Panel Methods (2): Difference-in-Differences

Northwestern Causal Inference Main Workshop

Yiqing Xu
Stanford University

July 31st, 2024

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?

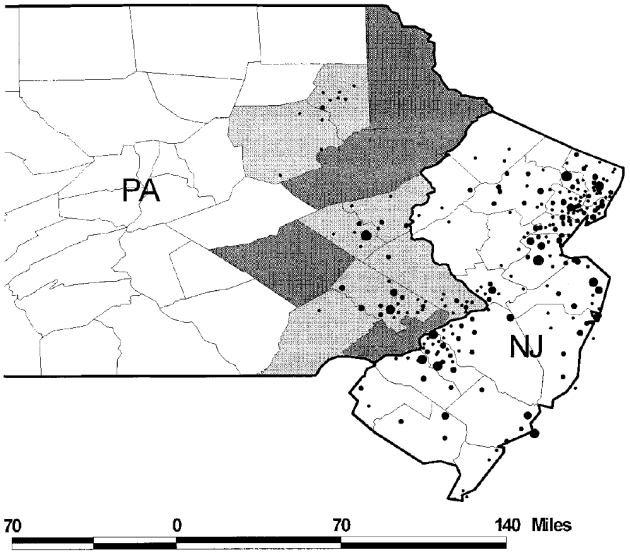
This Lecture

- Motivating Examples
- DID: Setup & Identification
- DID: Estimation
- DID: Extensions
- Connections to TWFE Models
- Threats to Validity

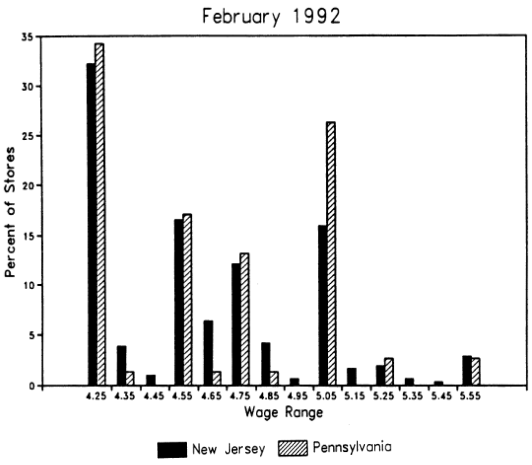
Example: Minimum Wage Laws and Employment

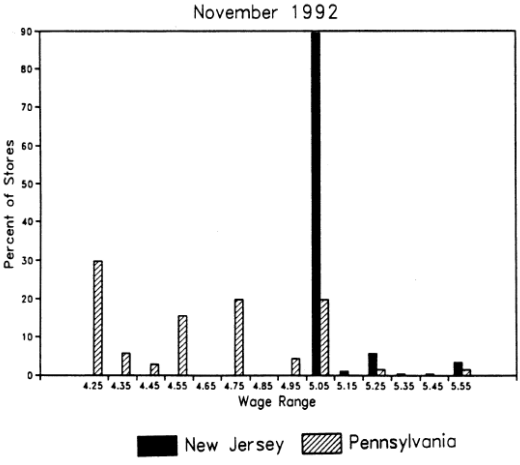
- Do higher minimum wages decrease low-wage employment?
- Card and Krueger (1994) consider impact of New Jersey's 1992 minimum wage increase from \$4.25 to \$5.05 per hour
- Compare employment in 410 fast-food restaurants in New Jersey and eastern Pennsylvania before and after the rise
- Survey data 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)

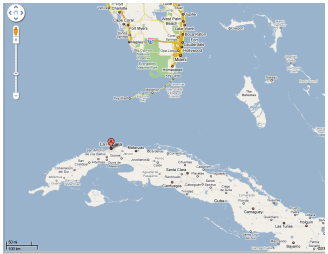


Wages Before & After Rise in Minimum Wage





Example: The Mariel Boatlift



- 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)

Example: The Mariel Boatlift

Table 7. Means of Log Wages of Cubans in Miami: Actual and Predicted, and by Quartile of Predicted Wages. (Standard Errors in Parentheses)

	Mean of Log Wages Log in Miami			*Mean of Log Wages by Quartile of Predicted Wages*				*Mean Log Wage of Cubans*	*Difference in Cuban Wages, Miami − Rest-of-U.S.*	
Year	*Actual*	*Pre-dicted*	*Actual-Pre-dicted*	*1st*	*2nd*	*3rd*	*4th*	*Outside Miami*	*Actual*	*Ad-justed*
1979	1.58	1.73	−.15	1.31	1.44	1.64	1.90	1.71	−.13	−.10
	(.02)	(.02)	(.03)	(.02)	(.03)	(.04)	(.05)	(.04)	(.04)	(.04)
1980	1.54	1.68	−.14	1.25	1.49	1.59	1.81	1.66	−.12	−.06
	(.02)	(.02)	(.03)	(.02)	(.05)	(.04)	(.05)	(.03)	(.04)	(.03)
1981	1.51	1.68	−.17	1.23	1.43	1.55	1.80	1.63	−.13	−.09
	(.02)	(.02)	(.03)	(.03)	(.03)	(.04)	(.05)	(.03)	(.04)	(.03)
1982	1.49	1.68	−.19	1.27	1.43	1.50	1.77	1.71	−.22	−.12
	(.02)	(.02)	(.03)	(.03)	(.04)	(.04)	(.06)	(.03)	(.04)	(.03)
1983	1.48	1.65	−.17	1.16	1.41	1.56	1.80	1.62	−.14	−.08
	(.03)	(.02)	(.03)	(.02)	(.04)	(.04)	(.06)	(.03)	(.04)	(.03)
1984	1.53	1.69	−.17	1.20	1.40	1.65	1.88	1.63	−.10	−.08
	(.03)	(.02)	(.03)	(.03)	(.04)	(.05)	(.06)	(.03)	(.04)	(.03)
1985	1.49	1.67	−.18	1.19	1.43	1.53	1.80	1.77	−.27	−.19
	(.04)	(.03)	(.05)	(.06)	(.06)	(.08)	(.09)	(.06)	(.07)	(.05)

Notes: Predicted wage is based on a linear prediction equation for the log wage fitted to individuals in four comparison cities; see text. Predicted wages for Cubans in Miami are based on coefficients for Hispanics in comparison cities. The adjusted wage gap between Cubans in Miami and Cubans in the rest of the U.S. are obtained from a linear regression model that includes education, potential experience, and other control variables; see text. Wages are deflated by the Consumer Price Index (1980 = 100).

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*

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:
$\tau_{ATT} = E[Y_{i,pst}(1) - Y_{i,pst}(0) | G_i = 1]$

Estimand (ATT)

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

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

Estimand (ATT)

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

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

Control Strategy: Before vs. After

- Use: $E[Y_{pst}|G = 1] - E[Y_{pre}|G = 1]$

Estimand (ATT)

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

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

Control Strategy: Before-After Comparison

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

Estimand (ATT)

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

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

Control Strategy: Treated-Control Comparison in Post-Period

- Use: $E[Y_{pst}|G = 1] - E[Y_{pst}|G = 0]$

Estimand (ATT)

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

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

Control Strategy: Treated-Control Comparison in Post-Period

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

Estimand (ATT)

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

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

Control Strategy: Difference-in-Differences (DD)

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

Estimand (ATT)

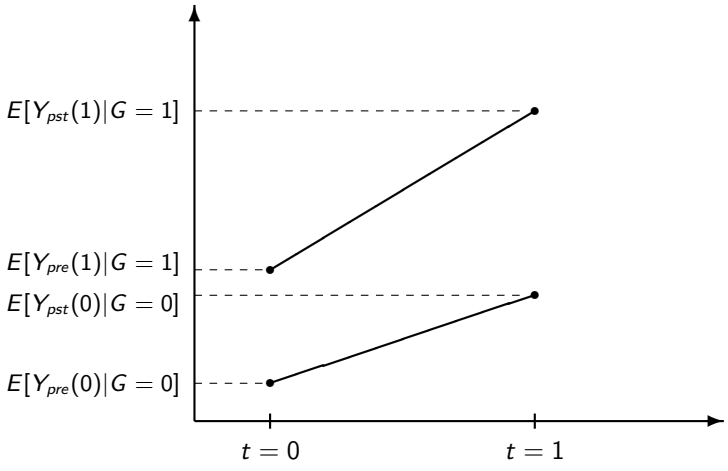
$\tau_{ATT} = E[Y_{pst}(1) - Y_{pst}(0)|G = 1]$

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

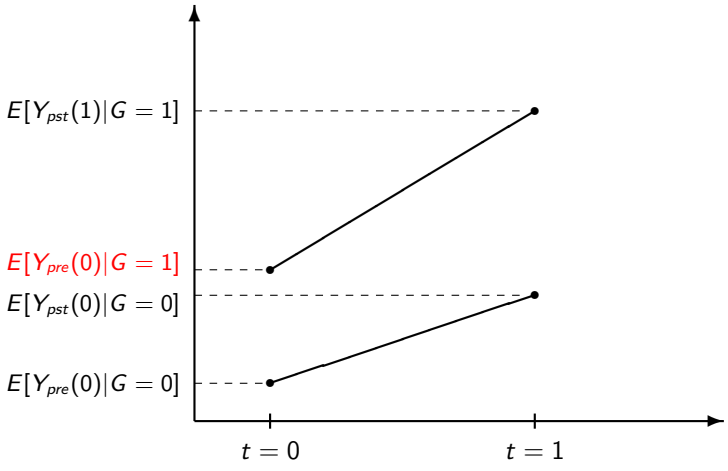
Control Strategy: Difference-in-Differences (DD)

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

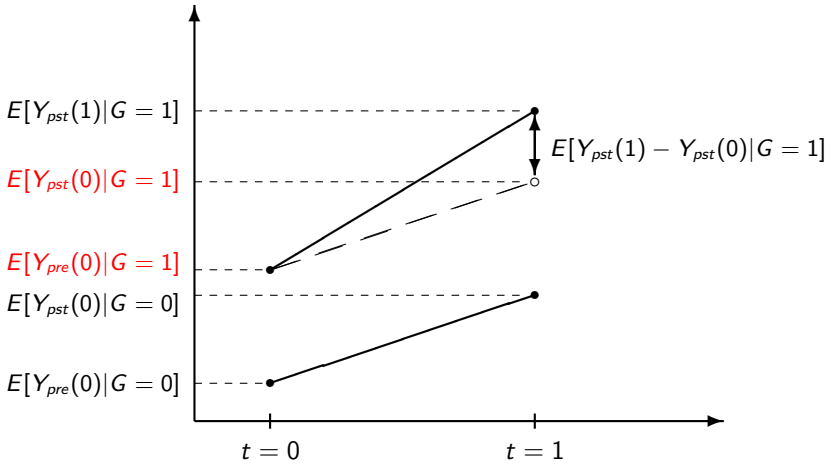
Graphical Representation



Graphical Representation



Graphical Representation



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:

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

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

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

□

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

Variable	Stores by state		
	PA (i)	NJ (ii)	Difference, NJ − PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	−2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	−0.14 (1.07)
3. Change in mean FTE employment	−2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

Estimator (Sample Means: Repeated Cross-Sections)

Let $\{Y_i, G_i, T_i\}_{i=1}^n$ be the pooled sample (the two different cross-sections merged) where T is a random variable that indicates the period (0 or 1) in which the individual is observed.

The difference-in-differences estimator is given by:

$$\left\{ \frac{\sum G_i \cdot T_i \cdot Y_i}{\sum G_i \cdot T_i} - \frac{\sum (1 - G_i) \cdot T_i \cdot Y_i}{\sum (1 - G_i) \cdot T_i} \right\} - \left\{ \frac{\sum G_i \cdot (1 - T_i) \cdot Y_i}{\sum G_i \cdot (1 - T_i)} - \frac{\sum (1 - G_i) \cdot (1 - T_i) \cdot Y_i}{\sum (1 - G_i) \cdot (1 - T_i)} \right\}$$

Estimator (Regression: Repeated Cross-Sections)

Alternatively, the same estimator can be obtained using regression techniques. Consider the linear model:

$$Y = \mu + \gamma \cdot G + \delta \cdot T + \tau \cdot (G \cdot T) + \varepsilon,$$

where $E[\varepsilon|G, T] = 0$.

Easy to show that τ estimates the DD effect:

$$\tau = \tau_{DID}$$

Estimator (Regression: Repeated Cross-Sections)

Alternatively, the same estimator can be obtained using regression techniques. Consider the linear model:

$$Y = \mu + \gamma \cdot G + \delta \cdot T + \tau \cdot (G \cdot T) + \varepsilon,$$

where $E[\varepsilon|G, T] = 0$.

	After (T=1)	**Before (T=0)**	**After - Before**
Treated G=1	$\mu + \gamma + \delta + \tau$	$\mu + \gamma$	$\delta + \tau$
Control G=0	$\mu + \delta$	μ	δ
Treated - Control	$\gamma + \tau$	γ	τ

Regression: Minimum Wage Laws and Employment

```
. g njXpostperiod = nj*postperiod

. reg emptot nj postperiod njXpostperiod
```

Source	SS	df	MS	Number of obs	=	794
				F(3, 790)	=	1.96
Model	521.116464	3	173.705488	Prob > F	=	0.1180
Residual	69887.878	790	88.4656683	R-squared	=	0.0074
				Adj R-squared	=	0.0036
Total	70408.9944	793	88.7881392	Root MSE	=	9.4056

emptot	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
nj	-2.891761	1.193524	-2.42	0.016	-5.234614	-.5489079
postperiod	-2.165584	1.515853	-1.43	0.154	-5.14116	.8099912
njXpostperiod	2.753606	1.688409	1.63	0.103	-.560693	6.067905
_cons	23.33117	1.07187	21.77	0.000	21.22712	25.43522

Estimator (Regression: Repeated Cross-Sections)

Can use regression version of the DD estimator to include covariates:

$$Y = \mu + \gamma \cdot G + \delta \cdot T + \tau \cdot (G \cdot T) + X'\beta + \varepsilon.$$

- *introducing time-invariant X's is not helpful (they get differenced-out)*
- *be careful with time-varying X's: they are often affected by the treatment and may introduce endogeneity (e.g. price of meal)*
- *always correct standard errors to account for temporal dependence*

Can interact time-invariant covariates with the time indicator:

$$Y = \mu + \gamma \cdot G + \delta \cdot T + \alpha \cdot (G \cdot T) + X'\beta_0 + (T \cdot X')\beta_1 + \varepsilon$$

$\Rightarrow X$ is used to explain differences in trends.

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* $u_i = \Delta \varepsilon_i$.

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

- 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

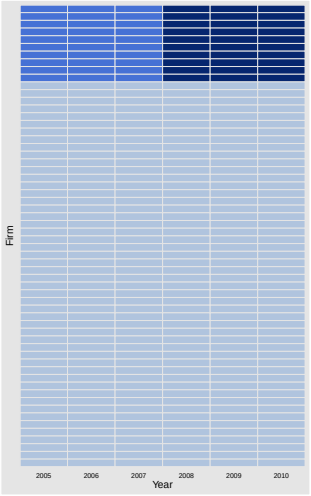
When Averages Fail and $\phi()$'s Thrive

Intuition: mean balancing is okay but may emphasize "wrong" features of the pre-treatment trend

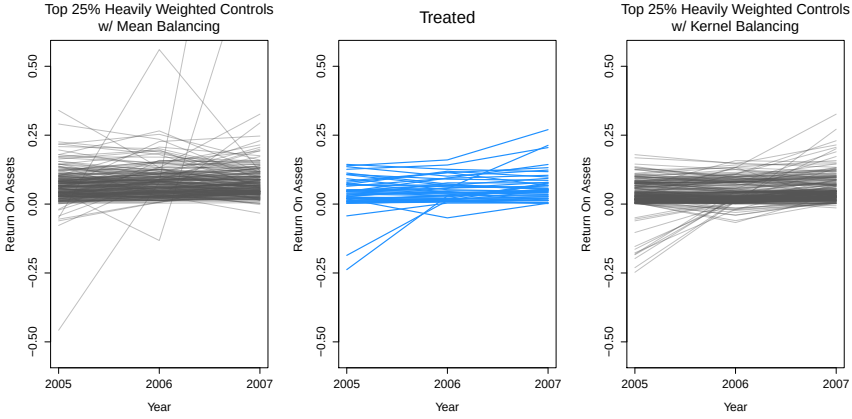
- Trajectory balancing gets you similarity of whole trajectories rather than just equal means at each time point $\rightarrow$ balance on "higher-order" features such as variance, curvature, etc.
- Approximately, trajectory balancing gets multivariate distribution of $\mathbf{V}_i$ for the controls equal to that of the treated, whereas mean balancing only gets equal marginals
- This can matter when non-linear functions of $\mathbf{V}_i$ are confounders, especially when T_0 short

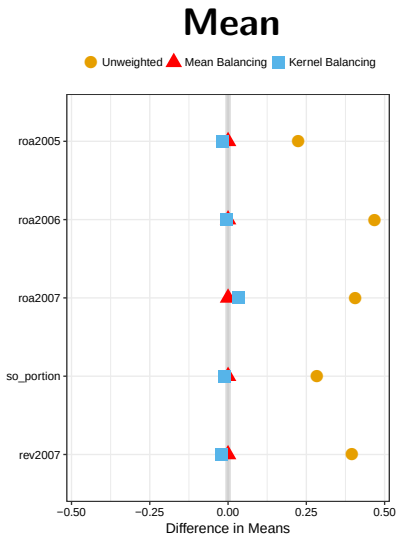
Truex (2014): Return to office in China's Parliament

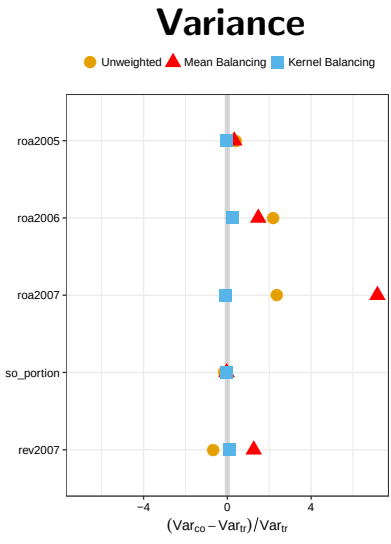
- **Treatment**: CEO taking a seat in the National People's Congress (NPC)
  Outcome: Return on assets (ROA)
- 48 treated firms, 984 controls
  Pre-treatment: 2005-2007
  Post-treatment: 2008-2010
- Two covariates: state ownership, revenue in 2007
- Balancing on: `roa2005, roa2006, roa2007, so_portion, rev2007` (and higher order terms through a kernel transformation)

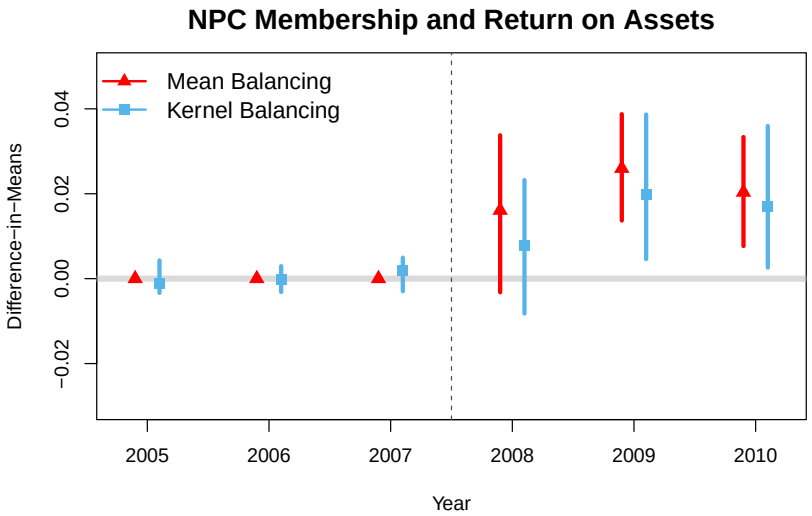


Balance on Pre-treatment Outcome Trajectories




Mean
Unweighted
Mean Balancing
Kernel Balancing
roa2005
roa2006
roa2007
so_portion
rev2007
−0.50
−0.25
0.00
0.25
0.50
Difference in Means


Variance
Unweighted
Mean Balancing
Kernel Balancing
roa2005
roa2006
roa2007
so_portion
rev2007
−4
0
4
(Varco − Vartr)/Vartr


NPC Membership and Return on Assets
Mean Balancing
Kernel Balancing
Difference−in−Means
0.04
0.02
0.00
−0.02
2005
2006
2007
2008
2009
2010
Year

This Lecture

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

- The parallel trends assumption implies additive fixed effects

$$Y_{it} = \tau_{it} D_{it} + \alpha_i + \xi_t + \varepsilon_{it}$$

or

$$\begin{cases} Y_{it}(0) &= \alpha_i + \xi_t + \varepsilon_{it} \\ Y_{it}(1) &= Y_{it}(0) + \tau_{it} \end{cases}$$

- In addition, TWFE model assumes constant and contemporaneous treatment effect:

$$Y_{it} = \tau D_{it} + X'\beta + \alpha_i + \xi_t + \varepsilon_{it}$$

$\Rightarrow$ the (negative) weighting problem

- Strict exogeneity is strong and complex in multi-period cases:

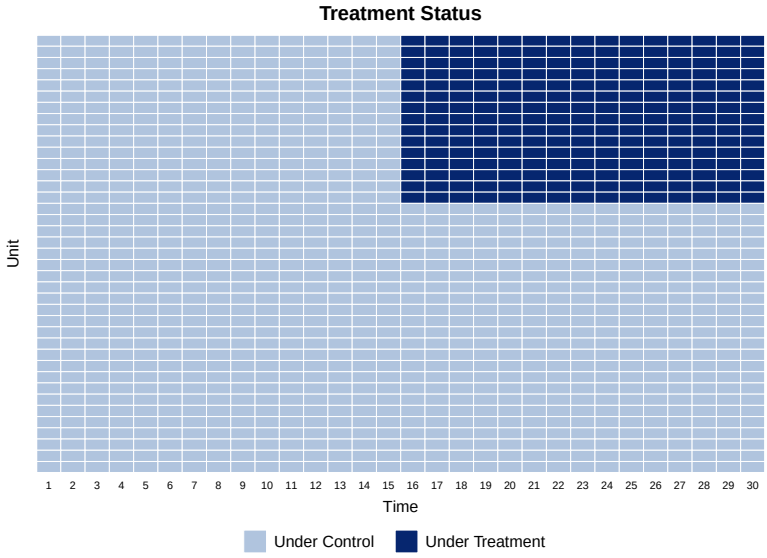
$$\varepsilon_{it} \perp\!\!\!\perp D_{js}, X_{js}, \alpha_j, \xi_s \qquad \forall i, j, t, s$$

$$\Rightarrow \quad \{Y_{it}(0), Y_{it}(1)\} \perp\!\!\!\perp D_{js} | \boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{\xi} \qquad \forall i, j, t, s$$

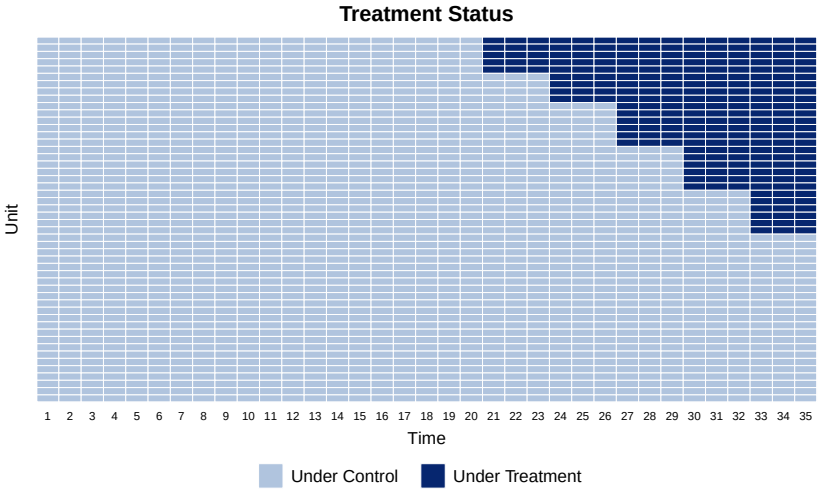
if only two groups, parallel trends:

$$\Rightarrow \quad \mathbb{E}[Y_{it}(0) - Y_{it'}(0) | \boldsymbol{X}] = \mathbb{E}[Y_{jt}(0) - Y_{jt'}(0) | \boldsymbol{X}] \quad i \in \mathcal{T},\ j \in \mathcal{C}, \forall t, t'$$

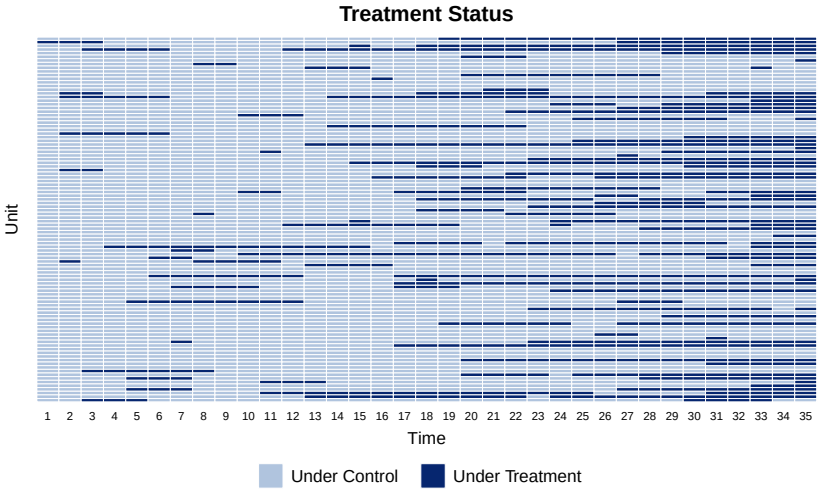
DID $>$ 2 Periods: Classic



DID $>$ 2 Periods: Staggered Adoption



The General Case: w/ Treatment Reversal



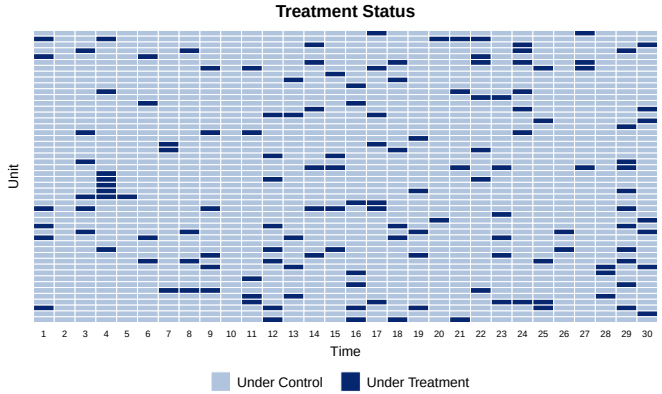
Hypothetical Experiment?

DGPs consistent with strict exogeneity:

$$\alpha_i, \boldsymbol{X_i} \rightarrow \boldsymbol{D_i} \rightarrow \boldsymbol{Y_i}$$

treatment status are assigned randomly or at one shot, not sequentially!

Examples: random assignment within units



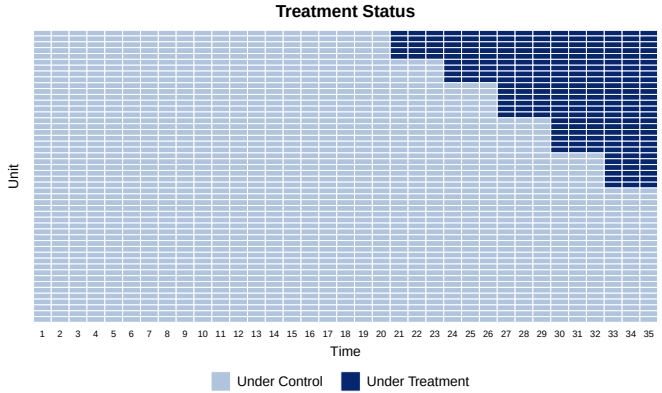
Hypothetical Experiment?

Strict exogeneity implies the following data generating processes:

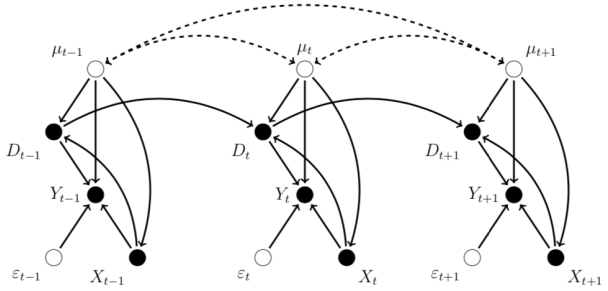
$$\alpha_i, \boldsymbol{X_i} \rightarrow A_i \rightarrow \boldsymbol{D_i} \rightarrow \boldsymbol{Y_i}$$

treatment status are assigned randomly or at one shot, not sequentially!

Examples: staggered adoption (Athey and Imbens 2018)



Reinterpreting Strict Exogeneity (Imai and Kim 2019)



Note: Unit indicies are dropped for simplicity. Vector μ_t represents unobserved time-invariant and decomposable time-varying (for IFEct and MC) confounders.

1. No unobserved time-varying confounder exists
2. No LDV $\rightarrow$ adding LDV is fine if T is large; not adding LDV is also
3. No "carryover effect" $\rightarrow$ adding lagged terms is fine
4. Past outcomes don't directly affect current treatment (no "feedback")

This Lecture

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

Threats to Validity

1. Non-parallel dynamics
2. Compositional differences
3. Long-term effects versus reliability
4. Scale dependence
5. Staggered DID: Heterogeneous treatment effects (HTE) & the weighting problem

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)

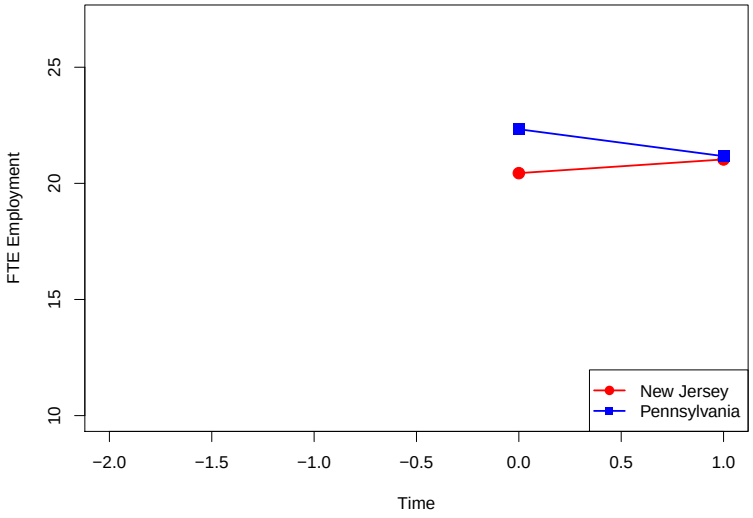
Checks for DID Design

1. Falsification test using data for prior periods
   - Given parallel trends during periods $t = T_0 - 1, T_0, T_0 + 1$, we have:

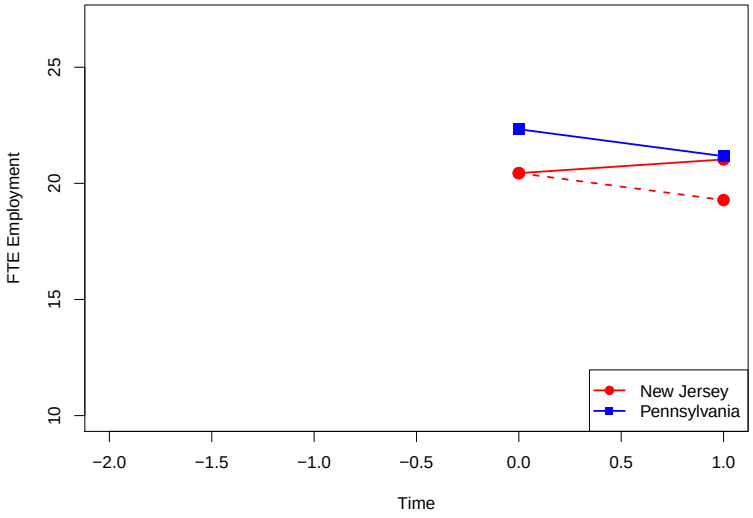
$$\mathbb{E}[Y_{T_0} - Y_{T_0-1}|G = 1] - \mathbb{E}[Y_{T_0} - Y_{T_0-1}|G = 0] = 0$$

   - run placebo DID on data from $t = T_0 - 1, T_0$ and test if $\alpha = 0$. If not, your estimate comparing $t = T_0$ and $t = T_0 + 1$ may be biased
2. Falsification test using data for alternative control group
   - if the DID with the alternative control is different from the DID with the original control, then the original DID may be biased
3. Falsification test using alternative placebo outcome that is not supposed to be affected by the treatment
   - if DID from placebo outcome is non-zero, then the DID estimate for original outcome may be biased
4. Tackling failure of parallel trends through reweighting or modeling

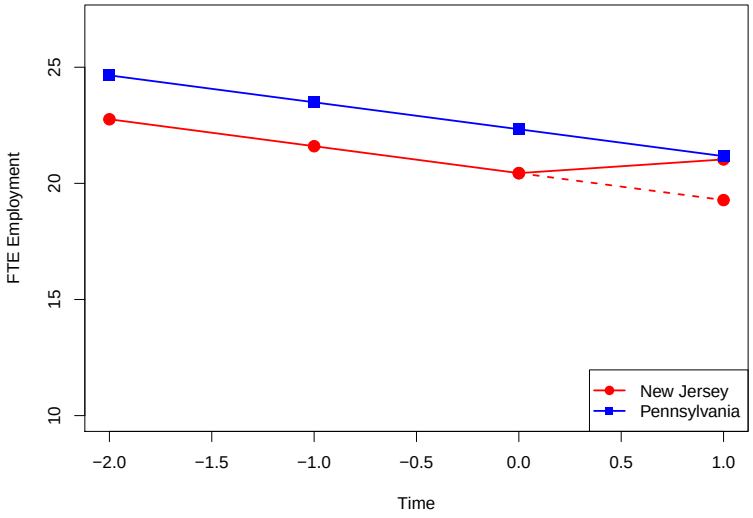
Falsification Test: Data for Prior Periods



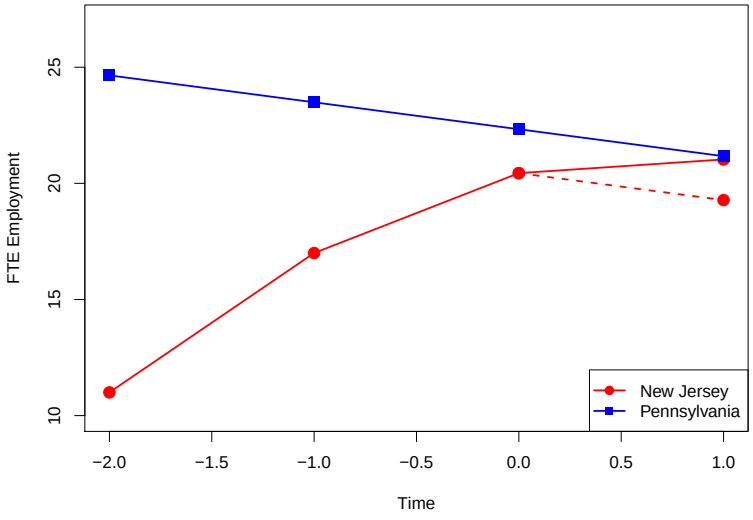
Falsification Test: Data for Prior Periods



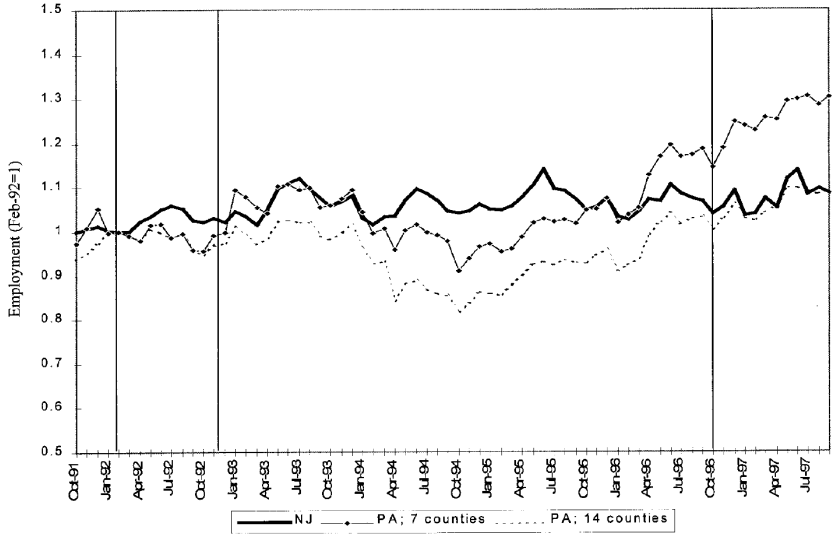
Falsification Test: Data for Prior Periods



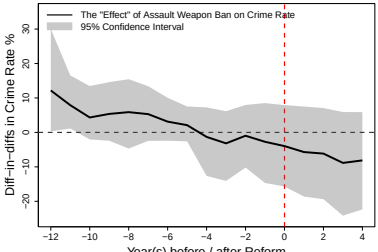
Falsification Test: Data for Prior Periods

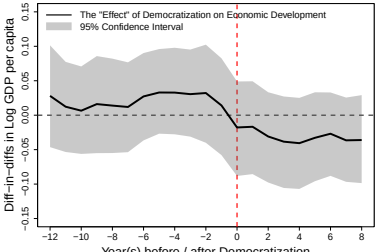


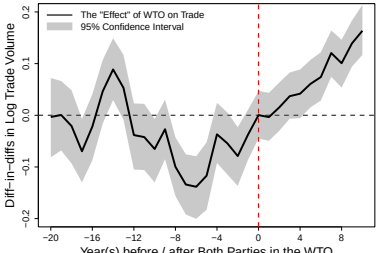
Longer Trends in Employment (Card and Krueger 2000)

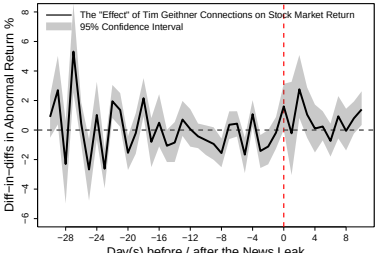


Failure of Parallel Trends









Falsification Test: Alternative Control Group

Variable	Stores by state			Stores in New Jersey[a]			Differences within NJ[b]	
	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

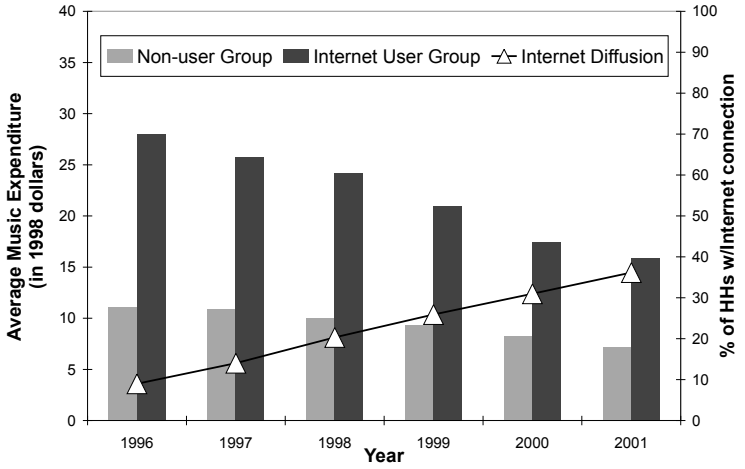
Compositional Differences

- 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]

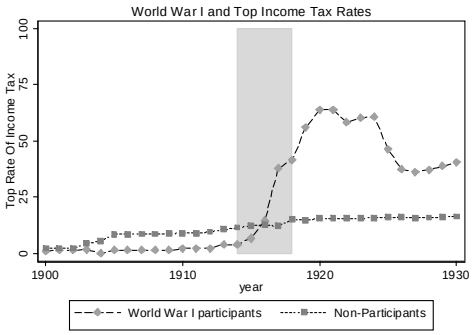
Year	1997		1998		1999	
	Internet User	Non-user	Internet User	Non-user	Internet User	Non-user
Average Expenditure						
Recorded Music	$25.73	$10.90	$24.18	$9.97	$20.92	$9.37
Entertainment	$195.03	$96.71	$193.38	$84.92	$182.42	$80.19
Zero Expenditure						
Recorded Music	.56	.79	.60	.80	.64	.81
Entertainment	.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?

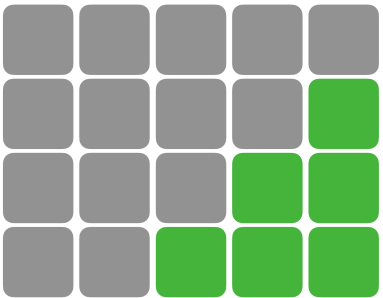
Roth and Sant'Anna (2021)

- 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

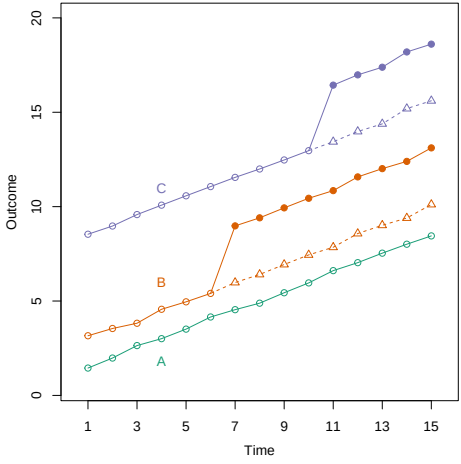
Staggered DID: HTE and the Weighting Problem

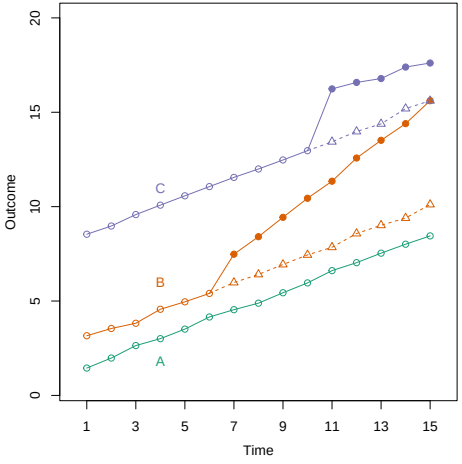


- Question: Can TWFE get at (some non-negatively weighted) ATT when the treatment effects are heterogeneous
- Probably not! e.g. Chernozhukov et al (2017), Borusyak et al (2022), Strezhnev (2018), Callaway & Sant'Anna (2020), de Chaisemartin & dHaultfoeuille (2020) Imai and Kim (2020)
- Early adopters (e.g. D) serves as controls for late adopters (e.g. B)

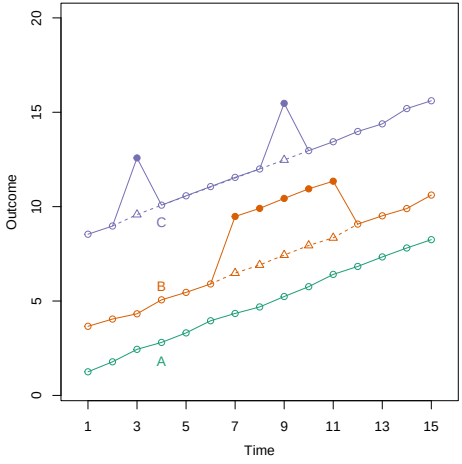
$\Rightarrow$ Some treated observations are being used as controls in some comparisons

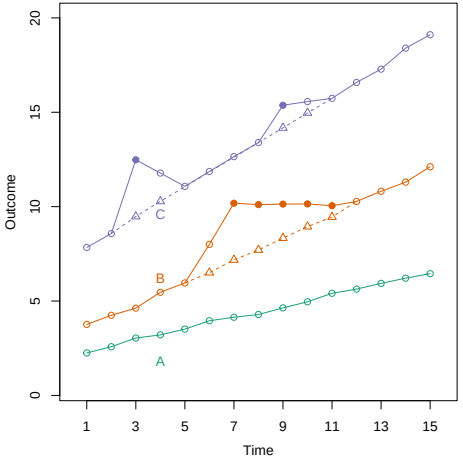
TWFE Assumptions vs. Reality – Staggered Adoption





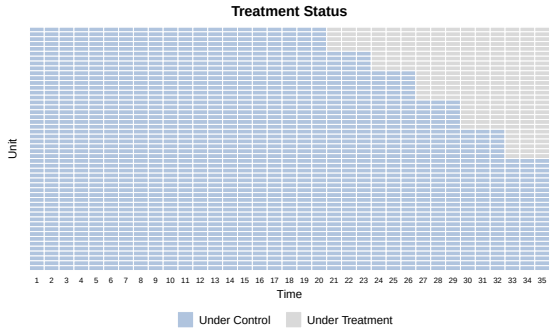
TWFE Assumptions vs. Reality – The General Case





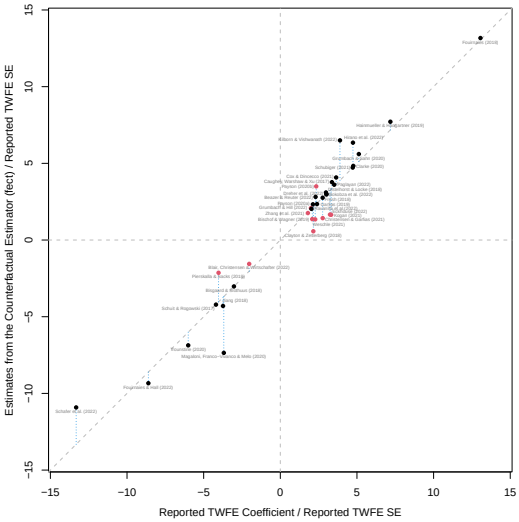
Solutions to the Weighting Problem

- **DID extensions**: Carefully construct estimators using valid 2×2 DID (advanced workshop)
- **Imputation**: Only use untreated data to estimate the model and impute Y_{pre} for the treated observations



How Bad is the Weighting Problem Caused by HTE? (Chiu et al. 2023)

- No too bad!
- We replicated 37 papers published in top political science journals and find:
  - There's almost no sign flipping
  - There's no systematic overestimation or underestimation of the treatment effects
  - Some marginally significant results become insignificant
  - However, parallel trends violations remain a major concern



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
- TWFE models are often used to estimate treatment effects in various DID settings; however, its constant treatment effect assumption induces a weighting problem, which may render the estimated parameters causally uninterpretable
- Solutions to the failure of the parallel trends assumption include reweighting and modeling (next lecture)