbiased as long as the bias is the same in both estimators."
- Practical tips:
  - Define h_i using ex-ante characteristics (before treatment) to avoid "bad control" problems.
  - Use robust standard errors (cluster by unit or relevant group) given repeated observations.
  - Ensure you interpret β_7 as the differential effect for the subgroup; total effect for subgroup $= \beta_6 + \beta_7$.

Triple-Difference: Indicator h_i vs Continuous Moderator S_i

In a DDD model you often include a third dimension h_i (e.g., sensitivity subgroup).

Option A: Indicator moderator

$$h_i = \begin{cases} 1 & \text{if unit is “high sensitivity” (above median size)} \\ 0 & \text{otherwise} \end{cases}$$

Option B: Continuous moderator

$$S_i = (\text{ex-ante size, assets in \$M, etc.})$$

Which form to use?

- Indicator $h_i \rightarrow$ you estimate different treatment effects for two discrete groups.
- Continuous $S_i \rightarrow$ you estimate how the treatment effect changes with the moderator value (slope).

Continuous S_i vs Indicator h_i: Advantages & Trade-Offs

Advantages of continuous moderator S_i:

- Uses more variation: avoids arbitrary cut-off, retains full information.
- Enables estimating a *dose-response* style slope: e.g., "for each \$10 M increase in assets, treatment effect rises by X."

Disadvantages / cautions:

- Imposes functional form (typically linear) between S_i and treatment effect; if the true effect is non-linear or threshold-based, results may mis-specify.
- More sensitive to outliers: very large values of S_i may drive the interaction estimate.
- Interpretation becomes more complex: you often need to compute marginal treatment effects at different S_i values (e.g., 10th percentile, median, 90th percentile).

Practical guidance:

- If theory suggests a smooth effect of the moderator (e.g., size matters linearly), continuous might work better.
- If you believe there is a threshold (e.g., only "very large firms" benefit) or few data at extremes, indicator may be safer.
- Always check with robustness: compare indicator specification to continuous, possibly test for non-linearity (e.g., quartile splines).

Example: Continuous Moderator in DDD

Firms receive a subsidy at time $t = 0$. Hypothesis: larger firms (higher assets) gain more.

- Let S_i = firm assets (in \$10 M) measured just before $t = 0$.
- Model:

$$y_{it} = \alpha_i + \delta_t + \beta_d(p_t \times d_i) + \beta_s(p_t \times d_i \times S_i) + u_{it}$$

- Suppose estimates: $\hat{\beta}_d = 3$, $\hat{\beta}_s = 0.5$.
- Interpretation: Among treated firms, a firm with $S_i = 2$ (assets \$20 M) sees effect: $3 + 0.5 \times 2 = 4$. A firm with $S_i = 5$ (assets \$50 M) sees effect: $3 + 0.5 \times 5 = 5.5$.

If you instead used an indicator $h_i = 1$ if assets $>$ \$30 M, you'd estimate one effect for "large" firms and another for "small" — but you'd lose nuance and might impose an arbitrary cut-off.

Generalized Triple-Difference (DDD) Specification

- As with standard DiD, we can add FE to soak up time-invariant unit heterogeneity and common time shocks — improving precision and controlling for omitted variables.

$$y_{it} = \beta_1 (p_t \times h_i) + \beta_2 (p_t \times d_i) + \beta_3 (p_t \times d_i \times h_i) + \alpha_i + \delta_t + u_{it}$$

  - $p_t = 1$ if period t is post-treatment, 0 otherwise.
  - $d_i = 1$ if unit i is in treated group, 0 otherwise.
  - $h_i = 1$ if unit i belongs to a "high-sensitivity" subgroup (the third difference dimension), 0 otherwise.
  - The constant, main effects of d_i and h_i, and two-way interactions that are time-invariant or unit-invariant are collinear with α_i or δ_t.
  - Therefore the specification focuses only on interactions that vary both across units & time.

Outline

Difference-in-difference continued
- When additional controls are appropriate

How to handle multiple events
- Why they are useful
- Simple estimation approach & its problems
- Better ways to handle multiple events

Falsification tests

Triple difference
- How to estimate & interpret it
- Subsample approach

Subsample Approach to Heterogeneity

- Rather than estimating the full triple-difference (DDD) specification, one can estimate separate DiD models for each subgroup defined by h_i:
  - DiD for the "less-sensitive" subgroup (units with $h_i = 0$).
  - DiD for the "more-sensitive" subgroup (units with $h_i = 1$).
- Advantages: simpler, intuitive, and you directly measure effect for each subgroup.
- Caveat: The estimates from these two subsample regressions will not necessarily match the coefficients from the combined regression (i.e., β_2 and $\beta_2 + \beta_3$ in the full model). Why? Because the baseline controls, FE structure, and time effects may differ across the two subsample models.

Why Subsample Estimates May Differ from DDD

- Key reason: When you run separate regressions on each subgroup, you allow for different time fixed effects (or other controls) in each regression implicitly.
  - In the subsample regression for $h_i = 0$, year effects reflect dynamics only for that subgroup.
  - In the subsample regression for $h_i = 1$, year effects may capture different macro/sectoral trends as applicable to high sensitivity units.
- In contrast, the combined DDD regression imposes a common* year fixed effect across both subgroups (unless you explicitly interact year FE by h_i).
- As a result, the estimated coefficients from separate subsample regressions may diverge from the combined regression's β_2 (low sensitivity) and $\beta_2 + \beta_3$ (high sensitivity) estimates.

Recovering Subsample Effects from One Regression

$$y_{it} = \beta_2(p_t \times d_i) + \beta_3(p_t \times d_i \times h_i) + \alpha_i + \delta_t + (\delta_t \times h_i) + u_{it}$$

- By interacting the year fixed effects δ_t with the subgroup indicator h_i, you allow each subgroup ($h_i = 0$ vs $h_i = 1$) to have its own time-effects structure — replicating what separate subsample regressions do.
- In this specification:
  - β_2 = DiD effect estimated for the less-sensitive subgroup ($h_i = 0$).
  - $\beta_2 + \beta_3$ = DiD effect for the more-sensitive subgroup ($h_i = 1$).
  - A t-test on β_3 tests whether the treatment effect differs significantly between the two subgroups.

Stacked Regression & Triple-Difference in Multiple Events

- When you have multiple treatment events (e.g., different cohorts or timing), the stacked regression approach (stack sub-samples for each event/cohort) remains popular.
- With this approach you can:
  - Create a separate "stack" (data subset) for each event/cohort.
  - Within each stack, apply DiD or DDD approaches, then combine stacks into one dataset.
  - Estimate one regression with interactions (and possibly fixed effects) to capture heterogeneity across events and sub-groups.
- Important caveat (Wing et al., 2024): The most basic unweighted stacked estimator "does not identify the target causal parameter or any convex combination of underlying causal effects" when there is variation in treatment timing and control/treatment shares across stacks.
- To correct for this bias, one must apply corrective sample weights as proposed by Wing et al. — or adopt alternative estimators (e.g., cohort-time ATT methods by Callaway & Sant'Anna (2021)

External Validity: Why Application Matters

- Even if your internal identification is strong (randomization or credibly parallel trends), the estimated treatment effect may not generalize beyond your sample/context.
- Key questions for external validity:
  - Are the treated firms/units representative of a broader population?
  - Does the policy context apply elsewhere, or was your setting unique?
  - Can you articulate the underlying mechanism and argue it applies beyond your particular setting?
- Being explicit about the scope of inference and the conditions under which the effect might differ helps make your findings more robust and interpretable.

Summary of Today [Part 1]

- Diff-in-diffs & control variables:
  - Don't add controls affected by treatment.
  - Controls shouldn't change estimates but can improve precision.
- Multiple events are helpful in mitigating concerns about parallel trends assumption.
  - But follow, e.g., Callaway and Sant'anna (2021) to avoid potential bias from dynamic treatment effects

Summary of Today [Part 2]

- Many falsification tests can help assess internal validity:
  - Compare ex-ante characteristics.
  - Check timing of observed effect.
- Triple difference is another way to check internal validity and mitigate concerns about identification