ECON 730: Causal Inference with Panel Data

Lecture 6: Incorporating Covariates into DiD

Pedro H. C. Sant'Anna

EMORY
UNIVERSITY

Spring 2026

Roadmap

- **Act I: Why Covariates?** — Two applications, conditional parallel trends, TWFE fragility
- **Act II: Three Estimation Strategies** — Regression adjustment, IPW, doubly robust
- **Act III: The Design Phase & Full Applications** — Balance diagnostics, Medicaid, Brazil CAPS
- **Act IV: Repeated Cross-Sections** — Compositional changes & new solutions
- **Act V: Machine Learning & DiD** — LASSO, cross-fitting, causal forests

Building on Lecture 5 (2×2 DiD without covariates)

Act I: Why Covariates?

Where We Left Off

Lecture 5 established the 2×2 DiD framework:

1. SUTVA + No-Anticipation + **Unconditional** Parallel Trends ⇒ ATT identified
2. DiD-by-hand = TWFE regression (numerically identical in 2×2)
3. Influence functions provide asymptotic theory; **always cluster**
4. Weights define the target parameter

- But is **unconditional** parallel trends realistic?
- What if treated and comparison units differ systematically in pre-treatment characteristics?

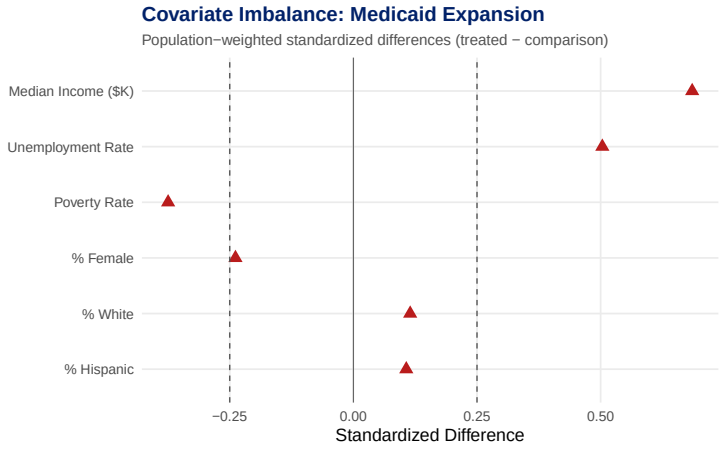
Meet the Applications

- **Application 1: ACA Medicaid Expansion** (Baker, Callaway, Cunningham, Goodman-Bacon and Sant'Anna, 2025)
  - Effect of Medicaid expansion on county-level mortality (2×2: 2013–2014)
  - Expansion states differ from non-expansion states in demographics, income, poverty
- **Application 2: Brazil Psychiatric Reform** (Dias and Fontes, 2024)
  - Community mental health centers (CAPS) replaced psychiatric hospitals
  - 5,180 municipalities, staggered rollout 2002–2016
  - For this lecture: 2×2 — CAPS adopters in 2006 vs. never-treated, pre/post = 2005/2007
  - Outcome: assault homicide rate per 10,000 population
- Both applications: treated and comparison groups differ in pre-treatment characteristics
- **Question:** Can we still use unconditional PT?

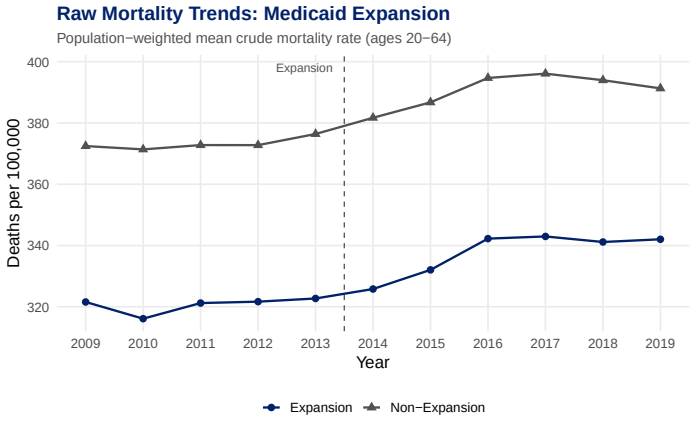
Medicaid: How Different Are the Groups?

- Treated states (expanded Medicaid in 2014) vs. comparison states (never expanded)
- Let's look at pre-treatment characteristics:
  - % below poverty line, median household income
  - % white, % Hispanic, urbanization rate
  - Pre-treatment mortality trends
- Do these groups look comparable?

Medicaid: Covariate Balance

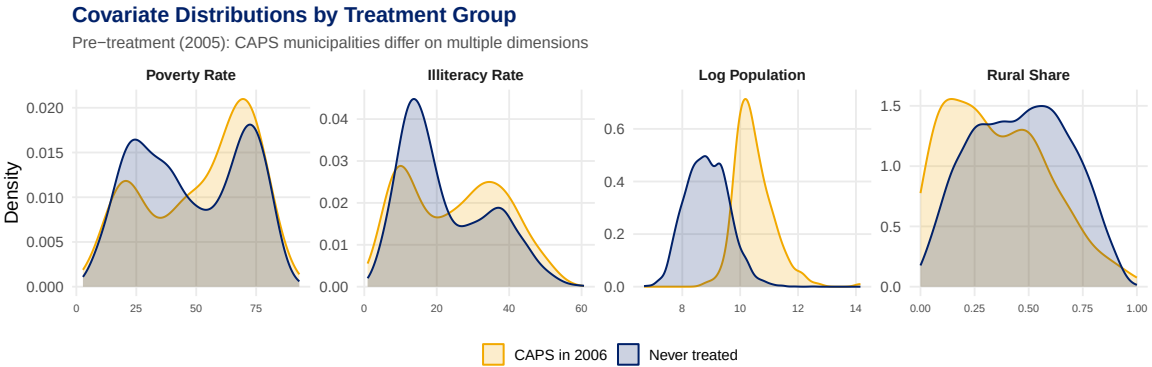


Medicaid: Raw Trends

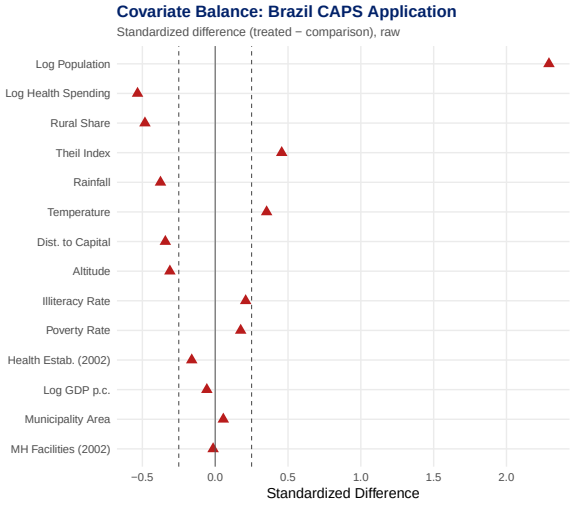


Pre-treatment trends look roughly parallel, but given the covariate differences we just saw — is unconditional PT enough?

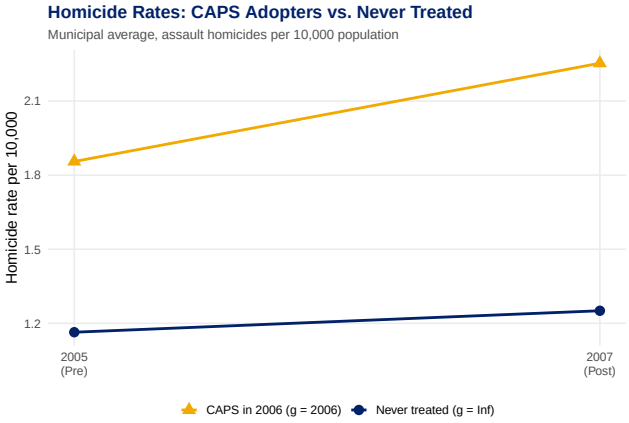
Brazil CAPS: Covariate Imbalance Across Dimensions



Brazil CAPS: Standardized Differences



Brazil CAPS: Raw Trends



Again, trends look roughly parallel — but given the covariate imbalance we just saw, is unconditional PT plausible here?

The Core Idea: Conditional Parallel Trends

- Sometimes the unconditional PT assumption is too strong
- But PT may be plausible **within subgroups** defined by pre-treatment characteristics X_i
- Intuition: **"Among counties with similar demographics, treated and comparison counties would have trended similarly absent treatment"**
- This is the **conditional parallel trends** assumption

Key insight: Covariates can make the PT assumption more credible, but we need appropriate estimation methods to exploit them.

Notation: Review from Lecture 5

- Two periods: $t = 1$ (pre-treatment) and $t = 2$ (post-treatment)
- Two groups: $G_i \in \{2, \infty\}$, with $D_i = \mathbf{1}\{G_i = 2\}$
- Potential outcomes: $Y_{i,t}(g)$ for each treatment timing g
- Target parameter:

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

Notation: New for Covariates

- X_i: vector of **pre-treatment covariates** (observed before treatment)
- $p(X_i) \equiv \mathbb{P}(D_i = 1 \mid X_i)$: **generalized propensity score**
- $p \equiv \mathbb{P}(D_i = 1) = \mathbb{E}[D_i]$: unconditional treatment probability
- $T_i \in \{1, 2\}$: **period indicator** for unit i (in panel data, each unit observed in both; in RCS, T_i is the sampling period)
- Later we will also define:
  - $m_\Delta^d(x) \equiv \mathbb{E}[\Delta Y_i \mid X_i = x, D_i = d]$: conditional mean outcome change
  - CATT(x): conditional ATT given $X_i = x$

Conditional Parallel Trends Assumption

We formalize the idea that parallel trends may hold only within covariate subgroups:

Assumption (Conditional Parallel Trends)

$$\mathbb{E}\left[Y_{i,t=2}(\infty) - Y_{i,t=1}(\infty) \mid X_i, D_i = 1\right] = \mathbb{E}\left[Y_{i,t=2}(\infty) - Y_{i,t=1}(\infty) \mid X_i, D_i = 0\right] \quad \text{a.s.}$$

- In words: conditional on X_i, the **average** evolution of $Y(\infty)$ is the same for treated and comparison units
- Allows for **covariate-specific trends**: outcome evolution can depend on X_i
- Remark: Caetano and Callaway (2024) condition on (X_{t^*}, X_{t^*-1}, Z), allowing time-varying covariates. We restrict to pre-determined baseline X_i to avoid "bad controls" concerns (Angrist and Pischke, 2009).

The Overlap Assumption

For identification, we also need treated units to have comparable controls at every covariate value:

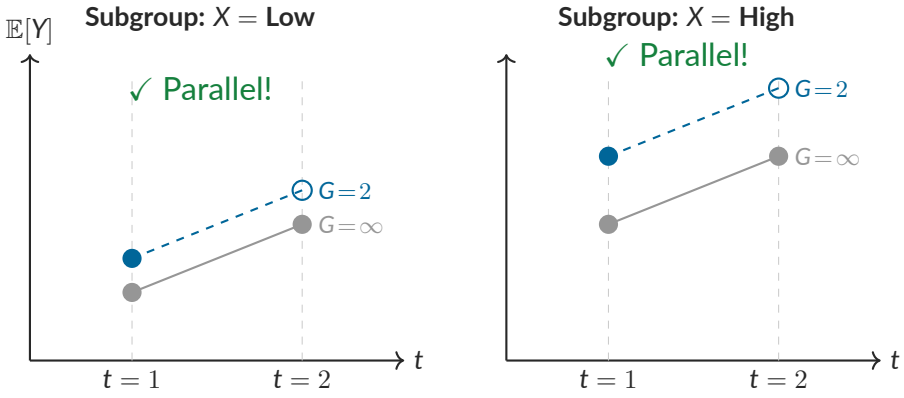
Assumption (Strong Overlap)

For some $\epsilon > 0$,

$$\mathbb{P}(D_i = 1 \mid X_i) < 1 - \epsilon \quad \text{almost surely.}$$

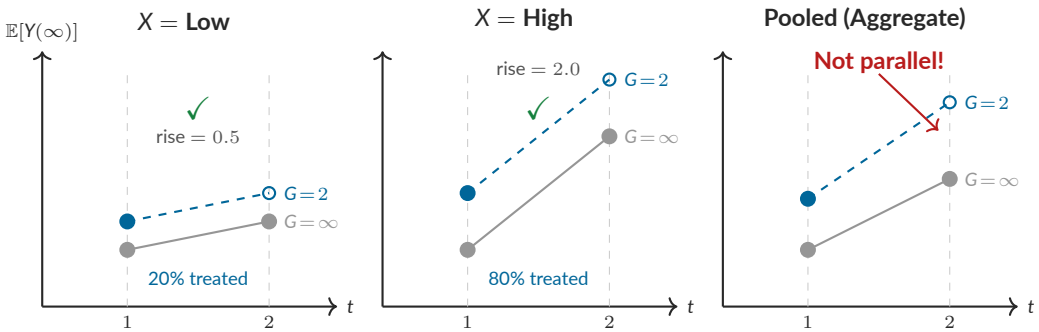
- Every treated unit must have comparison units with **similar** covariate values
- Without overlap: we cannot learn about the counterfactual for some treated units
- For **identification**: can relax to $\epsilon = 0$ (boundary case)
- For standard **inference**: need $\epsilon > 0$ to avoid irregularity (Khan and Tamer, 2010)
- Closely related to overlap conditions in the matching/weighting literature (Crump, Hotz, Imbens and Mitnik, 2009)
- Note: for ATT, we only need $p(X_i)$ bounded away from 1, not from 0 — unlike ATE, which

Conditional vs. Unconditional PT: A Visual



Solid = observed comparison trend. Dashed = PT counterfactual for the treated group. Within subgroups, these are parallel.

How Conditional PT Can Break Unconditional PT



Identification of ATT under Conditional PT

Assumptions: **A1.** SUTVA **A2.** No Anticipation: $Y_{i,t=1}(2) = Y_{i,t=1}(\infty)$ **A3.** Conditional PT **A4.** Overlap

Step 1: Identify the conditional ATT:

$$\text{CATT}(x) = \mathbb{E}[\Delta Y_i \mid X_i = x, D_i = 1] - \mathbb{E}[\Delta Y_i \mid X_i = x, D_i = 0]$$

where $\Delta Y_i \equiv Y_{i,t=2} - Y_{i,t=1}$.

Step 2: Integrate over the treated covariate distribution:

$$\text{ATT} = \mathbb{E}[\text{CATT}(X_i) \mid D_i = 1]$$

- We identify a very rich object: the **conditional ATT function** $\text{CATT}(x)$
- The unconditional ATT follows by averaging over treated units' covariates

The Practitioner's Instinct: Add X to TWFE

- The most common approach in applied work: "just add covariates to the regression"

$$Y_{i,t} = \alpha_i + \lambda_t + \tau D_{it} + X'_{i,t}\beta + \varepsilon_{i,t}$$

- Recall from Lecture 5: without covariates, TWFE $=$ DiD-by-hand in 2×2
- Many practitioners expect the same logic extends: "$\hat{\tau}$ should estimate the ATT after controlling for X"
- **This intuition is wrong.**

Adding X to TWFE is **not** the same as allowing for covariate-specific trends. The regression imposes strong — and often hidden — restrictions.

What Goes Wrong with TWFE + Covariates

- Consider the TWFE specification with pooled data:

$$Y_{i,t} = \tilde{\alpha}_0 + \tilde{\gamma}_0 D_i + \tilde{\lambda}_0 \mathbf{1}\{T_i = 2\} + \tilde{\beta}_0^{twfe}\big(D_i \cdot \mathbf{1}\{T_i = 2\}\big) + X_i'\tilde{\alpha}_1 + \tilde{\varepsilon}_{i,t}$$

- Write out the implied conditional means:

$$\begin{aligned}
\mathbb{E}[Y_{i,t} \mid D_i = 0, T = 1, X_i] &= \tilde{\alpha}_0 + X_i'\tilde{\alpha}_1 \\
\mathbb{E}[Y_{i,t} \mid D_i = 0, T = 2, X_i] &= \tilde{\alpha}_0 + \tilde{\lambda}_0 + X_i'\tilde{\alpha}_1 \\
\mathbb{E}[Y_{i,t} \mid D_i = 1, T = 1, X_i] &= \tilde{\alpha}_0 + \tilde{\gamma}_0 + X_i'\tilde{\alpha}_1 \\
\mathbb{E}[Y_{i,t} \mid D_i = 1, T = 2, X_i] &= \tilde{\alpha}_0 + \tilde{\gamma}_0 + \tilde{\lambda}_0 + \tilde{\beta}_0^{twfe} + X_i'\tilde{\alpha}_1
\end{aligned}$$

TWFE Imposes No Covariate-Specific Trends

- From the comparison group:

$$\mathbb{E}[Y \mid D = 0, T = 2, X] - \mathbb{E}[Y \mid D = 0, T = 1, X] = \tilde{\lambda}_0$$

- **The time trend does not depend on X!**

- Similarly for the treated group:

$$\mathbb{E}[Y \mid D = 1, T = 2, X] - \mathbb{E}[Y \mid D = 1, T = 1, X] = \tilde{\lambda}_0 + \tilde{\beta}_0^{twfe}$$

- This means:

$$\text{ATT}(X) = \tilde{\beta}_0^{twfe} \quad \text{for all } X$$

- **Treatment effects are forced to be homogeneous across covariate subgroups!**

- The very reason we introduced covariates — allowing for covariate-specific trends — is **assumed away** by the TWFE specification

TWFE forces homogeneous trends — but how bad is the bias in practice?

A controlled simulation where ATT $= 0$.

TWFE Bias: Monte Carlo Evidence

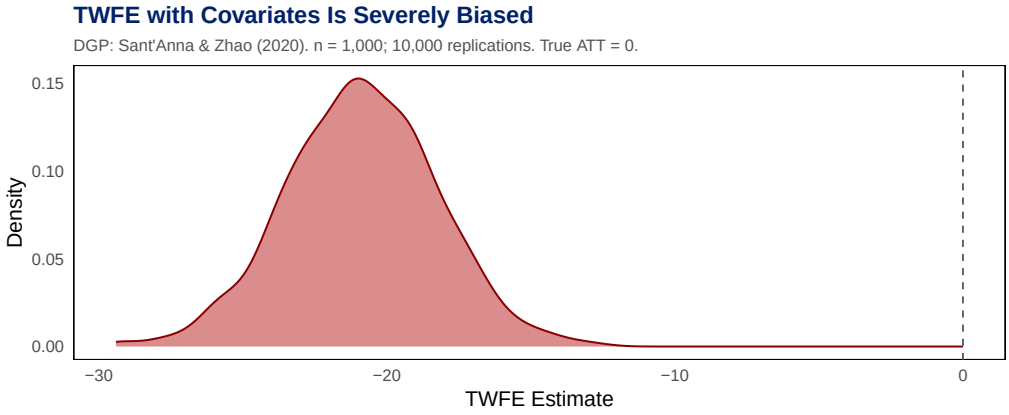
- Data generating process from Sant'Anna and Zhao (2020):
  - $X_j \sim N(0,1), j = 1, \ldots, 4$
  - Propensity score: logistic in $f_{ps}(X) = 0.75(-X_1 + 0.5X_2 - 0.25X_3 - 0.1X_4)$
  - Outcome regression: $f_{reg}(X) = 210 + 27.4\,X_1 + 13.7(X_2 + X_3 + X_4)$
  - Outcomes: $Y_{i,t}(\infty) = t \cdot f_{reg}(X_i) + v_i + \varepsilon_{i,t}$
  - True ATT$(X) = 0$ for all X

- TWFE regression: $Y_{i,t} = \alpha + \gamma D_i + \lambda \mathbf{1}\{T_i = 2\} + \tau\big(D_i \cdot \mathbf{1}\{T_i = 2\}\big) + X_i'\beta + \varepsilon_{i,t}$

- Results ($n = 1{,}000$, $1{,}000$ MC replications):

  - Average $\hat{\tau}^{twfe}$: -16.36 (true ATT $= 0$) — **severely biased!**
  - Coverage of 95% CI: 0% — **does not control size!**

TWFE Bias: Density Comparison



DGP 1 from Sant'Anna and Zhao (2020): covariates X observed. True ATT $= 0$, yet TWFE is severely biased even when all covariates are included in the regression.

The simulation used time-invariant X_i in a pooled regression.

What if covariates vary over time and we use fixed effects?

Caetano & Callaway (2024): a formal decomposition.

The FE Specification and Its Hidden Transformation

Caetano and Callaway (2024) analyze the standard fixed effects specification:

$$Y_{i,t} = \theta_t + \eta_i + \alpha D_{it} + X'_{i,t}\beta + e_{i,t}$$

- With two periods, the within/FD transformation eliminates η_i:

$$\Delta Y_i = \alpha D_i + \Delta X'_i \beta + \Delta e_i$$

- The transformation **also transforms the covariates**: only ΔX_i enters, not levels $X_{i,1}$
- Time-invariant covariates Z_i (e.g., race, region) are **completely absorbed** — cannot control for them
- This is the **hidden linearity bias**: the FE/FD form reveals restrictions that the levels specification obscures

Compare with the pooled specification on the previous slides, which controls for X_i in levels but forces homogeneous time trends.

Three Sources of Bias (Caetano and Callaway, 2024)

When conditional PT holds but TWFE is used with covariates:

$$\tilde{\beta}_0^{twfe} = \underbrace{\mathbb{E}[w(\Delta X) \cdot \text{ATT}(X) \mid D = 1]}_{\text{weighted ATT}} + \underbrace{\text{BIAS}_A}_{\text{time-invariant}} + \underbrace{\text{BIAS}_B}_{\text{levels vs. changes}} + \underbrace{\text{BIAS}_C}_{\text{nonlinearity}}$$

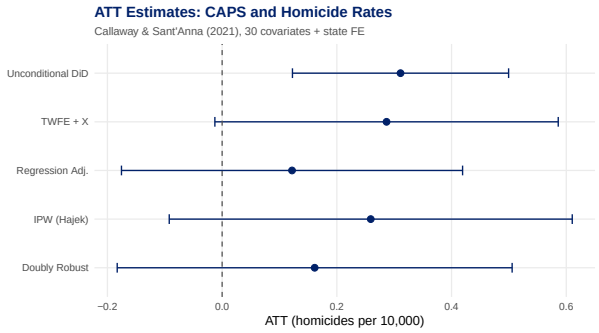
- **BIAS$_A$**: Time-invariant covariates Z_i absorbed by first-differencing — cannot control for them
- **BIAS$_B$**: TWFE only controls for **changes** ΔX, not **levels** X_{t-1}
- **BIAS$_C$**: Linear projection $\neq$ conditional expectation when the relationship is nonlinear
- Even the "weighted ATT" uses **non-transparent weights** $w(\Delta X)$ that can be **negative**

In Act II, we introduce three estimation strategies (RA, IPW, DR) that avoid all three biases by separating identification from estimation.

C&C showed TWFE with covariates introduces three sources of bias.
How do these biases play out in real data?

Back to the Brazil application.

Back to Brazil: All Estimators Compared



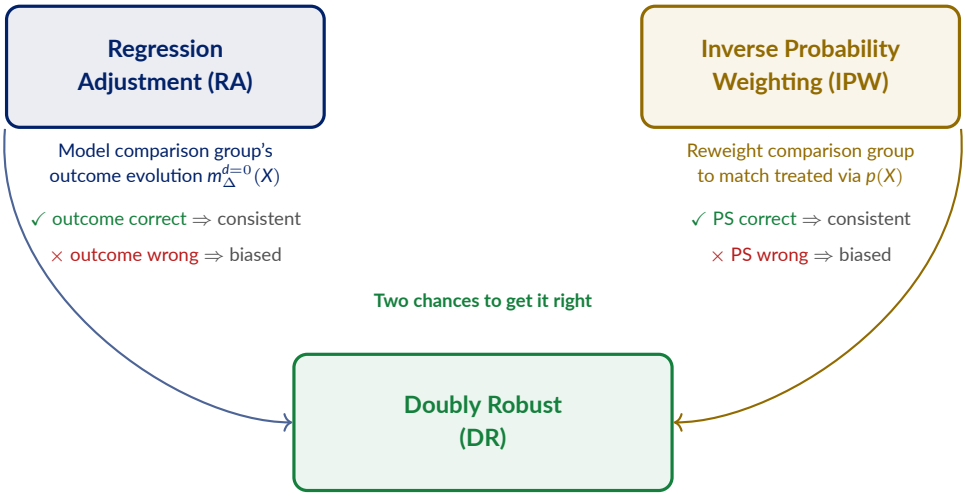
- 30 baseline covariates + state FE from Dias and Fontes (2024); RA, IPW, DR use levels, TWFE uses time-varying form (These three strategies are the subject of Act II.)
- **Overlap warning:** only 216 treated municipalities with 30 covariates + state FE

TWFE with covariates is fragile.

We need better tools: separate identification from estimation.

Act II: Three Estimation Strategies

Three Faces of DiD with Covariates



Act II: Three Estimation Strategies

Regression Adjustment

The First Face of DiD with Covariates: Regression Adjustment

- **Idea:** Model the comparison group's outcome evolution $\mathbb{E}[\Delta Y_i \mid X_i, D_i = 0]$, then impute for treated units
- With **panel data**, the ATT simplifies to:

$$\text{ATT} = \mathbb{E}[\Delta Y_i \mid D_i = 1] - \mathbb{E}\left[m_\Delta^{d=0}(X_i) \mid D_i = 1\right] = \mathbb{E}\left[m_\Delta^{d=1}(X_i) - m_\Delta^{d=0}(X_i) \mid D_i = 1\right]$$

  where $m_\Delta^{d=0}(x) \equiv \mathbb{E}[\Delta Y_i \mid X_i = x, D_i = 0]$

- Only need to model **one** conditional expectation: the comparison group's ΔY given X
- Originally proposed by Heckman, Ichimura and Todd (1997) and Heckman, Ichimura, Smith and Todd (1998)

Regression Adjustment: Estimation

- We need to estimate $m_{\Delta}^{d=0}(x) \equiv \mathbb{E}[\Delta Y_i \mid X_i = x, D_i = 0]$. A convenient choice is a **linear working model**: $m_{\Delta}^{d=0}(X_i) = X_i'\beta_0$
- **Step 1:** Estimate β_0 by OLS using comparison units only: $\hat{\beta}_n = \left(\sum_{i:D_i=0} X_i X_i'\right)^{-1} \sum_{i:D_i=0} X_i \Delta Y_i$
- **Step 2:** Impute for treated and average: $\hat{\theta}_n^{ra} = \frac{1}{n_1} \sum_{i:D_i=1} \left(\Delta Y_i - X_i'\hat{\beta}_n\right)$
- Any estimator of $m_{\Delta}^{d=0}(x)$ can be plugged in — kernel regression, random forests, LASSO, etc. Linear model is popular but **consistency requires correct specification**

Key difference from TWFE: regression is estimated on the **comparison group only**, then predictions are made for treated units. This allows covariate-specific trends.

Worked Example: RA with 5 Units

Unit i	D_i	X_i (poverty %)	ΔY_i (mortality change)
1	0 (comparison)	12	-2.0
2	0 (comparison)	18	-0.5
3	0 (comparison)	22	$+1.0$
4	1 (treated)	20	-3.0
5	1 (treated)	15	-2.5

Step 1: Regress ΔY_i on X_i using **comparison units only** $\Rightarrow \hat{\beta} = 0.30$

Step 2: Impute for treated: $\hat{m}_{\Delta}^{d=0}(20) = 0.30 \times 20 = 6.0$, $\hat{m}_{\Delta}^{d=0}(15) = 0.30 \times 15 = 4.5$
(intercept omitted for simplicity)

Step 3: $\widehat{\text{ATT}} = \frac{1}{2}\big[(-3.0 - 6.0) + (-2.5 - 4.5)\big] = -8.0$

The key: comparison group regression tells us "what would have happened to treated units if they hadn't been treated."

RA: Key Properties

- **Consistent** when the outcome model $m_{\Delta}^{d=0}(x)$ is correctly specified
- **Inconsistent** when $m_{\Delta}^{d=0}(x)$ is misspecified
- Works well when:
  - X is low-dimensional
  - Functional form is known or well-approximated
  - Good overlap (but does not explicitly reweight)

RA relies entirely on the researcher's ability to model the comparison group's outcome evolution.

Q: If RA is inconsistent under misspecification, why not always use a very flexible model for $m_{\Delta}^{d=0}(x)$?

RA in Practice: What Are We Actually Estimating?

Medicaid Expansion

- ΔY_i: change in county mortality
- X_i: poverty, income, % white, % Hispanic, urbanization
- $m_{\Delta}^{d=0}(X_i)$: **"How does mortality change in non-expansion counties with similar demographics?"**

Brazil CAPS Reform

- ΔY_i: change in homicide rate
- X_i: 30 municipal characteristics + state FE
- $m_{\Delta}^{d=0}(X_i)$: **"How do homicide rates change in non-CAPS municipalities with similar characteristics?"**

The RA recipe (same in both applications):

1. Estimate $\hat{m}_{\Delta}^{d=0}(x)$ using **comparison units only**
2. For each treated unit, plug in its X_i to get the predicted counterfactual change $\hat{m}_{\Delta}^{d=0}(X_i)$
3. $\widehat{\text{ATT}} = \frac{1}{n_1} \sum_{i:\, D_i=1} \left(\Delta Y_i - \hat{m}_{\Delta}^{d=0}(X_i)\right)$: mean observed change minus mean predicted counterfactual change among treated

RA models outcomes directly.

*What if we instead **reweight observations**?*

Act II: Three Estimation Strategies

Inverse Probability Weighting

The Second Face: Inverse Probability Weighting

- **Idea:** Instead of modeling outcomes, reweight the comparison group to "look like" the treated group in covariates
- Model the **propensity score**: $p(X_i) = \mathbb{P}(D_i = 1 \mid X_i)$
  With two groups (treated vs. never-treated), the PS is a single binary model
- Originally proposed by Abadie (2005):

$$\text{ATT}^{ipw} = \frac{\mathbb{E}\left[\left(D_i - \frac{(1-D_i)p(X_i)}{1-p(X_i)}\right)\Delta Y_i\right]}{\mathbb{E}[D_i]}$$

- The weights $\frac{p(X_i)}{1-p(X_i)}$ upweight comparison units that "resemble" treated units

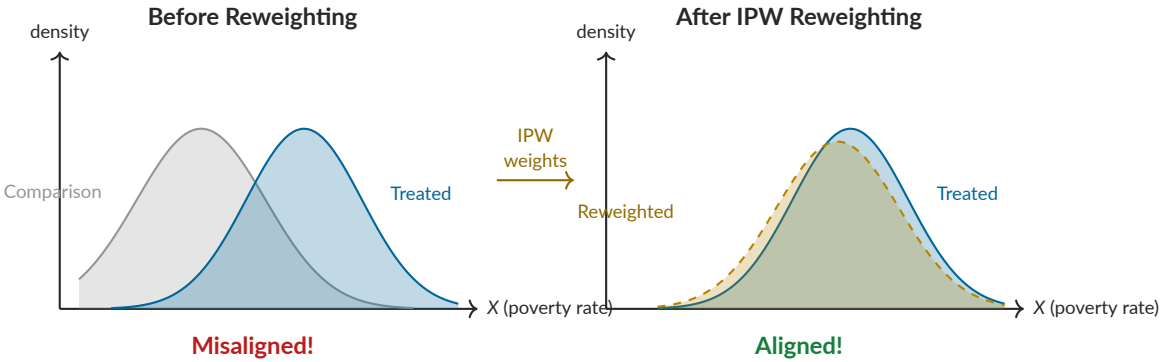
IPW: Normalized (Hájek) Weights

- Abadie (2005)'s IPW is of the **Horvitz–Thompson** type (weights do not sum to 1)
- Sant'Anna and Zhao (2020) proposed **Hájek-type** (normalized) weights:

$$\text{ATT}^{ipw}_{std} = \mathbb{E}\left[\left(\underbrace{\frac{D_i}{\mathbb{E}[D_i]}}_{w_1(D_i)} - \underbrace{\frac{\frac{p(X_i)\,(1-D_i)}{1-p(X_i)}}{\mathbb{E}\left[\frac{p(X_i)\,(1-D_i)}{1-p(X_i)}\right]}}_{w_0(D_i,X_i)}\right)\Delta Y_i\right]$$

- Normalized weights sum to 1 in each group $\Rightarrow$ more stable in finite samples
- Both versions are consistent under correct propensity score specification

IPW Reweighting: The Intuition



IPW uses $w = p(X)/(1 - p(X))$ to reweight comparison units so their covariate distribution matches the treated group.

IPW: Estimation and Key Properties

- **Working model:** Logistic $p(X_i; \gamma_0) = \Lambda(X_i'\gamma_0)$
- **Step 1:** Estimate γ_0 by logit MLE
- **Step 2:** Plug in $\hat{p}(X_i)$ and compute weighted averages
- Influence function accounts for estimation error in $\hat{\gamma}_n$
- **Consistent** when propensity score is correctly specified
- **Inconsistent** when propensity score is misspecified — even if outcome model is known!
- Overlap is **critical**: if $p(X_i) \approx 1$, weights explode

IPW in Practice: What Are We Actually Reweighting?

Medicaid Expansion

- $p(X_i)$: prob. county's state expands Medicaid, given demographics
- Counties "resembling" expansion states get **upweighted**; dissimilar ones **downweighted**
- 6 covariates — overlap manageable

Brazil CAPS Reform

- $p(X_i)$: prob. municipality adopts CAPS in 2006, given 30 covariates + state FE
- Non-CAPS municipalities resembling adopters get **upweighted**
- **Overlap concern:** only 216 treated with high-dimensional X

The IPW recipe (same in both applications):

1. Estimate $\hat{p}(X_i)$ using both treated and comparison units (e.g., logit)
2. Reweight comparison units by $\hat{p}(X_i)/(1-\hat{p}(X_i))$: units resembling treated get more weight
3. $\widehat{\text{ATT}}$: weighted mean ΔY_i among treated minus **reweighted** mean among comparison

RA vs. IPW: Complementary Strengths

	RA	IPW
Models	Outcome evolution	Treatment assignment
Consistent when	$m_{\Delta}^{d=0}(x)$ correct	$p(x)$ correct
Fails when	Outcome misspecified	PS misspecified
Sensitive to	Functional form	Overlap violations

RA and IPW have **complementary** failure modes. Can we combine them to get robustness against **either** type of misspecification?

RA and IPW each rely on one model being correct.

*Can we **combine them** for robustness against **either** type of misspecification?*

Act II: Three Estimation Strategies

Doubly Robust DiD

The Third Face: Doubly Robust Estimation

- **Key idea:** Combine outcome modeling (RA) with reweighting (IPW)
- Consistent if **either** the outcome model **or** the propensity score is correctly specified (but not necessarily both)

DR DiD Estimand (Sant'Anna and Zhao, 2020):

$$\text{ATT}^{dr} = \mathbb{E}\Big[\Big(w_1(D_i) - w_0(D_i, X_i)\Big)\Big(\Delta Y_i - m_{\Delta}^{d=0}(X_i)\Big)\Big]$$

Treated weight

$$w_1(D_i) = \frac{D_i}{\mathbb{E}[D_i]}$$

Comparison weight

$$w_0(D_i, X_i) = \frac{\frac{p(X_i)\,(1-D_i)}{1-p(X_i)}}{\mathbb{E}\left[\frac{p(X_i)\,(1-D_i)}{1-p(X_i)}\right]}$$

Why Is It Doubly Robust?

- The DR estimand has **two** equivalent decompositions:

$$\begin{aligned}\text{ATT}^{dr} &= \underbrace{\text{ATT}^{ipw}_{std}}_{\text{IPW}} - \underbrace{\mathbb{E}\Big[\big(w_1(D_i) - w_0(D_i, X_i)\big)\, m_{\Delta}^{d=0}(X_i)\Big]}_{\text{Outcome-based bias correction}} \\ &= \underbrace{\text{ATT}^{ra}}_{\text{RA}} - \underbrace{\mathbb{E}\Big[w_0(D_i, X_i)\,\Big(\Delta Y_i - m_{\Delta}^{d=0}(X_i)\Big)\Big]}_{\text{Reweighting-based bias correction}}\end{aligned}$$

- **If $p(x)$ correct:** w_0 rebalances $\Rightarrow$ first line's correction is mean-zero $\Rightarrow$ consistent
- **If $m_{\Delta}^{d=0}(x)$ correct:** residuals are mean-zero $\Rightarrow$ second line's correction vanishes $\Rightarrow$ consistent
- **If both wrong:** generally inconsistent, but bias is product of two errors

Double Robustness: A Scorecard

	PS Correct	PS Wrong
OR Correct	RA ✓ IPW ✓ DR ✓	RA ✓ IPW × DR ✓
OR Wrong	RA × IPW ✓ DR ✓	RA × IPW × DR ×

DR: consistent in 3 out of 4 scenarios — two chances to get it right

DR in Practice: Two Models, Two Chances

Medicaid Expansion

- **OR:** how does mortality change in non-expansion counties with similar demographics?
- **PS:** which counties look like expansion counties based on demographics?
- 6 covariates — both models tractable

Brazil CAPS Reform

- **OR:** how do homicide rates change in non-CAPS municipalities with similar characteristics?
- **PS:** which municipalities look like CAPS adopters, given 30 covariates + state FE?
- High-dimensional X — DR's insurance especially valuable

The DR recipe (same in both applications):

1. Estimate **both** $\hat{m}_{\Delta}^{d=0}(x)$ (on comparison units) and $\hat{p}(X_i)$ (on all units)
2. Combine: use IPW weights **and** outcome residuals $\Delta Y_i - \hat{m}_{\Delta}^{d=0}(X_i)$
3. If either model is correct, the other's errors wash out $\Rightarrow$ **two chances** to get it right

*DR is consistent if **either** model is correct.*

But what happens to precision when both are right?

The Semiparametric Efficiency Bound

Q: If DR gives us two chances, why care about getting **both** models right?

- Sant'Anna and Zhao (2020) derive the **semiparametric efficiency bound** for ATT under conditional PT
- The bound equals the variance of the **efficient influence function**:

$$\psi_i^{eff} = \big(w_1(D_i) - w_0(D_i, X_i; p_0)\big)\big(\Delta Y_i - m_\Delta^{d=0}(X_i)\big) - w_1(D_i) \cdot \text{ATT}$$

- The DR estimand's IF **equals** the efficient IF when both models are correct
- This means: **DR attains the semiparametric efficiency bound** when both $m_\Delta^{d=0}$ and p are correctly specified — it is **locally efficient**

Improved vs. Traditional DR

- Sant'Anna and Zhao (2020) propose two versions:
  - **Traditional DR** (`drdid_panel`): standard logit PS + OLS outcome model
  - **Improved DR** (`drdid_imp_panel`): **inverse probability tilted** PS (Graham, Pinto and Egel, 2012) + weighted OLS outcome model
- The improved version ensures the estimated PS satisfies an exact **balancing condition**, improving finite-sample performance
- Both are doubly robust and locally efficient under correct specification
- **Bonus:** The improved DR estimator is also **doubly robust for inference** — no need to adjust standard errors for first-step estimation of $p(X_i)$ or $m_{\Delta}^{d=0}(X_i)$

RA, IPW, and DR have different robustness and efficiency properties.

Do these properties hold in finite samples?

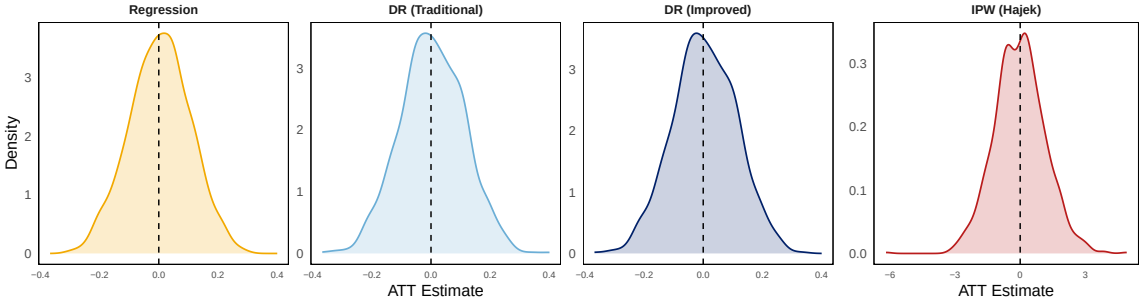
Monte Carlo: Comparing All Estimators

- DGP with true ATT $= 0$:
  - 4 DGPs: vary correct/incorrect outcome and PS models
  - **DGP 1:** Both correctly specified
  - **DGP 2:** PS misspecified, outcome correct
  - **DGP 3:** PS correct, outcome misspecified
  - **DGP 4:** Both misspecified
- 7 estimators: Oracle (infeasible), DR-Improved, DR-Traditional, IPW, IPW-Normalized, RA, TWFE
- $n = 500, 1{,}000$ MC replications

Monte Carlo: DGP 1 — Both PS and OR Correctly Specified

DGP 1: Both correct

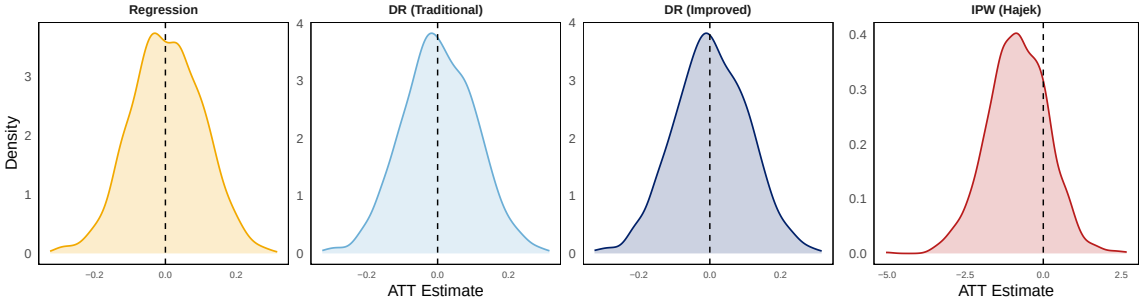
True ATT = 0. n = 1,000; 1,000 replications. TWFE omitted (off−scale). IPW x−axis differs.



Monte Carlo: DGP 2 — Propensity Score Misspecified

DGP 2: PS wrong, OR correct

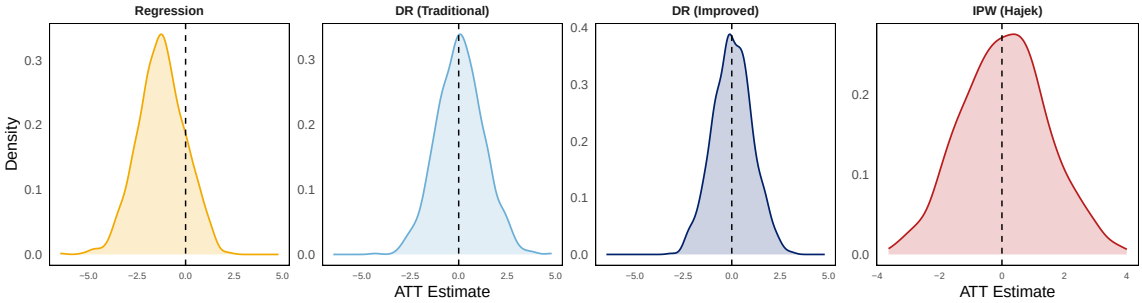
True ATT = 0. n = 1,000; 1,000 replications. TWFE omitted (off−scale). IPW x−axis differs.



Monte Carlo: DGP 3 — Outcome Regression Misspecified

DGP 3: PS correct, OR wrong

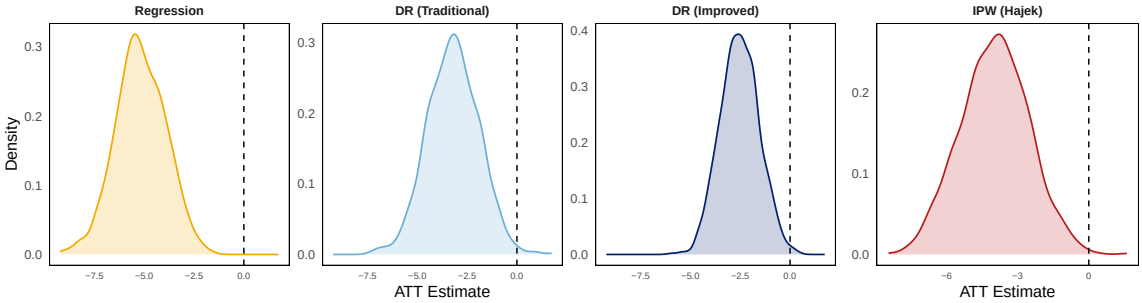
True ATT = 0. n = 1,000; 1,000 replications. TWFE omitted (off−scale). IPW x−axis differs.



Monte Carlo: DGP 4 — Both Misspecified

DGP 4: Both wrong

True ATT = 0. n = 1,000; 1,000 replications. TWFE omitted (off−scale). IPW x−axis differs.



Monte Carlo Summary: Bias and RMSE

	DGP 1 Both correct		DGP 2 PS wrong		DGP 3 OR wrong		DGP 4 Both wrong	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
TWFE	−20.9	21.1	−20.5	20.6	−28.2	28.3	−16.4	16.5
Regression	0.0	0.1	0.0	0.1	**−6.1**	6.2	**−5.2**	5.3
IPW (Hajek)	0.0	1.2	**−1.9**	2.2	0.0	1.3	**−4.0**	4.2
DR (Trad.)	0.0	0.1	0.0	0.1	0.0	1.0	**−3.2**	3.5
DR (Impr.)	0.0	0.1	0.0	0.1	0.0	1.0	**−1.0**	2.6

- DR unbiased whenever **at least one** model is correct (DGPs 1–3)
- TWFE severely biased in **all** DGPs — nonlinear X-dependence breaks linearity
- DGP 4: DR has smaller bias (product of two misspecification errors) — but is **not** consistent

Monte Carlo Summary: Coverage

	DGP 1 Both correct	DGP 2 PS wrong	DGP 3 OR wrong	DGP 4 Both wrong
TWFE	0.0%	0.0%	0.0%	0.0%
Regression	93.9%	94.9%	83.8%	1.1%
IPW (Hajek)	94.0%	83.7%	95.2%	22.0%
DR (Trad.)	95.3%	94.9%	94.6%	28.4%
DR (Impr.)	94.8%	94.4%	94.6%	26.8%

DR is the only estimator that performs well across all scenarios. TWFE should **not** be the default when covariates matter.

Two chances to get it right, and efficient when both models are correct.

Doubly robust is the default for DiD with covariates.

Act III: The Design Phase & Full Applications

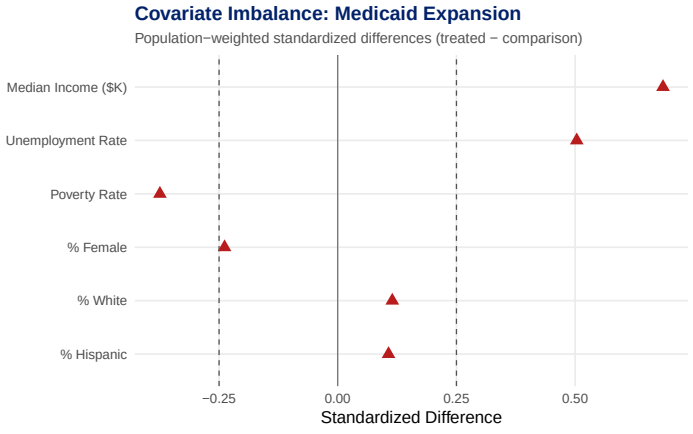
Covariate Balance and the Plausibility of PT

- If covariates that are important for outcome changes in the absence of treatment are **unbalanced** across treated and comparison groups, this raises serious concerns about unconditional PT (Abadie, 2005)
- Intuition: if groups differ in X, and X drives $\Delta Y(\infty)$, then $\mathbb{E}[\Delta Y(\infty) \mid D = 1] \neq \mathbb{E}[\Delta Y(\infty) \mid D = 0]$
- This motivates **covariate balance diagnostics** as part of any DiD analysis — following the broader principle that "design trumps analysis" (Rubin, 2008; Baker et al., 2025)
- Key diagnostics:
  - Unweighted standardized differences: $\frac{\bar{X}_1 - \bar{X}_0}{\sqrt{(s_1^2 + s_0^2)/2}}$
  - **IPW-weighted** standardized differences: does reweighting restore balance?
  - Propensity score overlap: are there regions of X with no comparison units?
- Good balance $\Rightarrow$ more credible results, less sensitivity to specification choices

Medicaid: Context and Covariates

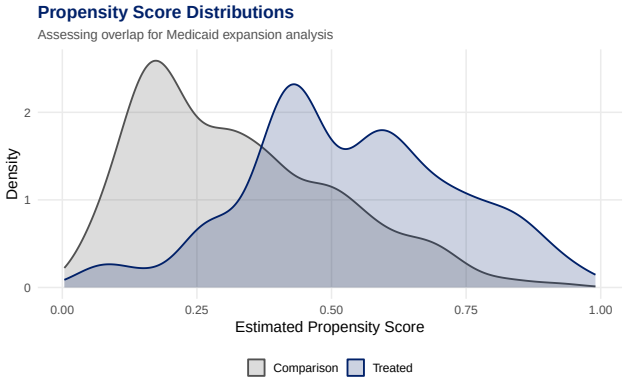
- **Setting:** Effect of Medicaid expansion on county-level mortality (Lecture 5 data)
- Now incorporate county-level covariates:
  - % white, % Hispanic, % female
  - Unemployment rate, poverty rate
  - Median household income
- **Why covariates matter:** Expansion states systematically differ from non-expansion states on these characteristics
- Conditional PT more plausible than unconditional PT: **"Among counties with similar demographics, mortality trends would be parallel absent expansion"**

Medicaid: Covariate Imbalance (Recap)



Expansion and non-expansion states differ systematically. Several covariates exceed the ± 0.25 threshold — motivating conditional PT.

Medicaid: Propensity Score Overlap



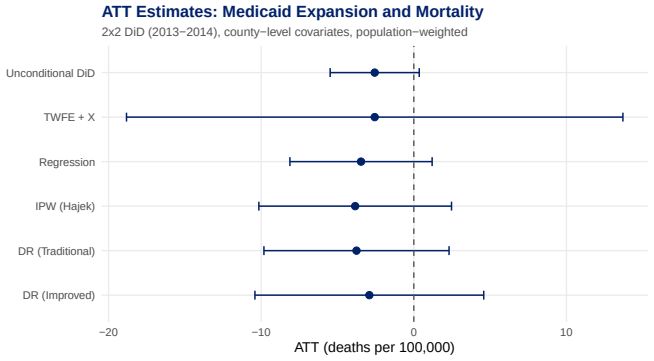
- Comparison counties' PS mostly within the support of expansion counties
- A few untreated counties have $\hat{p}(X_i)$ close to 1 — `did` and `DRDID` trim units with $\hat{p}(X_i) > 0.995$ by default

Medicaid: Covariate Balance Table (Population-Weighted)

Covariate	Treated	Comparison	Std. Diff.
% White	79.5	77.9	$+0.115$
% Hispanic	18.9	17.0	$+0.107$
% Female	50.1	50.5	-0.238
Unemp. Rate	8.0	7.0	$+0.503$
Poverty Rate	15.3	17.2	-0.375
Median Income ($K)	57.9	49.3	$+0.685$

- Expansion counties are wealthier, higher unemployment, less poverty
- Several covariates exceed ± 0.25 threshold — unconditional PT is questionable

Medicaid: All Estimators Compared (Population-Weighted)



- 6 baseline covariates, population-weighted; all estimates negative (mortality reduction)
- All estimators broadly agree; wide CIs reflect limited power at county level

Medicaid: What We Learned

- Wide confidence intervals — limited power with county-level data
- None of the estimates are statistically significant
- But the **sensitivity of estimates** to covariate inclusion is itself informative:
  - If results change dramatically with covariates, conditional PT is substantively different from unconditional PT
  - If results are stable, the unconditional DiD was already capturing the right comparison
- Good overlap and balance — the “design” checks out
- Covariates matter for **credibility** even if they do not dramatically change point estimates

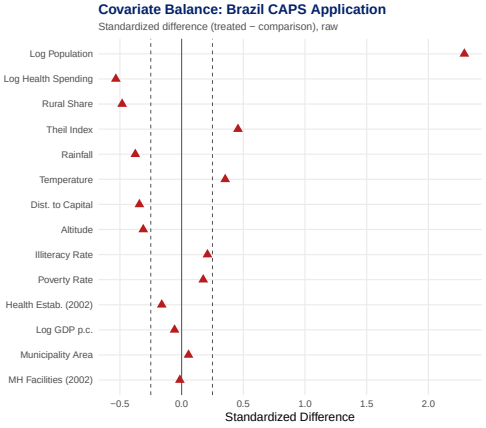
Medicaid had 6 covariates and good overlap.
What happens with a richer covariate set?

Brazil's psychiatric reform: 30 covariates + state FE.

Brazil CAPS: Full Context

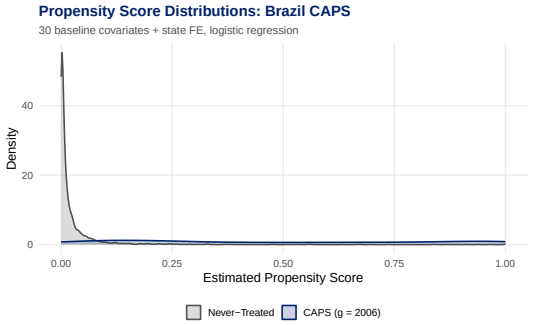
- Dias and Fontes (2024): Brazil's 2002 Psychiatric Reform created CAPS (community mental health centers) replacing psychiatric hospitals
- Staggered rollout across 5,180 municipalities (2002–2016)
- **Our 2×2 setup:** g = 2006 (early CAPS adopters) vs. never-treated; pre = 2005, post = 2007
- **Outcome:** Assault homicide rate per 10,000 population
- **30 covariates:** Demographics, income, transfers, poverty, geographic characteristics, health infrastructure (from 2000 census and administrative data) + state fixed effects
- Surprising finding: CAPS adoption **increases** homicides — consistent with the Penrose hypothesis (see Dias and Fontes, 2024) that deinstitutionalization reduces incapacitation

Brazil: Covariate Balance



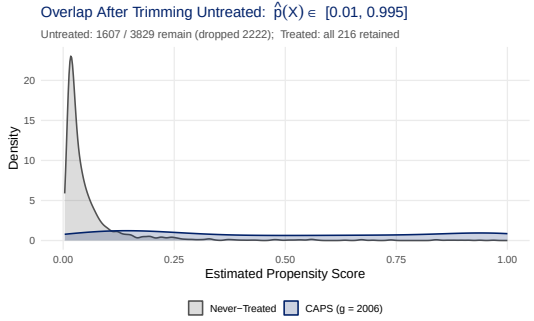
Standardized differences (unweighted). Several covariates exceed the ± 0.25 threshold.

Brazil: Propensity Score Overlap



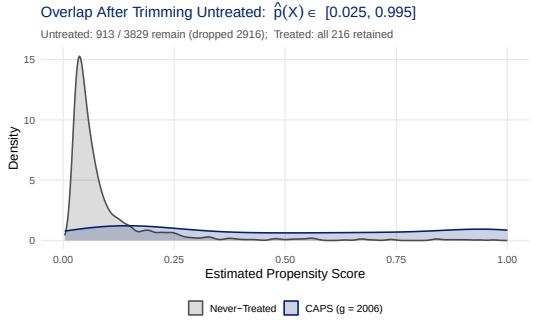
- Massive spike near 0: most untreated municipalities look nothing like CAPS adopters
- But $\hat{p}(X_i) \approx 0 \Rightarrow$ weights $\frac{p(X)}{1-p(X)} \approx 0$: these units **naturally drop out** of IPW
- The real concern: $\hat{p}(X_i) \approx 1$ among untreated, where weights **explode**

Brazil: Overlap After Trimming Untreated ($\hat{p}(X) \in [0.01, 0.995]$)



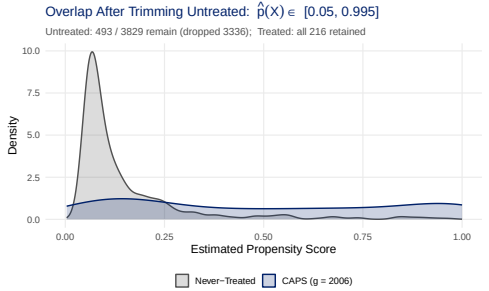
- Trimming untreated with $\hat{p}(X_i) < 0.01$ or > 0.995 drops 58% of comparison (2,222/3,829)
- All 216 treated units retained — trimming applies only to untreated
- Still a large spike near 0 among remaining untreated

Brazil: Overlap After Trimming Untreated ($\hat{p}(X) \in [0.025, 0.995]$)



- Trimming untreated at 0.025 drops 76% of comparison (2,916/3,829) — only 913 remain
- Overlap improves, but treated distribution still much more spread out
- Aggressive trimming changes the effective comparison group substantially

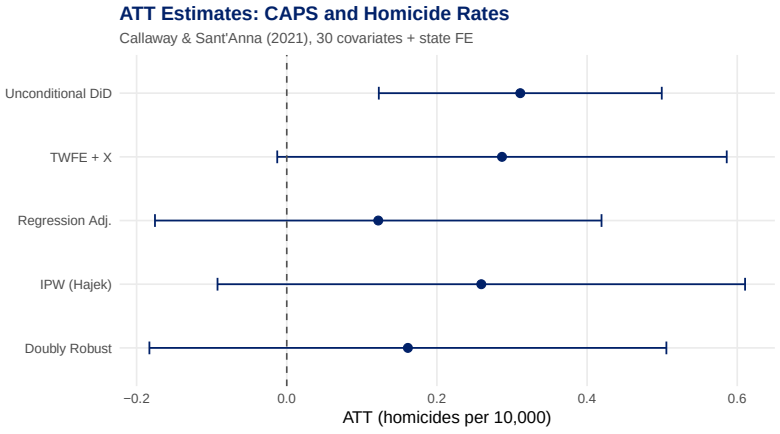
Brazil: Overlap After Trimming Untreated ($\hat{p}(X) \in [0.05, 0.995]$)



- Trimming untreated at 0.05 drops 87% — only 493 of 3,829 remain
- Overlap finally reasonable, but we lost most of the comparison group

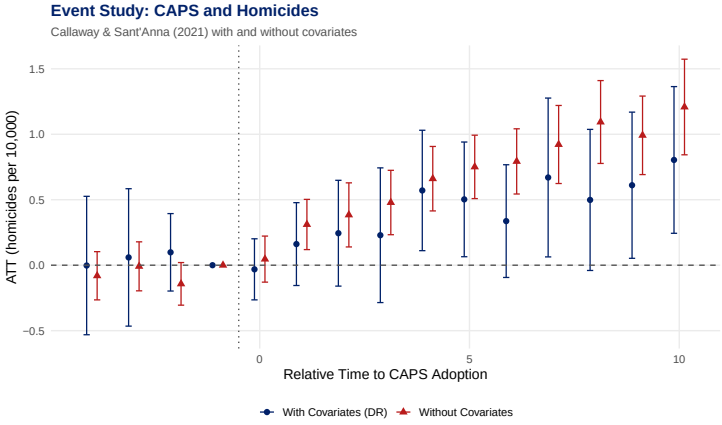
`did`/`DRDID` default: trim untreated with $\hat{p}(X_i) > 0.995$ only; units with $\hat{p}(X_i) \approx 0$ self-trim via vanishing weights

Brazil: Results — All Estimators



Same 30 covariates + state FE as Act I. All conditional estimators (RA, IPW, DR) use identical specification via `att_gt()`.

Brazil: Event Study with Covariates



Two groups (g = 2006 vs. never-treated), multiple periods. DR with covariates yields smaller pre-trend coefficients and more precise post-treatment estimates. We formalize event studies in later lectures.

Key Insight: Covariates and Credibility

- In both applications, covariates may not **dramatically change** point estimates
- But they **dramatically change credibility**:
  - Unconditional PT is a strong assumption when groups differ
  - Showing that conditional PT gives similar results **strengthens** the case
  - Showing that they differ **reveals** that the baseline was contaminated
- The "design phase" (balance diagnostics) is crucial for transparency
- **Bottom line:** Even when estimates are stable, the exercise of checking matters

In these data, covariates predict treatment adoption but DR and unconditional estimates are broadly similar—the exercise of checking is what matters.

We have seen the theory, diagnostics, and empirical results.

How do we implement this in practice?

Software: The `did` Package (Primary)

```r
library(did)

# Callaway & Sant'Anna (2021) with covariates
# Uses doubly robust estimation by default
result <- att_gt(
  yname  = "l_homicide",
  tname  = "year",
  idname = "sid",
  gname  = "first_treat",
  xformla = ~ x1 + x2 + x3,
  data   = my_data,
  control_group = "notyettreated",
  est_method = "dr",           # DR is the default
  base_period = "universal"    # use first period as base
)

# Aggregate to event study
es <- aggte(result, type = "dynamic")
```

Software: The DRDID Package (Low-Level)

```r
library(DRDID)

# Panel data: Doubly Robust DiD (improved)
result_dr <- drdid(yname = "y", tname = "post",
                   idname = "id", dname = "treat",
                   xformla = ~ x1 + x2 + x3,
                   data = panel_data, panel = TRUE)

# Also available: ipwdid(), ordid()
# For low-level functions, note the intercept convention:
#   drdid_imp_panel: needs cbind(1, X) explicitly
#   twfe_did_panel:  adds intercept internally (do NOT add)
```

The did package wraps DRDID and handles the intercept convention automatically. Use DRDID for more control.

Practitioner Checklist

Step-by-step guide for DiD with covariates:

1. **Specify:** Which **pre-determined** covariates make conditional PT plausible? Only condition on X_i measured before treatment and not affected by it.
2. **Check overlap:** Plot propensity score distributions. Trim if needed.
3. **Balance:** Compare covariate means across treated and comparison groups.
4. **Estimate:** Use DR as the default. Report RA and IPW as robustness.
5. **Sensitivity:** How much do results change with/without covariates?
6. **For repeated cross-sections:** Test for compositional changes (see Act IV).

What if we observe repeated cross-sections instead of a panel?

New challenges: compositional changes and stationarity.

Act IV: Repeated Cross-Sections

Panel vs. Repeated Cross-Sections: Recap

- **Panel data:** Observe same units in both periods
  - Can compute $\Delta Y_i = Y_{i,t=2} - Y_{i,t=1}$ directly
  - Only need one outcome model: $m_{\Delta}^{d=0}(x) = \mathbb{E}[\Delta Y_i \mid X_i = x, D_i = 0]$
- **Repeated cross-sections (RCS):** Different units sampled each period
  - Cannot first-difference
  - Need to model outcomes **separately** in each period
  - Requires additional assumptions about the sampling process
- From Lecture 5: panel data is **strictly more efficient** than RCS
- With covariates, the gap between panel and RCS has additional nuances

RCS Sampling Assumption

RCS Sampling

The pooled RCS data $\{Y_i, D_i, X_i, T_i\}_{i=1}^n$ consist of iid draws from the mixture distribution

$$\begin{aligned} P(Y \le y,\, D = d,\, X \le x,\, T = t) = {} & \mathbf{1}\{t = 2\} \cdot \lambda \cdot P(Y_2 \le y,\, D = d,\, X \le x \mid T = 2) \\ & + \mathbf{1}\{t = 1\} \cdot (1 - \lambda) \cdot P(Y_1 \le y,\, D = d,\, X \le x \mid T = 1) \end{aligned}$$

where $(y, d, x, t) \in \mathbb{R} \times \{0, 1\} \times \mathbb{R}^k \times \{1, 2\}$ and $\lambda = \mathbb{P}(T = 2) \in (0, 1)$.

- Each unit i is observed in **exactly one** period; $Y_i = \mathbf{1}\{T_i = 2\} \cdot Y_{i,2} + \mathbf{1}\{T_i = 1\} \cdot Y_{i,1}$
- In panel data, we observe $(Y_{i,1}, Y_{i,2})$ for each unit; in RCS, we observe Y_i for **one** period only
- Define outcome regressions per cell: $m_t^d(x) = \mathbb{E}[Y_i \mid D_i = d, T_i = t, X_i = x]$

Conditional PT for Repeated Cross-Sections

- The conditional PT assumption is the **same** as in the panel case (Sant'Anna and Zhao, 2020):

$$\mathbb{E}[Y_{i,t=2}(\infty) - Y_{i,t=1}(\infty) \mid D_i = 1, X_i] = \mathbb{E}[Y_{i,t=2}(\infty) - Y_{i,t=1}(\infty) \mid D_i = 0, X_i] \quad \text{a.s.}$$

- What changes with RCS is the **data structure**, not the assumption:
  - Different units sampled in each period $\Rightarrow$ cannot first-difference
  - Need to model $m_t^{d=0}(x) = \mathbb{E}[Y \mid D = 0, T = t, X = x]$ separately for $t = 1, 2$

- Key additional requirement: **stationarity** of the joint distribution of (D, X) across sampling periods (Assumption 2(b) in Sant'Anna and Zhao, 2020)

- With panel data, stationarity is automatic (same units observed twice)

- The conditional PT is stated for the superpopulation; the RCS sampling assumption ensures we can identify these quantities from the observed cross-sectional data.

The Stationarity Assumption

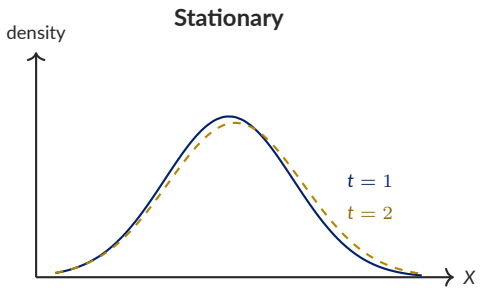
Assumption (Stationarity / No Compositional Changes)

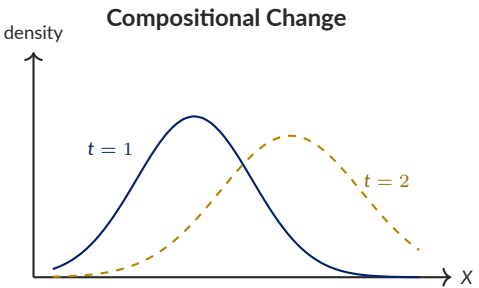
The joint distribution of (G_i, X_i) is the same across time periods:

$$(G_i, X_i) \mid T_i = 1 \overset{d}{=} (G_i, X_i) \mid T_i = 2$$

- In words: the "composition" of units sampled in each period is stable
- **Automatic in panel data** (same units observed each period)
- **Not automatic in RCS:** differential migration, attrition, survey redesigns can change who is sampled
- Standard DR DiD estimators for RCS **assume** stationarity

Compositional Changes: A Visual





When the covariate distribution shifts between periods (migration, attrition, survey redesign), standard RCS estimators are biased.

IPW for Repeated Cross-Sections

- With RCS, IPW must reweight across **both** treatment groups and time periods:

IPW for RCS (Abadie, 2005):

$$\text{ATT}^{ipw,rc} = \frac{1}{\mathbb{E}[D]}\,\mathbb{E}\left[\frac{D - p(X)}{1 - p(X)}\,\frac{T - \lambda}{\lambda(1-\lambda)}\,Y\right]$$

- Compared to panel IPW, the RCS version uses an additional reweighting factor $\frac{T-\lambda}{\lambda(1-\lambda)}$ that adjusts for the time dimension
- Still requires the same overlap condition: $p(X_i) < 1$ a.s.
- The propensity score $p(X) = \mathbb{P}(D = 1 \mid X)$ is estimated on **pooled data across both periods** — this is valid under the no-compositional-changes assumption

Efficient DR DiD for Repeated Cross-Sections: The Estimand

The efficient DR estimand, derived from the EIF (Sant'Anna and Zhao, 2020), models **all four** (d, t) cells:

$$\text{ATT}_{eff}^{dr,rc} = \mathbb{E}\Bigg[\underbrace{\frac{D}{\mathbb{E}[D]}\left(m_{\Delta}^{d=1}(X) - m_{\Delta}^{d=0}(X)\right)}_{\text{RA component}} + \underbrace{\left(w_{t=2}^{d=1}\left(Y - m_{t=2}^{d=1}(X)\right) - w_{t=1}^{d=1}\left(Y - m_{t=1}^{d=1}(X)\right)\right)}_{\text{treated bias correction}} - \underbrace{\left(w_{t=2}^{d=0}\left(Y - m_{t=2}^{d=0}(X)\right) - w_{t=1}^{d=0}\left(Y - m_{t=1}^{d=0}(X)\right)\right)}_{\text{comparison bias correction}} \Bigg]$$

Same DR logic as the panel case: RA plus IPW-based corrections for both groups. Key difference: each period needs its own outcome model and weight. (Weights defined on next slide.)

RCS Efficient DR DiD: The Four Hájek Weights

Where $m^d_\Delta(x) = m^d_{t=2}(x) - m^d_{t=1}(x)$. The estimand requires four Hájek weights — two for the treated group, two for the comparison:

Treated weights (simple: select treated at each period, normalize):

$$w_t^{d=1}(D_i, T_i) = \frac{D_i \cdot \mathbf{1}\{T_i = t\}}{\mathbb{E}[D \cdot \mathbf{1}\{T = t\}]} \quad \text{for } t = 1, 2$$

Comparison weights (reweight comparison to match treated covariate distribution):

$$w_t^{d=0}(D_i, T_i, X_i) = \frac{\frac{p(X_i)(1-D_i)\,\mathbf{1}\{T_i=t\}}{1-p(X_i)}}{\mathbb{E}\left[\frac{p(X)(1-D)\,\mathbf{1}\{T=t\}}{1-p(X)}\right]} \quad \text{for } t = 1, 2$$

Compare with panel DR DiD: instead of one weight pair, RCS needs a pair per period because we cannot track the same units across time.

Compositional Changes: The Problem

- **Compositional changes:** the distribution of (G, X) differs across periods
- Happens when:
  - Migration: people move in/out of regions across survey waves
  - Attrition: some types of units drop out differentially
  - Survey redesign: sampling frame changes between waves
  - Natural disasters, policy changes that affect who is "at risk"
- When stationarity fails, **standard RCS estimators are biased**
- The bias arises because the "comparison group trend" is contaminated by compositional shifts

Sant'Anna & Xu (2026): DiD with Compositional Changes

- Sant'Anna and Xu (2026) propose estimators that do **not** require stationarity
- Key innovation: **rate double robustness**
  - Consistent at $\sqrt{n}$ rate if both nuisance functions converge, even if each converges at a **slower** rate
  - Enables use of nonparametric/ML methods for nuisance estimation
- Additional nuisance parameter: model for compositional changes
  - How does $\mathbb{P}(T = 2 \mid X, D)$ vary with covariates?
  - Captures differential sampling across periods

Sant'Anna & Xu (2026): Estimation Strategy

- **Nonparametric nuisance estimation:** all nuisance functions (m_t^d, p, π) estimated nonparametrically — no need to assume linear/logistic models
- **DML-style procedures:**
  - Cross-fitting to avoid overfitting bias from flexible first-stage estimators
  - Enables use of ML methods (random forests, neural nets, LASSO, etc.) for nuisance estimation while maintaining valid inference
- **Leave-one-out estimation:**
  - Each unit's nuisance functions estimated **without** that unit's own data
  - Eliminates the "own observation" bias that arises with nonparametric estimators
  - Particularly useful with kernel-based or local polynomial methods
- These innovations ensure $\sqrt{n}$-consistent and asymptotically normal estimators even when nuisance functions are estimated at slower-than-$\sqrt{n}$ rates

DR DiD Robust to Compositional Changes

Sant'Anna and Xu (2026) derive a DR estimand that does **not** require stationarity. Same structure, different weights:

$$\begin{aligned}\tau_{dr}^{cc} = \mathbb{E}\Bigg[& w_{t=2}^{d=1,cc}\Big(m_{\Delta}^{d=1}(X) - m_{\Delta}^{d=0}(X)\Big) \\ & + \Big(w_{t=2}^{d=1,cc}\Big(Y - m_{t=2}^{d=1}(X)\Big) - w_{t=1}^{d=1,cc}\Big(Y - m_{t=1}^{d=1}(X)\Big)\Big) \\ & - \Big(w_{t=2}^{d=0,cc}\Big(Y - m_{t=2}^{d=0}(X)\Big) - w_{t=1}^{d=0,cc}\Big(Y - m_{t=1}^{d=0}(X)\Big)\Big)\Bigg]\end{aligned}$$

Key difference: weights use a **generalized PS** $\pi(d,t,x) = \mathbb{P}(D{=}d, T{=}t \mid X{=}x)$:

$$w_{t=2}^{d=1,cc}(D_i, T_i) = \frac{D_i \cdot \mathbf{1}\{T_i{=}2\}}{\mathbb{E}[D \cdot \mathbf{1}\{T{=}2\}]}, \quad w_t^{d,cc}(D_i, T_i, X_i) = \frac{\frac{\pi(1,2,X_i)}{\pi(d,t,X_i)} \cdot \mathbf{1}\{D_i{=}d,\ T_i{=}t\}}{\mathbb{E}\left[\frac{\pi(1,2,X)}{\pi(d,t,X)} \cdot \mathbf{1}\{D{=}d,\ T{=}t\}\right]}$$

Intuition: $\pi(1,2,X_i)/\pi(d,t,X_i)$ reweights each (d,t) cell to match the treated post-treatment covariate distribution — the same IPW logic of "make comparison look like treated," but now applied within each (d,t) cell separately.

Hausman-Type Test for Compositional Changes

- **Key diagnostic:** Compare estimators that use stationarity vs. those that do not
- Under stationarity: both should give the same answer
- Under compositional changes: they will diverge
- Sant'Anna and Xu (2026) formalize this as a Hausman-type test:

$$H_0 : \text{stationarity holds} \quad \text{vs.} \quad H_1 : \text{compositional changes}$$

- Test statistic based on the difference between two DR estimators
- Rejection $\Rightarrow$ use the estimator that allows for compositional changes

Application: South Africa–Mozambique Tariff Liberalization

- Sant'Anna and Xu (2026) revisit Sequeira (2016)'s study of South Africa's tariff liberalization on trade with Mozambique
- Data: repeated cross-sections of trade flows across product categories
- Compositional changes are plausible: product mix changes over time as trade patterns evolve
- Results:
  - Standard estimators (assuming stationarity): find large effects
  - Compositional-change-robust estimators: qualitatively similar but different magnitudes
  - Hausman test: rejects stationarity in some specifications

Key Takeaway: When to Worry about Compositional Changes

- **Panel data:** stationarity holds by construction — not an issue
- **RCS data:** always ask:
  - Is the sampling frame the same across periods?
  - Could migration, attrition, or policy changes affect who is observed?
  - Do covariate distributions shift across periods?
- **Practical advice:**
  1. Compare covariate distributions across periods
  2. Run the Hausman-type test
  3. If stationarity is rejected, use compositional-change-robust estimators

Q: In your own research, how would you diagnose whether compositional changes are present?

What if we have many potential covariates?

From low to high dimensions: machine learning meets DiD.

Act V: Machine Learning & DiD

What You Need to Know About This Section

- We now enter **high-dimensional territory**: many covariates, flexible estimation
- This section **sketches the key ideas**; a full treatment requires substantially more time than we have available
- **The practical takeaway:**
  - Use LASSO/Post-LASSO for covariate selection
  - Combine with cross-fitting to avoid overfitting bias
  - Wrap in the DR framework for robustness
- Focus on the **recipe and the intuition** — we will not derive convergence rates or prove oracle inequalities

If you plan to use these methods in your dissertation, see **?** for the full DML framework and Belloni, Chernozhukov and Hansen (2014) for Post-LASSO inference theory.

The High-Dimensional Challenge

- So far: X is low-dimensional and we specify parametric models (OLS, logit)
- But what if we have **many** potential confounders?
  - Administrative data with hundreds of variables
  - Interactions, polynomials, transformations
  - Researcher degrees of freedom in choosing which to include
- Standard approach: include "all reasonable" covariates $\Rightarrow$ overfitting, instability
- Can we do better?

Machine Learning to the Rescue?

- **Machine learning** offers principled ways to handle high-dimensional X:
  1. Select relevant covariates automatically (LASSO, elastic net)
  2. Estimate flexible functional forms (random forests, boosting)
  3. Avoid overfitting through regularization and cross-validation
- But: naively plugging ML into DR creates an **overfitting bias** problem
- Solution: **cross-fitting** — estimate nuisance functions on one sample, evaluate on another
- This lecture: LASSO $\rightarrow$ cross-fitting $\rightarrow$ DML-DiD

Why Not Just Use All Covariates?

- With k covariates and n observations:
  - OLS/logit requires $k < n$ (infeasible when k is large)
  - Even when $k < n$, overfitting degrades predictions
  - Variance of fitted values grows with k
- The bias-variance tradeoff:
  - **Underfitting** (too few covariates): misspecification bias
  - **Overfitting** (too many covariates): high variance, poor out-of-sample prediction
- We need methods that **regularize**: shrink or select to control complexity
- Key insight: the DR structure provides a natural framework for ML integration

Review: The DR Estimand Structure

Recall the DR DiD estimand (panel data):

$$\text{ATT}^{dr} = \mathbb{E}\left[(w_1(D) - w_0(D, X; p))(\Delta Y - m_{\Delta}^{d=0}(X))\right]$$

- Two **nuisance functions** to estimate:
  1. **Outcome model:** $m_{\Delta}^{d=0}(X) = \mathbb{E}[\Delta Y \mid X, D = 0]$
  2. **Propensity score:** $p(X) = \mathbb{P}(D = 1 \mid X)$

- The DR property means: bias from estimating these nuisance functions is a **product** of their respective errors

- This is exactly the right structure for ML: we can use **flexible** methods for nuisance estimation while maintaining valid inference for the ATT

LASSO: A Primer

- **LASSO** (Tibshirani, 1996) (Least Absolute Shrinkage and Selection Operator):

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i'\beta)^2 + \lambda \sum_{j=1}^{k} |\beta_j|$$

- The ℓ_1 penalty $\lambda\|\beta\|_1$ serves dual purpose:
  1. **Shrinkage:** Pulls coefficients toward zero (reduces variance)
  2. **Selection:** Sets some coefficients exactly to zero (selects variables)

- Useful when the true model is **sparse**: only $s \ll k$ covariates truly matter

- The tuning parameter λ controls the bias-variance tradeoff

Approximate Sparsity

- **Exact sparsity:** Only s coefficients are nonzero (restrictive)
- **Approximate sparsity:** Many small coefficients, but the best s-sparse approximation is close to the truth
- More realistic: covariates may all contribute, but most only marginally
- LASSO works well under approximate sparsity: it automatically finds the most important variables and provides a good approximation
- Key rate requirement: $s^2 \log(k)/n \to 0$ (sparsity grows slowly relative to n)

Post-LASSO: Correcting for Shrinkage Bias

- LASSO shrinks coefficients toward zero $\Rightarrow$ **downward bias** in fitted values
- **Post-LASSO** (Belloni and Chernozhukov, 2013; Belloni, Chernozhukov, Fernández-Val and Hansen, 2017): Two-step procedure
  1. Run LASSO to **select** variables (identify $\hat{S} = \{j : \hat{\beta}_j^{lasso} \neq 0\}$)
  2. Run OLS using only the selected variables $X_{\hat{S}}$ (no penalty)
- Post-LASSO reduces bias while maintaining the sparsity-driven variable selection
- Achieves the same rate of convergence as LASSO but with better constants

LASSO for the Propensity Score

- Model: $p(X_i) = \Lambda(X_i'\gamma)$ with **logistic LASSO**

$$\hat{\gamma}^{lasso} = \arg\min_{\gamma} \; -\frac{1}{n}\sum_{i=1}^{n}\Big[D_i \log \Lambda(X_i'\gamma) + (1-D_i)\log\big(1-\Lambda(X_i'\gamma)\big)\Big] + \lambda\|\gamma\|_1$$

- **Post-LASSO logit:** Select variables, then refit logit on selected set
- Key concern: logistic LASSO provides good **prediction** but not necessarily good **balancing**
- Solution: combine with DR estimation (bias correction handles PS errors)

LASSO for the Outcome Model

- Model: $m_{\Delta}^{d=0}(X_i) = X_i'\beta$ with **linear LASSO**

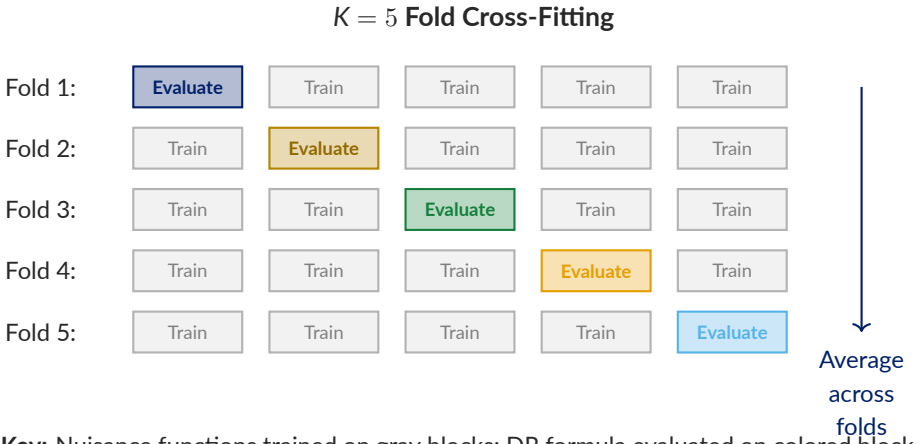
$$\hat{\beta}^{lasso} = \arg\min_{\beta} \frac{1}{n_0} \sum_{i:D_i=0} (\Delta Y_i - X_i'\beta)^2 + \lambda \|\beta\|_1$$

- Estimated using comparison group only (same as RA)
- **Post-LASSO OLS:** Select variables, then refit OLS on selected set
- LASSO automatically discovers which covariates predict ΔY among untreated

The Overfitting Problem: Why Naïve ML Can Fail

- Plugging ML-estimated nuisance functions into the DR formula **can** yield valid inference, but requires **Donsker-type conditions** on the nuisance function classes, plus stronger rate/smoothness/sparsity requirements
- ML estimators (LASSO, random forests, neural nets) typically **violate** these conditions — their complexity grows with n
- **Without these conditions:** using the **same** data to (i) fit nuisance functions and (ii) evaluate the DR formula creates **regularization bias** that may not vanish at $\sqrt{n}$ rate
- Intuition: ML adapts to noise in the training data, and this noise "leaks" into the DR estimator
- **Solution:** Sample splitting / cross-fitting (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins, 2018) — avoids Donsker conditions entirely

Cross-Fitting: A Visual



Key: Nuisance functions trained on gray blocks; DR formula evaluated on colored block.

Cross-Fitting: The DML Framework

Double/Debiased Machine Learning (DML) (Chernozhukov et al., 2018):

1. Split sample into K folds (e.g., $K = 5$)
2. For each fold k:
   - Estimate nuisance functions ($\hat{m}_{\Delta}^{d=0}$, $\hat{p}$) on all folds **except** k
   - Evaluate the DR formula on fold k using these estimates
3. Average across folds to get the final estimate

- **Key property:** Estimation and evaluation use **different** data $\Rightarrow$ no overfitting bias
- Allows $\sqrt{n}$-consistent and asymptotically normal ATT estimates even with slow-converging ML nuisance estimators (requires product of estimation errors $\|\hat{m} - m_0\| \cdot \|\hat{p} - p_0\| = o_p(n^{-1/2})$)
- Works with any ML method (LASSO, random forests, neural networks, ...)

DML Algorithm for DR-DiD

- **Input:** Data $\{Y_{i,t=1}, Y_{i,t=2}, D_i, X_i\}_{i=1}^n$, number of folds K
- **Step 1:** Randomly partition $\{1, \ldots, n\}$ into K folds $I_1, \ldots, I_K$
- **Step 2:** For each fold $k = 1, \ldots, K$:
  1. Estimate $\hat{m}_\Delta^{d=0,(-k)}(x)$ using data **outside** fold k (comparison units only)
  2. Estimate $\hat{p}^{(-k)}(x)$ using data **outside** fold k
  3. Compute DR estimate on fold k:

$$\hat{\theta}_k = \frac{1}{|I_k|} \sum_{i \in I_k} \hat{\psi}_i^{dr}\left(\hat{m}_\Delta^{d=0,(-k)}, \hat{p}^{(-k)}\right)$$

- **Step 3:** Aggregate: $\hat{\theta}^{DML} = \frac{1}{K} \sum_{k=1}^K \hat{\theta}_k$
- **Inference:** IF-based plug-in variance or multiplier bootstrap; see Belloni et al. (2017) and Chernozhukov et al. (2018) for details

We have the tools: LASSO + cross-fitting + DR.

How well does DML-DiD perform in finite samples?

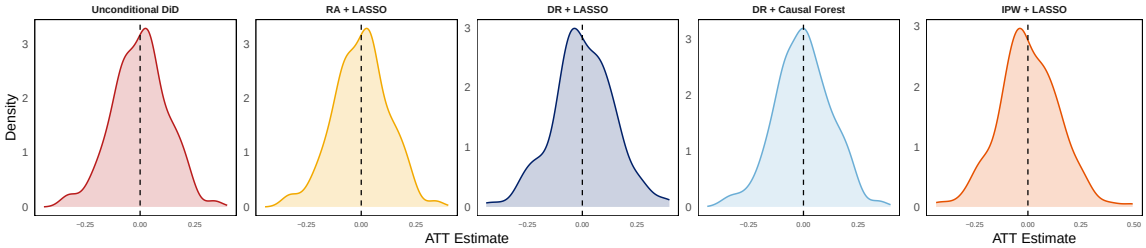
Monte Carlo: DML-DiD

- Three DGPs with $p = 100$ covariates ($s = 5$ active), $n = 500$:
  1. **DGP 1:** Unconditional PT valid (covariates irrelevant)
  2. **DGP 2:** Conditional PT + homogeneous ATT
  3. **DGP 3:** Conditional PT + heterogeneous ATT
- Compare: LASSO-DR (with cross-fitting), Linear DR, TWFE
- 200 MC replications (ML is computationally intensive)

DML-DiD: DGP 1 — Unconditional PT (Covariates Irrelevant)

DGP 1: Unconditional PT
(ATT = 0)

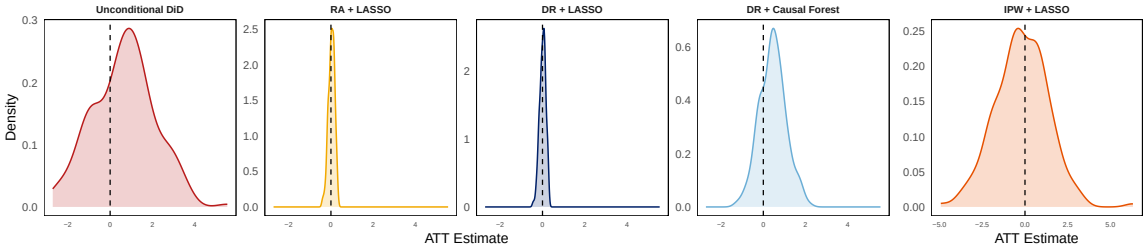
n = 500, p = 100, 200 replications. True ATT = 0.00. IPW x−axis differs.



DML-DiD: DGP 2 — Conditional PT, Homogeneous ATT

DGP 2: Conditional PT
(ATT = 0)

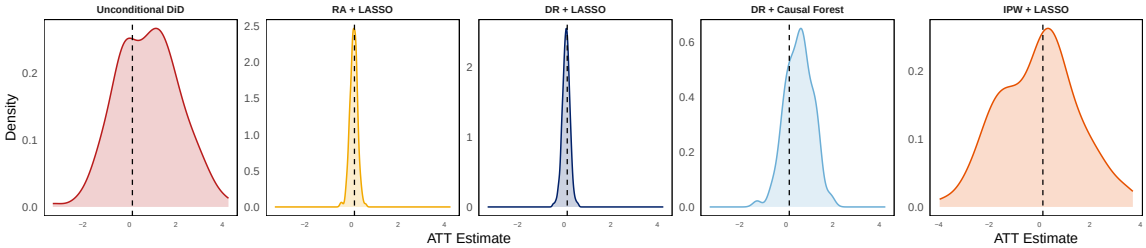
n = 500, p = 100, 200 replications. True ATT = 0.00. IPW x−axis differs.



DML-DiD: DGP 3 — Conditional PT, Heterogeneous ATT

DGP 3: Conditional PT
(ATT != 0)

n = 500, p = 100, 200 replications. True ATT = 0.13. IPW x−axis differs.



DML-DiD: Summary — Bias and RMSE

	DGP 1		DGP 2		DGP 3	
	Uncond. PT		Cond. PT, homog.		Cond. PT, heterog.	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
Unconditional DiD	−0.1	12.3	**64.0**	160.1	**67.0**	148.2
RA + LASSO	−0.1	12.3	2.9	14.9	−3.2	16.3
IPW + LASSO	0.4	14.2	**−19.2**	157.1	**−16.2**	155.4
DR + LASSO	0.4	14.0	0.9	14.7	−5.1	16.8
DR + Causal Forest	−0.2	12.5	**44.1**	76.0	**41.1**	70.9

- DGP 1: all estimators work — unconditional PT holds, covariates irrelevant
- DGPs 2–3: **DR + LASSO** and **RA + LASSO** dominate; unconditional DiD severely biased
- IPW + LASSO unstable: propensity score estimation alone insufficient with many covariates

DML-DiD: Summary — Coverage

	DGP 1 Uncond. PT	DGP 2 Cond. PT, homog.	DGP 3 Cond. PT, heterog.
Unconditional DiD	96.5%	91.0%	94.5%
RA + LASSO	96.5%	92.5%	93.0%
IPW + LASSO	95.0%	96.0%	96.0%
DR + LASSO	95.0%	94.5%	92.0%
DR + Causal Forest	96.0%	87.5%	92.0%

DR + LASSO with cross-fitting provides the best bias-variance tradeoff. Causal forests can estimate heterogeneous effects but are less reliable for average ATT.

DML-DiD estimates the average ATT with high-dimensional nuisance functions.

What if treatment effects vary across units?

Causal Forests: Heterogeneous Treatment Effects

- Beyond estimating the **average** ATT: what about CATT(x)?

- **Generalized Random Forests** (GRF) (Athey, Tibshirani and Wager, 2019):
  - Nonparametrically estimate conditional treatment effects
  - Provide valid pointwise confidence intervals
  - Handle high-dimensional covariates

- **Causal forests for DiD:**
  - Outcome: ΔY_i (first difference)
  - Treatment: D_i
  - Covariates: X_i
  - GRF estimates $\text{CATT}(x) = \mathbb{E}[Y_{i,t=2}(2) - Y_{i,t=2}(\infty) \mid X_i = x, D_i = 1]$

- Combines the DR framework with forest-based heterogeneity estimation

Open Question: Covariates for Identification vs. Interest

Suppose you care about treatment effects conditional on a small subset of covariates — say, gender, race, or income — but need many more covariates to justify conditional parallel trends. How should you proceed?

- This connects to CATT(x) — but the "x" you care about is low-dimensional, while the "X" for identification is high-dimensional
- Tension: averaging over nuisance covariates while conditioning on covariates of interest
- We leave this as an **open question** for you to think about

ML can estimate nuisance functions and uncover heterogeneity.

When does all this machinery actually help?

When Does ML Help?

ML is useful when:	ML is overkill when:
Many covariates ($k > 20$)	Few covariates ($k < 10$)
Unknown functional forms	Clear linear relationships
Complex interactions matter	Simple additive effects
Heterogeneity exploration	Only average effect needed
Administrative/big data	Small samples ($n < 500$)

ML is a **tool**, not a magic bullet. Use it when the complexity of the nuisance functions justifies the additional machinery. For simple settings, parametric DR is simpler.

Honest About Limitations

- ML does **not** fix identification problems:
  - If conditional PT does not hold, no amount of ML helps
  - ML estimates nuisance functions better, not the causal assumptions
- Computational cost: cross-fitting with LASSO/forests is slower than OLS
- Interpretability: harder to understand what's driving the estimates
- Finite-sample performance: ML guarantees are asymptotic; with $n = 200$, parametric methods often work better
- **Bottom line:** ML extends the toolkit, but the hard work is still in the identification assumptions and study design

Three estimation strategies (RA, IPW, DR), RCS extensions, and ML integration.

What should practitioners take away from all this?

Taking Stock

What We Accomplished Today

Key takeaways from DiD with covariates:

1. **Conditional PT** is often more credible than unconditional PT — but requires appropriate estimation methods
2. **TWFE with covariates is fragile:** imposes no covariate-specific trends, hidden linearity bias, potentially negative weights
3. **Doubly robust is the default:** consistent if either outcome model or propensity score is correct; efficient when both are
4. **Panel $>$ RCS:** strictly more efficient, no stationarity concerns. For RCS, test for compositional changes
5. **ML extends DR naturally:** LASSO + cross-fitting for high-dimensional settings; causal forests for heterogeneity

References

Abadie, Alberto, "Semiparametric Difference-in-Differences Estimators," *The Review of Economic Studies*, 2005, *72* (1), 1–19.

Angrist, Joshua D. and Jorn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton: Princeton University Press, 2009.

Athey, Susan, Julie Tibshirani, and Stefan Wager, "Generalized random forests," *The Annals of Statistics*, 2019, *47* (2), 1148 – 1178.

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*.

Belloni, Alexandre and Victor Chernozhukov, "Least Squares after Model Selection in High-Dimensional Sparse Models," *Bernoulli*, 2013, *19* (2), 521–547.

References ii

_ , _ , and Christian Hansen, "Inference on Treatment Effects after Selection among High-Dimensional Controls," *The Review of Economic Studies*, 2014, *81* (2), 608–650.

_ , _ , Iván Fernández-Val, and Christian Hansen, "Program Evaluation and Causal Inference With High-Dimensional Data," *Econometrica*, 2017, *85* (1), 233–298.

Caetano, Carolina and Brantly Callaway, "Difference-in-Differences with Time-Varying Covariates in the Parallel Trends Assumption," 2024. arXiv:2406.15288.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins, "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 2018, *21* (1), C1––C68.

Crump, Richard K, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik, "Dealing with limited overlap in estimation of average treatment effects," *Biometrika*, 2009, *96* (1), 187–199.

Dias, Mateus and Luiz Felipe Fontes, "The Effects of a Large-Scale Mental Health Reform: Evidence from Brazil," *American Economic Journal: Economic Policy*, 2024, *16* (3), 257–289.

References iii

Graham, Bryan, Cristine Pinto, and Daniel Egel, "Inverse Probability Tilting for Moment Condition Models with Missing Data," *The Review of Economic Studies*, 2012, *79* (3), 1053–1079.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd, "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 1998, *66* (5), 1017–1098.

Heckman, James J., Hidehiko Ichimura, and Petra Todd, "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme," *The Review of Economic Studies*, 1997, *64* (4), 605–654.

Khan, Shakeeb and Elie Tamer, "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 2010, *78* (6), 2021–2042.

Rubin, Donald B., "For Objective Causal Inference, Design Trumps Analysis," *The Annals of Applied Statistics*, 2008, *2* (3), 808–840.

Sant'Anna, Pedro H. C. and Jun Zhao, "Doubly robust difference-in-differences estimators," *Journal of Econometrics*, 2020, *219* (1), 101–122.

__ **and Qi Xu**, "Difference-in-Differences with Compositional Changes," *Working Paper*, 2026.

References iv

Sequeira, Sandra, "Corruption, Trade Costs, and Gains from Tariff Liberalization: Evidence from Southern Africa," *American Economic Review*, 2016, *106* (10), 3029–3063.

Tibshirani, Robert, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 1996, *58* (1), 267–288.