Causal Inference

Mixtape Session

Roadmap

Randomization inference

- Lady tasting tea
- Fisher's sharp null
- Alternative test statistics

Comparative case studies

- Synthetic control
- Prison expansion example

Randomization inference and causal inference

- “In randomization-based inference, uncertainty in estimates arises naturally from the random assignment of the treatments, rather than from hypothesized sampling from a large population.” (Athey and Imbens 2017)
- Athey and Imbens is part of growing trend of economists using randomization-based methods for doing causal inference
- Unclear (to me) why we are hearing more and more about randomization inference, but we are.
- Could be due to improved computational power and/or the availability of large data instead of samples?

Lady tasting tea experiment

- Ronald Aylmer Fisher (1890-1962)
  - → Two classic books on statistics: *Statistical Methods for Research Workers* (1925) and *The Design of Experiments* (1935), as well as a famous work in genetics, *The Genetical Theory of Natural Science*
  - → Developed many fundamental notions of modern statistics including the theory of randomized experimental design.

Lady tasting tea

- Muriel Bristol (1888-1950)
  - → A PhD scientist back in the days when women weren't PhD scientists
  - → Worked with Fisher at the Rothamsted Experiment Station (which she established) in 1919
  - → During afternoon tea, Muriel claimed she could tell from taste whether the milk was added to the cup before or after the tea
  - → Scientists were incredulous, but Fisher was inspired by her strong claim
  - → He devised a way to test her claim which she passed using randomization inference

Description of the tea-tasting experiment

- Original claim: Given a cup of tea with milk, Bristol claims she can discriminate the order in which the milk and tea were added to the cup
- Experiment: To test her claim, Fisher prepares 8 cups of tea – 4 **milk then tea** and 4 **tea then milk** – and presents each cup to Bristol for a taste test
- Question: How many cups must Bristol correctly identify to convince us of her unusual ability to identify the order in which the milk was poured?
- Fisher's sharp null: Assume she can't discriminate. Then what's the likelihood that random chance was responsible for her answers?

Choosing subsets

- The lady performs the experiment by selecting 4 cups, say, the ones she claims to have had the tea poured first.

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

- "8 choose 4" – $\binom{8}{4}$ – ways to choose 4 cups out of 8
  - → Numerator is $8 \times 7 \times 6 \times 5 = 1,680$ ways to choose a first cup, a second cup, a third cup, and a fourth cup, in order.
  - → Denominator is $4 \times 3 \times 2 \times 1 = 24$ ways to order 4 cups.

Choosing subsets

- There are 70 ways to choose 4 cups out of 8, and therefore a 1.4% probability of producing the correct answer by chance

$$\frac{24}{1680} = 1/70 = 0.014.$$

- For example, the probability that she would correctly identify all 4 cups is $\frac{1}{70}$

Statistical significance

- Suppose the lady correctly identifies all 4 cups. Then . . .
  1. Either she has no ability, and has chosen the correct 4 cups purely by chance, or
  2. She has the discriminatory ability she claims.
- Since choosing correctly is highly unlikely in the first case (one chance in 70), the second seems plausible.
- Bristol actually got all four correct
- I wonder if seeing this, any of the scientists present changed their mind

Null hypothesis

- In this example, the null hypothesis is the hypothesis that the lady has no special ability to discriminate between the cups of tea.
- We can never prove the null hypothesis, but the data may provide evidence to reject it.
- In most situations, rejecting the null hypothesis is what we hope to do.

Null hypothesis of no effect

- Randomization inference allows us to make probability calculations revealing whether the treatment assignment was “unusual”
- Fisher’s sharp null is when entertain the possibility that no unit has a treatment effect
- This allows us to make “exact” p-values which do not depend on large sample approximations
- It also means the inference is not dependent on any particular distribution (e.g., Gaussian); sometimes called nonparametric

Sidebar: bootstrapping is different

- Sometimes people confuse randomization inference with bootstrapping
- Bootstrapping randomly draws a percent of the total observations for estimation; “uncertainty over the sample”
- Randomization inference randomly reassigns the treatment; “uncertainty over treatment assignment”

(Thanks to Jason Kerwin for helping frame the two against each other)

6-step guide to randomization inference

1. Choose a sharp null hypothesis (e.g., no treatment effects)
2. Calculate a test statistic (T is a scalar based on D and Y)
3. Then pick a randomized treatment vector $\tilde{D}_1$
4. Calculate the test statistic associated with $(\tilde{D}, Y)$
5. Repeat steps 3 and 4 for all possible combinations to get $\tilde{T} = \{\tilde{T}_1, \ldots, \tilde{T}_K\}$
6. Calculate exact p-value as $p = \frac{1}{K} \sum_{k=1}^{K} I(\tilde{T}_k \geq T)$

Pretend experiment

Table: Pretend DBT intervention for some homeless population

Name	D	Y	Y^0	Y^1
Andy	1	10	.	10
Ben	1	5	.	5
Chad	1	16	.	16
Daniel	1	3	.	3
Edith	0	5	5	.
Frank	0	7	7	.
George	0	8	8	.
Hank	0	10	10	.

For concreteness, assume a program where we pay homeless people $15 to take dialectical behavioral therapy (DBT). Outcomes are some measure of mental health 0-20 with higher scores being improvements in mental health symptoms.

Step 1: Sharp null of no effect

Fisher's Sharp Null Hypothesis

$$H_0 : \delta_i = Y_i^1 - Y_i^0 = 0 \; \forall i$$

- Assuming no effect means any test statistic is due to chance
- Neyman and Fisher test statistics were different – Fisher was exact, Neyman was not
- Neyman's null was no average treatment effect (ATE=0). If you have a treatment effect of 5 and I have a treatment effect of -5, our ATE is zero. This is not the sharp null even though it also implies a zero ATE

More sharp null

- Since under the Fisher sharp null $\delta_i = 0$, it means each unit's potential outcomes under both states of the world are the same
- We therefore know each unit's missing counterfactual
- The randomization we will perform will cycle through all treatment assignments under a null well treatment assignment doesn't matter because all treatment assignments are associated with a null of zero unit treatment effects
- We are looking for evidence *against* the null

Step 1: Fisher's sharp null and missing potential outcomes

Table: Missing potential outcomes are no longer missing

Name	D	Y	Y^0	Y^1
Andy	1	10	**10**	10
Ben	1	5	**5**	5
Chad	1	16	**16**	16
Daniel	1	3	**3**	3
Edith	0	5	5	**5**
Frank	0	7	7	**7**
George	0	8	8	**8**
Hank	0	10	10	**10**

Fisher sharp null allows us to **fill in** the missing counterfactuals bc under the null there's zero treatment effect at the unit level.

Step 2: Choosing a test statistic

Test Statistic

A test statistic $T(D, Y)$ is a scalar quantity calculated from the treatment assignments D and the observed outcomes Y

- By scalar, I just mean it's a number (vs. a function) measuring some relationship between D and Y
- Ultimately there are many tests to choose from; I'll review a few later
- If you want a test statistic with high statistical power, you need large values when the null is false, and small values when the null is true (i.e., *extreme*)

Simple difference in means

- Consider the absolute SDO from earlier

$$\delta_{SDO} = \left| \frac{1}{N_T} \sum_{i=1}^{N} D_i Y_i - \frac{1}{N_C} \sum_{i=1}^{N} (1 - D_i) Y_i \right|$$

- Larger values of δ_{SDO} are evidence *against* the sharp null
- Good estimator for constant, additive treatment effects and relatively few outliers in the potential outcomes

Step 2: Calculate test statistic, $T(D, Y)$

Table: Calculate T using D and Y

Name	D	Y	Y^0	Y^1	δ_i
Andy	**1**	**10**	10	10	0
Ben	**1**	**5**	5	5	0
Chad	**1**	**16**	16	16	0
Daniel	**1**	**3**	3	3	0
Edith	**0**	**5**	5	5	0
Frank	**0**	**7**	7	7	0
George	**0**	**8**	8	8	0
Hank	**0**	**10**	10	10	0

We'll start with this simple the simple difference in means test statistic, $T(D, Y)$: $\delta_{SDO} = 34/4 - 30/4 = 1$

Steps 3-5: Null randomization distribution

- Randomization steps reassign treatment assignment for every combination, calculating test statistics each time, to obtain the entire distribution of counterfactual test statistics
- The key insight of randomization inference is that under Fisher's sharp null, the treatment assignment shouldn't matter
- Ask yourself:
  - → if there is no unit level treatment effect, can you picture a distribution of counterfactual test statistics?
  - → and if there is no unit level treatment effect, what must average counterfactual test statistics equal?

Step 6: Calculate “exact” p-values

- Question: how often would we get a test statistic as big or bigger as our “real” one if Fisher’s sharp null was true?
- This can be calculated “easily” (sometimes) once we have the randomization distribution from steps 3-5
  - → The number of test statistics ($t(D, Y)$) bigger than the observed divided by total number of randomizations

$$Pr(T(D, Y) \geq T(\tilde{D}, Y|\delta = 0)) = \frac{\sum_{D \in \Omega} I(T(D, Y) \leq T(\tilde{D}, Y)}{K}$$

First permutation (holding N_T fixed)

Name	$\tilde{D}_2$	Y	Y^0	Y^1
Andy	1	10	10	10
Ben	0	5	5	5
Chad	1	16	16	16
Daniel	1	3	3	3
Edith	0	5	5	5
Frank	1	7	7	7
George	0	8	8	8
Hank	0	10	10	10

$$\tilde{T}_1 = |36/4 - 28/4| = 9 - 7 = 2$$

Second permutation (again holding N_T fixed)

Name	$\tilde{D}_3$	Y	Y^0	Y^1
Andy	1	10	10	10
Ben	0	5	5	5
Chad	1	16	16	16
Daniel	1	3	3	3
Edith	0	5	5	5
Frank	0	7	7	7
George	1	8	8	8
Hank	0	10	10	10

$$T_{rank} = |36/4 - 27/4| = 9 - 6.75 = 2.25$$

Sidebar: Should it be 4 treatment groups each time?

- In this experiment, I've been using the same N_T *under the assumption* that N_T had been fixed when the experiment was drawn.
- But if the original treatment assignment had been generated by something like a Bernoulli distribution (e.g., coin flips over every unit), then you should be doing a complete permutation that is also random in this way
- This means that for 8 units, sometimes you'd have 1 treated, or even 8
- Correct inference requires you know the original data generating process

Randomization distribution

Assignment	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	$\lvert T_i \rvert$
True D	1	1	1	1	0	0	0	0	1
$\tilde{D}_2$	1	0	1	1	0	1	0	0	2
$\tilde{D}_3$	1	0	1	1	0	0	1	0	2.25
...									

Step 2: Other test statistics

- The simple difference in means is fine when effects are additive, and there are few outliers in the data
- But outliers create more variation in the randomization distribution
- A good test statistic is the one that best fits your data.
- Some test statistics will have weird properties in the randomization as we'll see in synthetic control.
- What are some alternative test statistics?

Transformations

- What if there was a constant multiplicative effect: $Y_i^1 / Y_i^0 = C$?
- Difference in means will have low power to detect this alternative hypothesis
- So we transform the observed outcome using the natural log:

$$T_{log} = \left| \frac{1}{N_T} \sum_{i=1}^{N} D_i ln(Y_i) - \frac{1}{N_C} \sum_{i=1}^{N} (1 - D_i) ln(Y_i) \right|$$

- This is useful for skewed distributions of outcomes

Difference in medians/quantiles

- We can protect against outliers using other test statistics such as the difference in quantiles
- Difference in medians:

$$T_{median} = |median(Y_T) - median(Y_C)|$$

- We could also estimate the difference in quantiles at any point in the distribution (e.g., 25th or 75th quantile)

Rank test statistics

- Basic idea is rank the outcomes (higher values of Y_i are assigned higher ranks)
- Then calculate a test statistic based on the transformed ranked outcome (e.g., mean rank)
- Useful with continuous outcomes, small datasets and/or many outliers

Rank statistics formally

- Rank is the domination of others (including oneself):

$$\tilde{R} = \tilde{R}_i(Y_1, \ldots, Y_N) = \sum_{j=1}^{N} I(Y_j \leq Y_i)$$

- Normalize the ranks to have mean 0

$$\tilde{R}_i = \tilde{R}_i(Y_1, \ldots, Y_N) = \sum_{j=1}^{N} I(Y_j \leq Y_i) - \frac{N+1}{2}$$

- Calculate the absolute difference in average ranks:

$$T_{rank} = |\overline{R}_T - \overline{R}_C| = \left| \frac{\sum_{i:D_i=1} R_i}{N_T} - \frac{\sum_{i:D_i=0} R_i}{N_C} \right|$$

- Minor adjustment (averages) for ties

Randomization distribution

Name	D	Y	Y^0	Y^1	Rank	R_i
Andy	1	10	**10**	10	6.5	2
Ben	1	5	**5**	5	2.5	-2
Chad	1	16	**16**	16	8	3.5
Daniel	1	3	**3**	3	1	-3.5
Edith	0	5	5	**5**	2.5	-2
Frank	0	7	7	**7**	4	-0.5
George	0	8	8	**8**	5	0.5
Hank	0	10	10	**10**	6.5	2

$$T_{rank} = |0 - 0| = 0$$

Effects on outcome distributions

- Focused so far on “average” differences between groups.
- Kolmogorov-Smirnov test statistics is based on the difference in the distribution of outcomes
- Empirical cumulative distribution function (eCDF):

$$\widehat{F}_C(Y) = \frac{1}{N_C} \sum_{i:D_i=0} 1(Y_i \leq Y)$$

$$\widehat{F}_T(Y) = \frac{1}{N_T} \sum_{i:D_i=1} 1(Y_i \leq Y)$$

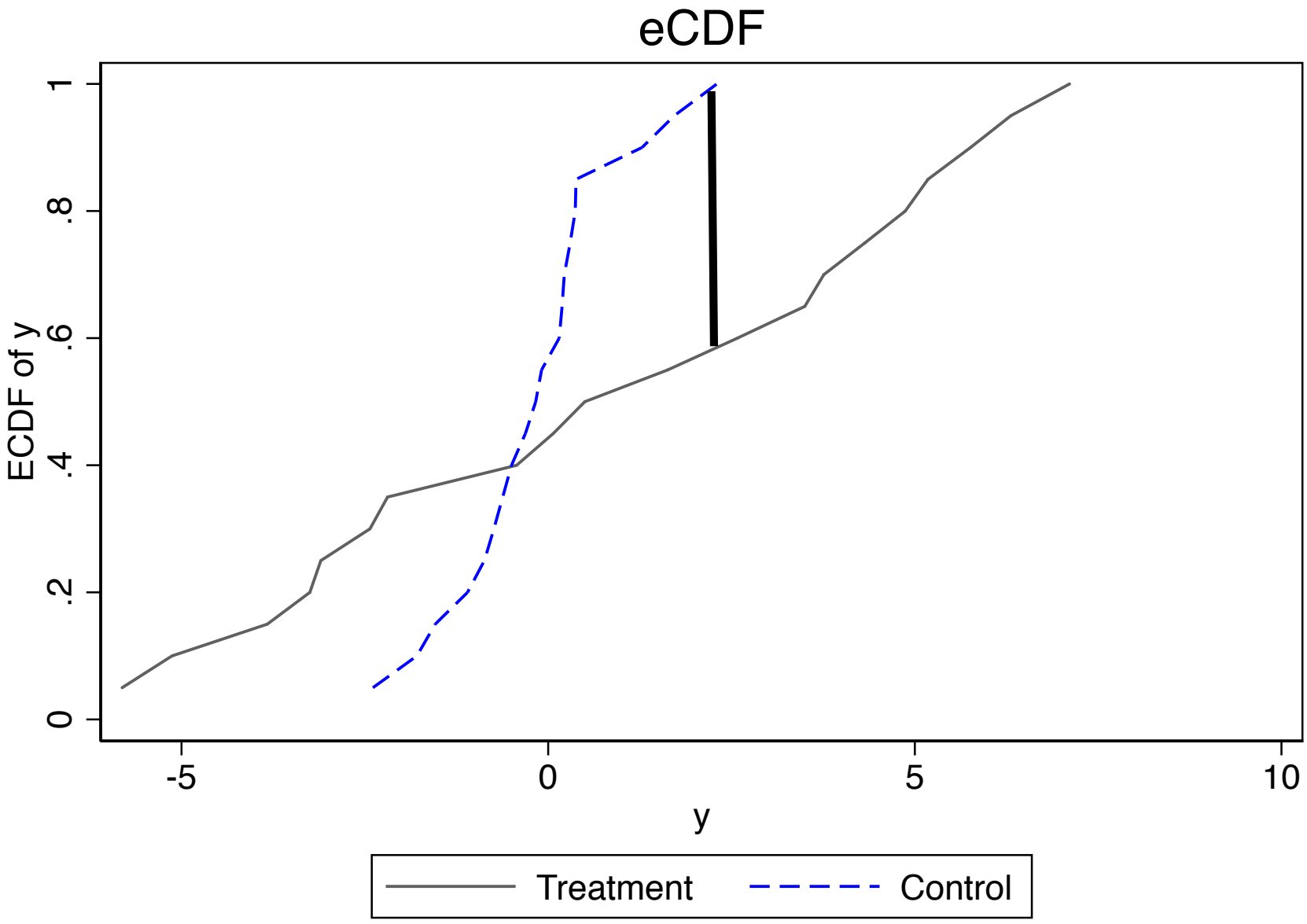
- Proportion of observed outcomes below a chosen value for treated and control separately
- If two distributions are the same, then $\widehat{F}_C(Y) = \widehat{F}_T(Y)$

Kolmogorov-Smirnov statistic

- Test statistics are scalars not functions
- eCDFs are functions, not scalars
- Solution: use the maximum discrepancy between the two eCDFs:

$$T_{KS} = max|\widehat{F}_T(Y_i) - \widehat{F}_C(Y_i)|$$

eCDFs by treatment status and test statistic



Small vs. Modest Sample Sizes are non-trivial

Computing the exact randomization distribution is not always feasible (Wolfram Alpha)

- $N = 6$ and $N_T = 3$ gives us 20 assignment vectors
- $N = 8$ and $N_T = 4$ gives us 70 assignment vectors
- $N = 10$ and $N_T = 5$ gives us 252 assignment vectors
- $N = 20$ and $N_T = 10$ gives us 184,756 assignment vectors
- $N = 50$ and $N_T = 25$ gives us 1.2641061×10^{14} assignment vectors

Exact p calculations are not realistic bc the number of assignments explodes at even modest size

Approximate p-values

These have been “exact” tests when they use every possible combination of D

- When you can’t use every combination, then you can get *approximate* p-values from a simulation (TBD)
- With a rejection threshold of α (e.g., 0.05), randomization inference test will falsely reject less than $100 \times \alpha\%$ of the time

Approximate p values

- Use simulation to get approximate p-values
  - $\rightarrow$ Take K samples from the treatment assignment space
  - $\rightarrow$ Calculate the randomization distribution in the K samples
  - $\rightarrow$ Tests no longer exact, but bias is under your control (increase K)
- Imbens and Rubin show that p values converge to stable p values pretty quickly (in their example after 1000 replications)

Thornton's experiment

ATE	Iteration	Rank	p	no. trials
0.45	1	1	0.01	100
0.45	1	1	0.002	500
0.45	1	1	0.001	1000

Table: Estimated p-value using different number of trials.

Including covariate information

- Let X_i be a pretreatment measure of the outcome
- One way is to use this as a gain score: $Y^{d'} = Y_i^d - X_i$
- Causal effects are the same $Y^{1i} - Y^{0i} = Y_i^1 - Y_i^0$
- But the test statistic is different:

$$T_{gain} = \left| (\overline{Y}_T - \overline{Y}_C) - (\overline{X}_T - \overline{X}_C) \right|$$

- If X_i is strongly predictive of Y_i^0, then this could have higher power
  - $\rightarrow$ Y_{gain} will have lower variance under the null
  - $\rightarrow$ This makes it easier to detect smaller effects

Regression in RI

- We can extend this to use covariates in more complicated ways
- For instance, we can use an OLS regression:

$$Y_i = \alpha + \delta D_i + \beta X_i + \varepsilon$$

- Then our test statistic could be $T_{OLS} = \widehat{\delta}$
- RI is justified even if the model is wrong
  - → OLS is just another way to generate a test statistic
  - → The more the model is "right" (read: predictive of Y_i^0), the higher the power T_{OLS} will have
- See if you can do this in Thornton's dataset using the loops and saving the OLS coefficient (or just use `ritest`)

Concluding remarks

- Randomization inference is very common, particularly useful you don’t want to make strong assumptions (parametric free)
- We’ll now explore its use in a popular observational method – the synthetic control

Roadmap

Randomization inference
- Lady tasting tea
- Fisher's sharp null
- Alternative test statistics

Comparative case studies
- Synthetic control
- Prison expansion example

What is synthetic control

- Synthetic control has been called the most important innovation in causal inference of the last 15 years (Athey and Imbens 2017)
- It's extremely useful for case studies, which is nice because that's often all we have
- Continues to also be methodologically a frontier for applied econometrics
- Consider this talk a starting point for you

What is a comparative case study

- Single treated unit – country, state, whatever
- Social scientists tackle such situations in two ways: qualitatively and quantitatively
- In political science, probably others, you see as a stark dividing line between camps
- Not so much in economics

Qualitative comparative case studies

- In qualitative comparative case studies, the goal is to reason *inductively* the causal effects of events or characteristics of a single unit on some outcome, oftentimes through logic and historical analysis.
  - → May not answer the causal questions at all because there is rarely a counterfactual, or if so, it's ad hoc.
  - → Classic example of comparative case study approach is Alexis de Toqueville's Democracy in America (but he is regularly comparing the US to France)

Traditional quantitative comparative case studies

- Quantitative comparative case studies are often explicitly causal designs.
- Usually a natural experiment applied to a single aggregate unit (e.g., city, school, firm, state, country)
- Method compares the evolution of an aggregate outcome for the unit affected by the intervention to the evolution of the same *ad hoc* aggregate control group (Card 1990; Card and Krueger 1994)

Pros and cons of traditional case study approaches

- Pros:
  - → Policy interventions often take place at an aggregate level
  - → Aggregate/macro data are often available
- Cons:
  - → Selection of control group is *ad hoc*
  - → Standard errors do not reflect uncertainty about the ability of the control group to reproduce the counterfactual of interest


EL DORADO


ELDORADO


SILVER SEA

Description of the Mariel Boatlift

- How do inflows of immigrants affect the wages and employment of natives in local labor markets?
- Card (1990) uses the Mariel Boatlift of 1980 as a natural experiment to measure the effect of a sudden influx of immigrants on unemployment among less-skilled natives
- The Mariel Boatlift increased the Miami labor force by 7%
- Individual-level data on unemployment from the Current Population Survey (CPS) for Miami and four comparison cities (Atlanta, Los Angeles, Houston, Tampa-St. Petersburg)

Why these four?

Tables 3 and 4 present simple averages of wage rates and unemployment rates for whites, blacks, Cubans, and other Hispanics in the Miami labor market between 1979 and 1985. For comparative purposes, I have assembled similar data for whites, blacks, and Hispanics in four other cities: Atlanta, Los Angeles, Houston, and Tampa–St. Petersburg. These four cities were selected both because they had relatively large populations of blacks and Hispanics and because they exhibited a pattern of economic growth similar to that in Miami over the late 1970s and early 1980s. A comparison of employment growth rates (based on establishment-level data) suggests that economic conditions were very similar in Miami and the average of the four comparison cities between 1976 and 1984.

Card's main results

Differences-in-differences estimates of the effect of immigration on unemployment[a]

	Group	Year		
		1979 (1)	1981 (2)	1981–1979 (3)
	Whites			
(1)	Miami	5.1 (1.1)	3.9 (0.9)	− 1.2 (1.4)
(2)	Comparison cities	4.4 (0.3)	4.3 (0.3)	− 0.1 (0.4)
(3)	Difference Miami-comparison	0.7 (1.1)	− 0.4 (0.95)	− 1.1 (1.5)
	Blacks			
(4)	Miami	8.3 (1.7)	9.6 (1.8)	1.3 (2.5)
(5)	Comparison cities	10.3 (0.8)	12.6 (0.9)	2.3 (1.2)
(6)	Difference Miami-comparison	− 2.0 (1.9)	− 3.0 (2.0)	− 1.0 (2.8)

[a] Notes: Adapted from Card (1990, Tables 3 and 6). Standard errors are shown in parentheses.

Can this ever lead to subjective biases?

- Card found that the Mariel boatlift reduced unemployment *compared to the four cities he chose*
- Is there anything principled we could do that doesn't give the researcher so much control over control group?
- Enter synthetic control (Abadie and Gardeazabal 2003; Abadie, Diamond and Hainmueller 2010)

Synthetic Control

- First appears in Abadie and Gardeazabal (2003) in a study of a terrorist attack in Spain (Basque) on GDP
- Revisited again in a 2011 JASA with Diamond and Hainmueller, two political scientists who were PhD students at Harvard (more proofs and inference)
- A combination of comparison units often does a better job reproducing the characteristics of a treated unit than single comparison unit alone

Researcher's objectives

- Our goal here is to reproduce the counterfactual of a treated unit by finding the combination of untreated units that best resembles the treated unit *before* the intervention in terms of the values of k relevant covariates (predictors of the outcome of interest)
- Method selects *weighted average of all potential comparison units* that best resembles the characteristics of the treated unit(s) - called the "synthetic control"

Synthetic control method: advantages

- Precludes extrapolation (unlike regression) because counterfactual forms a convex hull
- Does not require access to post-treatment outcomes in the “design” phase of the study - no peeking
- Makes explicit the contribution of each comparison unit to the counterfactual
- Formalizing the way comparison units are chosen has direct implications for inference

Synthetic control method: disadvantages

1. Subjective researcher bias kicked down to the model selection stage
2. Significant diversity at the moment as to how to principally select models - from machine learning to modifications - as well as estimation and software

Furman and Pinto (2018) recommend showing a few different results in their “cherry picking” JPAM

Synthetic control method: estimation

Suppose that we observe $J + 1$ units in periods $1, 2, \ldots, T$

- Unit "one" is exposed to the intervention of interest (that is, "treated") during periods $T_0 + 1, \ldots, T$
- The remaining J are an untreated reservoir of potential controls (a "donor pool")

Potential outcomes notation

- Let Y_{it}^0 be the outcome that would be observed for unit i at time t in the absence of the intervention
- Let Y_{it}^1 be the outcome that would be observed for unit i at time t if unit i is exposed to the intervention in periods $T_0 + 1$ to T.

Dynamic ATT

Treatment effect parameter is defined as dynamic ATT where

$$\begin{aligned} \delta_{1t} &= Y_{1t}^{1} - Y_{1t}^{0} \\ &= Y_{1t} - Y_{1t}^{0} \end{aligned}$$

for each post-treatment period, $t > T_0$ and Y_{1t} is the outcome for unit one at time t. We will estimate Y_{1t}^{0} using the J units in the donor pool

Estimating optimal weights

- Let $W = (w_2, \ldots, w_{J+1})'$ with $w_j \geq 0$ for $j = 2, \ldots, J+1$ and $w_2 + \cdots + w_{j+1} = 1$. Each value of W represents a potential synthetic control
- Let X_1 be a $(k \times 1)$ vector of pre-intervention characteristics for the treated unit. Similarly, let X_0 be a $(k \times J)$ matrix which contains the same variables for the unaffected units.
- The vector $W^* = (w_2^*, \ldots, w_{J+1}^*)'$ is chosen to minimize $||X_1 - X_0 W||$, subject to our weight constraints

Optimal weights differ by another weighting matrix

Abadie, et al. consider

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

where X_{jm} is the value of the m-th covariates for unit j and V is some $(k \times k)$ symmetric and positive semidefinite matrix

More on the V matrix

Typically, V is diagonal with main diagonal $v_1, \ldots, v_k$. Then, the synthetic control weights $w_2^*, \ldots, w_{J+1}^*$ minimize:

$$\sum_{m=1}^{k} v_m \left(X_{1m} - \sum_{j=2}^{J+1} w_j X_{jm} \right)^2$$

where v_m is a weight that reflects the relative importance that we assign to the m-th variable when we measure the discrepancy between the treated unit and the synthetic controls

Choice of V is critical

- The synthetic control $W^*(V^*)$ is meant to reproduce the behavior of the outcome variable for the treated unit in the absence of the treatment
- Therefore, the V^* weights directly shape W^*

Estimating the V matrix

Choice of $v_1, \ldots, v_k$ can be based on

- Assess the predictive power of the covariates using regression
- Subjectively assess the predictive power of each of the covariates, or calibration inspecting how different values for $v_1, \ldots, v_k$ affect the discrepancies between the treated unit and the synthetic control
- Minimize mean square prediction error (MSPE) for the pre-treatment period (default):

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

Cross validation

- Divide the pre-treatment period into an initial **training** period and a subsequent **validation** period
- For any given V, calculate $W^*(V)$ in the training period.
- Minimize the MSPE of $W^*(V)$ in the validation period

Suppose Y^0 is given by a factor model

What about unmeasured factors affecting the outcome variables as well as heterogeneity in the effect of observed and unobserved factors?

$$Y_{it}^0 = \alpha_t + \theta_t Z_i + \lambda_t u_i + \varepsilon_{it}$$

where α_t is an unknown common factor with constant factor loadings across units, and λ_t is a vector of unobserved common factors

With some manipulation

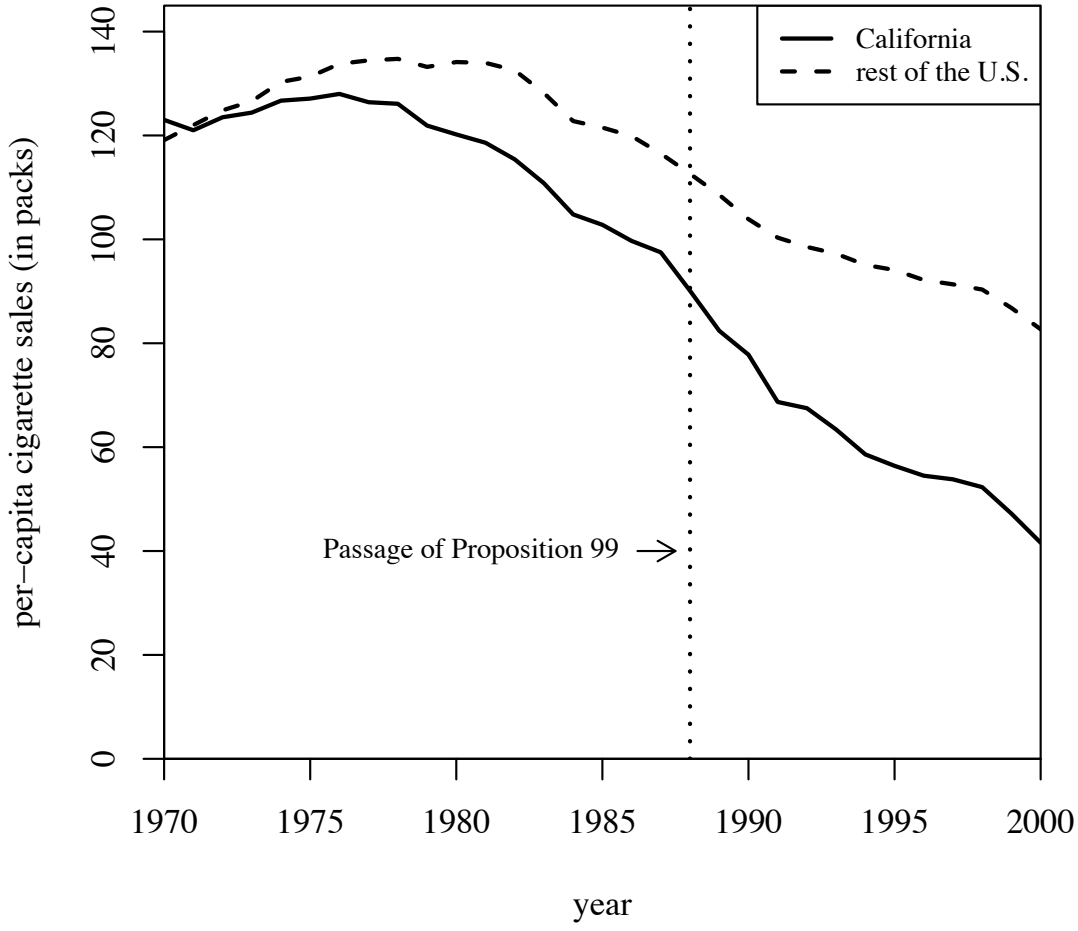
$$Y_{1t}^0 - \sum_{j=2}^{J+1} w_j^* Y_{jt} = \sum_{j=2}^{J+1} w_j^* \sum_{s=1}^{T_0} \lambda_t \left(\sum_{n=1}^{T_0} \lambda_n' \lambda_n \right)^{-1} \lambda_s' (\varepsilon_{js} - \varepsilon_{1s}) - \sum_{j=2}^{J+1} w_j^* (\varepsilon_{jt} - \varepsilon_{1t})$$

- If $\sum_{t=1}^{T_0} \lambda_t' \lambda_t$ is nonsingular, then RHS will be close to zero if number of preintervention periods is "large" relative to size of transitory shocks
- Only units that are alike in observables and unobservables should produce similar trajectories of the outcome variable over extended periods of time
- Proof in Appendix B of ADH (2011)

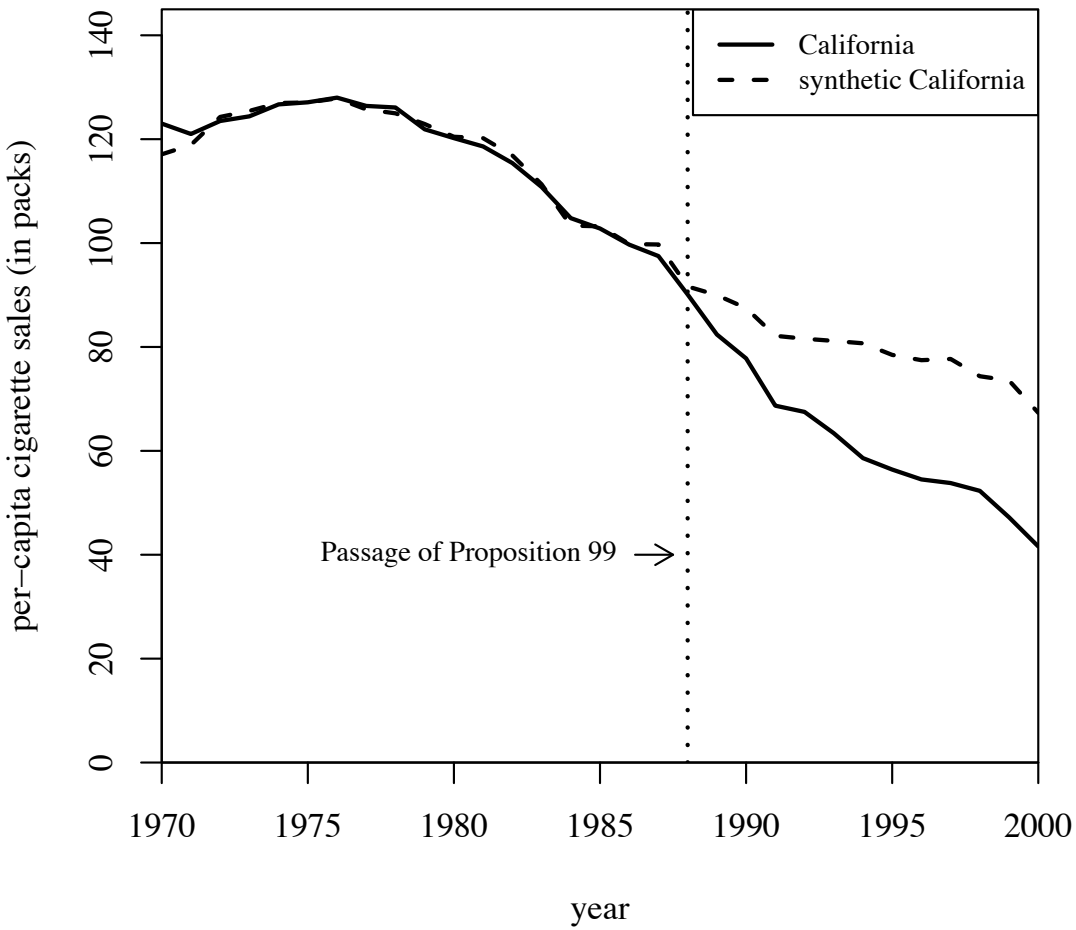
Example: California's Proposition 99

- In 1988, California first passed comprehensive tobacco control legislation:
  - → increased cigarette tax by 25 cents/pack
  - → earmarked tax revenues to health and anti-smoking budgets
  - → funded anti-smoking media campaigns
  - → spurred clean-air ordinances throughout the state
  - → produced more than $100 million per year in anti-tobacco projects
- Other states that subsequently passed control programs are excluded from donor pool of controls (AK, AZ, FL, HI, MA, MD, MI, NJ, OR, WA, DC)

Cigarette Consumption: CA and the Rest of the US



Cigarette Consumption: CA and synthetic CA

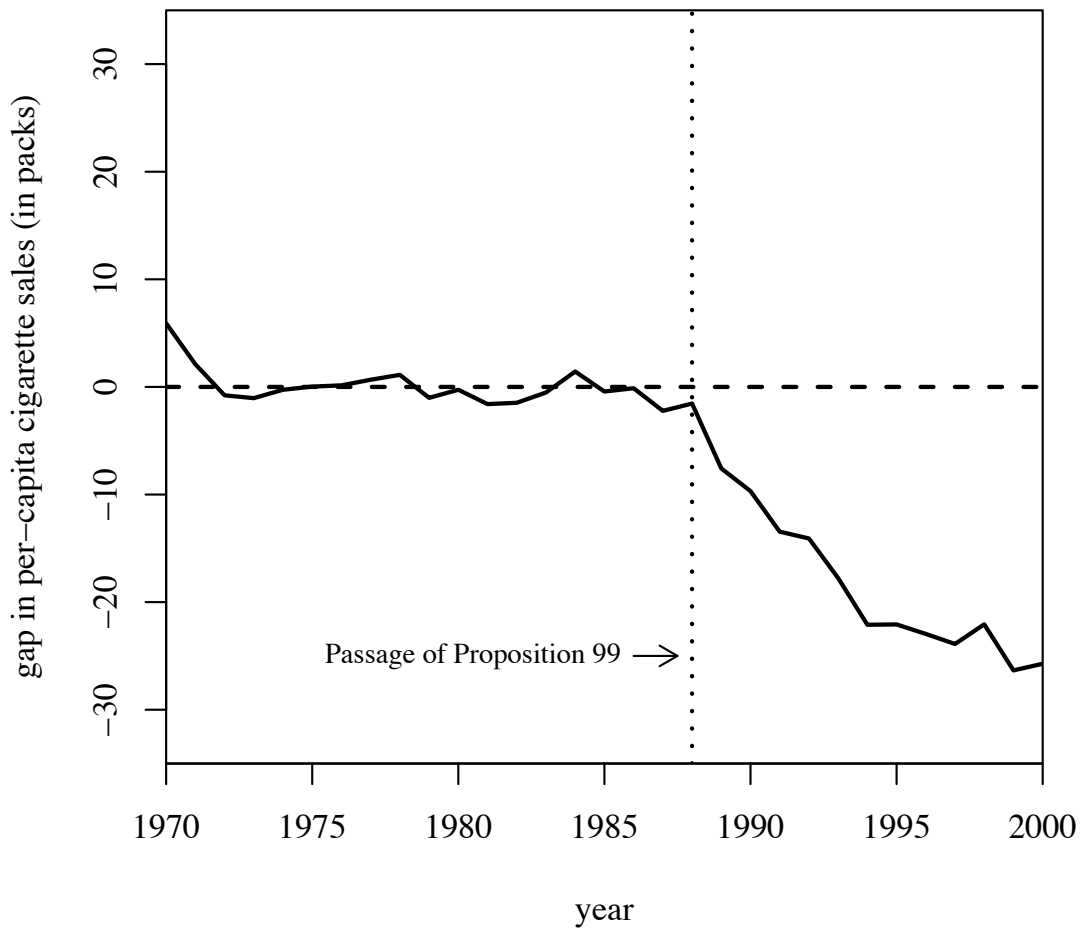


Predictor Means: Actual vs. Synthetic California

	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



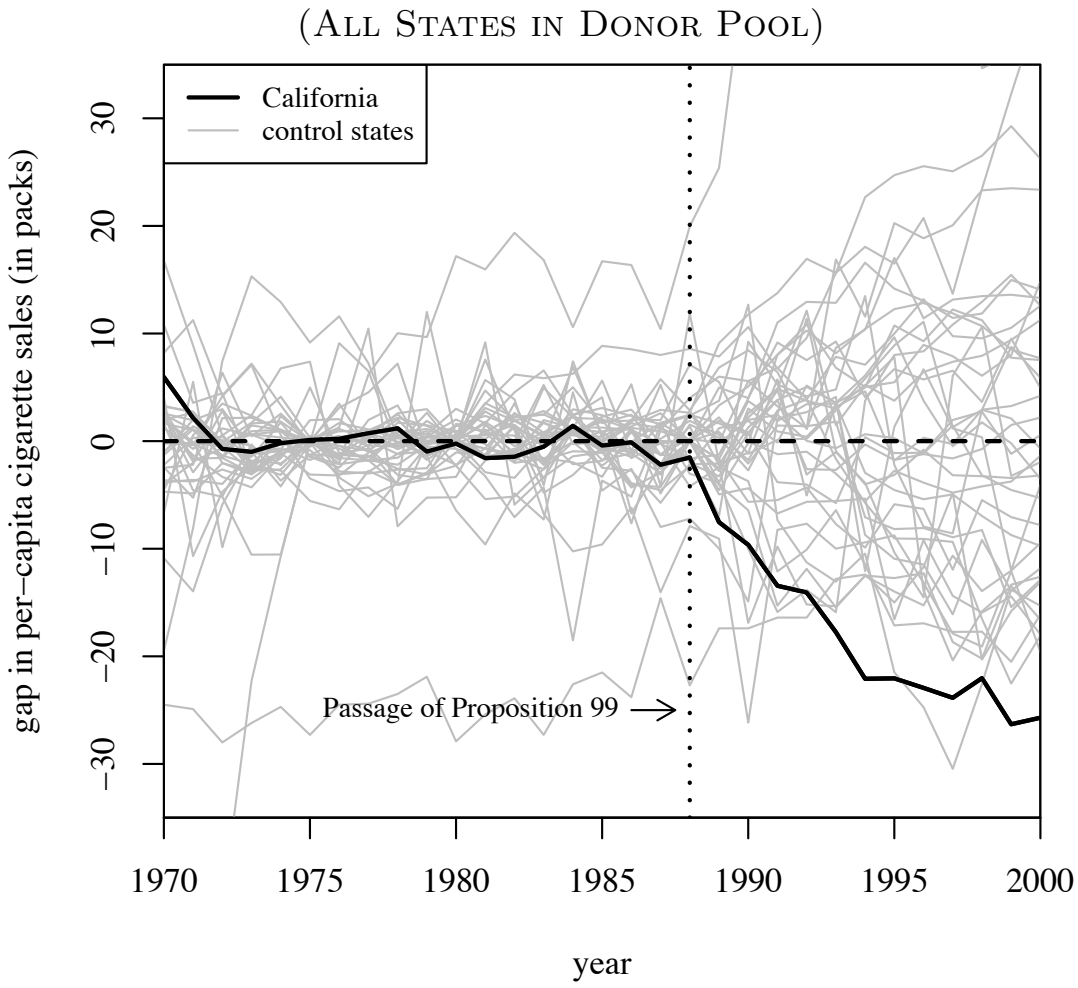
Inference

- To assess significance, we calculate exact p-values under Fisher's sharp null using a test statistic equal to after to before ratio of RMSPE
- Exact p-value method
  - → Iteratively apply the synthetic method to each country/state in the donor pool and obtain a distribution of placebo effects
  - → Compare the gap (RMSPE) for California to the distribution of the placebo gaps. For example the post-Prop. 99 RMSPE is:

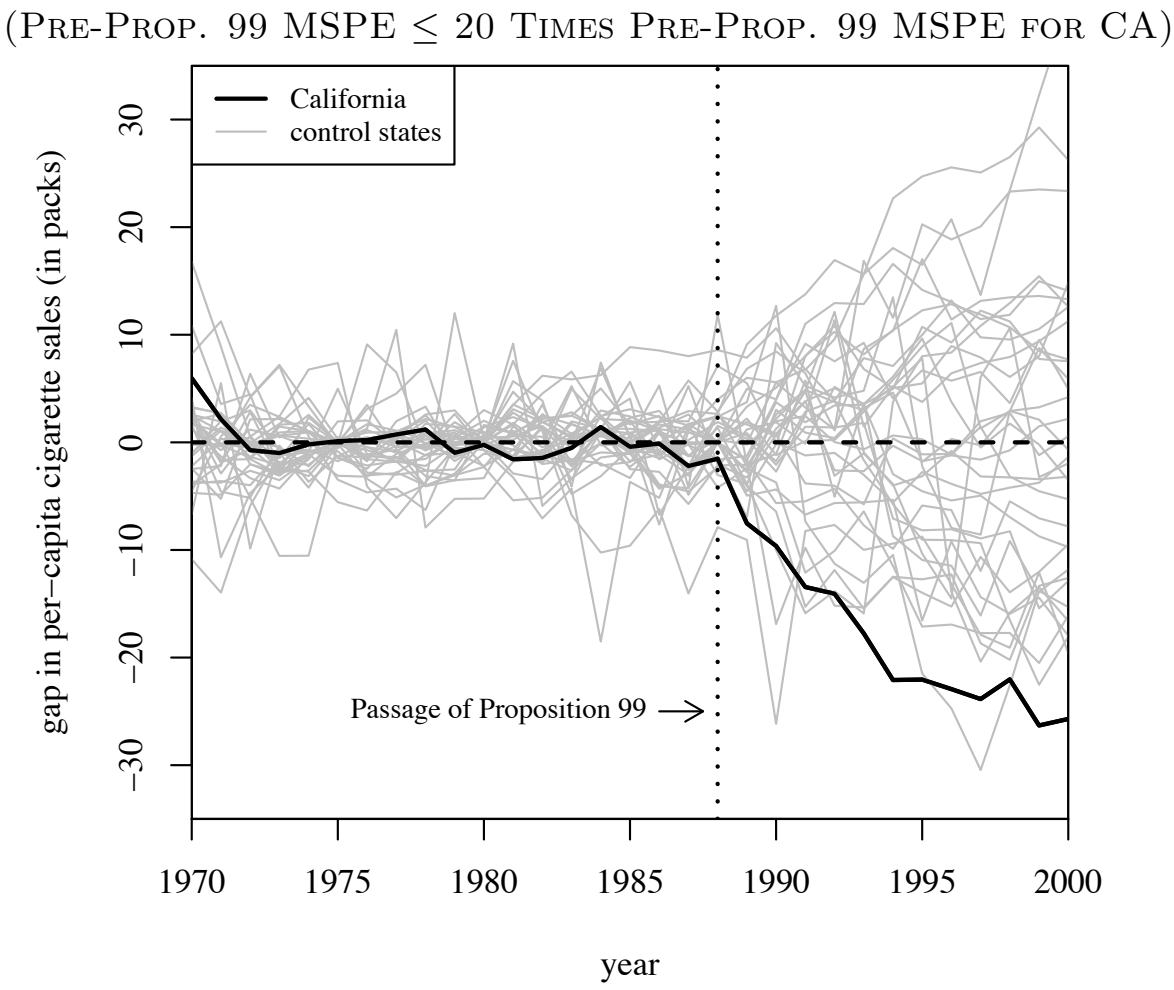
$$RMSPE = \left(\frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt} \right)^2 \right)^{\frac{1}{2}}$$

and the exact p-value is the treatment unit rank divided by J

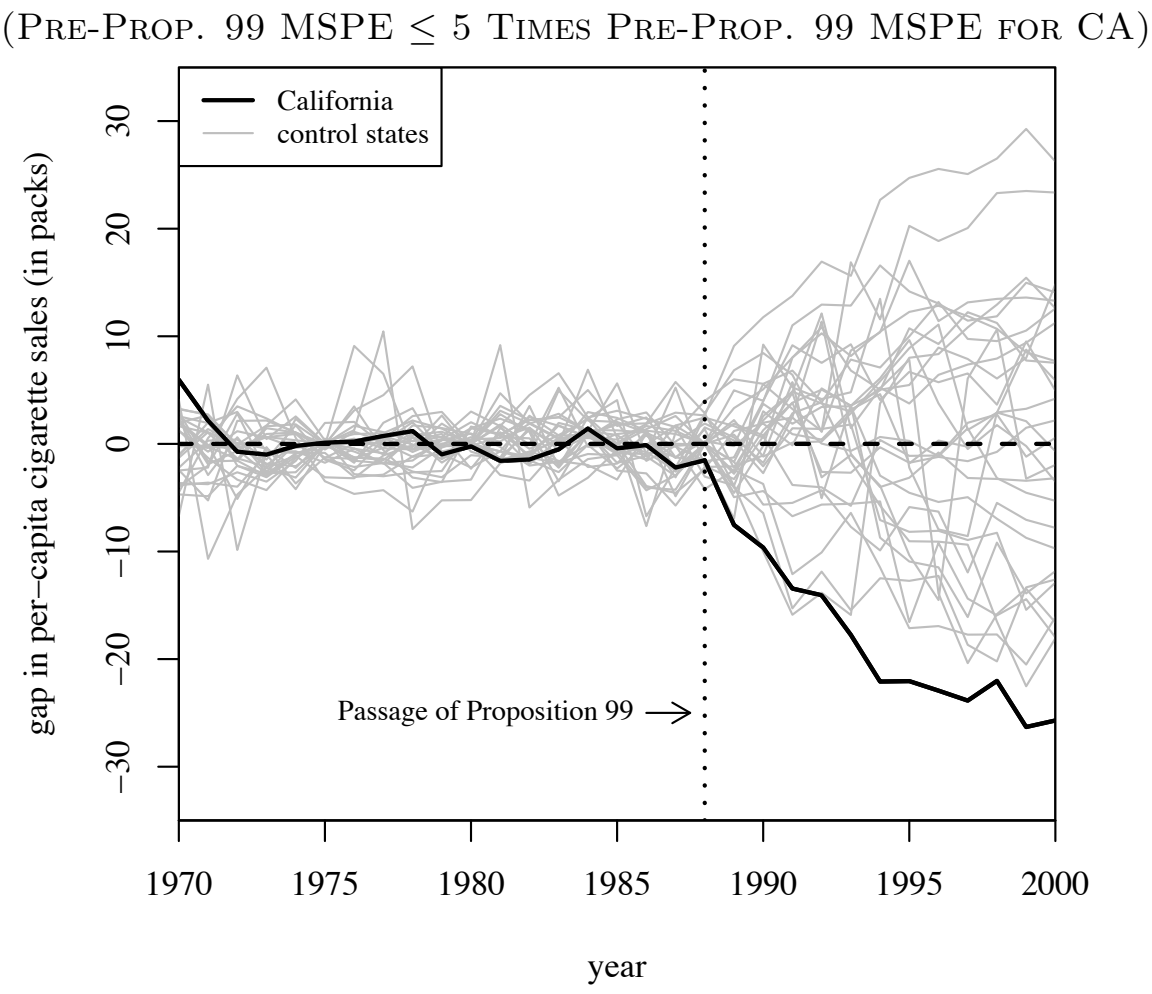
Smoking Gap for CA and 38 control states



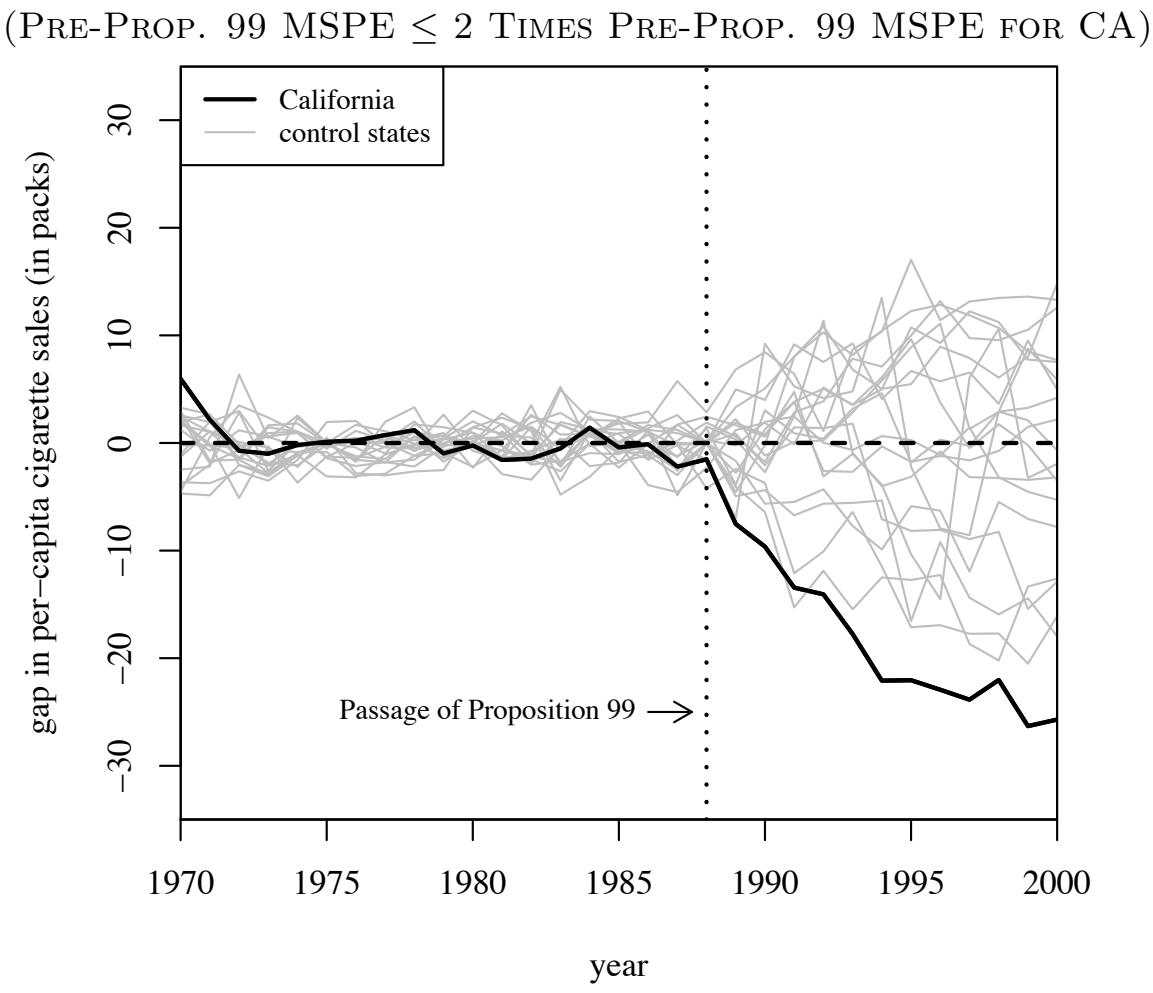
Smoking Gap for CA and 34 control states



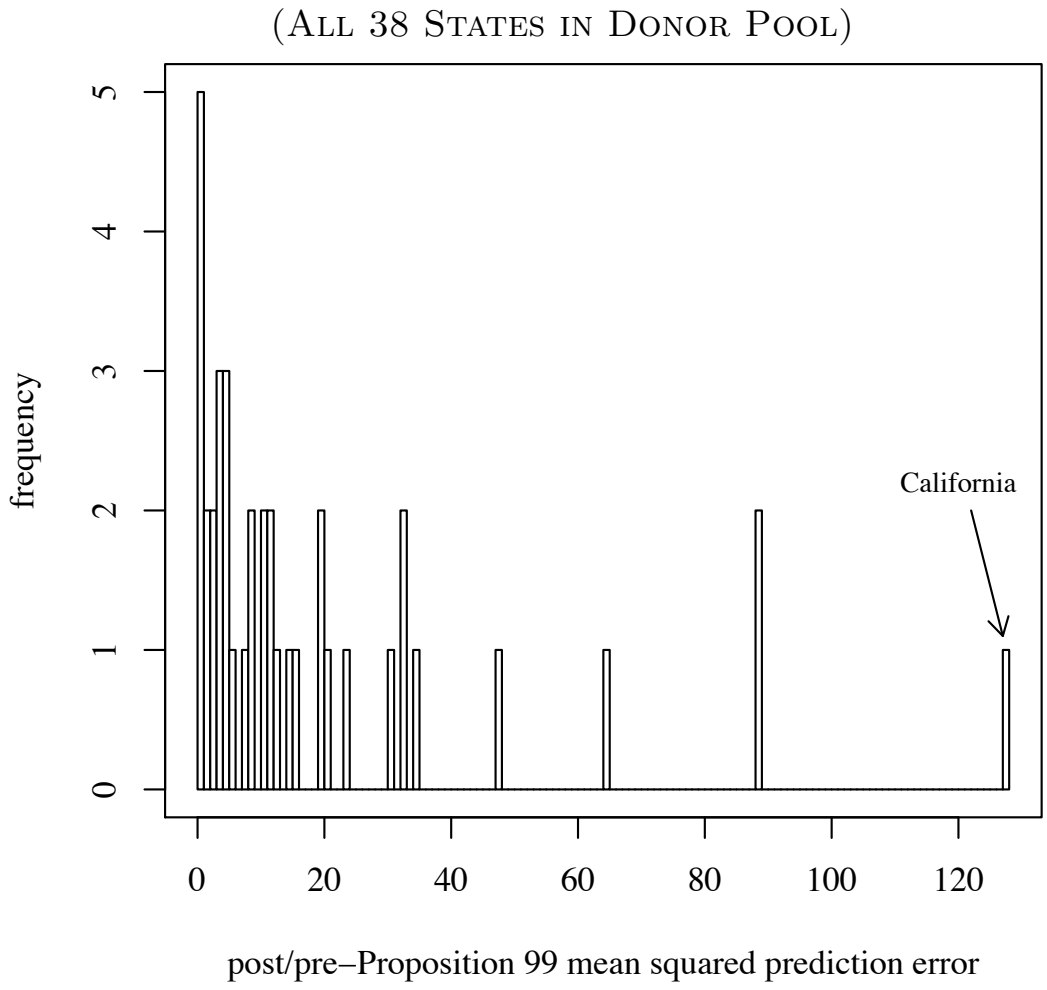
Smoking Gap for CA and 29 control states



Smoking Gap for CA and 19 control states



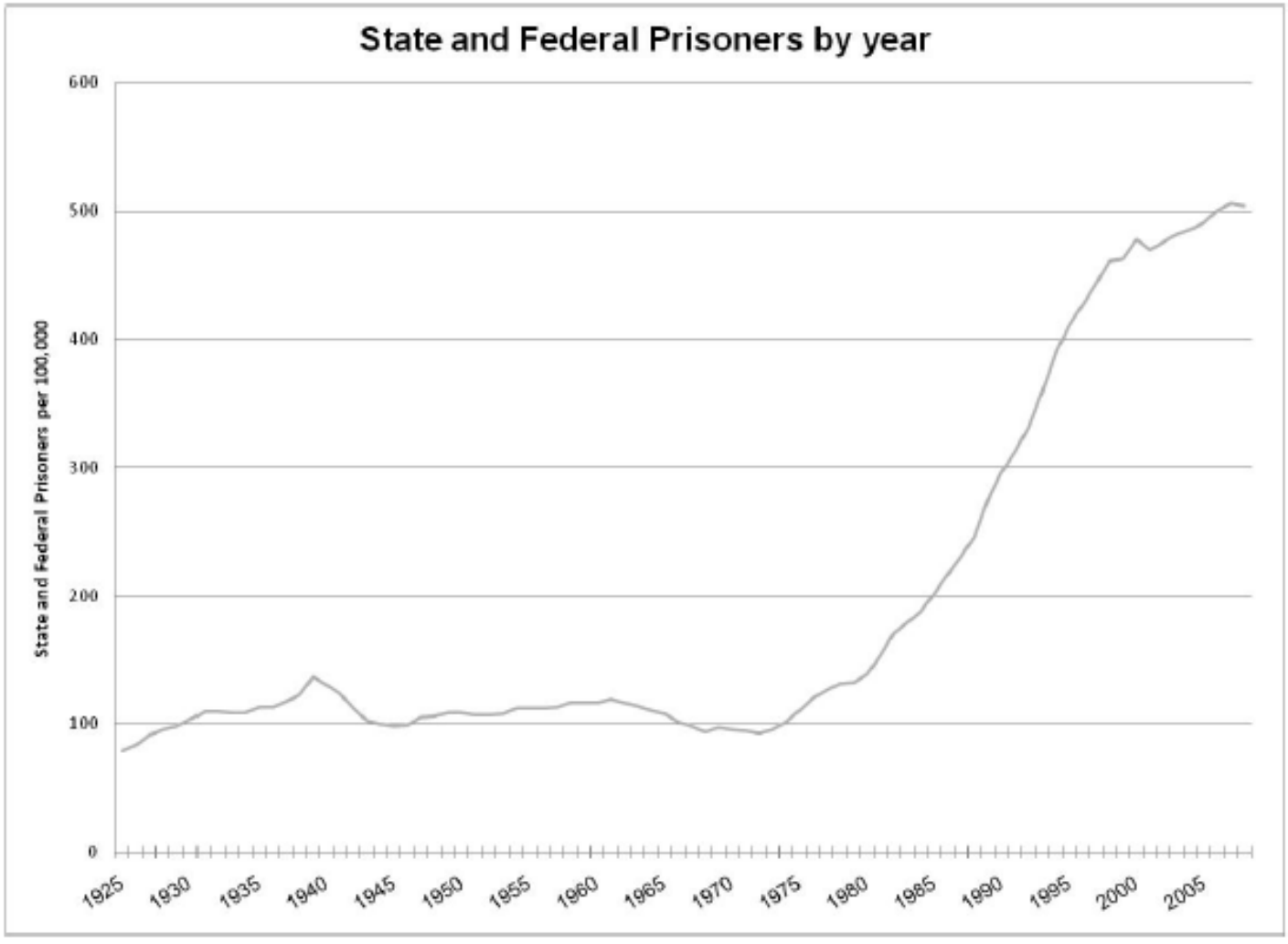
Ratio Post-Prop. 99 RMSPE to Pre-Prop. 99 RMSPE



Mass incarceration

- The US has the highest prison population of any OECD country in the world
- 2.3 million are currently incarcerated in US federal and state prisons and county jails
- Another 4.75 million are on parole
- From the early 1970s to the present, incarceration and prison admission rates quintupled in size

Figure 1
History of the imprisonment rate, 1925 - 2008



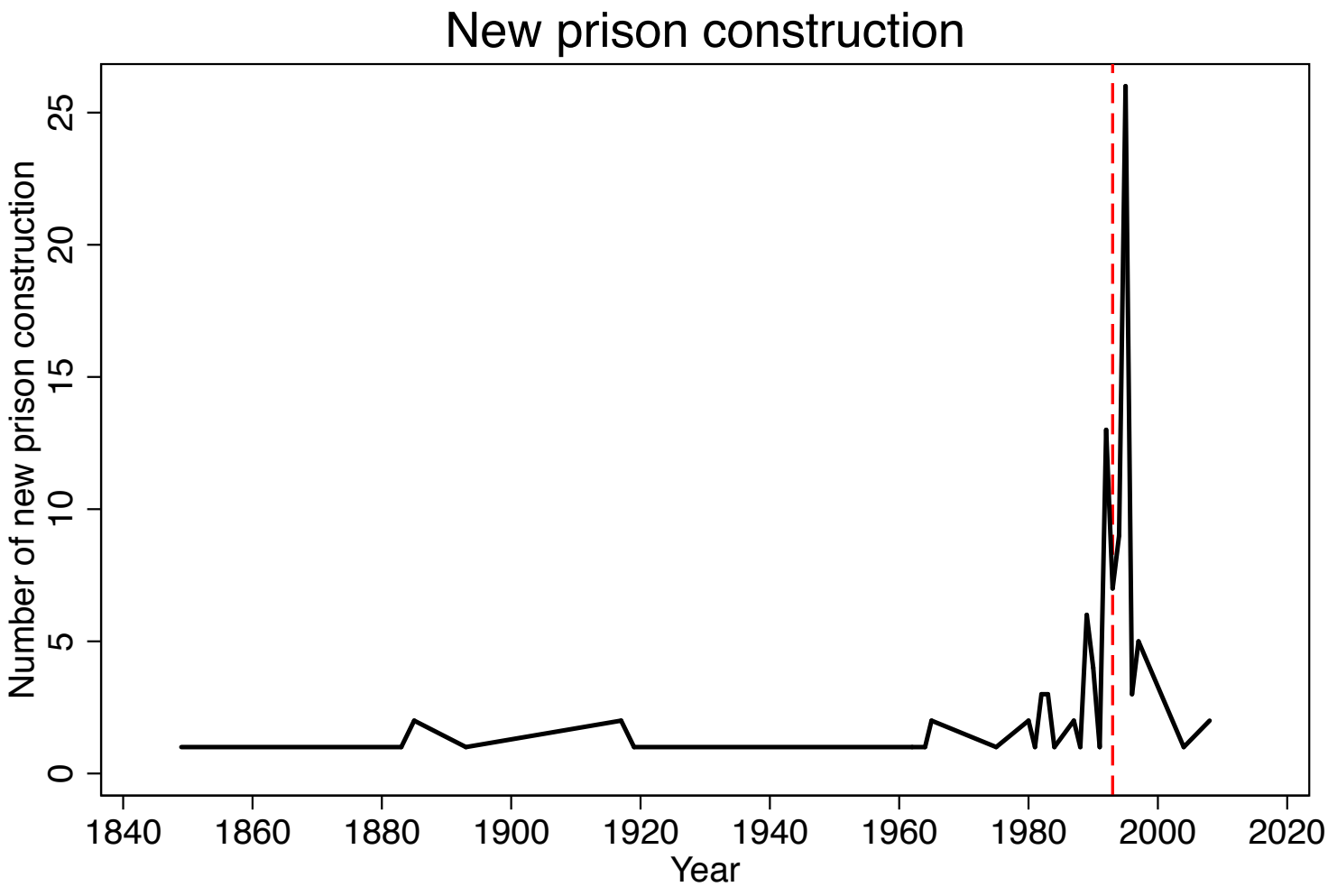
Source: www.albany.edu/sourcebook/tost_6/html

Overcrowding Lawsuit

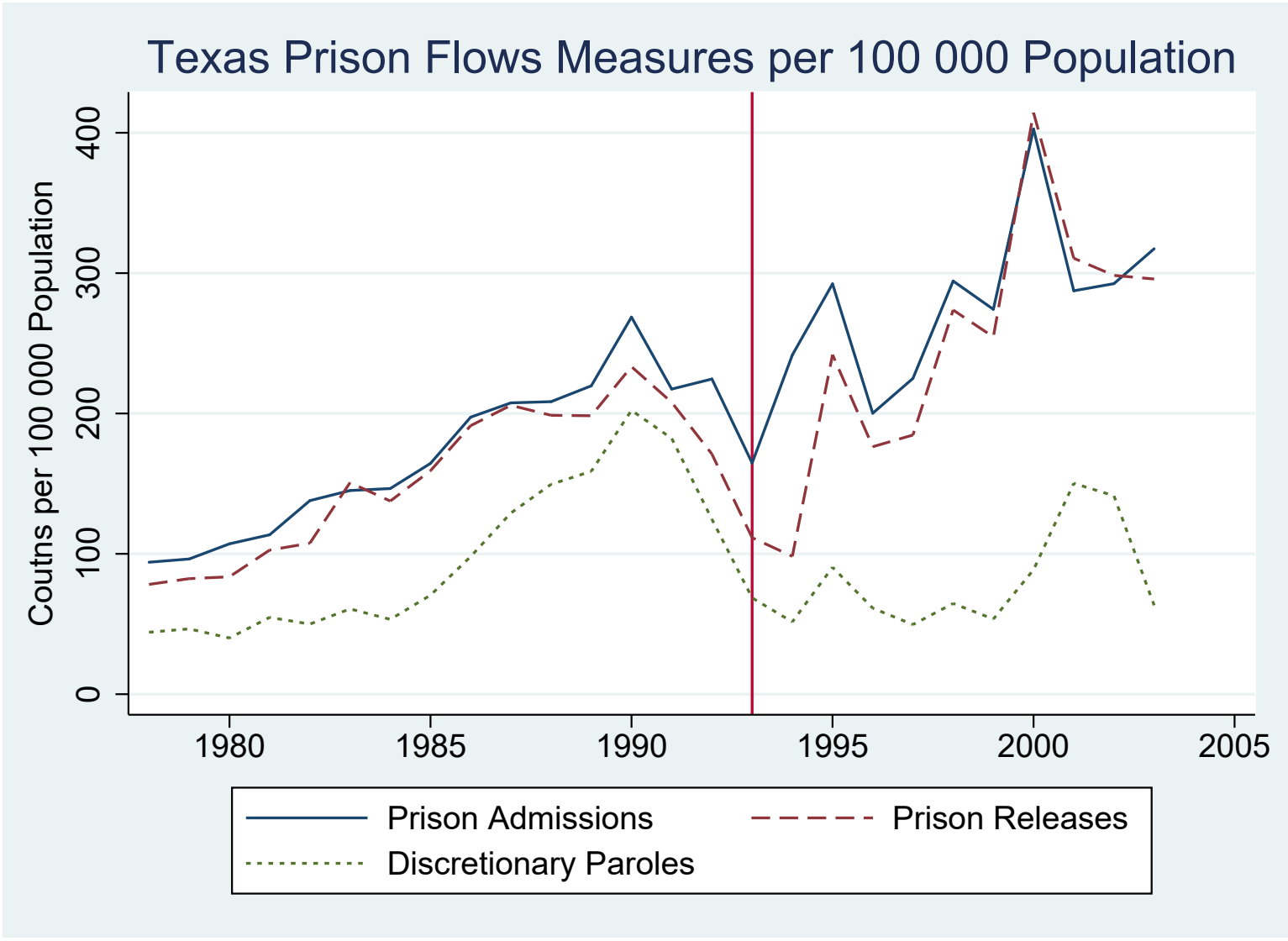
- Ruiz v. Estelle 1980
  - → Class action lawsuit against TX Dept of Corrections (Estelle, warden).
  - → TDC lost. Lengthy period of appeals and legal decrees.
  - → Lengthy period of time relying on paroles to manage flows
- Governor Ann Richards (D) 1991-1995
  - → Operation prison capacity increased 30-35% in 1993, 1994 and 1995.
  - → Prison capacity increased from 55,000 in 1992 to 130,000 in 1995.
  - → Building of new prisons (private and public)

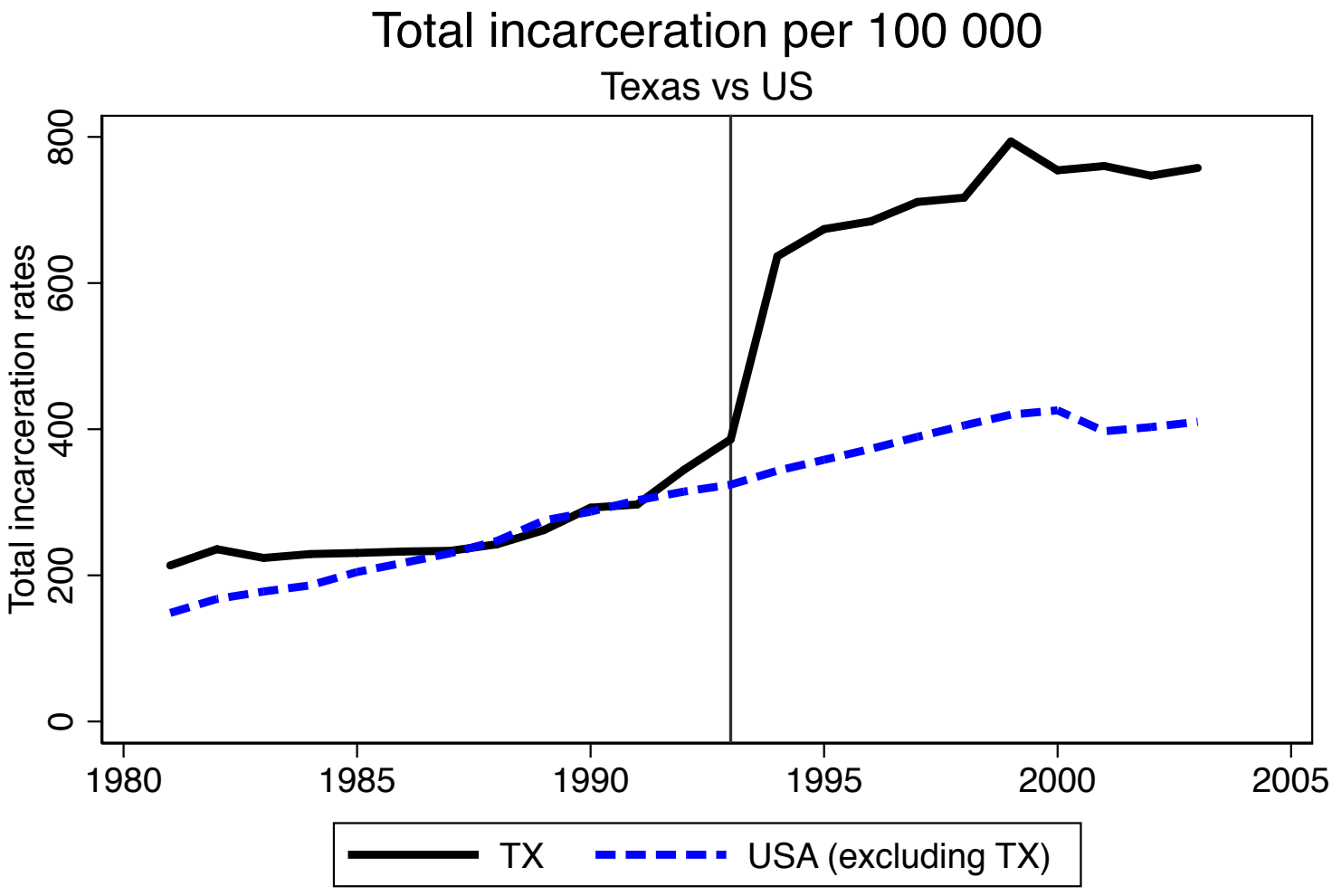
Prison constraints

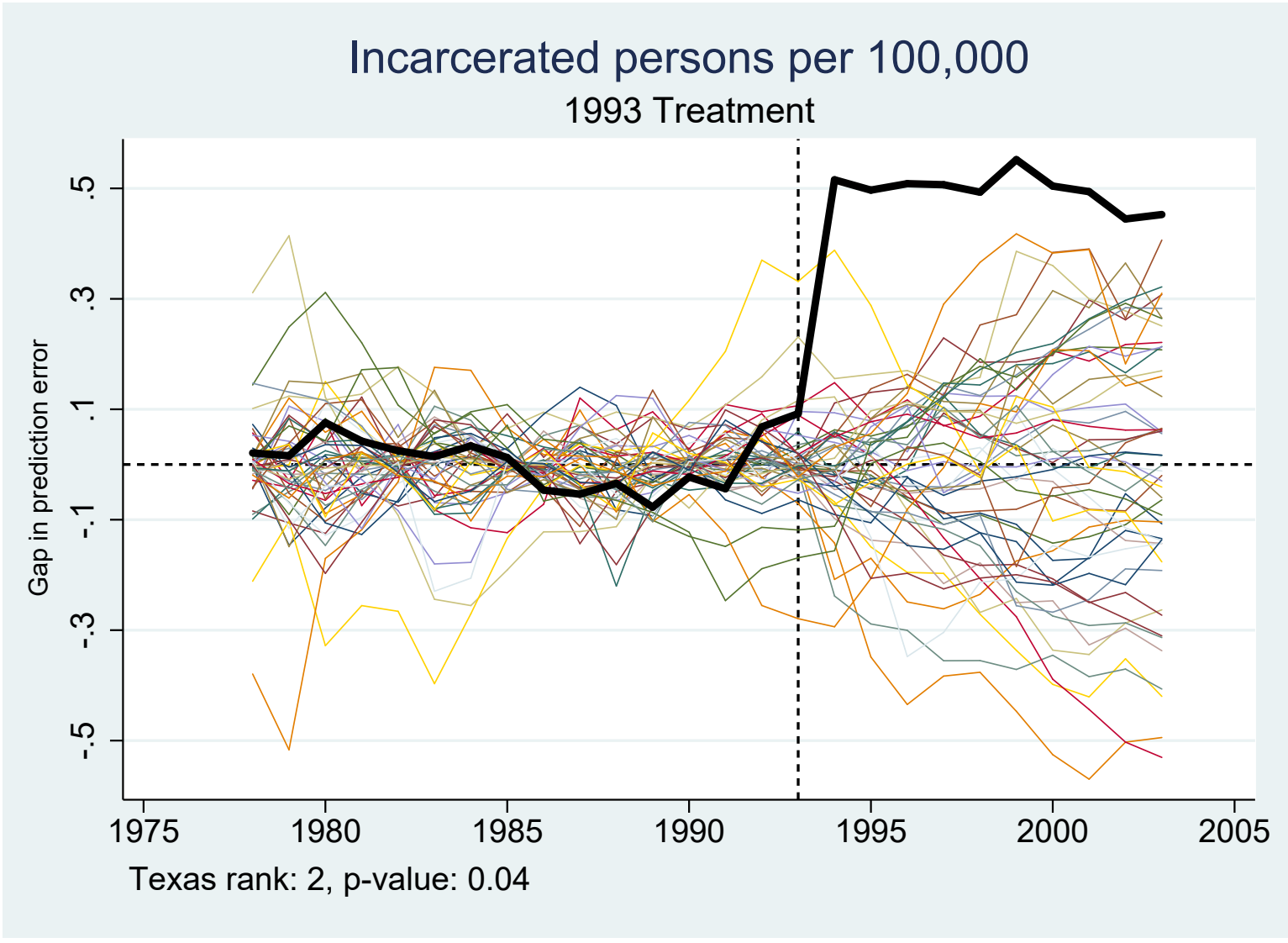
- Prisons are and have been at capacity for a long time.
- Requires managing flows through
  - → Prison construction
  - → Overcrowding
  - → Paroles



Red dashed line is 1993


Texas Prison Flows Measures per 100 000 Population
Coutns per 100 000 Population
0
100
200
300
400
1980
1985
1990
1995
2000
2005
Prison Admissions
Prison Releases
Discretionary Paroles


Total incarceration per 100 000
Texas vs US
Total incarceration rates
0
200
400
600
800
1980
1985
1990
1995
2000
2005
TX
USA (excluding TX)
1993 starts the prison expansion


Incarcerated persons per 100,000
1993 Treatment
Gap in prediction error
.5
.3
.1
-.1
-.3
-.5
1975
1980
1985
1990
1995
2000
2005
Texas rank: 2, p-value: 0.04