Causal inference
Difference-in-differences, Synthetic controls

Daniel Stegmueller

Duke University

Difference-in-differences II

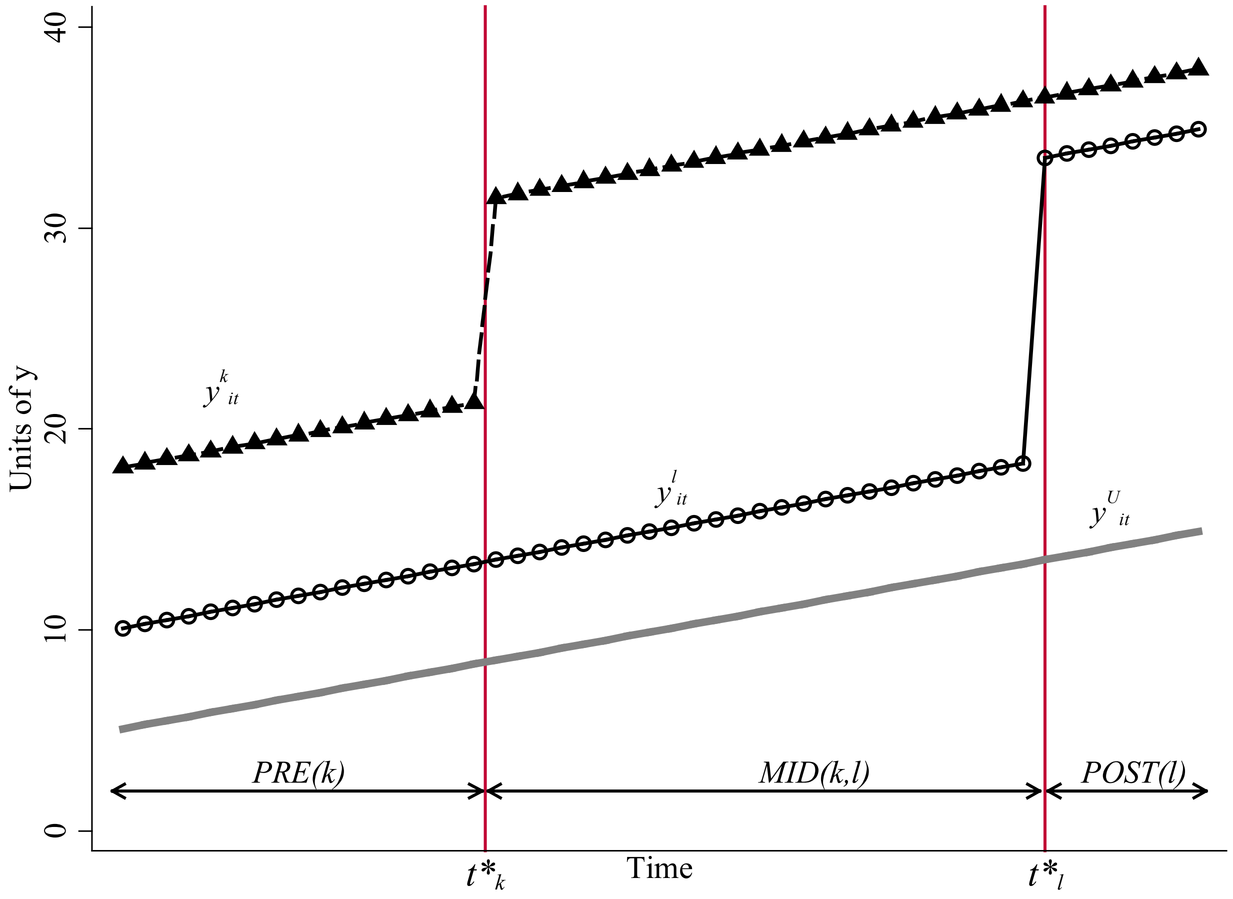
Staggered treatments

- So far we have discussed simple settings where some groups are treated at a point in time and the onset of the treatment is identical for all treated units

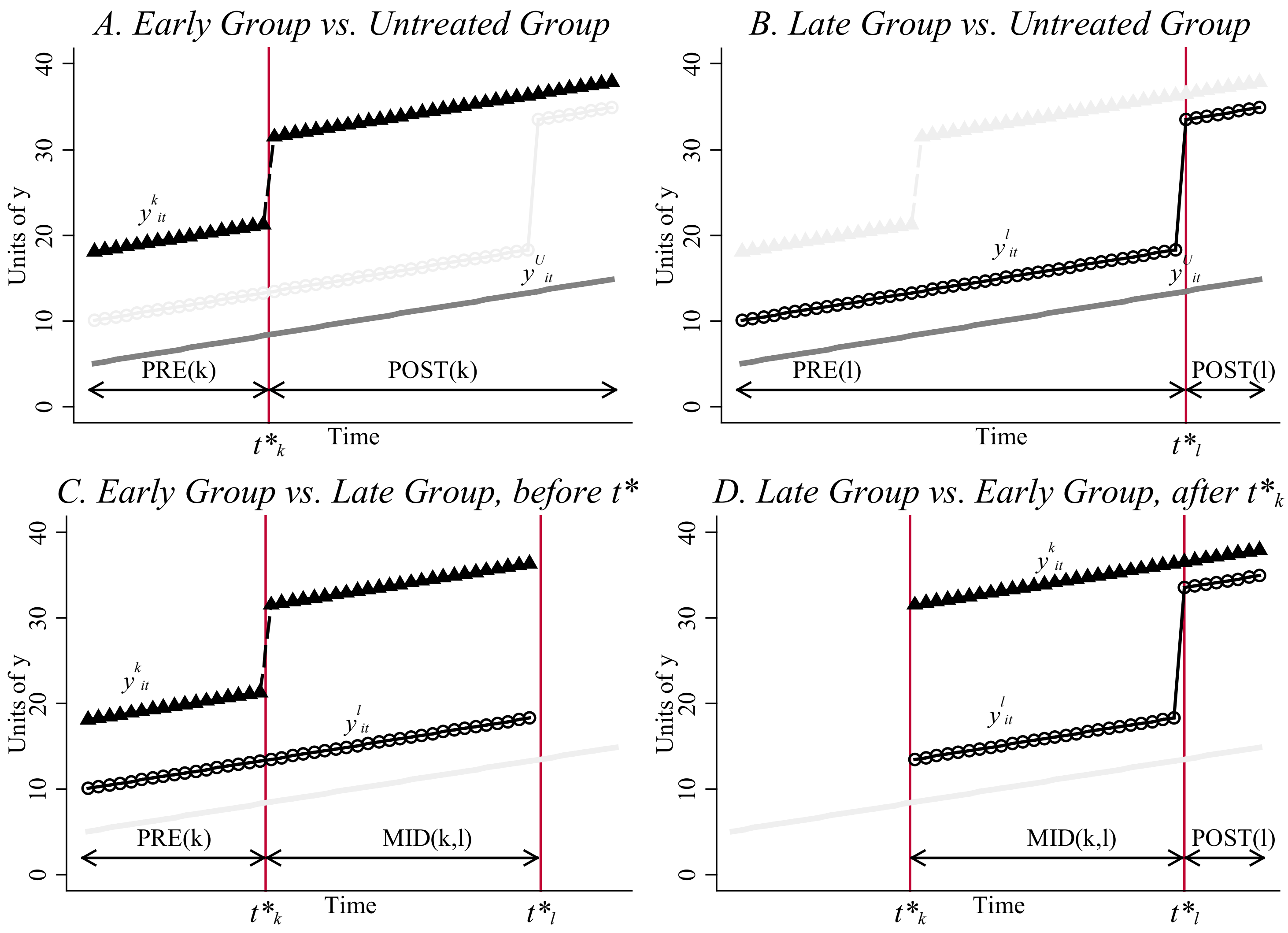
- However, in many settings treatment roll-out can be staggered (i.e., successive policy implementation in states)

- When treatment timing varies, the model above can be used ($Y_{ist} = \alpha + \delta D_{st} + \lambda_t + \xi_i + \epsilon_{ist}$), but
  - δ estimates a weighted average of all possible 2x2 DiD effects (weights are a function of group sizes and variance in treatment)
  - That estimate matches the ATT only under somewhat restrictive assumptions Goodman-Bacon, Andrew. 2021. Difference-in-Differences with Variation in Treatment Timing.
  - If there are time-varying treatment effects, applying the two-way FE model can lead to quite biased inferences Chaisemartin, C. de, and Xavier D'Haultfoeuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects."
  - This is an active area of research e.g., Callaway, B, and P. H. C. Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods" J Econometrics

Staggered treatments, 3 groups



Staggered treatments, 4 DiD estimates



A. Early Group vs. Untreated Group

B. Late Group vs. Untreated Group

C. Early Group vs. Late Group, before t^*

D. Late Group vs. Early Group, after t^*_k

Simulated data example

- Data with 4 time periods, N=2000
- Units in treated group are randomly (equal probability) assigned to first participate in the treatment (per group) in each time period
- 2000 units never receive tretment
- After dropping units treated in the first period we have

0	2	3	4
1656	1592	1532	1660

- DGP for untreated potential outcomes

$$Y_{it}(0) = \theta_t + \eta_i + X_i'\beta_t + \nu_{it}$$

NOTE:

- If η_i is distributed differently across groups, comparisons of outcomes (in levels) between treated and untreated will not yield ATT
- Because

$$\Delta Y_{it}(0) = (\theta_t - \theta_{t-1}) + X_i'(\beta_t - \beta_{t-1}) + \Delta \nu_{it}$$

the time path of the outcomes depends on covariates

  - Thus, (Unconditional) parallel trends assumption not valid (unless mean of covariates identical across groups or $\beta_t = \beta_{t-1} = \cdots = \beta_1$)

- DGP for treated potential outcomes

$$Y_{it}(g) = Y_{it}(0) + \mathbf{1}\{t \geq g\}(e+1) + (u_{it} - v_{it})$$

Here, $e := t - g$ is simply a variable in "event-time" metric (i.e., difference between current time and time when unit becomes treated)

- We have the following time-varying

$$ATT(g,t) = e + 1$$

in the post-treatment periods $t \geq g$

Other simulation parameters: $\theta_t = \beta_t = t$ for $t = 1, \ldots, 4$; $\eta_i \sim N(G_i, 1)$ with G_i the group an individual belongs to; $X_i \sim N(\mu_{D_i}, 1)$ with $\mu_{D_i} = 1$ for units that are never treated and 0 otherwise; $v_{it} \sim N(0,1)$, and $u_{it} \sim N(0,1)$

Heterogenous DiD estimates

Note: active area of research; (many) different proposals exist

- Estimate *ATT* for each combination of cohort and time
- Individuals / cases in sample are dneoted by i, $i = 1, \dots, N$
- Denote time by t, $t = 1, \dots, T$.
- Let d_{it} the treatment status of i at time t
- Cohorts g are defined by time group is treated. Let G_{ig} be an indicator equal to one if unit i is first treated at time g. Units in cohort g are denoted by $G_{ig} = 1$.
- For never treated units denote $G_{i0} = 1$; thus, cohort 0 indicates never treated units
- Note: one a unit is treated, it remains treated

Group-time treatment effects

- Denote by $\theta(g,t)$ the ATT for cohort g at time t
- It is defined as

$$\theta(g,t) = E\left(y_t(g) - y_t(0) \mid G_g = 1\right)$$

  - Here, $y_t(g)$ is the potential outcome at time t for those treated at time g
  - $y_t(0)$ is the potential outcome for the never treated
  - G_g equals 1 if a unit belongs to cohort g

Preliminaries

- Approach: transform problem into calssical 2x2 DID problem
- Restrict the data to an estimation sample with only two groups and only two periods based on g and t
- One group : all observations in cohort g; other group: control, i.e, untreated observations not in cohort g
- One time group: data in time t; other time group: period when cohort g is not treated ("base time")
- Defining the control group ($C^*_{g,t}$)
  - Never-treated: Let C^{NEV} be an indicator equal to one if a unit belongs to the never-treated group ($C^{\text{NEV}} = G_0$)
  - Use units not in cohort g and not yet treated at time t. Let $C^{\text{NY}}_{g,t}$ be an indicator equals to one if a unit belongs to the <u>not-yet-treated</u> group by time t.

$$C^{\text{NY}}_{g,t} = (1 - G_g)(1 - d_t).$$

- Note: also two ways to define base time t_0
  - Common base time Common base time $g-1$ for both pretreatment and posttreatment periods
  - Adaptive base time Choose the base time for the pretreatment periods; for pretreatment periods, base time is $t-1$; for posttreatment periods it is $g-1$

- For each unit, we observe $\{\tau_i, y_{i,\tau_i}, \mathbf{x}_{i,\tau_i}, d_{i,\tau_i}, \mathbf{z}_{i,\tau_i}, G_{i,\tau_i}\}$
  - y_i is the outcome
  - d_i is the treatment indicator
  - $\mathbf{x}_i$ are pretreatment covariates in outcome model
  - $\mathbf{z}_i$ are covariates in treatment assignment model
  - $\tau_i \in \{1, \ldots, T\}$ is a categorical variable indicating the time when unit i is observed (let T_t equal one if unit is observed at time t, zero otherwise)

Estimators

- Define the following notation, where the superscript denotes the group we condition on

$$m^{\text{treat}}_{g,s}(\mathbf{x}) = E(y \mid \mathbf{x}, G_g = 1, \tau = s)$$
$$m^{\text{comp}}_{g,s,t}(\mathbf{x}) = E(y \mid \mathbf{x}, C^*_{g,t} = 1, \tau = s)$$

and

$$w^{\text{treat}}_{g,s} = \frac{T_s G_g}{E(T_s G_g)}$$
$$w^{\text{comp}}_{g,s,t}(\mathbf{z}) = \frac{\dfrac{T_s p_{g,t}(\mathbf{z}) C^*_{g,t}}{1 - p_{g,t}(\mathbf{z})}}{E\left\{\dfrac{T_s p_{g,t}(\mathbf{z}) C^*_{g,t}}{1 - p_{g,t}(\mathbf{z})}\right\}}$$

- $p_{g,t}(\mathbf{z})$ is defined by

$$p_{g,t}(\mathbf{z}) = \Pr(G_g = 1 \mid \mathbf{z}, G_g + C^*_{g,t} = 1)$$

Estimators

- Estimating the ATT via regression adjustment

$$\theta(g,t) = E\left(\frac{G_g}{E(G_g)}\left[\{m^{\text{treat}}_{g,t}(\mathbf{x}) - m^{\text{treat}}_{g,g-1}(\mathbf{x})\} - \{m^{\text{comp}}_{g,t,t}(\mathbf{x}) - m^{\text{comp}}_{g,g-1,t}(\mathbf{x})\}\right]\right)$$

- Estimating the ATT via inverse probability weighting

$$\theta(g,t) = E\left\{\left(w^{\text{treat}}_{g,t} - w^{\text{treat}}_{g,g-1}\right) y\right\} - E\left[\left\{w^{\text{comp}}_{g,t,t}(\mathbf{z}) - w^{\text{comp}}_{g,g-1,t}(\mathbf{z})\right\} y\right]$$

- Extension: estimation via AIPW / DR

$$\begin{aligned}
\theta(g,t) = E&\left(\frac{G_g}{E(G_g)}\left[\{m^{\text{treat}}_{g,t}(\mathbf{x}) - m^{\text{treat}}_{g,g-1}(\mathbf{x})\} - \{m^{\text{comp}}_{g,t,t}(\mathbf{x}) - m^{\text{comp}}_{g,g-1,t}(\mathbf{x})\}\right]\right) \\
&+ \; E\left[w^{\text{treat}}_{g,t}\{y - m^{\text{treat}}_{g,t}(\mathbf{x})\} - w^{\text{treat}}_{g,g-1}\{y - m^{\text{treat}}_{g,g-1}(\mathbf{x})\}\right] \\
&- \; E\left[w^{\text{comp}}_{g,t,t}(\mathbf{z})\{y - m^{\text{comp}}_{g,t,t}(\mathbf{x})\} - w^{\text{comp}}_{g,g-1,t}(\mathbf{z})\{y - m^{\text{comp}}_{g,g-1,t}(\mathbf{x})\}\right]
\end{aligned}$$

Estimation steps

1. Restrict sample to time t and t_0. Keep only units either in cohort g or in control group $C^*_{g,t}$
2. Use a (parametric) model to estimate the “nuisance functions”
   (a) Outcomes: use linear regression to estimate $m^{\text{treat}}_{g,t}(\mathbf{x})$, $m^{\text{treat}}_{g,t_0}(\mathbf{x})$, $m^{\text{comp}}_{g,s,t}(\mathbf{x})$, and $m^{\text{comp}}_{g,s,t_0}(\mathbf{x})$
   (b) Propensity score: use logit to estimate $p_{g,t}(\mathbf{z})$.
   (c) Probability weights: $w^{\text{treat}}_{g,t}$, w^{treat}_{g,t_0}, $w^{\text{comp}}_{g,s,t}(\mathbf{z})$, and $w^{\text{comp}}_{g,s,t_0}(\mathbf{z})$ using propensity scores T_t and G_g.
3. Plug in estimates into equation on the previous slide

$E(\cdot)$ is replaced by the sample average

Read: Callaway, B, and P. H. C. Sant’Anna. 2021. “Difference-in-Differences with Multiple Time Periods” J Econometrics

Synthetic controls

Creating counterfactuals for a single unit

- Single unit experiences event (treatment) at point in time
- Qualitative case studies (Mills methods)
- Quantitative case studies
- Difficulty of obtaining comparable case for counterfactual
- Idea: create artificial case for comparising by weighting a set of cases
- Captures what would have happened had treatment not occurred
- Generalization of DiD strategies

Advantages

- Precludes extrapolation (uses convex hull of control; does not extrapolate beyong support like regression does (in extreme cases, e.g., King and Zeng 2006),)
- Processing: construction of counterfactual only requires pre-treatment data
- Explicit weights (remember: regression weights are implicit!)

Details

- Let Y_{jt} be outcome for unit j of $J+1$ aggregate units, treatment group is $j=1$
- Intervention at time T_0, effect on treatment group
- Causal effect in post-treatment period

$$Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}$$

here w_j^* is a vector of optimal weights

- Matching variables X_0 and X_1 are predictors of post-treatment outcomes, must be unaffected by intervention.
- Weights are chosen to minimize $\|X_1 - X_0 W\|$ under two constraints
  1. $W = (w_2, \ldots, w_{J+1})'$ with $w_j \geq 0$ for $j = 2, \ldots J+1$
  2. $w_2 + \cdots + w_{J+1} = 1$

There is work on relaxing nonnegativity constraint (Doudchenko Imbens 2017)

Details

- One possibility (as in Abadie et al 2010 JASA)

  $$\|X_1 - X_0 W\| = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)}$$

  with V a $k \times k$ matrix (pos. semidef.), typically diagonal with main diagonal $v_1, \ldots, v_k$
- Define X_{jm} as the value of the mth covariate.
- The synthetic control weights minimize

  $$\sum_{m=1}^{k} v_m \left(X_{1m} - \sum_{j=2}^{J+1} w_j X_{jm} \right)^2$$
- Think of v_m as the importance given to mth variable when assessing imbalance between treated and synthetic control units

Details

- Choice of V matters (W^* depends on it!)
- Synthetic control $W^*(V)$ is supposed to reproduce counterfactual outcome absent of treatment
- Weights $v_1, \ldots, v_k$ should reflect predictive value of covariates
- Different options for choice of V, in practice mostly

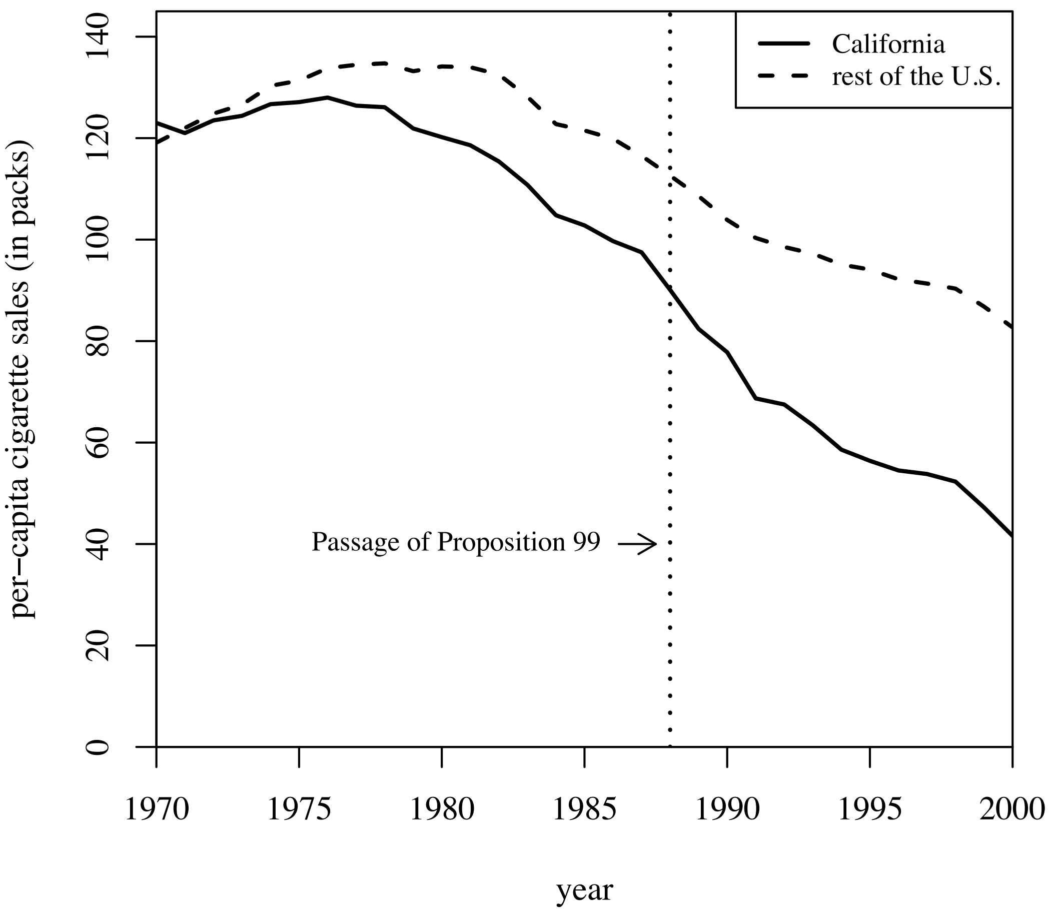
$$\sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=1}^{J+1} w_j^*(V) Y_{jt} \right)^2$$

(minimzed mean squared prediction error)

- What about unobservables? Abadie et al 2010 reason that length of pre-intervention period matters

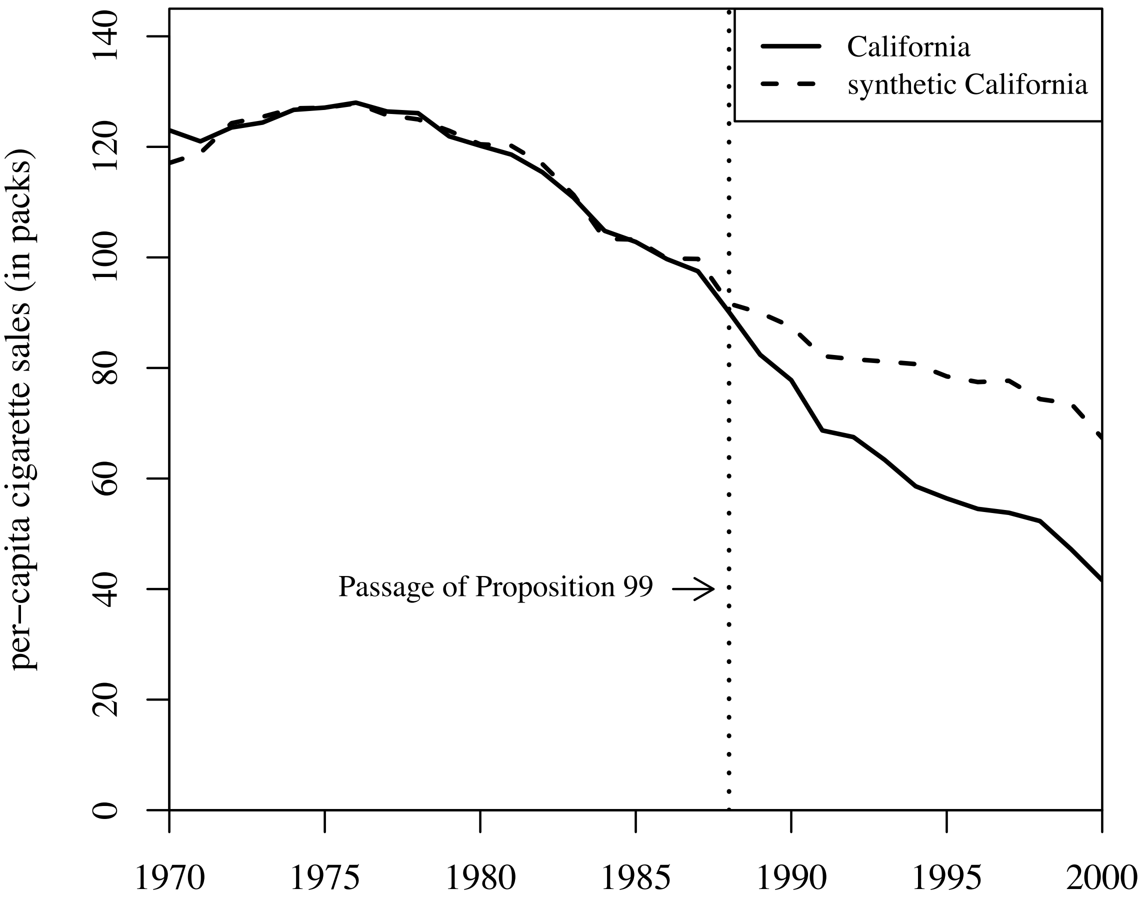
Example: tobacco control legislation

- Prop 99 in CA (cigarette taxes + 25 cent, ordinances, media campaigns...)



‘Synthetic’ California

- Comparison with synthetic control



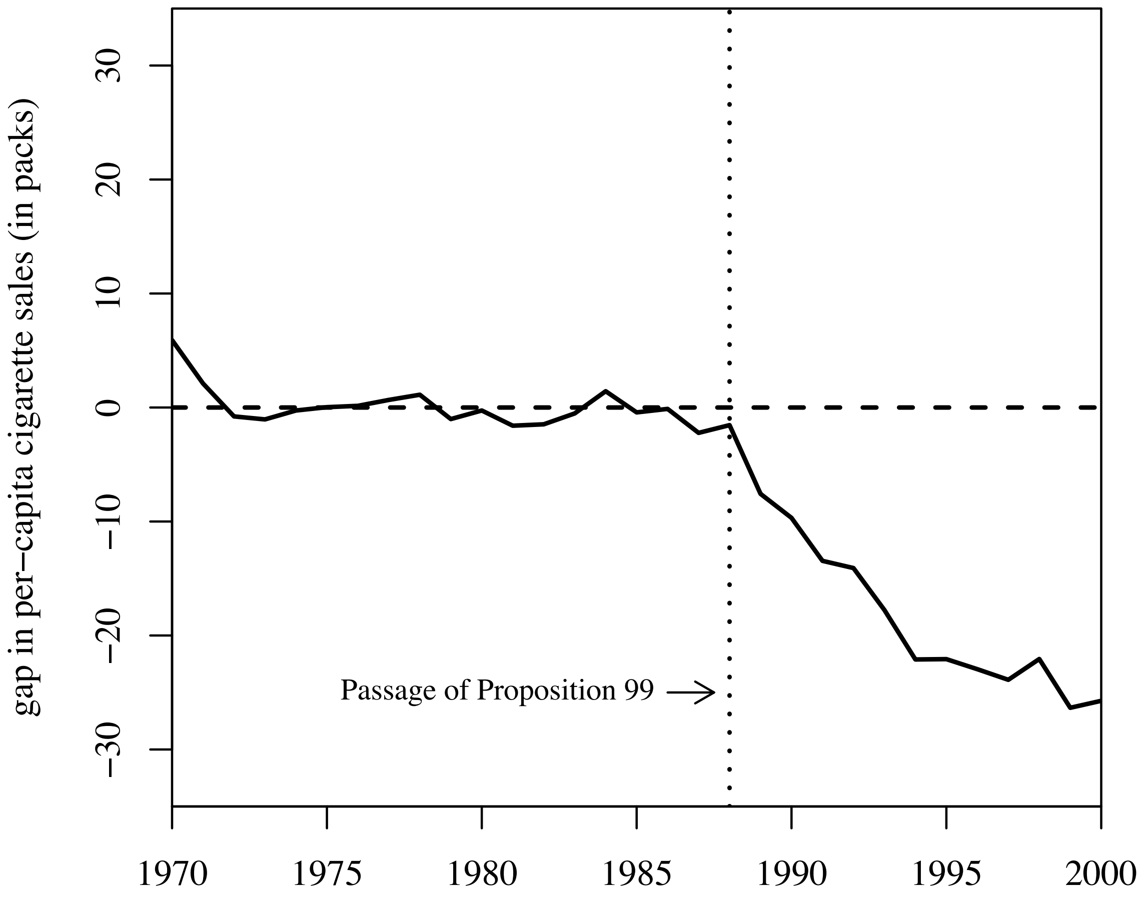
‘Synthetic’ California balance

	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).

Prop 99 ‘effect estimate’

Difference between both series (actual and counterfactual)



Is the difference significant?

Randomization inference in synth. control analyses

RI logic: randomize treatment to each unit, re-estimate model, check ‘tail position’ of estimate

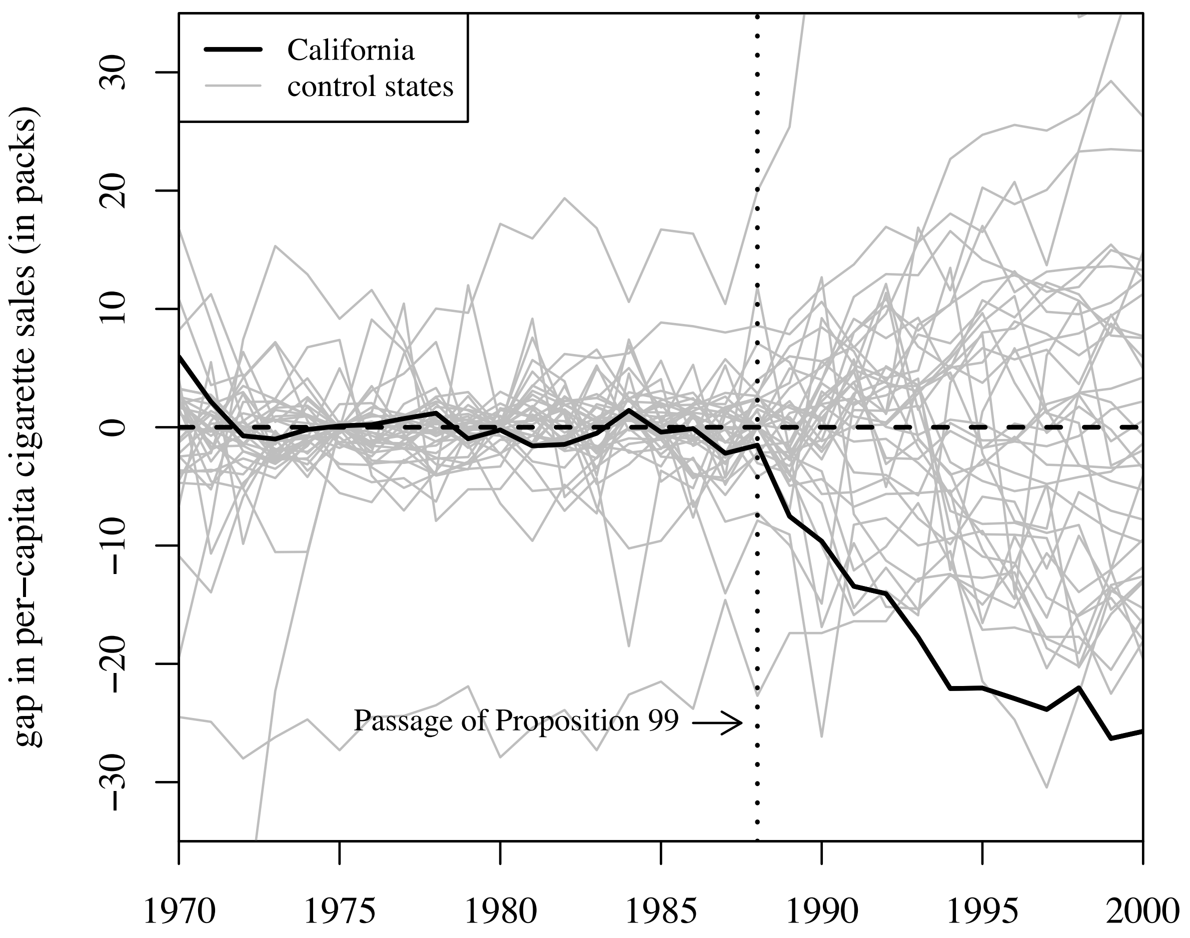
- Iteratively apply SC to each unit in donor pool. Obtain distribution of placebo effects
- Calculate RMSPE for each placebo in **pre-treatment** period

$$RMSPE = \left(\frac{1}{T - T_0} \sum_{t=T_0+t}^{T} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt} \right)^2 \right)^{1/2}$$

- Calculate RMSPE for each placebo for **post-treatment** period (mutatis mutandis)
- Compute ratio of post- to pre-treatment RMSPE
- Sort ratio in descending order
- Calculate treated unit’s ratio in the distribution: $p = rank/total$

Randomization inference in synth. control analyses

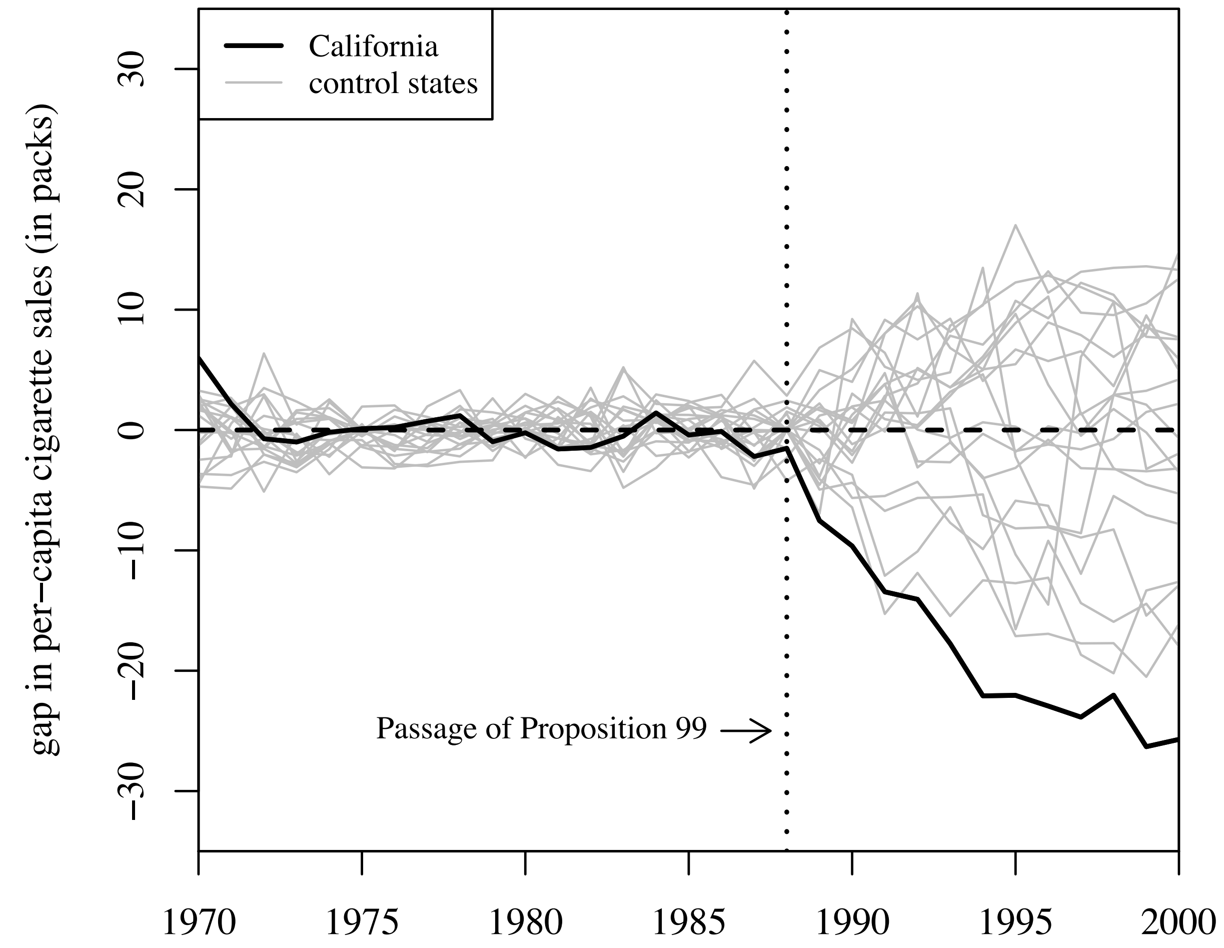
(All States in Donor Pool)



Randomization inference in synth. control analyses

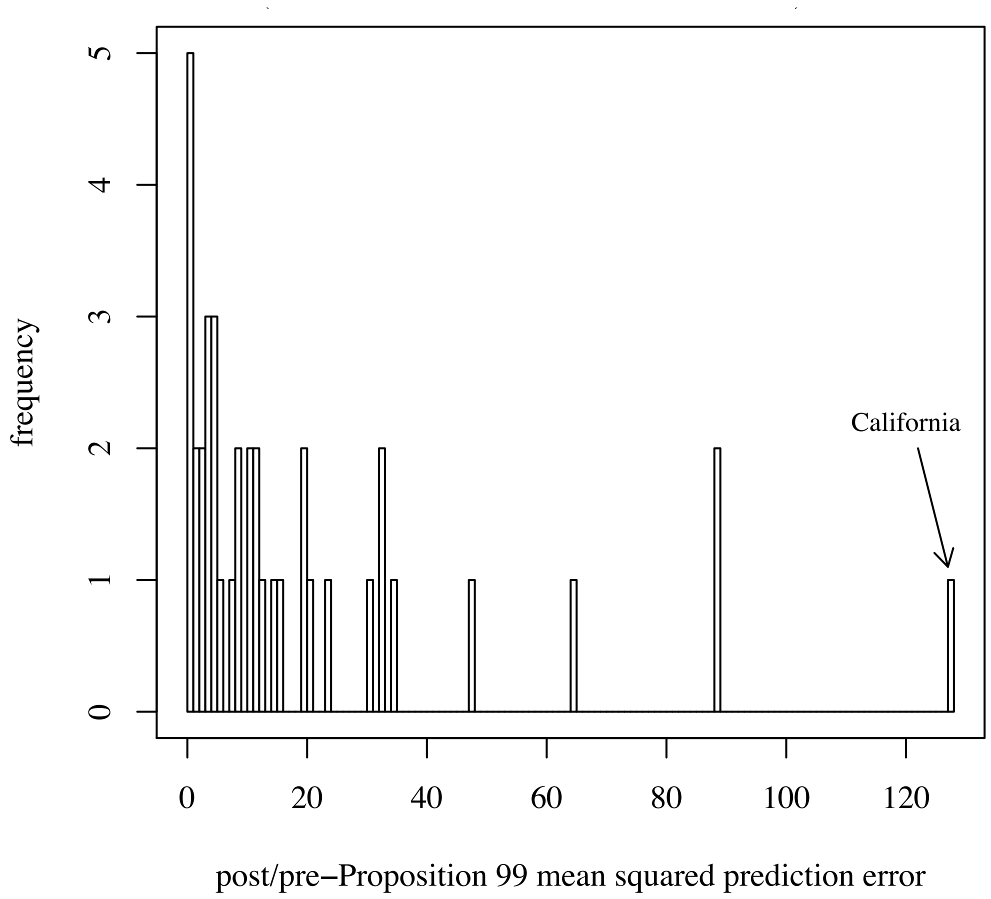
with extreme pre-treatment RMSPE excl.

(PRE-PROP. 99 MSPE ≤ 2 TIMES PRE-PROP. 99 MSPE FOR CA)



Randomization inference in synth. control analyses

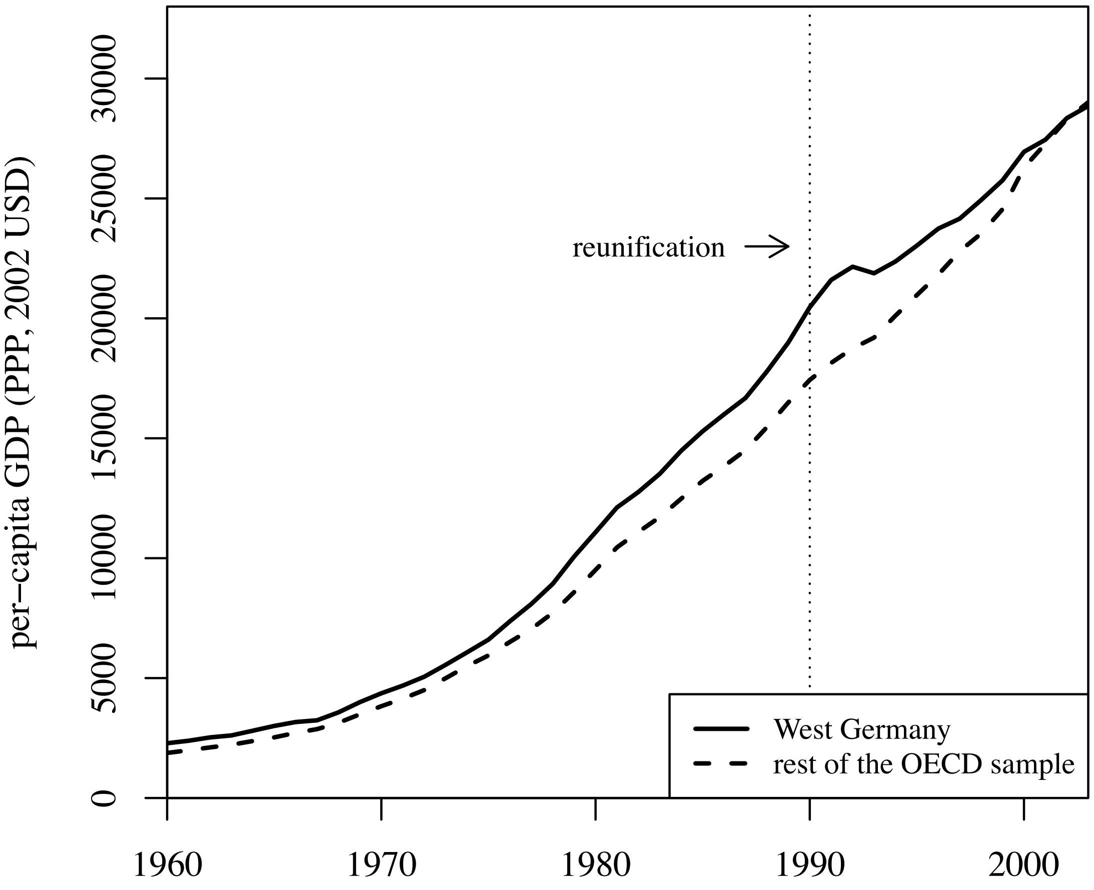
Histogram: post/pre-prop-99 mean squared prediction error ratios



CA rank: 1st out of 38 ⇒ Exact p-value $= 0.026$

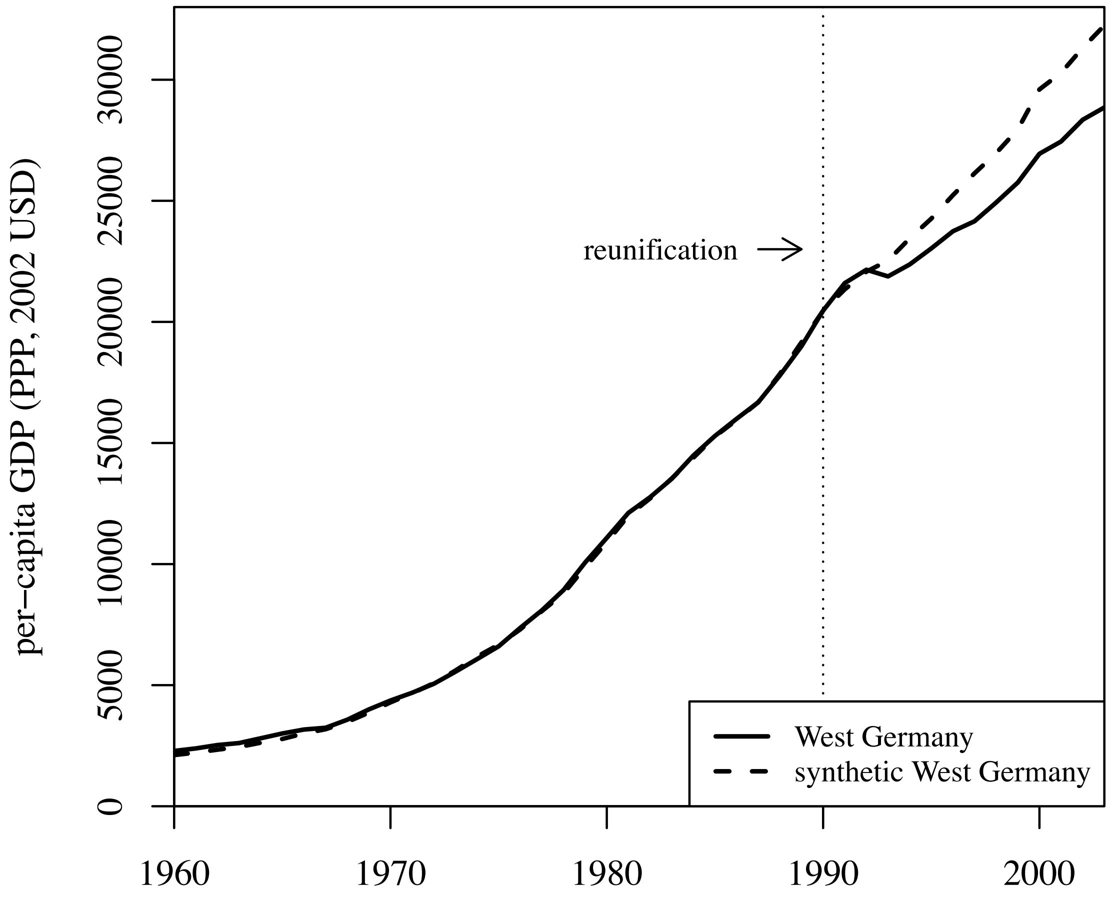
More placebo tests

- Example: GDP pc of West-Germany after reunification (Abadie et al 2015 AJPS)



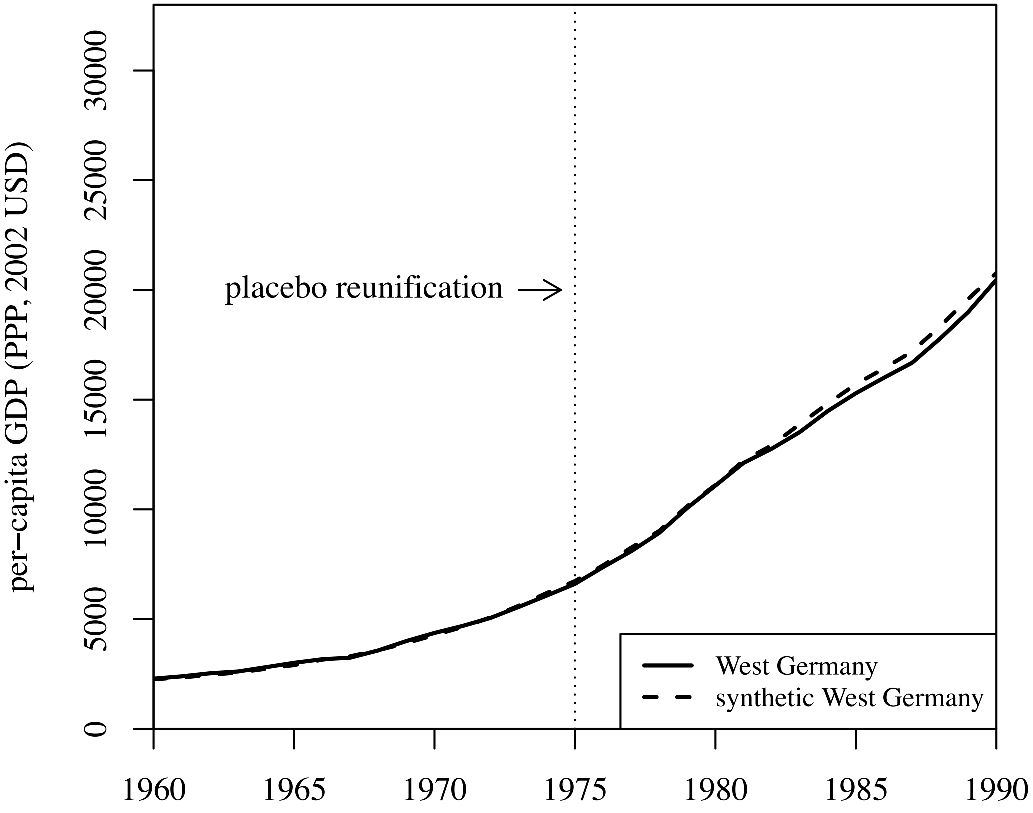
More placebo tests

- Synthetic control estimate



More placebo tests

- Placebo treatment



Extensions

- Beyond placebo p-values: confidence intervals
  - Hahn and Shi 2017. Synthetic controls and inference. Econometrics 5 (4).
  - Firo and Possebom. 2018.
  - Cattaneo et al. 2021. Prediction Intervals for Synthetic Control Methods. JASA 116.
- Regression-based estimators
  - Doudchenko and Imbens. 2016.
  - Chernozhukov, Wüthrich, and Zhu. 2019.
  - Arkhangelsky et al. 2019.
- Matrix completion methods, e.g., Athey et al. 2020.