# ECON 730: Causal Inference with Panel Data

## Lecture 5: Introduction to Difference-in-Differences

Pedro H. C. Sant'Anna


Spring 2026

# From Experiments to Observational Studies

# Roadmap

- **Lectures 3–4:** Causal inference with randomized panel experiments
  - Design-based identification; Horvitz–Thompson estimation
  - Known treatment assignment mechanism $\Rightarrow$ clear identification
- **Today:** What if treatment is not randomized?
- **Our approach:**
  - The canonical **2×2 difference-in-differences (DiD)** setup
  - Two groups (treated vs. untreated), two periods (pre vs. post)
  - Replace randomization with **parallel trends** assumption
- **The full journey today:**
  - Setup $\rightarrow$ Identification $\rightarrow$ Estimation $\rightarrow$ Inference $\rightarrow$ Application

# The Challenge: Selection into Treatment

- In observational studies, units **select into treatment**
  - States **choose** to expand Medicaid; firms **decide** to adopt new technology
- Simple pre–post or treated-vs.-untreated comparisons are biased
  - Treated units may differ from comparison units *even without treatment*
- **Difference-in-Differences (DiD):** Exploit the **time dimension**
  - Use pre-treatment periods to account for selection concerns, as long as additional assumptions (Parallel Trends, No-Anticipation) are met
- The most widely used identification strategy in applied microeconomics
  - See Lecture 1; Currie, Kleven and Zwiers (2020); Goldsmith-Pinkham (2024)

# The 2×2 DiD Setup

# Running Example: Medicaid Expansion and Mortality

- Baker, Callaway, Cunningham, Goodman-Bacon and Sant'Anna (2025): *A Practitioner's Guide to Difference-in-Differences*
- **Policy:** Affordable Care Act (ACA) Medicaid expansion
  - 2014: Some states expand eligibility $\rightarrow$ more residents gain health insurance
  - Other states never expand (through 2019)
- **Outcome:** County-level mortality rate (ages 20–64), deaths per 100,000
- **Question:** Did Medicaid expansion **reduce** mortality?
- **2$\times$2 Setup:**
  - Two periods: 2013 (pre) and 2014 (post)
  - Two groups: States expanding in 2014 vs. never-expanding states
  - Unit of observation: Counties within states

# Data Structure: Two Groups, Two Periods

- **Time periods:** $t \in \{1, 2\}$ (e.g., 2013 and 2014)
- **Treatment group indicator:** $G_i \in \{2, \infty\}$
  - $G_i = 2$: unit $i$ is first treated at period 2 (expansion states)
  - $G_i = \infty$: unit $i$ is **never treated** (never-expansion states)
- **Treatment indicator:** $D_{i,t} = \mathbf{1}\{G_i \leq t\}$
  - $D_{i,1} = 0$ for all units (no one is treated at $t = 1$)
  - $D_{i,2} = \mathbf{1}\{G_i = 2\}$ (only the treated group at $t = 2$)
- **Key features:**
  - Treatment is **binary** and **absorbing** (once treated, stay treated)
  - No one is treated in the first period
  - This is the **simplest** DiD setup — building block for everything else

# Potential Outcomes in the 2×2 Setup

- Recall from Lecture 2: potential outcomes indexed by treatment sequence
- **Specialize to 2×2:** Only two possible treatment paths
  - $Y_{i,t}(\infty)$: outcome if unit $i$ is **never** treated (untreated potential outcome)
  - $Y_{i,t}(2)$: outcome if unit $i$ is **first treated at** $t = 2$
- **SUTVA** (Stable Unit Treatment Value Assumption):
  - No interference: unit $i$'s outcome depends only on $i$'s own treatment path
  - No hidden versions of treatment
- **Observed outcome:**

$$Y_{i,t} = \mathbf{1}\{G_i = 2\} \cdot Y_{i,t}(2) + \mathbf{1}\{G_i = \infty\} \cdot Y_{i,t}(\infty)$$

- This is the **same** potential outcomes framework from Lectures 2–4, just specialized to two groups and two periods

# Target Parameter: The Average Treatment Effect on the Treated

- **Parameter of interest:** The ATT at period 2

$$ATT = \mathbb{E}\left[Y_{i,t=2}(2) - Y_{i,t=2}(\infty) \mid G_i = 2\right]$$

- **Interpretation:**
  - Average causal effect of treatment **for those who are actually treated**
  - **Not the ATE** — we condition on $G_i = 2$
- **In the Medicaid example:**
  - Average effect of Medicaid expansion on mortality **in counties that expanded**
  - Not: what would happen if **all** counties expanded
- **The fundamental problem:** We observe $Y_{i,t=2}(2)$ for the treated group, but we never observe $Y_{i,t=2}(\infty)$ for the treated group
- How do we **impute** the missing counterfactual?

# The DiD Estimand: A Preview

- The **DiD estimand** takes the form:

$$\theta^{DiD} = \Big(\mathbb{E}[Y_{i,t=2}|G_i = 2] - \mathbb{E}[Y_{i,t=1}|G_i = 2]\Big) - \Big(\mathbb{E}[Y_{i,t=2}|G_i = \infty] - \mathbb{E}[Y_{i,t=1}|G_i = \infty]\Big)$$

- Four observable group-period means; treated group's change minus comparison group's change

**Two fundamental questions for the rest of this lecture:**

1. **How do we arrive at this estimand?** Why this particular form?
2. **Under what assumptions does $\theta^{DiD} = ATT$?** Why?

# Selection Bias and the Missing Data Problem

# The Missing Data Problem: Under SUTVA

- **SUTVA** $\Rightarrow$ each unit reveals outcomes from its own treatment path only
- What do we **observe** vs. what is **missing**?

| | **Period 1** ($t = 1$) | | **Period 2** ($t = 2$) | |
|---|---|---|---|---|
| | $Y_{i,t=1}(\infty)$ | $Y_{i,t=1}(2)$ | $Y_{i,t=2}(\infty)$ | $Y_{i,t=2}(2)$ |
| $G_i = 2$ (Treated) | ? | ✓ | ? | ✓ |
| $G_i = \infty$ (Comparison) | ✓ | ? | ✓ | ? |

- Treated units reveal $Y_{i,t}(2)$; comparison units reveal $Y_{i,t}(\infty)$
- **Problem:** The ATT requires $\mathbb{E}[Y_{i,t=2}(\infty)|G_i = 2]$ — a missing cell!
- Can we learn more using **No-Anticipation**?

# Filling In: Adding No-Anticipation

- **No-Anticipation** $\Rightarrow Y_{i,t=1}(2) = Y_{i,t=1}(\infty)$ for all $i$ (can be relaxed to hold on average among treated)
- So for treated units at $t = 1$: $Y_{i,t=1} = Y_{i,t=1}(2) = Y_{i,t=1}(\infty)$ — **both** potential outcomes observed!

| | **Period 1** ($t = 1$) | | **Period 2** ($t = 2$) | |
|---|---|---|---|---|
| | $Y_{i,t=1}(\infty)$ | $Y_{i,t=1}(2)$ | $Y_{i,t=2}(\infty)$ | $Y_{i,t=2}(2)$ |
| $G_i = 2$ (Treated) | ✓* | ✓ | ? | ✓ |
| $G_i = \infty$ (Comparison) | ✓ | – | ✓ | – |

*Same as $Y_{i,t=1}(2)$ under No-Anticipation

- The **one remaining missing cell**: $\mathbb{E}[Y_{i,t=2}(\infty)|G_i = 2]$ — the **counterfactual** for the treated group at $t = 2$

# Selection Bias in the Post-Treatment Comparison

- **Naive approach:** Compare treated and comparison at $t = 2$

$$\begin{aligned}
&\mathbb{E}[Y_{i,t=2}|G_i = 2] - \mathbb{E}[Y_{i,t=2}|G_i = \infty] \\
&= \mathbb{E}[Y_{i,t=2}(2)|G_i = 2] - \mathbb{E}[Y_{i,t=2}(\infty)|G_i = \infty] \\
&= \underbrace{\mathbb{E}[Y_{i,t=2}(2) - Y_{i,t=2}(\infty)|G_i = 2]}_{ATT} + \underbrace{\mathbb{E}[Y_{i,t=2}(\infty)|G_i = 2] - \mathbb{E}[Y_{i,t=2}(\infty)|G_i = \infty]}_{\text{Selection Bias}}
\end{aligned}$$

- The selection bias reflects differences in **untreated potential outcomes** between groups at $t = 2$

# Worked Example: Job Training Program

- **Setting:** Job training for disadvantaged workers — workers with lower baseline wages **select into** the program

| | Period 1 (Pre) | | Period 2 (Post) | |
|---|---|---|---|---|
| Unit | $Y_{i,t=1}(\infty)$ | Wage | $Y_{i,t=2}(\infty)$ | $Y_{i,t=2}(2)$ |
| A (Trained) | 20 | 20 | 22 | **27** |
| B (Trained) | 18 | 18 | 20 | **24** |
| C (Not trained) | 30 | 30 | 32 | — |
| D (Not trained) | 28 | 28 | 30 | — |

- **True ATT** $= \frac{(27-22)+(24-20)}{2} = 4.5$
- **Naive:** $\frac{27+24}{2} - \frac{32+30}{2} = 25.5 - 31 = -5.5$

Selection bias $= -10$

Selection bias contaminates naive comparisons.

**What assumptions let us recover the ATT?**

# Assumptions: No-Anticipation and Parallel Trends

# Assumption: No Anticipation

Our first assumption ensures that future treatment does not contaminate pre-treatment outcomes:

**No-Anticipation**

For all units $i$: $Y_{i,t=1}(2) = Y_{i,t=1}(\infty)$. Treatment at $t = 2$ has no effect on outcomes at $t = 1$.

For ATT identification, we only need this on average across treated units: $\mathbb{E}[Y_{i,t=1}(2)|G_i = 2] = \mathbb{E}[Y_{i,t=1}(\infty)|G_i = 2]$

- **Interpretation:** Agents do not change behavior **before** treatment begins
- **When it holds:** Unexpected policy changes; units unaware of future treatment
- **When it fails:** Pre-announced policies $\rightarrow$ behavioral adjustments before implementation (Malani and Reif, 2015)
- **Notational simplification:** Under No-Anticipation, $Y_{i,t=1} = Y_{i,t=1}(\infty)$ for all units

# Assumption: Parallel Trends

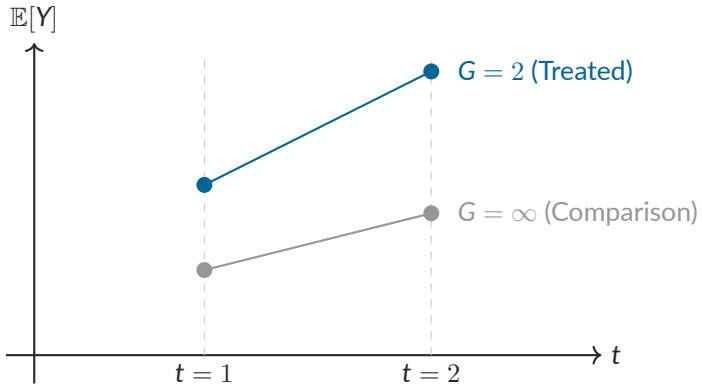
**Parallel Trends (PT)**

$$\mathbb{E}[Y_{i,t=2}(\infty) - Y_{i,t=1}(\infty) \mid G_i = 2] = \mathbb{E}[Y_{i,t=2}(\infty) - Y_{i,t=1}(\infty) \mid G_i = \infty].$$

- **What PT says:**
  - In the **absence of treatment**, both groups would have followed the same **trend**
  - Allows for permanent **level** differences between groups
- **What PT does NOT require:**
  - Same **levels** of outcomes: $\mathbb{E}[Y_{i,t}(\infty)|G_i = 2] \neq \mathbb{E}[Y_{i,t}(\infty)|G_i = \infty]$ is fine!
  - Random treatment assignment or no selection into treatment
- **What PT rules out:** Differential *time-varying* selection — trends in untreated outcomes differ by group

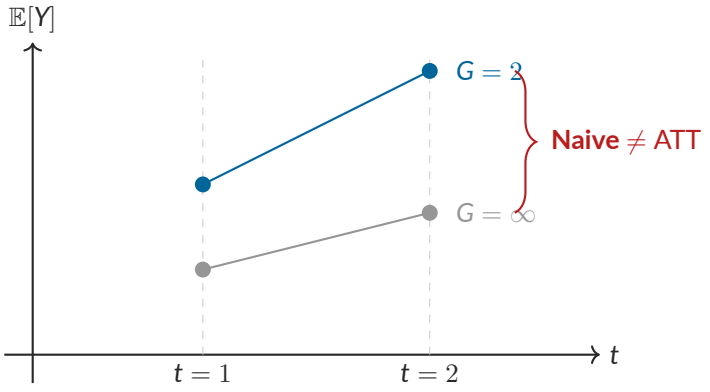
This PT involves counterfactual outcomes $\Rightarrow$ fundamentally untestable. We can assess plausibility using pre-treatment data (in a few lectures).

# Parallel Trends: Graphical Intuition (1/4)



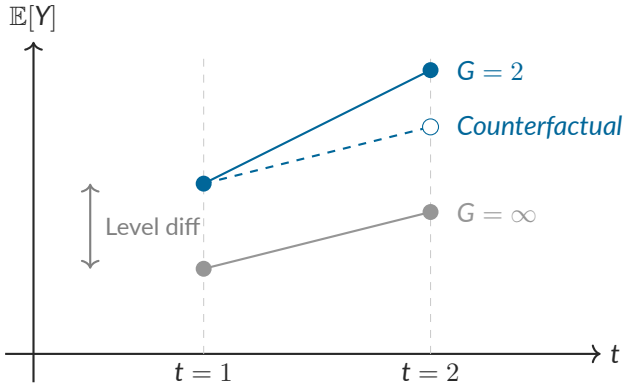

- Treated group has **higher levels** and rises **more** from $t = 1$ to $t = 2$
- But how much of the rise is due to **treatment**?

# Parallel Trends: Graphical Intuition (2/4)



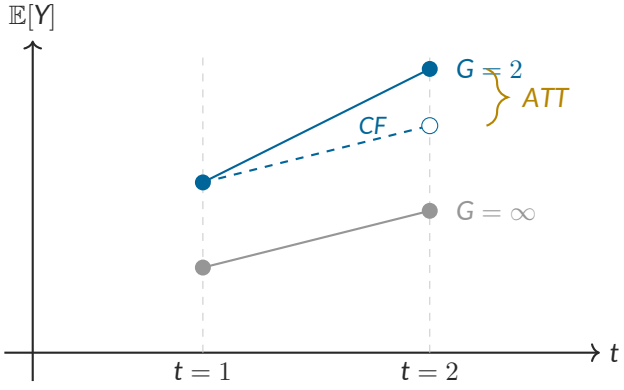

- The naive comparison at $t = 2$ includes the **level difference** that existed before treatment
- This is selection bias + treatment effect, mixed together

# Parallel Trends: Graphical Intuition (3/4)




- Under PT: the treated group's **counterfactual** trajectory is parallel to the comparison group
- The **hollow circle** is the missing counterfactual $\mathbb{E}[Y_{i,t=2}(\infty)|G_i = 2]$

# Parallel Trends: Graphical Intuition (4/4)




- **ATT** = (Treated group's change) − (Comparison group's change) = **Difference-in-Differences**
- **Think:** What would this diagram look like if Parallel Trends **fails**?

# When Is Parallel Trends Plausible?

- PT is **fundamentally untestable** — it concerns **counterfactual** trends, not observed trends
- **Suggestive evidence:** Pre-treatment trends can provide support
  - If expansion and non-expansion states had similar mortality trends before 2014, PT is more credible
  - But **parallel pre-trends** $\neq$ **parallel trends** (Roth, 2022; Ghanem, Sant'Anna and Wüthrich, 2026)
- **Potential violations in the Medicaid example:**
  - Opioid crisis differentially affected states — could confound mortality trends
  - Expansion states may have had different health infrastructure investments
- **Best practice:** Argue for PT using institutional knowledge, not just pre-trend tests. We will revisit this in future lectures

# Identification of the ATT

# Constructive Imputation: Step by Step

**Goal:** Show that $\theta^{DiD} = ATT$ under SUTVA + No-Anticipation + PT.

**Start from the ATT:**

$$ATT = \underbrace{\mathbb{E}[Y_{i,t=2}(2)|G_i = 2]}_{\text{observable}} - \underbrace{\mathbb{E}[Y_{i,t=2}(\infty)|G_i = 2]}_{\text{counterfactual}}$$

**Use PT to impute the counterfactual:**

$$\mathbb{E}[Y_{i,t=2}(\infty)|G_i = 2] = \mathbb{E}[Y_{i,t=1}(\infty)|G_i = 2] + \mathbb{E}[Y_{i,t=2}(\infty)|G_i = \infty] - \mathbb{E}[Y_{i,t=1}(\infty)|G_i = \infty]$$

**Under No-Anticipation + SUTVA, all terms are observable:**

$$\mathbb{E}[Y_{i,t=2}(\infty)|G_i = 2] = \underbrace{\mathbb{E}[Y_{i,t=1}|G_i = 2]}_{\checkmark \text{ No-Anticipation}} + \underbrace{\mathbb{E}[Y_{i,t=2}|G_i = \infty] - \mathbb{E}[Y_{i,t=1}|G_i = \infty]}_{\checkmark \text{ all observable}}$$

# The DiD Identification Result

**Theorem (DiD Identification).** Under SUTVA, No-Anticipation, and Parallel Trends:

$$ATT = \theta^{DiD} = \Big(\mathbb{E}[Y_{i,t=2}|G_i = 2] - \mathbb{E}[Y_{i,t=1}|G_i = 2]\Big) - \Big(\mathbb{E}[Y_{i,t=2}|G_i = \infty] - \mathbb{E}[Y_{i,t=1}|G_i = \infty]\Big)$$

- **Three assumptions** $\Rightarrow$ observable formula with **four population means**
- **Note:** This is an **identification** result, not an estimation result
  - The formula involves *population* expectations, not sample averages
  - Estimation comes next

# Weighted Expectations: Who Defines the Parameter?

Baker et al. (2025) emphasize: the **weights** define the ATT

**Unweighted DiD:**

$$ATT = \mathbb{E}[Y_{i,t=2}(2) - Y_{i,t=2}(\infty) | G_i = 2]$$

- Each unit gets **equal** weight
- "Average effect per **county**"

**Population-weighted DiD:**

$$ATT_\omega = \mathbb{E}_\omega[Y_{i,t=2}(2) - Y_{i,t=2}(\infty) | G_i = 2]$$

- Weight by population $\omega_i$
- "Average effect per **person**"

**Key message:** Both are valid ATTs, but they answer **different questions**

- The choice of weights is a **substantive** decision, not a statistical one
- The choice of weights also impacts the PT assumption!

We know **what** to estimate.

Now: **how** to estimate and do inference.

# Estimation

# The Analogy Principle

- **Identification:** $ATT = f(\text{population means})$
- **Estimation:** Replace population means with sample analogs
- **Notation:**
  - $n$: total sample size; $n_2$: treated units; $n_\infty$: comparison units
  - $\bar{Y}_{g,t} = \frac{1}{n_g} \sum_{i:\, G_i = g} Y_{i,t}$: sample mean for group $g \in \{2, \infty\}$ at time $t$
- **The DiD estimator ("DiD-by-hand"):**

$$\widehat{\theta}^{DiD} = \left(\bar{Y}_{g=2,t=2} - \bar{Y}_{g=2,t=1}\right) - \left(\bar{Y}_{g=\infty,t=2} - \bar{Y}_{g=\infty,t=1}\right)$$

- Simply the difference of two **within-group time changes**
- With **panel data**, this simplifies further...

# Panel Data Simplification: First-Differencing

With **panel data** (same units in both periods), define $\Delta Y_i = Y_{i,t=2} - Y_{i,t=1}$

**The DiD estimator as a two-sample difference in means:**

$$\widehat{\theta}^{DiD} = \overline{\Delta Y}_2 - \overline{\Delta Y}_\infty = \frac{1}{n_2} \sum_{i:\, G_i=2} \Delta Y_i - \frac{1}{n_\infty} \sum_{i:\, G_i=\infty} \Delta Y_i$$

**Two steps:**

1. First-difference $\Rightarrow$ removes **time-invariant** unit effects
2. Compare treated vs. comparison **changes**

**Requires panel data:**

- Same units observed in both periods
- With repeated cross-sections, use the four-means formula

# Notation: Treatment Indicator for Regressions

- We use $G_i$ notation consistently for **conditioning and parameters**:
  $ATT = \mathbb{E}[Y_{i,t=2}(2) - Y_{i,t=2}(\infty) \mid G_i = 2]$
- For **regression and influence function formulas**, define a binary shorthand:

$$D_i = \mathbf{1}\{G_i = 2\} = \begin{cases} 1 & \text{if unit } i \text{ is in the treated group} \\ 0 & \text{if unit } i \text{ is in the comparison group} \end{cases}$$

- **Why this convention?**
  - $G$-notation generalizes naturally to staggered adoption
  - $D_i$ is convenient in regression equations where we need 0/1 arithmetic
- **Rule:** All conditioning, parameters, and potential outcomes use $G_i$; $D_i$ appears only in regression specifications and influence function formulas

# TWFE Regression

A common way to estimate DiD: **Two-Way Fixed Effects (TWFE)** regression

**Pooled OLS form:**

$$Y_{i,t} = \alpha_0 + \gamma_0 D_i + \lambda_0 T_t + \beta^{TWFE}(D_i \times T_t) + \varepsilon_{i,t}$$

where $T_t = \mathbf{1}\{t = 2\}$, $D_i = \mathbf{1}\{G_i = 2\}$

**Unit & time FE form:**

$$Y_{i,t} = \alpha_i + \lambda_t + \beta^{TWFE} D_{i,t} + \varepsilon_{i,t}$$

where $D_{i,t} = \mathbf{1}\{G_i \leq t\}$

- **Both are equivalent** in the $2 \times 2$ case with **balanced panel data** — same $\hat{\beta}^{TWFE}$
- **Key question:** Does $\hat{\beta}^{TWFE} = \widehat{\theta}^{DiD}$?
- **Answer:** Yes! In the $2 \times 2$ case, they are **numerically identical**

# TWFE = DiD-by-Hand: The Equivalence

- **Claim:** $\hat{\beta}^{TWFE} = \widehat{\theta}^{DiD}$ (exact numerical equality)
- **Proof sketch:** OLS solves four moment conditions:

$$\begin{aligned}
\mathbb{E}_n[\varepsilon_{i,t}] = 0 &\quad\Rightarrow \hat{\alpha}_0 = \bar{Y}_{g=\infty,t=1} \\
\mathbb{E}_n[D_i \cdot \varepsilon_{i,t}] = 0 &\quad\Rightarrow \hat{\gamma}_0 = \bar{Y}_{g=2,t=1} - \bar{Y}_{g=\infty,t=1} \\
\mathbb{E}_n[T_t \cdot \varepsilon_{i,t}] = 0 &\quad\Rightarrow \hat{\lambda}_0 = \bar{Y}_{g=\infty,t=2} - \bar{Y}_{g=\infty,t=1} \\
\mathbb{E}_n[(D_i \cdot T_t) \cdot \varepsilon_{i,t}] = 0 &\quad\Rightarrow \hat{\beta}^{TWFE} = \widehat{\theta}^{DiD}
\end{aligned}$$

- The four moment conditions **uniquely pin down** the four group-time means
- **Bottom line:** In the 2×2 case, TWFE regression is just a **convenient way to compute** the DiD estimator. Nothing more, nothing less.
- **Every student in this class should be able to derive these equivalence results!**
- **Think:** Will this equivalence survive when we move to **staggered adoption**? Why or why not?

# Three Equivalent Specifications

Baker et al. (2025): Three OLS specifications, same $\hat{\beta}$

| Spec | Regression | Data Used |
|---|---|---|
| (1) Pooled OLS | $Y_{i,t} = \alpha + \gamma D_i + \lambda T_t + \beta(D_i \cdot T_t) + \varepsilon_{i,t}$ | Panel ($2n$ obs) |
| (2) First-diff | $\Delta Y_i = \delta + \beta D_i + u_i$ | Panel, FD ($n$ obs) |
| (3) Unit & time FE | $Y_{i,t} = \alpha_i + \lambda_t + \beta D_{i,t} + \varepsilon_{i,t}$ | Panel ($2n$ obs) |

- All three yield **identical** point estimates for $\hat{\beta}$ in the $2\times2$ case with balanced panel
- **BUT:** Standard errors differ unless you cluster appropriately!
  - Spec (1) and (3) have $2n$ observations; Spec (2) has $n$ observations
  - *Clustering at the unit level* in Specs (1) and (3) reconciles the SEs

# OLS: Regression as a Means to an End

- **Important conceptual point:** The regression is a **computational device**
- In the 2×2 case, OLS does not add any statistical content
  - Same estimate as computing four means and subtracting
  - The **interpretation** comes from the DiD identification argument, not from the regression
- **Why use regression then?**
  - Convenience: standard software handles SEs and clustering
  - Reporting: tables with coefficients and SEs are standard in economics
- **Warning for later:** In more complex settings (staggered adoption, heterogeneous effects), TWFE ≠ "the" DiD estimator
  - The 2×2 equivalence is **special** and does not generalize
  - Recall Lecture 3: FE can be biased under carryover effects. Here, with no carryover in the 2×2 case, TWFE is well-behaved

# Empirical Application: Medicaid Expansion

# Application: The ACA and County Mortality

- **Data:** County-level mortality rates (ages 20–64), deaths per 100,000
  - Source: Baker et al. (2025), replication data from `pedrohcgs/JEL-DiD`
- **Treatment:** State-level Medicaid expansion under the ACA
  - 24 states + DC expanded in January 2014
  - 19 states never expanded (through 2019)
  - Drop DC and pre-2014 adopters (DE, MA, NY, VT) for clean $2\times2$
- **$2\times2$ Setup:**
  - Pre: 2013, Post: 2014
  - Treated: Counties in states expanding in 2014
  - Comparison: Counties in never-expanding states
- **Sample:** $\sim$2,300 counties, roughly 900 treated and 1,400 comparison
- **Key feature:** Counties vary enormously in population size — weighting matters!

# Simple 2×2 Results: Unweighted vs. Weighted

Baker et al. (2025), Table 2: Simple means and DiD

| | Unweighted | | Pop-Weighted | |
|---|---|---|---|---|
| | Pre (2013) | Post (2014) | Pre (2013) | Post (2014) |
| Treated ($G = 2$) | 419.2 | 428.5 | 322.7 | 326.5 |
| Comparison ($G = \infty$) | 474.0 | 483.1 | 376.4 | 382.7 |
| $\Delta$ Treated | +9.3 | | +3.7 | |
| $\Delta$ Comparison | +9.1 | | +6.3 | |
| **DiD** | **+0.1** | | **−2.6** | |

- **Unweighted:** DiD $\approx +0.1$ (SE $= 3.7$) (essentially zero, "wrong sign")
- **Population-weighted:** DiD $\approx -2.6$ (SE $= 1.5$) (meaningful reduction in mortality)
- **Why the difference?** **Different parameters!** "Per county" vs. "per person" ATT

# Regression Equivalence: Three Specifications

- All three specifications yield the same point estimate (county-clustered SEs)

| | (1) Pooled OLS | (2) First-diff | (3) Unit & time FE |
|---|---|---|---|
| $\hat{\beta}$ (unweighted) | 0.1 | 0.1 | 0.1 |
| SE (county-clustered) | (3.7) | (3.7) | (3.7) |
| $\hat{\beta}$ (pop-weighted) | $-2.6$ | $-2.6$ | $-2.6$ |
| SE (county-clustered) | (1.5) | (1.5) | (1.5) |
| Observations | $2n$ | $n$ | $2n$ |

- Point estimates are **identical** across all three specifications
- County-clustered SEs are also identical (as expected from theory)
- Neither result is statistically significant at conventional levels

# R Code: DiD by Hand vs. Regression

**DiD by hand:**

```r
# Four group-time means
means <- short_data %>%
  group_by(Treat, Post) %>%
  summarise(
    m = mean(crude_rate_20_64))

# DiD estimate
did_hat <- (means$m[4] - means$m[3]) -
           (means$m[2] - means$m[1])
```

**TWFE regression:**

```r
library(fixest)

# TWFE with county + year FE
reg <- feols(
  crude_rate_20_64 ~ Treat:Post
  | county_code + year,
  data = short_data,
  cluster = ~stateID)

coef(reg)  # Same as did_hat!
```

- Both approaches give **identical** estimates
- `short_data`: balanced panel with 2013–2014 only (the 2×2 subset)
- Full replication code: `github.com/pedrohcgs/JEL-DiD`

# Key Takeaway: Weights Matter

**Lesson from the Medicaid application:**

- The choice of weights (unweighted vs. population-weighted) changes the **target parameter**
- Unweighted DiD: $\hat{\theta} = +0.1$ ("per county" ATT)
- Weighted DiD: $\hat{\theta} = -2.6$ ("per person" ATT)
- Both are **valid**, but they answer **different questions**

- **Why this matters going forward:**
  - When we add **covariates**, conditioning variables implicitly change weights
  - Different estimators (regression adjustment, IPW, doubly robust) target the same parameter but may use different implicit weights
- **Researcher's responsibility:** Be explicit about what parameter you are estimating

We have seen DiD work in practice.

Now: **what are its statistical properties?**

# Influence Functions and Asymptotic Theory

# Why Influence Functions?

Having seen DiD work in practice, we now formalize the statistical properties of the estimator.

- We have an estimator $\widehat{\theta}^{DiD}$. What are its **large-sample properties**?
- **Goals:** Consistency, asymptotic normality, variance estimation, and valid bootstrap inference
- **Influence functions** provide a **unified** approach:
  - Decompose the estimator as a **sum of iid terms** (+ remainder)
  - The influence function $\psi_i$ captures unit $i$'s contribution to the estimator
  - Variance of the IF $=$ asymptotic variance of $\widehat{\theta}^{DiD}$
- This framework generalizes naturally to **more complex estimators** (covariates, staggered designs) in later lectures

# Panel Data Sampling Scheme

**Panel Data Sampling**

We observe an iid random sample $\{Z_i\}_{i=1}^n$ where $Z_i = (Y_{i,t=1}, Y_{i,t=2}, G_i)$ is drawn from the joint distribution of $(Y_{i,t=1}, Y_{i,t=2}, G_i)$.

- The **same** $n$ **units** are observed in both periods ($t = 1$ and $t = 2$)
- **Key:** Units are iid across $i$, but $Y_{i,t=1}$ and $Y_{i,t=2}$ are **not** independent within unit
- This allows us to compute $\Delta Y_i = Y_{i,t=2} - Y_{i,t=1}$ for each unit
- Let $D_i = \mathbf{1}\{G_i = 2\}$ and $p = \Pr(G_i = 2) \in (0, 1)$
- **Notation:** $\mathbb{E}_n[f(Z_i)] = \frac{1}{n}\sum_{i=1}^n f(Z_i)$ denotes the sample average

# The DiD Estimator as a Function of Means

- Write the DiD estimator using empirical expectations:

$$\widehat{\theta}^{DiD} = \frac{\mathbb{E}_n[\Delta Y_i \cdot D_i]}{\mathbb{E}_n[D_i]} - \frac{\mathbb{E}_n[\Delta Y_i \cdot (1 - D_i)]}{\mathbb{E}_n[1 - D_i]}$$

- This is a **smooth function of sample means:** $\widehat{\theta}^{DiD} = g(\mathbb{E}_n[\mathbf{m}(Z_i)])$
- where $\mathbf{m}(Z_i) = (\Delta Y_i \cdot D_i,\ D_i,\ \Delta Y_i \cdot (1 - D_i),\ 1 - D_i)'$
- **By the Law of Large Numbers:** $\mathbb{E}_n[\mathbf{m}(Z_i)] \xrightarrow{p} \mathbb{E}[\mathbf{m}(Z_i)]$
- **By the Continuous Mapping Theorem:** $\widehat{\theta}^{DiD} = g(\mathbb{E}_n[\mathbf{m}(Z_i)]) \xrightarrow{p} g(\mathbb{E}[\mathbf{m}(Z_i)]) = \theta^{DiD}$

$\Rightarrow$ **Consistency** follows from LLN + CMT

# Asymptotic Normality: Setup

- Apply the **delta method** to $g(\mathbb{E}_n[\mathbf{m}(Z_i)])$:

$$\sqrt{n}(\widehat{\theta}^{DiD} - \theta^{DiD}) = \sqrt{n} \cdot \nabla g(\boldsymbol{\mu})' \big(\mathbb{E}_n[\mathbf{m}(Z_i)] - \boldsymbol{\mu}\big) + o_p(1)$$

  where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{m}(Z_i)]$

- By the CLT:

$$\sqrt{n}\big(\mathbb{E}_n[\mathbf{m}(Z_i)] - \boldsymbol{\mu}\big) \xrightarrow{d} N(\mathbf{o}, \text{Var}(\mathbf{m}(Z_i)))$$

- Combining (Slutsky):

$$\sqrt{n}(\widehat{\theta}^{DiD} - \theta^{DiD}) \xrightarrow{d} N\big(0,\ \nabla g(\boldsymbol{\mu})' \text{Var}(\mathbf{m}(Z_i)) \nabla g(\boldsymbol{\mu})\big)$$

- This can be written more compactly using the **influence function**...

# The Influence Function: Panel Data Case

**Panel data IF.** Under the panel data sampling scheme:

$$\psi_i^p = \underbrace{\frac{D_i}{p}}_{w_1(D_i)} (\Delta Y_i - \mu_{\Delta,2}) \; - \; \underbrace{\frac{1-D_i}{1-p}}_{w_0(D_i)} (\Delta Y_i - \mu_{\Delta,\infty})$$

where $p = \Pr(G_i = 2)$, $\mu_{\Delta,g} = \mathbb{E}[\Delta Y_i | G_i = g]$ for $g \in \{2, \infty\}$.

- **Key property:** $\sqrt{n}(\widehat{\theta}^{DiD} - \theta^{DiD}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_i^p + o_p(1)$
- **Asymptotic variance:** $V_p = \text{Var}(\psi_i^p)$
- Full derivation in the appendix

# Computing the IF: Medicaid Example

**Suppose:** $p = 0.4$, $\mu_{\Delta,2} = -4.0$, $\mu_{\Delta,\infty} = -1.4$

- **Expansion county** with $\Delta Y_i = -6.5$:

$$\psi_i^p = \tfrac{1}{0.4}(-6.5 - (-4.0)) - 0 = -6.25$$

This county "pulls" the estimate toward a larger mortality reduction

- **Non-expansion county** with $\Delta Y_i = -0.5$:

$$\psi_i^p = 0 - \tfrac{1}{0.6}(-0.5 - (-1.4)) = -1.5$$

Its mortality fell less than average $\Rightarrow$ supports the treatment effect

- **Takeaway:** The IF assigns each unit a signed "credit" for the overall estimate

# Understanding the Influence Function

- **Two terms:** treated group's contribution and comparison group's contribution
- Each term is a **demeaned** quantity, weighted by group proportion
- $\mathbb{E}[\psi_i^p] = 0$ by construction — the IF is centered
- **Intuition for the weights:**
  - $D_i/p$: up-weights treated units (rarer group gets more weight)
  - $(1 - D_i)/(1 - p)$: up-weights comparison units
- **Why this matters:** The IF gives us everything for inference — consistency, asymptotic normality, variance estimation, and bootstrap validity all follow from this representation

# Asymptotic Normality Result

**Theorem.** Under the panel data sampling scheme, SUTVA, No-Anticipation, and $\text{Var}(\Delta Y_i | G_i = g) < \infty$:

$$\sqrt{n}\left(\widehat{\theta}^{DiD} - \theta^{DiD}\right) \xrightarrow{d} N(0, V_p)$$

where $V_p = \text{Var}(\psi_i^p) = \frac{\sigma^2_{\Delta,2}}{p} + \frac{\sigma^2_{\Delta,\infty}}{1-p}$

- $\sigma^2_{\Delta,g} = \text{Var}(\Delta Y_i | G_i = g)$: within-group variance of the first-differenced outcome
- **Intuition:** Two components — treated group uncertainty ($\sigma^2_{\Delta,2}/p$) and comparison group uncertainty ($\sigma^2_{\Delta,\infty}/(1-p)$)
- **Design matters:** $p$ small $\Rightarrow$ first term dominates $\Rightarrow$ large variance. Minimized at $p \approx 0.5$ (balanced design)
- **Think:** If only 5% of counties expanded, what does this formula say about **power**?

# Repeated Cross-Sections: Brief Overview

- With **repeated cross-sections (RCS)**, different units sampled at $t = 1$ and $t = 2$
  - Let $T_i = \mathbf{1}\{\text{unit } i \text{ sampled at } t = 2\}$ (distinct from $T_t$ in TWFE regression)
- Cannot first-difference $\Rightarrow$ must estimate **four means** separately
- The IF has **four components** instead of two:

$$\psi_i^{rc} = w_1(D_i, T_i) \cdot (Y_i - \mu_{2,T_i}) - w_0(D_i, T_i) \cdot (Y_i - \mu_{\infty,T_i})$$

  where each weight depends on both group **and** period membership
- **Requires additional assumption:** Stationarity of group composition
  - $\Pr(G_i = 2|T_i = 1) = \Pr(G_i = 2|T_i = 2)$ (no compositional changes)
- **Panel is strictly more efficient than RCS** — panel exploits within-unit correlation

# Inference: Standard Errors and Clustering

# How to Conduct Inference

- **Given:** $\sqrt{n}(\widehat{\theta}^{DiD} - \theta^{DiD}) \xrightarrow{d} N(0, V_p)$
- **Variance estimation (analogy principle):**

$$\widehat{V}_p = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\psi}_i^p\right)^2$$

  where $\widehat{\psi}_i^p$ replaces population quantities with sample analogs

- **Standard error:** $\widehat{SE} = \sqrt{\widehat{V}_p/n}$
- **Confidence interval:** $\widehat{\theta}^{DiD} \pm z_{\alpha/2} \cdot \widehat{SE}$
- $t$**-statistic:** $t = \widehat{\theta}^{DiD}/\widehat{SE}$. Reject $H_0 : \theta^{DiD} = 0$ if $|t| > z_{\alpha/2}$
- **But wait:** Should we **cluster** the standard errors?

# Why Cluster Standard Errors?

- **DiD-by-hand** operates on $n$ units (first-differenced panel)
  - Natural degrees of freedom: $n$ independent observations
- **TWFE regression** uses $2n$ observations ($n$ units $\times$ 2 periods)
  - Without clustering: SE formula assumes $2n$ independent observations
  - Artificially inflates sample size by a factor of 2!
- **Clustering at the unit level** accounts for within-unit correlation
  - Two observations from the same unit are *not* independent
  - Corrected SE matches the DiD-by-hand SE
- Bertrand, Duflo and Mullainathan (2004): Ignoring clustering in DiD leads to massive over-rejection (up to 40% rejection at 5% nominal level)

# Multiplier Bootstrap Using Influence Functions

- **Key advantage of IF-based inference:** The multiplier bootstrap
- **Idea:** Perturb the IF with random weights instead of resampling data

$$\widehat{\theta}^{*,b} = \widehat{\theta}^{DiD} + \frac{1}{n}\sum_{i=1}^{n} U_i^{(b)} \cdot \widehat{\psi}_i^p$$

  where $U_i^{(b)} \sim N(0,1)$ are iid random weights

- **No re-estimation needed:** Each bootstrap draw is a simple weighted sum
  - Extremely fast compared to traditional bootstrap (which re-estimates $\widehat{\theta}$ each time)
- **For cluster-level inference:** Draw $U_s^{(b)}$ at the **cluster level**
  - All units in cluster $s$ share the same weight $U_s^{(b)}$
  - This preserves the within-cluster correlation structure

# Bootstrap Algorithm

**Multiplier Bootstrap for DiD (Cluster-Robust)**

1. Compute $\widehat{\theta}^{DiD}$ and the influence function $\widehat{\psi}_i^p$ for each unit $i$
2. For $b = 1, \ldots, B$ (e.g., $B = 999$):
   - 2.1 Draw $U_s^{(b)} \sim N(0,1)$ for each cluster $s = 1, \ldots, S$
   - 2.2 Assign $U_i^{(b)} = U_{s(i)}^{(b)}$ for all units $i$ in cluster $s$
   - 2.3 Compute: $T^{*,b} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i^{(b)} \cdot \widehat{\psi}_i^p$
3. Compute bootstrap critical value: $c_\alpha = \text{quantile}_{1-\alpha}(|T^{*,1}|, \ldots, |T^{*,B}|)$
4. Reject $H_0 : \theta^{DiD} = 0$ if $|\sqrt{n} \cdot \widehat{\theta}^{DiD}| > c_\alpha$

- Bootstrap CIs can also be constructed from quantiles of $\widehat{\theta}^{*,b} = \widehat{\theta}^{DiD} + T^{*,b}/\sqrt{n}$
- Rademacher weights ($U_s^{(b)} \in \{-1, +1\}$ with equal probability) also valid and sometimes preferred for few clusters

# The Few-Clusters Problem

- **Standard cluster-robust inference** relies on $S \to \infty$ (number of clusters)
- **Problem:** In many DiD applications, $S$ is small (e.g., 50 states, 10 provinces)
  - CLT approximation may be poor with few clusters
  - Cluster-robust SEs can be severely biased downward
- **Approaches in the literature:**
  - Donald and Lang (2007): ***t*-distribution** with $S - 2$ degrees of freedom (assumes homoskedasticity)
  - Conley and Taber (2011): **Large untreated group** to inform inference (fixed $S_1$, growing $S_0$)
  - Ferman and Pinto (2019): Allow for **heteroskedasticity** across clusters
- **No silver bullet:** Each approach requires additional assumptions; this is an active area of research (see Alvarez, Ferman and Wüthrich, 2025, for a recent survey)

# Panel Data vs. Repeated Cross-Sections

We have assumed **panel data**. But what if we do not have it?

---

What changes with **repeated cross-sections**?

# Repeated Cross-Section (RCS) Sampling Scheme

**Repeated Cross-Section Sampling**

Period $t$ data $\{(Y_{i,t}, G_i)\}$ is an iid sample from $F_{Y,G|T=t}$. Observations across periods are independent.

- **Different units** sampled at $t = 1$ and $t = 2$
- **Requires additional assumption:** Stationarity of group composition
  - $\Pr(G_i = 2|T_i = 1) = \Pr(G_i = 2|T_i = 2)$
  - Violation: **compositional changes** (Sant'Anna and Xu, 2026)
- **IF has four components:** one per group-period cell
- **Examples:** CPS microdata, Census data, polling data before/after events

# Comparison: Panel vs. Repeated Cross-Sections

**Panel Data**

- Same units both periods
- Can first-difference
- 2 IF components
- More efficient
- Risk: attrition, survivorship bias

**Repeated Cross-Sections**

- Different units each period
- Must estimate 4 means
- 4 IF components
- Less efficient
- Risk: compositional changes

- **Key result:** Panel data is **strictly more efficient** than RCS
- **Intuition:** Panel exploits within-unit correlation $\Rightarrow \text{Var}(\Delta Y_i | G_i)$ can be much smaller than $\text{Var}(Y_{i,t=2}|G_i) + \text{Var}(Y_{i,t=1}|G_i)$
- **Unbalanced panels:** Some units observed once, some twice — use panel structure where available, but gains depend on strength of within-unit correlation

# Taking Stock

# What We Accomplished Today

**Key takeaways from the 2×2 DiD framework:**

1. **Identification:** SUTVA + No-Anticipation + Parallel Trends $\Rightarrow$ DiD identifies the ATT using four observable group-time means
2. **Estimation:** DiD-by-hand and TWFE regression are **numerically identical** in the 2×2 case — regression is just a convenient computational device
3. **Inference:** Influence functions provide consistency, asymptotic normality, and variance estimation. **Always cluster** at least at the unit level; ideally at the treatment-assignment level
4. **Weights matter:** The choice of weights (unweighted vs. population-weighted) defines a **different** target parameter

# References

**Alvarez, Luis, Bruno Ferman, and Kaspar Wüthrich**, "Inference with Few Treated Units," 2025. arXiv:2504.19841.

**Baker, Andrew, Brantly Callaway, Scott Cunningham, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna**, "Difference-in-Differences Designs: A Practitioner's Guide," *Journal of Economic Literature*, 2025, *Forthcoming*.

**Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan**, "How Much Should We Trust Differences-In-Differences Estimates?," *Quarterly Journal of Economics*, 2004, *119* (1), 249–275.

**Conley, Timothy G. and Christopher R. Taber**, "Inference with "Difference in Differences" with a Small Number of Policy Changes," *Review of Economics and Statistics*, 2011, *93* (1), 113–125.

**Currie, Janet, Henrik Kleven, and Esmée Zwiers**, "Technology and Big Data Are Changing Economics: Mining Text to Track Methods," *AEA Papers and Proceedings*, 2020, *110*, 42–48.

# References ii

**Donald, Stephen G. and Kevin Lang**, "Inference with Difference-in-Differences and Other Panel Data," *Review of Economics and Statistics*, 2007, *89* (2), 221–233.

**Ferman, Bruno and Cristine Pinto**, "Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity," *The Review of Economics and Statistics*, 2019, *101* (3), 452–467.

**Ghanem, Dalia, Pedro H. C. Sant'Anna, and Kaspar Wüthrich**, "When Should Pre-trends Be Parallel?," *AEA Papers and Proceedings*, 2026, *116*, forthcoming.

**Goldsmith-Pinkham, Paul**, "Tracking the Credibility Revolution across Fields," *arXiv:2405.20604*, 2024.

**Malani, Anup and Julian Reif**, "Interpreting Pre-Trends as Anticipation: Impact on Estimated Treatment Effects from Tort Reform," *Journal of Public Economics*, 2015, *124*, 1–17.

**Roth, Jonathan**, "Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends," *American Economic Review: Insights*, 2022, *4* (3), 305–322.

**Sant'Anna, Pedro H. C. and Qi Xu**, "Difference-in-Differences with Compositional Changes," *Working Paper*, 2026.

# Appendix

# Appendix: Panel Data — Consistency (Detailed)

- **Setup:** $Z_i = (Y_{i,t=1}, Y_{i,t=2}, G_i)$ iid, $D_i = \mathbf{1}\{G_i = 2\}$, $\Delta Y_i = Y_{i,t=2} - Y_{i,t=1}$
- Define population quantities:

$$\mu_1 = \mathbb{E}[\Delta Y_i \cdot D_i], \quad p = \mathbb{E}[D_i]$$
$$\mu_0 = \mathbb{E}[\Delta Y_i \cdot (1 - D_i)], \quad 1 - p = \mathbb{E}[1 - D_i]$$

- The DiD estimand: $\theta^{DiD} = \frac{\mu_1}{p} - \frac{\mu_0}{1-p}$
- **Sample analogs:**

$$\widehat{\theta}^{DiD} = \frac{\mathbb{E}_n[\Delta Y_i \cdot D_i]}{\mathbb{E}_n[D_i]} - \frac{\mathbb{E}_n[\Delta Y_i \cdot (1 - D_i)]}{\mathbb{E}_n[1 - D_i]}$$

- By **LLN**: $\mathbb{E}_n[\Delta Y_i \cdot D_i] \xrightarrow{p} \mu_1$, etc.
- By **CMT** (continuous mapping theorem): $\widehat{\theta}^{DiD} \xrightarrow{p} \theta^{DiD}$ $\square$

# Appendix: Panel Data — Asymptotic Linearity (1/3)

- **Goal:** Show $\sqrt{n}(\widehat{\theta}^{DiD} - \theta^{DiD}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i^p + o_p(1)$
- **Step 1:** Write the estimator as a function of sample means.
  Let $\hat{\mu}_1 = \mathbb{E}_n[\Delta Y_i D_i]$, $\hat{p} = \mathbb{E}_n[D_i]$, $\hat{\mu}_0 = \mathbb{E}_n[\Delta Y_i(1 - D_i)]$, $\hat{q} = \mathbb{E}_n[1 - D_i]$

$$\widehat{\theta}^{DiD} = \frac{\hat{\mu}_1}{\hat{p}} - \frac{\hat{\mu}_0}{\hat{q}}$$

- **Step 2:** Linearize each ratio. For the first term:

$$\begin{aligned}
\frac{\hat{\mu}_1}{\hat{p}} - \frac{\mu_1}{p} &= \frac{\hat{\mu}_1 p - \mu_1 \hat{p}}{\hat{p} \cdot p} \\
&= \frac{(\hat{\mu}_1 - \mu_1)p - \mu_1(\hat{p} - p)}{\hat{p} \cdot p} \\
&= \frac{1}{p}(\hat{\mu}_1 - \mu_1) - \frac{\mu_1}{p^2}(\hat{p} - p) + o_p(n^{-1/2})
\end{aligned}$$

  where the last step uses $\hat{p} \xrightarrow{p} p > 0$.

# Appendix: Panel Data — Asymptotic Linearity (2/3)

- **Step 3:** Similarly for the second ratio:

$$\frac{\hat{\mu}_0}{\hat{q}} - \frac{\mu_0}{1-p} = \frac{1}{1-p}(\hat{\mu}_0 - \mu_0) + \frac{\mu_0}{(1-p)^2}(\hat{p} - p) + o_p(n^{-1/2})$$

- **Step 4:** Combine:

$$\begin{aligned}\widehat{\theta}^{DiD} - \theta^{DiD} = &\frac{1}{p}(\hat{\mu}_1 - \mu_1) - \frac{\mu_{\Delta,2}}{p}(\hat{p} - p) \\ &- \frac{1}{1-p}(\hat{\mu}_0 - \mu_0) - \frac{\mu_{\Delta,\infty}}{1-p}(\hat{p} - p) + o_p(n^{-1/2})\end{aligned}$$

  where $\mu_{\Delta,2} = \mu_1/p = \mathbb{E}[\Delta Y_i | G_i = 2]$ and $\mu_{\Delta,\infty} = \mu_0/(1-p) = \mathbb{E}[\Delta Y_i | G_i = \infty]$.

- Note: $\hat{p} - p = \mathbb{E}_n[D_i] - p$ and $\hat{\mu}_1 - \mu_1 = \mathbb{E}_n[\Delta Y_i D_i] - \mu_1$

# Appendix: Panel Data — Asymptotic Linearity (3/3)

- **Step 5:** Express as average of iid terms. Each sample mean is an average:

$$\widehat{\theta}^{DiD} - \theta^{DiD} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_i\Delta Y_i - \mu_1}{p} - \frac{\mu_{\Delta,2}(D_i - p)}{p} - \frac{(1-D_i)\Delta Y_i - \mu_0}{1-p} - \frac{\mu_{\Delta,\infty}(D_i - p)}{1-p}\right] + o_p(n^{-1/2})$$

- **Step 6:** Simplify the treated group term:

$$\frac{D_i\Delta Y_i - \mu_1}{p} - \frac{\mu_{\Delta,2}(D_i - p)}{p} = \frac{D_i(\Delta Y_i - \mu_{\Delta,2})}{p}$$

- Similarly for the comparison term. Thus:

$$\widehat{\theta}^{DiD} - \theta^{DiD} = \frac{1}{n}\sum_{i=1}^{n}\underbrace{\left[\frac{D_i}{p}(\Delta Y_i - \mu_{\Delta,2}) - \frac{1-D_i}{1-p}(\Delta Y_i - \mu_{\Delta,\infty})\right]}_{\psi_i^p} + o_p(n^{-1/2})$$

# Appendix: Panel Data — Asymptotic Variance

- From the IF representation: $\sqrt{n}(\widehat{\theta}^{DiD} - \theta^{DiD}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i^p + o_p(1)$
- By CLT: $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i^p \xrightarrow{d} N(0, \text{Var}(\psi_i^p))$
- **Computing $\text{Var}(\psi_i^p)$:**

$$V_p = \text{Var}\left( \frac{D_i}{p}(\Delta Y_i - \mu_{\Delta,2}) - \frac{1 - D_i}{1 - p}(\Delta Y_i - \mu_{\Delta,\infty}) \right)$$

$$= \frac{1}{p^2}\text{Var}(D_i(\Delta Y_i - \mu_{\Delta,2})) + \frac{1}{(1-p)^2}\text{Var}((1 - D_i)(\Delta Y_i - \mu_{\Delta,\infty}))$$

(cross-term is zero since $D_i(1 - D_i) = 0$)

- Simplifying: $\text{Var}(D_i(\Delta Y_i - \mu_{\Delta,2})) = p \cdot \sigma^2_{\Delta,2}$ where $\sigma^2_{\Delta,2} = \text{Var}(\Delta Y_i | G_i = 2)$

$$\boxed{V_p = \frac{\sigma^2_{\Delta,2}}{p} + \frac{\sigma^2_{\Delta,\infty}}{1-p}}$$

# Appendix: Variance Estimation

- **Plug-in estimator:** Replace population quantities with sample analogs

$$\widehat{\psi}_i^p = \frac{D_i}{\hat{p}}(\Delta Y_i - \overline{\Delta Y}_2) - \frac{1 - D_i}{1 - \hat{p}}(\Delta Y_i - \overline{\Delta Y}_\infty)$$

- **Variance estimate:**

$$\widehat{V}_p = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\psi}_i^p)^2$$

- By LLN: $\widehat{V}_p \xrightarrow{p} V_p$
- **Standard error:** $\widehat{SE}(\widehat{\theta}^{DiD}) = \sqrt{\widehat{V}_p/n}$
- **Alternative:** Direct plug-in

$$\widehat{V}_p^{alt} = \frac{\hat{\sigma}^2_{\Delta,2}}{\hat{p}} + \frac{\hat{\sigma}^2_{\Delta,\infty}}{1 - \hat{p}}$$

where $\hat{\sigma}^2_{\Delta,g} = \frac{1}{n_g}\sum_{i:G_i=g}(\Delta Y_i - \overline{\Delta Y}_g)^2$

# Appendix: RCS — Influence Function Derivation (1/2)

- **RCS data:** $Z_i = (Y_i, G_i, T_i)$, $D_i = \mathbf{1}\{G_i = 2\}$, $T_i = \mathbf{1}\{\text{sampled at } t = 2\}$ (distinct from $T_t$ in regression)
- Cannot first-difference. The DiD estimand uses four conditional means:

$$\theta^{DiD} = (\mu_{2,2} - \mu_{2,1}) - (\mu_{\infty,2} - \mu_{\infty,1})$$

where $\mu_{g,t} = \mathbb{E}[Y_i | G_i = g, T_i = t]$

- Under **stationarity**: $\Pr(G_i = 2 | T_i = t)$ is constant across $t$
- **Linearization:** Apply the delta method to each of the four ratios
- The IF takes the form:

$$\psi_i^{rc} = \frac{D_i T_i}{p \cdot \lambda}(Y_i - \mu_{2,2}) - \frac{D_i(1 - T_i)}{p \cdot (1 - \lambda)}(Y_i - \mu_{2,1}) - \frac{(1 - D_i)T_i}{(1 - p) \cdot \lambda}(Y_i - \mu_{\infty,2}) + \frac{(1 - D_i)(1 - T_i)}{(1 - p)(1 - \lambda)}(Y_i - \mu_{\infty,1})$$

where $\lambda = \Pr(T_i = 1)$ and $\mu_{g,t} = \mathbb{E}[Y_i | G_i = g, T_i = t]$

# Appendix: RCS — Influence Function Derivation (2/2)

- **Asymptotic variance:** $V_{rc} = \text{Var}(\psi_i^{rc})$
- Since $(D_i, T_i) \in \{0,1\}^2$ creates four disjoint groups, cross-products vanish:

$$V_{rc} = \frac{\sigma^2_{2,2}}{p \cdot \lambda} + \frac{\sigma^2_{2,1}}{p \cdot (1-\lambda)} + \frac{\sigma^2_{\infty,2}}{(1-p) \cdot \lambda} + \frac{\sigma^2_{\infty,1}}{(1-p)(1-\lambda)}$$

  where $\sigma^2_{g,t} = \text{Var}(Y_i | G_i = g, T_i = t)$

- **Comparison with panel:**

$$V_p = \frac{\sigma^2_{\Delta,2}}{p} + \frac{\sigma^2_{\Delta,\infty}}{1-p} \quad < \quad V_{rc}$$

  **Two sources of RCS inefficiency:**

  (i) Cannot exploit within-unit correlation: $\sigma^2_{\Delta,g} = \sigma^2_{g,2} + \sigma^2_{g,1} - 2\text{Cov}(Y_{i,2}, Y_{i,1} | G_i = g)$

  (ii) Must split sample across periods: each cell has $\leq n$ observations (factors $1/\lambda$, $1/(1-\lambda)$)

- **Conclusion:** Panel data is **strictly more efficient** than RCS for DiD ($V_p < V_{rc}$ always)

# Appendix: Optimal Sample Allocation for RCS

- **Question:** Given a total budget of $n$ observations, how should we split between $t = 1$ and $t = 2$ in an RCS?
- Let $\lambda = n_2/(n_1 + n_2)$ be the fraction sampled at $t = 2$
- The asymptotic variance is:

$$V_{rc}(\lambda) = \frac{1}{\lambda}\left(\frac{\sigma_{2,2}^2}{p} + \frac{\sigma_{\infty,2}^2}{1-p}\right) + \frac{1}{1-\lambda}\left(\frac{\sigma_{2,1}^2}{p} + \frac{\sigma_{\infty,1}^2}{1-p}\right)$$

- **Optimal allocation:**

$$\lambda^* = \frac{\sqrt{A_2}}{\sqrt{A_1} + \sqrt{A_2}}$$

where $A_t = \frac{\sigma_{2,t}^2}{p} + \frac{\sigma_{\infty,t}^2}{1-p}$

- **Practical implication:** If the outcome is more variable in the post-period (e.g., due to treatment effect heterogeneity), sample more observations in the post-period