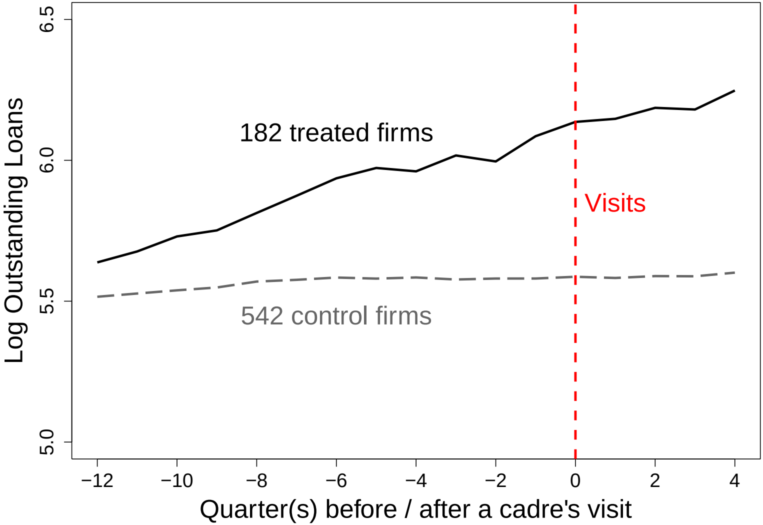
Causal Inference with Panel Data

Yiqing Xu
University of California, San Diego

Northwestern-Duke Causal Inference
Advanced Workshop
June 26, 2018

Cadre Visits on Loans



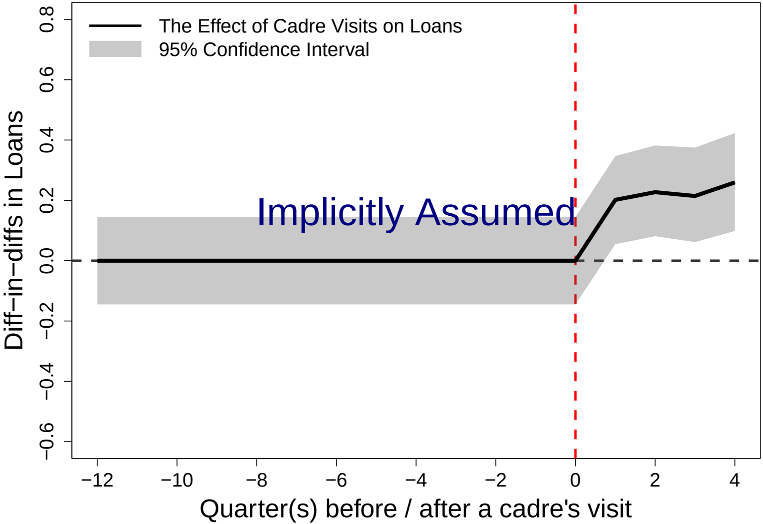
Cadre Visits on Loans

The Effect of Cadre Visits on Loans

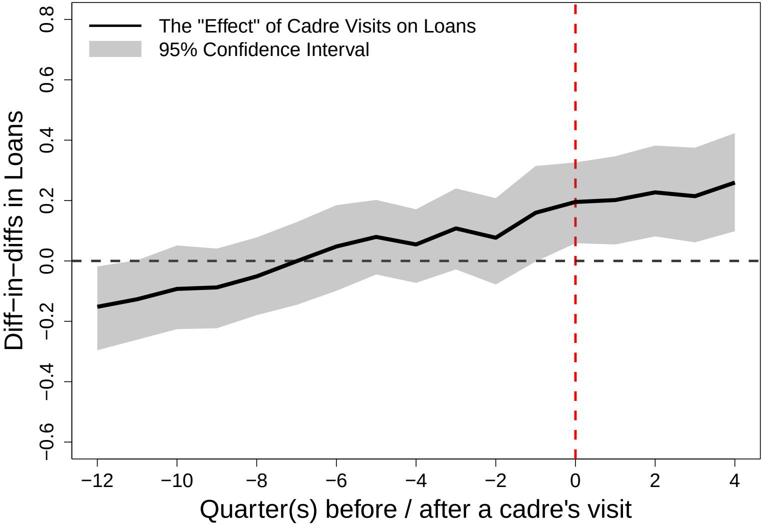
Outcome Variable	Log Outstanding Loans	
	(1)	(2)
Treated Firms × Post-Visit	0.47***	0.53***
	(0.12)	(0.14)
Treated Firms × One Year Before Visit		0.23**
		(0.10)
Quarter Fixed Effects	Yes	Yes
Firm Fixed Effects	Yes	Yes
#Firms	724	724
Treated	182	182
Controls	542	542
Observations	23,168	23,168

Notes: Robust standard errors clustered at the firm level are in the parentheses. *** $p < 0.01$, ** $p < 0.05$.

Cadre Visits on Loans

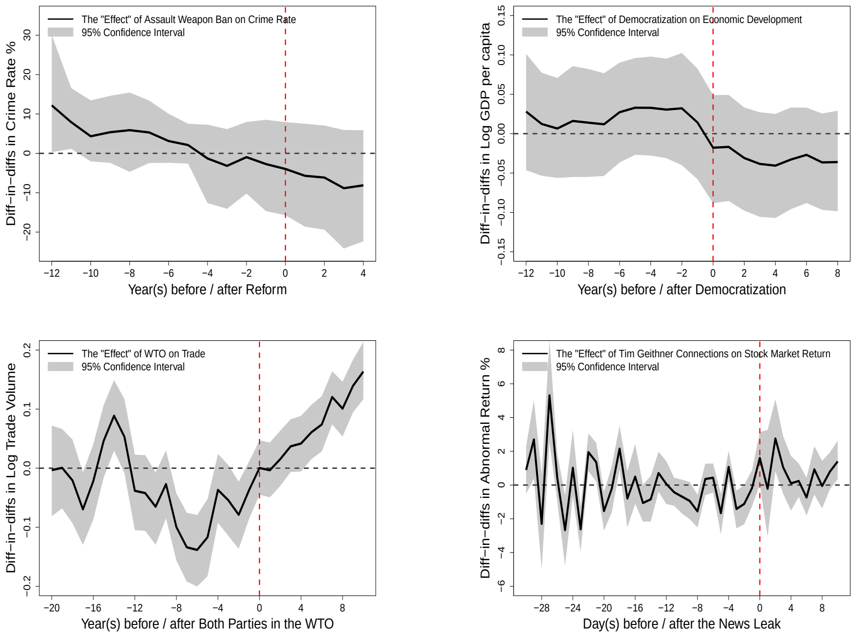


Cadre Visits on Loans



Challenge to the Conventional DiD Approach

- The “parallel trends" assumption often appears to fail
- Equivalently, presence of unobserved time-varying confounders



The "Effect" of Assault Weapon Ban on Crime Rate
95% Confidence Interval
Diff−in−diffs in Crime Rate %
Year(s) before / after Reform
The "Effect" of Democratization on Economic Development
95% Confidence Interval
Diff−in−diffs in Log GDP per capita
Year(s) before / after Democratization
The "Effect" of WTO on Trade
95% Confidence Interval
Diff−in−diffs in Log Trade Volume
Year(s) before / after Both Parties in the WTO
The "Effect" of Tim Geithner Connections on Stock Market Return
95% Confidence Interval
Diff−in−diffs in Abnormal Return %
Day(s) before / after the News Leak

Causal Inference with Panel Data

- **Conventional wisdom**: we can deal with time-invariant confounders, but not time-varying confounders
  - Diff-in-Diffs (`DiD`): difference out time invariant confounder
  - Two-way Fixed Effects: “absorb” time invariant confounders
- Any hope for time-varying confounders?
  $\rightarrow$We explore several possibilities to address this challenge
- We focus on panel data with dichotomous treatments

What's Special about Panel Data?

- The fundamental problem of causal inference

$$\tau_i = Y_{1i} - Y_{i0}$$

- A statistical solution makes use of others' information

$$\text{e.g. } ATE = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$$

- A scientific solution exploits homogeneity or invariance assumptions

e.g. A rock stays a rock.
e.g. The long-run growth rate of the US economy is 2.5%.

- Panel data allow us to construct treated counterfactuals using information from both the past and the others

Causal Inference with Panel Data

- It's all about predicting treated counterfactuals
- "Scientific" solution: modeling (but all models are wrong...)
- Statistical solution: similar to the Selection-on-Observable (SOO) approach, e.g., matching/reweighting
- Panel data make both easier
    - Pre-trends are observable $\rightarrow$ more information for modeling
    - Allowing intercept shift relaxes the conventional ignorability assumption
- And we can do more...

Difference-in-Differences: Setup

- Data structure:
  - Two waves of randomly sampled cross-sectional observations
  - Either a panel or repeated cross sections
- Cross-sectional units: $i \in \{1, ..., N\}$
- Time periods: $t \in \{0 \text{ (pre-treatment)}, 1 \text{ (post-treatment)}\}$
- Group indicator: $G_i = \begin{cases} 1 & \text{(treatment group)} \\ 0 & \text{(control group)} \end{cases}$
- Treatment indicator: $D_{it} \in \{0, 1\}$
- Units in the treatment group receive treatment in $t = 1$

Difference-in-Differences: Setup

	Time Period	
Group	$t = 0$	$t = 1$
$G_i = 1$ (treatment group)	$D_{i0} = 0$ (untreated)	$D_{i1} = 1$ (treated)
$G_i = 0$ (control group)	$D_{i0} = 0$ (untreated)	$D_{i0} = 0$ (untreated)

Difference-in-Differences: Setup

Potential outcomes $Y_{it}(d)$:

- $Y_{it}(0)$: potential outcome for unit i in period t when not treated
- $Y_{it}(1)$: potential outcome for unit i in period t when treated

Causal effect for unit i at time t is

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

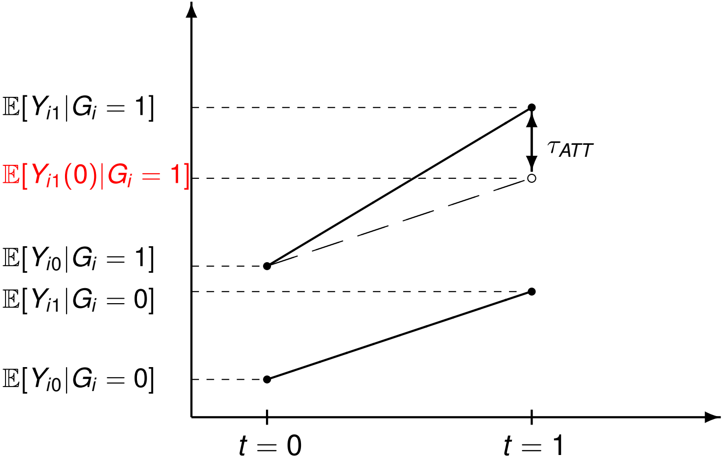
Observed outcomes Y_{it} are realized as

$$Y_{it} = Y_{it}(0)(1 - D_{it}) + Y_{it}(1)D_{it}$$

Because $D_{i1} = G_i$ in the post-treatment period, we can also write

$$Y_{i1} = Y_{i1}(0)(1 - G_i) + Y_{i1}(1)G_i$$

Difference-in-Differences: Graphical Representation



Identification Strategies

Estimand: ATT in the post-treatment period

$$\begin{aligned}\tau_{ATT} &= \mathbb{E}[Y_{i1}(1) - Y_{i1}(0)|G_i = 1] \\ &= \mathbb{E}[Y_{i1}(1)|G_i = 1] - \mathbb{E}[Y_{i1}(0)|G_i = 1]\end{aligned}$$

	Pre-Period ($t = 0$)	Post-Period ($t = 1$)
Treatment Group ($G_i = 1$)	$\mathbb{E}[Y_{i0}(0)\|G_i = 1]$	$\mathbb{E}[Y_{i1}(1)\|G_i = 1]$
Control Group ($G_i = 0$)	$\mathbb{E}[Y_{i0}(0)\|G_i = 0]$	$\mathbb{E}[Y_{i1}(0)\|G_i = 0]$

Problem: Missing potential outcome: $\mathbb{E}[Y_{i1}(0)|G_i = 1]$, i.e. what is the average post-period outcome for the treated group in the absence of the treatment?

Identification Strategies

Estimand: ATT in the post-treatment period

$$\begin{aligned} \tau_{ATT} &= \mathbb{E}[Y_{i1}(1) - Y_{i1}(0)|G_i = 1] \\ &= \mathbb{E}[Y_{i1}(1)|G_i = 1] - \mathbb{E}[Y_{i1}(0)|G_i = 1] \end{aligned}$$

	Pre-Period ($t = 0$)	Post-Period ($t = 1$)
Treatment Group ($G_i = 1$)	$\mathbb{E}[Y_{i0}(0)\|G_i = 1]$	$\mathbb{E}[Y_{i1}(1)\|G_i = 1]$
Control Group ($G_i = 0$)	$\mathbb{E}[Y_{i0}(0)\|G_i = 0]$	$\mathbb{E}[Y_{i1}(0)\|G_i = 0]$

Control Strategy: Before vs. After

- Use $\mathbb{E}[Y_{i1}|G_i = 1] - \mathbb{E}[Y_{i0}|G_i = 1]$ for τ_{ATT}
- Assumes $\mathbb{E}[Y_{i1}(0)|G_i = 1]$ = $\mathbb{E}[Y_{i0}(0)|G_i = 1]$
  (No change in average potential outcome over time)

Identification Strategies

Estimand: ATT in the post-treatment period

$$\begin{aligned}\tau_{ATT} &= \mathbb{E}[Y_{i1}(1) - Y_{i1}(0)|G_i = 1] \\ &= \mathbb{E}[Y_{i1}(1)|G_i = 1] - \mathbb{E}[Y_{i1}(0)|G_i = 1]\end{aligned}$$

	Pre-Period ($t = 0$)	Post-Period ($t = 1$)
Treatment Group ($G_i = 1$)	$\mathbb{E}[Y_{i0}(0)\|G_i = 1]$	$\mathbb{E}[Y_{i1}(1)\|G_i = 1]$
Control Group ($G_i = 0$)	$\mathbb{E}[Y_{i0}(0)\|G_i = 0]$	$\mathbb{E}[Y_{i1}(0)\|G_i = 0]$

Control Strategy: Treated vs. Control in Post-Period

- Use $\mathbb{E}[Y_{i1}|G_i = 1] - \mathbb{E}[Y_{i1}|G_i = 0]$ for τ_{ATT}
- Assumes $\mathbb{E}[Y_{i1}(0)|G_i = 1]$ = $\mathbb{E}[Y_{i1}(0)|G_i = 0]$ (Mean ignorability of treatment assignment)

Identification Strategies

Estimand: ATT in the post-treatment period

$$\begin{aligned} \tau_{ATT} &= \mathbb{E}[Y_{i1}(1) - Y_{i1}(0)|G_i = 1] \\ &= \mathbb{E}[Y_{i1}(1)|G_i = 1] - \mathbb{E}[Y_{i1}(0)|G_i = 1] \end{aligned}$$

	Pre-Period ($t = 0$)	Post-Period ($t = 1$)
Treatment Group ($G_i = 1$)	$\mathbb{E}[Y_{i0}(0)\|G_i = 1]$	$\mathbb{E}[Y_{i1}(1)\|G_i = 1]$
Control Group ($G_i = 0$)	$\mathbb{E}[Y_{i0}(0)\|G_i = 0]$	$\mathbb{E}[Y_{i1}(0)\|G_i = 0]$

Control Strategy: Difference-in-Differences (DiD)

- Use: $\left\{\mathbb{E}[Y_{i1}|G_i = 1] - \mathbb{E}[Y_{i1}|G_i = 0]\right\} - \left\{\mathbb{E}[Y_{i0}|G_i = 1] - \mathbb{E}[Y_{i0}|G_i = 0]\right\}$
- Assumes: $\mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|G_i = 1] = \mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|G_i = 0]$ (Parallel trends)

Identification with Difference-in-Differences

Assumption ("parallel trends")

$$\mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|G_i = 1] = \mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|G_i = 0]$$

The ATT can be nonparametrically identified as:

$$\begin{aligned} \tau_{ATT} &= \Big\{\mathbb{E}[Y_{i1}|G_i = 1] - \mathbb{E}[Y_{i1}|G_i = 0]\Big\} \\ &\quad - \Big\{\mathbb{E}[Y_{i0}|G_i = 1] - \mathbb{E}[Y_{i0}|G_i = 0]\Big\} \end{aligned}$$

Proof:

$$\begin{aligned} & \{\mathbb{E}[Y_{i1}|G_i = 1] - \mathbb{E}[Y_{i1}|G_i = 0]\} - \{\mathbb{E}[Y_{i0}|G_i = 1] - \mathbb{E}[Y_{i0}|G_i = 0]\} \\ = \; & \{\mathbb{E}[Y_{i1}(1)|G_i = 1] - \mathbb{E}[Y_{i1}(0)|G_i = 0]\} - \{\mathbb{E}[Y_{i0}(0)|G_i = 1] - \mathbb{E}[Y_{i0}(0)|G_i = 0]\} \\ = \; & \underbrace{\mathbb{E}[Y_{i1}(1)|G_i = 1] - \mathbb{E}[Y_{i1}(0)|G_i = 1]}_{= \tau_{ATT}} + \mathbb{E}[Y_{i1}(0)|G_i = 1] \\ & \qquad - \mathbb{E}[Y_{i1}(0)|G_i = 0] - \mathbb{E}[Y_{i0}(0)|G_i = 1] + \mathbb{E}[Y_{i0}(0)|G_i = 0] \\ = \; & \tau_{ATT} + \underbrace{\{\mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|G_i = 1] - \mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|G_i = 0]\}}_{= 0 \text{ under parallel trends}} \end{aligned}$$

Notes on the Parallel Trends Assumption

- What type of confounding does DiD make the estimator robust to?
  - Unobserved confounding is time-invariant and additive
  - Violated if there is unobserved time-varying confounding

- Parallel trends may be more plausible with pre-treatment covariates:

$$\mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|G_i = 1, X_i = x] \;=\; \mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|G_i = 0, X_i = x]$$

  This assumes parallel trends within strata

- Under the conditional parallel trends assumption, the ATT is identified as

$$\tau_{ATT} \;=\; \sum_x \Big[\{\mathbb{E}[Y_{i1}|G_i = 1, X_i = x] - \mathbb{E}[Y_{i1}|G_i = 0, X_i = x]\} \\ - \{\mathbb{E}[Y_{i0}|G_i = 1, X_i = x] - \mathbb{E}[Y_{i0}|G_i = 0, X_i = x]\}\Big] \Pr(X_i = x \mid G_i = 1)$$

- Note the parallel trends assumption is not invariant to nonlinear transformation of the outcome scale, e.g. parallel trends in $Y_{it}(d)$ implies non-parallel trends in $\log Y_{it}(d)$ and vice versa

Difference-in-Differences: Baseline Model

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

or

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

- τ_{it} is the treatment effect for unit i at time t
- Y_{it}^0 is a combination of two additive fixed effects and idiosyncratic errors
- $\mathbb{E}[\varepsilon_{it}] = 0$ and $\varepsilon_{it} \perp\!\!\!\perp D_{is}$, for all i, t, s (strict exogeneity)
- $ATT = \mathbb{E}[\tau_{it} | D_{it} = 1]$ can be non-parametrically identified if there are only two periods (or two treatment histories)

Difference-in-Differences: Baseline Model

$$\begin{pmatrix} Y^0_{\mathcal{T},pre} & ?? \\ Y^0_{\mathcal{C},pre} & Y^0_{\mathcal{C},post} \end{pmatrix}$$

- τ_{it} is the treatment effect for unit i at time t
- Y^0_{it} is a combination of two additive fixed effects and idiosyncratic errors
- $\mathbb{E}[\varepsilon_{it}] = 0$ and $\varepsilon_{it} \perp\!\!\!\perp D_{is}$, for all i, t, s (strict exogeneity)
- $ATT = \mathbb{E}[\tau_{it} | D_{it} = 1]$ can be non-parametrically identified if there are only two periods (or two treatment histories)

Difference-in-Differences: An Extended Model

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

or

$$\begin{cases} Y_{it}^0 & = X_{it}'\beta + \alpha_i + \xi_t + \varepsilon_{it} \\ Y_{it}^1 & = Y_{it}^0 + \tau_{it} \end{cases}$$

- $\mathbb{E}[\varepsilon_{it}] = 0$ and $\varepsilon_{it} \perp\!\!\!\perp \{X_{is}, D_{is}\}$, for all i, t, s
- Two-way fixed effect models cannot identify ATT unless $\tau_{it} = \tau, \ \forall i, t$
- Imai and Kim (2018) proposes a matching estimator (`wfe`)
- Liu, Wang & Xu (2018) (re-)introduce a fixed-effect counterfactual (`FEct`) estimator

Two-way Fixed Effects: Identification Assumptions

What are assumptions of two-way fixed effects models (Imai and Kim 2018)?

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

1. No unobserved time-varying confounders
2. Past outcomes do not directly affect current treatment (no feedback)
3. Past treatment do no directly affect current and future outcomes (no carryover effect)
4. On the top of that: constant treatment effect, i.e. linearity

* (2) and (3) are the cost we choose to pay to remove time-invariant confounders; in the rest of the talk, we focus on (1)

A Deeper Question: Hypothetical Experiment?

Where is the hypothetical experiment happening?

- The DiD framework implies that it happens cross-sectionally, i.e. the baseline growth rate is ignorable to treatment assignment
- But how about the fixed-effect setup with more than two periods?
- The strict exogeneity assumption seems to suggest that the experiment is happening both within and across units, but this is not always true with real data
- Clustering and block bootstrap help but are not perfect solutions
- Probably we should model the probability of receiving the treatment for each unit over time
- That's why DiD designs in which a non-reversible treatment kicks in at a given time is a cleaner design

Addressing Time-varying Confounders

$$Y_{it} = \tau_{it} D_{it} + X'_{it}\beta + \alpha_i + \xi_t + \lambda'_i f_t + \varepsilon_{it}$$

or

$$\begin{cases} Y^0_{it} & = X'_{it}\beta + \alpha_i + \xi_t + \lambda'_i f_t + \varepsilon_{it} \\ Y^1_{it} & = Y^0_{it} + \tau_{it} \end{cases}$$

- Suppose there are R time-varying signals f_t out there
- Each unit (e.g. country, participant) picks up a fixed linear combination of these signals based on factor loadings λ_i
- Since these "confounders" are evidenced in the outcomes we measure pre-treatment for both treated and control, we can try to use this information to "balance on" or model out these confounders.

The Semi-parametric Approaches

Interactive Fixed Effects (IFE): (Bai 2009; Xu 2017; Athey et al. 2018)

$$Y_{it}^{0} = \lambda_i' f_t + \xi_t + \varepsilon_{it}, \qquad \text{and} \quad \mathbb{E}[\varepsilon_{it}] = 0,\ \forall i, t$$

- Fit the model — essentially decomposing error into factors and loadings
- Need large T_0 and N_{co}

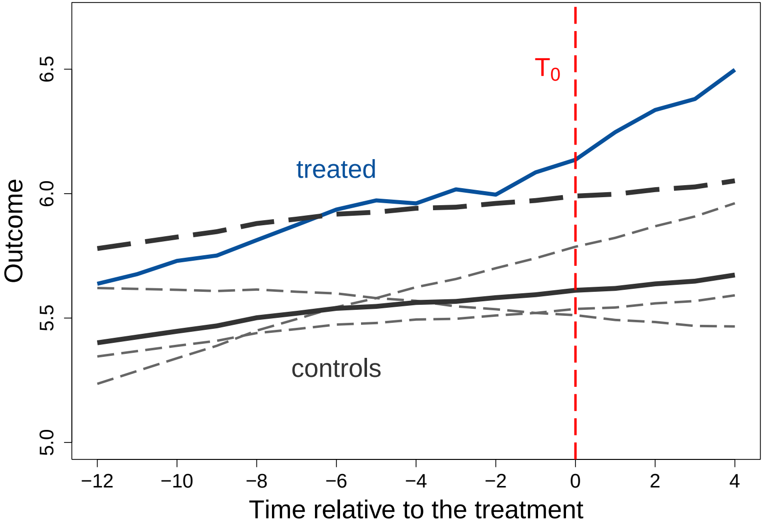
The Non-parametric Approach (Matching/Reweighting)

Synthetic control (`Synth`): (Abadie & Gardeazabal 2003; ADH 2010)

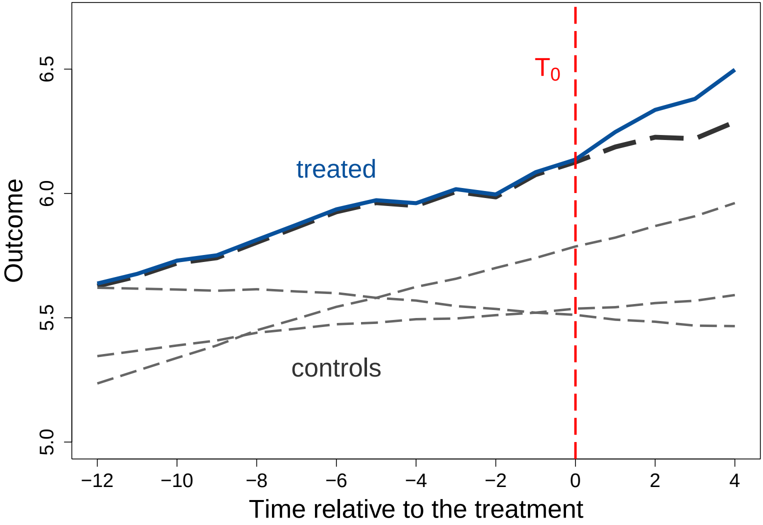
$$Y_{it}^0 = \lambda_i' f_t + \xi_t + \varepsilon_{it}, \qquad \text{and} \quad \mathbb{E}[\varepsilon_{it}] = 0, \ \forall i, t$$

- Chooses weights on control units s.t. weighted average of controls ("synthetic control") looks like the treated unit(s) in the pre-treatment period
- Essentially, make sure $\lambda_i = \sum w_j^* \lambda_j$

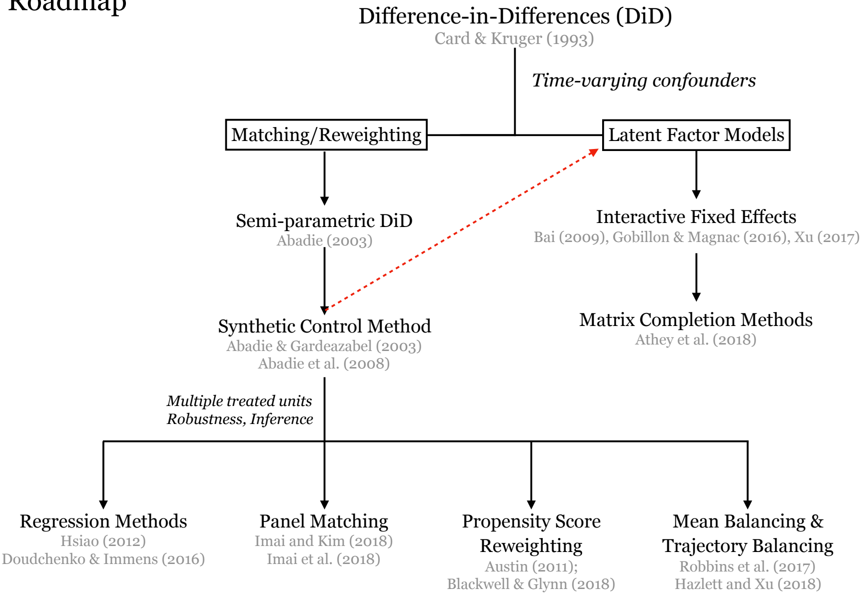
Difference-in-Differences (DiD)



Customized Reweighting



Roadmap


Difference-in-Differences (DiD)
Card & Kruger (1993)
Time-varying confounders
Matching/Reweighting
Latent Factor Models
Semi-parametric DiD
Abadie (2003)
Interactive Fixed Effects
Bai (2009), Gobillon & Magnac (2016), Xu (2017)
Synthetic Control Method
Abadie & Gardeazabel (2003)
Abadie et al. (2008)
Matrix Completion Methods
Athey et al. (2018)
Multiple treated units
Robustness, Inference
Regression Methods
Hsiao (2012)
Doudchenko & Immens (2016)
Panel Matching
Imai and Kim (2018)
Imai et al. (2018)
Propensity Score Reweighting
Austin (2011);
Blackwell & Glynn (2018)
Mean Balancing & Trajectory Balancing
Robbins et al. (2017)
Hazlett and Xu (2018)

Outline

Interactive Fixed Effects

Xu (2017) proposes a two-step approach based on interactive fixed effects (IFE) models :

$$
\begin{array}{llcll}
\text{Control} & Y_{it}(0) & = & X_{it}'\beta + \alpha_i + \xi_t + \lambda_i' f_t + \varepsilon_{it} & \\
\text{Treated} & Y_{jt}(0) & = & X_{jt}'\beta + \alpha_j + \xi_t + \lambda_j' f_t + \varepsilon_{jt} & \text{(pre)} \\
 & Y_{jt}(1) & = & X_{jt}'\beta + \alpha_j + \xi_t + \lambda_j' f_t + \varepsilon_{jt} + \tau_{jt} & \text{(post)}
\end{array}
$$

1. Estimate using controls
2. Predict on treated
3. EM helps

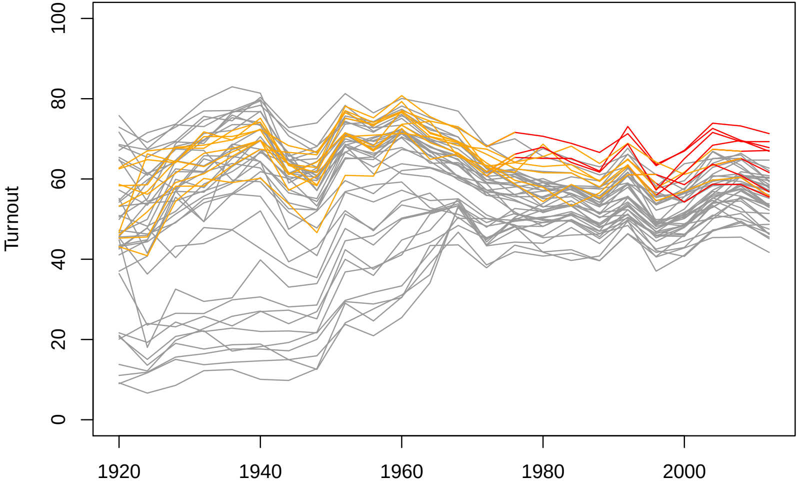
A few technical issues:

- Identification?
- Additive fixed effects?
- How many factors?
- Inference?

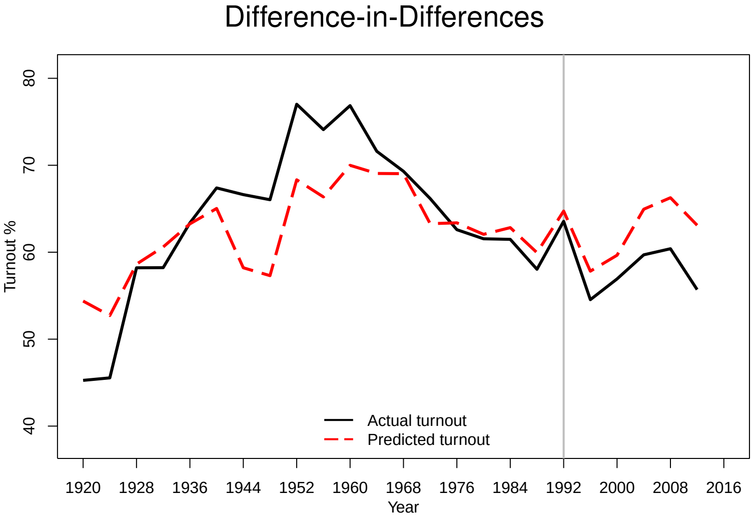
Two Examples

- Election-Day Registration on Voter Turnout
- Cadre Visits on Firms' Access to Loans

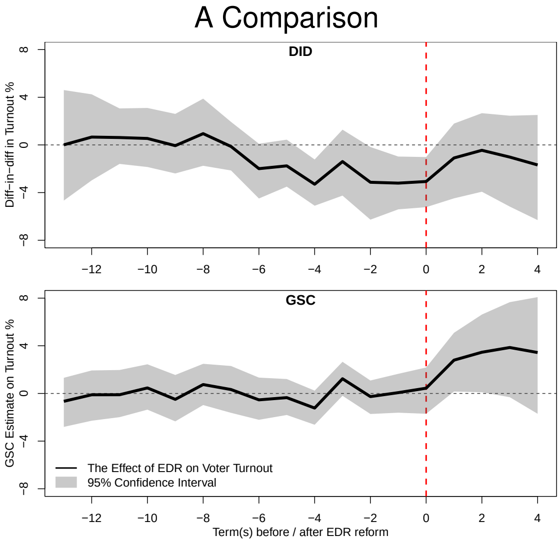
Voter Turnout in US Presidential Elections: 1920-2012



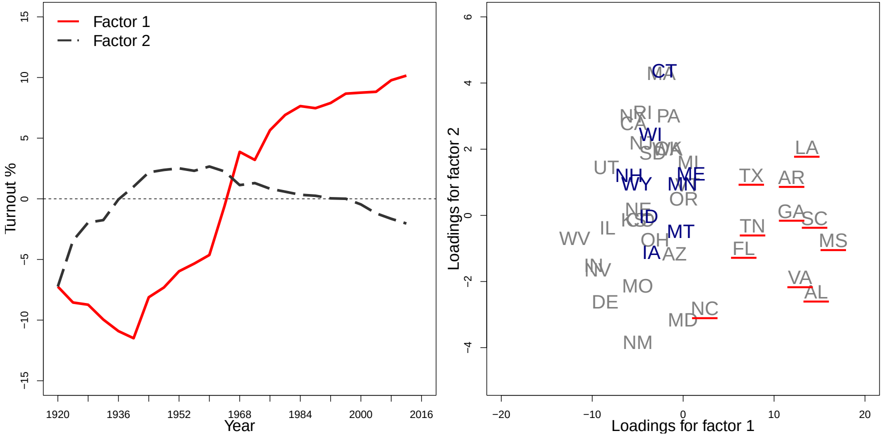
The Case of Connecticut



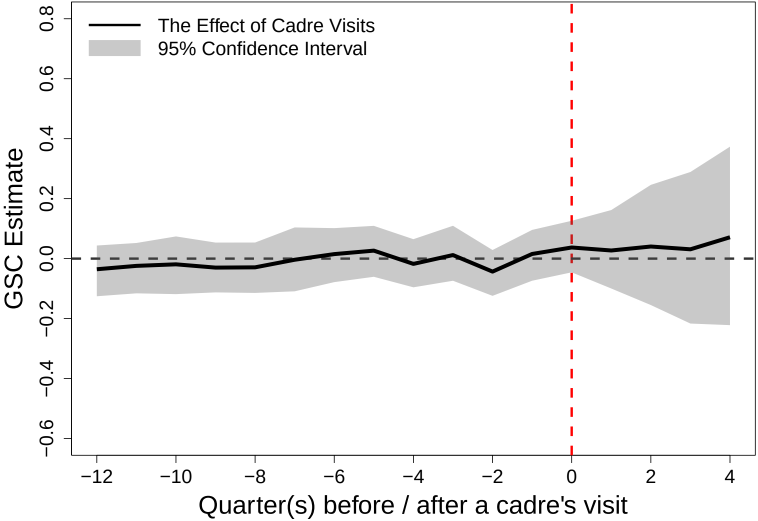
Main Results



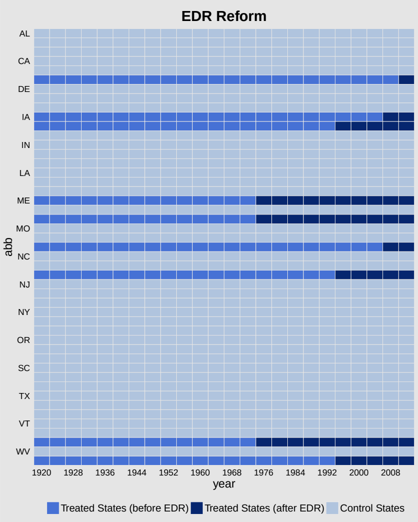
Factors and Factor Loadings



Cadre Visits on Loans



Matrix Completion Methods



- Recall that our main goal is to predict treated counterfactuals
- Taking advantage of the matrix structure, matrix completion methods use non-treated data to achieve this goal
- The basic idea to find a lower-rank representation of the matrix to impute the "missing data"
- Xu (2017) is a special case of this approach

Matrix Completion Methods

- Recall in the baseline DiD setup:

$$\mathbf{Y} = \begin{pmatrix} Y^0_{\mathcal{T},pre} & ?? \\ Y^0_{\mathcal{C},pre} & Y^0_{\mathcal{C},post} \end{pmatrix}$$

- Matrix completion (MC) methods attempt to find a lower-rank representation of **Y**, which we call **L**, that makes predictions of missing values in **Y**
- Athey et al. (2018) generalize Xu (2017) with different ways of constructing **L**
- Plus, missingness can be arbitrary→accommodate reversible treatments

Matrix Completion Methods

- Mathematically,

$$Y_{it} = L_{it} + \alpha_i + \xi_t + X_{it}'\beta + \varepsilon_{it}$$

in which L_{it} is an element of $\mathbf{L}$, an $(N \times T)$ matrix

- We need regularization on $\mathbf{L}$ because of too many parameters:

$$\min_{\mathbf{L}} \frac{1}{\#Controls} \sum_{D_{it}=0} (Y_{it} - L_{it})^2 + \lambda_L \|\mathbf{L}\|_*$$

- The nuclear norm $\|.\|_*$ generally leads to a low-rank solution for $\mathbf{L}$

$$\|\mathbf{L}\|_* = \sum_{i=1}^{\min(N,T)} \sigma_i(\mathbf{L})$$

in which $\sigma_i(\mathbf{L})$ that the singular values of $\mathbf{L}$

IFE vs. MC

- Singular value decomposition of L $\mathbf{L}_{N\times T} = \mathbf{S}_{N\times N}\mathbf{\Sigma}_{N\times T}\mathbf{R}_{T\times T}$
- Difference in how $\mathbf{\Sigma}_{N\times T}$ is regularized

IFE

best subset

$$\begin{pmatrix} \sigma_1 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

MC

nuclear norm

$$\begin{pmatrix} |\sigma_1 - \lambda_L|_+ & 0 & 0 & \cdots & 0 \\ 0 & |\sigma_2 - \lambda_L|_+ & 0 & \cdots & 0 \\ 0 & 0 & |\sigma_3 - \lambda_L|_+ & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ & & & & |\sigma_T - \lambda_L|_+ \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

in which $|a|_+ = \max(a, 0)$

Example: Democracy and Education

- Stasavage (2005) investigates the effect of multiparty competition on education spending in African countries
- Without looking at the data more closely, it is hard to evaluate whether identification assumption is likely to be valid

TABLE 2 Electoral Competition and Overall Government Spending on Education

Spending Measure	% GDP		% Total Govt. Spending	
	OLS	Fixed Effects	OLS	Fixed Effects
	(1)	(2)	(3)	(4)
Multiparty competition	**1.10*** **	**.358** **	**4.41*** **	**3.10*** **
	(0.21)	**(.168)**	**(0.68)**	**(0.92)**
Election year	−.085	.065	−0.50	−0.12
	(.388)	(.206)	(1.44)	(1.12)
Per capita GDP (log)	1.49***	.591***	2.32***	5.65***
	(0.12)	(.214)	(0.64)	(1.17)
Aid (% GDP)	−.0004	−.021**	−.175***	−.067
	(.007)	(.009)	(.037)	(.050)
% population rural	.035***	.012	.170***	.188**
	(.010)	(.015)	(.032)	(.081)
% population under 15	.049	−.272***	−.190*	−.561
	(.039)	(.077)	(.102)	(.418)
Constant	−10.32***	11.84***	−18.8***	−7.31
	(1.84)	(3.70)	(6.73)	(20.2)
N	365	365	365	365
R^2	0.37	0.26	0.13	0.12

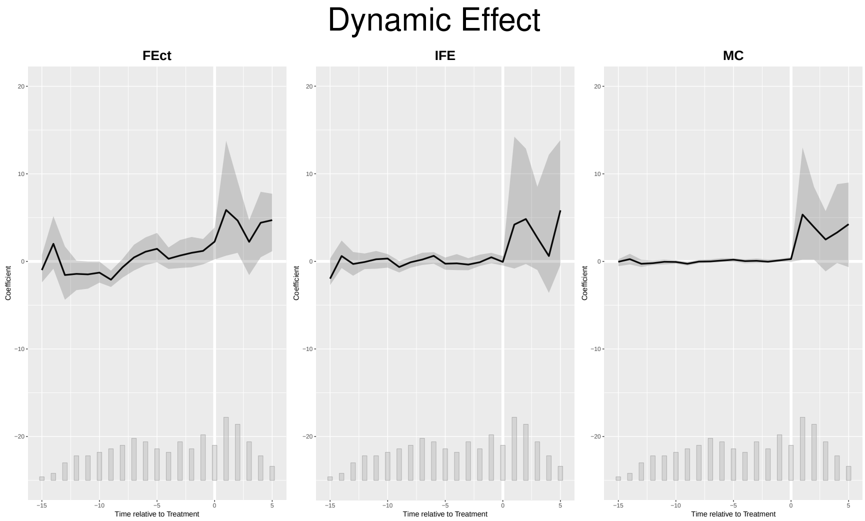
Standard errors in parentheses (panel corrected standard errors for OLS. *, **, and *** refer to significance at the 10%, 5%, and 1% levels, respectively).

Example: Democracy and Education

What Can We Do? Liu, Wang & Xu (2018) suggest:

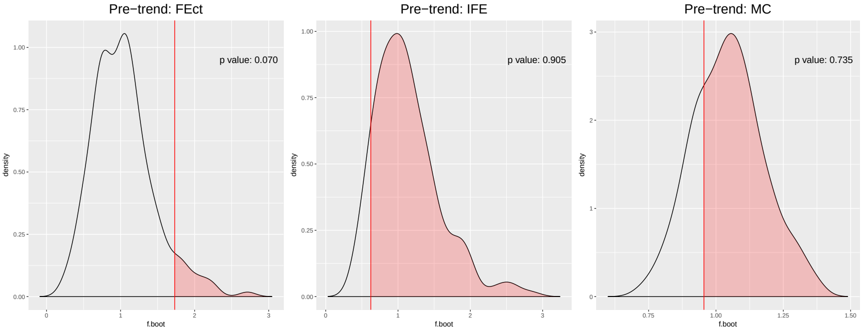
1. Plot the estimated dynamic effect
2. Compare estimates from FEct (DiD), IFE, and MC
3. Test the Null Hypothesis that a “pre-trend” does not exist
4. Conduct a placebo test with different models

FEct (DiD) vs. IFE vs. MC

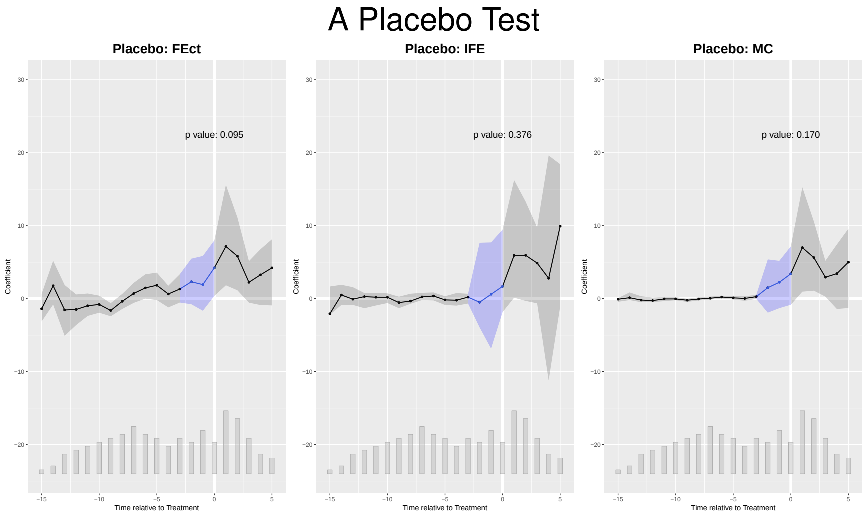


FEct (DiD) vs. IFE vs. MC

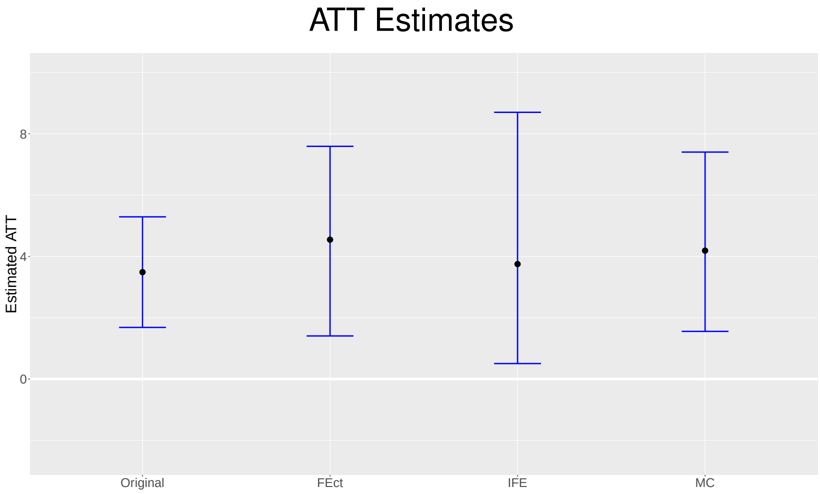
A Ward Test of Pre-trend



FEct (DiD) vs. IFE vs. MC



FEct (DiD) vs. IFE vs. MC



Outline

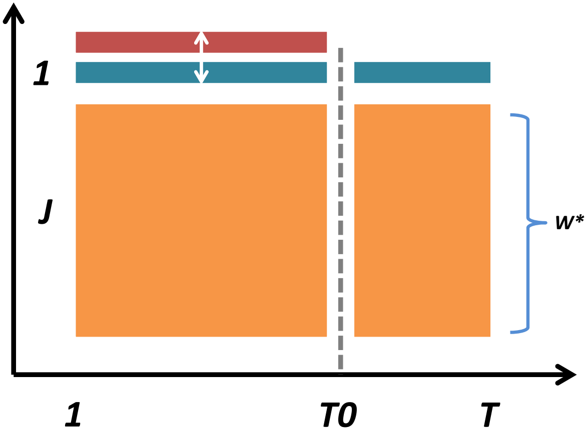
Motivated by a Factor Model

- $J + 1$ units in periods $1, 2, \ldots, T$; one treated "1", J controls.
- Region "1" is exposed to the intervention after period T_0
- The potential outcomes are defined as:

$$\begin{cases} Y_{it}^0 & = f_t'\lambda_i + \theta_t' Z_i + \xi_t + \varepsilon_{it} \\ Y_{it}^1 & = Y_{it}^0 + \tau_{it} \end{cases}$$

- We aim to estimate the effect of the intervention on Region "1": τ_{1t}, $t = T_0 + 1, T_0 + 2, \cdots, T$.

Intuition



Key Idea

$$\begin{cases} Y_{it}^0 & = & f_t'\lambda_i + \theta_t' Z_i + \xi_t + \varepsilon_{it} \\ Y_{it}^1 & = & Y_{it}^0 + \tau_{it} \end{cases}$$

- Let $W = (w_2, \ldots, w_{J+1})'$ with $w_j \geq 0$ and $w_2 + \cdots + w_{J+1} = 1$.
- Let $\bar{Y}_i^{K_1}, \ldots, \bar{Y}_i^{K_M}$ be $M > R$ linear functions of pre-intervention outcomes
- Suppose that we can choose W^* such that:

$$Z_1 = \sum_{j=2}^{J+1} w_j^* Z_j, \quad \bar{Y}_1^k = \sum_{j=2}^{J+1} w_j^* \bar{Y}_j^k,\ k \in \{K_1, \ldots, K_M\}$$

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

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

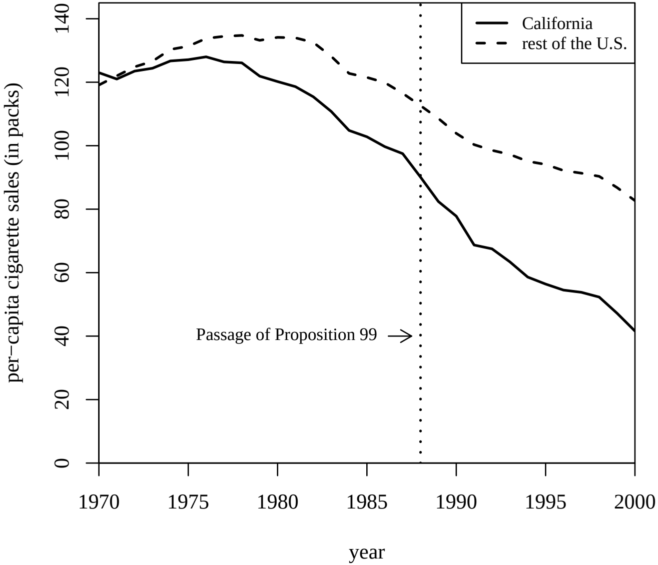
Implementation

- Let $X_1 = (Z_1, \bar{Y}_1^{K_1}, \ldots, \bar{Y}_1^{K_M})'$ be a $(k \times 1)$ vector of pre-intervention characteristics for the treated and X_0, a $(k \times J)$ matrix, for the controls.
- The vector W^* is chosen to minimize $\|X_1 - X_0 W\|$, subject to our weight constraints.
  - We consider $\|X_1 - X_0 W\|_V = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)}$, where V is some $(k \times k)$ symmetric and positive semidefinite matrix.
  - Various ways to choose V (subjective assessment of predictive power of X, regression, minimize MSPE, cross-validation, etc.).

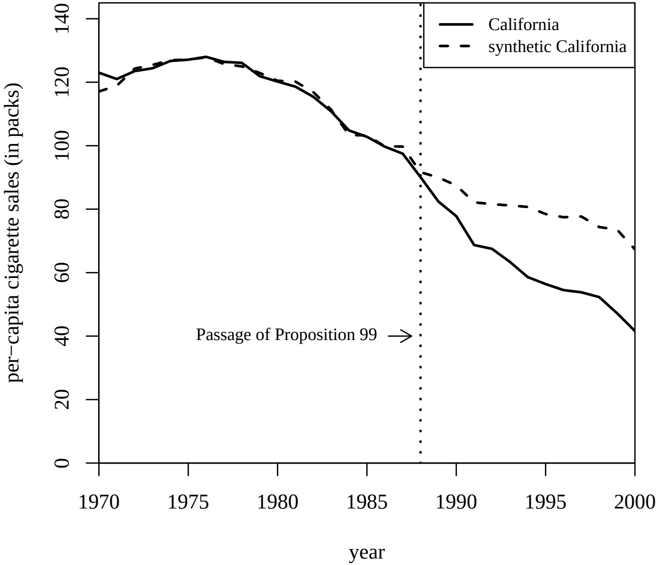
Proposition 99 on Cigarette Consumption

- In 1988, California first passed comprehensive tobacco control legislation (cigarette tax, media campaign etc.)
- Using 38 states that had never passed such programs as controls

Cigarette Consumption: CA and the Rest of the U.S.



Cigarette Consumption: CA and Synthetic CA

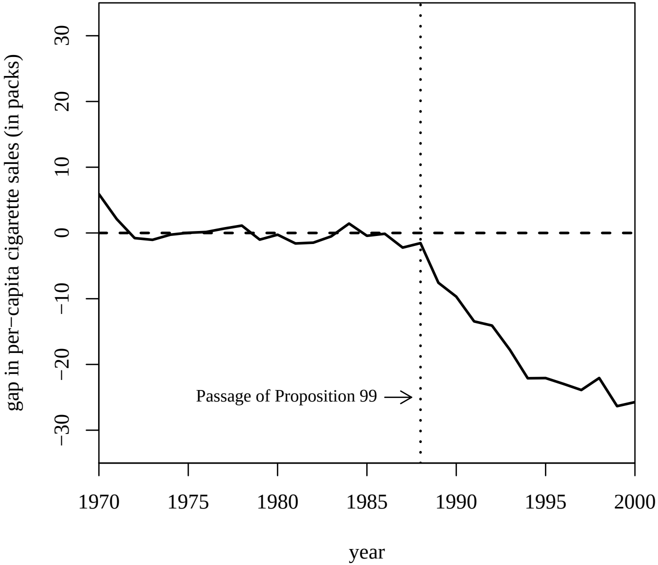


Predictor Means: Actual vs. Synthetic California

	California		Average of
Variables	Real	Synthetic	38 control states
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15-24	17.40	17.40	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.20	23.75
Cigarette sales per capita 1988	90.10	91.62	114.20
Cigarette sales per capita 1980	120.20	120.43	136.58
Cigarette sales per capita 1975	127.10	126.99	132.81

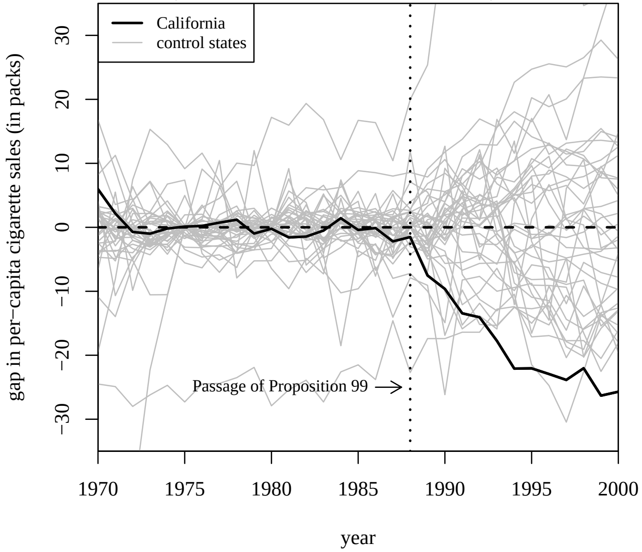
Note: All variables except lagged cigarette sales are averaged for the 1980-1988 period (beer consumption is averaged 1984-1988).

Smoking Gap Between CA and Synthetic CA



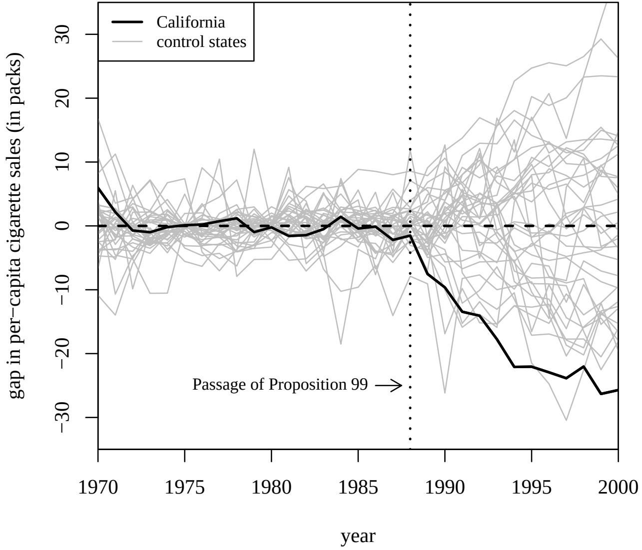
Smoking Gap for CA and 38 Control States

(All States in Donor Pool)



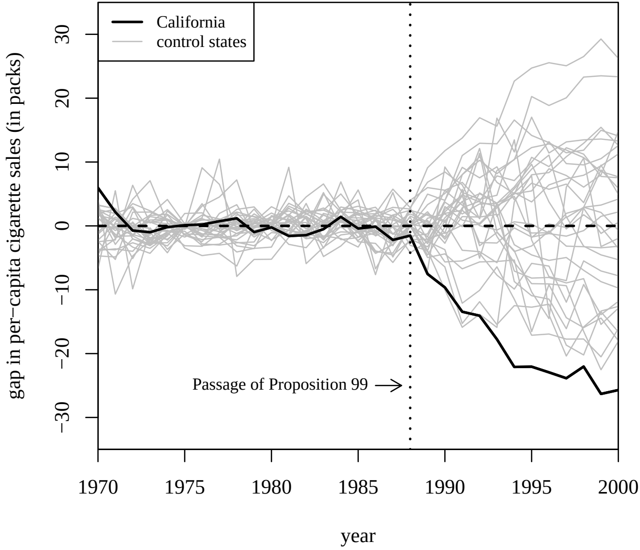
Smoking Gap for CA and 34 Control States

(Pre-Prop. 99 MSPE $\leq$ 20 Times Pre-Prop. 99 MSPE for CA)



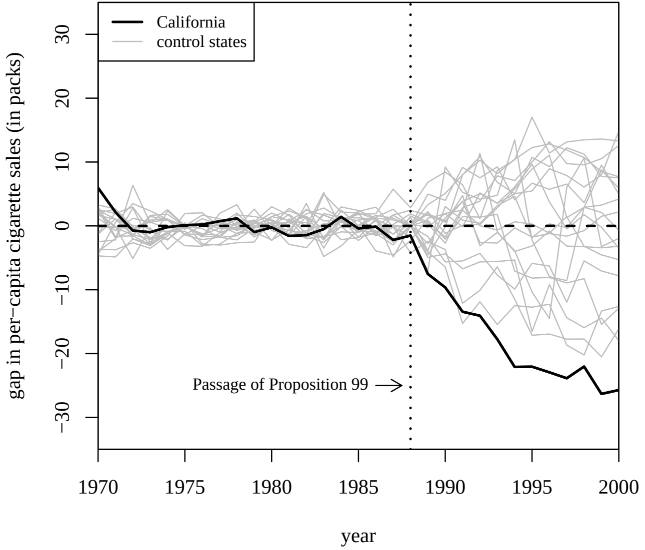
Smoking Gap for CA and 29 Control States

(Pre-Treatment MSPE $\leq$ 5 Times Pre-Treatment MSPE for CA)



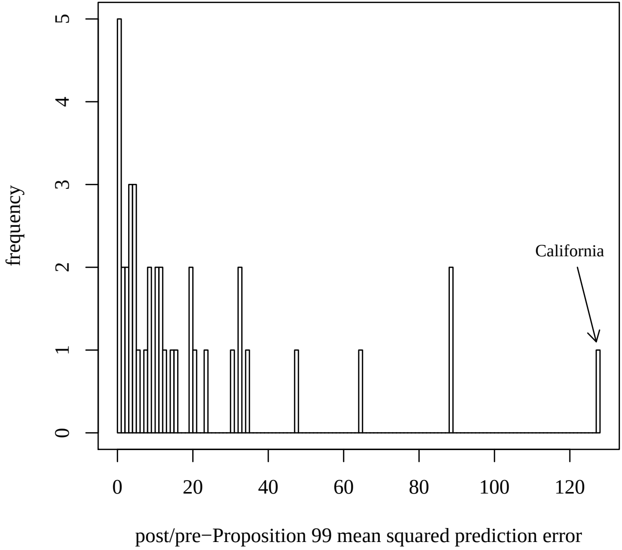
Smoking Gap for CA and 19 Control States

(PRE-TREATMENT MSPE $\leq$ 2 TIMES PRE-TREATMENT MSPE FOR CA)



Ratio Post-Treatment MSPE to Pre-Treatment MSPE

(ALL 38 STATES IN DONOR POOL)



Limitations

- Deal with only one treated unit at a time
- Inference is hard
- Slow to implement
- Sometimes hard to find a solution
- Allow too much user discretion, e.g. cherry-picking $\bar{Y}_i^k$ results in over-rejection (Ferman et al. 2017)

Recent Development

- Multiple treated units, e.g. Acemoglu et al. (2017)
- Permutation inference and sensitivity analysis, e.g. Hahn and Shi (2016); Sergio et al. (2017)
- Other reweighting approaches...
    - Allow w^* to be negative or bigger than 1 and intercept shift, e.g. Hsiao al. (2012)
    - Regularization on w^*, e.g. ridge/lasso Doudchenko and Imbens (2016)
    - Calibration reweighting$\rightarrow$ Reweight to satisfy sample moment conditions (more...)

A Unified "Balancing" Framework

Almost all commonly used panel data models imply common function space with Y^0_{post} linear in Y_{pre}.

Assumption (Linearity in Pre-Treatment Outcome – LPO)

$$\mathbb{E}[Y^0_{it}|D_i = 1, \mathbf{Y}_{i,pre}, X_i] = (1\ \mathbf{Y}^0_{i,pre}\ X_i)'\theta_t, \qquad T_0 < t \leq T.$$

- Diff-in-Diffs, Two-way fixed effects
- Time-series models, e.g. ARMA
- Latent factor models and the synthetic control method

Mean Balancing

Robbins et al. (2017) suggest a balancing approach: mean balancing

- Objective: choose weights on controls to get same average pre-treatment trajectory for weighted controls as treated while maximize entropy of the weights:

$$\min_{\mathbf{w}_c} H(\mathbf{w}_c) = \sum_{j \in \mathcal{C}} w_j log(w_j)$$

$$s.t. \sum_{j \in \mathcal{C}} w_j \mathbf{Y}_{j,pre} = \sum_{i \in \mathcal{T}} \mathbf{Y}_{i,pre} / N_{tr}; \quad \sum_{j \in \mathcal{C}} w_j X_j = \sum_{i \in \mathcal{T}} X_i / N_{tr}$$

$$\sum_{j \in \mathcal{C}} w_j = 1$$

- Challenge: $T0$ may be comparable to or even bigger than N_{co}
- Hazelett and Xu (2018) suggest to seek approximate balance, working from largest toward smallest principal components of $\mathbf{Y}_{pre}$ with a stopping rule based on actual L_1 imbalance.

Mean Balancing

Conceptually,

- Similar to `Synth`: choosing weights to get a good counterfactual
- Easy to see the LPO assumption: groups with equal means on $\mathbf{Y}_{pre}$ have guaranteed equal means on $\mathbf{Y}'_{pre}\theta$ for any θ.

Practical advantages:

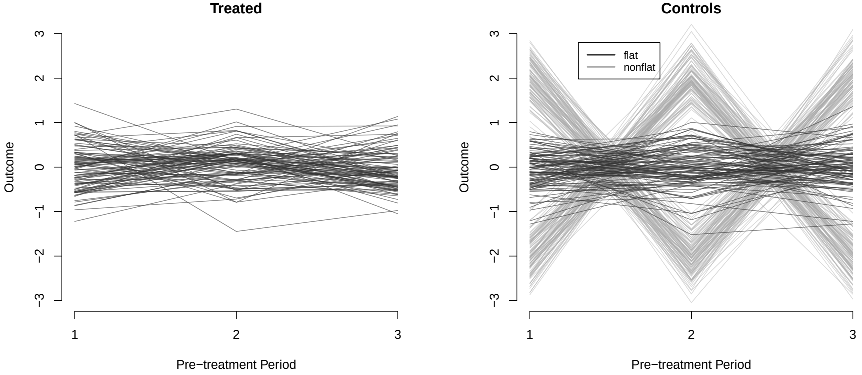
- `Synth` gives users a lot of discretion, assumes one or few treated units, often fails to produce any solution.
- `IFE/MC` requires large T_0 and large N; incurs risks of severe extrapolation
- The mean balancing approach, combined with intercept shift, works in most scenarios (examples will follow)

But Why Stop There?

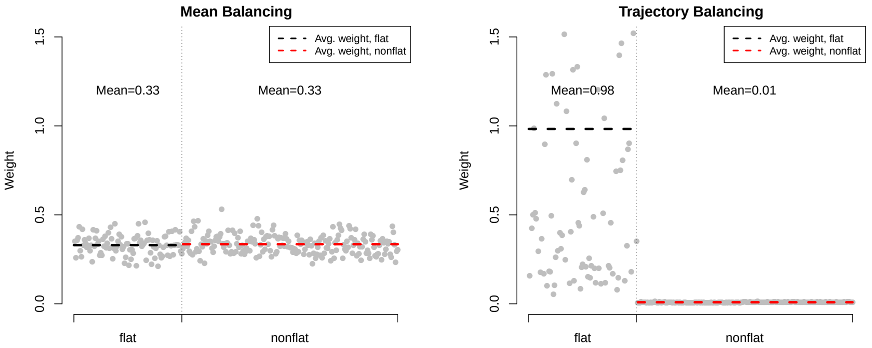
Intuition: Mean balancing limited by the number of pre-treatment points

- Few pre-treatment periods $=$ few constraints unless you make more
- Weights that achieve mean balancing can leave treated and control different on non-linear functions of $\mathbf{Y}_{pre}$
- With enough periods, anything that matters to $Y(0)$ will appear in $Y(0)$ and can be balanced on – but with fewer periods, no guarantees
- Trajectory balancing seeks balance on higher-order terms of $\mathbf{Y}_{pre}$

Example: Similar Average Trajectories $\neq$ Similar Trajectories



Who Should Contribute to Counterfactual?



Trajectory Balancing

- We apply a feature expansion, $\phi(\mathbf{Y}_{i,pre}, X_i)$, $\phi : \mathbb{R}^P \mapsto \mathbb{R}^{P'}$, and get mean balance on this instead.
- A good choice of $\phi()$ is one that:
  - can balance on more than T_0 functions ($P' > P$)
  - requires little or no user guessing
  - includes all continuous functions (at the limit)
  - allows covariates to play a role
  - perhaps, prioritizes low frequency, smoother functions

Trajectory Balancing

Implementation: Gaussian kernel then principal components

- Instead of $\mathbf{Y}_{pre}$, form kernel matrix $\mathbf{K}_{i,j} = k([Y_i, X_i], [Y_j, X_j]) = exp(-||[Y_i, X_i] - [Y_j, X_j]||^2/h)$
- Intuition: replaces each unit's $[Y_{i,pre}, X_i]$ with a vector k_i encoding how similar observation i is to observation 1, 2, ...
- SVD this matrix to obtain components/eigenvectors
- Choose weights to get mean balance on these principal components, starting from the largest
- We use an L_1 measure on remaining imbalance on $\mathbf{K}$ to determine how many components to include.

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: balance on “higher-order” features such as variance, curvature, etc.
- We can show that, to an approximation, trajectory balancing gets multivariates distribution of $\mathbf{Y}_{pre}$ for the controls equal to that of the treated, whereas *mean balancing* only gets margins equal.

A Severe Example

- $N = 150$ countries with simulated GDP over years $T \in \{1, 2, ..., 24\}$
- Two "types" of countries:
  Volatile with no growth:

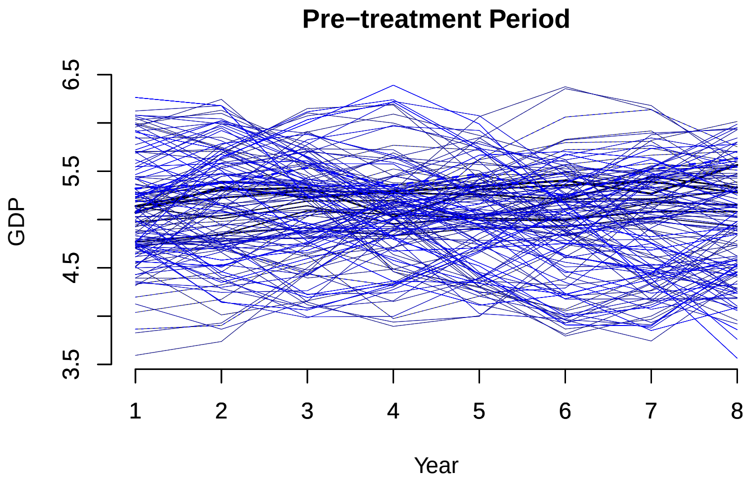
$$GDP_{it} = 5 + a_i sin(.2\pi t) + b_i cos(.2\pi t) + .1\varepsilon_{it}$$
$$\varepsilon_{it} \sim N(0,1), \quad a_i, b_i \sim U(-1,1)$$

  Or steady growing:

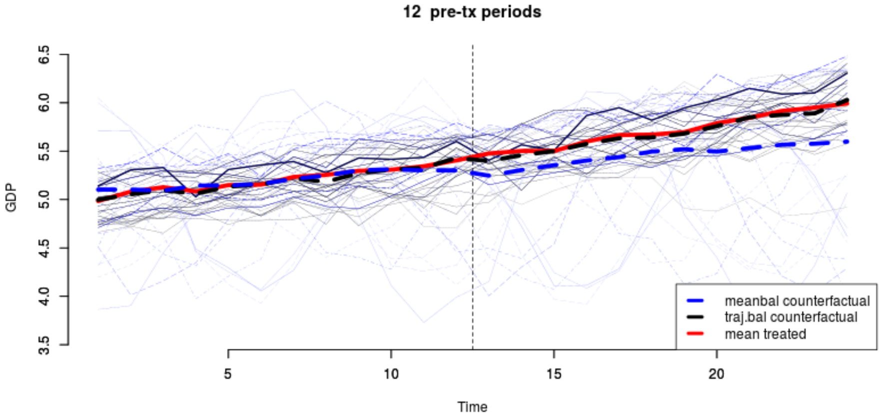
$$GDP_{it} = 4 + c_i 1.03^t + .1\varepsilon_{it}$$
$$\varepsilon_{it} \sim N(0,1), \quad c_i \sim U(0.9, 1.1)$$

- A randomly selected 25% of the stable type take the treatment.

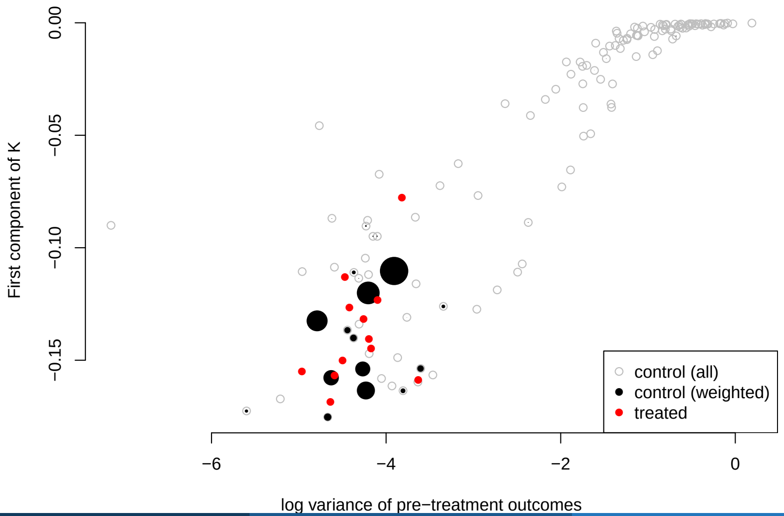
A Severe Example



Results:


12 pre-tx periods
GDP
3.5
4.0
4.5
5.0
5.5
6.0
6.5
5
10
15
20
Time
meanbal counterfactual
traj.bal counterfactual
mean treated

What information is in this kernel matrix?



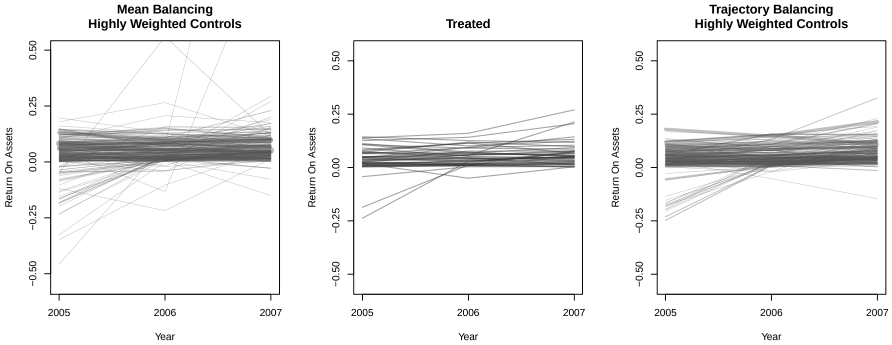
Empirical Examples

1. Truex (2014)
   - Re-weighting using `ebalance`
   - Small T_0, relatively large N
2. ADH (2010)
   - Classic example for `Synth`
   - Large T_0, small N_{co}, $N_{tr} = 1$
3. Xu (2017)
   - Causal inference based on IFE models
   - Modest T_0 and N

Truex (2014): Return to office in National People's Congress

- Treatment: a seat in the Chinese parliament by a firm's CEO
- Outcome: profitability measures, e.g. return on assets (ROA)
- 48 treated firms, 984 controls
- 3 pre-treatment periods (2005-2007)
  3 post-treatment periods (2008-2010)

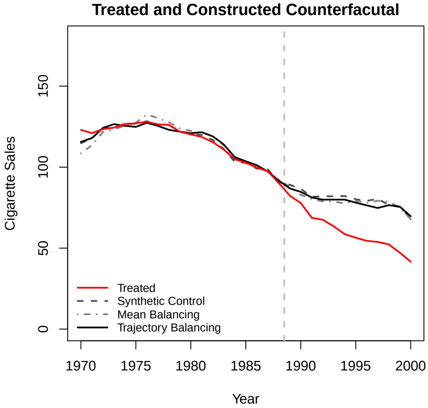
Truex (2014): Return to office in National People's Congress

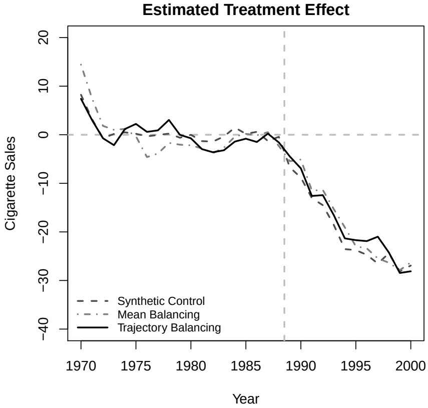


ADH (2010): Effect of Prop 99 on tobacco consumption in California

- Treatment: Proposition 99, a comprehensive tobacco control program implemented in California since 1989
- Outcome: Cigarette sales per person
- 1 treated state, 38 control states
- 19 pre-treatment periods (1970–1988)
  12 post-treatment periods (1989-2000)
- Covariates: {*beer consumption, retail price, age15to24, log income*}
  - included here for replication purposes.
  - we balance on $\phi([\mathbf{Y}_{pre}, X])$
  - with covariates, finds poor balance unless we allow intercept shifts

ADH (2010): Effect of Prop 99 on tobacco consumption in California

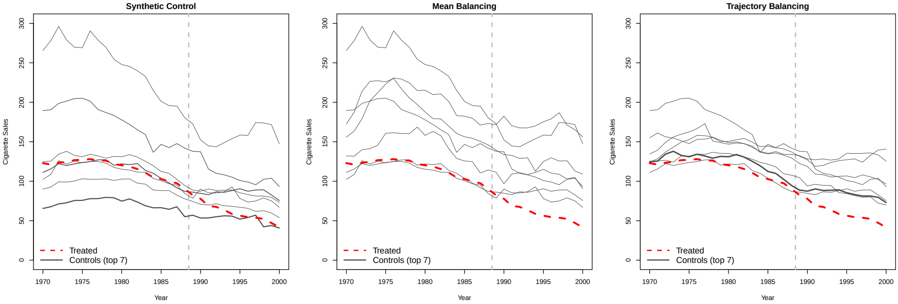




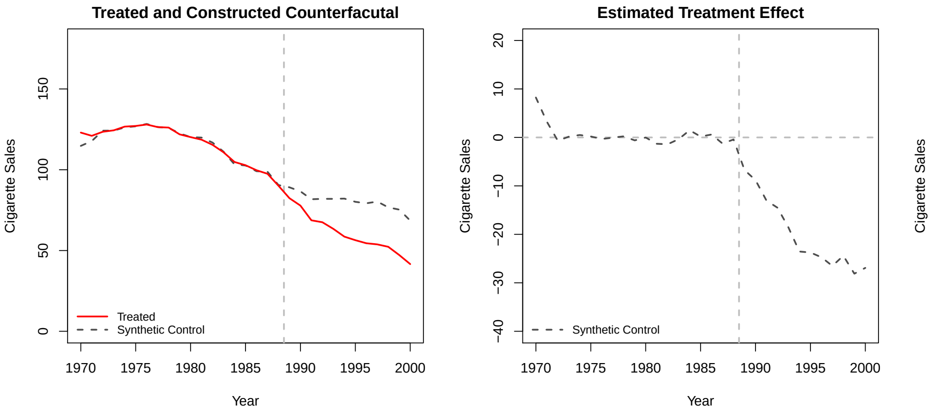
Note: we allow intercept shift with trajectory balancing

ADH (2010): Differences in weights chosen

Gray lines: 7 most heavily weighted controls



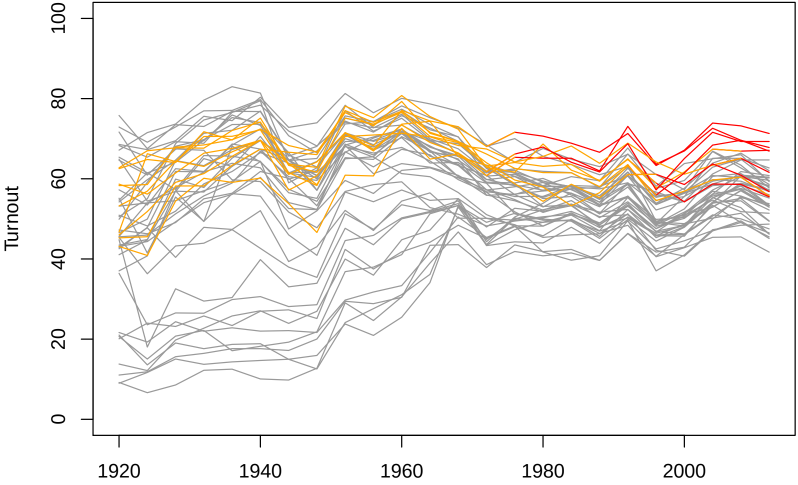
Minimizing specification searches



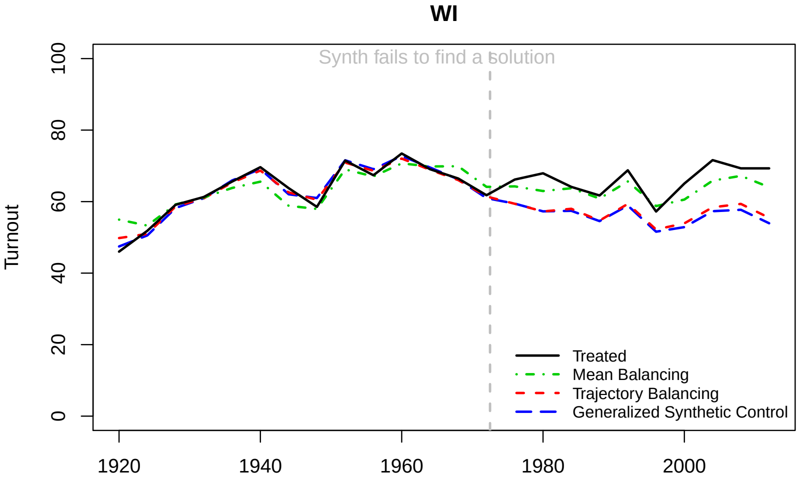
Xu (2017): Election Day Registration on Voter Turnout

- 47 states, 24 election years (1920-2012)
- 9 states started EDR before 2012 (treated); 38 controls
- Based on an IFE model (the `GSC`), Xu (2017) finds EDR increases turnout modestly; the effects are very heterogeneous

Xu (2017): Election Day Registration on Voter Turnout

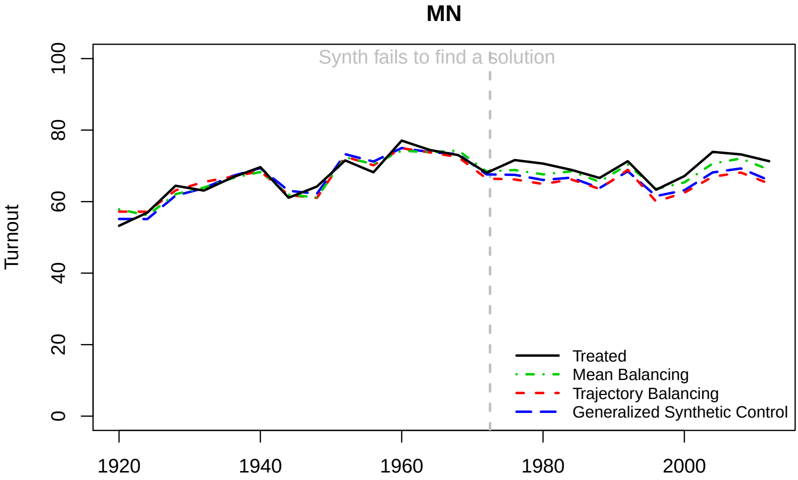


Xu (2017); Election Day Registration on Voter Turnout



* Trajectory balancing mostly agree with the latent factor model.
* In 5 out of 9 cases, `Synth` cannot find a solution.

Xu (2017); Election Day Registration on Voter Turnout



* Trajectory balancing mostly agree with the latent factor model.

* In 5 out of 9 cases, `Synth` cannot find a solution.

In Summary

- Removing time-invariant confounders is costly, i.e., we assume no carryover effect, no feedback from past Y to current D
  $\rightarrow$ There are alternatives, e.g. Blackwell & Glynn (2018)
- The parallel trends assumption can very well be wrong; when T is relative large, we assess the assumption by looking at the "pre-trend"
- Causal inference is a missing data problem
- Fixed-effect counterfactual models relaxes the constant treatment effect assumption
- Both latent factor models and the matching/reweighting approach can help with time-varying confounders with sufficient data
- Inference is hard

Practical Recommendations

- Plot, plot, plot $\rightarrow$ Plots of raw data help us see obvious problems
- Start from fixed effect counterfactual (i.e. DiD) estimators and check the “pre-trend”
- If DiD doesn’t work, try easy fixes, e.g. trimming the data to make the treated and controls more alike
- If that doesn’t work, either, we need more complex models, e.g. trajectory balancing or matrix completion methods

Packages

- `panelView`: panel data visualization
- `fastplm`: fast panel linear fixed effects estimation (coming soon)
- `gsynth`: the IFE/MC approach with non-reversible treatments
- `fect`: general IFE/MC methods with diagnostic tests (coming soon)
- `tjbal`: trajectory balancing

Thank you!

yiqingxu@ucsd.edu
yiqingxu.org
github.com/xuyiqing