Causal Inference and Covariate Balance with Observational Data: A Discussion and Some Examples

Estimating Treatment Effects in the Potential Outcomes Framework

Presentation by Mathew McCubbins

With materials borrowed from Daniel Enemark, UCSD; Guido Imbens, Harvard; Colin McCubbins, Stanford, Jas Sekhon, Berkeley

The Points (relax there are only 5):

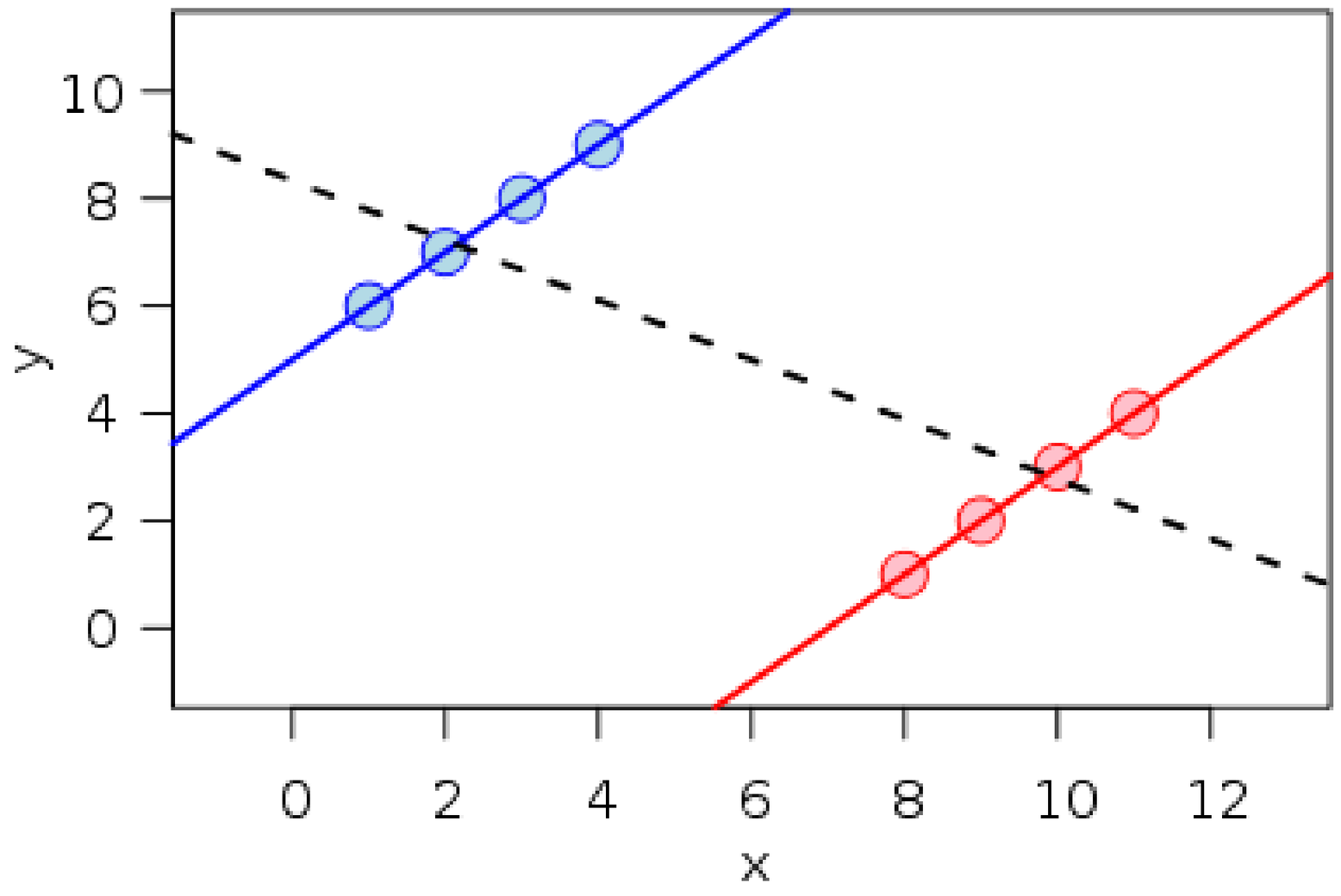
Many Papers We Read:

---

1. Lack of Clarity as to the Treatment and/or Control Groups
2. Lack of Clear Definition of What is The Counterfactual
3. The Treatment is Confounded With the Outcome
4. There is a Lack of Overlap in the Sample Between Treatment and Control, and
5. There is no Covariate Balance Between T & C, and Thus they do not Look Identical

1. A Lack of Clarity in the Definitions of Treatment and Control.

Simpson’s Paradox



Running regressions is easy. In this example, a correlation across groups (is negative) is reversed in aggregate. To quote Don Rubin, “Design Trumps Analysis.”

Example: Were Berkeley graduate admissions biased against women in 1973?

Men		Women	
Applicants	Admitted	Applicants	Admitted
8442	44%	4321	35%

Were Berkeley graduate admissions biased against women in 1973?

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

P.J. Bickel, E.A. Hammel and J.W. O'Connell. 1975. "Sex Bias in Graduate Admissions: Data From Berkeley." *Science* 187 (4175): 398–404.

2. A Lack of Clarity in Defining the Counterfactual

A Salary Study at a Large Research Unit

In the table that follows:

The Dependent Variable is the subject's real salary in 2000 dollars (called salary in the table) in each year;

The main Independent Variables are salary at appointment (Appt_salary), which is the subject's real salary of the subject in 2000 dollars at time of appointment; and Years since PhD (Yrs_since_deg), which is the subject's years since PhD for each year;

The Treatment Variable is Gender_code, which is equal to 0 for men and 1 for women subjects.

```
Random-effects ML regression                    Number of obs      =        72
Group variable: ism                             Number of groups   =         6
Random effects u_i ~ Gaussian                   Obs per group: min =        12
                                                               avg =      12.0
                                                               max =        12
                                                LR chi2(15)        =    230.85
Log likelihood  = -311.75688                    Prob > chi2        =    0.0000
```

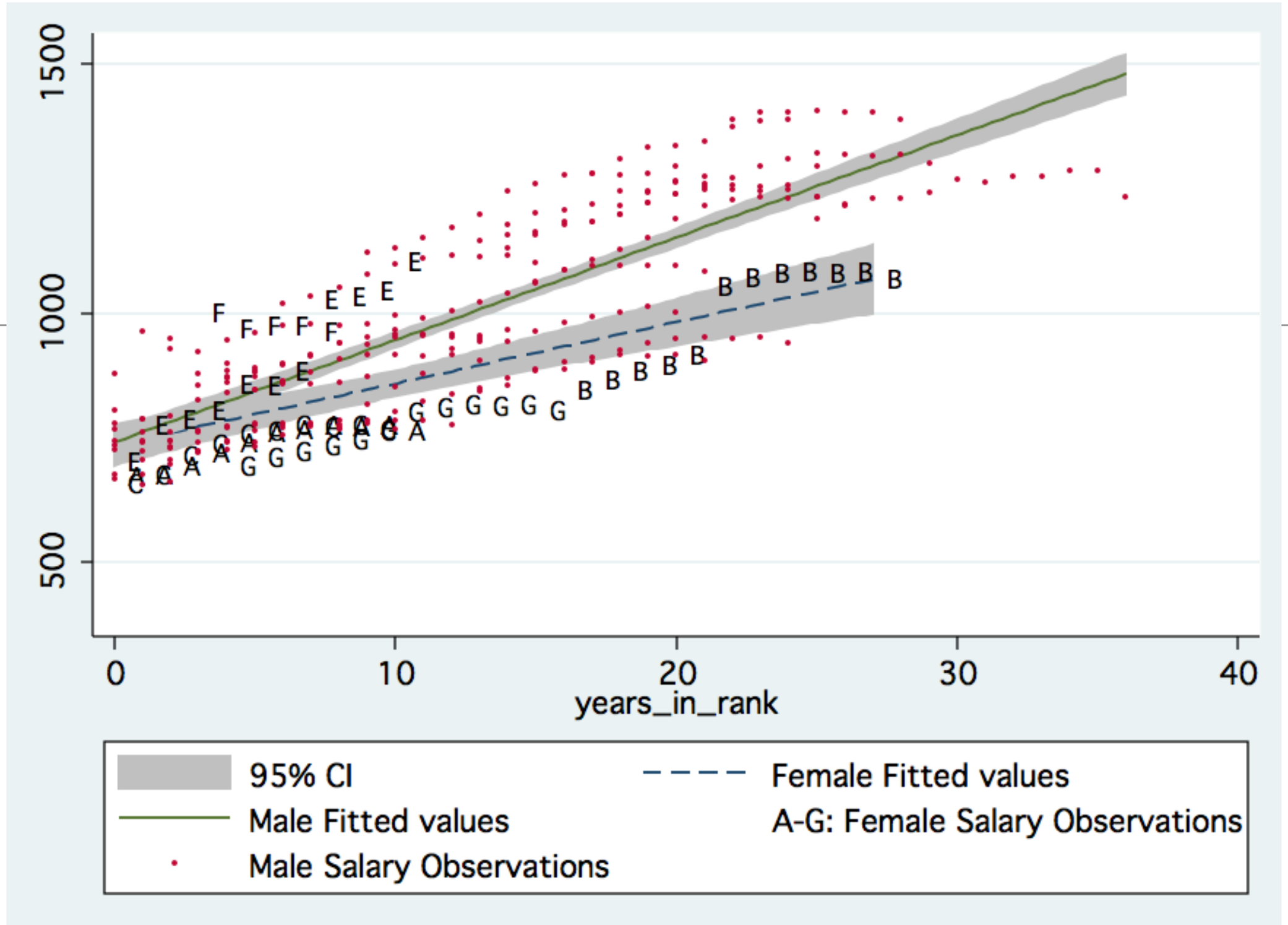
salary	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
appt_salary	-3.297377	.4854905	-6.79	0.000	-4.248921	-2.345833
yrs_since_deg	48.4478	4.360772	11.11	0.000	39.90084	56.99475
gender_code	-123.5581	37.16056	-3.32	0.001	-196.3914	-50.72474
professor	79.4396	9.888771	8.03	0.000	60.05797	98.82124
_Iyear_1998	-46.98194	11.79058	-3.98	0.000	-70.09105	-23.87283
_Iyear_1999	-107.1882	17.70109	-6.06	0.000	-141.8817	-72.49472
_Iyear_2000	-165.4684	25.64805	-6.45	0.000	-215.7376	-115.1991
_Iyear_2001	-208.2177	32.42948	-6.42	0.000	-271.7783	-144.657
_Iyear_2002	-225.6228	38.01764	-5.93	0.000	-300.136	-151.1096
_Iyear_2003	-268.6547	44.36851	-6.06	0.000	-355.6154	-181.694
_Iyear_2004	-329.0877	51.10362	-6.44	0.000	-429.249	-228.9265
_Iyear_2005	-395.8666	58.50241	-6.77	0.000	-510.5292	-281.204
_Iyear_2006	-469.3563	65.6741	-7.15	0.000	-598.0752	-340.6375
_Iyear_2007	-536.2346	72.43002	-7.40	0.000	-678.1948	-394.2743
_Iyear_2008	-623.3683	80.3218	-7.76	0.000	-780.7961	-465.9405
_cons	664.2797	92.06111	7.22	0.000	483.8432	844.7161
/sigma_u	44.26374	12.98688			24.90621	78.66628
/sigma_e	15.14551	1.319032			12.76886	17.96452
rho	.8951935	.0575353			.7368428	.9696461

```
Likelihood-ratio test of sigma_u=0: chibar2(01)=  119.79 Prob>=chibar2 = 0.000
```

In the analyses that follows we have given unique letter labels to each of the women full professors in our analysis.

The labels (A-G) correspond to the following unique identification numbers:

The lines of best fit show only the *bivariate* relationship between years in rank and real salary for men and women. Each individual subject's real salary, each year, is plotted in the figure as are the 95% confidence intervals around this simple regression line (in gray). Each female subject is identified in the plot.



Salary Differentials

A comparison of actual versus predicted (using the full model) salaries for female full professors A to G versus the male counterfactual (if the women were men) in 2008 dollars.

2008 Salaries – Actual, Predicted, Counterfactual

Label	ISM	Actual Salary	Predicted	Predicted Counterfactual (if Male)
A	22053	$166,100.00	$176,724.57	$191,105.87
B	37206	$231,850.00	$247,173.41	$261,554.72
C	43963	$166,100.00	$172,786.82	$187,168.12
E	78851	$239,000.00	$177,425.98	$191,807.28
F	26095	$209,150.00	$191,915.91	$206,297.22
G	85601	$174,850.00	$195,010.45	$209,391.76

Satisfaction:
As the philosopher Mick Jagger once said, you can’t always get what you want.

The research design above is not very satisfying. To get some more satisfaction, we need to think a bit more about research design.

We know the counterfactuals to which we are implicitly comparing our subjects do not and cannot exist (we cannot go back make the women into men and vice-versa) and therefore we may make very large inferential errors.

This Slides takes up Space Whilst I Clear My Throat (Bernie: Slit my Throat?)

Some Famous Law Scholars Wrote That Republicans Were More Likely to vote for the Civil Rights Act?

Democrats		Republicans	
House	Senate	House	Senate
61%	69%	80%	82%

Were Republicans More Likely to vote for the Civil Rights Act?

Region	Democrats		Republicans	
	House	Senate	House	Senate
North	94%	98%	85%	84%
South	7%	5%	0%	0%

- **Larger Questions:** What can we safely call a treatment? Were legislators treated with party affiliation or region?
- Can we imagine a counterfactual? “What would Strom Thurmond have done if he were a Northern Republican?”

Point 3: For whom can we make inferences?

Imagine that ideology is primarily a function of income and age

My theory is that aging and getting richer makes you more conservative:

$$\text{Conservativeness} = \beta_0 + \beta_1 (\text{Income}) + \beta_2 (\text{Age}) + \varepsilon$$

As is typical in the literature, let me say my hypotheses are that $\beta_1 > 0$ and $\beta_2 > 0$

If my theory is supported, for whom can I make inferences about ideology?

For whom can we make inferences?

So, How Do we Get to the Point Where Design Trumps Analysis?

Mill’s Methods

- Method of difference: identical units (i.e., covariates) but different treatments
- Method of agreement: different covariates but identical treatments
- Joint method of agreement and difference
- Method of residue
- Method of concomitant variations (regression)

METHOD OF TIME MACHINE

Experiments aka RCT

- In any experiment you need either
  - multiple groups (e.g., one treatment and one control) or
  - multiple observations (e.g., a within subjects design where we observe the same subject under different conditions).

Randomization

- Randomized Control Trial (RCT) is the best available study design to explore causal effect

  $(Y_1, Y_0) \perp\!\!\!\perp T$ (Ex ante, the Outcome and Treatment Assignment are independent)

$$E(Y_{1i} - Y_{0i}) = E(Y_{1i} - Y_{0i} \mid T)$$

$$= E(Y_{1i} \mid T) - E(Y_{0i} \mid T) = E(Y_i \mid T=1) - E(Y_i \mid T=0)$$

No confounding effect in RCT

Quasi-experiments

- In quasi-experiments you typically need both multiple groups and multiple observations. We do not usually have this, so any quasi-experiment we do here is fraught with threats to validity.
- Ex. The biggest threats to validity in studying the effects of gender on salary will be that we could not randomize the pool of employees across the two treatment groups: female and male.
  - We do not observe what would have happened to the men if they were women and vice-versa, we do not observe the counterfactuals.

Potential Outcomes

1. Each case i is one of N random draws from a large population

2. These draws collectively constitute the sample

3. W is a binary treatment; $W_i = 0$ if control, $W_i = 1$ if treatment

4. $Y_i(W_i)$ is the outcome for case i given its treatment status

5. For each draw, we postulate $Y_i(1)$ and $Y_i(0)$, the outcomes that would obtain under treatment and control conditions

6. For each draw there is also a vector of exogenous variables X_i
   Generally, X_i (the covariates) can include lagged outcomes

Potential Outcomes

7. We observe (W_i, Y_i, X_i), where Y_i is the *realized* outcome, i.e. $Y_i \equiv Y_i(W_i)$.

8. Propensity Score is the probability of receiving treatment given the vector of covariates:
   $e(x) = Pr(W_i = 1 | X_i = x) = E[W_i | X_i = x]$

9. Conditional regression and and variance functions
   $\mu_w(x) = E[Y_i(w) | X_i = x]$ $\sigma_w^2(x) = V[Yi(w) | Xi = x]$

An Example of the Fundamental Problem

Y_{1i} denotes the outcome of individual i given being treated

Y_{0i} denotes the outcome of individual i given being control

$\Delta_i = Y_{1i} - Y_{0i}$ is the treatment effect on i

Sub.	Y_1	Y_0	Δ
A	15		
B	13		
C		8	
D		4	

The Same Example Continued

Y_{1i} denotes the outcome of individual i given being treated

Y_{0i} denotes the outcome of individual i given being control

$\Delta_i = Y_{1i} - Y_{0i}$ is the treatment effect on i

Sub.	Y_1	Y_0	Δ
A	15	10	5
B	13	8	5
C	13	8	5
D	9	4	5

The Same Example Continued

Suppose we also know the covariate X, which is associated with the treatment reception

Sub.	X	Y_1	Y_0	Δ
A	40	15	10	5
B	30	13	8	5
C	30	13	8	5
D	20	9	4	5

What do we *want* to estimate?

- Depending on your design, you may be able to measure one of the following:
  - ATE = average treatment effect
  - ATT = average treatment effect for the treated
  - ATC = average treatment effect for the control
  - ITT = intent-to-treat effect

- Treatment effects as population parameters:
  PATE: $\tau_P = \mathrm{E}\,[Y_i(1) - Y_i(0)]$
  PATT: $\tau_{P,T} = \mathrm{E}\,[Y_i(1) - Y_i(0) \mid W = 1]$

- Treatment effects as sample statistics:
  SATE: $\tau_S = \frac{1}{N} \sum_{i=1}^{N} [Y_i(1) - Y_i(0)]$
  SATT: $\tau_{S,T} = \frac{1}{N_T} \sum_{i:W_i=1} [Y_i(1) - Y_i(0)]$

Problem: we *never* observe both $Y_i(0)$ and $Y_i(1)$

How do we identify ATE (τ) with only one outcome per case?

- We are forced to *assume* "Strongly Ignorable Treatment Assignment" (SITA). This is actually two assumptions.

  - Unconfoundedness: $[Y_i(0), Y_i(1)] \perp\!\!\!\perp W_i \mid X_i$
    ⫫ or ⊥ means "is independent of," so this equation means "treatment assignment [W] and response [Y(0),Y(1)] are known to be conditionally independent given [X]" (Rosenbaum and Rubin, 1983).

- Overlap: $0 < \Pr(W_i = 1 \mid X_i) < 1$
  No cases are in a region of the covariates in which all cases are in the same treatment group. That is, the propensity score is always greater than 0 and less than 1.

4. Overlap

- If all cases in a certain region of the joint distribution of covariates receive the treatment, there is no way to estimate the outcome that would have obtained for that type without treatment (and vice versa).

- A propensity score is the probability of receiving treatment given covariates. $e(x) \equiv Pr(W=1 \mid X=x)$

- Usually p-scores are estimated using a logit model to regress treatment status on the covariates

  - ROT: Usually exclude cases where $e(x)<.1$ or $e(x)>.9$

5. Unconfoundedness

- This assumption is also referred to as “selection on observables” or “conditional independence”

- It is analogous to the “missing at random” assumption for missing data

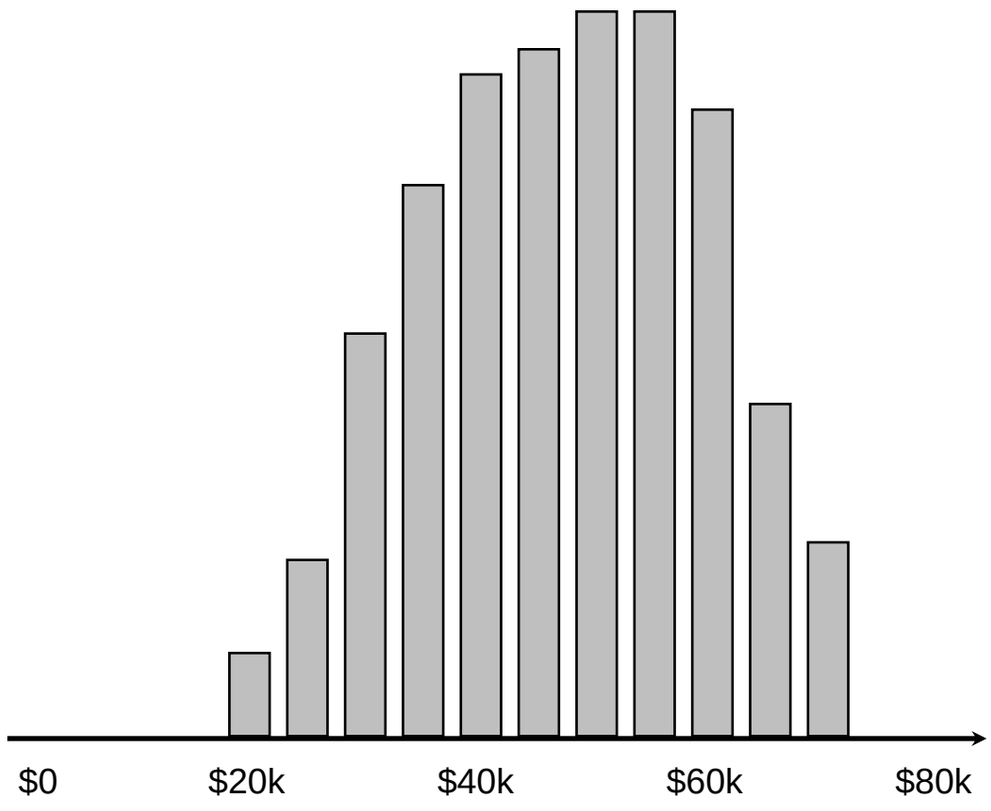
- To achieve unconfoundedness, you need either

    - random assignment of the treatment, or

    - a near-perfect understanding of the assignment process and the ability to observe all relevant variables, so that you can condition Y on the right covariates—with the right functional form!

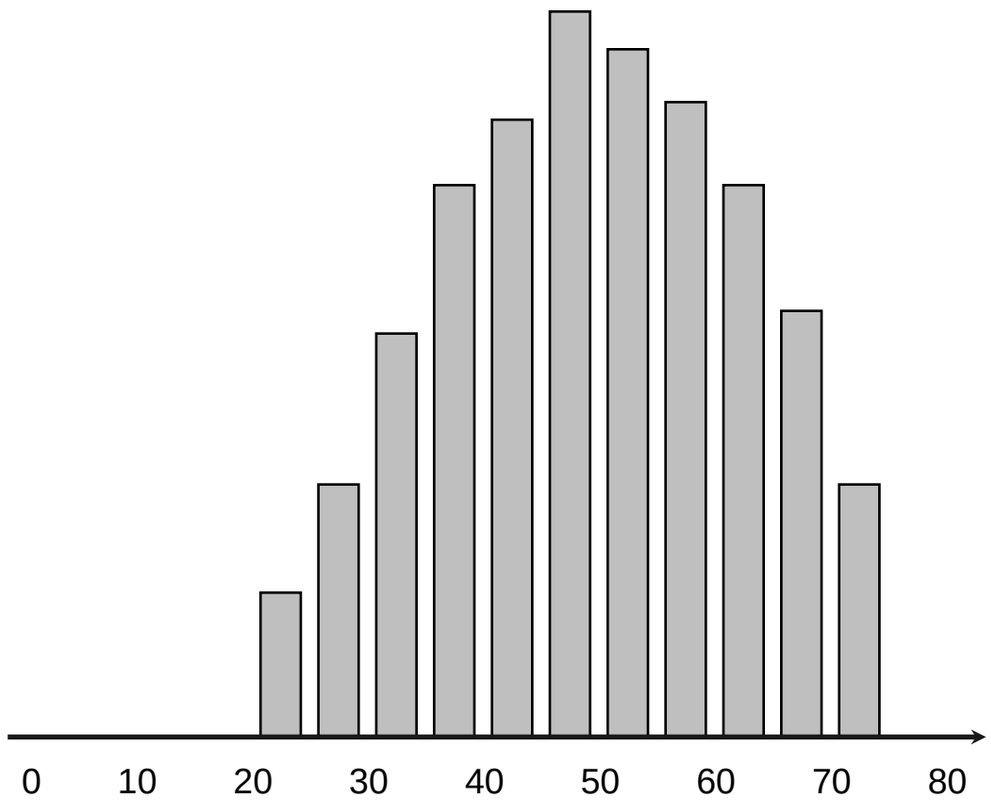
Overlap: For whom can we make inferences?

anyone who makes
between $20k and $75k?



Income

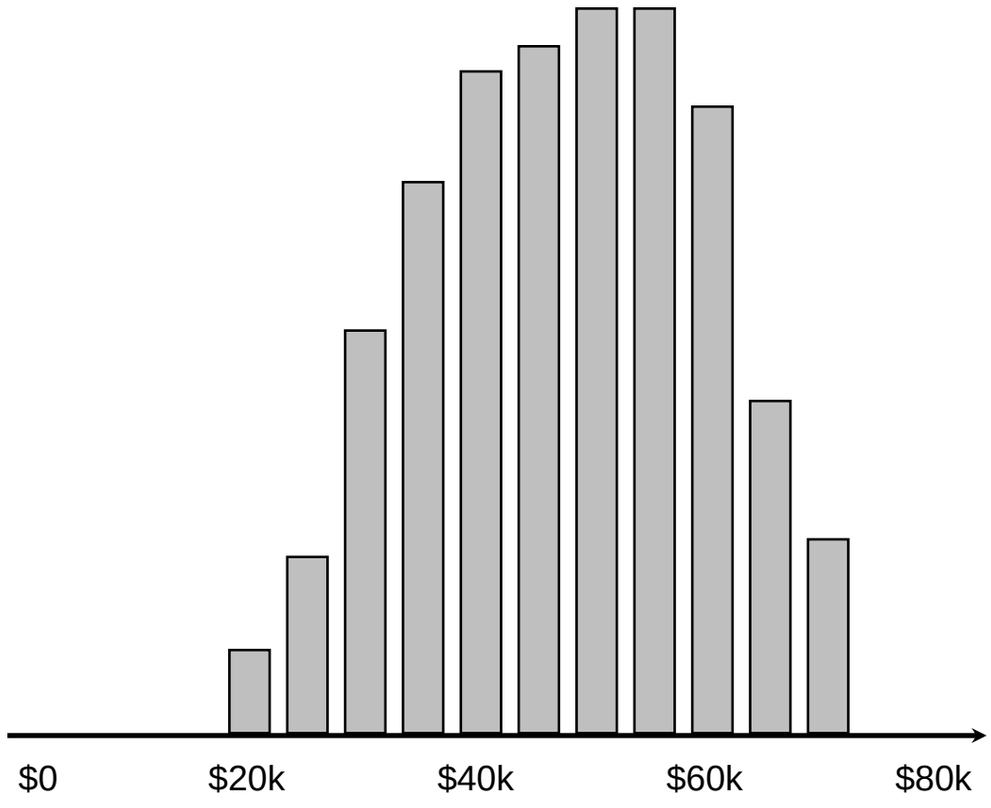
anyone between
18 and 77 years old?



Age

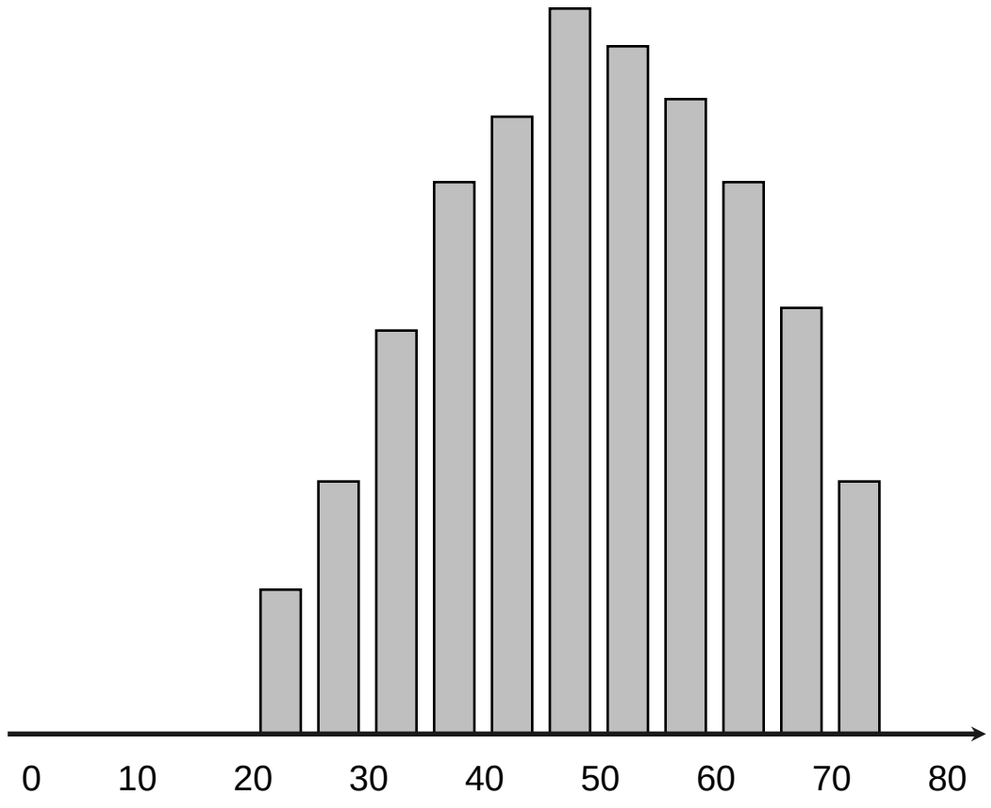
For whom can we make inferences?

anyone who makes
between $20k and $75k?



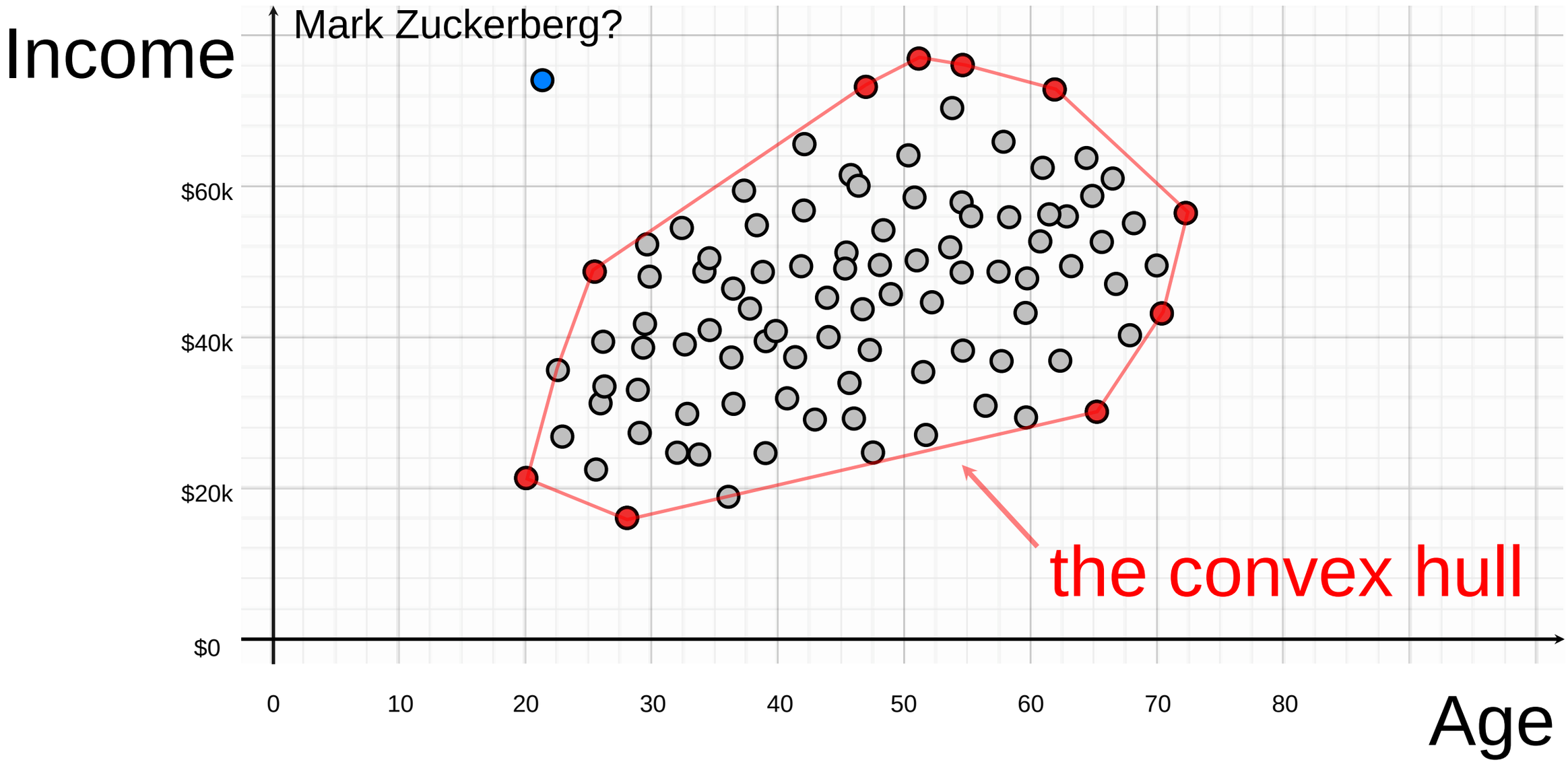
Income

anyone between
18 and 77 years old?



Age

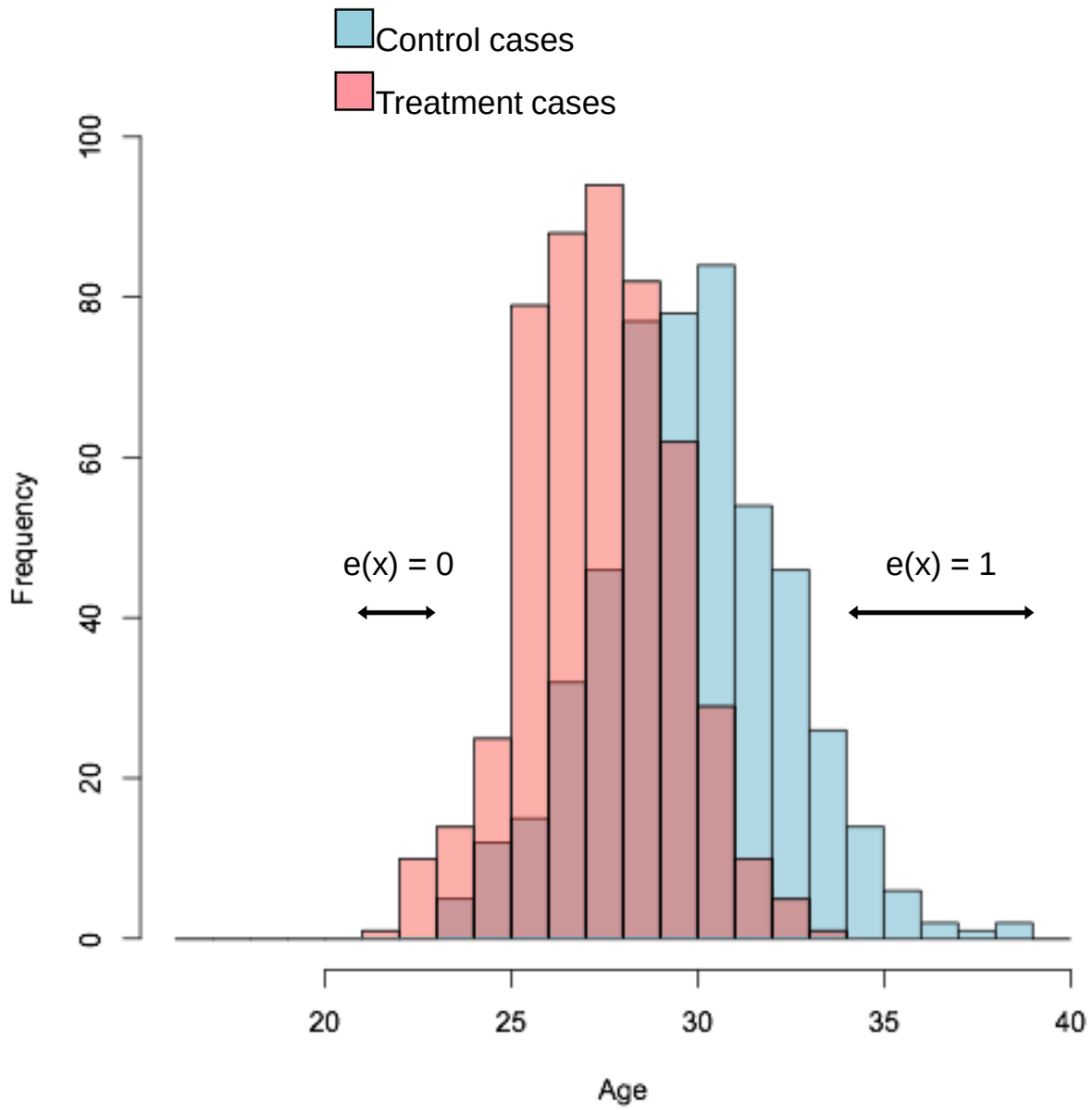
For whom can we make inferences?



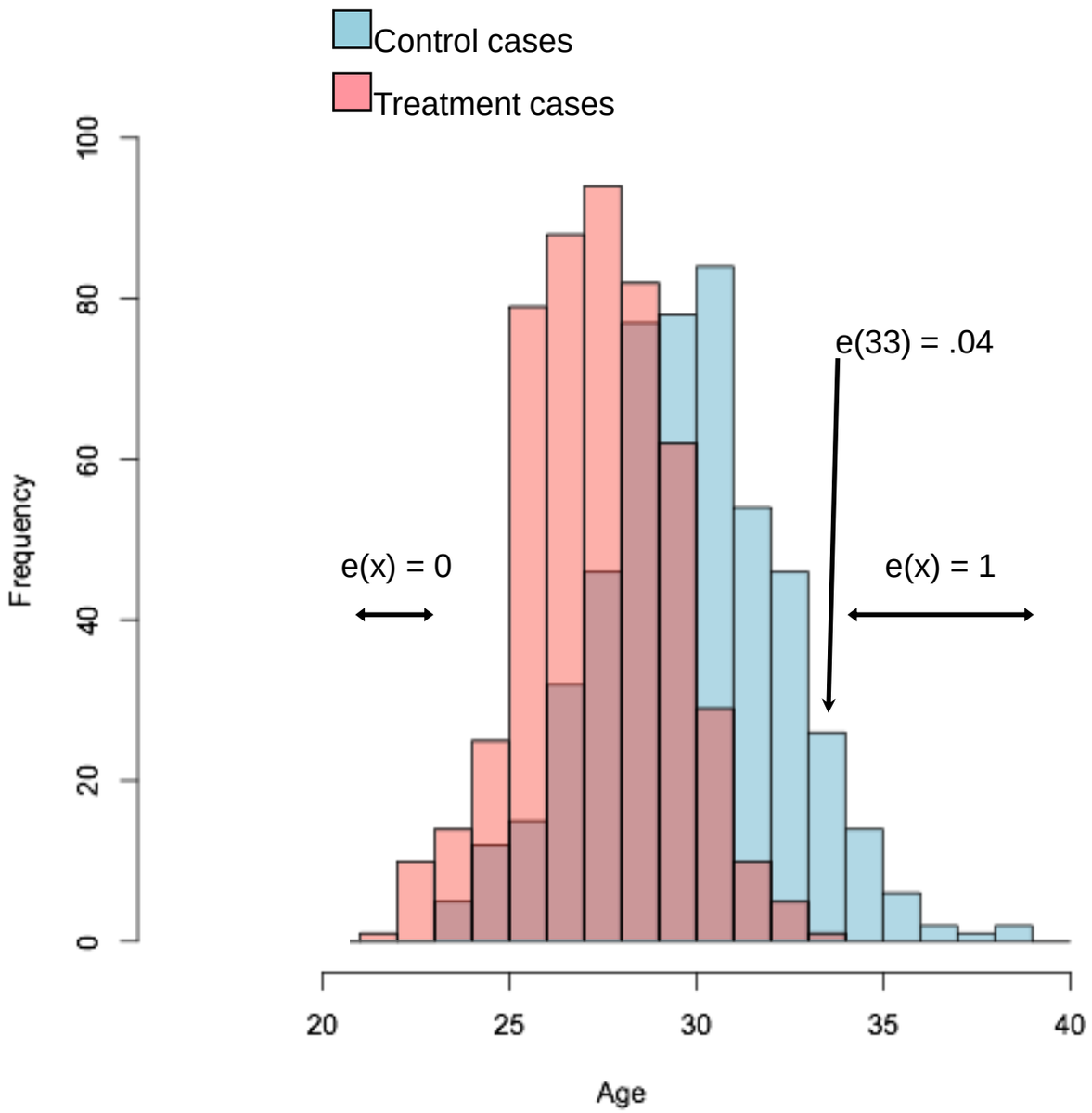
Back to the potential outcomes framework

- Consider a binary treatment non-randomly assigned to a population for which we have one covariate, age.

of



Back to the potential outcomes framework



- To ensure overlap, you must discard cases very likely or unlikely to receive treatment—cases for which $\alpha \leq e(X) \leq 1-\alpha$. Imbens says $\alpha = 0.1$ is in practice the optimal set for inference.

- Should we exclude 33-year-olds?

- The answer requires a *theory*. Are people whose ages vary by one year (or one day) distinct? Should we recode age into 4-year cohorts based on first presidential election?

What does SITA (Strongly Ignorable Treatment Assignment) get us?

---

1. Conditional regression is identified:

$$\mu_w(x) = \mathbb{E}[Y_i(w)|X_i = x] = \mathbb{E}[Y_i(w)|W_i = w, X_i = x] = \mathbb{E}[Y_i|W_i = w, X_i = x]$$

   (This means we can calculate the mean outcome for both treatment values within a specific vector of covariates.)

2. The ATE can be found by analyzing each sub-population with covariates $X_i = x$:

$$\tau(x) \equiv \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] = \mathbb{E}[Y_i|X_i, W_i = 1] - \mathbb{E}[Y_i|X_i, W_i = 0]$$

   (This relies on the overlap assumption, since it would be impossible to find $Y_i(1)$-$Y_i(0)$ for any x where $Pr(W_i) = 1$ or 0.)

3. Given identification of $\tau(x)$, $\tau_P = E[\tau(X_i)]$

What does SITA get us? (part 2)

4. Estimating ATE doesn't require conditioning simultaneously on all covariates. Conditioning on propensity score (a scalar function of the covariates) removes all biases from observable covariates. (Though it may be less efficient than conditioning on the full X_i.)

5. Efficiency bounds & asymptotic variances for PATE:

$$V \geq \mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)} + \frac{\sigma_0^2(X_i)}{1-e(X_i)} + (\tau(X_i) - \tau_P)^2\right]$$

That's the lower bound of V for a regular estimator of τ_P, given

$$\sqrt{N}(\hat{\tau} - \tau_P) \to N[0, V]$$

Variance of PATE decreases as the variance of either treatment group decreases or as the likelihood of being assigned to that group increases, so if you have a small treatment or control group with high variance, the variance in your estimator is going to be very high. The more constant the treatment effect across X, the lower your estimator's variance.

“But a tenuous assumption”

- Imbens argues that SITA is a defensible assumption for two reasons:
  - (1) it is necessary for working with observational data
  - (2) even if agents choose their treatment, they may still be comparable if the unobserved variables driving their different treatment choices are ⊥ Y. (Though this is not knowable.)

- Rosembaum argues that you should test this assumption, by crafting an “elaborate theory” of treatment assignment from which you can deduce testable hypotheses. The Nonequivalent Dependent Variables design is well-suited for such tests (c.f. Cook & Campbell; Trochim)

One other essential assumption: SUTVA

- SUTVA requires that the potential outcome for any particular unit i following treatment t is stable, "in the sense that it would take the same value for all other treatment allocations such that unit i receives treatment t (Rubin 1990, p. 282)

- Most common violations:
  - there are versions of each treatment varying in effectiveness (heterogeneous treatment effects)
  - there exists interference between units (aka, spillover)

Summary: Assumptions behind a regression

① Unconfoundedness: $[Y_i(0), Y_i(1)] \perp\!\!\!\perp W_i \mid X_i$
This subsumes BLUE assumption that error is uncorrelated with treatment

② Overlap: $e(x) \equiv \Pr(W=1 \mid X=x)$, and $\alpha \leq e(X) \leq 1-\alpha$, with α generally = 0.1.

③ SUTVA: $Y_i(1)$ is invariant across all possible distributions of treatment

④ Homoskedasticity: variance of ε constant across all values of X
(i.e. model's predictions are of equal quality at all levels of X).

One other essential assumption: SUTVA

Three treatments are distributed among six subjects. SUTVA requires that the the outcome under treatment for Policeman is the same for any distribution of treatments in which Policeman receives a treatment.

One other essential assumption: SUTVA

1: Cowboy
2: Policeman
3: Serviceman

All receive the treatment.

One other essential assumption: SUTVA

1: Workman
2: Policeman
3: Biker??

So far this may
Seem plausible

One other essential assumption: SUTVA

1: Workman
2: Biker?
3: Policeman

Now Policeman Gets Needle #3

Estimating ATE: Regression Boils Down to Estimating a Difference in Means

---

1. Given $\hat{\mu}_w(x)$ for $w = 0,1$, a regression estimates the PATE or SATE by averaging the difference between these quantities over the empirical distribution of covariates:

$$\hat{\tau}_{reg} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\right)$$

2. Generally, in a regression, average predicted outcome = average observed outcome, so $\sum_i W_i \cdot \hat{\mu}_1 X_i = \sum_i W_i Y_i$

   so that $\hat{\tau}_{reg} = \frac{1}{N}\sum_{i=1}^{N}\left[W_i\left(Y_i - \hat{\mu}_0 X_i\right) + (1 - W_i)\left(\hat{\mu}_1 X_i - Y_i\right)\right]$

3. In a regression,
   $\mu_w(x) = \beta' x + \tau w$, where τ is the ATE. In an OLS, $Y_i = \alpha + \beta' X_i + \tau W_i + \varepsilon_i$

4. A simple regression is sensitive to differences in the covariate distributions of treated and control groups, because of the way it extrapolates. As in point (2) above, the regression function for the controls is used to predict missing outcomes for the treated, and vice-versa. Ideally, we want to predict the control outcome at the average covariate value for treatment cases, $\bar{X}_T = \sum_i W_i \cdot X_i / N_t$.

   Since the average prediction for $\hat{Y}_T$ is $\bar{Y}_C + \hat{\beta}'(\bar{X}_T - \bar{X}_C)$, the specification of the model is more robust the smaller the difference in average covariate values for treatment and control (because β-hat has less influence on prediction the smaller its multiplicand).

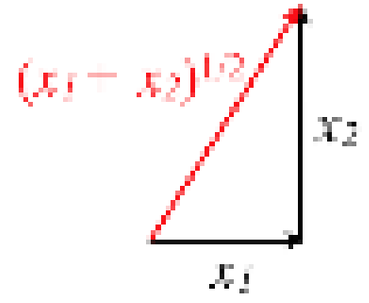
Estimating ATE: Matching

1. Notation

   a) $\ell_m(i)$ is the mth-closest match to case i, an index l such that $W_l \neq W_i$, and
   $m = \sum_{j|W_j \neq W_i} 1\{\|X_j - X_i\| \leq \|X_l - X_i\|\}$

      (1) $1\{\bullet\}$ is an indicator function. $1\{\bullet\} = 1$ if the expression in brackets is true, else 0.

      (2) $\|\mathbf{x}\|$ is the length of vector $\mathbf{x}$. $\|\mathbf{x}\| = (x_1 + x_2 + \ldots x_n)^{1/2}$
      You can think of this as a multidimensional extension of the Pythagorean Theorem. For a vector $\mathbf{x}$ with two elements x_1 and x_2, $\|\mathbf{x}\|$ is the hypotenuse of a triangle with sides x_1 and x_2.



      (3) Obviously, since we're going to minimize the distance across multiple dimensions (i.e. multiple covariates), we need some way to normalize the scales of the covariates. If each person has (cars, dollars) such that Dan=(1,1), Ben=(1,2) and Steve=(2,1), Dan is closer to Ben than he is to Steve.

      (4) So people use Mahalanobis distance, $d_M(x,z) = (x-z)'(\sum_X^{-1})(x-z)$ where Σ is the covariance matrix of the covariates.

   b) $J_M(i)$ is the set of indices for the first M matches for case i. $J_M(i) = \{\ell_1(i), \ldots \ell_M(i)\}$

Estimating ATE: Matching part 2

---

2. Method of estimation

a) The imputed values are estimated thus:

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \frac{1}{M}\sum_{j \in \mathcal{J}_M(i)} Y_j & \text{if } W_i = 1, \end{cases} \qquad \hat{Y}_i(1) = \begin{cases} \frac{1}{M}\sum_{j \in \mathcal{J}_M(i)} Y_j & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1. \end{cases}$$

b) Using the imputed values, we calculate the "simple matching estimator:"

$$\hat{\tau}_M^{sm} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{Y}_i(1) - \hat{Y}_i(0)\right)$$

3. Bias

a) Bias of the simple matching estimator is of order $O(N^{-1/K})$, where K is # of covariates.

(1) But for large enough samples, only continuous variables add to bias.

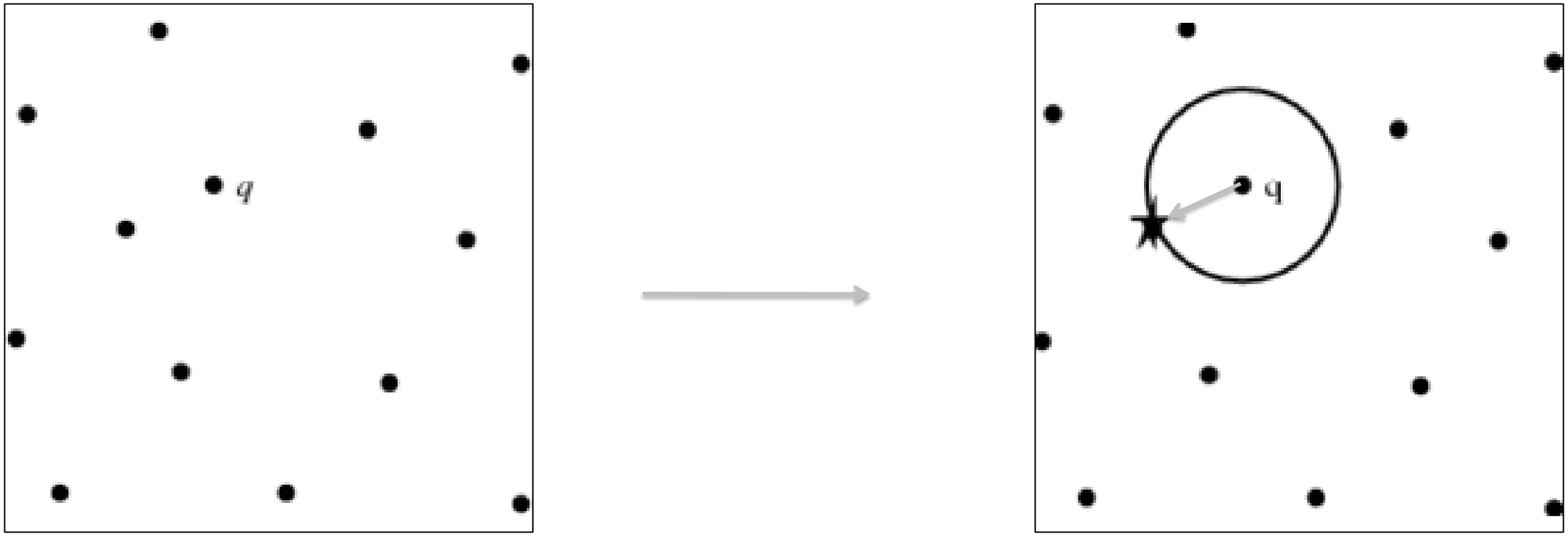
(2) If there are many more control cases than treated, match only treated, and it's OK.

(3) Or you could just rely on heroic assumptions (p. 16). Yay Imbens!

b) Abadie and Imbens recommend matching + regression adjustment to handle bias

4. Efficiency: matching can seriously harm efficiency (because you're throwing out data)

The Idea of Matching



We have a point q that we are interested in. What are we matching with it?

The starred point is the closest “neighbor” to q by distance. We assume that this point is, thus, most likely to be like q. Thus we “match” q to the starred point.

Matching, In General

- Matching to balance the covariate distribution

  To make the treated and control subject look alike before treatment

  To produce a study regime which resembles a randomized experiment most, in terms of the observed covariates

Improving on OLS – Standard Matching

- One way to improve OLS is to match (Rubin 1983).
    - Matching deals with the overlap problem but also unconfoundedness, which indicates that the treatment assignment behaves as if it were randomly assigned (i.e. it is like a controlled experiment).
    - This goes a long way to improving our ability to make causal inference.

Estimating ATE: Propensity-Score Methods

1. *Imbens' Warning:* "In practice the propensity score is rarely known, and in that case the advantages of the estimators discussed below are less clear. Although they avoid the high-dimensional nonparametric estimation of the two conditional expectations $\mu_w(x)$, they require instead the equally high-dimensional nonparametric estimation of the p-score."
2. Basic p-score method
   a) A simple difference in means is biased: $\hat{\tau} = \frac{\sum W_i Y_i}{\sum W_i} - \frac{\sum (1 - W_i) Y_i}{\sum (1 - W_i)}$
   b) If you weight by the inverse of the p-score, it's unbiased (*assuming accurate p-scores*)
   $$\tau_P = \mathrm{E}\left[\frac{(W)Y}{e(X)} - \frac{(1 - W)Y}{1 - e(X)}\right]$$
   c) Normalize the weights: In expectation, the weights add up to 1 = $[\Sigma W_i / e(X_i)]/N$, but with positive variance most samples will yield some number other than 1. Hirano, Imbens and Ridder (2003) have a solution to this problem, not worth going into.
3. Blocking on p-score: divide data into M blocks based on p-score then just pretend you have strongly ignorable assignment. (A "crude form of nonparametric regression")
4. Regression on p-score: bad idea
5. Matching on p-score: variance is unknown when you match on *estimated* p-score.

Estimating ATE: Mixed Methods (matching + regression)

---

(1) Matching compares $\hat{Y}_i(0)$ to $\hat{Y}_i(1)$, derived from unit i and its match $\ell(i)$.

(2) The covariate values for i and $\ell(i)$ are close but not equal, i.e. $X_i \neq X_{\ell(i)}$
This produces bias $= \mathbb{E}[\hat{Y}_i(1) - \hat{Y}_i(0)] - [Y_i(1) - Y_i(0)]$

(3) Suppose $W_i = 1$. So $\hat{Y}_i(1) = Y_i(1)$, while $\hat{Y}_i(0)$ is imputed.
$\hat{Y}_i(0)$ is unbiased for $\mu_0(X_{\ell(i)})$, but not for $\mu_0(X_i)$, so adjust $\hat{Y}_i(0)$ by $\mu_0(X_i) - \mu_0(X_{\ell(i)})$

(4) Abadie and Imbens do this by taking the control cases used as matches for the treated units, weighted according to the number of times each is used as a match, and estimating the regression $Y_i = \alpha_0 + \beta_0' X_i + \varepsilon_i$

(Then run a second regression for the treated cases used as matches for controls?)

(5) A&I show that the resulting estimator is consistent and asymptotically normal, with bias dominated by the variance.

Back to Mill’s methods

- Mill’s method of agreement (treatment status is the same but covariates are different) sucks—there is no overlap! How do we rule out interactions?

- Concomitant variation also sucks, if treatment assignment is not random

- The potential outcomes framework is designed to be like using Mill’s method of difference many times, and averaging the outcomes.

- But what if you only have one treated case, and you want to know τ?
  This is where qualitative people feel they have the edge, using Mill’s methods

- But case studies are very fragile with regard to case selection

Assessing Unconfoundedness

A. Only 1 potential outcome is observable, so the unconfoundedness assumption is not testable!

B. Method 1: Compare two different control groups to see if, controlling for covariates, $\tau = 0$, as it should. (Imbens cites an example of using "ineligible for treatment" and "eligible nonparticipants," but this comparison poses overlap problems, as in the Lalonde data.) Remember, finding no treatment effect doesn't mean unconfoundedness is appropriate, but finding a treatment effect indicates it isn't.

C. Method 2: Use a covariate (especially a lagged outcome) as an outcome. Clearly you should not find a treatment effect, since all covariates must be exogenous. In the case of a lagged outcome, estimate the treatment effect on earnings one year prior to the program.

This is often called a "placebo test,"
but "robustness check" is probably a better term.

Placebo Tests (Cont.)

- This is almost more art than science.
- The number of donor pool males in the gender-salary states will determine your significance level:
  - If you don’t have enough donors, you will never hit your assigned value of statistical significance!

Returning to Our Salary Study at the Research Unit

- A First Look at Covariate Balance

Matching Test

- Matches were made using the person's
  - (1) initial appointment salary,
  - (2) title,
  - (3) and years since degree.

- As we suspected, the matching algorithm determined that there were only 3 sets of acceptable matches. In what follows we study these three matches.

Returning to the Salary Study at the Research Unit

Characteristics of the pre-test matches:

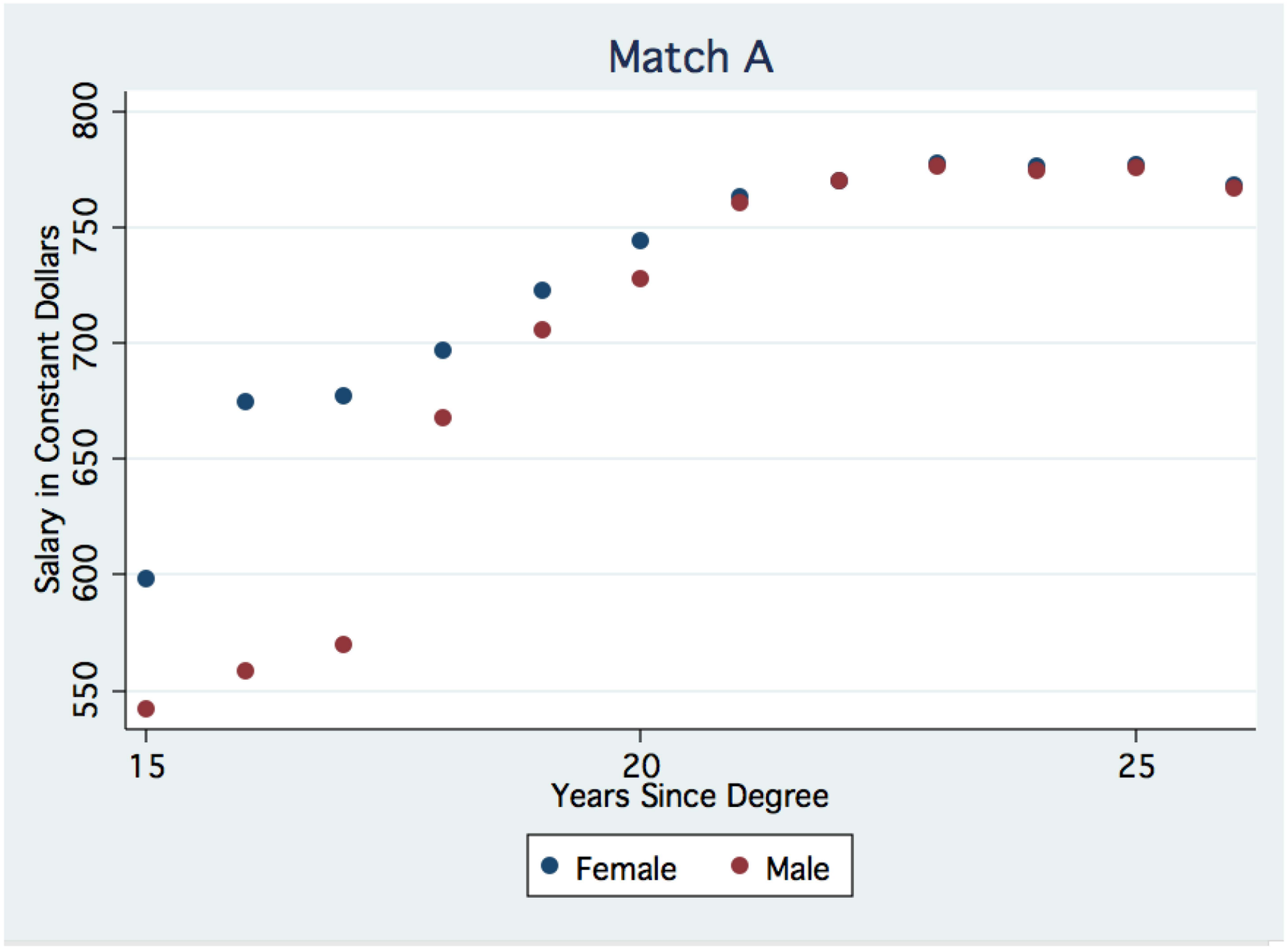
Match	ISM	Title in Year	Yrs Since Degree	Gender	Appt. Salary
Match A	22053	Associate Professor	15	F	224.2991
	27066	Associate Professor	15	M	249.2212
Match B	37206	Professor	30 (16 in Rank)	F	348.9097
	95964	Professor	28 (15 in rank)	M	336.4486
Match C	43963	Associate Professor	16	F	224.2991
	32041	Associate Professor	16	M	249.2212

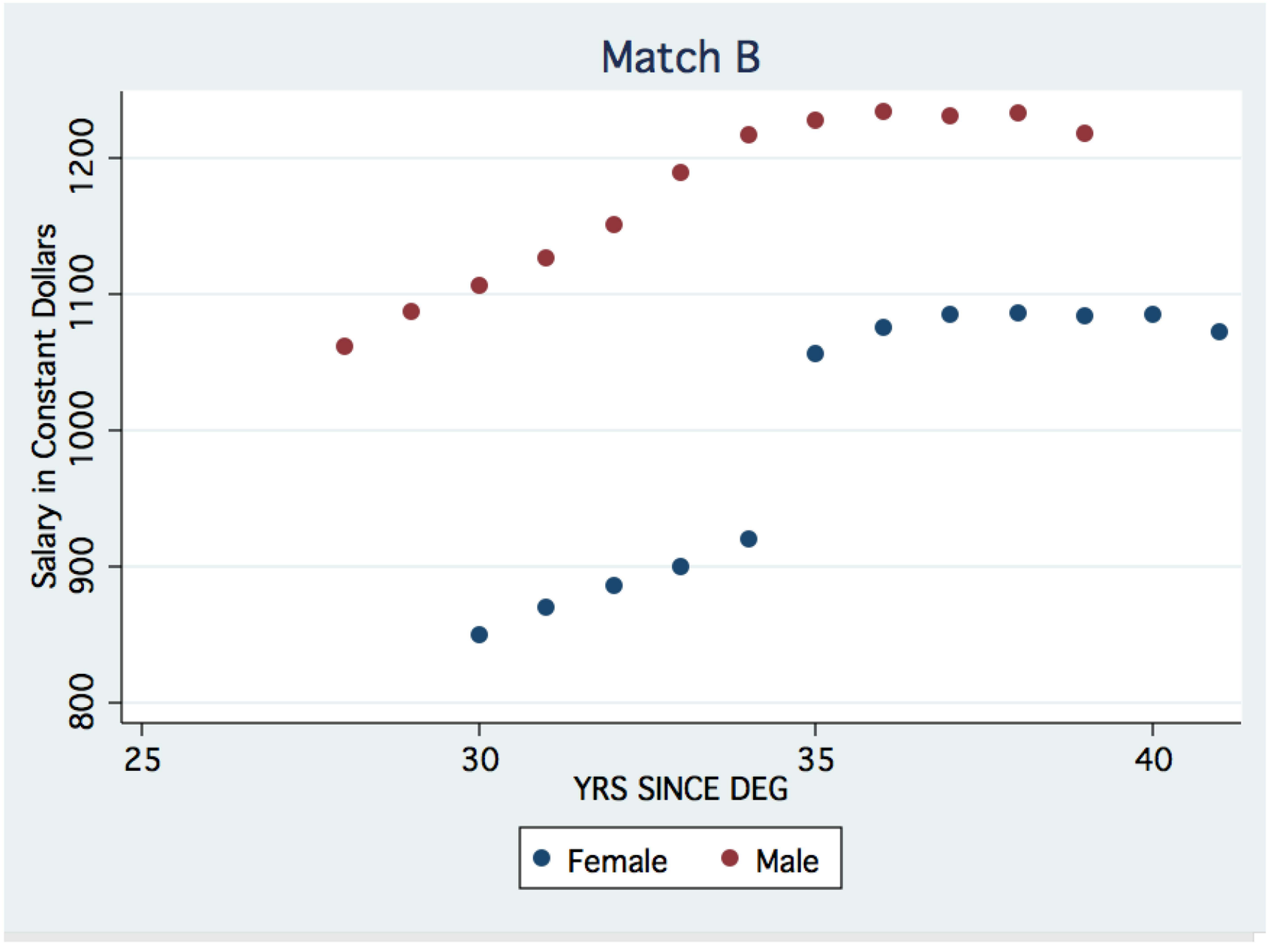
Note a nearest neighbor algorithm was used to match on initial appointment salaries and years since degree at the time the subject enters the dataset. Exact matches were made on title (ie if subject was an Associate Professor at the time they entered the dataset, then the subject was matched with another Associate professor).

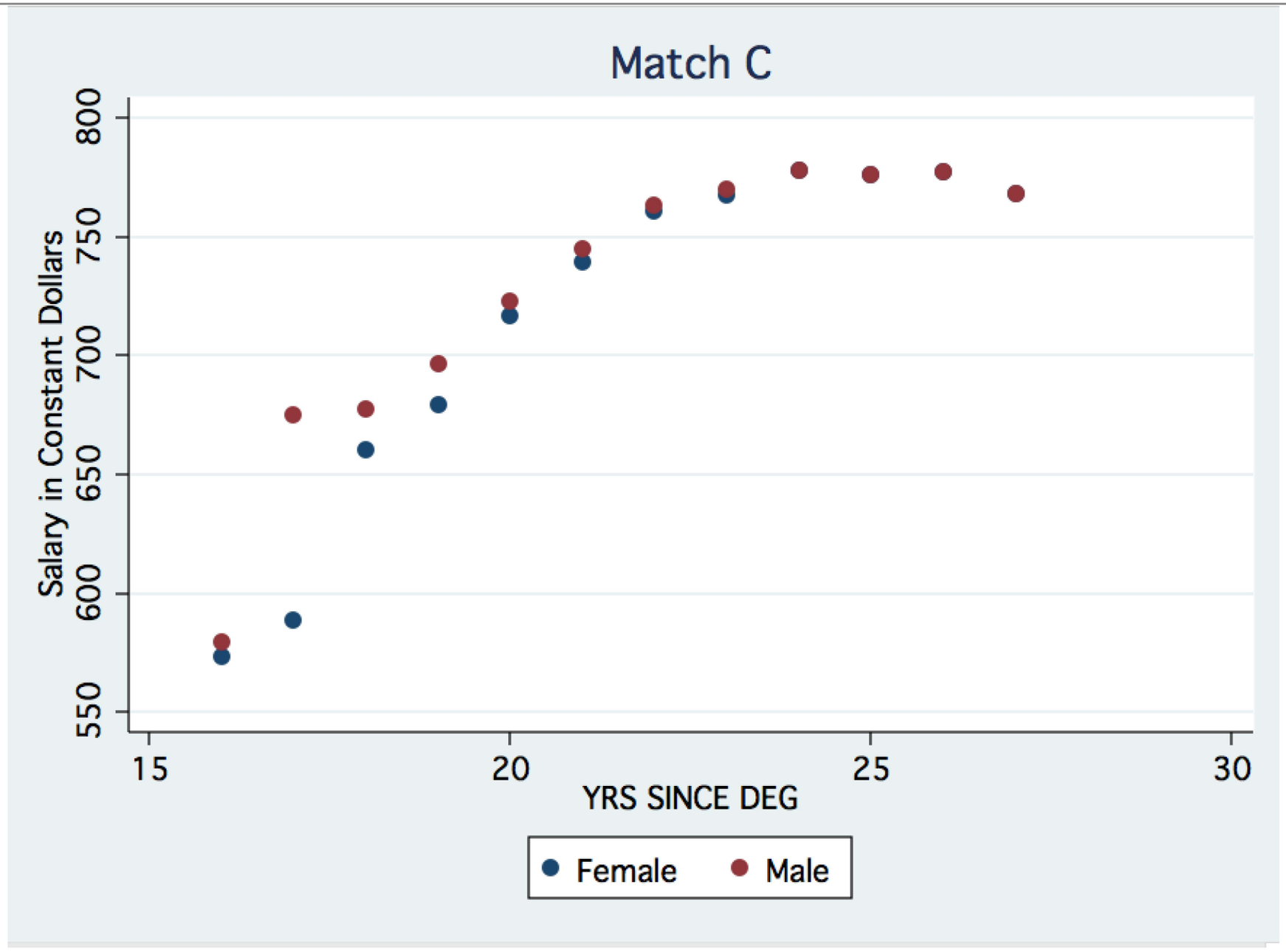
The following three scatter plots depict the salaries of the matched professors (Our DV) over our time series. The matches (A, B, C) correspond to the letter labels used to identify female observations in the other graphs.

The Same Subject Continued

- The graphs plot real salaries for A, B and C (the y-axis) and their matches for the years since degree (x-axis). Note the x-axis changes from one plot to another; for B the plot starts at 28 years for the male match and 30 years for the female B.
- What the graphs show is that the younger full professor women (A and C) are treated roughly equally to the males they were matched with.
- It also shows that Female B is treated much worse than the male she is matched to, given the variables we have to match on.
- We believe it is this difference for B that drives the matching regression later results on slide 12.


Match A
Salary in Constant Dollars
800
750
700
650
600
550
15
20
25
Years Since Degree
Female
Male


Match B
Salary in Constant Dollars
800
900
1000
1100
1200
25
30
35
40
YRS SINCE DEG
Female
Male

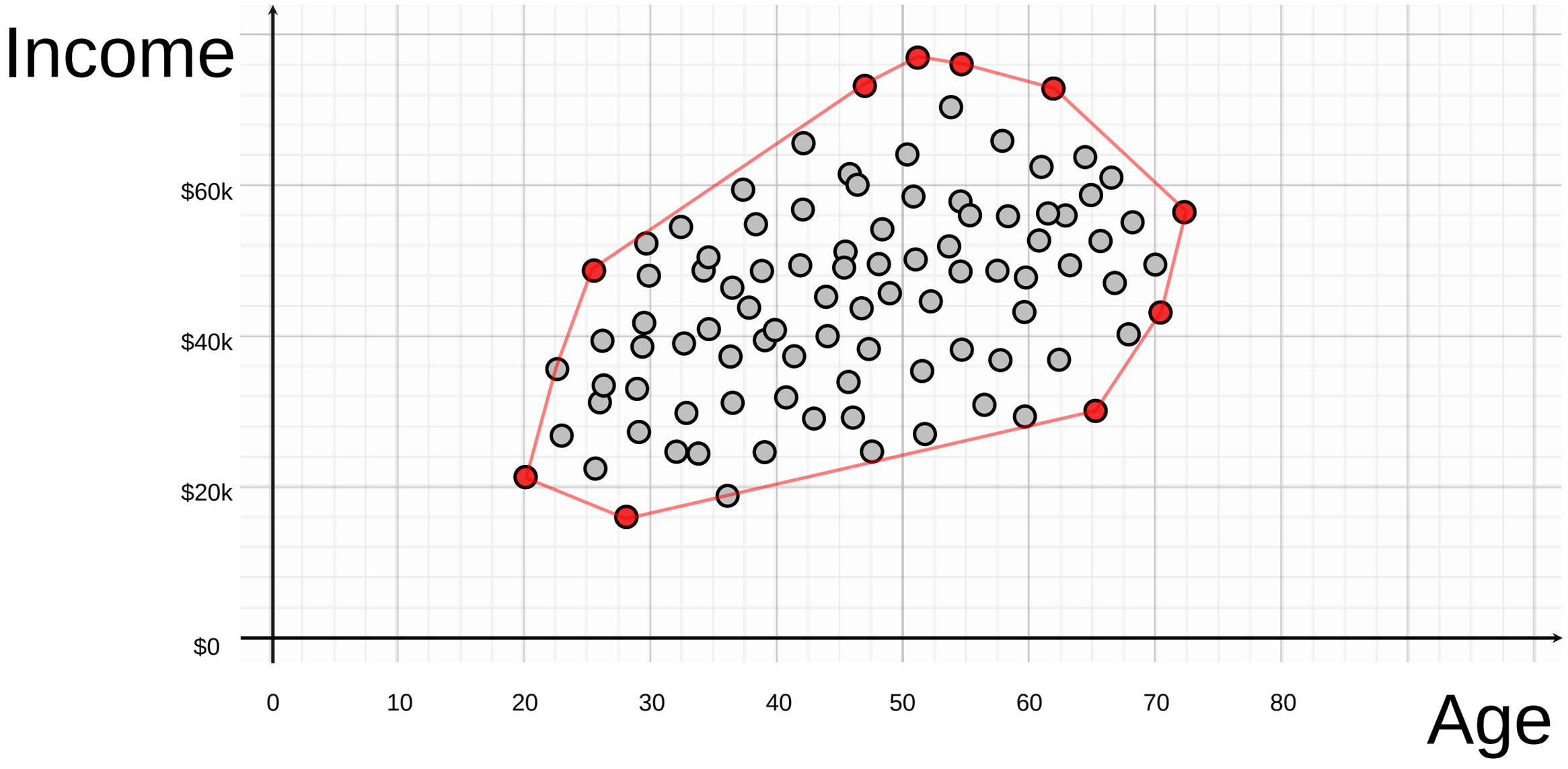

Match C
Salary in Constant Dollars
550
600
650
700
750
800
15
20
25
30
YRS SINCE DEG
Female
Male

A random-effects time-series, cross-sectional regression, using maximum likelihood, was conducted on the three pairs matched above. The effect of gender is negative and significant. Looking at the scatter plots of the matches above, it is obvious that these results are being driven solely by the differences in match B.

Synthetic Controls

Synthetic controls (Abadie et al 2009)

- It’s Mill’s method of difference where the control case is a synthesis of many other cases such that the synthesis is as similar as possible on the covariates
- Requires the treated case to be within (not lying on) the convex hull.



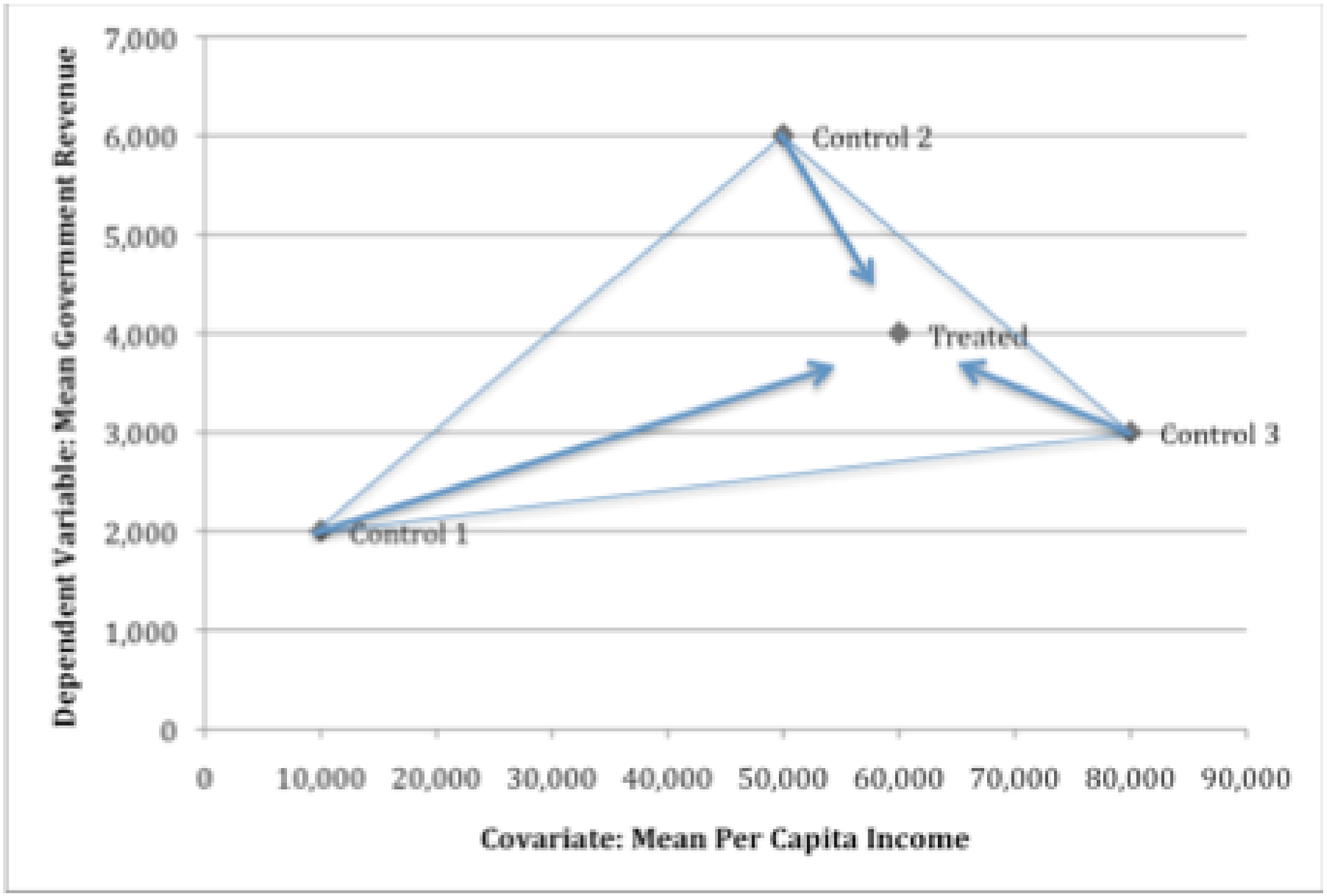
What are Synthetic Controls?

- Synthetic Control Methods are Matching Methods
  - Work similarly to other matching methods explained earlier
  - Instead of matching one-to-one, the algorithm creates a new control
    - Control is created after a search of units in a “donor pool”.
    - Control is created to minimize the distance between it’s trajectory and that of the treated unit.
      - This is analogous to a weighted average.
      - Average is taken pre-treatment.

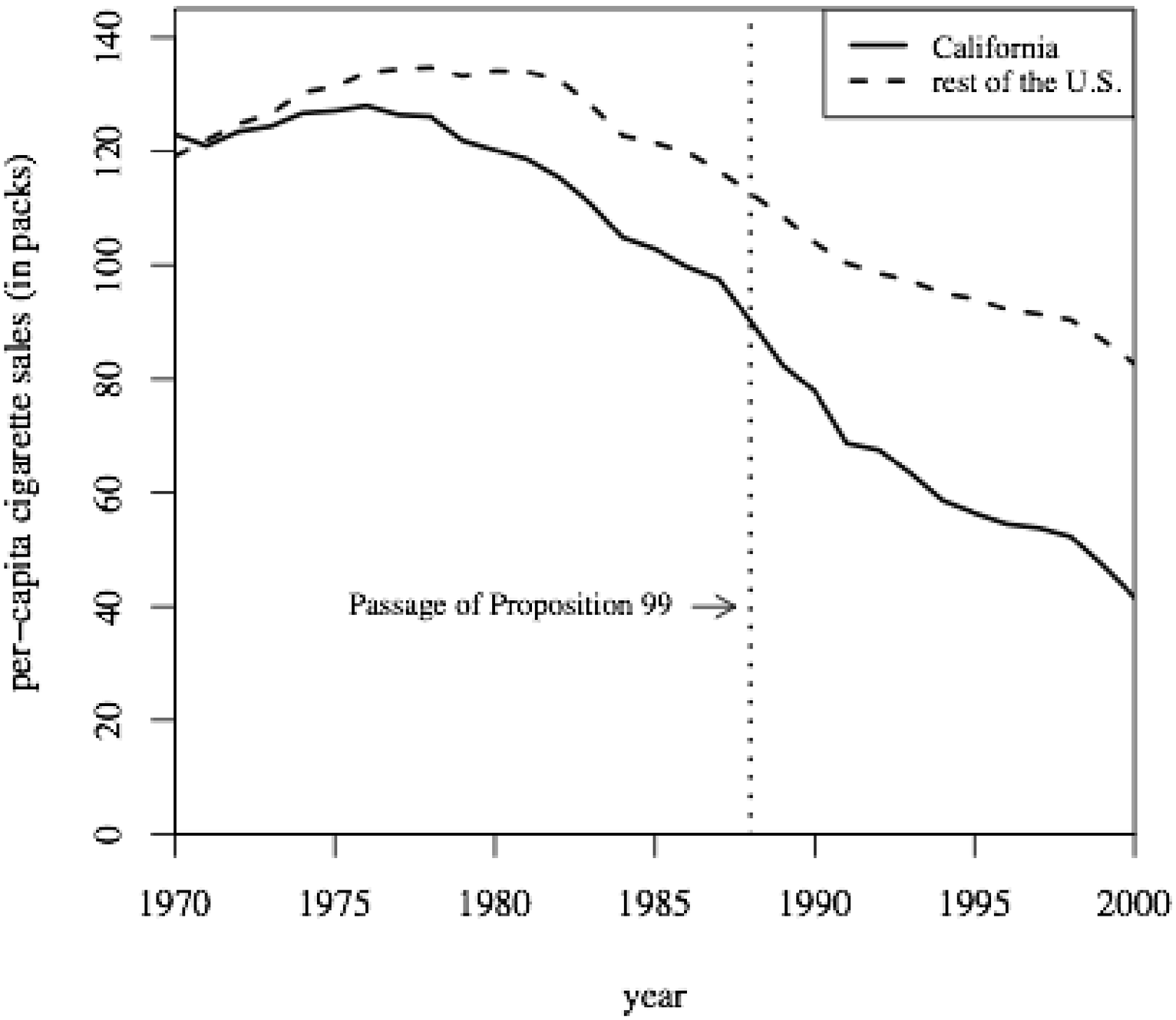
How do Synthetic Controls improve on typical matching algorithms?

- Synthetic Controls are completely data driven
    - The only theoretical constructs needed are covariate and donor pool selections.
- Synthetic Controls deal with Treatment Heterogeneity
    - They are event studies, so you are comparing Apples to Apples
- Using Synthetic Controls, covariate balance is maintained.
    - Balance is incorporated automatically into the weighting function (Hainmueller 2010).

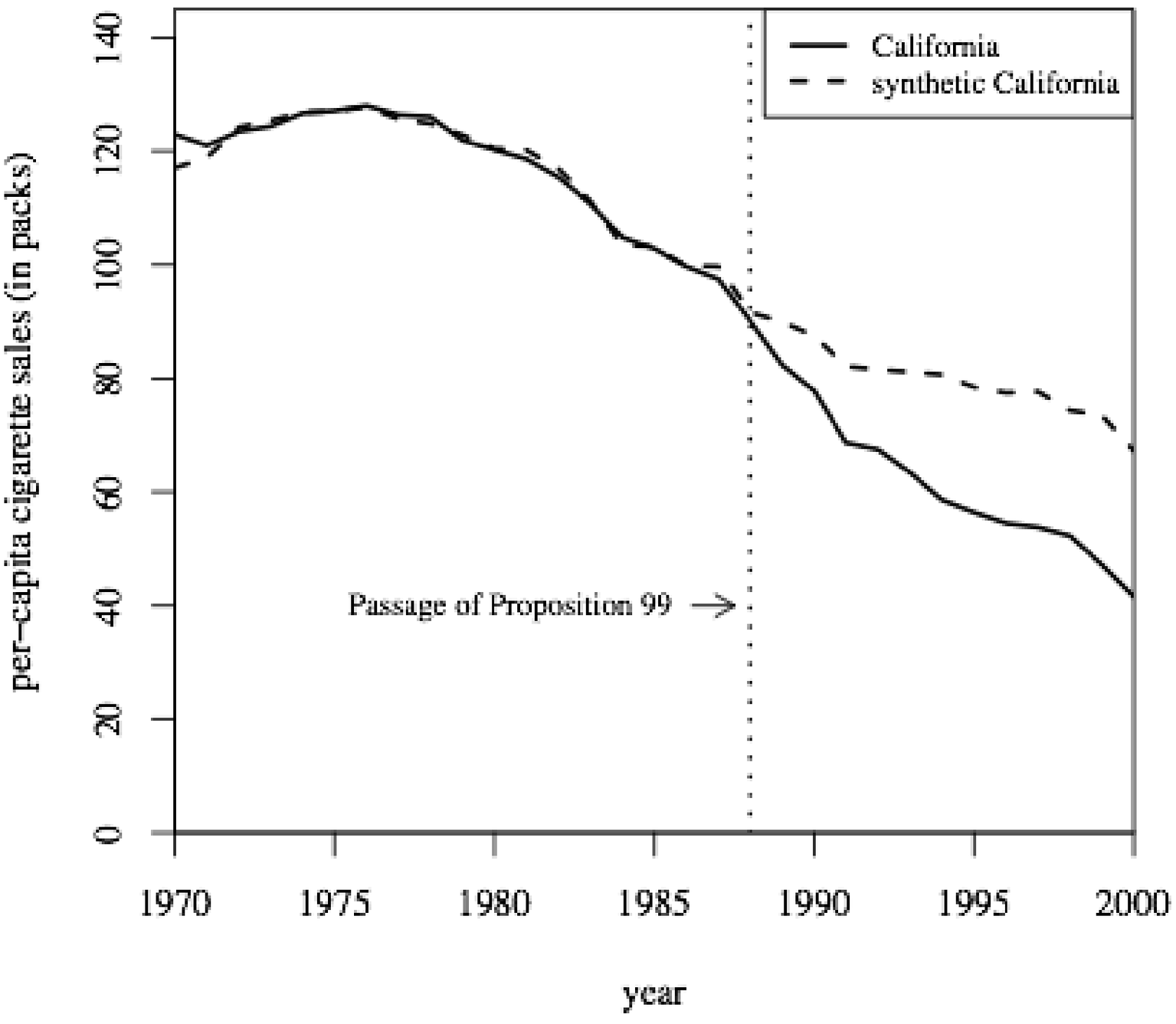
Synthetic Controls: A Simple Example



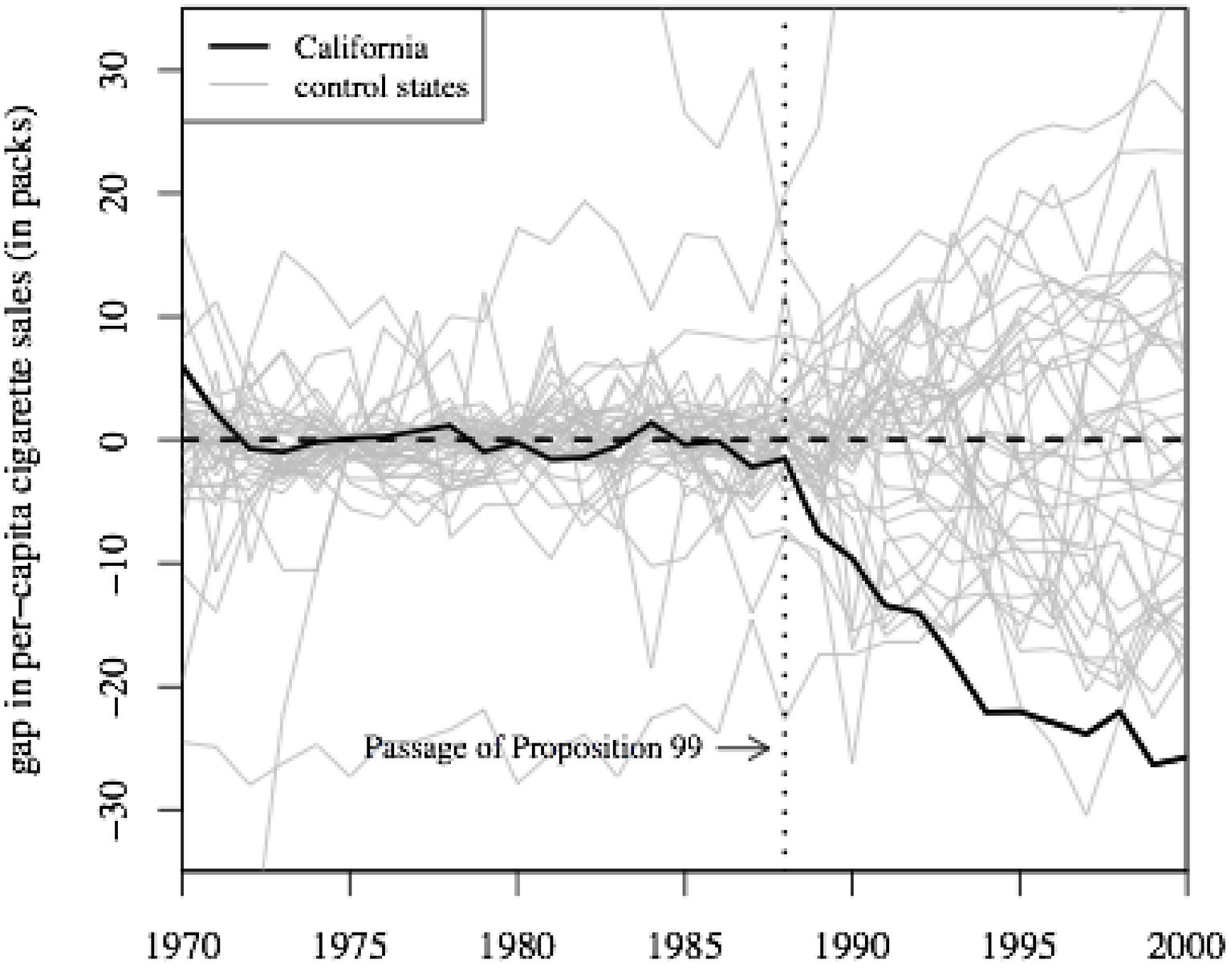
Synthetic controls (Abadie et al 2009)



Synthetic controls (Abadie et al 2009)



Synthetic controls (Abadie et al 2009)



Synthetic Controls and Covariate Balance

Keele, Malhotra & c. McCubbins forthcoming

OLS vs. Synthetic Control Methods

- KMM are trying to estimate a Causal Effect

- What advantages do you get out of this method that you can’t get out of standard OLS?

Well, let’s look at an example: Term Limits in the American States

OLS w/PCSE Results

Table 1 Effects of term limits on state expenditures, 1977–2001

	Exp. per capita	Exp. as a % of income
Term limits	59.8	0.004
	(20.9)***	(.001)***
Divided government	42.6	.001
	(13.4)***	(.0007)
Governor (1 = democrat)	32.8	.0009
	(12.7)***	(0.0007)
Grants	0.830	0.00003
	(0.115)***	(0.000006)***
Income	0.0000001	
	(0.0000001)***	
Population density	0.0008	0.00000004
	(0.0004)**	(0.00000002)**
Unemployment rate	0.276	0.00004
	(0.392)	(0.00002)***
Observations	1175	1175
Number of states	47	47
Number of years	25	25
R-squared	0.94	0.92

All estimates include fixed effects for states and years (not reported)

Entries are regression coefficients from panel data regression models. Figures in parenthesis are panel corrected standard errors

* $p < 0.1$

** $p < 0.05$

*** $p < 0.01$

Some Problems with the Panel Method

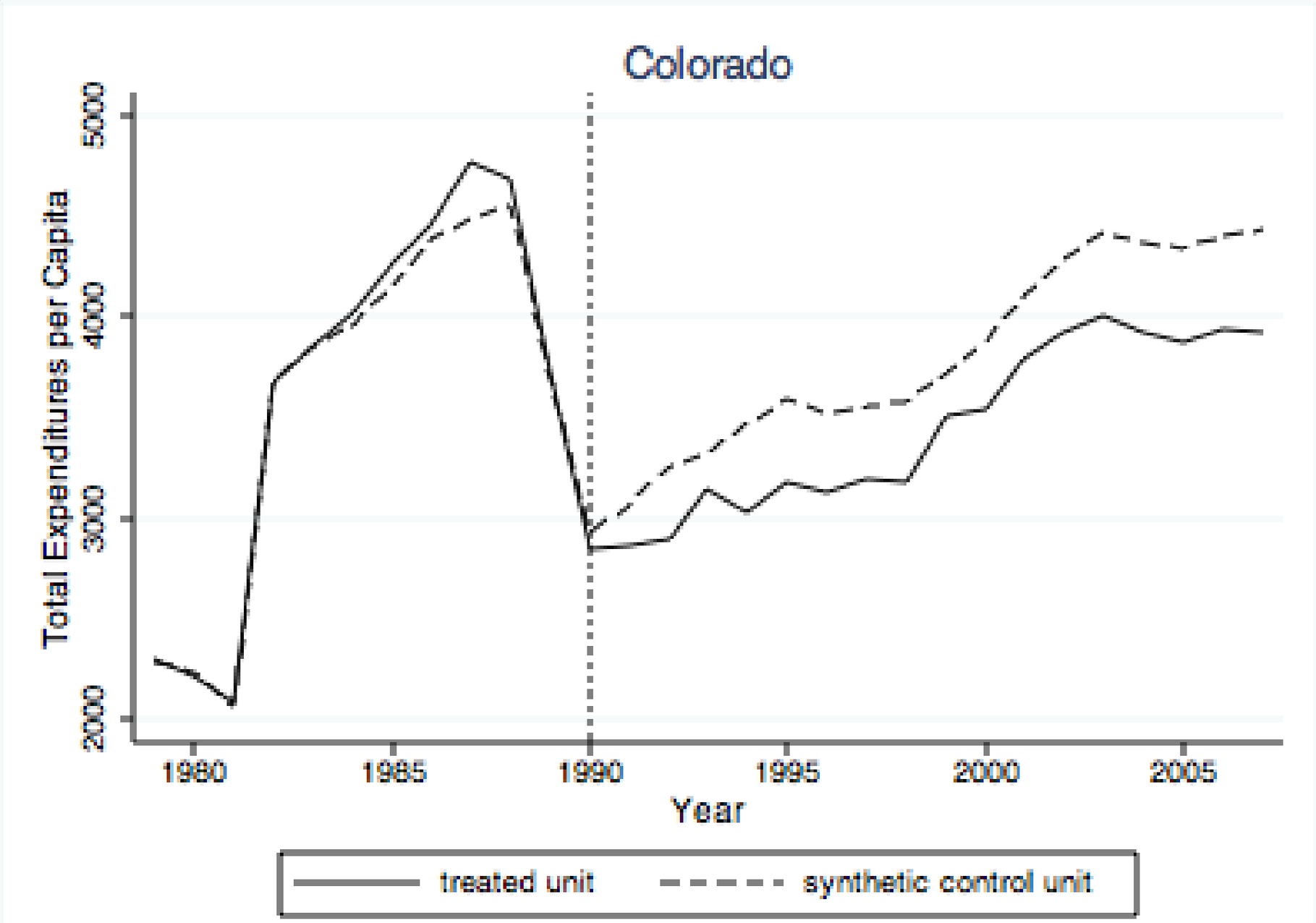
- Treatment Heterogeneity
  - Are term limits actually the same in every state? They are treated as such in the model.
- Overlap
  - “In every state with the initiative process, except one, voters have passed some form of term limit legislation. Similarly, no state without the initiative process currently has term limits for their state representatives.” – Erler pg. 484
  - This indicates that there is **no overlap**, which in turn indicates that you cannot make a causal inference because the treatment is not as if randomly assigned. OLS does not deal effectively with any of these problems.

Regression of all Pertinent Variables

.regress ipoltotalexppc income unemployment schoolage ovr65age gsp popgrowth population populationdensity governors dividedgov seats_h seat_s grants emptot empwagesalary emppriv fedempciv fedempmil govtempsandl I_GovLimits II_GovLimits III_GovLimits I_termlimits_h II_termlimits_h III_termlimits_h I_termlimits_s II_termlimits_s III_termlimits_s

```
•
•      Source |       SS       df       MS              Number of obs =    1206
•-------------+------------------------------           F( 27,  1178) =  237.15
•       Model |  .001352494    27  .000050092           Prob > F      =  0.0000
•    Residual |  .000248827  1178  2.1123e-07           R-squared     =  0.8446
•-------------+------------------------------           Adj R-squared =  0.8411
•       Total |  .001601321  1205  1.3289e-06           Root MSE      =  .00046
•
•------------------------------------------------------------------------------
•ipoltotale~c |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
•-------------+----------------------------------------------------------------
•      income |   1.45e-07   3.33e-09    43.57   0.000     1.39e-07    1.52e-07
•unemployment |  -.0000102   .0000102    -1.00   0.316    -.0000301    9.73e-06
•   schoolage |  -.0065554   .0013921    -4.71   0.000    -.0092868   -.0038241
•    ovr65age |  -.0054618   .0012105    -4.51   0.000    -.0078368   -.0030868
•         gsp |  -2.81e-09   5.86e-10    -4.79   0.000    -3.95e-09   -1.66e-09
•   popgrowth |  -.0227101   .0015114   -15.03   0.000    -.0256754   -.0197448
•  population |   2.80e-10   5.35e-11     5.22   0.000     1.75e-10    3.85e-10
•population~y |  -6.36e-07   7.49e-08    -8.49   0.000    -7.83e-07   -4.89e-07
•I_GovLimits |    -.00019   .0000432    -4.40   0.000    -.0002747   -.0001052
•II_GovLimits |  -.0001447   .0000351    -4.12   0.000    -.0002136   -.0000758
•III_GovLim~s |  -.0002824    .000058    -4.87   0.000    -.0003963   -.0001686
•I_termlimi~h |   .0003074   .0002353     1.31   0.192    -.0001544    .0007691
•II_termlim~h |  -.0005899   .0001418    -4.16   0.000    -.0008682   -.0003116
•III_termli~h |    .000308   .0001705     1.81   0.071    -.0000266    .0006425
•I_termlimi~s |  -.0005022   .0002559    -1.96   0.050    -.0010043   -8.19e-08
•II_termlim~s |  (dropped)
•III_termli~s |   .0000811   .0002008     0.40   0.686    -.0003129    .0004752
•       _cons |   .0029022   .0004429     6.55   0.000     .0020332    .0037712
•
```

Term Limits: Synthetic Controls (KMM forthcoming)



Case: Weights and Balance Output

Weights:

Unit Weights:

Co_No	Unit_Weight
5	0
9	.115
10	0
12	.238
15	0
17	0
18	0
19	0
20	.08
23	0
24	0
25	0
27	.293
29	0
30	0
31	0
33	0
35	0
39	0
41	0
45	.274
48	0

Balance:

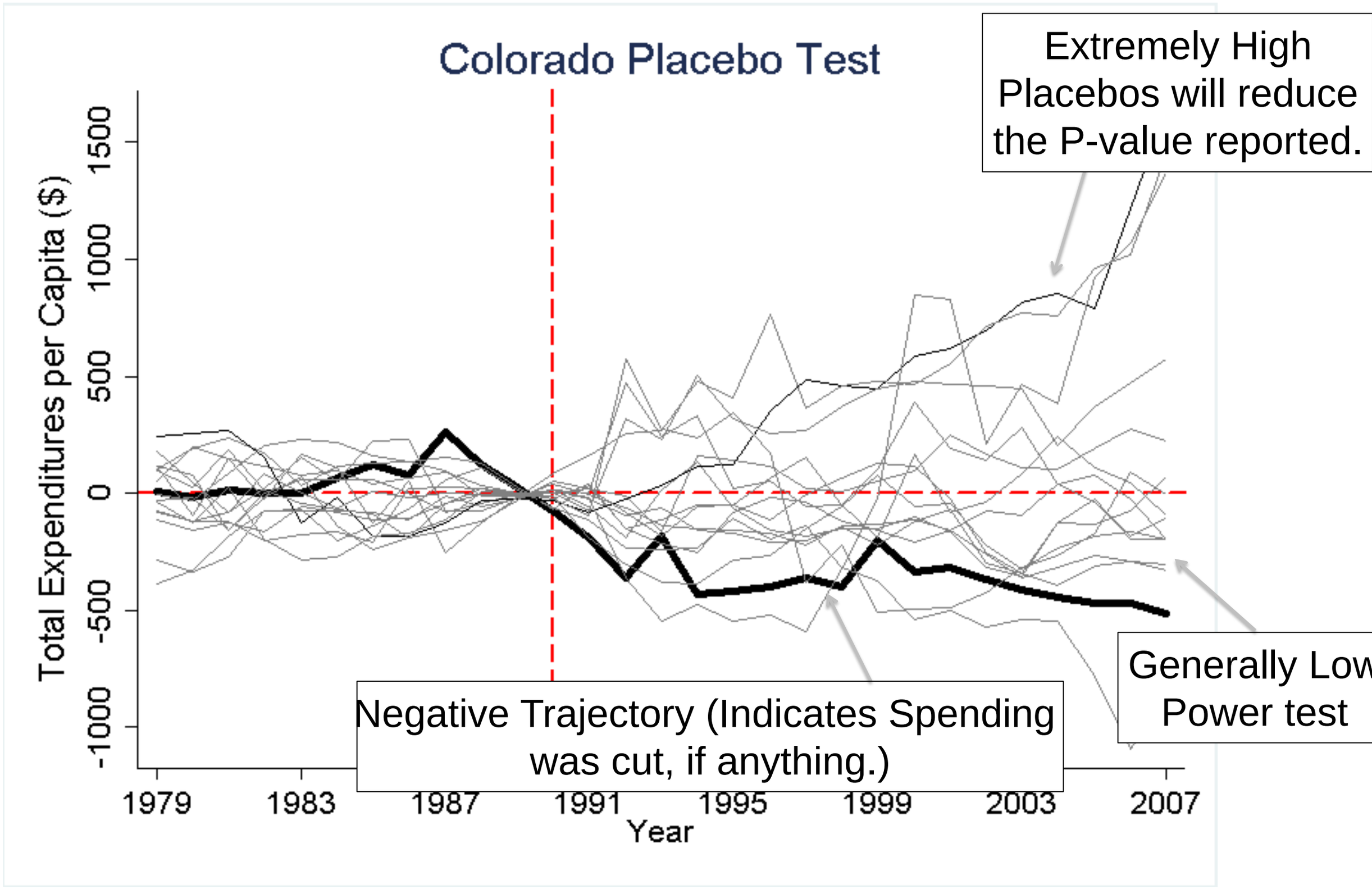
Predictor Balance:

	Treated	Synthetic
totalpopulation	2831045	2338199
pop5to17_pc	.2104775	.2122493
pop65plus_pc	.1141638	.1220863
govdummy	.9411765	.7261176
demhousecontrol	.0588235	.1391176
demsenatecontrol	.1764706	.1541176
percapinc_cpi	114.7965	114.5907

Making Inference - What are Placebo Tests?

- Type of Randomization Inference
- Iteratively applying the Synthetic Control Method to the donor pool units
    - What would have happened to these units if the now counterfactual treatment had been applied?
        - This is a comparison of trajectories.
- You are comparing the treated case to the donor pool counterfactual cases and seeing where in the distribution of cases the treated case is.
    - If the treated case is significantly different than the other cases, we have evidence that the treatment had an effect.
    - E.g.: (Abadie et al 2007) California was the lowest out of 39 states in the study. The probability of seeing a California-like result was thus 1/39 (0.02, which is significant).

Term Limits: Placebo Tests



Term Limits: Reported P-value

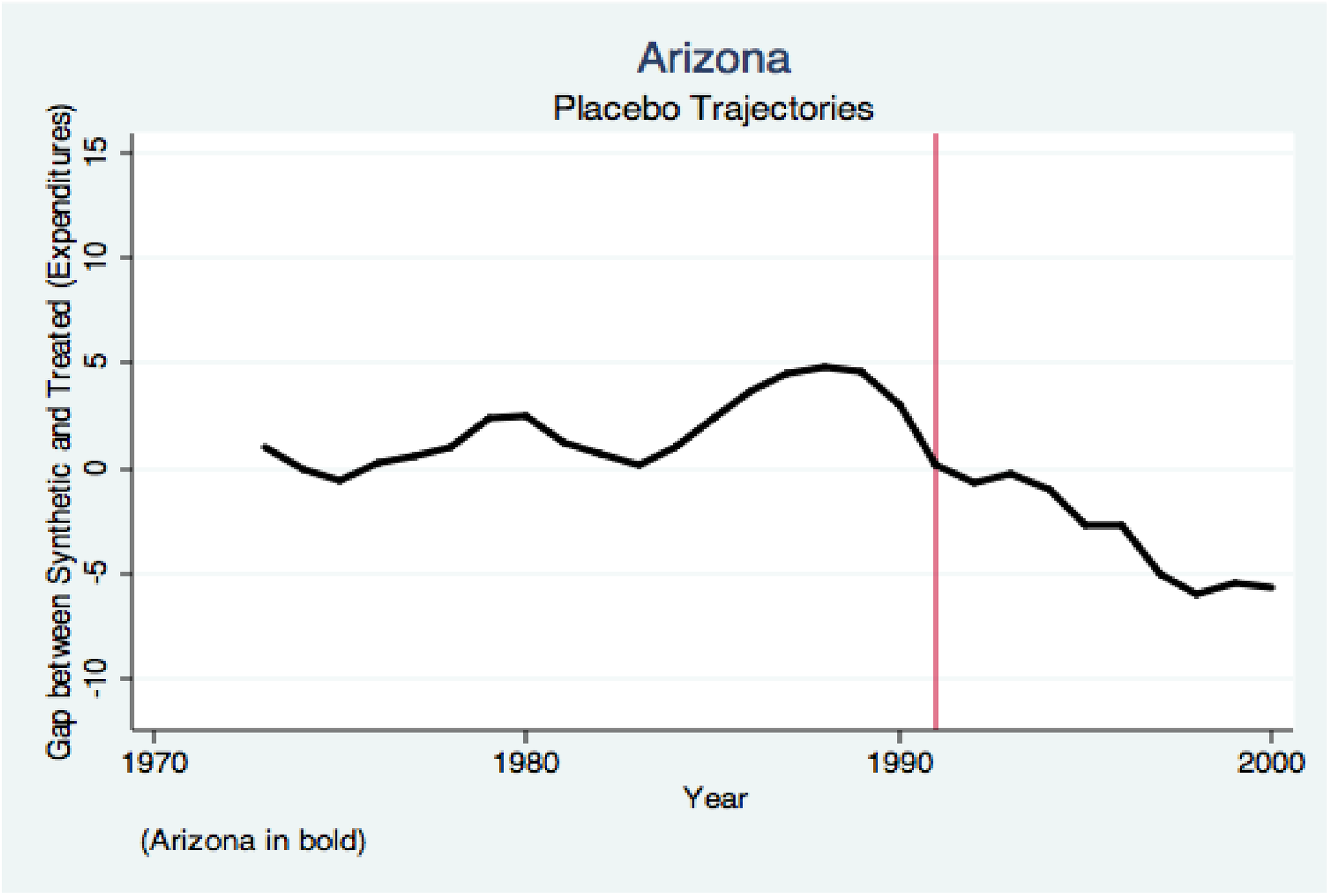
- The reported p-value for Colorado is 0.18, which is not statistically significant (even at the more generous 0.1 level).

- When we perform this analysis, the effects from the OLS disappear. We have no evidence that term limits had any effect on spending one way or the other.

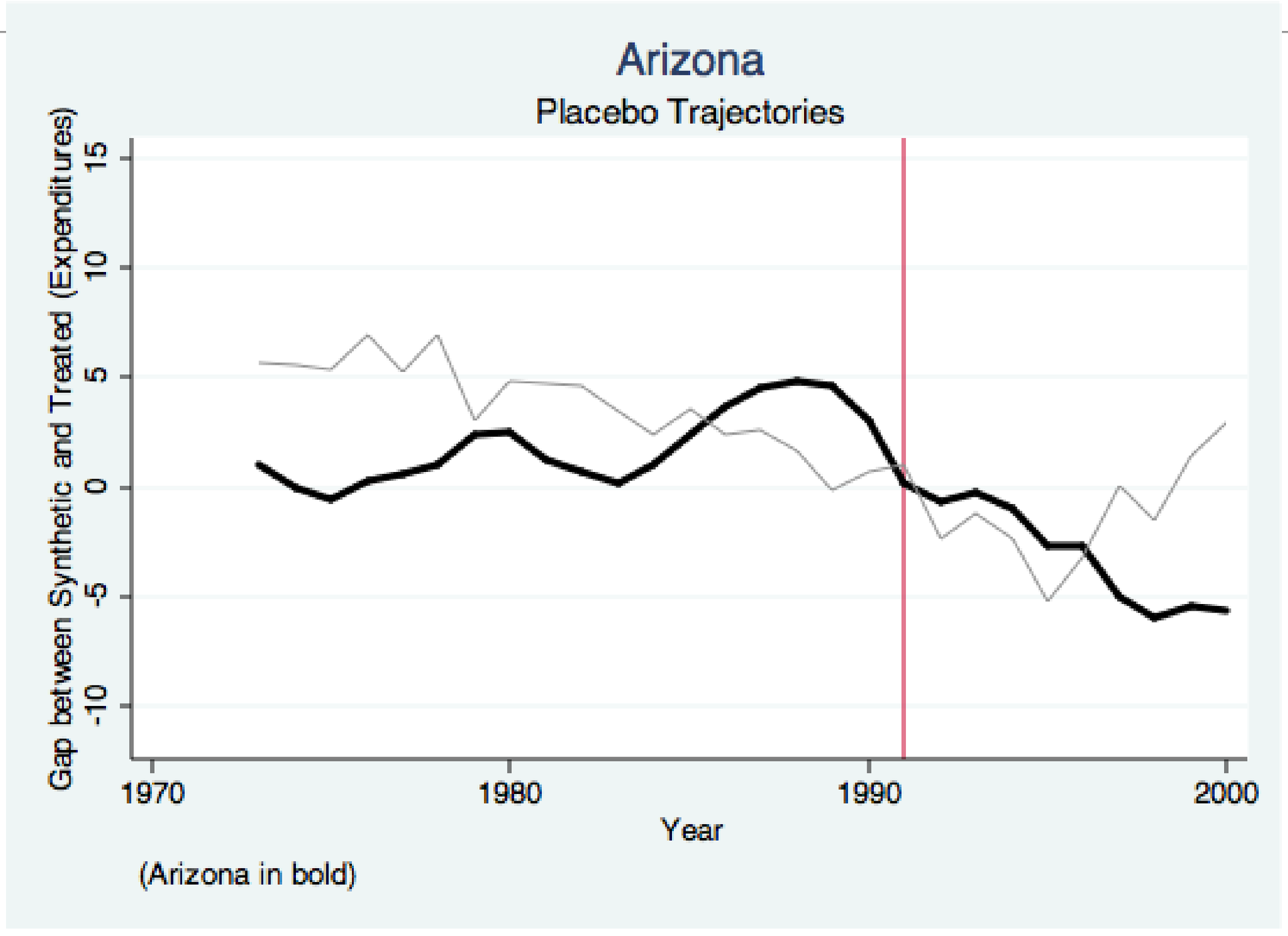
A Related Study: Budget Stabilization Funds

- Studying these policies have all the same problems

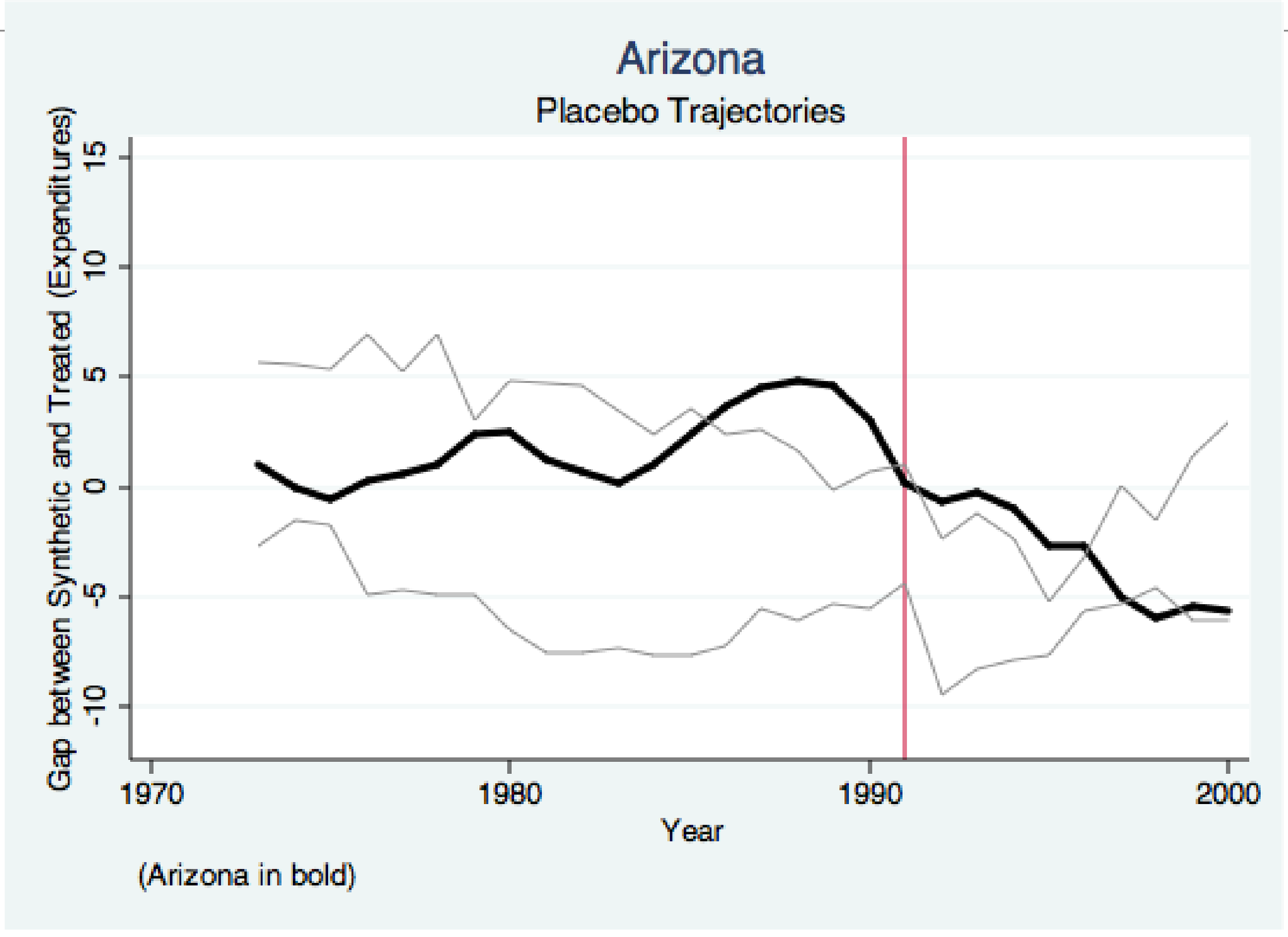
Case: Arizona – Constructing Placebos



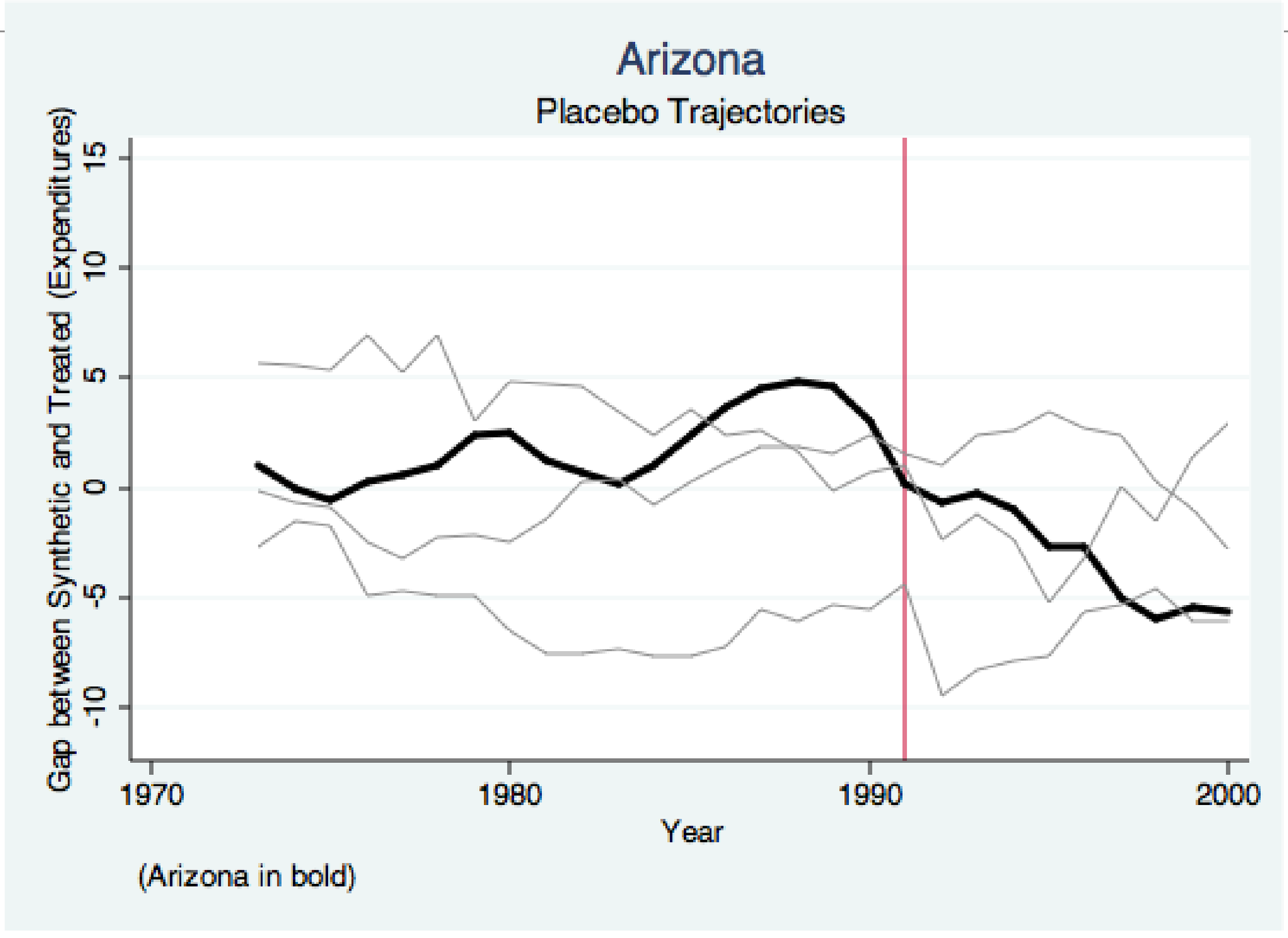
Case: Arizona – Constructing Placebos



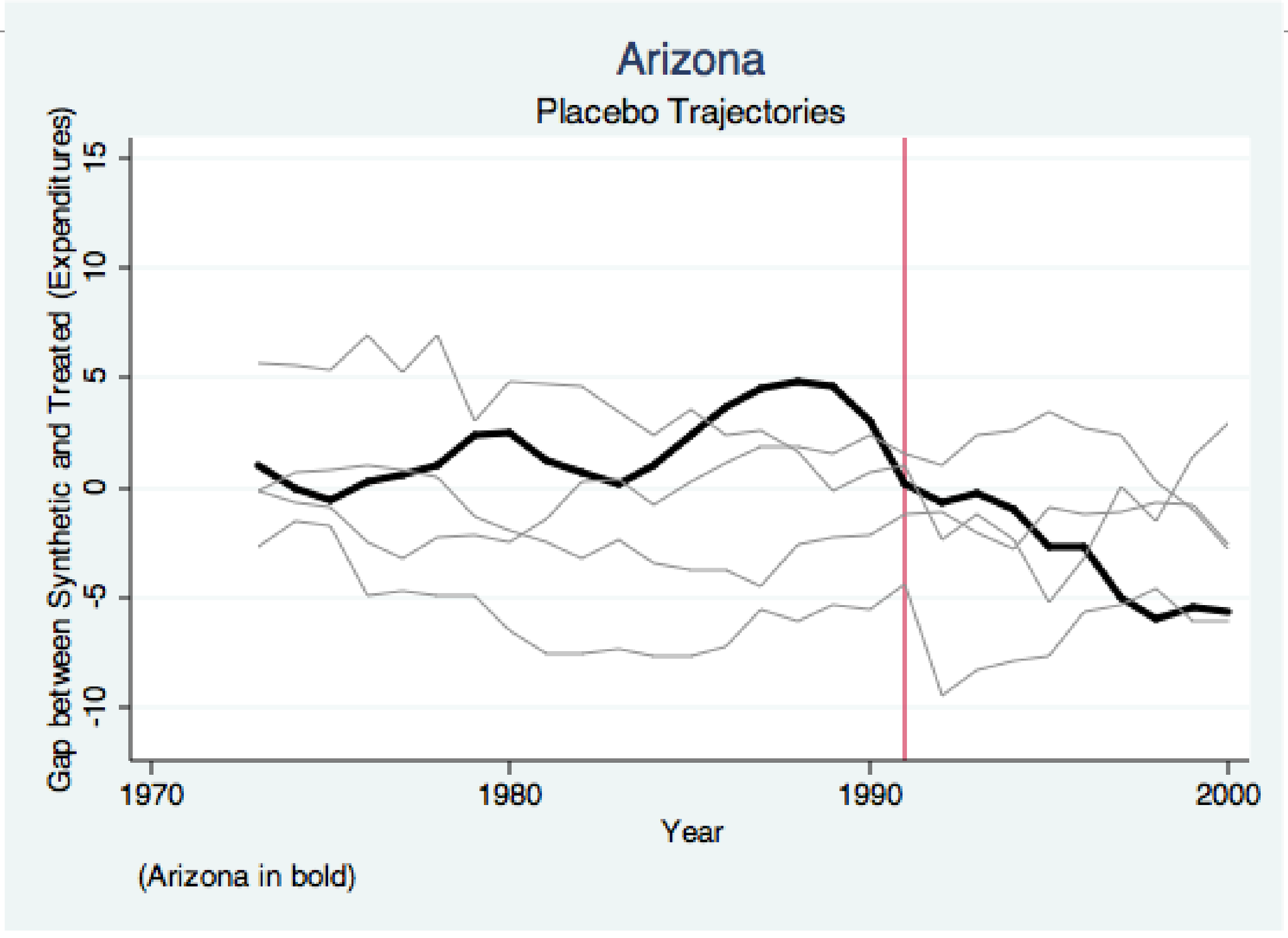
Case: Arizona – Constructing Placebos



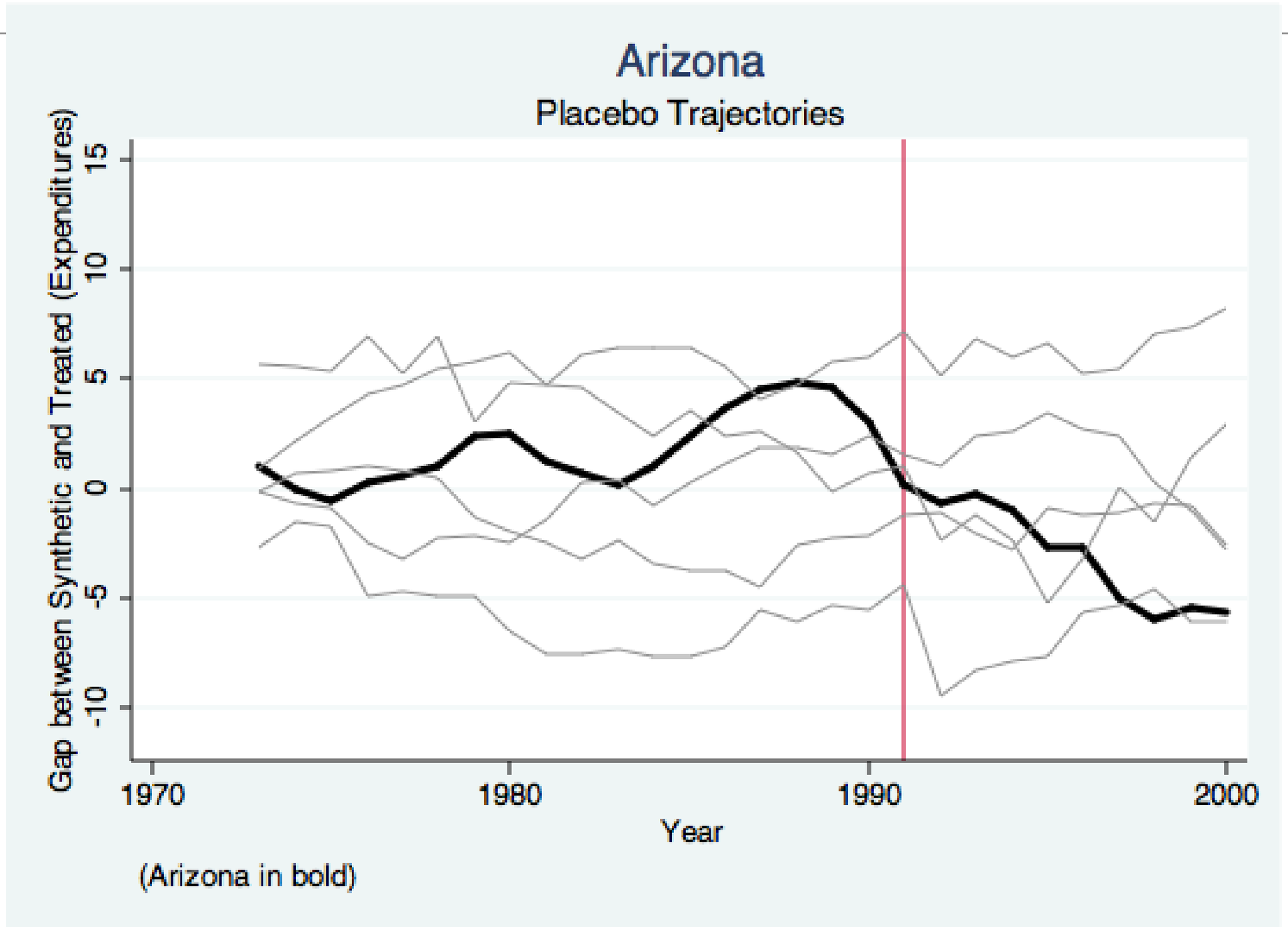
Case: Arizona – Constructing Placebos



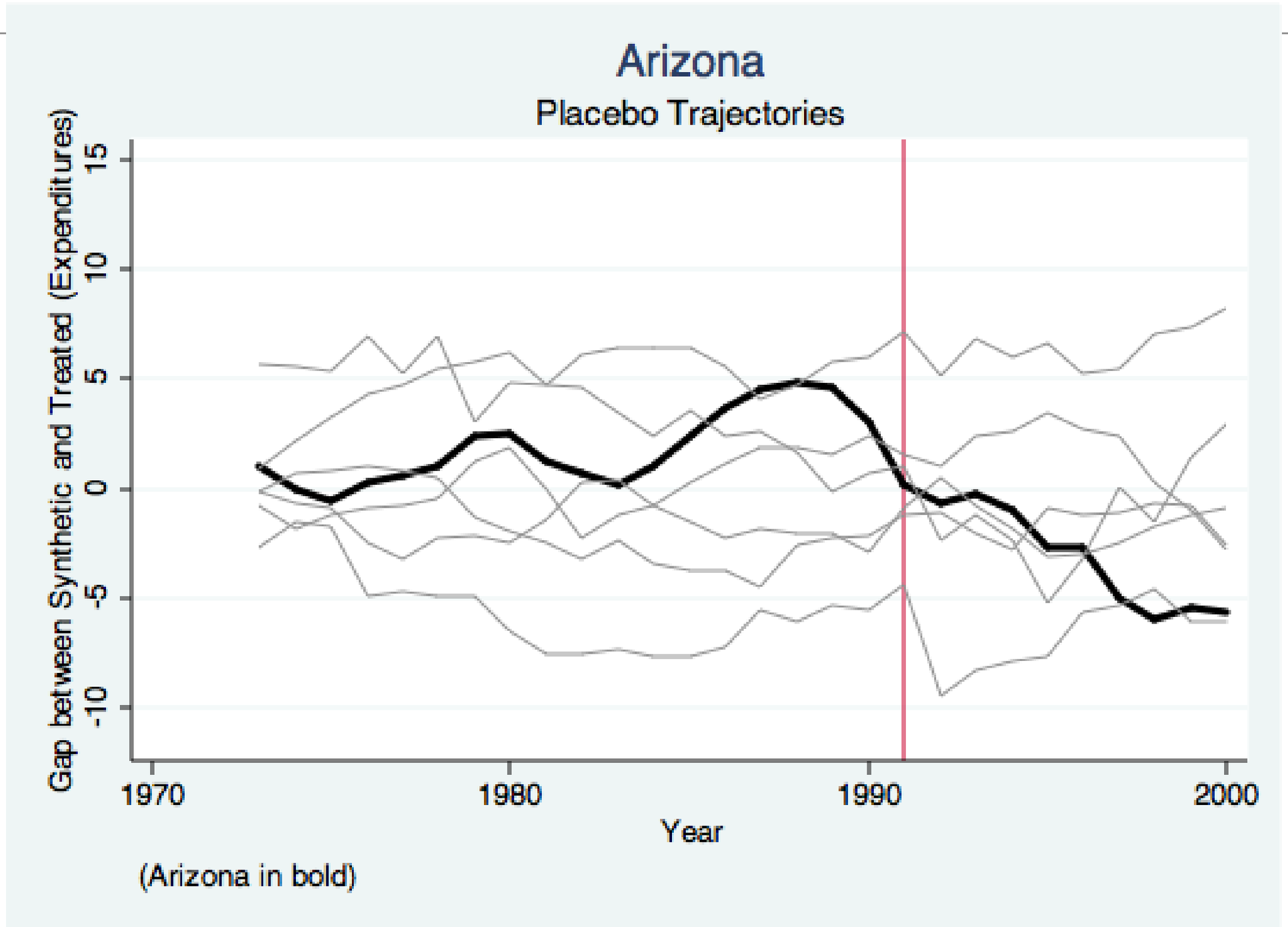
Case: Arizona – Constructing Placebos



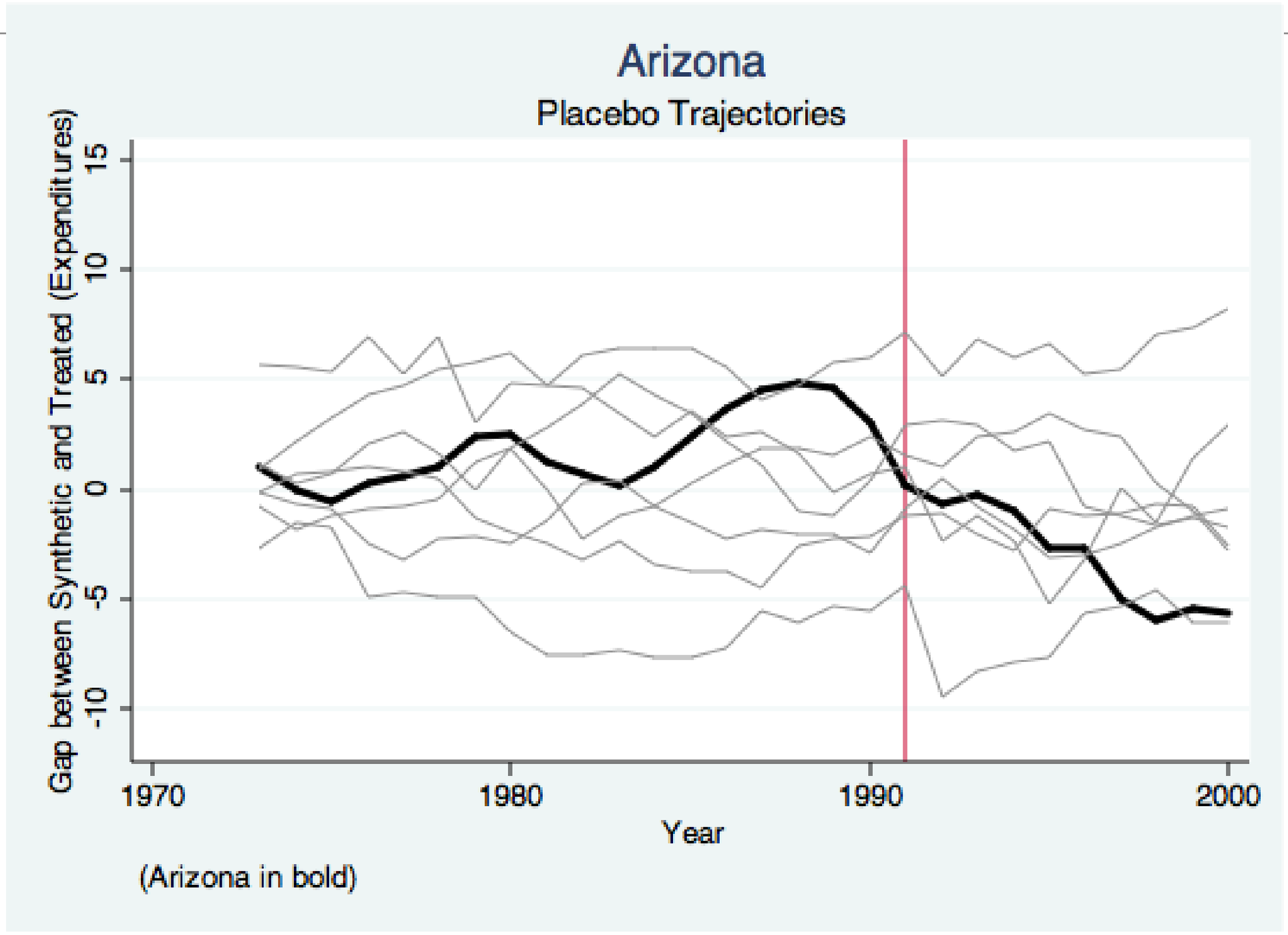
Case: Arizona – Constructing Placebos



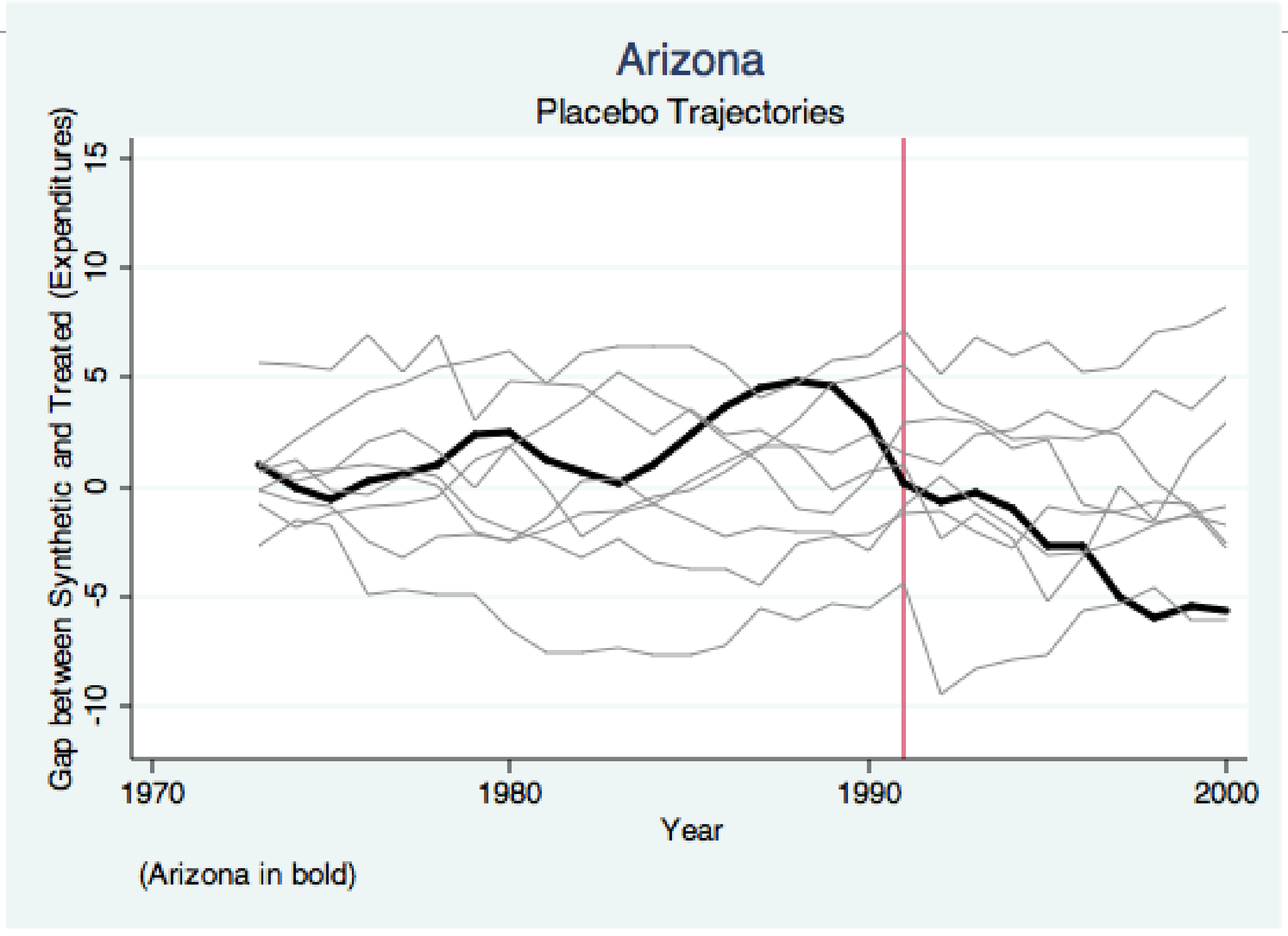
Case: Arizona – Constructing Placebos



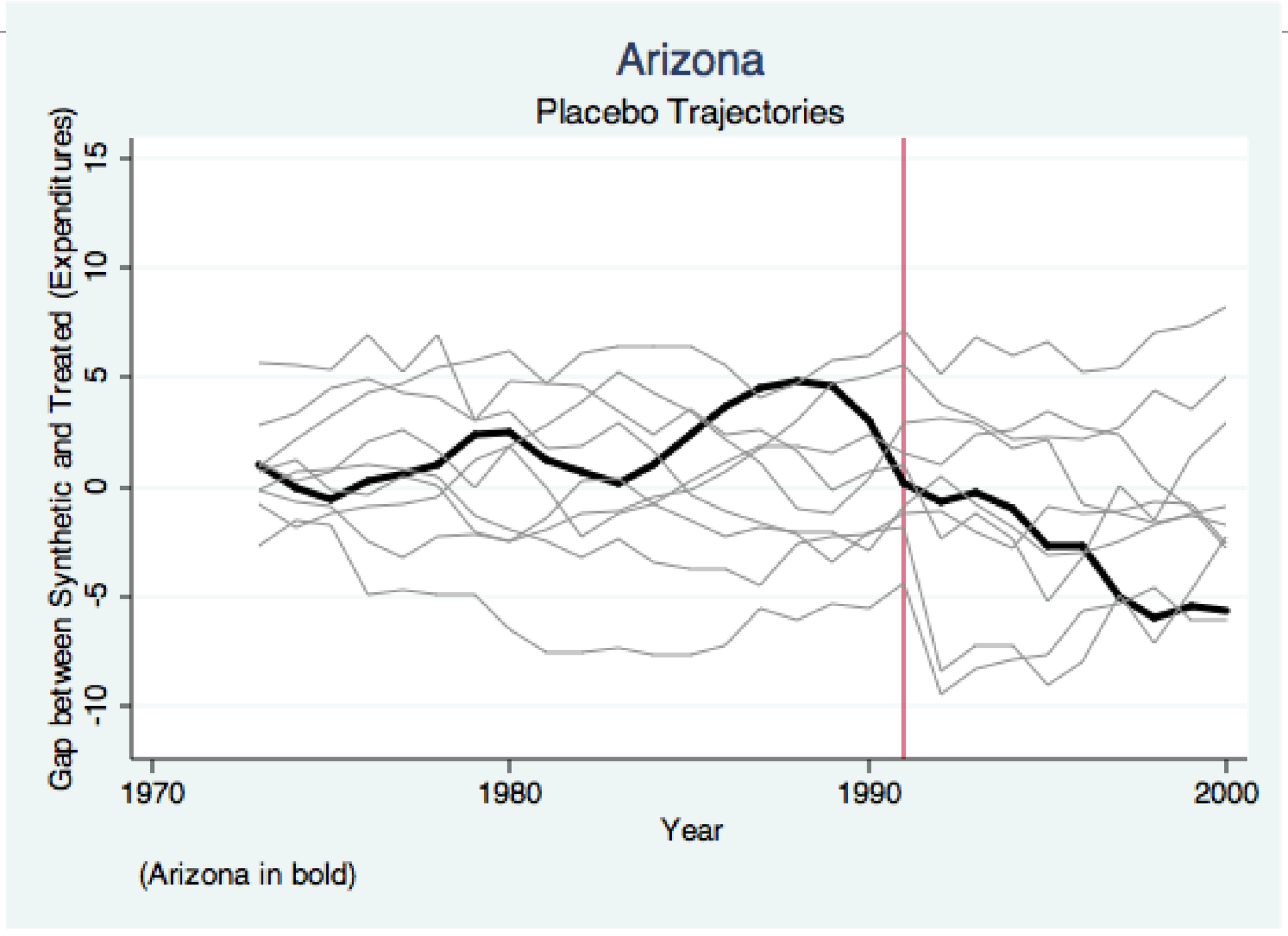
Case: Arizona – Constructing Placebos



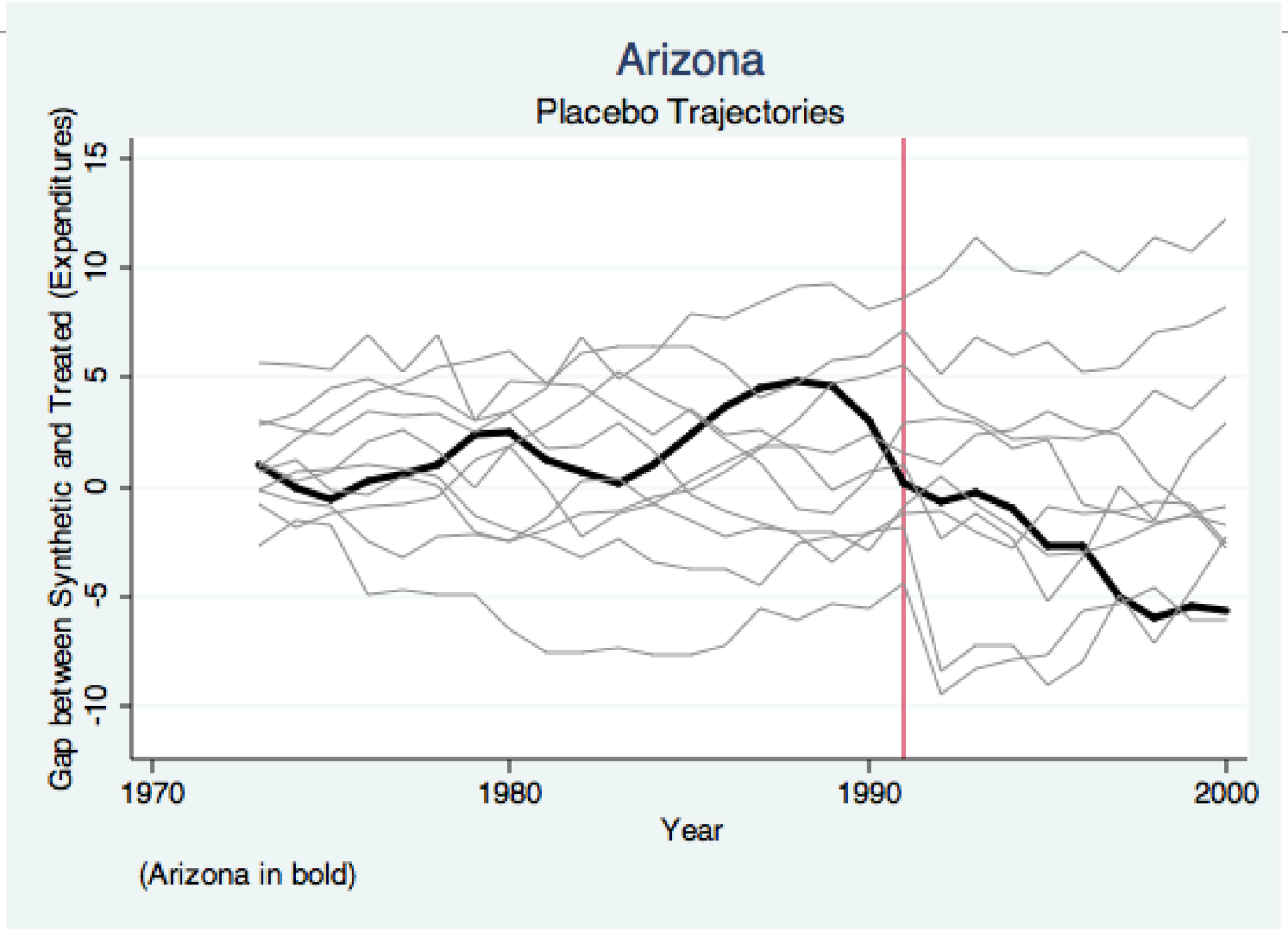
Case: Arizona – Constructing Placebos



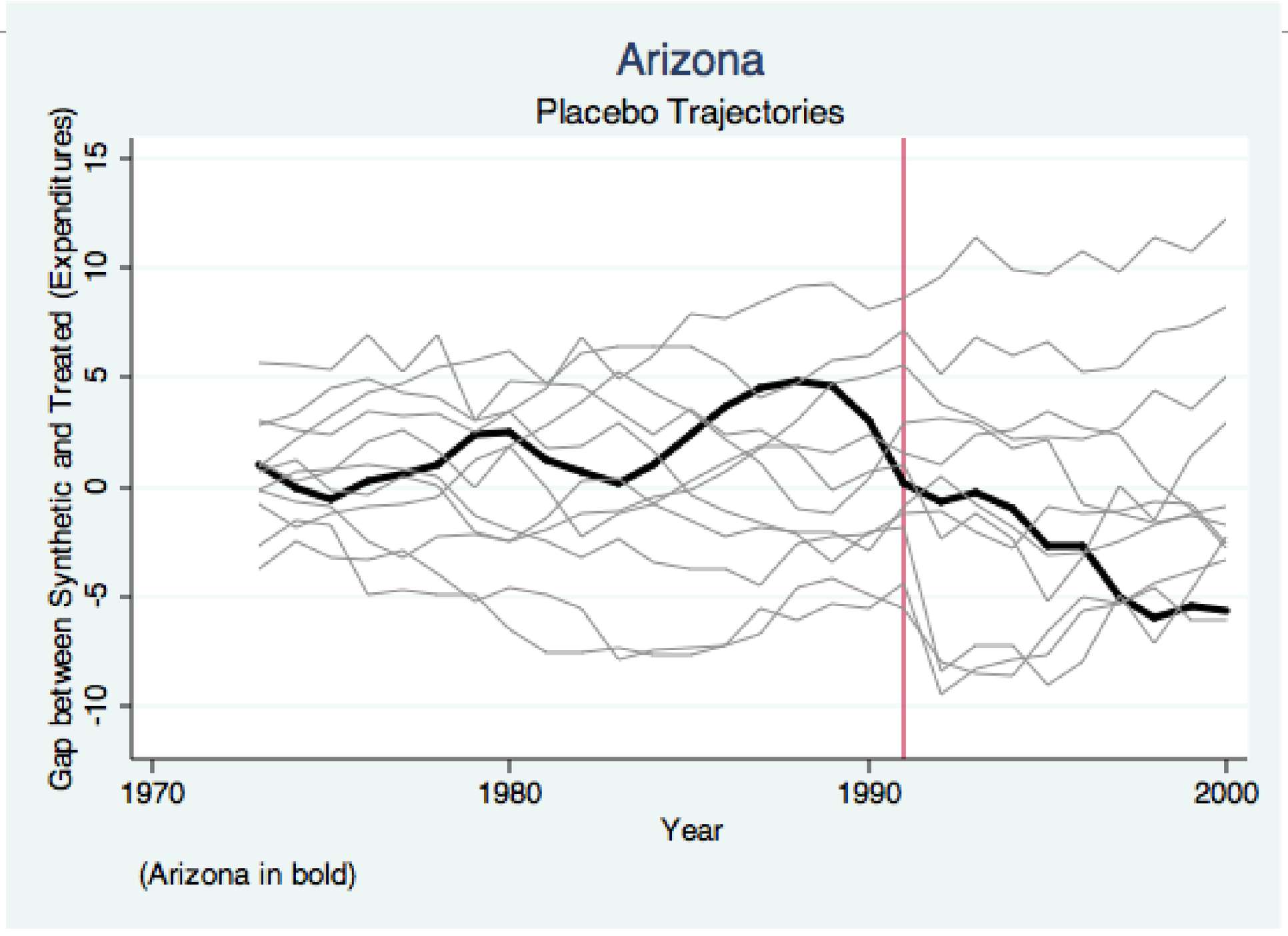
Case: Arizona – Constructing Placebos



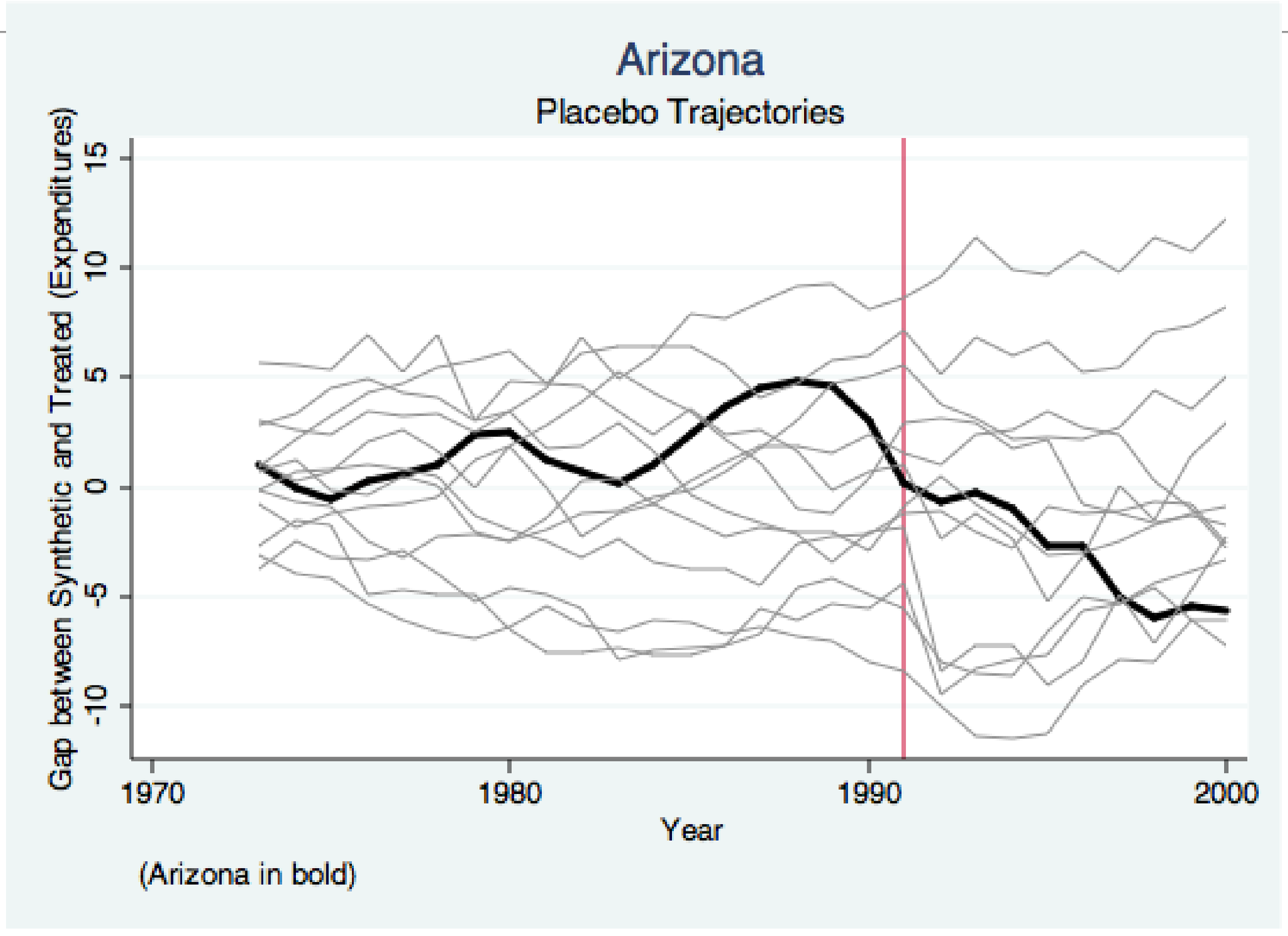
Case: Arizona – Constructing Placebos



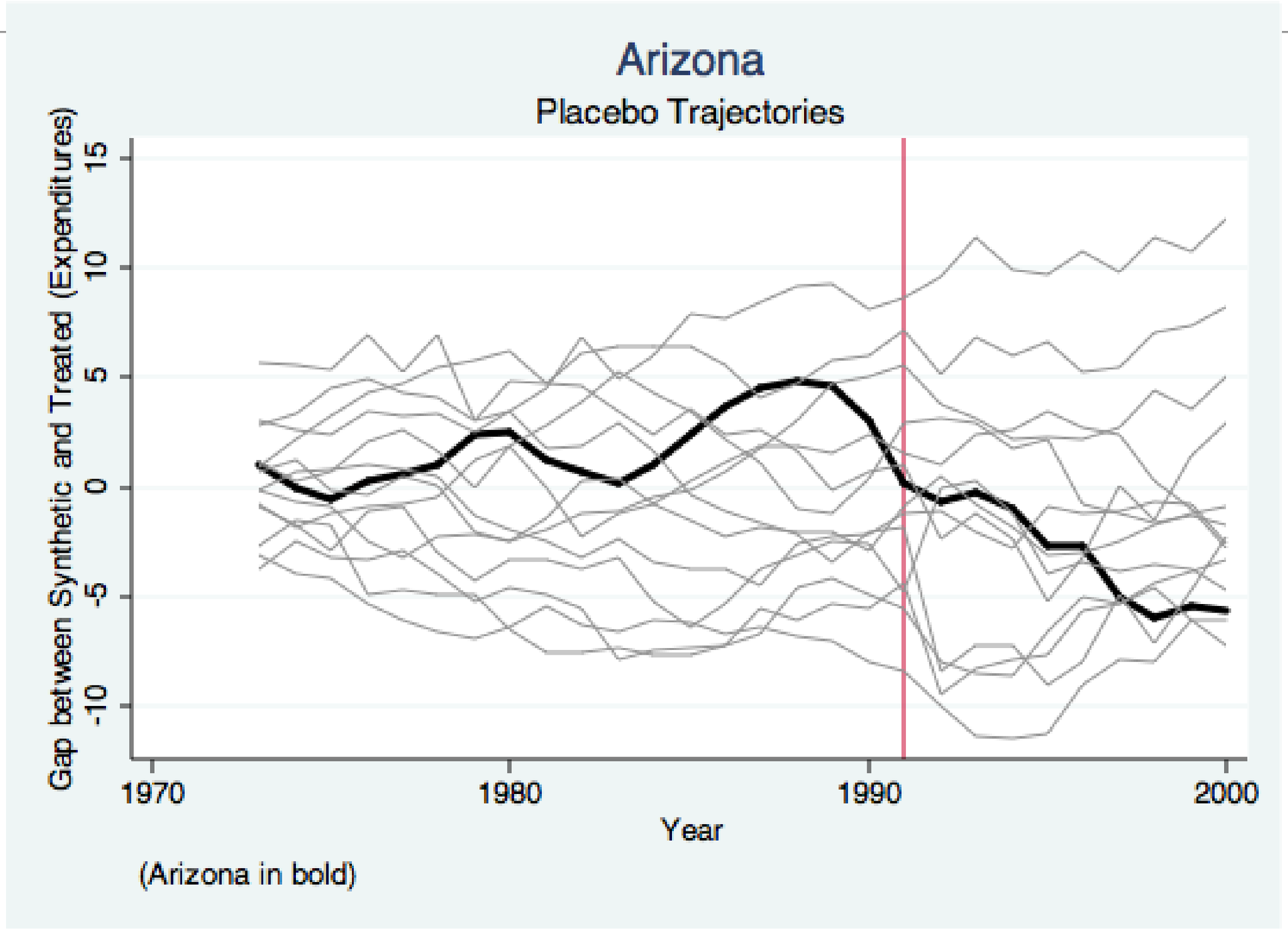
Case: Arizona – Constructing Placebos



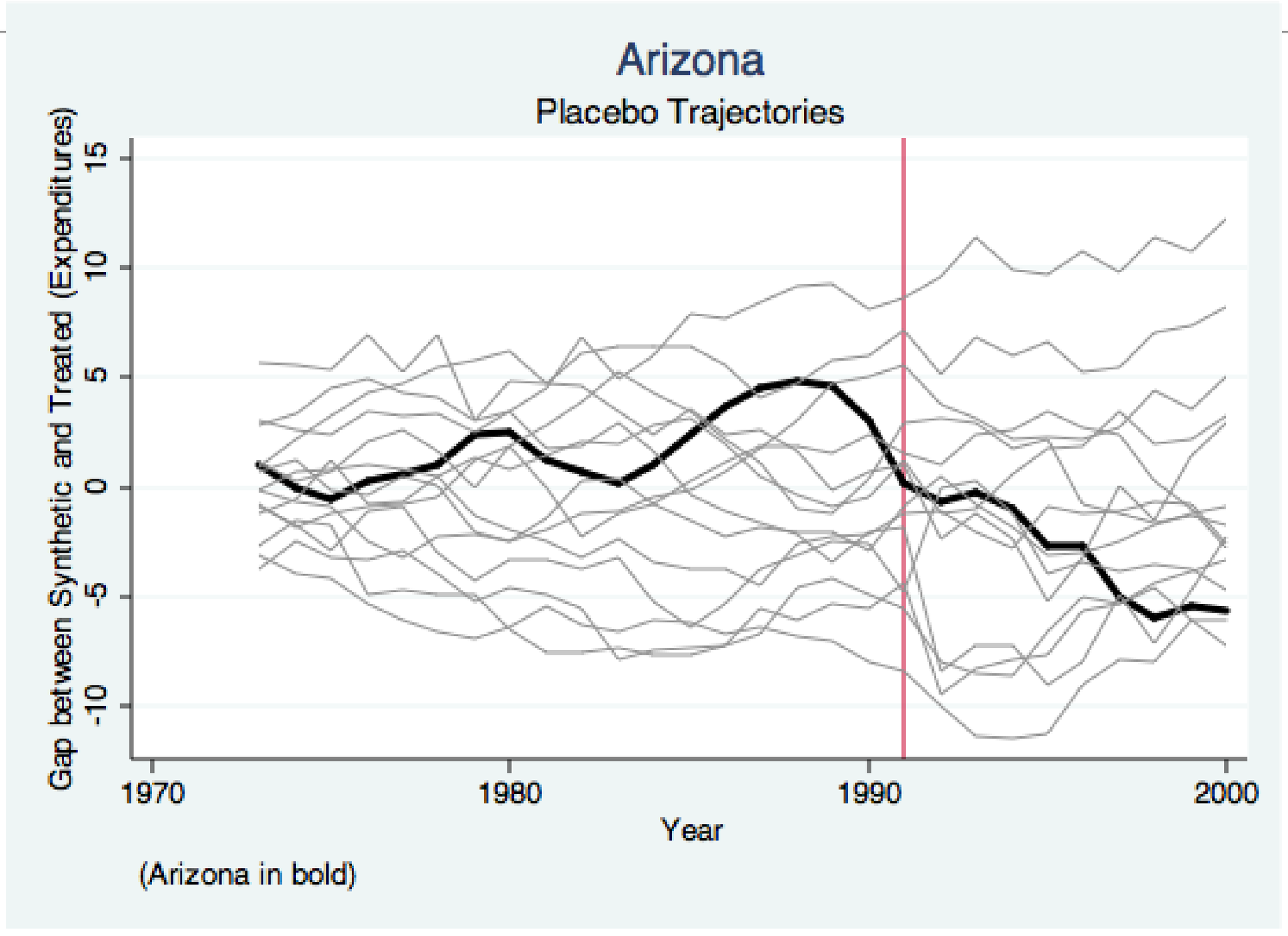
Case: Arizona – Constructing Placebos



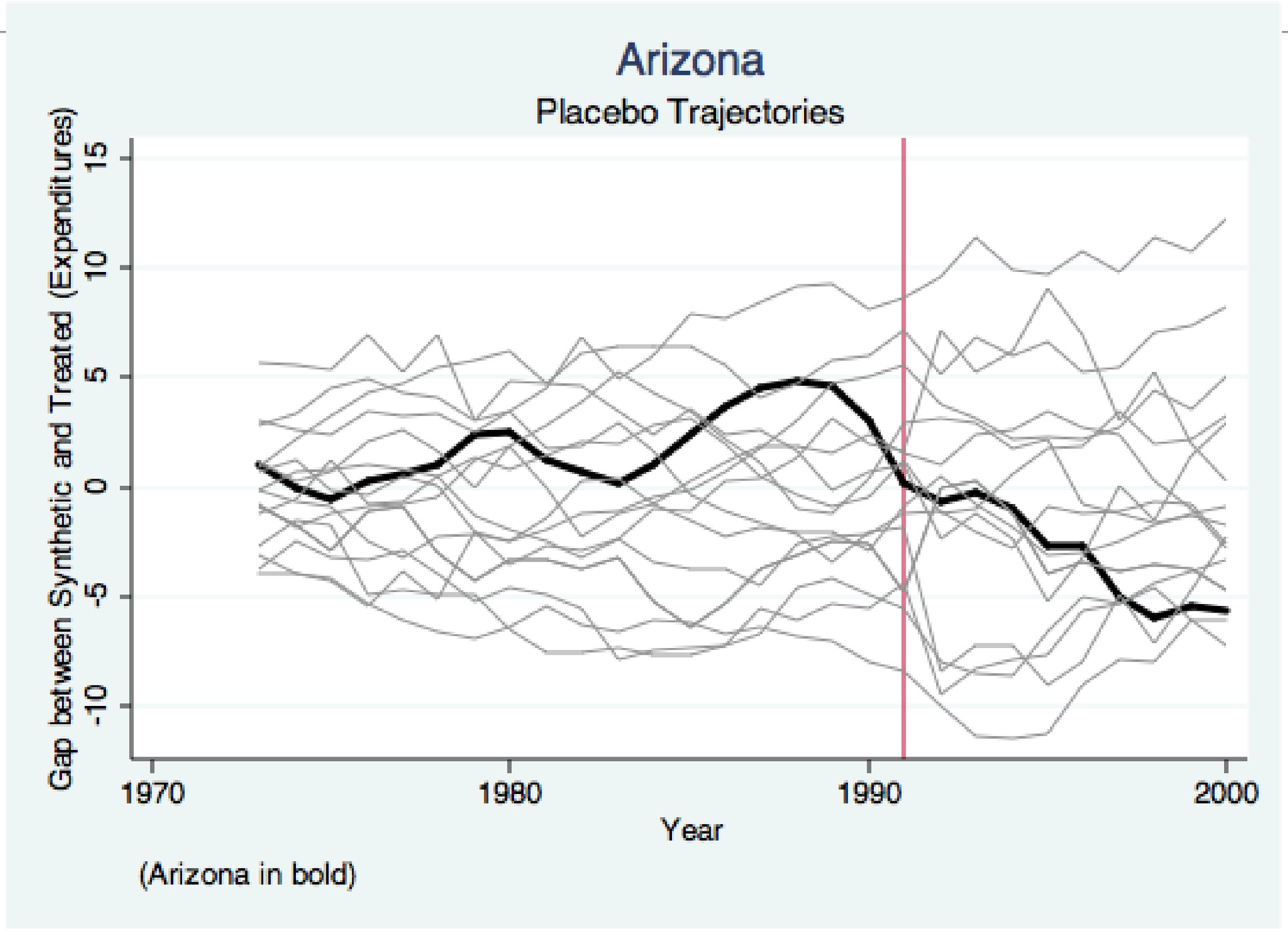
Case: Arizona – Constructing Placebos



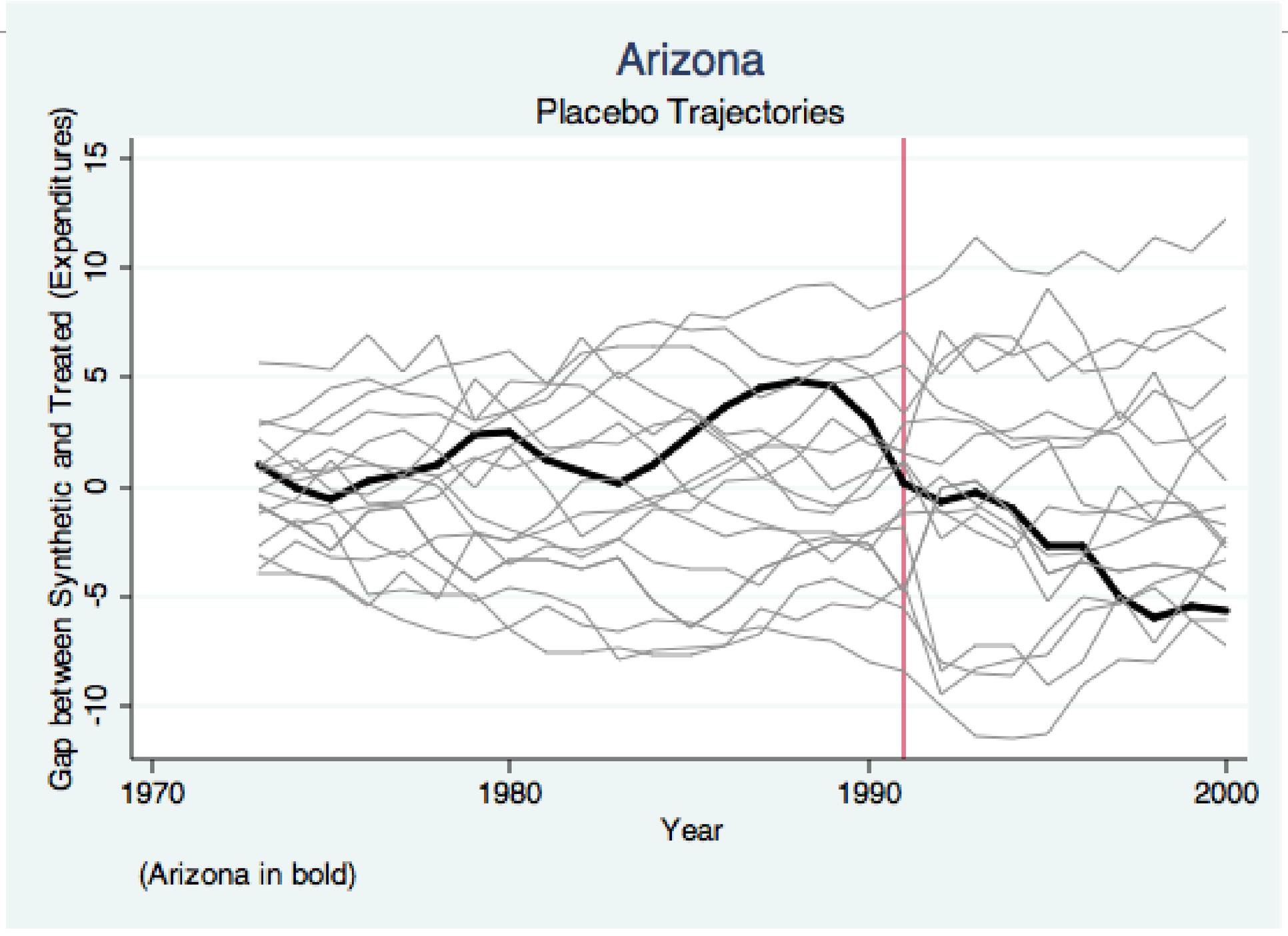
Case: Arizona – Constructing Placebos



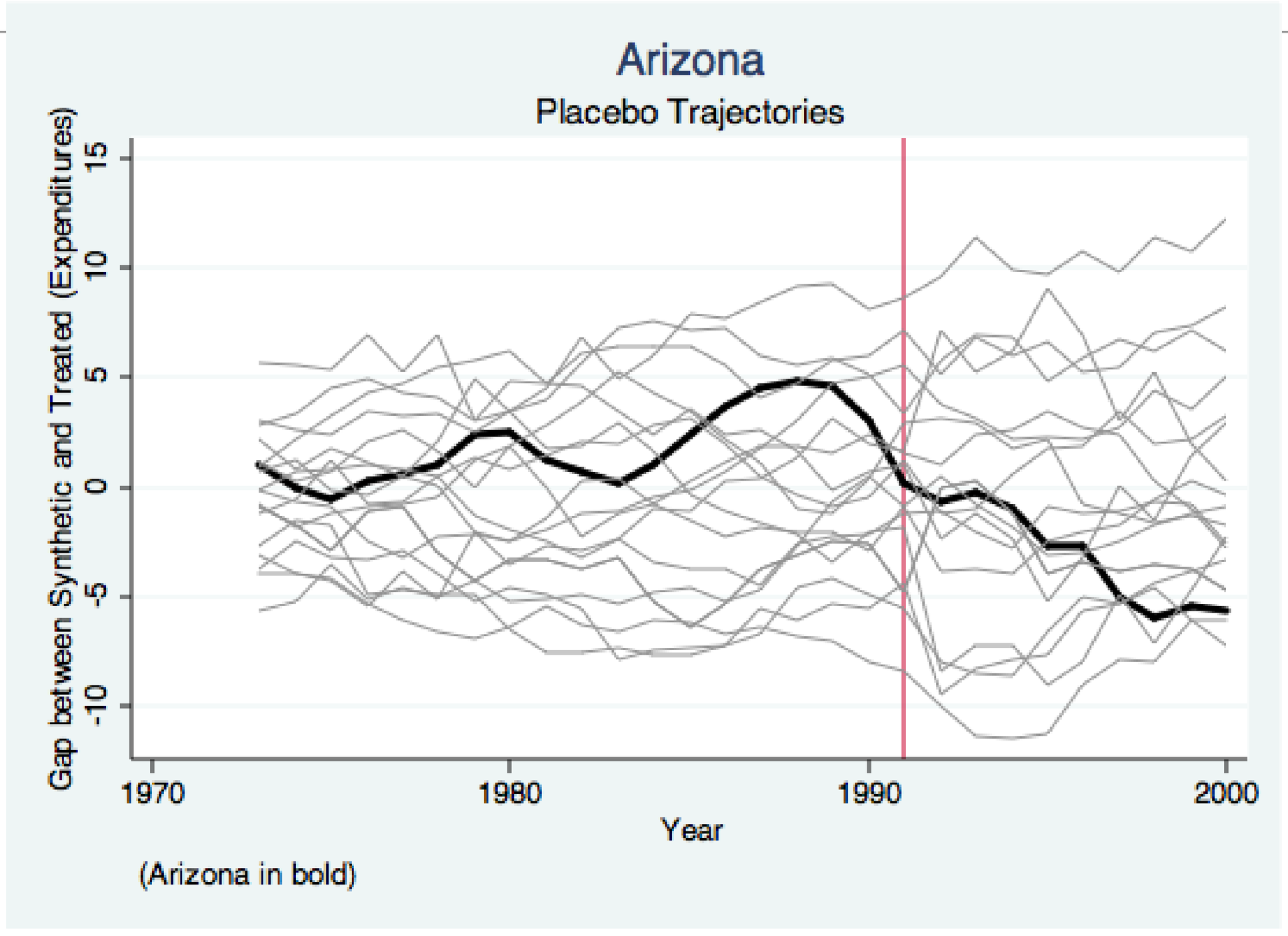
Case: Arizona – Constructing Placebos



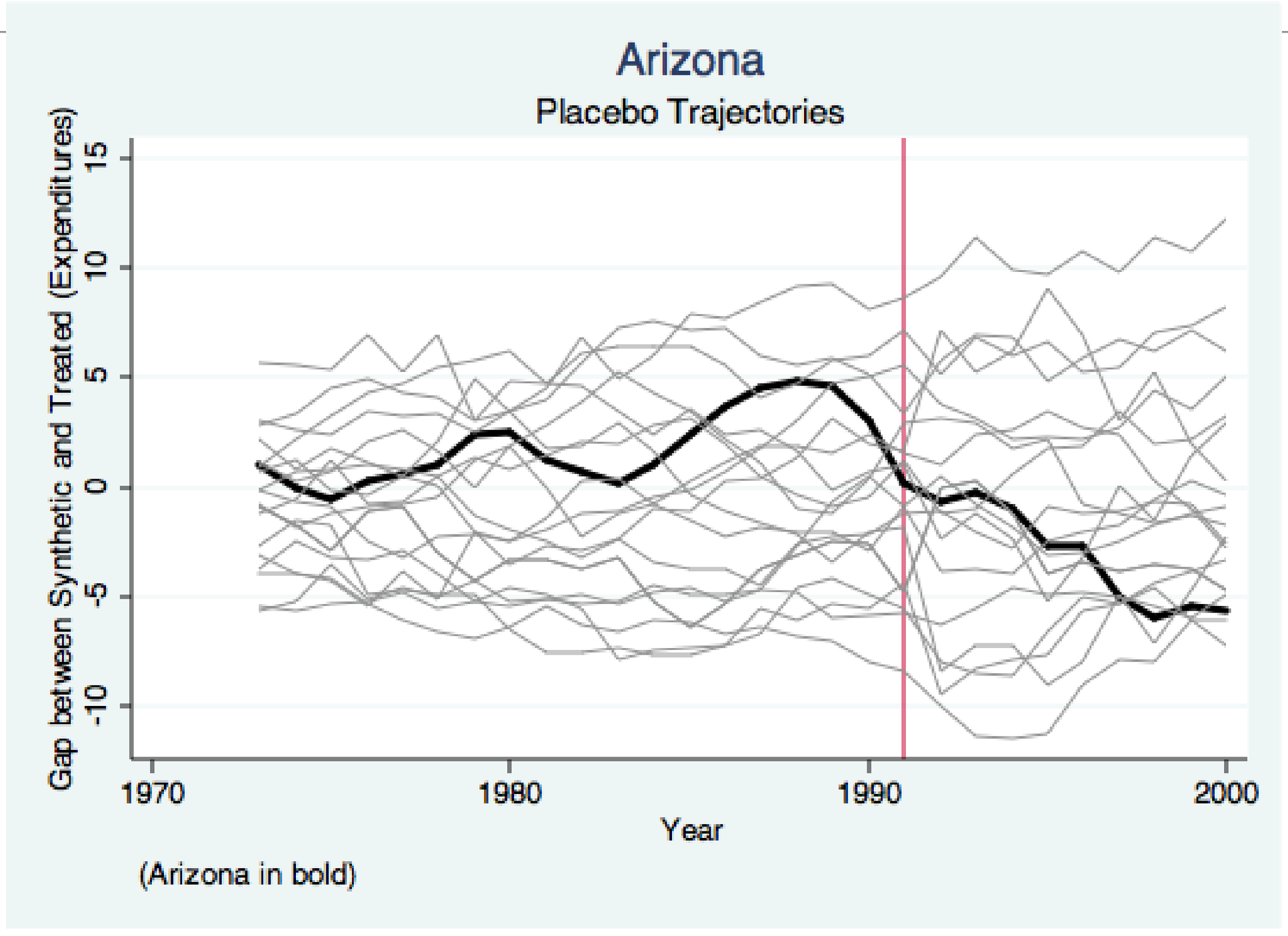
Case: Arizona – Constructing Placebos



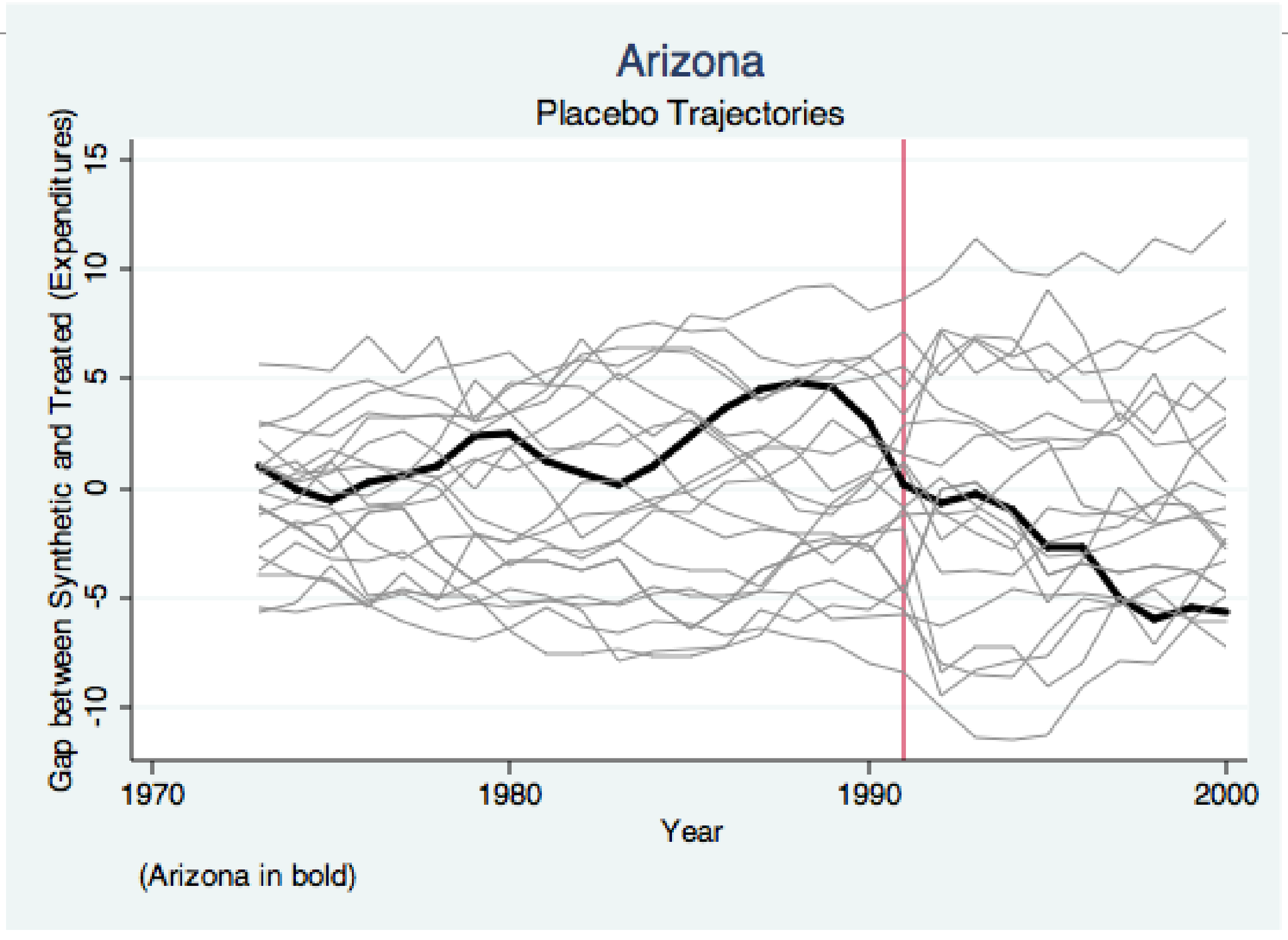
Case: Arizona – Constructing Placebos



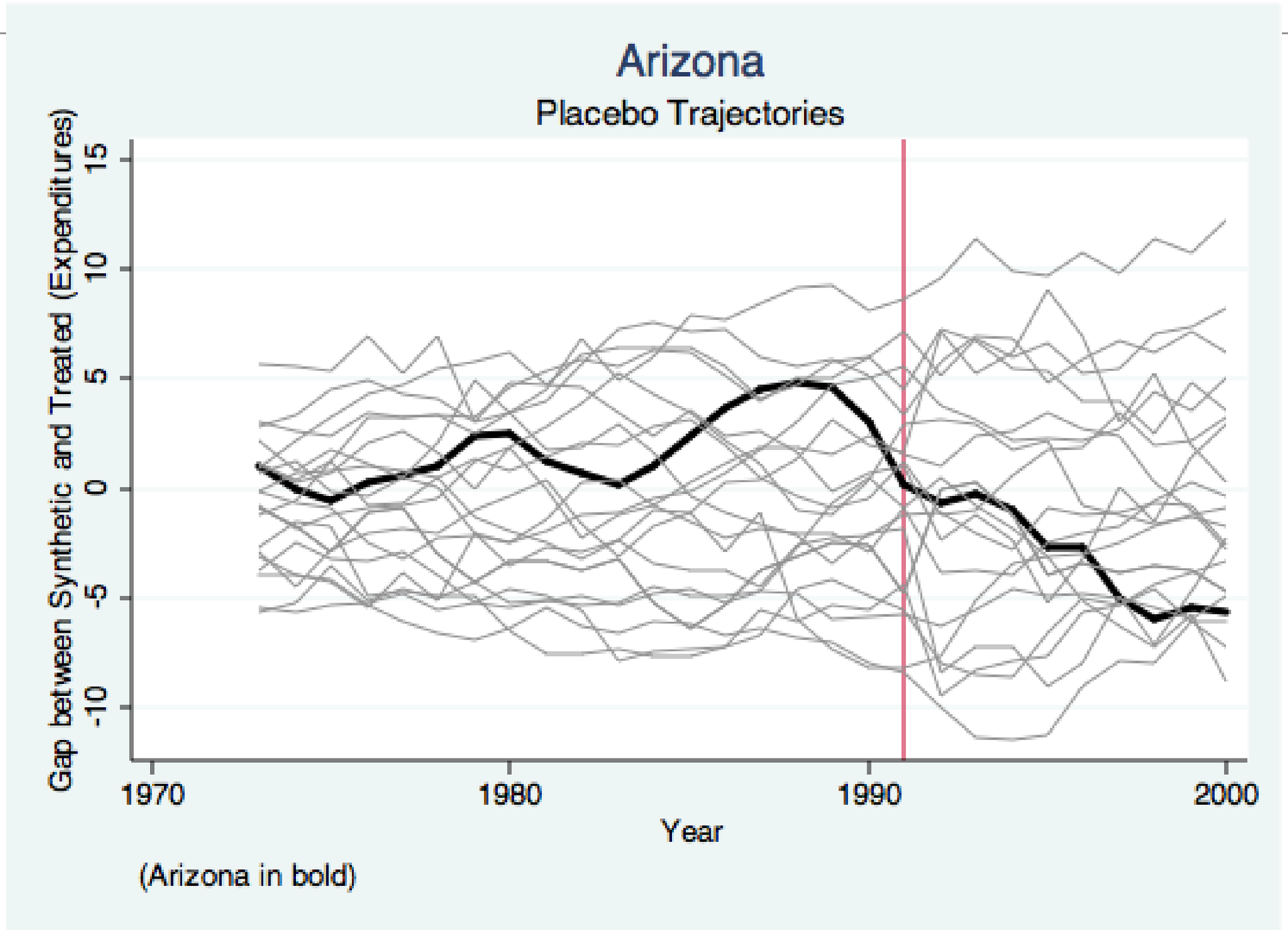
Case: Arizona – Constructing Placebos



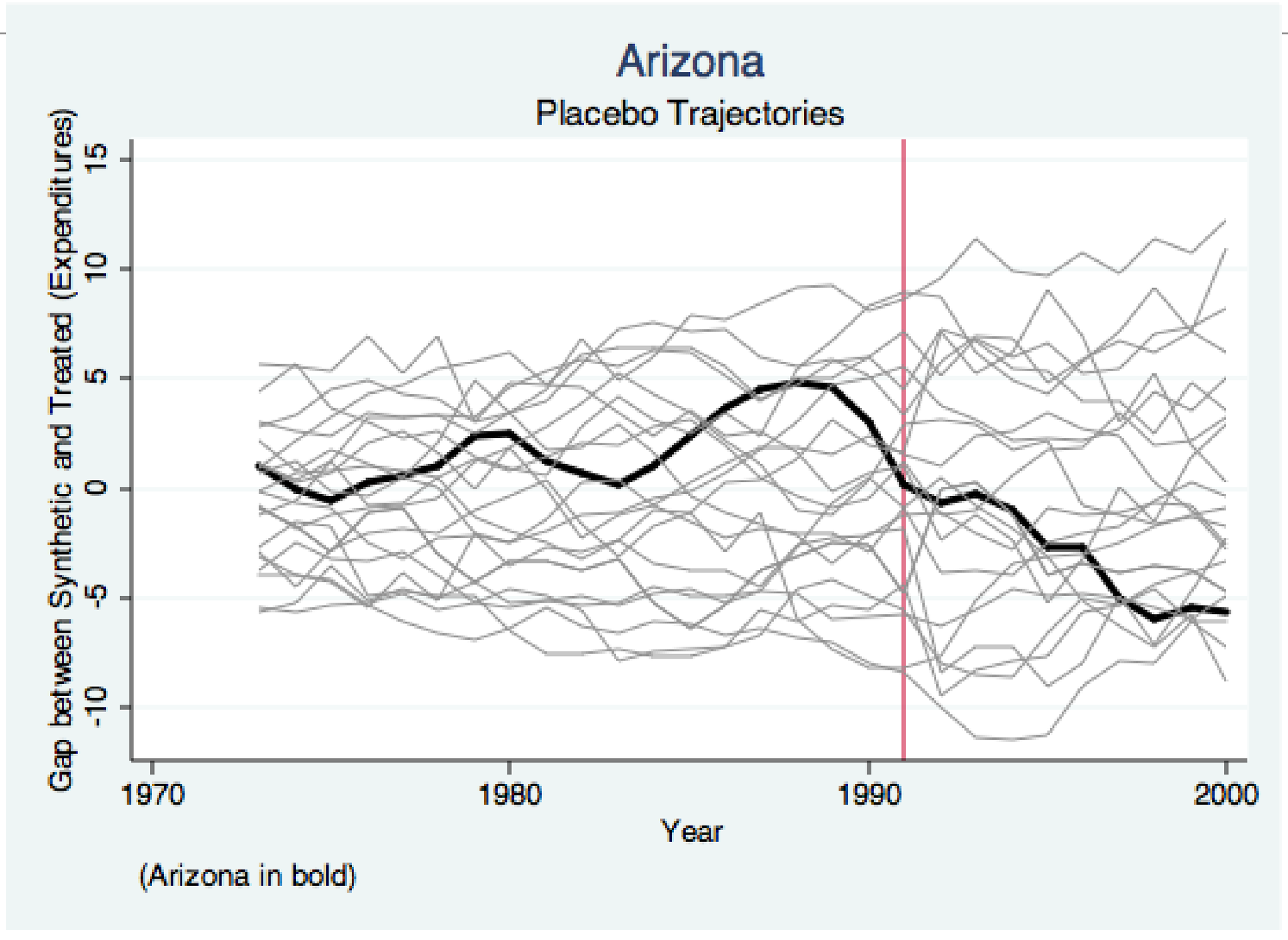
Case: Arizona – Constructing Placebos



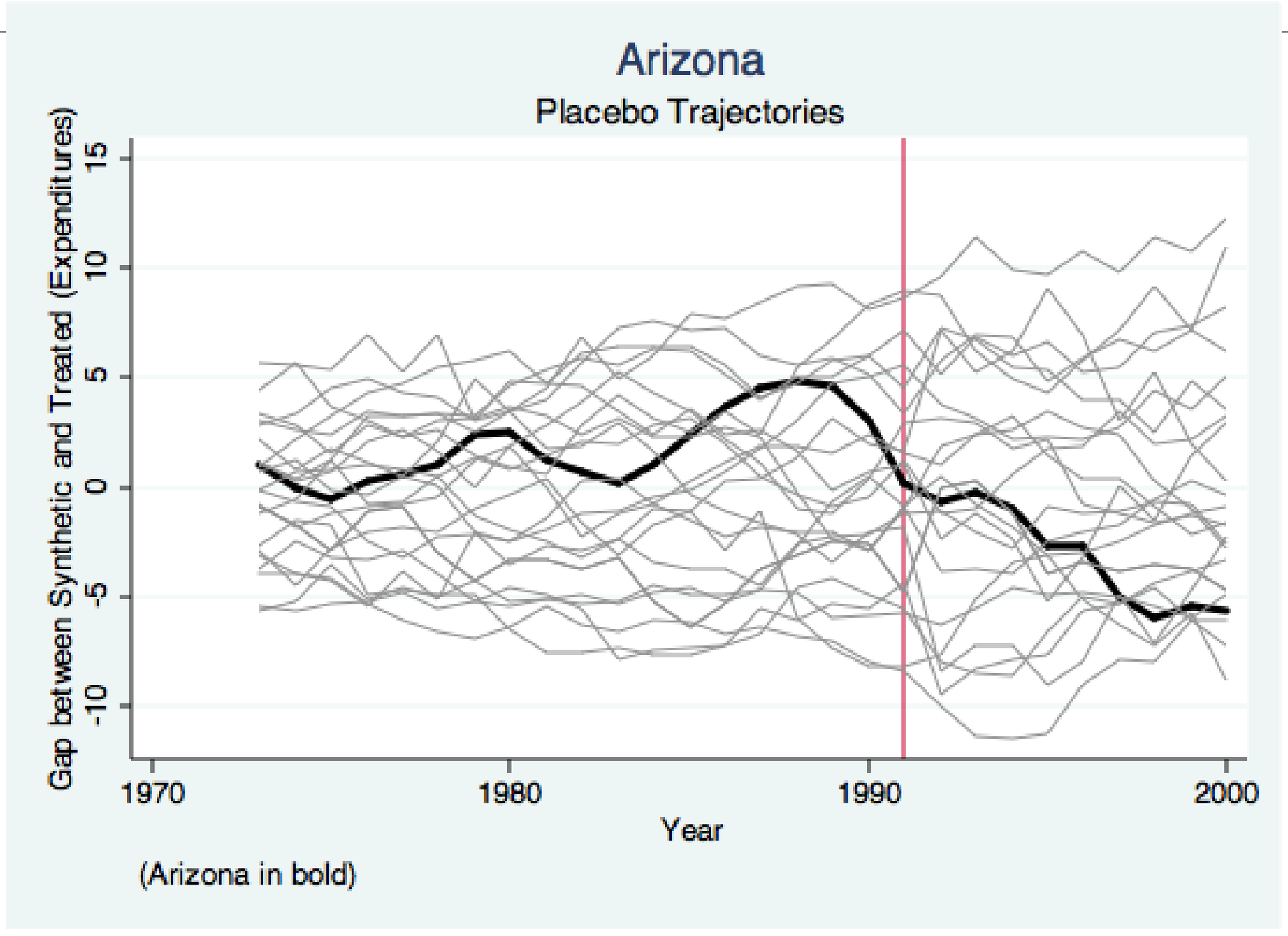
Case: Arizona – Constructing Placebos



Case: Arizona – Constructing Placebos



Case: Arizona – Final Placebo Graph



```
Random-effects ML regression                    Number of obs      =        72
Group variable: ism                             Number of groups   =         6
Random effects u_i ~ Gaussian                   Obs per group: min =        12
                                                               avg =      12.0
                                                               max =        12
                                                LR chi2(15)        =    230.85
Log likelihood  = -311.75688                    Prob > chi2        =    0.0000
------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  appt_salary |  -3.297377   .4854905    -6.79   0.000    -4.248921   -2.345833
 yrs_since_deg|    48.4478   4.360772    11.11   0.000     39.90084    56.99475
  gender_code |  -123.5581   37.16056    -3.32   0.001    -196.3914   -50.72474
    professor |    79.4396   9.888771     8.03   0.000     60.05797    98.82124
  _Iyear_1998 |  -46.98194   11.79058    -3.98   0.000    -70.09105   -23.87283
  _Iyear_1999 |  -107.1882   17.70109    -6.06   0.000    -141.8817   -72.49472
  _Iyear_2000 |  -165.4684   25.64805    -6.45   0.000    -215.7376   -115.1991
  _Iyear_2001 |  -208.2177   32.42948    -6.42   0.000    -271.7783    -144.657
  _Iyear_2002 |  -225.6228   38.01764    -5.93   0.000     -300.136   -151.1096
  _Iyear_2003 |  -268.6547   44.36851    -6.06   0.000    -355.6154    -181.694
  _Iyear_2004 |  -329.0877   51.10362    -6.44   0.000     -429.249   -228.9265
  _Iyear_2005 |  -395.8666   58.50241    -6.77   0.000    -510.5292    -281.204
  _Iyear_2006 |  -469.3563    65.6741    -7.15   0.000    -598.0752   -340.6375
  _Iyear_2007 |  -536.2346   72.43002    -7.40   0.000    -678.1948   -394.2743
  _Iyear_2008 |  -623.3683    80.3218    -7.76   0.000    -780.7961   -465.9405
        _cons |   664.2797   92.06111     7.22   0.000     483.8432    844.7161
-------------+----------------------------------------------------------------
    /sigma_u |   44.26374   12.98688                      24.90621    78.66628
    /sigma_e |   15.14551   1.319032                      12.76886    17.96452
         rho |   .8951935   .0575353                      .7368428    .9696461
------------------------------------------------------------------------------
Likelihood-ratio test of sigma_u=0: chibar2(01)=  119.79 Prob>=chibar2 = 0.000
```

RDD

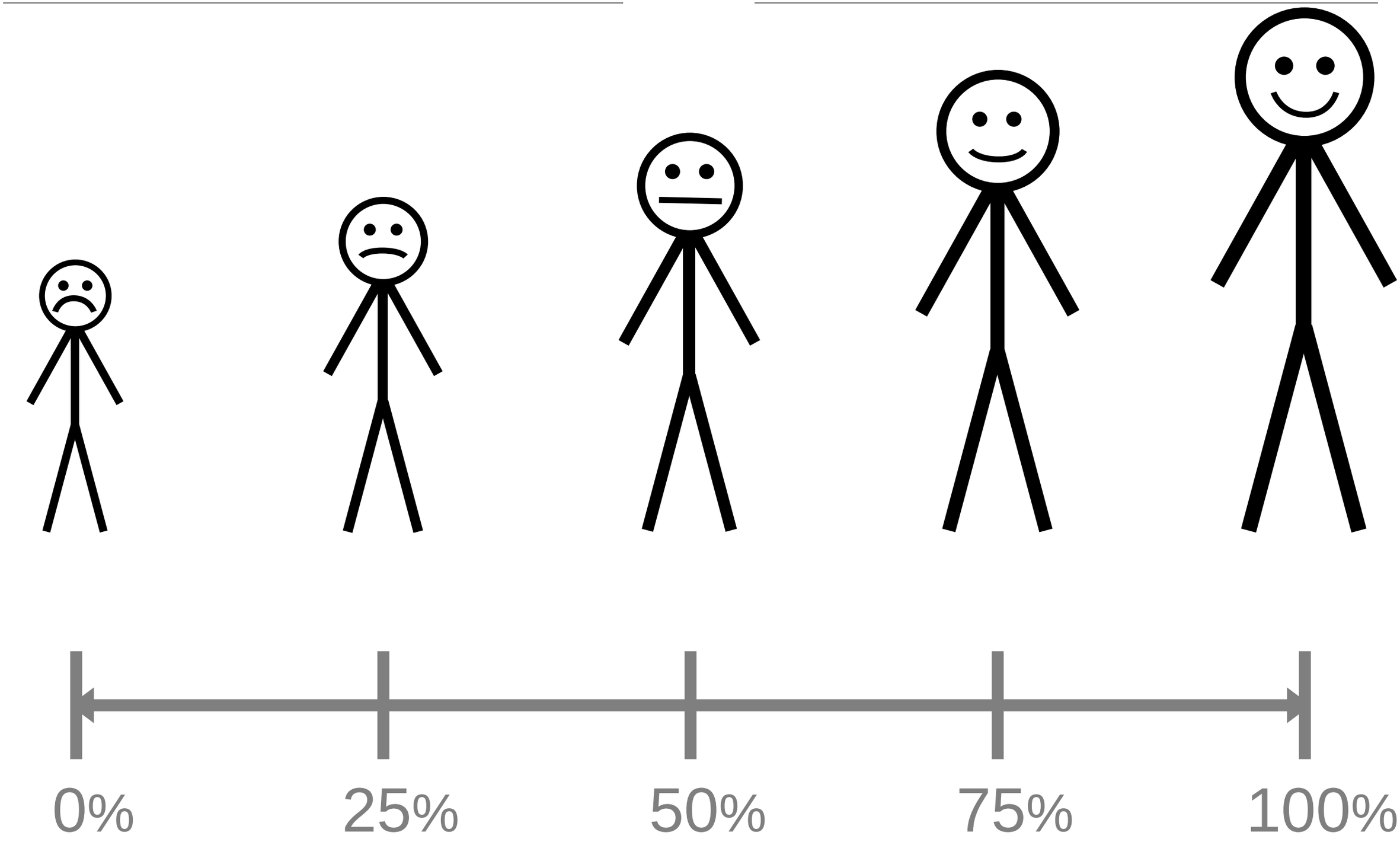
Estimating LATE: Regression discontinuity (RDD)

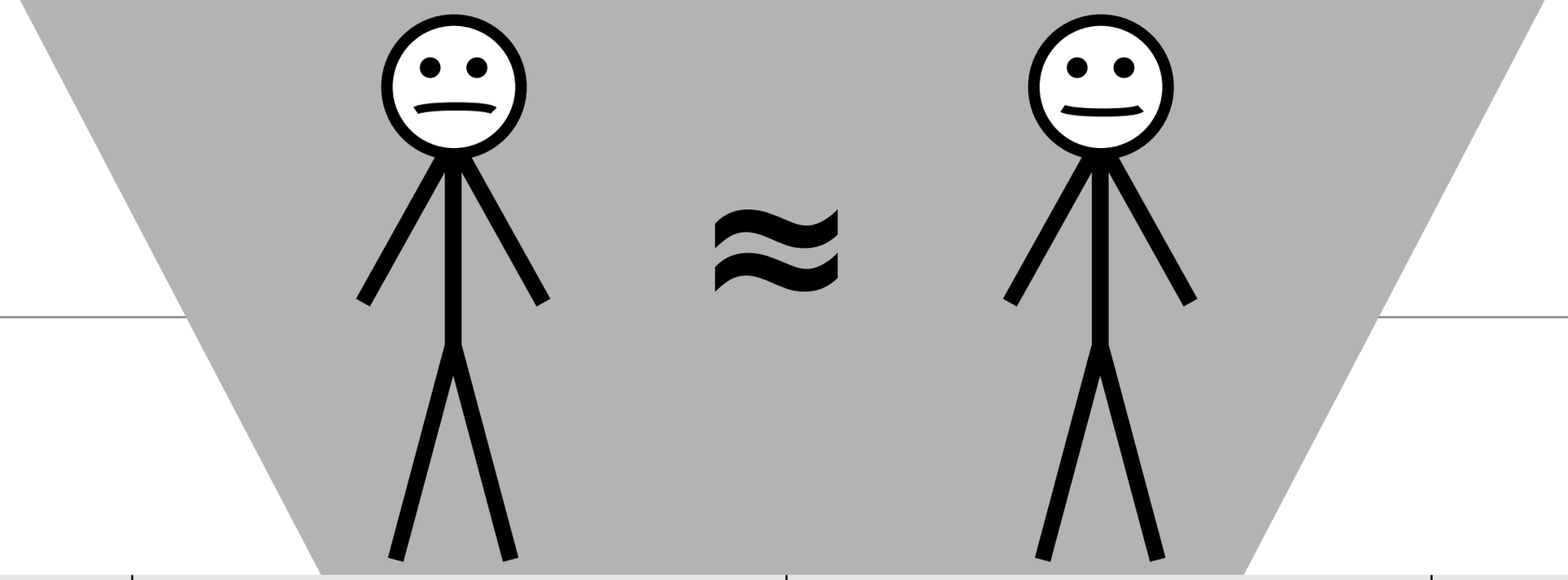
- How do we determine the effect on behavior of holding office in Zambia?
  (in this case we have a battery of economic games for measuring politicians' behavior)

- Candidates are self-selected
  the selection mechanism is a function of unobservables

- Office-holders are selected by voters
  again, selection is a function of unobservables

- Can't compare office holders to undergraduates,
  or an "average Zambian"
  or even to failed candidates


0%
25%
50%
75%
100%



Covariate	Loser Mean/Proportion	Winner Mean/Proportion	KS p-value
Bemba Dummy	66.60%	60.70%	0.5
Province			
Copperbelt	30.20%	25.00%	0.53
Eastern	12.70%	8.90%	0.51
Lusaka	9.50%	8.90%	0.91
Luapula	1.20%	3.40%	0.5
Northern	14.30%	12.50%	0.78
Northwestern	11.10%	12.50%	0.82
Southern	7.90%	14.30%	0.23
Western	1.60%	5.40%	0.27
Female	7.9%	8.9%	0.85
Age	48.1	50.4	0.27
Education	5.965	5.96	0.87
Income	5.306	4.41	0.69
Ownership			
Business	71.30%	71.30%	1
House	92.50%	92.50%	0.87
Car	35.90%	14.20%	0.008
TV	77.80%	66.10%	0.16

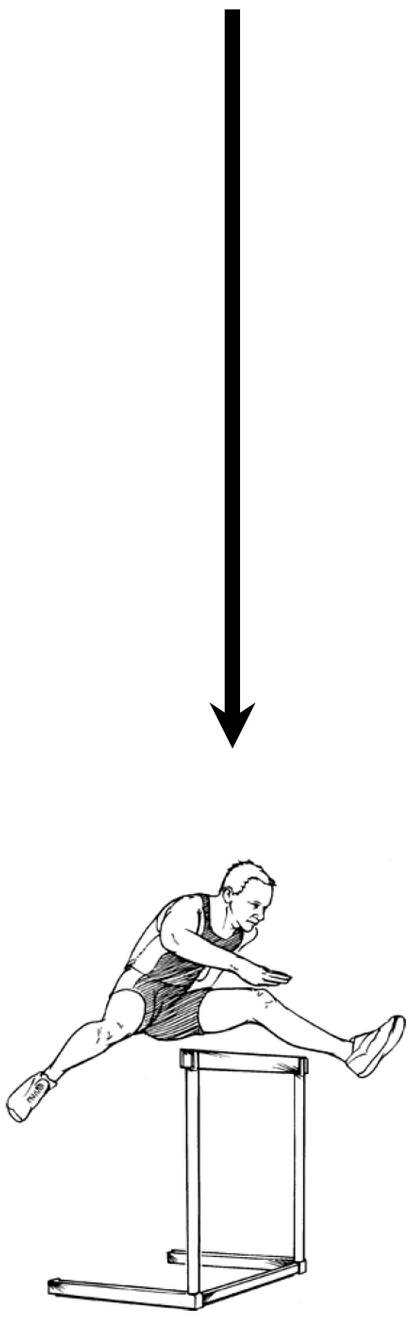
Regression Discontinuity

- Analysis: regress behavior on treatment, controlling for forcing variable
  Forcing variable = margin of victory
  Treatment variable = winner (positive margin of victory)
  $y = \beta T + f(V) + \varepsilon$, for $V \in [-5, +5]$ where β is the LATE

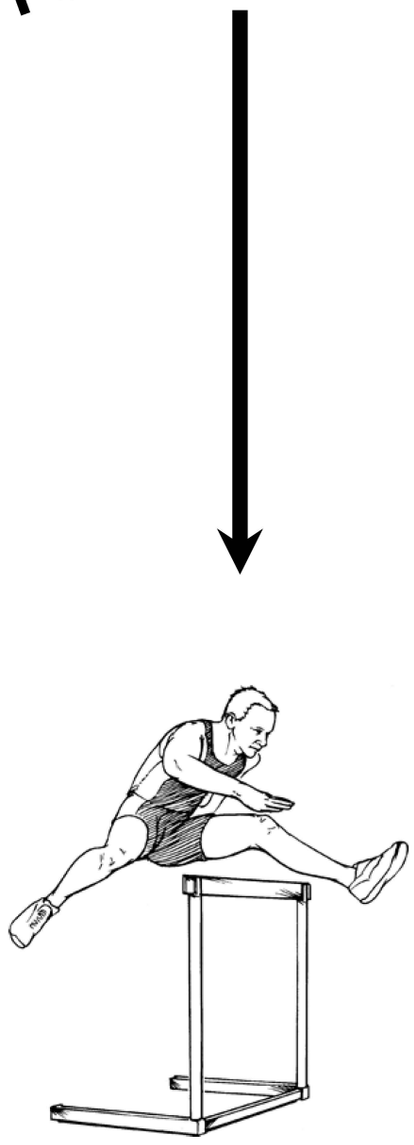
- RDD requires no discontinuity other than treatment
  - We showed balance in observed covariates
  - Is there discontinuity in unobservables?
    - Very low information = candidates cannot strategically attain 51% of vote as in US federal elections (see Caughey & Sekhon 2010)
    - Zambia is the only country in Africa in which citizens of all parties have high confidence in electoral institutions, and it has the highest proportion of people stating they would fight for democracy if they did not trust the election results (Moehler 2005)

In conclusion

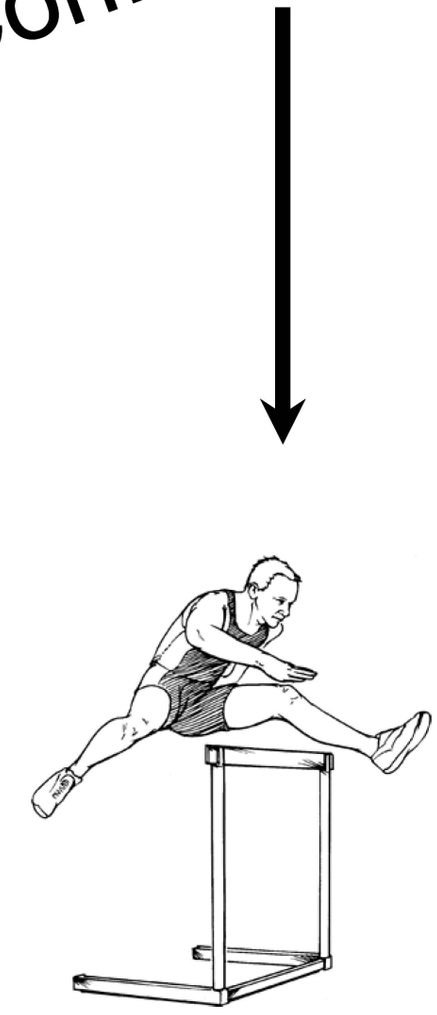
Correlation



Temporal Precedence



No Confounds



Fundamental Problem of Causal Inference