Causal Inference with Panel Data

Lecture 2: Synthetic Control and Extensions

Yiqing Xu (Stanford University)
Washington University in St. Louis

25 August 2021

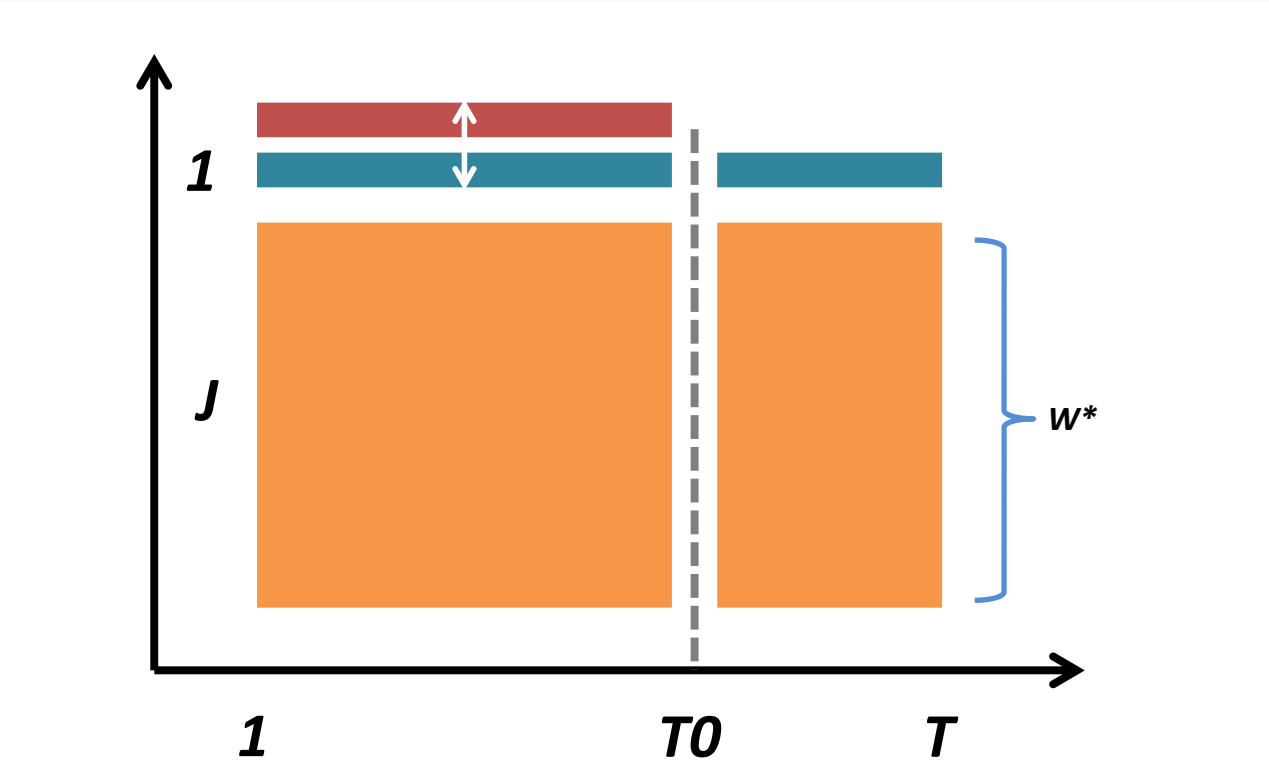
Today's Plan

- The synthetic control method: a review
- Alternative algorithms
- Next lecture
  - The interactive fixed effect model
  - The matrix completion method
  - Diagnostics
  - Bayesian multi-factor models

The Synthetic Control Method (SCM)

SCM: Basic Idea

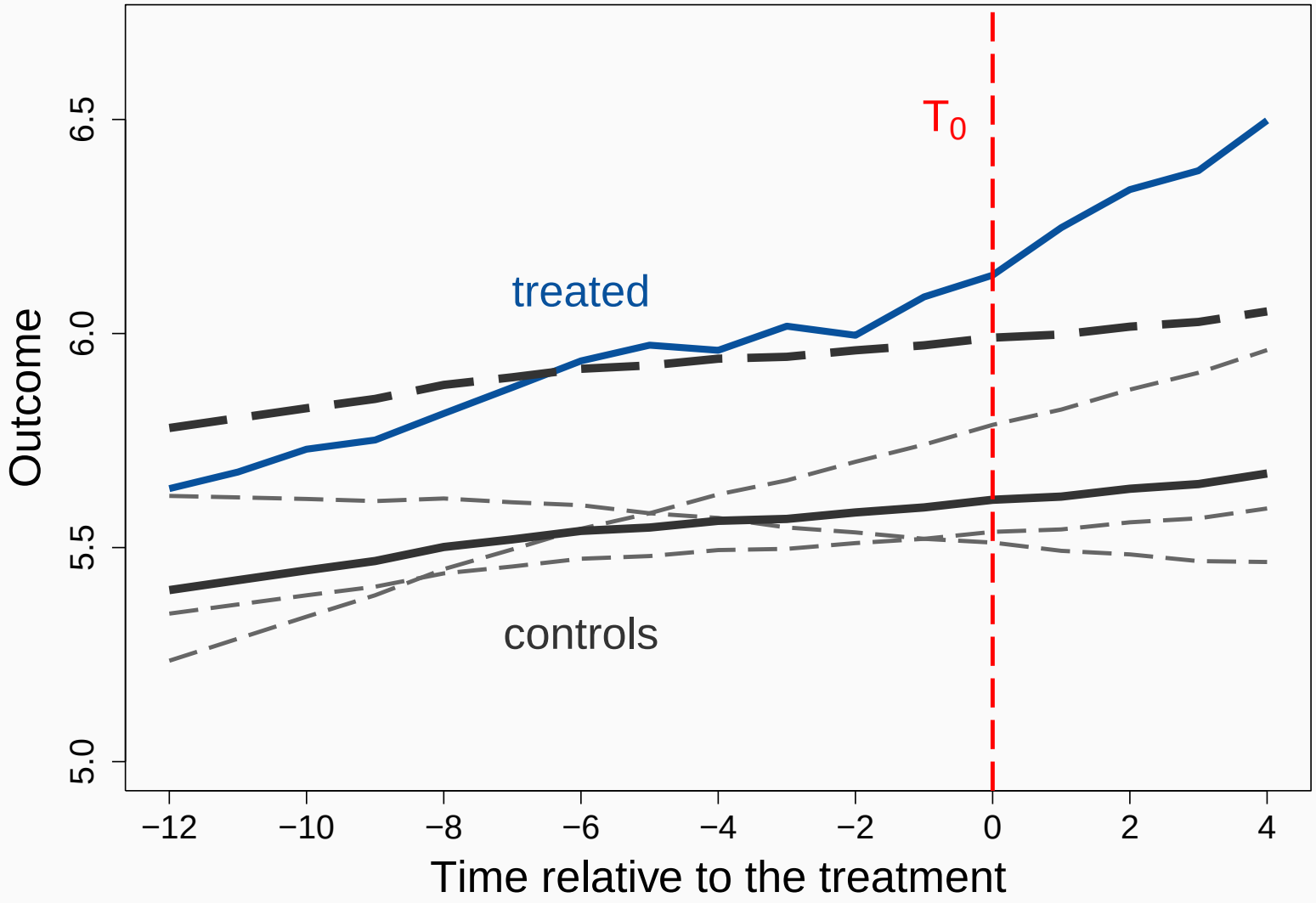
- $J+1$ units in periods $1, 2, \ldots, T$; one treated "1", J controls
- Region "1" is exposed to the intervention after period T_0
- We aim to estimate the effect of the intervention on Region "1"



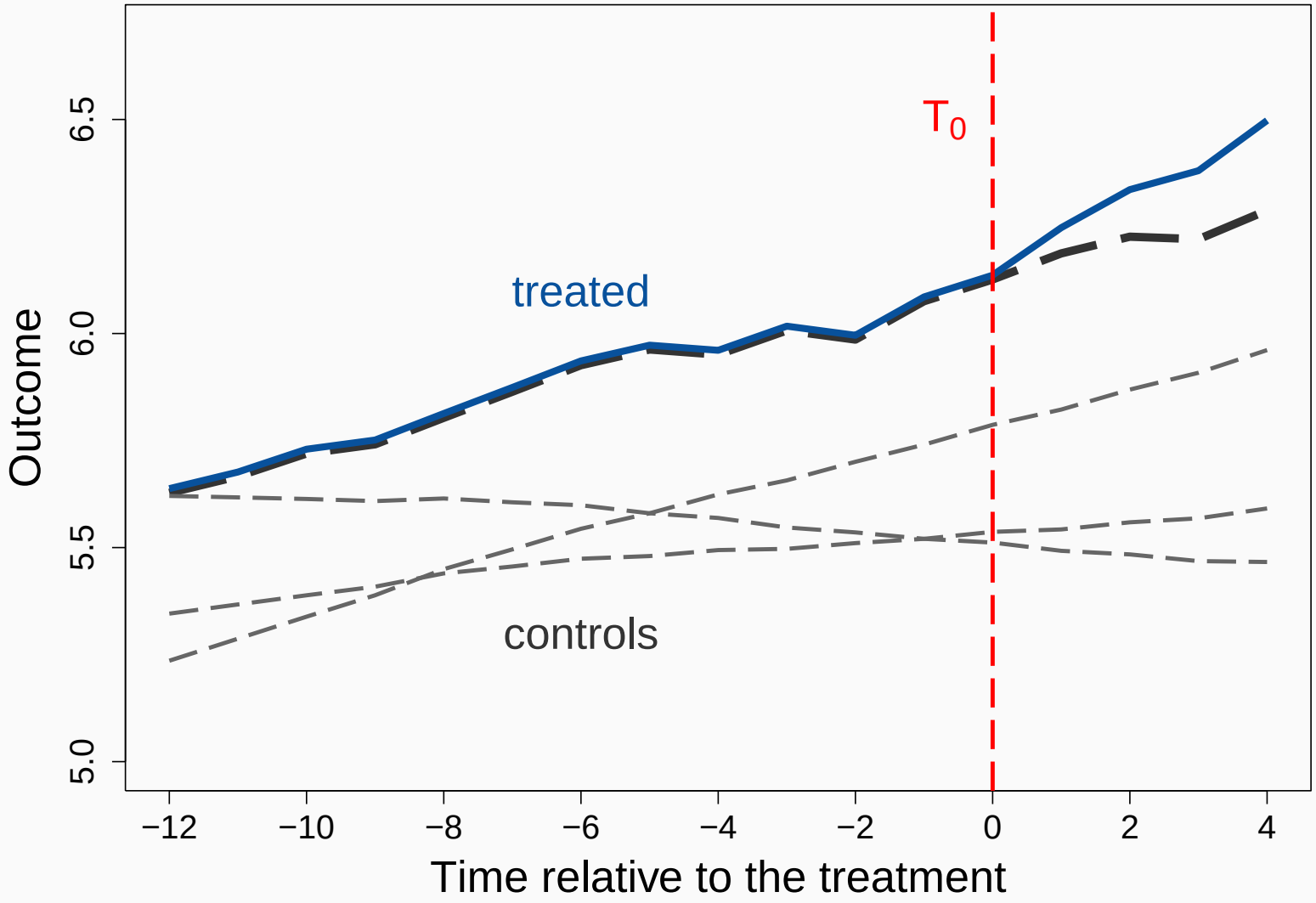
SCM: Insights

- Athey and Imbens (2016): "[a]rguably the most important innovation in the policy evaluation literature in the last 15 years."
- A combo of many innovations
  - Take advantage of pre-treatment outcomes
  - Use cross-sectional instead of temporal correlations in data
  - Construct a convex combination of donors to construct a counterfactual
  - Reserve some pre-treatment periods for testing

Difference-in-Differences (DiD)



SCM (and Many Extensions)



Theoretical Justification

$$Y_{it} = \tau_{it} D_{it} + \theta_t' Z_i + \xi_t + \lambda_i' f_t + \varepsilon_{it}$$

or

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

- 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 pre-treatment outcomes for both treated and controls, we can try to use this information to "balance on" these confounders
- We will discuss the model-based approach later

Theoretical Justification

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

- 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 τ_{1t} is:

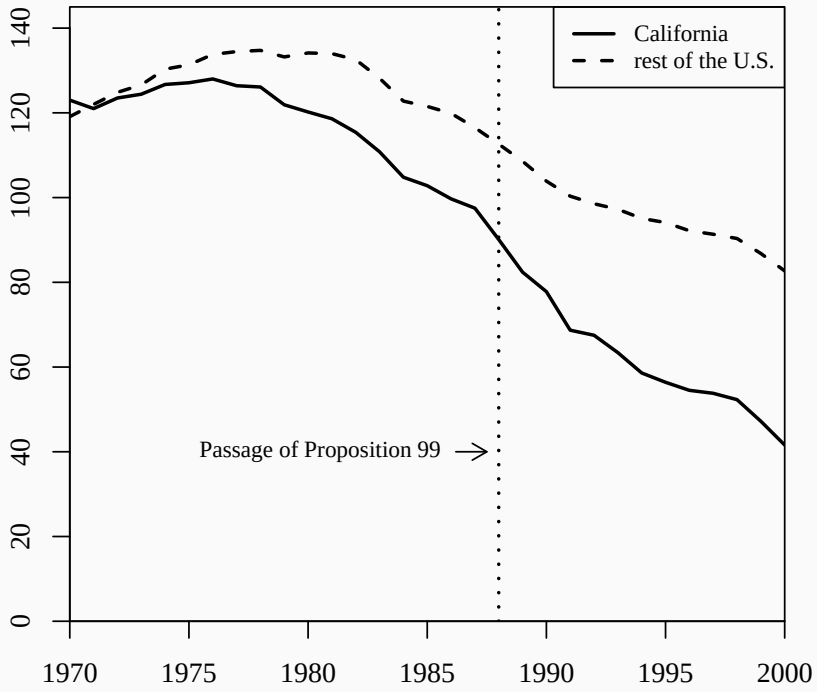
$$\widehat{\tau}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}, \quad t \in \{T_0 + 1, \ldots, T\}$$

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.).

Example: Proposition 99 on Cigarette Consumption

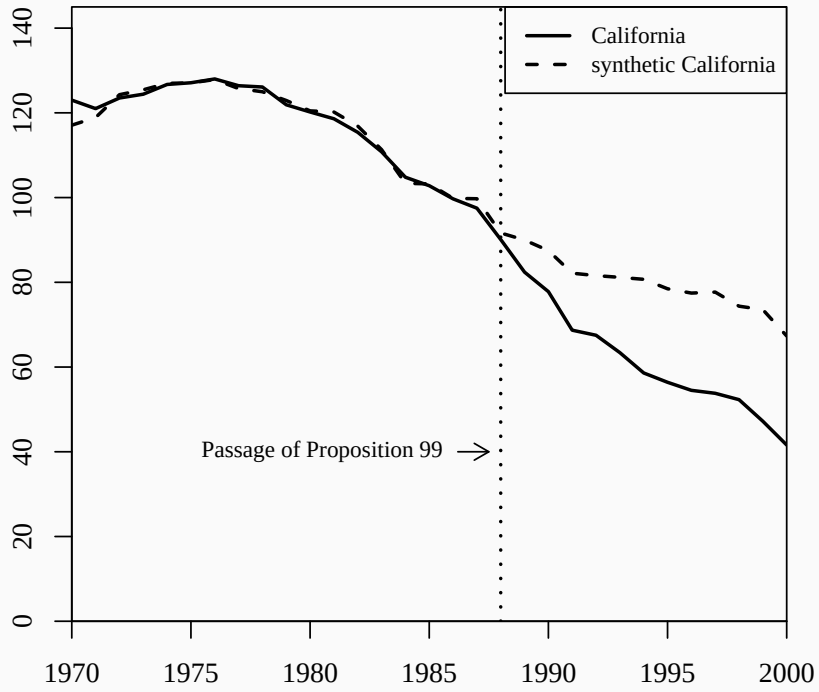
- In 1988, California first passed comprehensive tobacco control legislation (cigarette tax, media campaign etc.)
- Using 38 states that had never passed such programs as controls



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

Example: 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



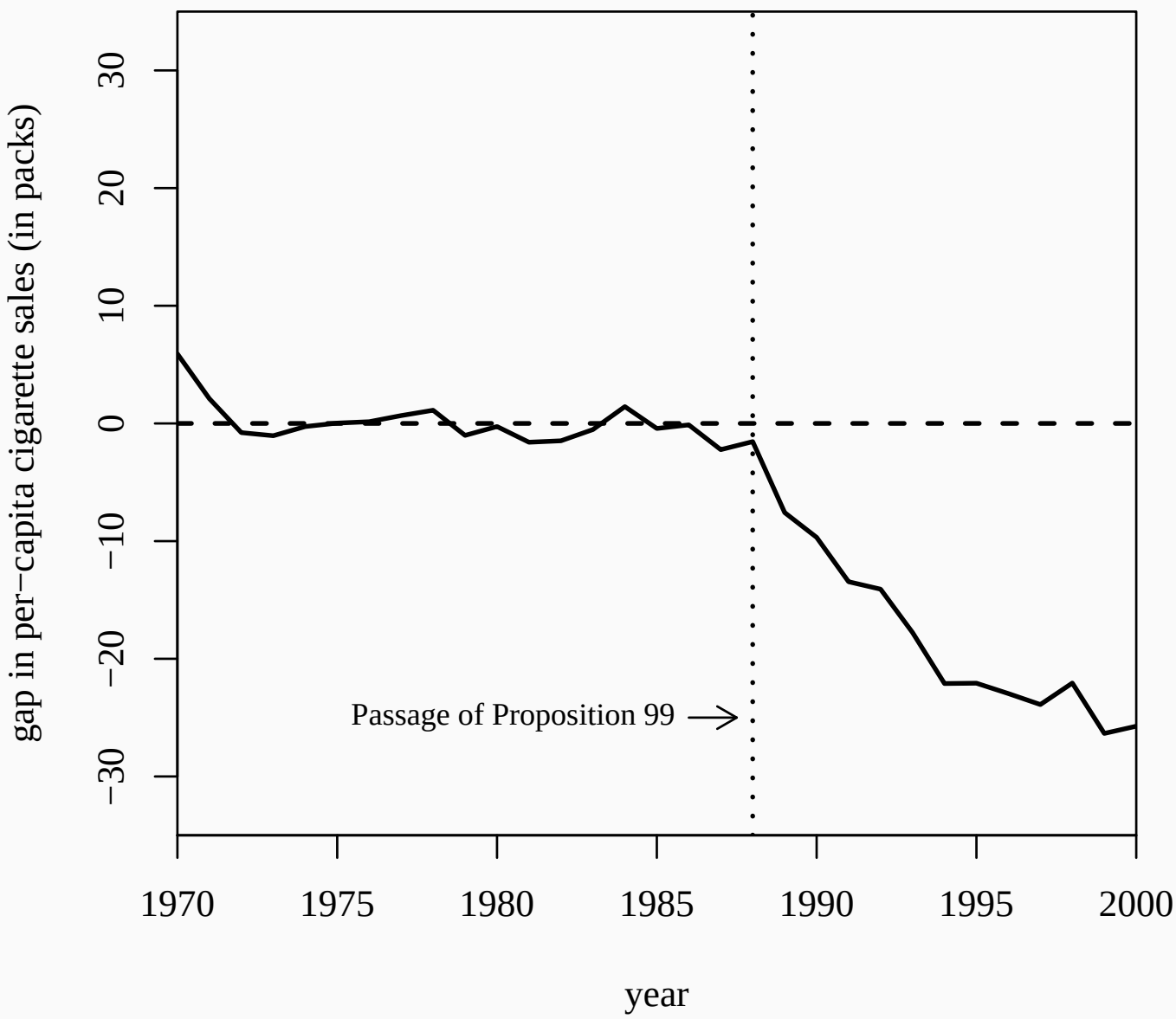
Cigarette Consumption: CA and Synthetic CA

Predictor Means: Actual vs. Synthetic California

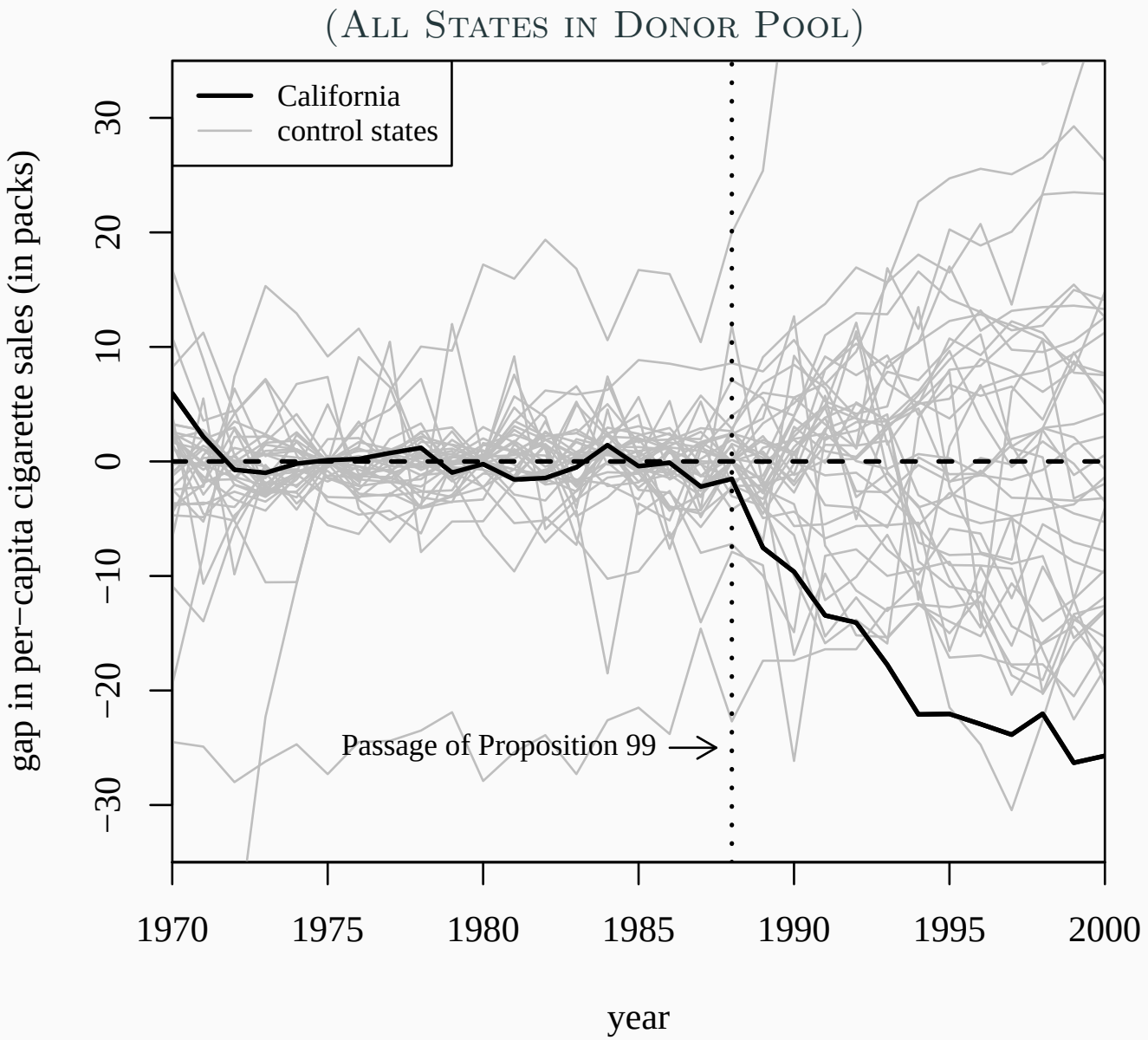
Variables	California		Average of
	Real	Synthetic	38 control states
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15-24	17.40	17.40	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.20	23.75
Cigarette sales per capita 1988	90.10	91.62	114.20
Cigarette sales per capita 1980	120.20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

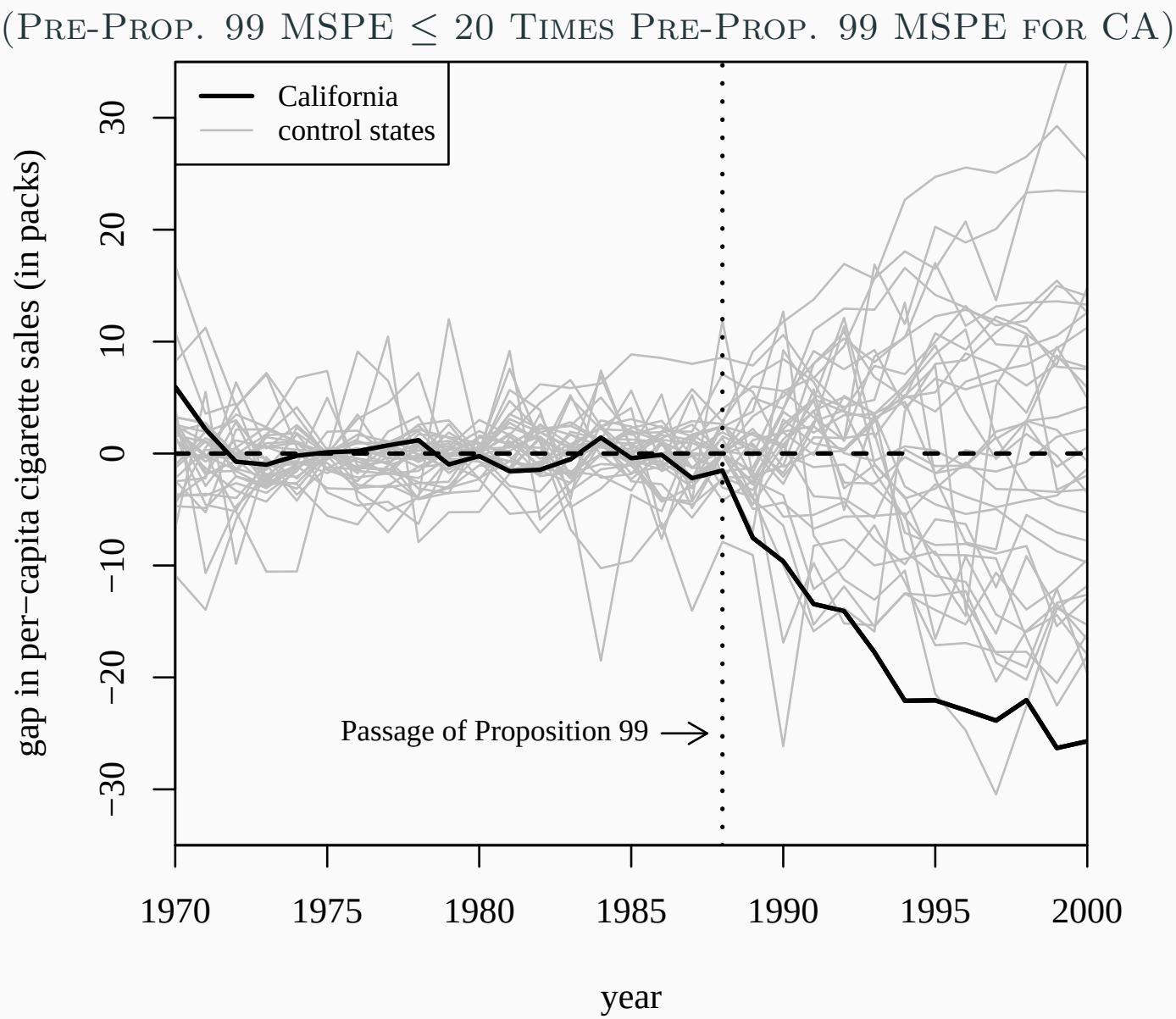
Smoking Gap Between CA and Synthetic CA



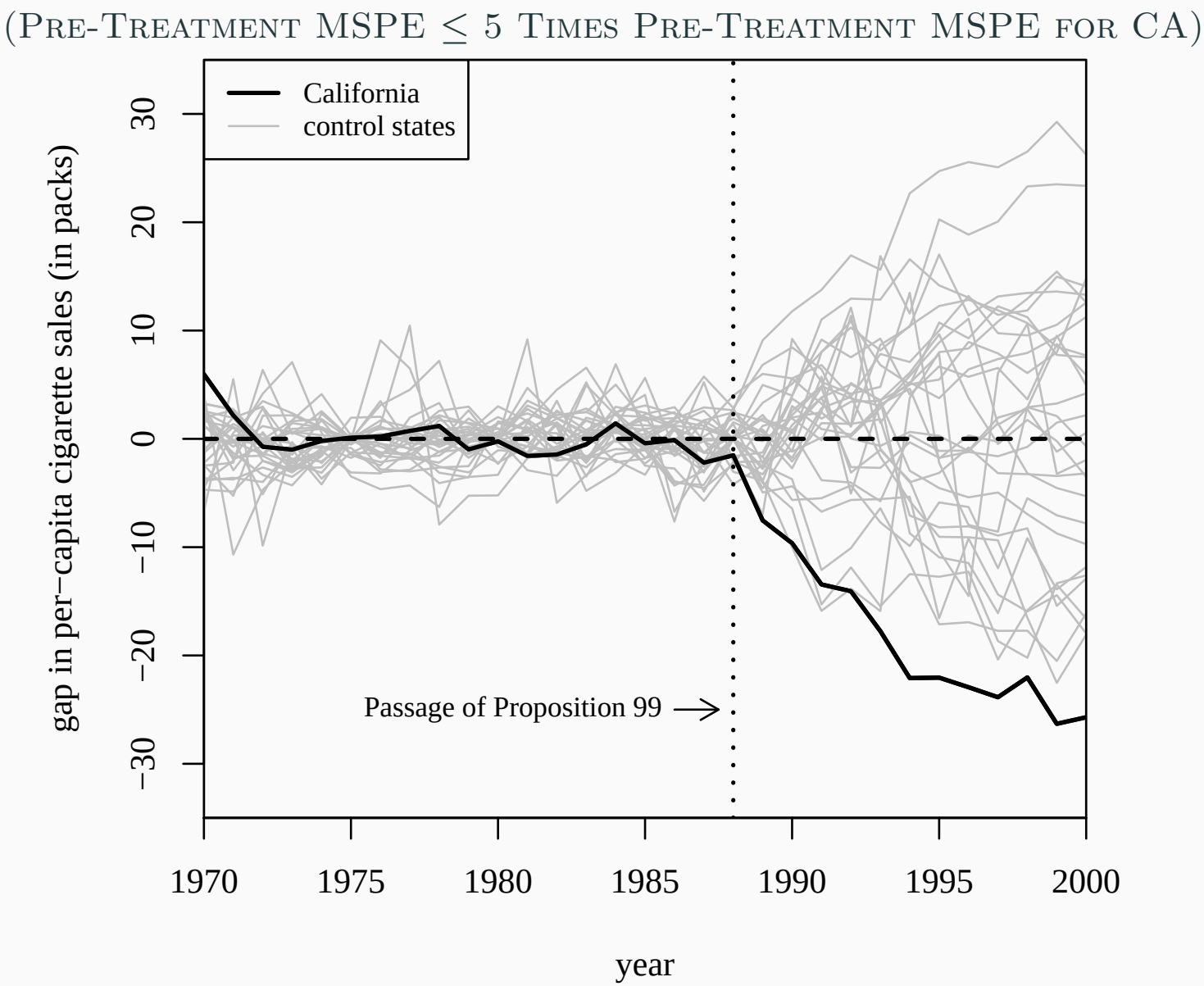
Smoking Gap for CA and 38 Control States



Smoking Gap for CA and 34 Control States

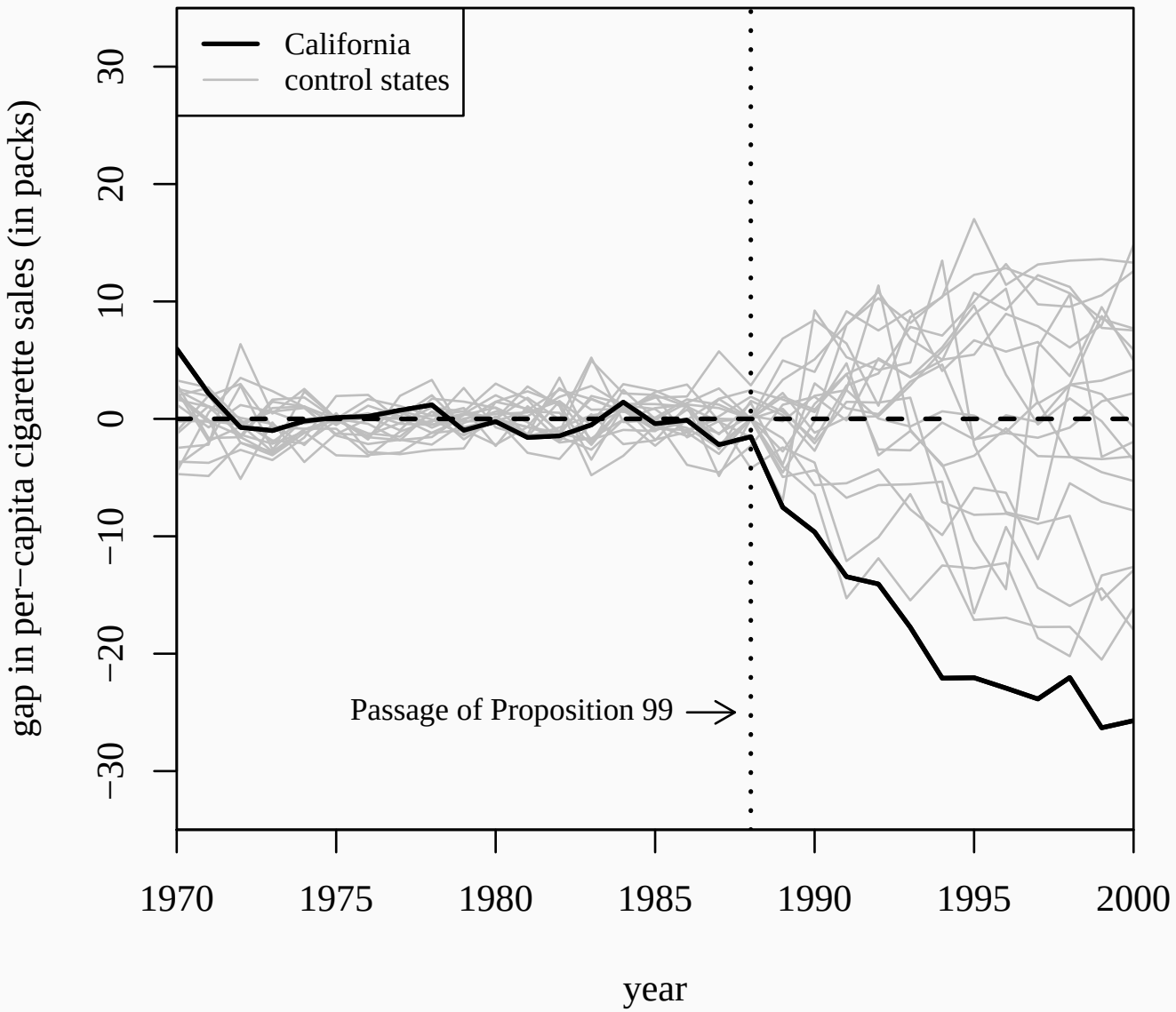


Smoking Gap for CA and 29 Control States

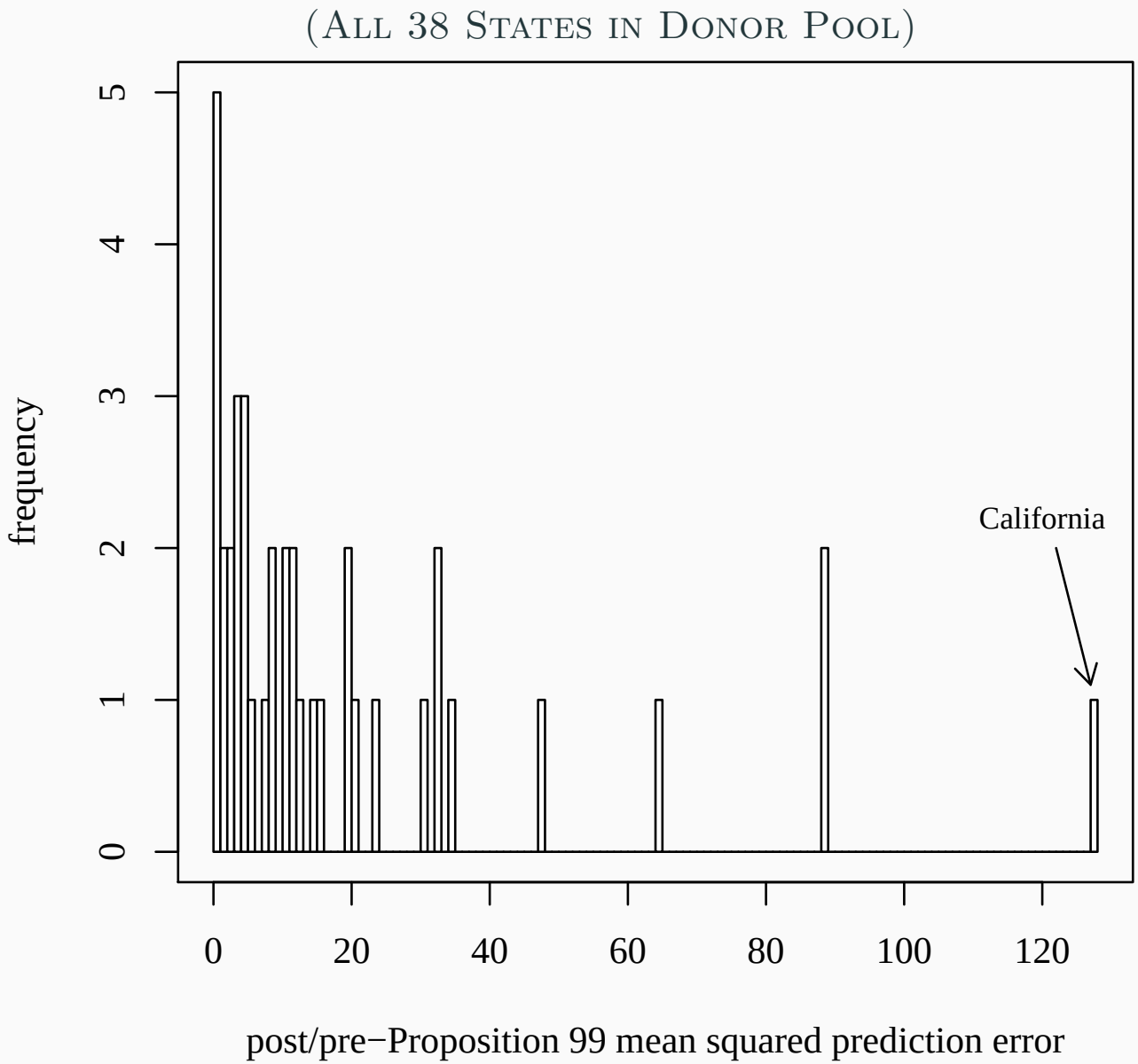


Smoking Gap for CA and 19 Control States

(Pre-Treatment MSPE $\leq$ 2 Times Pre-Treatment MSPE for CA)



Ratio Post-Treatment MSPE to Pre-Treatment MSPE



Limitations

- Algorithmic
  - Deal with one treated unit at a time
  - Deal with one outcome at a time
  - Slow to implement and sometimes difficult 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)
- Inference
  - Permutation inference and sensitivity analysis, (e.g. Hahn and Shi 2016; Firpo et al. 2017; Chernochukov 2017)
  - Inflated precision with nonstationary data (Cattaneo et al. 2019)
- Identification
  - My opinion: intrinsically, a method based on strict exogeneity (fixed timing)

Alternative Algorithms

Balancing, Regression, DiD, and SCM (Doudchenko & Imbens 2016)

- Panel methods can be characterized into three broad groups:
  - DiD: $\Delta Y^{\text{post}} - \Delta Y^{\text{pre}}$
  - Matching: on both pre-treatment outcomes and other covariates
  - SCM: For each treated unit, a "synthetic control" is constructed as a weighted average of control units s.t. the weighted average matches pre-treatment outcomes and covariates

- Doudchenko & Imbens (2016) provide a framework to nest existing approaches and estimators

Notation

- $N+1$ units observed for T periods, with a subset of treated units (for simplicity, unit 1) treated from T_0+1 and onwards
- Treatment : $D_{i,t} = \mathbb{1}_{i=1 \wedge t \in T_0+1,\ldots,T}$
- Potential outcomes for unit 0 define the treatment effect: $\tau_{0,t} := Y_{0,t}(1) - Y_{0,t}(0)$ for $t = T_0+1,\ldots,T$
- Observed outcome: $Y_{i,t}^{obs} = Y_{i,t}(D_{i,t})$

Outcome Matrices

$$\mathbf{Y}^{\text{obs}} = \begin{bmatrix} \mathbf{Y}^{\text{obs}}_{\text{t, pre}} & \mathbf{Y}^{\text{obs}}_{\text{t, post}} \\ \mathbf{Y}^{\text{obs}}_{\text{c, pre}} & \mathbf{Y}^{\text{obs}}_{\text{c, post}} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{\text{t, pre}}(0) & \mathbf{Y}_{\text{t, post}}(1) \\ \mathbf{Y}_{\text{c, pre}}(0) & \mathbf{Y}_{\text{c, post}}(0) \end{bmatrix} \qquad (N+1) \times T$$

$$\mathbf{Y}(0) = \begin{bmatrix} \mathbf{Y}_{\text{t, pre}}(0) & ? \\ \mathbf{Y}_{\text{c, pre}}(0) & \mathbf{Y}_{\text{c, post}}(0) \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{\text{t, pre}}(0) & ? \\ \mathbf{Y}_{\text{c, pre}}(0) & \mathbf{Y}_{\text{c, post}}(0) \end{bmatrix}$$

Relative magnitudes of T and N might dictate whether we impute the missing potential outcome ? using this or this comparison

- $N \gg T_0$, $\mathbf{Y}(0)$ is "tall", and red comparison becomes appealing relative to blue. So matching methods are attractive.
- $T_0 \gg N$, $\mathbf{Y}(0)$ is "fat", and matching becomes infeasible. So it might be easier to estimate blue dependence structure.
- Finally, if $T_0 \approx N$, regularization strategy for limiting the number of control units that enter into the estimation of $Y_{1,T_0+1}(0)$ may be important

Common Structure: Four Constraints

- Focus on last period. Many estimators impute $Y_{1,T}(0)$ with the linear structure

$$\widehat{Y}_{1,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y_{i,T}^{\text{obs}}$$

  while differ in how μ and ω are chosen as a function of $\mathbf{Y}^{\text{obs}}_{\text{c, post}}$, $\mathbf{Y}^{\text{obs}}_{\text{t,pre}}$, $\mathbf{Y}^{\text{obs}}_{\text{c, pre}}$

- Impose four constraints
  1. No Intercept: $\mu = 0$. Stronger than Parallel trends in DiD.
  2. Adding up : $\sum_{i=1}^{n} \omega_i = 1$. Common to DiD, SCM.
  3. Non-negativity: $\omega_i \geq 0 \; \forall \; i$. Ensures uniqueness via 'coarse' regularisation + precision control. Negative weights may improve out-of-sample prediction.
  4. Constant Weights: $\omega_i = \bar{\omega} \; \forall \; i$

- DiD imposes 2-4;

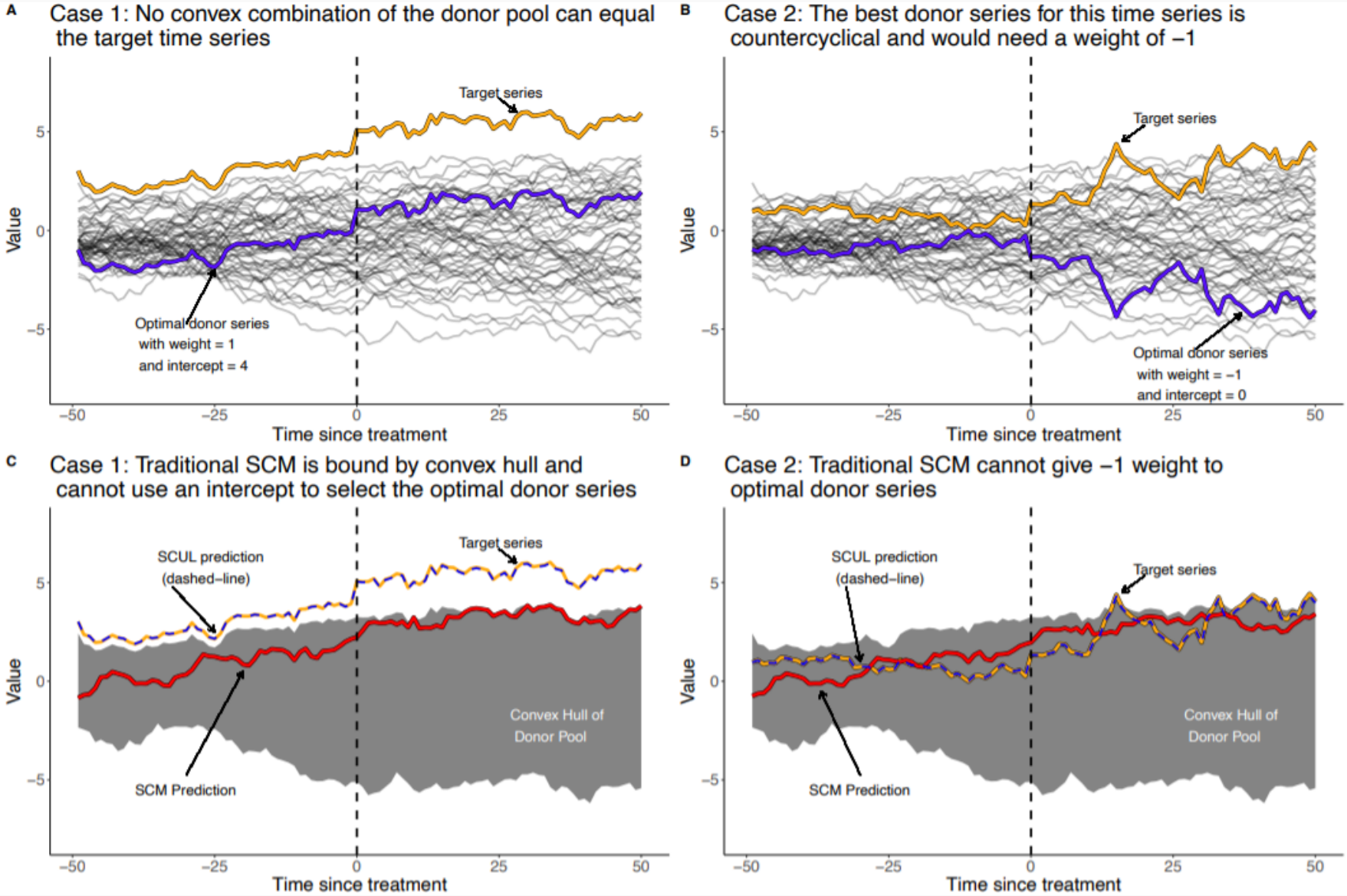
- SCM imposes 1-3: "convex hull" (no extrapolation)

Relaxing the Constraints

- Negative weights
  - If treated units are outliers on important covariates, allowing negative weights may improve fit
  - Bias reduction: negative weights increase bias-reduction rate
- When $N \gg T_0$, (1-3) alone might not result in a unique solution. Choose by
  - Matching on pre-treatment outcomes: one good control unit is better than synthetic one comprised of disparate units
  - Constant weights: implicit in DiD
- Given many pairs of (μ, ω), prefer values s.t.
  - Synthetic control unit is similar to treated units in terms of lagged outcomes
  - Low dispersion of weights
  - Few control units with non-zero weights (sparsity)

Case for Nonconvex or Negative Weights

Hollingsworth and Wing (2020)



Re-Expressing Existing Methods

DiD

- Assume (2-4)
- Fix $\omega^{\text{did}} = \frac{1}{N}$
- $\widehat{\mu}^{\text{did}} = \frac{1}{T_0}\sum_{s=1}^{T_0} Y_{0,s} - \frac{1}{NT_0}\sum_{s=1}^{T_0}\sum_{i=1}^{N} Y_{i,s}$

SCM

- Assume (1-3): convex hull
- For $M \times M$ PSD diagonal matrix $\mathbf{V}$

$$(\widehat{\omega}, \widehat{\mu}) = \underset{\omega,\mu}{\arg\min}\{(\mathbf{X}_t - \mu - \omega'\mathbf{X})'\mathbf{V}(\mathbf{X}_t - \mu - \omega'\mathbf{X})\}$$

$$\widehat{\mathbf{V}} = \underset{\mathbf{V}=\text{diag}(v_1,\ldots,v_M)}{\arg\min}\{(\mathbf{Y}_{\text{t, pre}} - \widehat{\omega}'\mathbf{Y}_{\text{c, pre}})'(\mathbf{Y}_{\text{t, pre}} - \widehat{\omega}'\mathbf{Y}_{\text{c, pre}})\}$$

Constrained Regression

- Assume (1-3): convex hull
- Controls as regressors
- Solve

$$\hat{\omega}^{constr} = \arg\min_{\omega} \sum_{t \leq T_0} \left(Y_{1t} - \omega' Y_{\mathcal{C}t} \right)^2$$
$$s.t. \sum_{i \in \mathcal{C}} \omega_i = 1 \text{ and } \omega_i \geq 0, \forall i \in \mathcal{C}$$

- Limitation: $T_0 > N$

Best Subset

Hsiao et al. (2012)

- Controls as regressors
- Bottom-up approach: search for the best 1, then the best 2, then the best 3 ... (greedy)
- Optimize the weights after taking out the intercepts

$$\left(\hat{\mu}^{subset}, \hat{\omega}^{subset}\right) = \arg\min_{\mu,\omega} \sum_{t \leq T_0} (Y_{1t} - \mu - \omega' Y_{Ct})^2$$

$$s.t. \sum_{i \in C} 1_{\omega_i \neq 0} \leq k$$

- Weights can be negative and do not need to add up to 1

The Optimization Problem

Ingredients of objective function

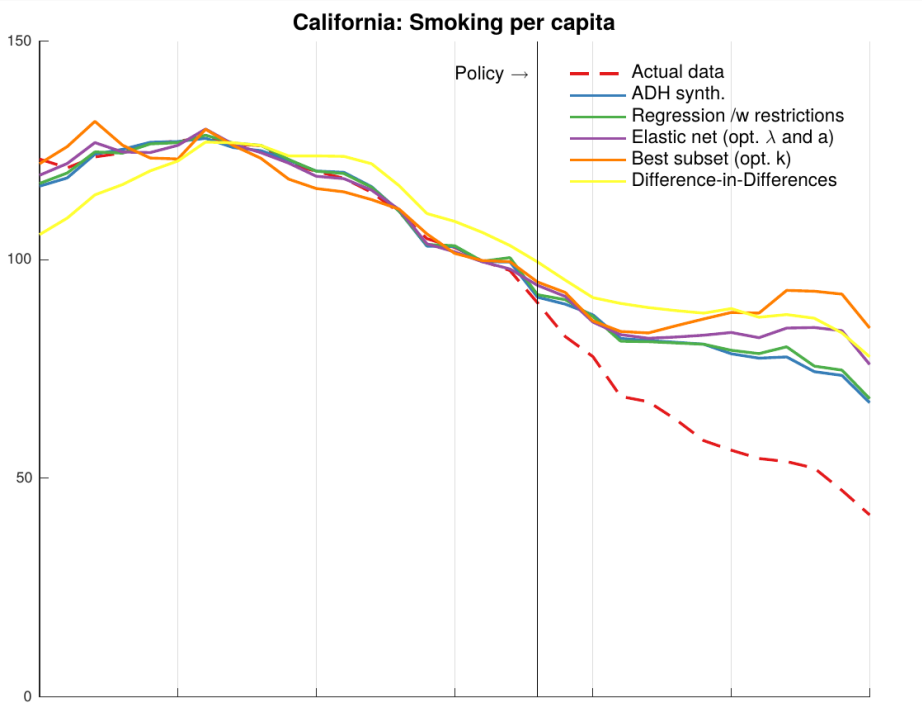
- Balance: difference between pre-treatment outcomes for treated and linear-combination of pre-treatment outcomes for control
  - $\|\mathbf{Y}_{\text{t, pre}} - \mu - \omega'\mathbf{Y}_{\text{c, pre}}\|_2^2 = (\mathbf{Y}_{\text{t, pre}} - \mu - \omega'\mathbf{Y}_{\text{c, pre}})'(\mathbf{Y}_{\text{t, pre}} - \mu - \omega'\mathbf{Y}_{\text{c, pre}})$
- Sparse and small weights:

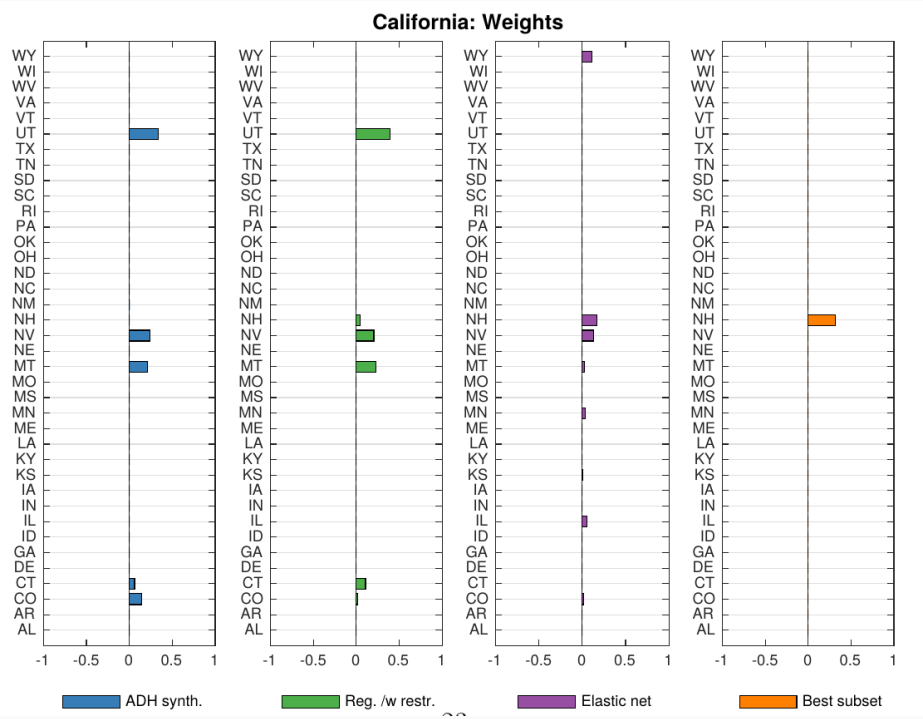
$$(\widehat{\mu}^{en}, \widehat{\omega}^{en}) = \arg\min_{\mu,\omega} \; Q(\mu, \omega | \mathbf{Y}_{\text{t, pre}}, \mathbf{Y}_{\text{c, pre}}; \lambda, \alpha)$$

where $$Q(\mu, \omega | \mathbf{Y}_{\text{t, pre}}, \mathbf{Y}_{\text{c, pre}}; \lambda, \alpha) = \|\mathbf{Y}_{\text{t, pre}} - \mu - \omega'\mathbf{Y}_{\text{c, pre}}\|_2^2 + \lambda\left(\frac{1-\alpha}{2}\|\omega\|_2^2 + \alpha\|\omega\|_1\right)$$

  - sparsity : $\|\omega\|_1$
  - magnitude: $\|\omega\|_2$
  - tuning parameters: λ, α

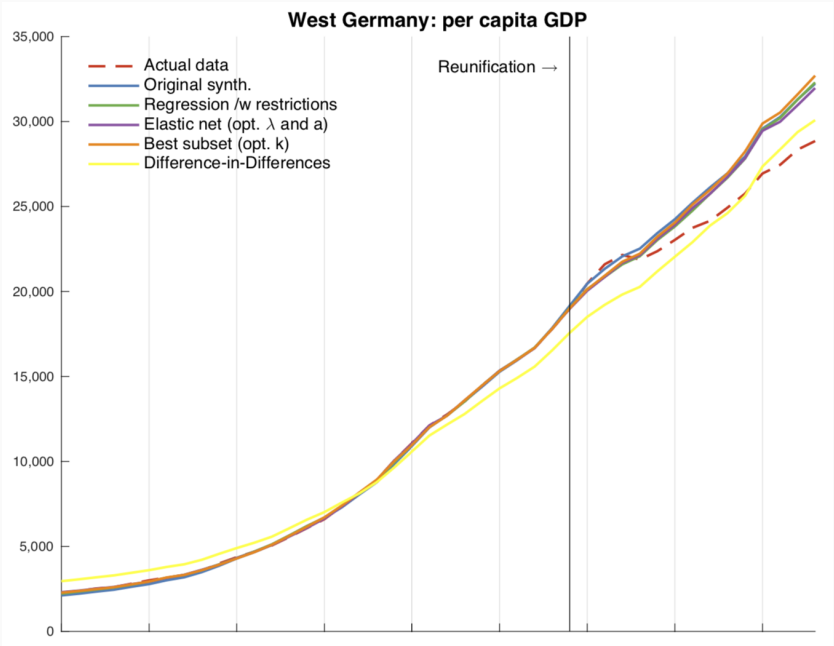
Revisiting ADH California smoking example

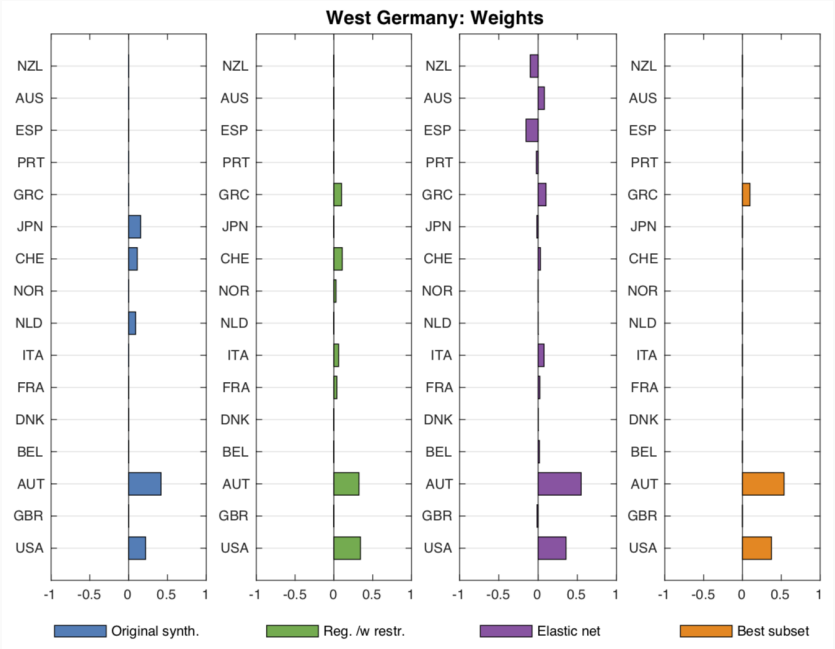




Model	$\sum_i \omega_i$	μ	$\widehat{\tau}$	s.e.
Original Synth	1	0	-22.1	16.1
Constrained	1	0	-22.9	12.8
Elastic Net	.55	18.5	-26.9	16.8
Best Subset	.32	37.6	-31.9	20.3
Diff-in-Diff	1	-14.4	-32.4	18.9

Example: German Reunification





What's Missing? The Balancing Approach

Robbins et al. (2017)

- SCM can be reinterpreted as an algorithm to achieve mean-balancing

$$\min_{\mathbf{w}_\mathcal{C}} L(\mathbf{w}_\mathcal{C})$$
$$s.t. \sum_{i \in \mathcal{T}} q_i \mathbf{Y}_{i,pre} = \sum_{j \in \mathcal{C}} w_j \mathbf{Y}_{j,pre}$$

  in which q_i is the base weight for treated unit i

- A popular choice for the loss function is the opposite of entropy, which is the Kullback-Leibler divergence between the distributions of the base weights and solution weights:

$$L(\mathbf{w}_\mathcal{C}) = -\sum_{i \in \mathcal{C}} w_i log(w_i/q_i)$$

- We will revisit this idea in Lecture 4

Methods Comparison

	SCM	CstrReg	Hsiao	ElasNet	IPW	Bal
Intercept shift			✓	✓	✓	(✓)
Weights add up to 1	✓	✓				✓
Non-negative weights	✓	✓			✓	✓
Allow short T_0	✓					✓
Multiple treated units					✓	✓
Computational efficiency		✓	✓	✓	✓	✓

- Alternatively, we can adopt an outcome model-based approach
- Recall that ADH (2010) use a factor-augmented model to motivate the SCM: $Y_{it}(0) = \theta_t' Z_i + \xi_t + \lambda_i' f_t + \varepsilon_{it}$
- What if we take the model more seriously? Next lecture.

References

- Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010). *Journal of the American Statistical Association*. June 1, 2010, 105(490): 493–505.
- Athey, Susan, and Guido Imbens. 2016. "The State of Applied Econometrics - Causality and Policy Evaluation." arXiv [stat.ME]. arXiv. http://arxiv.org/abs/1607.00699.
- Hahn, Jinyong, and Ruoyao Shi. 2017. "Synthetic Control and Inference." *Econometrics* 5 (4): 52.
- Firpo, Sergio, and Vitor Possebom. 2018. "Synthetic Control Method: Inference, Sensitivity Analysis and Confidence Sets." *Journal of Causal Inference* 6 (2).
- Chernozhukov, Victor, Kaspar Wuthrich, and Yinchu Zhu. 2017. "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls." arXiv [econ.EM]. arXiv. http://arxiv.org/abs/1712.09089.
- Ferman, B., C. Pinto, and V. Possebom. 2020. "Cherry Picking with Synthetic Controls." *Journal of Policy Analysis and Management*.
- Cattaneo, Matias D., Yingjie Feng, and Rocio Titiunik. 2019. "Prediction Intervals for Synthetic Control Methods." arXiv [stat.ME]. arXiv. http://arxiv.org/abs/1912.07120.
- Doudchenko, Nikolay and Guido Imbens (2016). "Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis." Working Paper Stanford University.
- Hollingsworth, Alex, and Coady Wing. 2020. "Tactics for Design and Inference in Synthetic Control Studies: An Applied Example Using High-Dimensional Data." https://doi.org/10.2139/ssrn.3592088.
- Hsiao, Cheng, H. Steve Ching and Shui Ki Wan (2012). "A Panel Data Approach for Program Evaluation: Measuring the Benefits of Political and Economic Integration of Hong Kong with Mainland China," *Journal of Applied Econometrics*, Vol. 27, Iss. 5, August 2012, pp. 705–740.
- Robbins, Michael W., Jessica Saunders, and Beau Kilmer. 2017. "A Framework for Synthetic Control Methods With High-Dimensional, Micro-Level Data: Evaluating a Neighborhood-Specific Crime Intervention." *Journal of the American Statistical Association* 112 (517): 109–26.