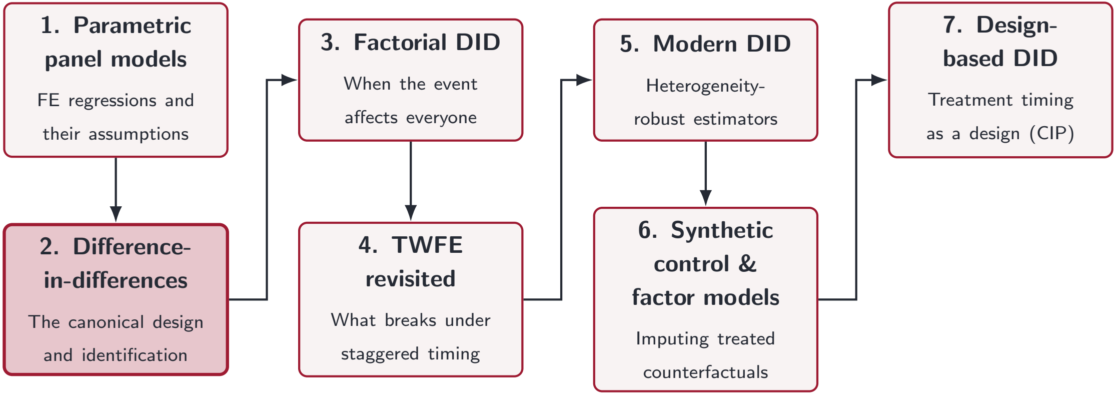
Panel Methods (2): Difference-in-Differences

Waseda University Workshop

Yiqing Xu
Stanford University

September 2026

The Seven Lectures



From Outcome Modeling to "Design"-Based Inference

- In the last lecture, we discussed how to estimate the treatment effect using a parametric, outcome modeling approach
- In this lecture, we will focus on the Difference-in-Differences (DID), a research design-based approach to estimate the treatment effect
- DID is closely connected to TWFE models; like TWFE, The goal is to account for unobserved confounding
- There are other research designs for causal panel analyses, e.g. interrupted time series, synthetic control.

Selection on Unobservables

Problem

Often there are reasons to believe that treated and untreated units differ in unobservable characteristics that are associated with potential outcomes even after controlling for differences in observed characteristics.

In such cases, treated and untreated units are not directly comparable. What can we do then?

This Lecture

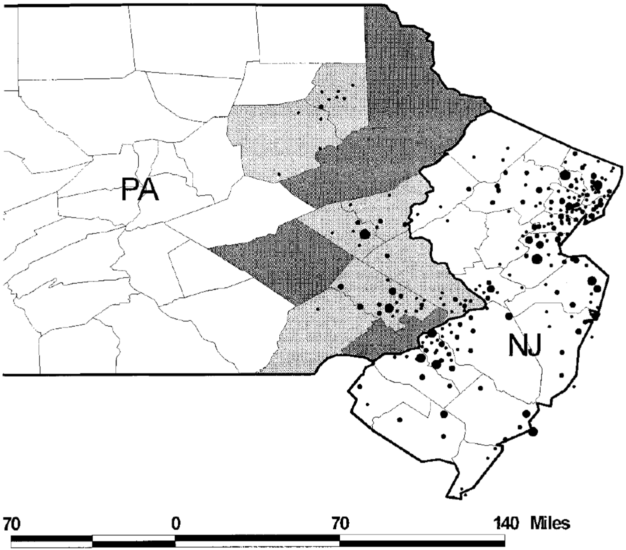
- Motivating Examples
- DID: Setup & Identification
- DID: Estimation and Extensions
- Threats to Validity

Source: Organization informed by Sant'Anna's 2025 PolMeth DID workshop.

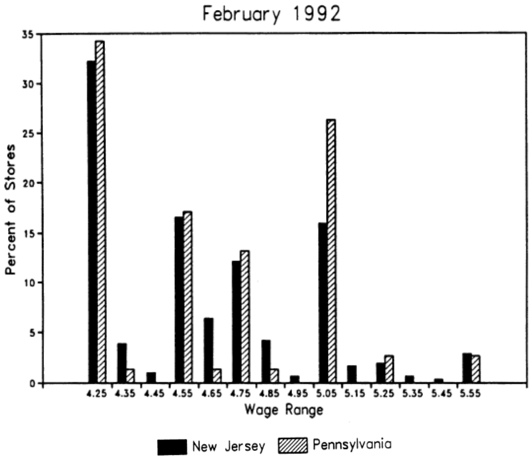
Example: Minimum Wage Laws and Employment

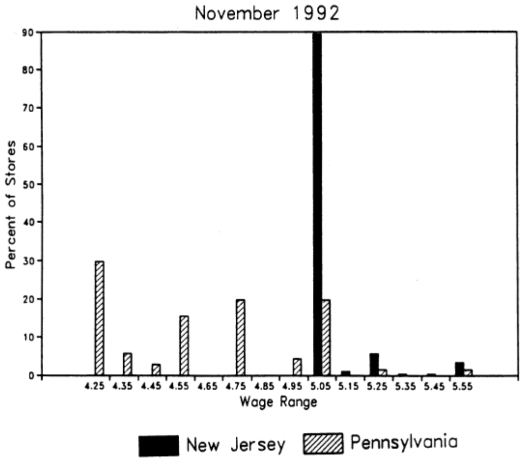
- Do higher minimum wages decrease low-wage employment?
- Card and Krueger (1994) consider impact of New Jersey's 1992 minimum wage increase from $4.25 to $5.05 per hour
- Compare employment in 410 fast-food restaurants in New Jersey and eastern Pennsylvania before and after the rise
- Survey data on wages and employment from two waves:
  - Wave 1: March 1992, one month before the minimum wage increase
  - Wave 2: December 1992, eight months after increase

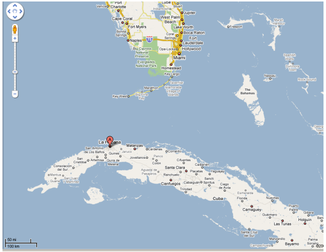
Locations of Restaurants (Card and Krueger 2000)



Wages Before & After Rise in Minimum Wage







- How do inflows of immigrants affect the wages and employment of natives in local labour markets?
- Card (1990) uses the Mariel Boatlift of 1980 as a natural experiment to measure the effect of a sudden influx of immigrants on unemployment among less-skilled natives
- The Mariel Boatlift increased the Miami labor force by 7%
- Individual-level data on unemployment from the Current Population Survey (CPS) for Miami and four comparison cities (Atlanta, Los Angeles, Houston and Tampa-St. Petersburg)

A DID Is a Claim about the Missing Trend

	Before the increase	After the increase
New Jersey	observed	never observed
Pennsylvania	observed	observed

Every cell is $Y_{it}(0)$: employment *without* the increase.

Question

Absent the minimum-wage increase, would New Jersey employment have moved like Pennsylvania's?

- The comparison can be another state, market, firm, or product.
- Here, Pennsylvania must be unaffected by New Jersey's law but exposed to the same regional shocks.

Two Groups and Two Periods

Definition

Two groups:

- $G = 1$ Treated units
- $G = 0$ Control units

Two periods:

- $t = pre$: Pre-Treatment period
- $t = pst$: Post-Treatment period

Treatment assignment:

- $D_{i,t} = 1$ if $G_i = 1$ and $t = pst$; $D_{i,t} = 0$ otherwise

Potential outcomes $Y_t(g)$:

- $Y_{i,t}(1)$ potential outcome unit i attains in period t w/ treatment received between *pre* and *pst*
- $Y_{i,t}(0)$ potential outcome unit i attains in period t w/o treatment received between *pre* and *pst*

Two Groups and Two Periods

Definition

Causal effect for unit i at time t is

- $\tau_{it} = Y_{i,t}(1) - Y_{i,t}(0)$

Observed outcomes $Y_{i,t}$ are realized as

- $Y_{i,t} = Y_{i,t}(0) \cdot (1 - G_i) + Y_{i,t}(1) \cdot G_i$

In the *pst* period, we have,

- $Y_{i,pst} = Y_{i,pst}(0) \cdot (1 - G_i) + Y_{i,pst}(1) \cdot G_i$

Estimand (ATT)

Focus on estimating the average effect of the treatment on the treated: $\tau_{ATT} = E[Y_{i,pst}(1) - Y_{i,pst}(0) | G_i = 1]$

Two Groups and Two Periods

Estimand (ATT)

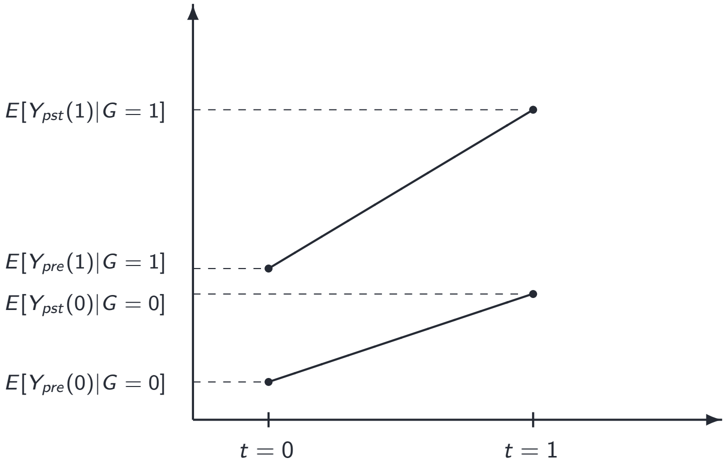
$\tau_{ATT} = E[Y_{pst}(1) - Y_{pst}(0)|G = 1]$

	Pre-Period (T=*pre*)	**Post-Period (T=*pst*)**
Treated G=1	$E[Y_{pre}(1)\|G = 1]$	$E[Y_{pst}(1)\|G = 1]$
Control G=0	$E[Y_{pre}(0)\|G = 0]$	$E[Y_{pst}(0)\|G = 0]$

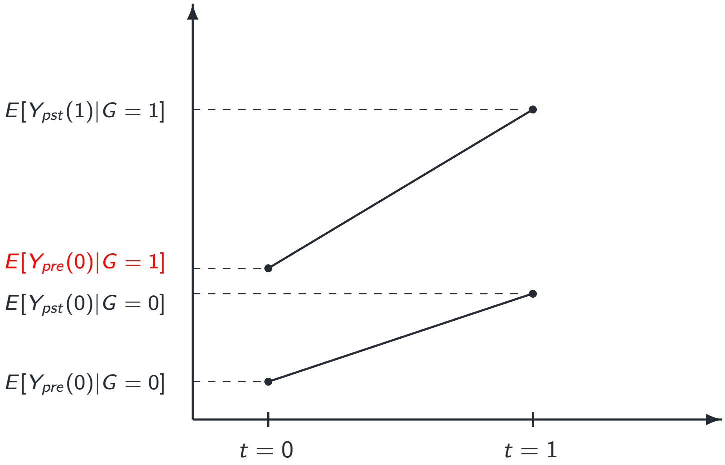
Problem

Missing potential outcome: $E[Y_{pst}(0)|G = 1]$, *ie. what is the average post-period outcome for the treated in the absence of the treatment?*

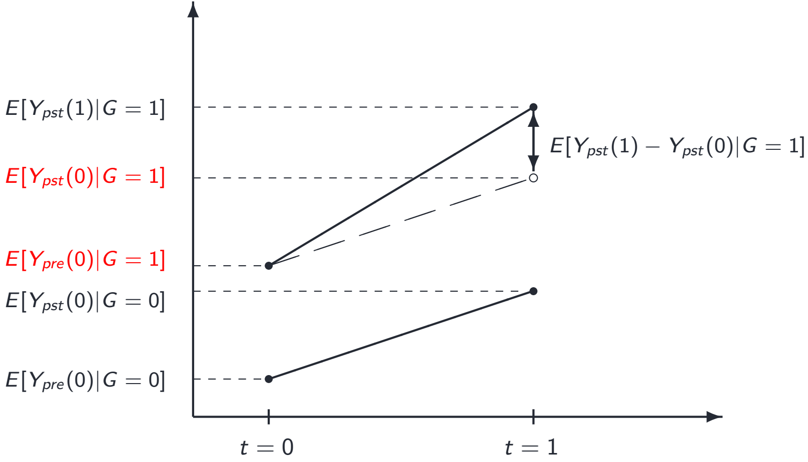
Graphical Representation



Graphical Representation



Graphical Representation



DID: Identification

Identification Assumption (no anticipation)

$$Y_{i,pre}(0) = Y_{pre}(1), i = 1, 2, \cdots, n$$

Identification Assumption (parallel trends)

$$E[Y_{pst}(0) - Y_{pre}(0)|G = 1] = E[Y_{pst}(0) - Y_{pre}(0)|G = 0]$$

Identification Result

Given no anticipation and parallel trends, the ATT is identified as:

$$E[Y_{pst}(1) - Y_{pst}(0)|G = 1] = \Big\{E[Y_{pst}|G = 1] - E[Y_{pst}|G = 0]\Big\} - \Big\{E[Y_{pre}|G = 1] - E[Y_{pre}|G = 0]\Big\}$$

DID: Identification

Identification Assumption (no anticipation)

$$Y_{i,pre}(0) = Y_{pre}(1), i = 1, 2, \cdots, n$$

Identification Assumption (parallel trends)

$$E[Y_{pst}(0) - Y_{pre}(0)|G = 1] = E[Y_{pst}(0) - Y_{pre}(0)|G = 0]$$

Proof.

Plug in $E[Y_{pre}(1)|G = 1] = E[Y_{pre}(0)|G = 1]$ and $E[Y_{pst}(0)|G = 0] = E[Y_{pst}(0)|G = 1] - E[Y_{pre}(0)|G = 1] + E[Y_{pre}(0)|G = 0]$ we get

$$\begin{aligned}
\tau_{DID} &\equiv \{E[Y_{pst}|G = 1] - E[Y_{pst}|G = 0]\} - \{E[Y_{pre}|G = 1] - E[Y_{pre}|G = 0]\} \\
&= \{E[Y_{pst}(1)|G = 1] - E[Y_{pst}(0)|G = 0]\} - \{E[Y_{pre}(0)|G = 1] - E[Y_{pre}(0)|G = 0]\} \\
&= \{E[Y_{pst}(1)|G = 1] - (E[Y_{pst}(0)|G = 1] - E[Y_{pre}(0)|G = 1] + E[Y_{pre}(0)|G = 0])\} \\
&- \{E[Y_{pre}(0)|G = 1] - E[Y_{pre}(0)|G = 0]\} \\
&= E[Y_{pst}(1) - Y_{pst}(0)|G = 1]
\end{aligned}$$

$\square$

DID: Estimators

Estimand (Identifying the ATT)

$$\tau_{DID} = E[Y_{pst}(1) - Y_{pst}(0)|G = 1]$$

Estimator (Sample Means: Panel)

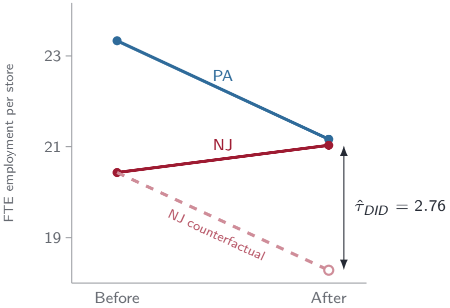
$$\hat{\tau}_{DID} = \left\{ \frac{1}{N_1} \sum_{G_i=1} Y_{i,pst} - \frac{1}{N_0} \sum_{G_i=0} Y_{i,pst} \right\} - \left\{ \frac{1}{N_1} \sum_{G_i=1} Y_{i,pre} - \frac{1}{N_0} \sum_{G_i=0} Y_{i,pre} \right\}$$
$$= \left\{ \frac{1}{N_1} \sum_{G_i=1} \{Y_i(1) - Y_{i,pre}\} - \frac{1}{N_0} \sum_{G_i=0} \{Y_i(1) - Y_{i,pre}\} \right\},$$

where N_1 and N_0 are the number of treated and control units respectively, and

$$\text{plim}\, \hat{\tau}_{DID} = \tau_{DID}$$

Sample Means: Minimum Wage Laws and Employment

FTE employment	PA	NJ	NJ−PA
Before	23.33	20.44	−2.89
After	21.17	21.03	−0.14
Change	−2.16	0.59	**2.76**

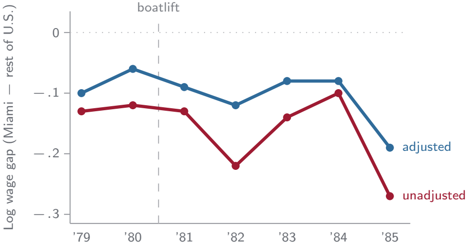


Source: Card and Krueger (1994), Table 3, rows 1–3.

Sample Means: The Mariel Boatlift

	Mean log wage		Miami − outside	
	Miami	Outside	Unadj.	Adj.
1979	1.58	1.71	−.13	−.10
1980	1.54	1.66	−.12	−.06
1981	1.51	1.63	−.13	−.09
1982	1.49	1.71	−.22	−.12
1983	1.48	1.62	−.14	−.08
1984	1.53	1.63	−.10	−.08
1985	1.49	1.77	−.27	−.19

Cubans in Miami vs. Cubans elsewhere; 1979 is the last full pre-boatlift year. Adjusted controls for education, potential experience, gender, marital status, part-time status, and public vs. private sector.



Source: Card (1990), Table 7; the two difference columns are his ninth and tenth.

DID: Estimators

Estimator (Regression: Panel Data)

With panel data we can estimate the difference-in-differences effect using a fixed effects regression with unit and period fixed effects:

$$Y_{it} = \mu + \gamma_i + \delta T + \tau D_{it} + X_{it}'\beta + \varepsilon_{it}$$

- *One intercept for each unit* γ_i
- D_{it} *coded as 1 for treated in post-period and 0 otherwise*

Or equivalently we can use regression with the dependent variable in first differences:

$$\Delta Y_i = \delta + \tau \cdot G_i + u_i,$$

where $\Delta Y_i = Y_i(1) - Y_{i,pre}$ *and* $u_i = \Delta \varepsilon_i$.

R Demo: Card–Krueger in One Regression

```r
library(haven)
library(fixest)

ck <- read_dta("figs/CK1994_longformat.dta")
m <- feols(emptot ~ nj * postperiod | ID,
           data = ck, cluster = ~ID)
etable(m)
```

	Estimate	Clustered SE
NJ $\times$ post	2.750	1.338

The interaction coefficient is the DID estimate.

IPW (reweighting) + DID

- **Assumption**: non-parallel outcome dynamics between treated and controls caused by observed characteristics, or conditional parallel trends
- Abadie (2005) proposes a two-step strategy:
  1. estimate the propensity score based on observed covariates; compute the fitted value
  2. run a weighted DID model
- $\tau_{IPW} = \mathbb{E}\left\{\frac{G_i \cdot \Delta Y_i}{\pi(X_i)}\right\} - \mathbb{E}\left\{\frac{(1-G_i)\cdot \Delta Y_i}{1-\pi(X_i)}\right\}$
- The idea of using pre-treatment variables to adjust trends is a precursor to synthetic control
- Balancing is an alternative reweighting strategy and has better finite sample properties

Balancing + DID

- Mean balancing: on original features (Robbins et al. 2017)

$$\sum_{i\in\mathcal{T}} q_i \mathbf{Y}_{i,pre} = \sum_{j\in\mathcal{C}} w_j \mathbf{Y}_{j,pre}$$

- Trajectory balancing: feature mapping $\mathbf{Y}_{i,pre} \mapsto \phi(\mathbf{Y}_{i,pre})$, then balance on the expanded features (Hazlett and Xu 2018):

$$\phi : \mathbb{R}^P \mapsto \mathbb{R}^{P'}$$
$$\sum_{i\in\mathcal{T}} q_i \phi(\mathbf{Y}_{i,pre}) = \sum_{j\in\mathcal{C}} w_j \phi(\mathbf{Y}_{j,pre})$$

- In practice: seek approximate balance, working from largest toward smallest principal components of $\mathbf{Y}_{pre}(\mathbf{Y}_{pre})'$ with a stopping rule of minimizing the upper bound of biases

This Lecture

- Motivating Examples
- DID: Setup & Identification
- DID: Estimation
- DID: Extensions
- **Threats to Validity**

Threats to Validity

1. Non-parallel dynamics
2. Compositional differences
3. Long-term effects versus reliability
4. Scale dependence
5. Staggered adoption: heterogeneous treatment effects & the weighting problem ($\rightarrow$ Lecture 4)

Bias is a matter of degree. Small violations of the identification assumptions may not matter much as the bias may be rather small. However, biases can sometimes be so large that the estimates we get are completely wrong, even of the opposite sign of the true treatment effect.

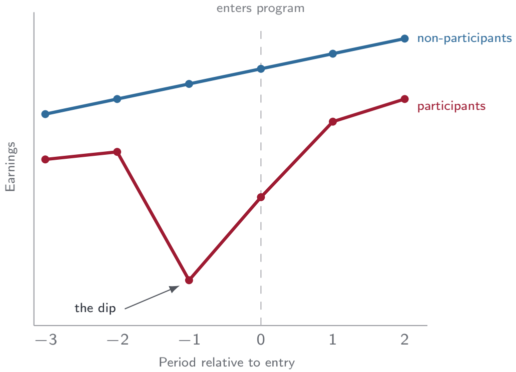
Helpful to avoid overly strong causal claims for difference-in-differences estimates.

Non-parallel Dynamics

Often treatments/programs are targeted based on pre-existing differences in outcomes. "Feedback" or time-varying confounding cause failure of the parallel trends assumption.

- "Ashenfelter dip": participants in training programs often experience a dip in earnings just before they enter the program (that may be *why* they participate). Since wages have a natural tendency to mean reversion, comparing wages of participants and non-participants using DD leads to an upward biased estimate of the program effect
- Regional targeting: NGOs may target villages that appear most promising (or worst off)

The Ashenfelter Dip



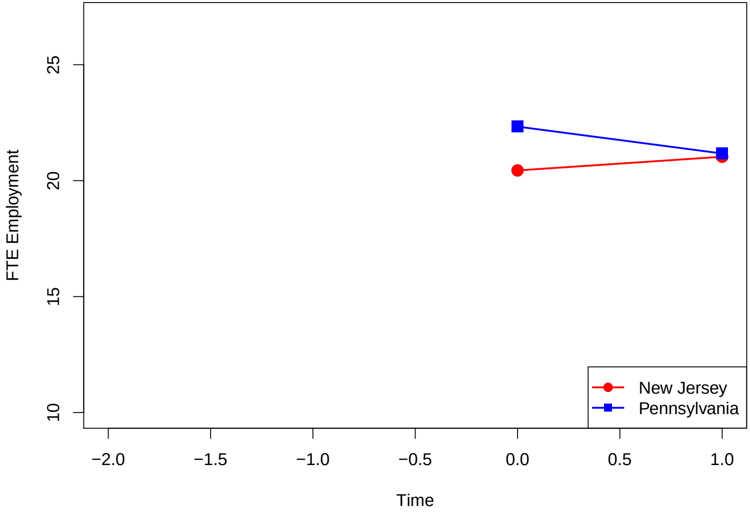
What happens

Earnings fall just before entry, which is often *why* people enter, then recover on their own.

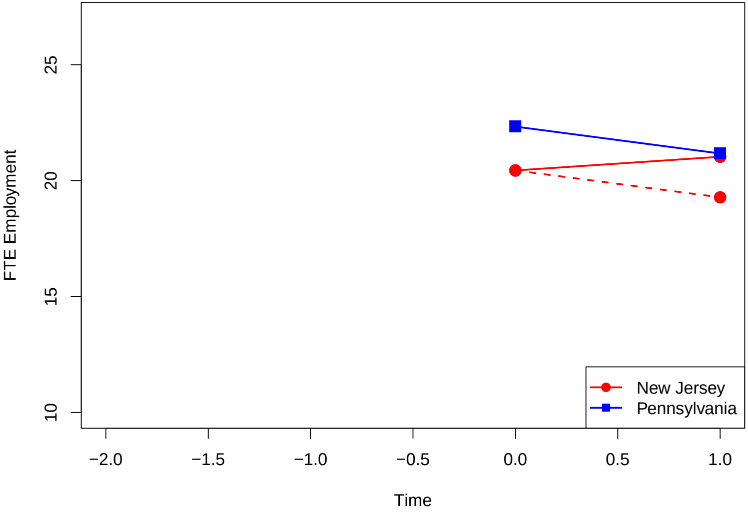
Why DID is biased

Using $t = -1$ as the baseline counts ordinary recovery as program effect, so the estimate is too large.

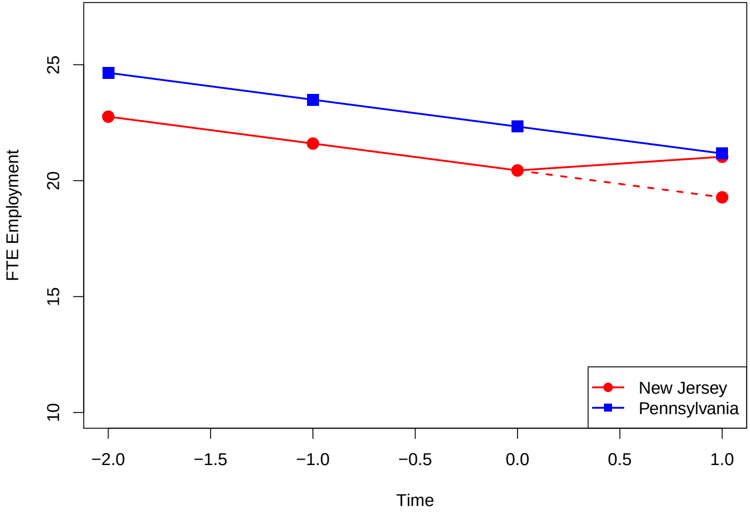
Falsification Test using Data from Pretreatment Periods



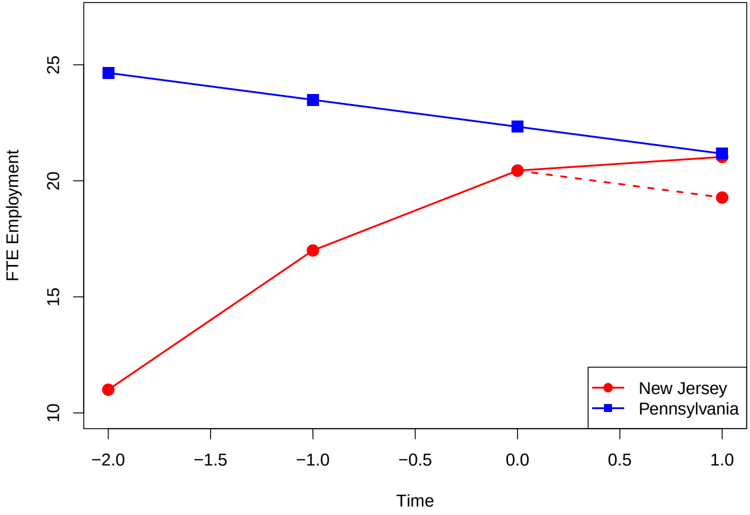
Falsification Test using Data from Pretreatment Periods



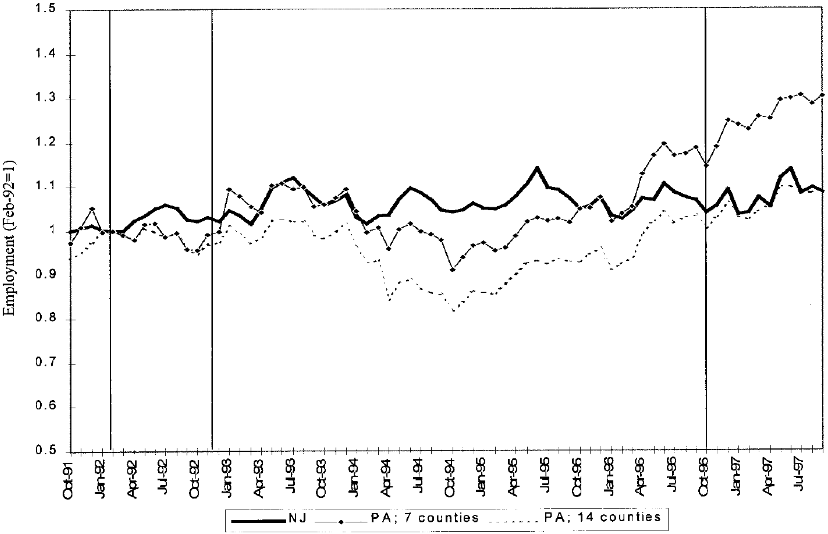
Falsification Test using Data from Pretreatment Periods



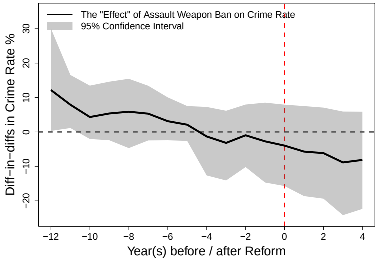
Falsification Test using Data from Pretreatment Periods

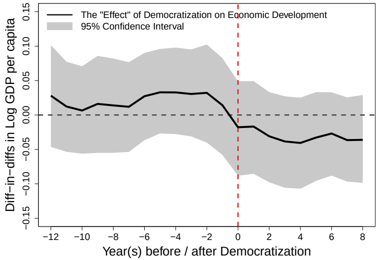


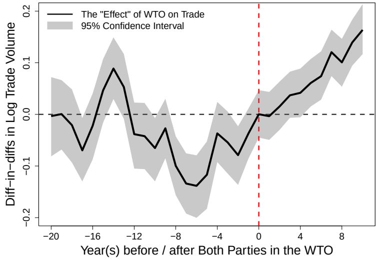
Longer Trends in Employment (Card and Krueger 2000)

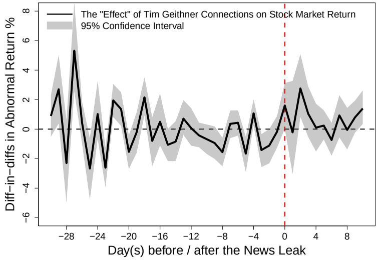


Signs of Parallel Trends Violations

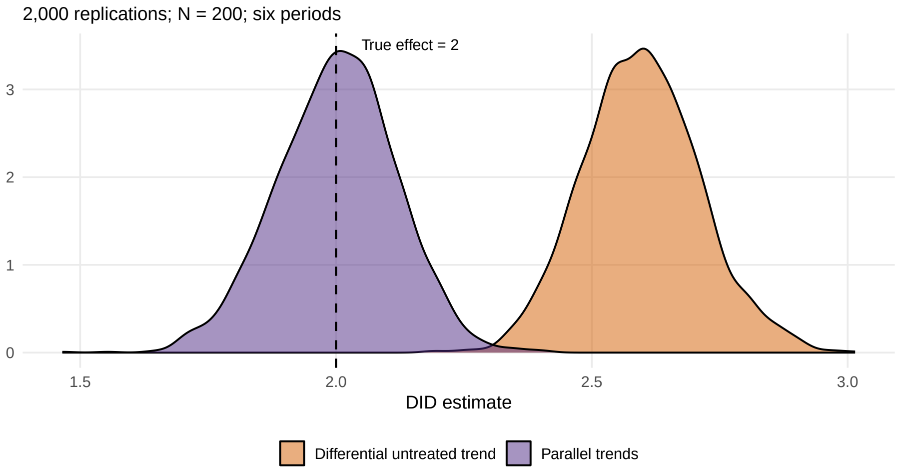








Simulation: Parallel Trends and Bias



- When untreated trends are parallel, the estimates center on the true effect.
- Even a modest differential trend biases DID.

Falsification Test: Alternative Control Group

	Stores by state			Stores in New Jersey[a]			Differences within NJ[b]	
Variable	PA (i)	NJ (ii)	Difference, NJ − PA (iii)	Wage = \$4.25 (iv)	Wage = \$4.26–\$4.99 (v)	Wage ≥ \$5.00 (vi)	Low–high (vii)	Midrange–high (viii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	−2.89 (1.44)	19.56 (0.77)	20.08 (0.84)	22.25 (1.14)	−2.69 (1.37)	−2.17 (1.41)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	−0.14 (1.07)	20.88 (1.01)	20.96 (0.76)	20.21 (1.03)	0.67 (1.44)	0.75 (1.27)
3. Change in mean FTE employment	−2.16 (1.25)	0.59 (0.54)	2.76 (1.36)	1.32 (0.95)	0.87 (0.84)	−2.04 (1.14)	3.36 (1.48)	2.91 (1.41)

If the DID with the alternative control is different from the DID with the original control, then the original DID may be biased

Triple DID: Mandated Maternity Benefits (Gruber, 1994)

TABLE 3—DDD ESTIMATES OF THE IMPACT OF STATE MANDATES ON HOURLY WAGES

Location/year	Before law change	After law change	Time difference for location
A. *Treatment Individuals: Married Women, 20–40 Years Old:*			
Experimental states	1.547 (0.012) [1,400]	1.513 (0.012) [1,496]	−0.034 (0.017)
Nonexperimental states	1.369 (0.010) [1,480]	1.397 (0.010) [1,640]	0.028 (0.014)
Location difference at a point in time:	0.178 (0.016)	0.116 (0.015)	
Difference-in-difference:	−0.062 (0.022)		
B. *Control Group: Over 40 and Single Males 20–40:*			
Experimental states	1.759 (0.007) [5,624]	1.748 (0.007) [5,407]	−0.011 (0.010)
Nonexperimental states	1.630 (0.007) [4,959]	1.627 (0.007) [4,928]	−0.003 (0.010)
Location difference at a point in time:	0.129 (0.010)	0.121 (0.010)	
Difference-in-difference:	−0.008: (0.014)		
DDD:	**−0.054** (0.026)		

How Useful is Triple DID?

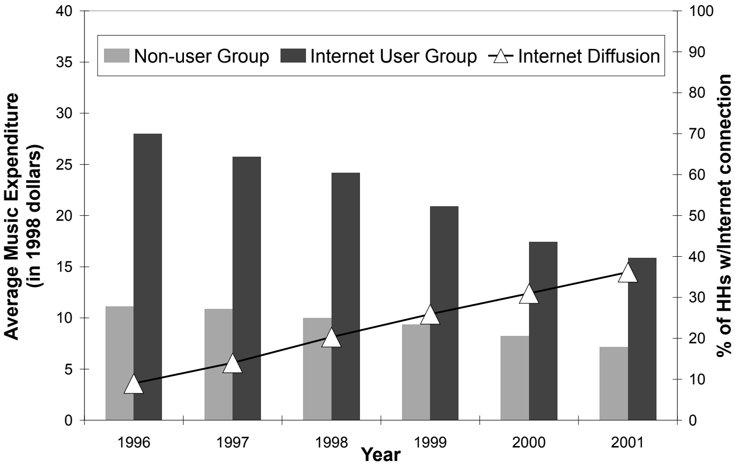
- The triple DID estimate is the difference between the DID of interest and the placebo DID (that is supposed to be zero)
  - If the placebo DID is nonzero, it might be difficult to convince reviewers that the triple DID removes all the bias
  - If the placebo DID is zero, then DID and triple DID give the same results but DID is preferable because standard errors are smaller for DD than for triple DID

Compositional Differences

- In repeated cross-sections, we do not want that the composition of the sample changes between periods.
- Examples:
  - Hong (2011) uses repeated cross-sectional data from Consumer Expenditure Survey (CEX) containing music expenditures and internet use for random samples of U.S. households
  - Study exploits the emergence of Napster (the first sharing software widely used by Internet users) in June 1999 as a natural experiment.
  - Study compares internet users and internet non-users, before and after emergence of Napster

Compositional Differences?

Figure 1: Internet Diffusion and Average Quarterly Music Expenditure in the CEX



Compositional Differences?

Table 1: Descriptive Statistics for Internet User and Non-user Groups[a]

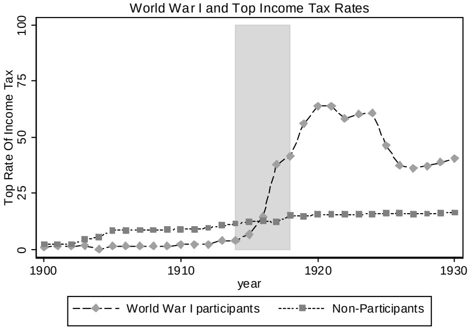
Year	1997		1998		1999	
	Internet User	Non-user	Internet User	Non-user	Internet User	Non-user
Average Expenditure						
Recorded Music	$25.73	$10.90	$24.18	$9.97	$20.92	$9.37
Entertainment	$195.03	$96.71	$193.38	$84.92	$182.42	$80.19
Zero Expenditure						
Recorded Music	.56	.79	.60	.80	.64	.81
Entertainment	.08	.32	.09	.35	.14	.39
Demographics						
Age	40.2	49.0	42.3	49.0	44.1	49.4
Income	$52,887	$30,459	$51,995	$28,169	$49,970	$26,649
High School Grad.	.18	.31	.17	.32	.21	.32
Some College	.37	.28	.35	.27	.34	.27
College Grad.	.43	.21	.45	.21	.42	.20
Manager	.16	.08	.16	.08	.14	.08

Diffusion of the internet changes samples (e.g. younger music fans are early adopters)

Long-term Effects versus Reliability

- Parallel trends assumption for DID is more likely to hold over a shorter time-window
- In the long-run, many other things may happen that could confound the effect of the treatment
- Should be cautious to extrapolate short-term effects to long-term effects

Effect of War on Tax Rates (Scheve and Stasavage 2010)



Scale Dependence

Magnitude or even sign of the DID effect may be sensitive to the functional form, when average outcomes for controls and treated are very different at baseline

- Training program for the young:
    - Employment for the young increases from 20% to 30%
    - Employment for the old increases from 5% to 10%
    - Positive DID effect: $(30 - 20) - (10 - 5) = 5\%$ increase
    - But if you consider log changes in employment, the DD is, $[log(30) - log(20)] - [log(10) - log(5)] = log(1.5) - log(2) < 0$
- DD estimates may be more reliable if treated and controls are more similar at baseline
- More similarity may help with parallel trends assumption

When the Parallel Trends Assumption is More Defensible?

Roth and Sant'Anna (2021)

- The parallel trends assumption is scale-dependent
- When is the assumption not sensitive to strictly monotonic transformation of the outcome?
- A "stronger parallel trends" for the entire distribution of $Y_{it}(0)$

$$F^{Y(0)}_{G=1,t=1}(y) - F^{Y(0)}_{G=1,t=0}(y) = F^{Y(0)}_{G=0,t=1}(y) - F^{Y(0)}_{G=0,t=0}(y), \qquad \text{for all } y \in \mathcal{R}$$

- It holds when the population are consists of
  - A subgroup in which the treatment is as-if randomly assigned
  - A subgroup in which the distribution of $Y_{it}(0)$ is stable over time

Conclusion

- DID is a powerful research design for causal inference with panel or repeated cross-sectional data
- The parallel trends assumption embeds a functional form requirement, which is scale dependent
- Researchers should be concerned about its validity and use various falsification tests to probe its plausibility