MY457/MY557: Causal Inference for Observational and Experimental Studies

Week 11: Difference-in-Differences 2

Daniel de Kadt

Department of Methodology
LSE

Winter Term 2025

Course Outline

- **Week 1:** The potential outcomes framework
- **Week 2:** Randomized experiments
- **Week 3:** Estimation under selection on observables I
- **Week 4:** Estimation under selection on observables II
- **Week 5:** Estimation under selection on observables III
- *Week 6: Reading week*
- **Week 7:** Instrumental variables I
- **Week 8:** Instrumental variables II
- **Week 9:** Regression discontinuity
- **Week 10:** Difference-in-differences I
- **Week 11:** Difference-in-differences II

Difference-in-Differences So Far

Last week we studied the canonical 2-period difference-in-differences (DiD) design, with a brief foray into a special 3-period case.

Identification and estimation was reasonably straightforward:

- Key identification assumption is parallel trends, plus no anticipation
- Use either a plug-in or regression-based estimator

However, we often encounter DiD settings that are more complex:

- More than 2 time periods
- Treatment is assigned variably over time
- Treatment effects are heterogeneous
- Treatment is non-binary

Today we will consider identification and estimation in such settings.

1. Motivating Example
2. Fixed Effects
3. Variable Treatment Timing
4. Multi-Period Designs with Heterogeneous Treatment Effects
5. Synthetic Control Method Primer

1. Motivating Example
2. Fixed Effects
3. Variable Treatment Timing
4. Multi-Period Designs with Heterogeneous Treatment Effects
5. Synthetic Control Method Primer

Minimum Wages and Employment

Do higher minimum wages decrease employment?

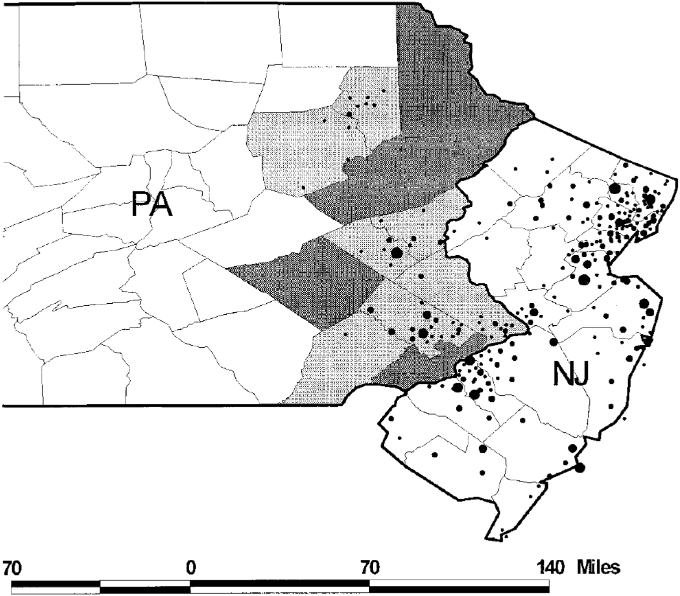
Card and Krueger (1994) consider impact of New Jersey's 1992 minimum wage increase from \$4.25 to \$5.05 per hour.

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 month after increase

Minimum Wages and Employment



Minimum Wages 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)

Treatment: Increase of minimum wage in New Jersey

Difference 1: Pre-minimum wage - Post-minimum wage

Difference 2: New Jersey- Pennsylvania

Minimum Wage and Employment

Card & Krueger (1994) found a positive effect of minimum wage on employment.

Card & Krueger (2000) revisited this design and setting with better data, and found no effect either way.

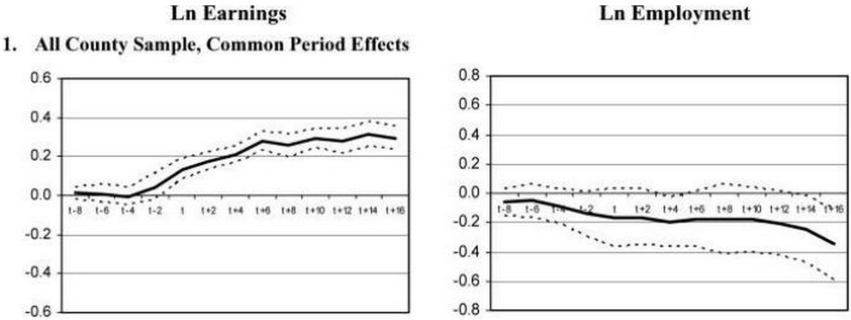
Lots of debate – many papers reconsidered this question using a more general approach: Leveraging cross- and within-state variation throughout the USA. They largely find negative effects on employment.

Dube, Lester, and Reich (2010) revisit this debate:

- Find all cross-state-border changes in MW policies (1990 - 2006)
- Collect earnings and employment data for every county in the USA in this time period.
- Generalize the DiD case study approach.

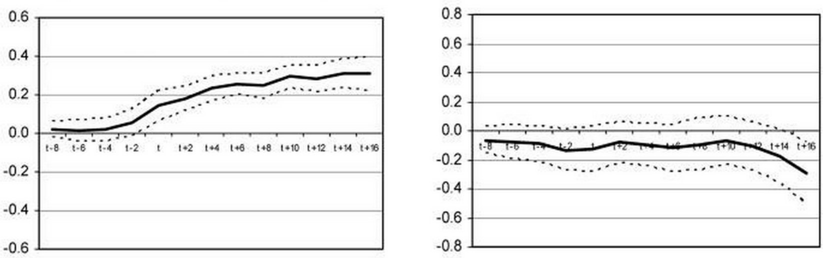
Variation in Space

Estimated Dynamic Effects – Entire Sample



Estimated Dynamic Effects – Border Sample Only

5. Contiguous Border County-Pair Sample, Common Period Effects



1. Motivating Example
2. Fixed Effects
3. Variable Treatment Timing
4. Multi-Period Designs with Heterogeneous Treatment Effects
5. Synthetic Control Method Primer

Fixed Effects Estimation and Difference in Differences

Recall the additive linear model for panel data with 2 periods:

$$Y_{it}(z) \ = \ \alpha_i + \gamma t + \tau z + \varepsilon_{it}$$

where

- $i \in \{0, ..., N\}$: Unit indicator
- $t \in \{0, 1\}$: Time indicator
- $Y_{it}(z)$: potential outcome under treatment status $Z \in \{0, 1\}$
- α_i: time-invariant unobserved effect
- ε_{it}: idiosyncratic error term
- τ: (homogeneous, constant) treatment effect of interest

In a 2-period design, we saw that the first-difference regression:

- Unbiasedly estimates τ under parallel trends and no anticipation assumptions
- τ will coincide with τ_{ATE} and τ_{ATT} if assumed model is correct

Fixed effects estimation generalises this to $t > 2$

Panel Data Notation and Setup

Let's introduce some new-ish notation for panel data:

y_{it}: Vector of observed outcomes for unit i in period t

$\boldsymbol{x}_{it} \equiv [z_{it}, x_{it1}, ..., x_{it(K-1)}]^\top$: Matrix of explanatory variables for unit i in period i

Note: The matrix $\boldsymbol{x}_{it}$ includes both treatment indicator (z_{it}) and other observed covariates (x_{itk}).

In a given panel, we observe a sample of $i = 1, 2, ..., N$ distinct units at $t = 1, 2, ...T$ time periods (a "balanced panel")

Panel Data Notation and Setup

Collect variables for unit $\boldsymbol{i}$:

$$\boldsymbol{y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{it} \\ \vdots \\ y_{iT} \end{pmatrix}_{T\times 1} \qquad \boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{x}_{i1}^\top \\ \vdots \\ \boldsymbol{x}_{it}^\top \\ \vdots \\ \boldsymbol{x}_{iT}^\top \end{pmatrix} = \begin{pmatrix} z_{i1} & x_{i11} & \cdots & x_{i1(K-1)} \\ \vdots & \vdots & & \vdots \\ z_{it} & x_{it1} & \cdots & x_{it(K-1)} \\ \vdots & \vdots & & \vdots \\ z_{iT} & x_{iT1} & \cdots & x_{iT(K-1)} \end{pmatrix}_{T\times K}$$

And stack them for all units (a "long panel"):

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_i \\ \vdots \\ \boldsymbol{y}_N \end{pmatrix}_{NT\times 1} \qquad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_i \\ \vdots \\ \boldsymbol{X}_N \end{pmatrix}_{NT\times K}$$

Pooled OLS Model

Let's start by ignoring the panel structure entirely, and assume:

$$y_{it} = \boldsymbol{x}_{it}^\top \boldsymbol{\tau} + v_{it}, \qquad t = 1, 2, ..., T$$

with composite error $v_{it} \equiv \alpha_i + \varepsilon_{it}$

The estimator $\widehat{\boldsymbol{\tau}_{OLS}}$ will be unbiased and consistent when:

$$E[v_{it}|\boldsymbol{x}_{it}] = 0 \text{ for } t = 1, 2, ..., T$$

That is, when $\boldsymbol{x}_{it}$ is strictly exogenous.

Read: the composite error v_{it} in each time period is uncorrelated with the past, current, and future regressors.

Equivalent to strict conditional ignorability of potential outcomes

Unit Fixed Effects Model

Now consider a "unit fixed effects" assumed model:

$$y_{it} = \boldsymbol{x}_{it}^\top \boldsymbol{\tau} + \alpha_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

We can estimate both $\boldsymbol{\tau}$ and α_i via OLS:

$$(\widehat{\boldsymbol{\tau}_{FE}}, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_N) = \underset{\boldsymbol{\tau}, \alpha_1, \ldots, \alpha_N}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \boldsymbol{x}_{it}^\top \boldsymbol{\tau} - \alpha_i)^2$$

This is called the least squares dummy variables (LSDV) estimator

If the assumed model is correct, then $\widehat{\boldsymbol{\tau}_{FE}}$ is a generalization of the pre-post design we discussed last week.

(Unit) Fixed Effects Estimation

1. We have already seen the LSDV estimator for $\boldsymbol{\tau}$:
   a. Regress $\boldsymbol{y}_{it}$ on $\boldsymbol{x}_{it}$ and unit dummies
   Note 1: Here we directly estimate α_i
   Note 2: With large $\boldsymbol{N}$, this can be computationally expensive

2. Can also be obtained via within estimation:
   a. Create demeaned variables: $\ddot{\boldsymbol{x}}_{it} \equiv \boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i$ and $\ddot{\boldsymbol{y}}_{it} \equiv \boldsymbol{y}_{it} - \bar{\boldsymbol{y}}_i$
   b. Regress $\ddot{\boldsymbol{y}}_{it}$ on $\ddot{\boldsymbol{x}}_{it}$
   Note 1: By within-demeaning we "purge" the fixed effects α_i
   Note 2: The point estimate $\widehat{\boldsymbol{\tau}_{FE}}$ here is exactly equivalent to case (1)

3. Finally, can be obtained via first differences estimation (like last week):
   a. Create differenced variables: $\Delta \boldsymbol{x}_{it} = \boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1}$ and $\Delta \boldsymbol{y}_{it} = \boldsymbol{y}_{it} - \boldsymbol{y}_{i,t-1}$
   b. Regress $\Delta \boldsymbol{y}_{it}$ on $\Delta \boldsymbol{x}_{it}$
   Note 1: First differencing purges the fixed effects α_i
   Note 2: Can be more efficient under serial correlation

All consistent with $\boldsymbol{T}$ fixed and $\boldsymbol{N} \to \infty$ under the same assumptions.

Fixed Effects Estimators: Assumptions and Uncertainty

Assumptions:

1. Strict exogeneity conditional on the unobserved effect
   - $\mathbb{E}[\varepsilon_{it}|\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT}, \alpha_i] = 0, \ t = 1, 2, ..., T$
   - allows $\boldsymbol{x}_{it}$ to be arbitrarily related to α_i
   - SUTVA is implicitly assumed both across units and time periods
2. No carryover effects
   - Treatment status for any Z_{it} does not directly affect outcome $Y_{it'>t}$
3. Rank condition
   - Regressors vary over time for at least some i and are not perfectly collinear

Under these assumptions, $\widehat{\tau_{FE}}$ is unbiased and consistent as $N \to \infty$
(But note that $\hat{\alpha}_i$ via LSDV is *inconsistent* for fixed T and $N \to \infty$)

Uncertainty Estimation:

- Usually SEs should be clustered by unit
- If N is small use block bootstrap

Adding Time Effects

Consider again our assumed "unit fixed effects" model:

$$y_{it} = \boldsymbol{x}_{it}^{\top}\boldsymbol{\tau} + \alpha_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

Typical violation of strict exogeneity assumption: Common shocks that affect all units' $\boldsymbol{y}_{it}$ in the same way and are correlated with $\boldsymbol{x}_{it}$.

Examples include typical history or maturation effects.

More realistic models might include time effects:

- linear time trends
- non-linear time trends
- unit-specific time trends
- time fixed effects

Two-Way Fixed Effects Regression

Let's add time fixed effects to our assumed model:

$$y_{it} = \boldsymbol{x}_{it}^{\top}\boldsymbol{\tau} + \alpha_i + \delta_t + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

where

- α_i represents the unit effect
- δ_t represents common shocks in each time period

This is the two-way fixed effects (TWFE) model.

If our model is correct and $\boldsymbol{x}_{it}$ includes binary $\boldsymbol{z}_{it}$, then the $\widehat{\tau_{TWFE}}$ is generalized difference-in-differences.

Use typical FE estimators (FD, within, LSDV) with both unit and time effects; in `R`:

- `lm` (slow!)
- `plm`
- `fixest`

Dynamic Two-Way Fixed Effects

We can specify a TWFE model allowing dynamic (time-varying) treatment effects:

$$y_{it} = \alpha_i + \delta_t + \sum_{r \neq 0} \mathbf{1}[R_{it} = r]\tau_r + \varepsilon_{it}$$

where

- α_i represents the unit effect
- δ_t represents common shocks in each time period
- R_{it} is the period relative to treatment for unit i
- τ_r is a relative-period treatment effect

Read: This "event study" estimator allows for heterogeneous treatment effects of a specific form: constant τ_r across treatment cohorts.

1 Motivating Example

2 Fixed Effects

3 Variable Treatment Timing

4 Multi-Period Designs with Heterogeneous Treatment Effects

5 Synthetic Control Method Primer

Variable Treatment Timing

Multi-period treatment regimes usually vary over two dimensions:

- Uniform vs. Staggered: Does treatment occur simultaneously, or over time?
- Absorbing vs. Non-absorbing: Once treatment occurs, can it switch off?

With anything other than uniform and absorbing treatment timing, TWFE for DiD may not behave well. For synthesis, see:

- Baker, Larcker, and Wang (2022)
- Roth, Sant'Anna, Bilinski, and Poe (2023)

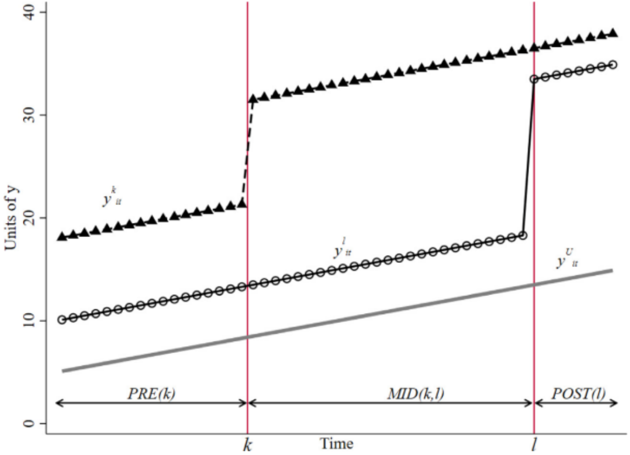
Short of further assumptions, the estimand targeted by TWFE is not easily interpretable $\rightsquigarrow$ it is a weighted average of many different treatment effects.

These weights can be negative (!), are generally non-intuitive, and can potentially severely mislead (e.g. sign-flips).

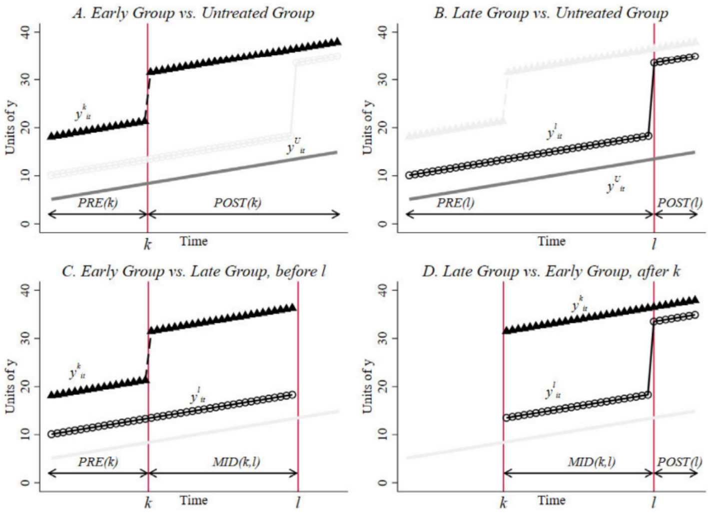
$\widehat{\tau_{TWFE}}$ Decomposition

To see this, we can decompose $\widehat{\tau_{TWFE}}$. We focus on Goodman-Bacon (2021).

Define three groups: never treated (U), early treated (k), and late treated (l)



$\widehat{\tau TWFE}$ Decomposition



$\widehat{\tau TWFE}$ is the weighted average of these four 2x2 treatment effects.

$\widehat{\tau_{TWFE}}$ Decomposition

$$\widehat{\tau_{TWFE}} = \sum_{k \neq U} s_{kU}\, \hat{\tau}_{kU}^{2x2} + \sum_{k \neq U} \sum_{\ell > k} [s_{kl}^{k}\, \hat{\tau}_{k\ell}^{2x2,k} + s_{kl}^{l}\, \hat{\tau}_{k\ell}^{2x2,\ell}]$$

where

- $\hat{\tau}^{2x2}$ are different 2x2 estimators
- s are estimator-specific weights

The weights are a function of three things:

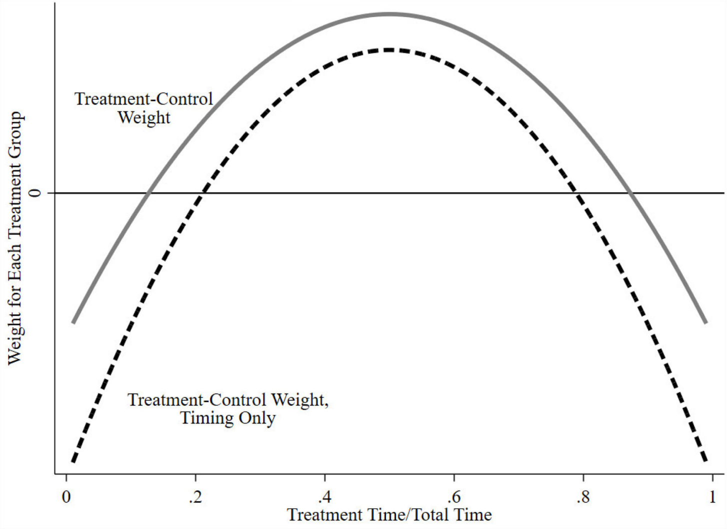
1. 2x-2 subsample size (higher, $\uparrow$ weight)
2. Ratio of treated to control (closer to 0.5, $\uparrow$ weight)
3. Relative timing of treatment (more central, $\uparrow$ weight)

Problem: Some of these comparisons may be "forbidden" such that already-treated units used as controls after they are treated.

Intuition: We subtract off changes in untreated outcomes and treatment effects.

τ_{TWFE} Decomposition

Weighting is heavily dependent on timing:



Key Assumptions for TWFE as Generalised DID

This decomposition reveals the assumptions under which traditional TWFE might be trusted with multi-period panel data and a DiD design.

A causal DiD interpretation with TWFE requires either:

1. Parallel trends, no anticipation, and homogenous τ

   Or:

2. Parallel trends, no anticipation, and uniform timing (constant τ over time)

Quickly explore how much of a problem this may be using `bacondecomp` in `R`.

1 Motivating Example

2 Fixed Effects

3 Variable Treatment Timing

4 Multi-Period Designs with Heterogeneous Treatment Effects

5 Synthetic Control Method Primer

Two Classes of 'Modern' Estimators

Multi-period DiD with non-uniform (staggered or non-staggered) treatment timing should be approached with caution.

Two general types of modern estimators that can help:

1. Flexible matching and re-weighting estimators:
   - Make the 'right' comparisons only, weight appropriately, and recover τ_{ATT}.
   - Many estimators exist: Strezhnev (2018), de Chaisemartin and D'Haultfœuille (2020), Sun and Abraham (2021), Imai and Kim (2021), Dube et al. (2023), de Chaisemartin and D'Haultfœuille (2024)
   - We will focus on Callaway and Sant'Anna (2021)
2. Counterfactual estimators:
   - Estimate only $\boldsymbol{Y_0}$, thus avoiding forbidden comparisons, and recover τ_{ATT}.
   - Many estimators exist: Gobillon and Magnac (2016), Xu (2017), Borusyak et al. (2021), Gardner (2021), Wooldridge (2021)
   - We will focus on Liu, Wang, and Xu (2022)

Callaway and Sant'Anna (2021): Setup

Callaway and Sant'Anna (2021) study a DiD setting with multiple time periods, staggered treatment timing, there may be heterogeneous τ, and parallel trends may hold only conditional on X.

They begin by defining a new estimand, the group-time ATT:

$$\tau_{g,t}^{ATT} = \mathbb{E}[Y_t(1) - Y_t(0)|G_g = 1]$$

where

- there are $T = t \in \{1, ..., T\}$ time periods
- $G_g \in \{0, 1\}$ indicates whether a unit is first treated in period g
- $Y_t(1)$ and $Y_t(0)$ are potential outcomes under treatment and control, at $T = t$

Intuitively, this has already brought us a long way. We can now reason in abstraction about every 2x2 comparison in our data.

Callaway and Sant'Anna (2021): Identification

If parallel trends holds, this group-time ATT can be identified as:

$$\tau_{g,t}^{ATT} = \mathbb{E}[\underbrace{Y_t - Y_{g-1}}_{\text{Long difference}} | G_g = 1] - \mathbb{E}[\underbrace{Y_t - Y_{g-1}}_{\text{Long difference}} | C = 1]$$

where $C \in \{0, 1\}$, taking 1 if never treated (no forbidden comparisons!)

Read: For any treated cohort, at any time period, conduct a diff-in-diff where:

- Difference 1: $T = t$ vs. $T = g - 1$ (long difference)
- Difference 2: G_g (cohort) vs. C (never treated)

Long difference in Y: change in Y between any time period t and the last pre-treatment period, $g - 1$.

Callaway and Sant'Anna (2021): Identification

If conditional parallel trends holds:

$$\tau_{g,t}^{ATT} = \mathbb{E}\left[\underbrace{\left(\frac{G_g}{\mathbb{E}[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{\mathbb{E}\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)}_{\text{Weights}}\underbrace{\left(Y_t - Y_{g-1}\right)}_{\text{Long difference}}\right]$$

$p_g(X) = P(G_g = 1|X, G_g + C = 1)$ is a propensity score

Read: Up-weight control units where $p_g(X)$ is similar to the group-specific treated units

Callaway and Sant'Anna (2021): Estimation

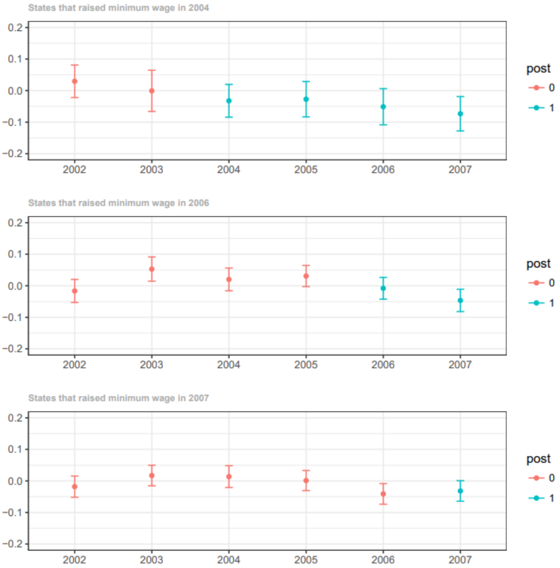
Estimation proceeds as follows:

1. Estimate $\hat{p_g}$ for each group g
2. Estimate $\tau_{g,t}^{\hat{ATT}}$ by plugging in fitted values and observed Y into the (estimator-version) of the expression on the previous slide
3. Combine the estimated values of $\tau_{g,t}^{\widehat{ATT}}$ to retrieve quantities of interest

Some quantities of interest:

- Simple average of $\tau_{g,t}^{\widehat{ATT}}$ across t and g
- Weighted average of $\tau_{g,t}^{\widehat{ATT}}$ weighting by group sizes
- Any other principled summary measure!

Callaway and Sant'Anna (2021) on the Minimum Wage



Liu, Wang, and Xu (2022): Setup

Liu, Wang, and Xu (2022) consider DiD settings with multiple time periods, staggered treatment timing that may or may not be absorbing, and there may be heterogeneous τ.

Define the estimand of interest as:

$$\tau_{ATT} = \mathbb{E}[Y_{it}(1) - Y_{it}(0) \mid Z_{it} = 1, C_i = 1]$$

where

- Z_{it} is our DiD treatment indicator
- C_i is an indicator for 'ever treated' units
- $Y_{it}(1)$ and $Y_{it}(0)$ are potential outcomes under treatment and control

Idea: Estimate only $Y_{it}(0)$ using pre-treatment data, taking $Y_{it}(1)$ as missing.

Estimate τ_{ATT} by taking differences between $Y(1)$ and $Y(\hat{0})$.

Note: This is a philosophical departure from TWFE!

Liu, Wang, and Xu (2022): Estimation

Authors offer three estimators:

- FEct Estimator:

$$Y_{it}(0) = \boldsymbol{x}_{it}^{\top}\boldsymbol{\tau} + \alpha_i + t_t + \varepsilon_{it}$$

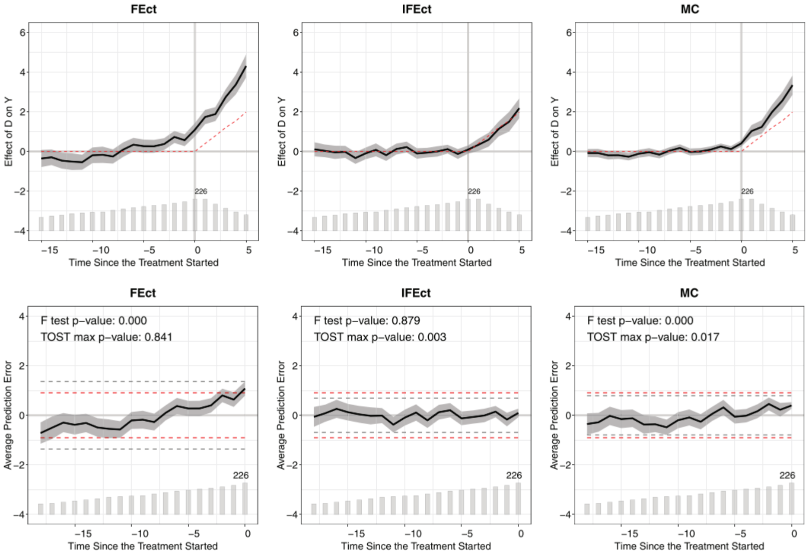
- IFEct Estimator:

$$Y_{it}(0) = \boldsymbol{x}_{it}^{\top}\boldsymbol{\tau} + \alpha_i + t_t + \lambda_i' f_t + \varepsilon_{it}$$

- MC Estimator:

$$\mathbf{Y}(\mathbf{0}) = \boldsymbol{x}_{it}^{\top}\boldsymbol{\tau} + \mathbf{L} + \boldsymbol{\varepsilon}$$

Liu, Wang, and Xu (2022): Simulated Example



Some Practical Advice

For computation, most of these routines are well packaged in most software:

1. For Callaway & Sant'Anna package `{did}` in R.
2. For Liu et al, package `{fect}` in R.
3. See Asjad Naqvi's summary: `https://asjadnaqvi.github.io/DiD/`

Be aware, however:

- TWFE is often okay! Try a number of estimators, report honestly.
- "Modern" estimators have low power (see Chiu et al, 2025, and Weiss, 2025)
- Low power may mean quite variable point estimates.
- Focus on research design and falsification first!

Design precedes estimation!

1. Motivating Example
2. Fixed Effects
3. Variable Treatment Timing
4. Multi-Period Designs with Heterogeneous Treatment Effects
5. Synthetic Control Method Primer

Synthetic Control Method: Primer

DiD requires parallel trends in the expected value of potential outcomes. Generally cannot help where there are time- and unit-varying confounders.

Synthetic Control Methods (SCM) take a different approach:

1. Find $\mathbf{W}^*$, an $N-1$ length vector of unit weights that minimizes $||\mathbf{X_1} - \mathbf{X_0}w||$ for $\mathbf{X_1}$ a matrix of pre-treatment outcomes and covariates for the treated unit, and $\mathbf{X_0}$ likewise for the control.

   Read: Find the weighted combination of control units that best matches – in both levels and trend – the treated unit in the pre-treatment period.

   This weighted set of control units is the synethic control.

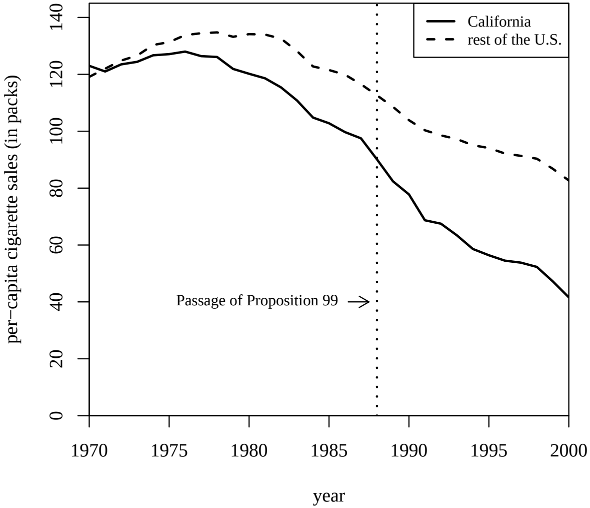
2. An approximately unbiased estimator of the unit-specific effect in the post-treatment period is:

$$\widehat{\tau}_{1t} \ = \ Y_{1t} - \textstyle\sum_{j=2}^{J+1} w_j^* Y_{jt} \quad \text{for} \quad t \in \{T_0 + 1, \dots, T\}$$

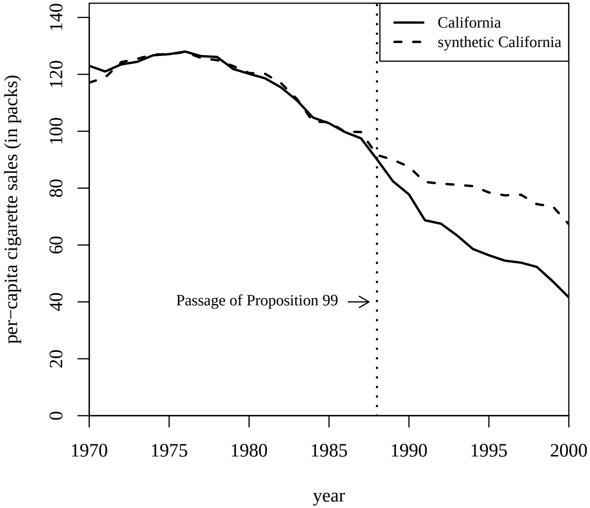
3. Inference uses placebos (roughly, randomization inference).

For primer on use, see Abadie (2020).

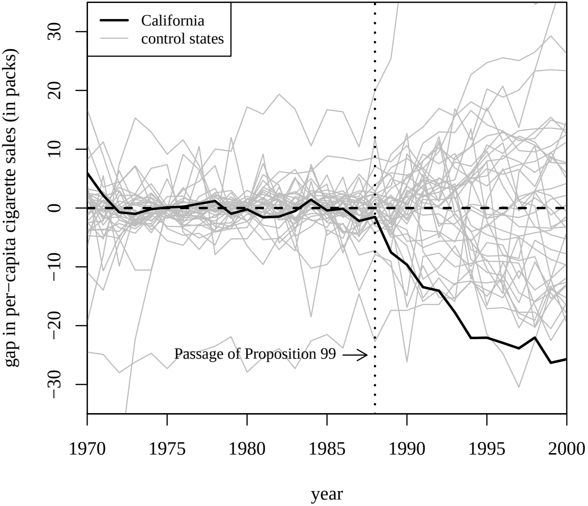
SCM Example: California Prop 99 (Abadie et al. 2010)



SCM Example: California Prop 99 (Abadie et al. 2010)



SCM Example: California Prop 99 (Abadie et al. 2010)



"Be very very quiet, we're hunting identifying variation in D"