Difference-in-Differences: Extension and Application

Huan Deng

April 29, 2025

Outline

Extensions

Extensions

Clustering

Clustered Sampling

- So far we have discussed identification and estimation of the DID design. Now let's turn to inference.
- Recall the large-sample asymptotics of OLS where we assume i.i.d. sampling and let N go to infinity.
- But it might not be desirable to assume independent sampling in some settings.
  - Suppose we want to study the effect of an educational program on students, is it reasonable to assume that students from the same class are mutually independent?
  - Consider a policy implemented at the state level, should we treat individuals from the same state as mutually independent?
- In the above cases, it is more desirable to assume clustered sampling: allowing dependence within clusters while maintaining independence across clusters.

Why Does Clustering Matter in DID?

- **Intra-cluster Correlation:**
  - Observations within a cluster (e.g., state, school, firm) can share unobserved shocks.
  - Naive standard errors assume independence, underestimating the true variability.
- **Conventional Fix: Cluster-Robust SEs**
  - Valid under many clusters (asymptotics where number of clusters $\to \infty$).
  - But can break down when the number of *independent* clusters is small, which is quite common for DID applications.
- **Implication:**
  - Even if each cluster has a large number of units, the *cluster-level shocks* may not average out with only a few clusters, jeopardizing standard inference.

Model-Based Approaches (TWFE Framework)

- Here we consider an example from Roth et al. (2023)
- **Consider a Generic Model:**

$$Y_{i,j,t} = \alpha_j + \phi_t + D_{j,t}\,\beta \;+\; \nu_{j,t} \;+\; \epsilon_{i,j,t}, \tag{1}$$

  - j indexes clusters, t indexes time, i indexes units within cluster.
  - $\nu_{j,t}$: cluster-time shock (common to all units in cluster j).
  - $\epsilon_{i,j,t}$: idiosyncratic error.
- **Cluster-Averaging:**

$$\bar{Y}_{j,t} = \alpha_j + \phi_t + D_{j,t}\,\beta + \eta_{j,t}, \quad \text{where } \eta_{j,t} = \nu_{j,t} + \frac{1}{n_j}\sum_{i=1}^{n_j} \epsilon_{i,j,t}.$$

- **Small-J Problem:**
  - With few clusters J, normal approximations for $\Delta\nu_j$ (the cluster-time shocks) are dubious.

Examples of Model-based Methods

- Model-based methods impose assumptions on these structural error terms to make inferences.
- **Donald & Lang (2007)**:
  - Assume $\nu_{j,t}$ are i.i.d. normal, homoskedastic across clusters.
  - Then treat cluster means as random draws from this distribution and use a t-distribution with $(J-2)$ degrees of freedom.
  - *Limitation:* Homoskedasticity can be violated if treatment-effect heterogeneity correlates with cluster membership.
- **Conley & Taber (2011)**:
  - Suppose few treated clusters, many untreated.
  - Use the untreated-cluster distribution of shocks to infer the treated-cluster shock.
  - *Limitation:* Assumes i.i.d. cluster-level shocks across clusters, no cluster-level heterogeneity in potential outcomes.

Examples of Model-based Methods

- **Ferman & Pinto (2019)**:
  - A bootstrap-based approach to allow certain types of heteroskedasticity (e.g., different cluster sizes).
  - Still relies on assumptions about how cluster shocks vary across j.
- **Hagemann (2020)**:
  - Proposes a permutation-based test for a single (large) treated cluster vs. a few untreated clusters.
  - Must assume that each untreated cluster has the same potential outcome evolution, so any single cluster can stand in as a counterfactual for the treated.

Conditional on Shocks

- Treat $\nu_{j,t}$ as fixed rather than random draws.
- Then the variation is mostly at the individual level, effectively ignoring the sampling at the cluster level.
- This violates the Parallel Trend assumption since $\nu_{j,t}$ is a time-varying unobservable.
- We can only hope that the violation may be relatively small if the cluster-specific shocks are small relative to the idiosyncratic variation.

The Fisherian Approach

- Permute (or "re-randomize") the treatment assignment across clusters to build a reference distribution of the test statistic (the DID estimate).
- Exactly valid if treatment is truly random at the cluster level and we test a *sharp* null (no effect for all units).
- *Limitation:* Real-world DID often doesn't have random assignment; also a difference between "sharp" vs. "weak" null.
- Rambachan & Roth, (2022) extend the above idea to settings where there is staggered adoption and (quasi-)random timing of treatment.

At What Level to Cluster?

- The level of clustering would affect the effective sample size, which is the number of clusters.
- So, if we cluster at a very aggregate level, we are left with a small effective sample size. If outcomes within clusters are positively correlated, then the clustered standard error might be too conservative.
- Designed-based approach suggests clustering at the level of treatment assignment.
- Abadie et al. (2023) offer a latest discussion on clustering.

Extensions

Conditional Parallel Trend

Relaxing the Parallel Trend Assumption

- Remember how we relax the random assignment to unconfoundedness.
- Actually, we can also relax the unconditional Parallel Trend assumption (PTA) to its conditional version.
- PTA might be too strong if units with different pre-treatment covariates X have different trends.
- Conditional PTA states that, in the absence of treatment, conditional on X, the evolution of the outcome among the treated units is, on average, the same as the evolution of the outcome among the untreated units.
- Below we discuss identification and estimation with the Conditional PTA using the cononical DID.

Conditional Parallel Trends & Strong Overlap

Conditional PT Assumption:

$$E\left[Y_{t=2}(\infty) - Y_{t=1}(\infty) \,\middle|\, G = 2, X\right] = E\left[Y_{t=2}(\infty) - Y_{t=1}(\infty) \,\middle|\, G = \infty, X\right]. \tag{2}$$

- Allows the *evolution* of outcomes to differ by X.
- More flexible than requiring *everyone*, regardless of covariates, to follow parallel trends.

Strong Overlap Assumption:

$$0 < P(G = 2 \mid X) < 1 \quad \text{almost surely.} \tag{3}$$

- Ensures every X-type has a positive probability of being in treatment or control.
- Helps to avoid "extrapolation" from segments of X-space where the control group is absent.

Identification of the ATT Under Conditional PT

Under **No Anticipation** and **Conditional PT**, we can show:

$$\begin{aligned}\text{ATT}(X) = &\big[E[Y_{t=2} \mid G = 2, X] - E[Y_{t=1} \mid G = 2, X]\big] - \\ &\big[E[Y_{t=2} \mid G = \infty, X] - E[Y_{t=1} \mid G = \infty, X]\big]. \qquad (4)\end{aligned}$$

The **unconditional ATT** is then the expectation of $\text{ATT}(X)$ over the distribution of X among the treated group:

$$\text{ATT} \;=\; E\big[\text{ATT}(X) \mid G = 2\big].$$

- It is tempting to think that we can also use TWFE to estimate ATT by adding X as controls.
- However, merely *adding* X linearly to a TWFE regression is **not** guaranteed to work.

Problems with TWFE + Covariates

- **Standard TWFE Model with Covariates**:

$$Y_{it} = \gamma_t + \delta_i + \beta D_{it} + X_i'\theta + \epsilon_{it}.$$

- Key Restriction from a linear-additive form: the evolution of outcomes among both the treated units and control units don't vary with X
- This also implies that $ATT(X)$ won't change with X, which means the model is misspecified if we have treatment heterogeneity by X. Bias can be huge.
- We need to use other estimators.

1) Regression Adjustment

- **Idea**:

$$\widehat{m}_{G=g,t=s}(X) \;\approx\; E\big[Y \mid G=g, T=s, X\big]$$

  via parametric, semi-parametric, or nonparametric methods.

- Then,

$$\widehat{\mathrm{ATT}} \;=\; E\Big[\big(\widehat{m}_{2,2}(X)-\widehat{m}_{2,1}(X)\big)-\big(\widehat{m}_{\infty,2}(X)-\widehat{m}_{\infty,1}(X)\big) \;\Big|\; G=2\Big].$$

- **Pros**:
  - Flexible modeling of outcome process given X.
  - Straightforward to implement if dimension of X is not too large.
- **Cons**:
  - Must model outcome evolution accurately.
  - Potential bias if outcome model is mis-specified.

2) Propensity Score (PS) Weighting

- **Inverse Probability Weighting** for panel data (Abadie, 2005; Sant'Anna & Zhao, 2020):

$$ATT^{ipw,Abadie} = \frac{\mathbb{E}\left[\left(D - \frac{(1-D)p(X)}{1-p(X)}\right)(Y_{t=2} - Y_{t=1})\right]}{\mathbb{E}[D]} \quad (5)$$

$$ATT^{ipw,SZ} = \mathbb{E}\left[\left(\frac{D}{\mathbb{E}[D]} - \frac{\frac{p(X)(1-D)}{1-p(X)}}{\mathbb{E}\left[\frac{p(X)(1-D)}{1-p(X)}\right]}\right)(Y_{t=2} - Y_{t=1})\right] \quad (6)$$

Where:

- D is the treatment indicator ($D = 1$ if treated, $D = 0$ if control).
- Y_t is the outcome at time t (before and after treatment).
- $p(X) = P(D = 1 \mid X)$ is the propensity score.
- $\mathbb{E}[D]$ normalizes to ensure ATT interpretation.
- Need to build a good model for the propensity score

3) Doubly Robust (DR) Estimators

- **Combine** outcome regression + propensity score weighting.
- **Key Feature**:

$$\widehat{\tau}_{\text{DR}} = \underbrace{\text{IPW piece}}_{\text{PS model}} + \underbrace{\text{Regression ``correction'' piece}}_{\text{Outcome model}}$$

$\Longrightarrow$ consistent if *either* the PS or the outcome model is correct.

- Sant'Anna and Zhao (2020) prpose the below DR estimator:

$$ATT^{dr} = \mathbb{E}\left[\left(\frac{D}{\mathbb{E}[D]} - \frac{\frac{p(X)(1-D)}{1-p(X)}}{\mathbb{E}\left[\frac{p(X)(1-D)}{1-p(X)}\right]}\right)\left((Y_{t=2} - Y_{t=1}) - \left(m_{t=2}^{G=\infty}(X) - m_{t=1}^{G=\infty}(X)\right)\right)\right] \tag{7}$$

- "Insurance" against mis-specification of *one* of the models.
- Some DR estimators are *semiparametrically efficient*.

Synthetic Control

Synthetic Control

Basics

Introduction

- The synthetic control method (SCM) was originally proposed in Abadie and Gardeazabal(2003) and Abadie et al. (2010) to estimate the effects of aggregate interventions.
- Many events/interventions happen at aggregate level (cities, regions, countries).
- SCM doesn't rely on Parallel Trend, instead, SCM uses a weighted average of control units to best match the characteristics of the treated unit.
- Here I follow the NBER Method lecture given by Abadie on SCM.
- The lecture takes around 1 hour and is definitely worth the time!

Setup

- We observe $J+1$ units over T time periods.
- Unit 1 is treated in periods $T_0+1,\ldots,T$.
- The remaining J units form the donor pool (untreated).
- Let Y_{it}^I be the outcome for unit i at time t if exposed to treatment in periods $T_0+1,\ldots,T$.
- Let Y_{it}^N be the outcome for unit i at time t if never treated.
- Goal: Estimate

$$\tau_{1t} \;=\; Y_{1t}^I - Y_{1t}^N \;=\; Y_{1t} - Y_{1t}^N, \quad t > T_0.$$

Synthetic Control Estimator

- Let $W = (w_2, \ldots, w_{J+1})'$ with $w_j \geq 0$ and $\sum_{j=2}^{J+1} w_j = 1$.
- Let X_1 be the (k-dimensional) vector of pre-treatment characteristics for the treated unit.
- Let X_0 be the $(k \times J)$ matrix of the same characteristics for donor units.
- Note that X may include pre-intervention values of Y_{it}
- W^* is chosen to minimize $\|X_1 - X_0 W\|$ subject to the weight constraints.
- For $t \geq T_0 + 1$, the synthetic control estimator of τ_{1t} is

$$\hat{\tau}_{1t} \;=\; Y_{1t} \;-\; \sum_{j=2}^{J+1} w_j^* \, Y_{jt}.$$

Predictor Weighting

A common choice of norm:

$$\|X_1 - X_0 W\| = \left(\sum_{h=1}^{k} v_h \left(X_{h1} - \sum_{j=2}^{J+1} w_j X_{hj} \right)^2 \right)^{1/2},$$

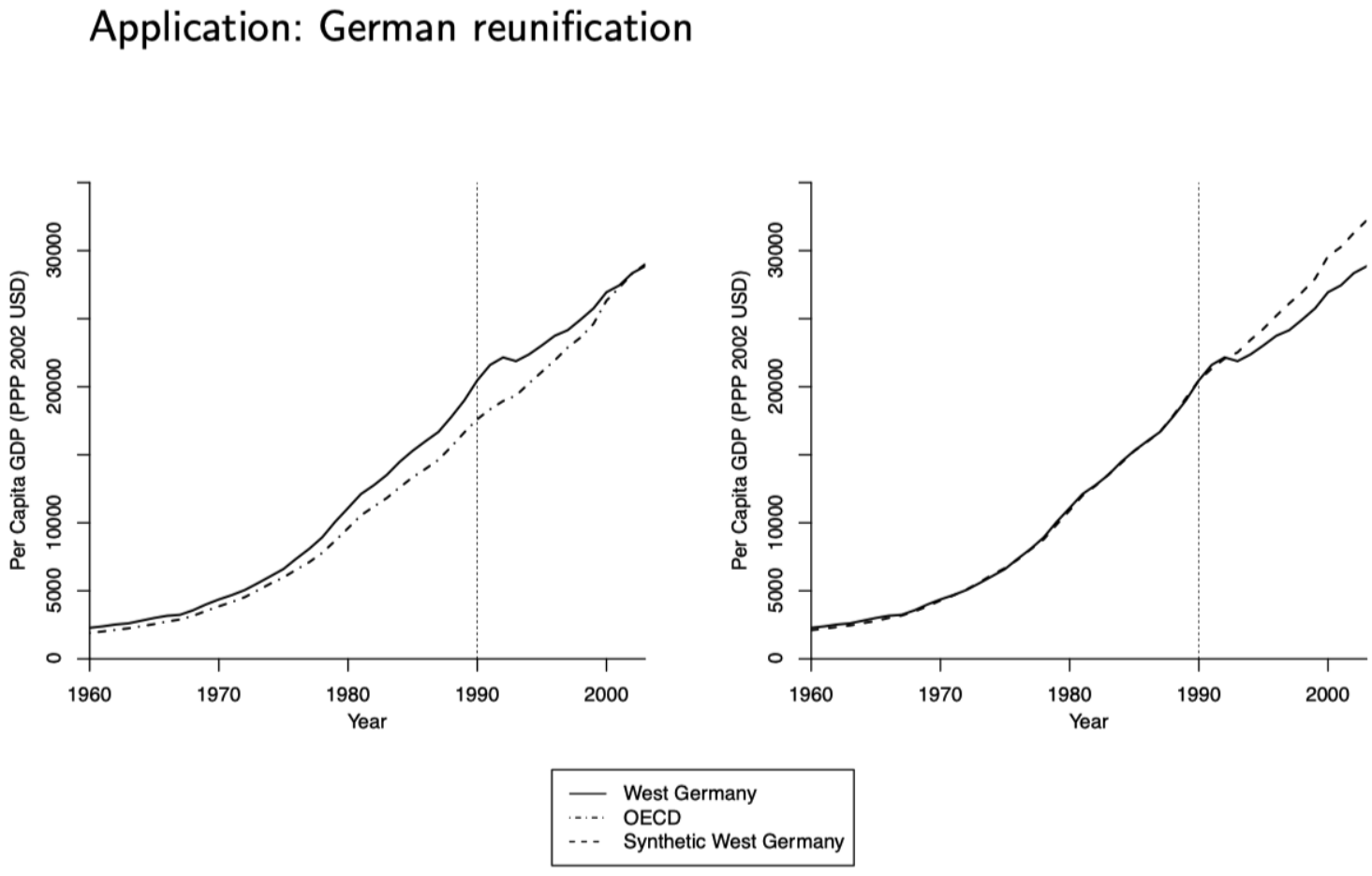
where $v_1, \ldots, v_k > 0$ reflect the relative predictive power of each predictor for the outcome.

Data-driven or user-specified v_h can be used.

Application: German Reunification

- Illustrative example from Abadie, Diamond, and Hainmueller (2010).
- West Germany is "treated" by the impact of reunification.
- Donor pool: Other OECD countries.
- Synthetic West Germany is constructed as a weighted combination of these countries to match pre-treatment characteristics.

Application: German Reunification



Application: German Reunification

Application: German reunification

	West Germany (1)	Synthetic West Germany (2)	OECD Sample (3)
GDP per-capita	15808.9	15802.24	13669.4
Trade openness	56.8	56.9	59.8
Inflation rate	2.6	3.5	7.6
Industry share	34.5	34.5	34.0
Schooling	55.5	55.2	38.7
Investment rate	27.0	27.0	25.9

Note: First column reports $\boldsymbol{X}_1$, second column reports $\boldsymbol{X}_0\boldsymbol{W}^*$, and last column reports a simple average for the 16 OECD countries in the donor pool. GDP per capita, inflation rate, and trade openness are averages for 1981–1990. Industry share (of value added) is the average for 1981–1989. Schooling is the average for 1980 and 1985. Investment rate is averaged over 1980–1984.

Application: German Reunification

Application: German reunification

country j	W_j^*	country j	W_j^*
Australia	0	Netherlands	0.10
Austria	0.42	New Zealand	0
Belgium	0	Norway	0
Denmark	0	Portugal	0
France	0	Spain	0
Greece	0	Switzerland	0.11
Italy	0	United Kingdom	0
Japan	0.16	United States	0.22

Why SCM Works?

- Under the factor model:

$$Y_{it}^N = \theta_t Z_i + \eta_t + \lambda_t \mu_i + \varepsilon_{it},$$

where Z_i are observed features, μ_i unobserved features, and ε_{it} random noise.

- Suppose that we can choose W^* such that:

$$Z_1 = \sum_{j=2}^{J+1} w_j^* Z_j, \quad Y_{1t} = \sum_{j=2}^{J+1} w_j^* Y_{jt}, \quad t \in \{1, \ldots, T_0\}$$

- When T_0 is large, an approximately unbiased estimator of τ_{1t} is:

$$\hat{\tau}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}, \quad t \in \{T_0 + 1, \ldots, T\}$$

Inference via Permutation (Placebo) Tests

- Reassign treatment ("placebo interventions") to each unit in the donor pool.
- Compute synthetic control estimates for these placebo cases.
- Compare the estimated "treatment" effect of the actual treated unit to the distribution of placebo effects.
- This approach does not rely on classical sampling-based arguments but rather on the permutation of units within the sample.

Why Use Synthetic Controls?

- **No extrapolation**: Weighted average must lie in the convex hull of the donor units.
- **Transparency of fit**: We see exactly how well the synthetic matches the treated pre-intervention.
- **Safeguard against specification searches**: Weights can be chosen without seeing post-treatment outcomes.
- **Interpretability**: Sparse weights reveal which donor units matter most.

Synthetic Controls for Experimental Design

- When a firm wants to pilot a policy (treatment) in one or few aggregate units:
  - Must pick which unit(s) to treat and which to leave as controls.
- (Abadie & Zhao, 2021): Propose constructing a "synthetic treated unit" that matches the *overall population* of interest, then constructing a synthetic control for that synthetic treated unit.
- Also used extensively by business analytics teams (e.g., for marketing or pricing pilots).

Synthetic Control

Synthetic DID

Overview

- Many applied settings use panel data to estimate causal effects of a policy or intervention.
- **Difference-in-Differences (DID)**: Commonly used when multiple units are treated and parallel trends are plausible.
- **Synthetic Control (SC)**: Commonly used for a single (or few) large treated units; reweights donor units to match pre-treatment trajectory.
- **Synthetic DID (SDID)**: Proposed by Arkhangelsky et al. (2021). Combines ideas from DID and SC:
  - Matches pre-treatment trends via data-driven weights (like SC).
  - Includes unit & time fixed effects (like DID).
  - Goal: Achieve robustness to bias while retaining reliable inference in large panels.

Model Setup

Data: Balanced panel with N units, T time periods.

- Y_{it}: outcome of interest for unit i in period t.
- $W_{it} \in \{0, 1\}$: treatment indicator (1 if unit i treated in period t, 0 otherwise).
- Often assume a *block* treatment: after some $t = T_{\text{pre}}$, certain units become treated for all subsequent t. (can be generalized to staggered adoption)

Potential Outcomes:

$$Y_{it}^N \quad \text{(not treated)}, \quad Y_{it}^I \quad \text{(treated)}.$$

Observed outcome:

$$Y_{it} = (1 - W_{it})\, Y_{it}^N \;+\; W_{it}\, Y_{it}^I.$$

Estimand: average treatment effect for the treated (ATT),

$$\tau = \frac{1}{N_{\text{tr}}\, T_{\text{post}}} \sum_{i \in \text{treated}} \sum_{t = T_{\text{pre}}+1}^{T} \left(Y_{it}^I - Y_{it}^N\right).$$

Difference-in-Differences (DID)

Classical setup: parallel trends assumption.

$$Y_{it}^{N} = \alpha_i + \beta_t + \varepsilon_{it}.$$

Two-way fixed effects regression:

$$(\hat{\tau}_{\text{did}}, \hat{\alpha}, \hat{\beta}) = \arg\min_{\tau,\alpha,\beta} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(Y_{it} - \alpha_i - \beta_t - \tau\, W_{it}\right)^2.$$

In the simplest block-assignment case:

$$\hat{\tau}_{\text{did}} = \left(\overline{Y}_{\text{tr, post}} - \overline{Y}_{\text{tr, pre}}\right) - \left(\overline{Y}_{\text{co, post}} - \overline{Y}_{\text{co, pre}}\right).$$

Limitations:

- Requires additive fixed effects.
- Sensitive if pre-treatment trends are *not* parallel.

Synthetic Control (SC)

Reweight the untreated units to match the treated unit's outcome path in pre-treatment periods.

$$(\hat{\tau}_{\text{sc}}, \hat{\mu}, \hat{\beta}) = \arg\min_{\tau,\mu,\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \mu - \beta_t - \tau W_{it}\right)^2 \hat{\omega}_i^{\text{sc}},$$

where weights $\hat{\omega}_i^{\text{sc}} \geq 0$ sum to 1, chosen to make $\sum_{i=1}^{N_{\text{co}}} \hat{\omega}_i^{\text{sc}} Y_{i,\text{pre}} \approx Y_{\text{tr, pre}}$.

- **Pro:** Helps "force" parallel trends by focusing on units whose observed *trajectories* match the treated unit(s).
- **Con:** Not invariant to additive unit-level shifts (α_i).
- Often used in "comparative case studies" with small N_{tr} (like 1 or 2).

Motivation for Synthetic DID (SDID)

- Both DID and SC have limitations.
- SC can handle non-parallel trends by matching carefully, but can be thrown off by additive unit-level shifts and can have large variance with small donor pools.
- DID includes unit fixed effects (α_i) but can be biased if trends are not parallel.
- **SDID:** Combine reweighting from SC (to improve parallel-trends plausibility) *and* invariance to additive unit/time fixed effects from DID.

SDID Estimator Formulation

High-level algorithm (Arkhangelsky et al. (2021)):

1. *Choose unit weights* $\hat{\omega}^{\text{sdid}}$ to align pre-treatment *outcomes* between treated and control.
2. *Choose time weights* $\hat{\lambda}^{\text{sdid}}$ to align pre- vs. post-treatment periods among control units.
3. *Run a weighted two-way fixed-effects regression*:

$$(\hat{\tau}_{\text{sdid}}, \hat{\mu}, \hat{\alpha}, \hat{\beta}) = \arg \min_{\tau,\mu,\alpha,\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \mu - \alpha_i - \beta_t - \tau W_{it}\right)^2 \hat{\omega}_i^{\text{sdid}} \hat{\lambda}_t^{\text{sdid}}. \quad (8)$$

Why both weights?

- *Unit weights:* ensure that control units with "similar" pre-treatment trends to the treated units are emphasized.
- *Time weights:* ensure that pre-treatment time periods more like post-treatment periods get higher weight.

Comparison of DID, SC, and SDID

Feature	DID	SC	SDID
Uses unit FE (α_i)?	Yes	No (not invariant)	Yes (invariance)
Uses time FE (β_t)?	Yes	Yes	Yes
Balances pre-trends via weighting?	No	Yes (unit weighting)	Yes (unit + time weighting)
Robust to non-parallel trends?	Weak	Better	Best
Large-sample inference standard?	Yes	Less standard	Yes (like DID)

Key points:

- DiD is simplest but fragile if parallel trends fail.
- SC reweights control units, but lacks invariance to additive unit shifts.
- SDID *double-weights* and is *invariant to row/column shifts*.
- Empirically performs well in both "many-treated" (DiD-like) and "few-treated" (SC-like) scenarios.

Inference for SDID

Challenges:

- Weighted regression means standard Eicker-Huber-White formula needs adjustments.
- Within-unit serial correlation of errors is a concern.

Approaches:

- *Cluster-robust variance estimators* at the unit level, using the final weighted residuals.
- *Placebo or randomization-based inference* (e.g., apply the same method to artificially "assigned" treatments).
- Authors show consistency of a block-robust variance estimator where one clusters at the unit level.

Applications

Applications

Minimum Wage (Cengiz et al. (2019))

Motivation

- The minimum wage has been studied for decades since Card and Krueger (1994).
- Much of prior empirical literature focused on:
  - Teen employment.
  - Certain industries like restaurants or retail.
  - Estimating net effects on average or aggregate employment with little focus on the *distribution of wages*.
- **Key Question:** Does raising the minimum wage overall *reduce the number of low-wage jobs*?
- **Paper's Contribution:**
  - Proposes a new approach that looks at **changes in the entire wage distribution**.
  - Uses **difference-in-differences** on *138* prominent state-level minimum wage changes (1979–2016).
  - Introduces the idea of "missing" and "excess" jobs to isolate impacts at the bottom of the wage scale.

Data

- **Primary Data:** Current Population Survey (CPS) Merged Outgoing Rotation Groups (1979–2016).
  - Quarterly, state-level wage distribution; 4.7 million individual records.
  - Focus on *hourly wages* (reported or weekly earnings / weekly hours).
- **Minimum Wage Policy Series:**
  - Uses daily state-level minimum wage values compiled by Vaghul & Zipperer (2016).
  - Identifies 138 "prominent" state-level min wage events:
    - At least \$0.25 real increase,
    - Affects at least 2% of workers.

Empirical Strategy

- **Difference-in-Differences (DID) Event Study:**
  - Compare states that raise their minimum wage (*treated*) with those that do not (*controls*), focusing on an 8-year window: 3 years before, 5 after.
  - Regress the *frequency* of jobs in each \$0.25 wage bin on event-time indicators.
  - Control for:
    - State-by-wage-bin fixed effects,
    - Wage-bin-by-period fixed effects,
    - (Sometimes) state-specific linear or quadratic trends, etc.

Bunching-Based Logic

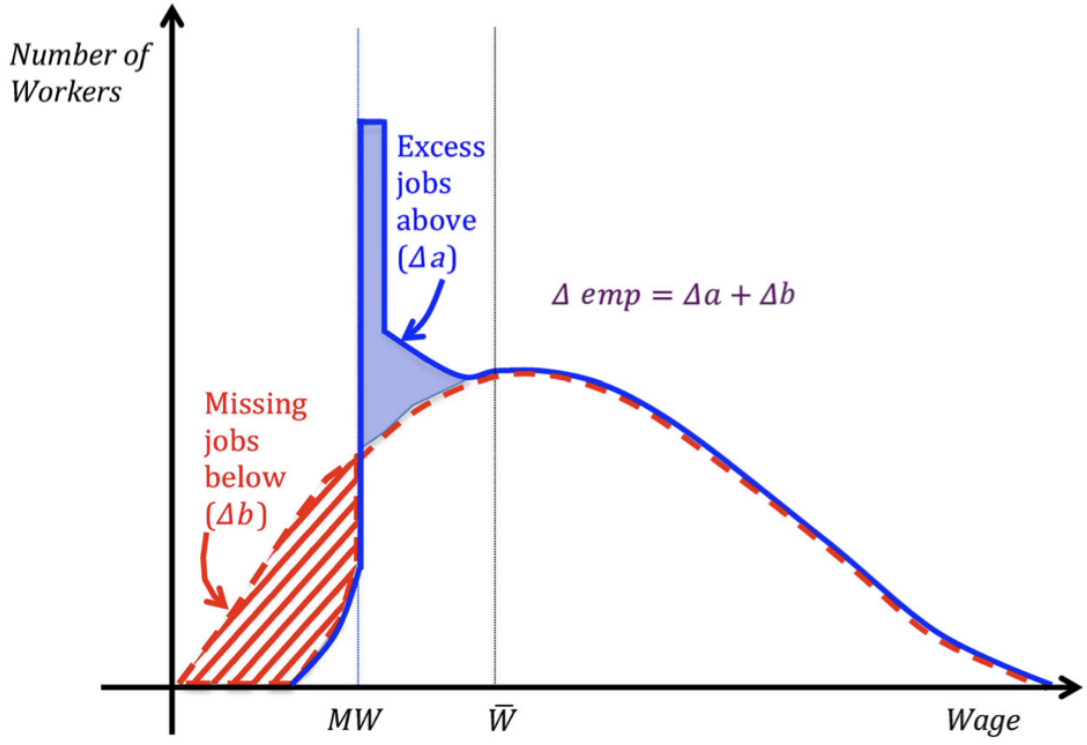


FIGURE I

The Impact of Minimum Wages on the Frequency Distribution of Wages

The figure shows the effect of the minimum wage on the frequency distribution of hourly wages. The red dashed line (color version available online) shows the wage distribution before the introduction of the minimum wage, and the blue solid line shows the distribution afterwards. Because compliance is less than perfect, some workers are paid below the minimum wage, and the post-treatment distribution starts below the minimum wage. For other workers, the introduction of the minimum wage produces "missing jobs" (Δb), as shown by the striped red shaded area (under the red dashed line) between the origin and MW. These missing jobs may either reflect workers getting a raise, or their jobs being destroyed. The former group creates the "excess jobs above" (Δa), as shown by the solid blue shaded area (under the blue solid line) between MW and $\overline{W}$, the upper limit for any effect of the minimum wage on the earnings distribution. The overall change in employment due to the minimum wage (Δe) is the sum of the two areas ($\Delta a + \Delta b$).

Key Results

- On average, *no net job loss*:
  - Missing jobs below MW $\approx$ Excess jobs just above it.
  - Within 5 years after the change, total number of low-wage jobs remains roughly unchanged.
- **Significant wage gains for affected workers** (6–7% on average).
- **Spillovers** up to \$2 or \$3 above the min wage.
- Very little evidence that min wage changes in the US lead to large negative total employment effects, *even for bigger hikes* in the sample.

Why This Methodology is Unique

- **It focuses on the wage distribution** rather than *just* on a narrow group (teens) or aggregated outcome (total employment).
- **Allows direct measurement of "missing" vs. "excess" jobs**:
  - Clarifies whether missing jobs are truly destroyed or just moved above the new minimum wage.
- **Uses event-based approach** (3 years before, 5 after each state policy change) to mitigate confounding from broader time-series trends.
- Use the upper-tail employment changes as a falsification test.

Applications

Expansion of Medicaid (Goodman-Bacon (2018))

Research Question & Motivation

- **Key Question:** Did the introduction of Medicaid (1965–70) improve health outcomes for low-income and minority children?
- **Policy Context:**
  - Medicaid intended to provide health coverage to poor families, especially those on welfare (AFDC).
  - Large expansion of coverage for children via "categorical eligibility".
  - Pre-1965: Many low-income children lacked health insurance and had poor health outcomes.
- **Contribution:** Examines mortality and infant/child health, using state-level variation in welfare (AFDC) rates to identify Medicaid's effects.

Background: Medicaid Implementation

- Enacted in 1965 as part of SSA Amendments (alongside Medicare).
- States were required to implement Medicaid by 1970 (many by 1967–68).
- Federal requirement: All welfare (AFDC) recipients must be covered ("categorical eligibility").
- This generated cross-state variation in eligibility, driven by differences in AFDC take-up and policies.
- Nonwhite children often had higher AFDC rates $\rightarrow$ More Medicaid exposure.

Data Sources

- **AFDC/Eligibility:** State-level counts of AFDC recipients by race, matched with population denominators.
- **Mortality:** Vital Statistics (Multiple-Cause-of-Death) from 1950–79.
  - Infant mortality rates (IMR) per 1,000 live births.
  - Child mortality (ages 1–14) per 100,000 population.
- **Public Insurance:** Department of Health, Education, and Welfare (DHEW) reports (1963–76).
  - Measures children (0–19) using Medicaid or other means-tested insurance.
- **Socioeconomic Controls:** Census, per capita income, hospital capacity, etc.

Empirical Strategy

- **Key Variation:** State differences in AFDC-based Medicaid eligibility at time of implementation.
- **Identification:** Difference-in-differences using (i) pre vs. post Medicaid and (ii) high vs. low AFDC states.
- **Event Study Specification:**

$$\ln(\text{Mortality}_{st}^{k}) = \alpha_s + \delta_t + \beta \cdot \text{AFDC}_s^* \times \mathbf{1}\{\text{post-}t\} + \text{controls} + \varepsilon_{st},$$

  - AFDC_s^* = AFDC rate in state s the year Medicaid started.
  - "Post" = after Medicaid introduction in state s.
  - Includes region-by-year fixed effects and state fixed effects.
- **Interpretation:** Captures how mortality changes more in states with higher AFDC coverage.

Potential Confounders & Common Trends

- **Parallel Trends Assumption:**
  - States with higher AFDC rates must not differ systematically in pre-Medicaid mortality trends, other than through eligibility.
  - Author tests for pre-trends using an event-study approach: No difference prior to Medicaid.
- **Checks:**
  - AFDC rates reflect long-standing state institutions, uncorrelated with new policy changes.
  - In extensive balancing tests, $AFDC_s^*$ is not correlated with prior mortality levels/trends, hospital capacity, or private coverage.
- **Inference:** Standard errors clustered at the state level.

Mechanisms and Demographic Effects

- **Mechanism:** Gains in public insurance coverage $\Longrightarrow$ more hospital care, earlier treatment of infection, pneumonia, etc.
- **By Causes:** Largest declines in treatable conditions (pneumonia, anemia, etc.).
- **Aggregate Impact:**
  - Nonwhite child mortality fell by 11% overall.
  - Narrowed black-white mortality gap in infancy and childhood.
  - Suggests large health returns to expansions of public insurance for children.

Conclusions

- **Medicaid's Introduction** led to significant reductions in mortality for children, especially nonwhite children.
- **Policy Implication:** Expansions of public insurance can yield sizable mortality benefits and reduce racial/SES disparities in child health.
- **Broader Relevance:** Debates on modern Medicaid expansions highlight potentially large health improvements from coverage for low-income populations.

Applications

Social Media and Mental Health (Braghieri et al. (2022))

Motivation and Background

- **Rapid growth** of social media usage in the past decade.
  - "In 2021, 4.3 billion people—more than half of the world population—had a social media account, and the average user spent around two and a half hours per day on social media platforms (GWI 2021; We Are Social 2021)."
- Increasing concerns about social media's impact on mental health.
- Causal evidence is rare.

Research Question

Key Question

Does increased social media use *cause* a measurable change in mental health outcomes?

- Quasi-experimental design by leveraging a unique natural experiment: the staggered introduction of Facebook across US colleges in the mid-2000s.
- Survey data on college students' mental health collected in the years around Facebook's expansion

Difference-in-Differences (DID)

- As a baseline specification, we estimate the following two-way fixed-effect (TWFE) model:

$$Y_{icgt} = \alpha_g + \delta_t + \beta \times Facebook_{gt} + \mathbf{X}_i \times \gamma + \mathbf{X}_c \times \psi + \epsilon_{icgt}, \quad (9)$$

- where Y_{icgt} represents an outcome for individual i who participated in survey wave t and attends college c that belongs to expansion group g
- α_g (or sometimes α_c) indicates expansion-group (or college) fixed effects
- δ_t indicates survey-wave fixed effects; $Facebook_{gt}$ is an indicator for whether, in survey wave t, Facebook was available at colleges in expansion group g
- $\mathbf{X}_i$ and $\mathbf{X}_c$ are vectors of individual-level and college-level controls, respectively.

Results

TABLE 1—BASELINE RESULTS: INDEX OF POOR MENTAL HEALTH

	Index of poor mental health			
	(1)	(2)	(3)	(4)
Post-Facebook introduction	0.137	0.124	0.085	0.077
	(0.040)	(0.022)	(0.033)	(0.032)
Observations	374,805	359,827	359,827	359,827
Survey-wave fixed effects	✓	✓	✓	✓
Facebook-expansion-group fixed effects	✓	✓		
Controls		✓	✓	✓
College fixed effects			✓	✓
FB-expansion-group linear time trends				✓

Results

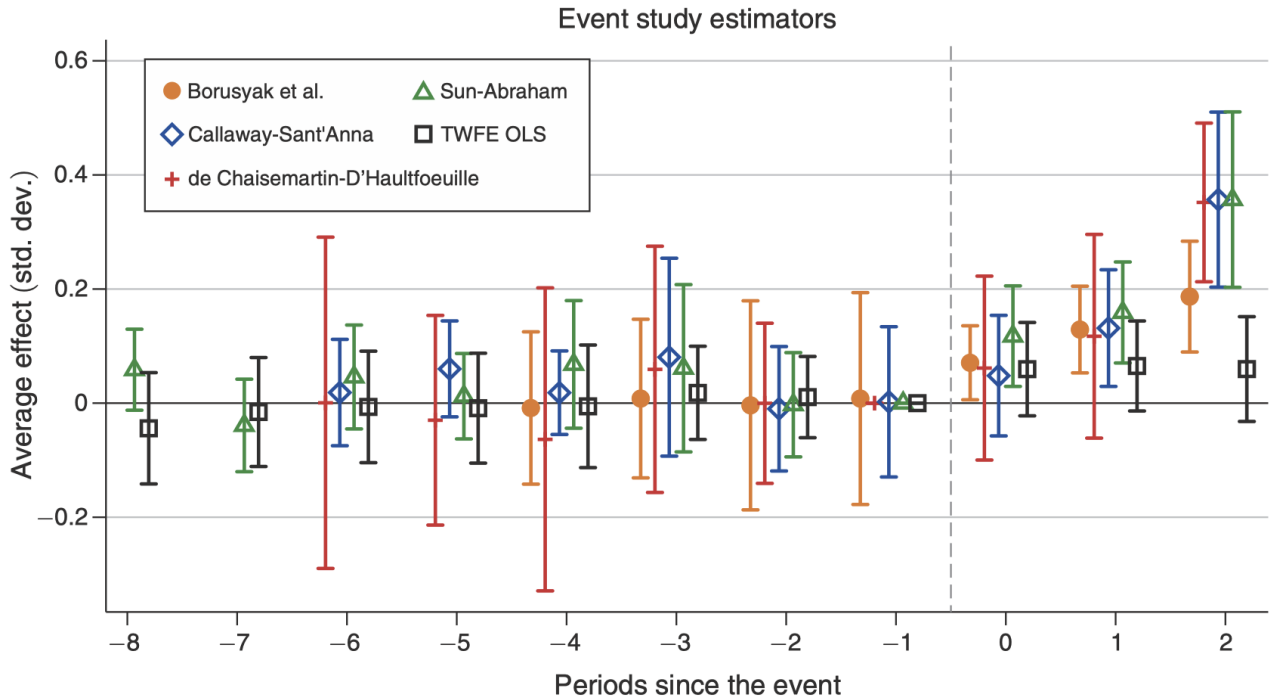


FIGURE 2. EFFECTS OF FACEBOOK ON THE INDEX OF POOR MENTAL HEALTH BASED ON DISTANCE TO/FROM FACEBOOK INTRODUCTION

Further Reading

DID Materials by Sant'Anna

Sant'Anna provides a DID checklist for practitioners on his website `https://psantanna.com/did-resources/#/`

Difference-in-Differences Checklist

1. Start plotting the treatment rollout (e.g., use panelView R package)
2. Document how many units are treated in each cohort.
3. Plot the evolution of average outcomes across cohorts.
4. Choose the comparison groups and the PT assumption carefully:
   Who decides treatment? What do they know? What type of selection is allowed?
5. Do event-study analysis without any covariates and assess if PT is plausible.
6. If unconditional PT is not plausible, incorporate covariates into the analysis.
7. When using covariates, check for overlap: If control groups are small, problems with overlap will probably arise. If you are OK with extrapolation, use regression adjustment DiD procedures.
8. Do event-study analysis after adjusting for covariates and assess if conditional PT is plausible.
9. Conduct some sensitivity analysis for violations of PT (e.g., use honestDiD R package).
10. If conditional PT is not plausible, look for other methods.

In addition, this lecture greatly benefits from the course materials posted by Sant'Anna.

Useful Packages

Roth et. al. (2023) provide a list of useful packages. Of course, when you need to implement a DID estimator, always check for the most up-to-date package.

Table 2
Statistical packages for recent DiD methods.

Heterogeneity Robust Estimators for Staggered Treatment Timing		
Package	Software	Description
did, csdid	R, Stata	Implements Callaway and Sant'Anna (2021)
did2s	R, Stata	Implements Gardner (2021), Borusyak et al. (2021), Sun and Abraham (2021), Callaway and Sant'Anna (2021), Roth and Sant'Anna (2021)
didimputation, did_imputation	R, Stata	Implements Borusyak et al. (2021)
DIDmultiplegt, did_multiplegt	R, Stata	Implements de Chaisemartin and D'Haultfoeuille (2020)
eventstudyinteract	Stata	Implements Sun and Abraham (2021)
flexpaneldid	Stata	Implements Dettmann (2020), based on Heckman et al. (1998)
fixest	R	Implements Sun and Abraham (2021)
stackedev	Stata	Implements stacking approach in Cengiz et al. (2019)
staggered	R	Implements Roth and Sant'Anna (2021), Callaway and Sant'Anna (2021), and Sun and Abraham (2021)
xtevent	Stata	Implements Freyaldenhoven et al. (2019)
DiD with Covariates		
Package	Software	Description
DRDID, drdid	R, Stata	Implements Sant'Anna and Zhao (2020)
Diagnostics for TWFE with Staggered Timing		
Package	Software	Description
bacondecomp, ddtiming	R, Stata	Diagnostics from Goodman-Bacon (2021)
TwoWayFEWeights	R, Stata	Diagnostics from de Chaisemartin and D'Haultfoeuille (2020)
Diagnostic/ Sensitivity for Violations of Parallel Trends		
Package	Software	Description
honestDiD	R, Stata	Implements Rambachan and Roth (2022b)
pretrends	R	Diagnostics from Roth (2022)

Note: This table lists R and Stata packages for recent DiD methods, and is based on Asjad Naqvi's repository at https://asjadnaqvi.github.io/DiD/. Several of the packages listed under "Heterogeneity Robust Estimators" also accommodate covariates.

Useful Packages

Xu also provides many good packages for conducting inference with panel data on his website: `https://yiqingxu.org/software/#panel-data-methods`.

	gsynth	fect	bpCausal	tjbal
Staggered adoption of treatment	✓	✓	✓	✓
General treatment patten (e.g. switch on-and-off)		✓		
Allow additive fixed effects	✓	✓	✓	(✓)
Allow feedback from past outcomes to treatment				✓
Allow short pretreatment period				✓
Semiparametric & distributional effect				✓
Accommodate a small number of treated units	✓		✓	
Easily interpretable (Bayesian) uncertainty estimates			✓	
Placebo tests for pretrend		✓	✓	
Latent factor structure for unobservables	✓	✓	✓	

You can also find many other useful resources on this website, such as Xu's lecture slides on panel data methods.