

The Statistics of Causal Inference in the Social Sciences¹

Political Science 236A
Statistics 239A

Jasjeet S. Sekhon

UC Berkeley

October 29, 2014

¹ © Copyright 2014

Background

- We are concerned with causal inference in the social sciences.
- We discuss the **potential outcomes** framework of causal inference in detail.
- This framework originates with **Jerzy Neyman** (1894-1981), the founder of the Berkeley Statistics Department. [@](#)
- A key insight of the framework is that causal inference is a missing data problem.
- The framework applies regardless of the method used to estimate causal effects, whether it be quantitative or qualitative.

The Problem

- Many of the great minds of the 18th and 19th centuries contributed to the development of social statistics: De Moivre, several Bernoullis, Gauss, Laplace, Quetelet, Galton, Pearson, and Yule.
- They searched for a method of statistical calculus that would do for social studies what Leibniz's and Newton's calculus did for physics.
- It quickly came apparent this is would be most difficult. @
- For example: deficits crowding out money versus chemotherapy

The Experimental Model

- In the early 20th Century, Sir Ronald Fisher (1890-1962) helped to establish randomization as the “**reasoned basis for inference**” (Fisher, 1935)
- Randomization: method by which systematic sources of bias are made random
- Permutation (Fisherian) Inference @

Historical Note: Charles Sanders Peirce

- Charles Sanders Peirce (1839–1914) independently, and before Fisher, developed permutation inference and randomized experiments
- He introduced terms “confidence” and “likelihood” @
- Work was not well known until later. Bertrand Russell wrote (1959): “Beyond doubt [...] he was one of the most original minds of the later nineteenth century, and certainly the greatest American thinker ever.”
- Karl Popper (1972): “[He is] one of the greatest philosophers of all times”

There are Models and then there are Models

- Inference based on observational data remains a difficult challenge. Especially, without rigorous mathematical theories such as Newtonian physics

Note the difference between the following two equations:

- $F = m \times a$
usually measured as F (force) Newtons, N; m (mass) kg; a (acceleration) as m/s^2
- $Y = X\beta$

Regression is Evil

- It was hoped that statistical inference through the use of multiple regression would be able to provide to the social scientist what experiments and rigorous mathematical theories provide, respectively, to the micro-biologist and astronomer. @
- OLS has unfortunately become a black box which people think solves hard problems it actually doesn't solve. It is in this sense **Evil**.
- Let's consider a simple sample to test intuition

A Simple Example

- Let Y be a IID standard normal random variable, indexed by t
- Define $\Delta Y_t = Y_t - Y_{t-1}$
- Let's estimate via OLS:

$$\Delta Y_t = \alpha + \beta_1 \Delta Y_{t-1} + \beta_2 \Delta Y_{t-2} + \beta_3 \Delta Y_{t-3}$$

- Question: What are the values of the betas as $n \rightarrow \infty$?

A Simple Example II

- What about for:

$$\Delta Y_t = \alpha + \beta_1 \Delta Y_{t-1} + \beta_2 \Delta Y_{t-2} + \cdots + \beta_k \Delta Y_{t-k}$$

?

OLS: Sometimes Mostly Harmless

Notwithstanding the forgoing, as we shall see

- OLS, and some other estimators, have some nice properties under certain **identification** assumptions
- These assumptions are distinct from the usual statistical assumptions (e.g., those required for Gauss-Markov)
- Relevant theorems do not assume that OLS is correct, that the errors are IID.
- e.g., what happens when we don't assume $\mathbb{E}(\epsilon|X) = 0$?

Early and Influential Methods

- John Stuart Mill (in his *A System of Logic*) devised a set of five methods (or canons) of inference.
- They were outlined in Book III, Chapter 8 of his book.
- Unfortunately, people rarely read the very next chapter entitled “Of Plurality of Causes: and of the Intermixture of Effects.”

Mill's Methods of Inductive Inference

- **Of Agreement:** “If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree is the cause (or effect) of the given phenomenon.”
@
- **Of Difference:** “If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ is the effect, or the cause, or an indispensable part of the cause, of the phenomenon.”

Three Other Methods

- **Joint Method of Agreement and Difference:** “If an antecedent circumstance is invariably present when, but only when, a phenomenon occurs, it may be inferred to be the cause of that phenomenon.”
- **Method of Residues:** “If portions of a complex phenomenon can be explained by reference to parts of a complex antecedent circumstance, whatever remains of that circumstance may be inferred to be the cause of the remainder that phenomenon.”
- **Method of Concomitant Variation:** “If an antecedent circumstance is observed to change proportionally with the occurrence of a phenomenon, it is probably the cause of that phenomenon.”

Uses of Mills Methods

These methods have been used by a vast number of researchers, including such famous ones as Durkheim and Weber. They are known as the “most similar” and “most different” research designs in some fields (Przeworski and Teune, 1970):

Here are some examples:

- The Protestant Ethic @
- Deficits and interest rates @
- Health care systems and life expectancy @
- Gun control
- Three strikes
- The list goes on, and on....

Uses of Mills Methods

Mill himself thought they were inappropriate for the study of social questions (Sekhon, 2004).

“Nothing can be more ludicrous than the sort of parodies on experimental reasoning which one is accustomed to meet with, not in popular discussion only, but in grave treatises, when the affairs of nations are the theme. “How,” it is asked, “can an institution be bad, when the country has prospered under it?” “How can such or such causes have contributed to the prosperity of one country, when another has prospered without them?” Whoever makes use of an argument of this kind, not intending to deceive, should be sent back to learn the elements of some one of the more easy physical sciences” (Mill, 1873, pp. 346–7).

Basic Statistical Inference

Let's look at an example Mill himself brought up:

"In England, westerly winds blow during about twice as great a portion of the year as easterly. If, therefore, it rains only twice as often with a westerly as with an easterly wind, we have no reason to infer that any law of nature is concerned in the coincidence. If it rains more than twice as often, we may be sure that some law is concerned; either there is some cause in nature which, in this climate, tends to produce both rain and a westerly wind, or a westerly wind has itself some tendency to produce rain."

Conditional Probability

$$H : P(\text{rain} | \text{westerly wind}, \Omega) > P(\text{rain} | \text{not westerly wind}, \Omega),$$

where Ω is a set of background conditions we consider necessary for a valid comparison.

A lot is hidden in the Ω . The issue becomes clearer with the potential outcomes framework.

Potential Outcomes

- Fundamental problem: not observing all of the potential outcomes or counterfactuals
- This is the [Neyman-Rubin Causal Model](#) (Neyman 1923, Rubin 1974, Holland 1986). @
- Can be used to describe problems of causal inference for both experimental work and observational studies.

Observational Study

An observational study concerns

- cause-and-effect relationships
- treatments, interventions or policies and
- the effects they cause

@

The design stage of estimating the causal effect for treatment T is common for all Y .

A Thought Experiment

An observational study could in principle have been an experiment but for ethical concerns or logistical issues.

@

You are probably not estimating a causal effect if you can't answer Dorn's (1953) Question: "what experiment would you have run if you were dictator and has infinite resources?"

@

E.G.: Can we estimate the causal effect of race on SAT scores?

@

Descriptive and predictive work is something else and can be interesting.

ford

Web

Images

Maps

Shopping

News

More ▾

Search tools

About 868,000,000 results (0.30 seconds)

Ads related to **ford**

[Ford.com - Ford Focus Official Site](#)

www.ford.com/Focus ↗

Be Everywhere @ Once w/ Up to 40MPG **Ford** Focus. Learn More @ **Ford**.com.

[Build and Price](#)

Build & Price the 2014 Ford Focus.
A Small Car That's Big On Features!

[Photo Gallery](#)

See Interior & Exterior Photos
of the 2014 Focus @ Ford.com.

[2013 Toyota Camry - BuyAToyota.com](#)

www.buyatoyota.com/Camry ↗

0% for 60 mos PLUS \$1,000 Trade-in Cash on a new Camry. Learn more.

[Ford – New Cars, Trucks, SUVs, Hybrids & Crossovers | Ford Vehicles](#)

www.ford.com/ ↗

The Official **Ford** Site to research, learn and shop for all new **Ford** Vehicles. View photos, videos, specs, compare competitors, build and price, search inventory ...

[All Vehicles - 2013 F-150 - Mustang - 2014 Ford Focus](#)

77,507 people +1'd this

Ford Motor Company

2,620,994 follower

Ford Motor Company
multinational automaker
Dearborn, Michigan
founded by Henry Ford
June 16, 1903. Wik

CEO: Alan Mulally

Headquarters: De

Founder: Henry F

Founded: June 16

Awards: Car and

Recent posts

ebay

Web

Images

Maps

Shopping

News

More ▾

Search tools

About 1,730,000,000 results (0.19 seconds)

[eBay: Electronics, Cars, Fashion, Collectibles, Coupons and More ...](http://www.ebay.com)

www.ebay.com 

Buy and sell electronics, cars, fashion apparel, collectibles, sporting goods, digital cameras, baby items, coupons, and everything else on eBay, the world's ...

[Motors](#)

Cars Trucks - Parts & Accessories -
Collector Cars - Motorcycles

[Women](#)

Tops & Blouses - Coats & Jackets -
Swimwear - Shoes - Sweaters

[Daily Deals](#)

Deals are updated daily, so check
back for the deepest discounts ...

[More results from ebay.com »](#)

[Cars Trucks](#)

Salvage - Truck - Rat rod - eBay
Motors - Project - No Reserve

[Men](#)

Casual Shirts - Jeans - Coats &
Jackets - Shorts - Dress Shirts

[Cell Phone, Accessories](#)

Cell Phones & Accessories. Cell
Phones & Smartphones · Cell ...

eBay

Corporation

eBay Inc. is an American consumer-to-consumer headquartered in

Customer service
(Consumer)

Stock price: [EBA](#)

Oct 4, 4:00 PM EDT

Founder: Pierre Omidyar

Founded: September 1995

Headquarters: San Jose, California

CEO: John Donahoe

[News for ebay](#)



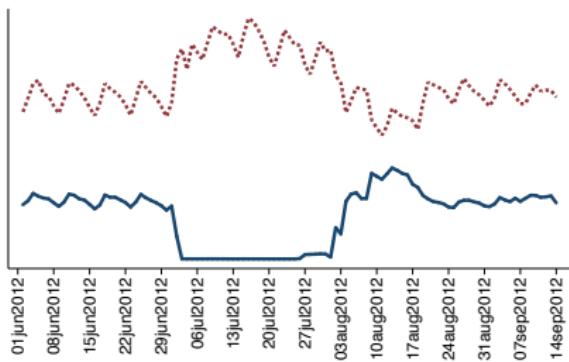
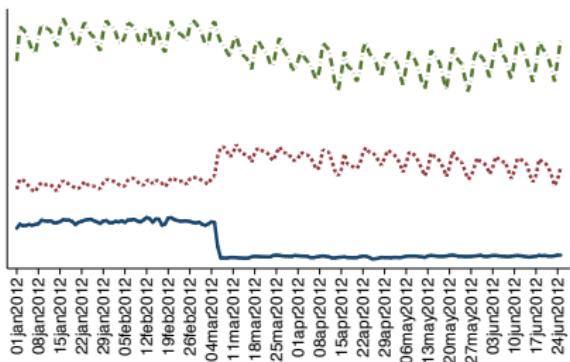
[How the eBay of Illegal Drugs Came Undone](#)

New Yorker (blog) - 12 hours ago

People also search for

227778

A Modest Experiment



From: Blake, Nosko, and Tadelis, 2014

General Problems with Observational Studies

- Prediction accuracy is uninformative @
- No randomization, the “reasoned basis for inference” @
- Selection problems abound
- Data/Model Mining—e.g., not like vision
 @
- Found data, usually retrospective

@

Note: it is far easier to estimate the effect of a cause than the cause of an effect. Why?

Difficult Example: Does Information Matter?

- We want to estimate the effect on voting behavior of paying attention to an election campaign. @
- Survey research is strikingly uniform regarding the ignorance of the public (e.g., Berelson, Lazarsfeld, and McPhee, 1954; Campbell et al., 1960; J. R. Zaller, 1992). @
- Although the fact of public ignorance has not been forcefully challenged, the meaning of this observation has been (Sniderman, 1993). @
- Can voters use information such as polls, interest group endorsements and partisan labels to vote like their better informed compatriots (e.g., Lupia, 2004; McKelvey and Ordeshook, 1985a; McKelvey and Ordeshook, 1985b; McKelvey and Ordeshook, 1986)?

Fundamental Problem of Causal Inference

- Fundamental problem: not observing all of the potential outcomes or counterfactuals @
- Let Y_{i1} denote i 's vote intention when voter i learns during the campaign (i.e., is in the treatment regime). @
- Let Y_{i0} denote i 's vote intention when voter i does not learn during the campaign (i.e., is in the control regime). @
- Let T_i be a treatment indicator: 1 when i is in the treatment regime and 0 otherwise. @
- The observed outcome for observation i is
$$Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}. @$$
- The treatment effect for i is
$$\tau_i = Y_{i1} - Y_{i0}.$$

Probability Model I: Repeated Sampling

- Assume that $Y_i(t)$ is a simple random sample from an infinite population, where Y is the outcome of unit i under treatment condition t .
- Assume the treat T is randomly assigned following either:
 - Bernoulli assignment to each i
 - Complete randomization

Probability Model II: Fixed Population

- Assume that $Y_i(t)$ is a fixed population
- Assume the treat T is randomly assigned following either:
 - Bernoulli assignment to each i
 - Complete randomization

Assignment Mechanism, Classical Experiment

- individualistic
- probabilistic
- unconfounded: given covariates, not be dependent on any potential outcomes

Notation

Let:

- N denote the number of units [in the sample], indexed by i
- number of control units: $N_c = \sum_{i=1}^N (1 - T_i)$; treated units: $N_t = \sum_{i=1}^N (T_i)$, $N_c + N_t = N$
- X_i is a vector of K covariates, where X is a $N \times K$ matrix
- $Y(0)$ and $Y(1)$ denote N -component vectors

Notation

- Let T be the N -dimensional vector with elements $T_i \in \{0, 1\}$ with positive probability
- Let all possible values be $\mathbb{T} = \{0, 1\}^N$, with cardinality 2^N
- $\{0, 1\}^N$ denotes the set of all N vectors with all elements equal to 0 or 1
- Let the subset of values for \mathbb{T} with positive probability be denoted by \mathbb{T}^+

Assignment: Bernoulli Trials

- $Pr(T|X, Y(0), Y(1)) = 0.5^N$,
where $\mathbb{T}^+ = \{0, 1\}^N = \mathbb{T}$
- $Pr(T|X, Y(0), Y(1)) = q^{N_t} \cdot (1 - q)^{N_c}$

Notation Note

- $Pr(T|X, Y(0), Y(1))$ is **not** the probability of a particular unit receiving the treatment
- it reflects a measure across the full population of N units of a particular assignment vector occurring
- The unit-level assignment for unit i is:

$$p_i(X, Y(0), Y(1)) = \sum_{\mathbb{T}: T_i=1} Pr(T|X, Y(0), Y(1))$$

where we sum the probabilities across all possible assignment vectors T for which $T_i = 1$

Assignment: Complete Randomization

- An assignment mechanism that satisfies:

$$\mathbb{T}^+ = \left\{ T \in \mathbb{T} \left| \sum_{i=1}^N T_i = N_t \right. \right\},$$

for some preset $N_t \in \{1, 2, \dots, N - 1\}$

- Number of assignment vectors in this design: $\binom{N}{N_t}$,

$$q = \frac{N_t}{N} \quad \forall i$$

Assignment: Complete Randomization

$$Pr(T | X, Y(0)), Y(1)) =$$

$$\begin{cases} \left(\frac{N_c! \cdot N_t!}{N!}\right) \cdot \left(\frac{N_t}{N}\right)^{N_t} \cdot \left(\frac{N_c}{N}\right)^{N_c} & \text{if } \sum_{i=1}^N T_i = N_t \\ 0 & \text{otherwise} \end{cases}$$

Experimental Data

- Under classic randomization, the inference problem is straightforward because: $T \perp\!\!\!\perp (Y(1), Y(0))$ @
- Observations in the treatment and control groups are not exactly alike, but they are comparable—i.e., they are exchangeable @
- With exchangeability plus noninterference between units, for $j = 0, 1$ we have:

$$\mathbb{E}(Y(j)|T=1) = \mathbb{E}(Y(j)|T=0)$$

Average Treatment Effect

- The Average Treatment Effect (ATE) can be estimated simply:

$$\bar{\tau} = \text{Mean Outcome for the treated} - \text{Mean Outcome for the control}$$

@

- In notation:

$$\begin{aligned}\bar{\tau} &\equiv \bar{Y}_1 - \bar{Y}_0 \\ \bar{\tau} &= \mathbb{E}(Y(1) - Y(0)) \\ &= \mathbb{E}(Y|T=1) - \mathbb{E}(Y|T=0)\end{aligned}$$

Observational Data

- With observational data, the treatment and control groups are not drawn from the same population
- Progress can be made if we assume that the two groups are comparable once we condition on observable covariates denoted by X
- This is the conditional independence assumption:
 $\{Y(1), Y(0) \perp\!\!\!\perp T|X\},$
the reasonableness of this assumption depends on the section process

Average Treatment Effect for the Treated

- With observational data, the treatment and control groups are not drawn from the same population.
- Thus, we often want to estimate the average treatment effect for the treated (ATT):

$$\bar{\tau}|(T = 1) = \mathbb{E}(Y(1)|T = 1) - \mathbb{E}(Y(0)|T = 1)$$

@

- Progress can be made if we assume that the selection process is the result of only observable covariates denoted by X @
- We could alternatively estimate the Average Treatment Effect for the Controls (ATC).

- Under random assignment

$$ATT = \mathbb{E}[Y(1) - Y(0) | T = 1] = E[Y(1) - Y(0)] = ATE$$

@

- Note:

$$\begin{aligned}\mathbb{E}[Y | T = 1] &= \mathbb{E}[Y(0) + T(Y(1) - Y(0)) | T = 1] \\ &= \mathbb{E}[Y(1) | T = 1], \text{ by } \perp\!\!\!\perp \\ &= \mathbb{E}[Y(1)]\end{aligned}$$

@

- Same holds for $\mathbb{E}[Y | T = 0]$

ATE=ATT

- Since

$$\begin{aligned}\mathbb{E}[Y(1) - Y(0) | T = 1] &= \mathbb{E}[Y(1) - Y(0)] \\ &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\ &= \mathbb{E}[Y | T = 1] - \mathbb{E}[Y | T = 0]\end{aligned}$$

- Then,

$$ATE = ATT = \mathbb{E}[Y | T = 1] - \mathbb{E}[Y | T = 0]$$

Review and Details

- More details on ATE, ATT, and potential outcomes: [LINK]
- Some review of probability: [LINK]

Selection on Observables

- Then, we obtain the usual exchangeability results:

$$\mathbb{E}(Y(j)|X, T = 1) = \mathbb{E}(Y(j)|X, T = 0)$$

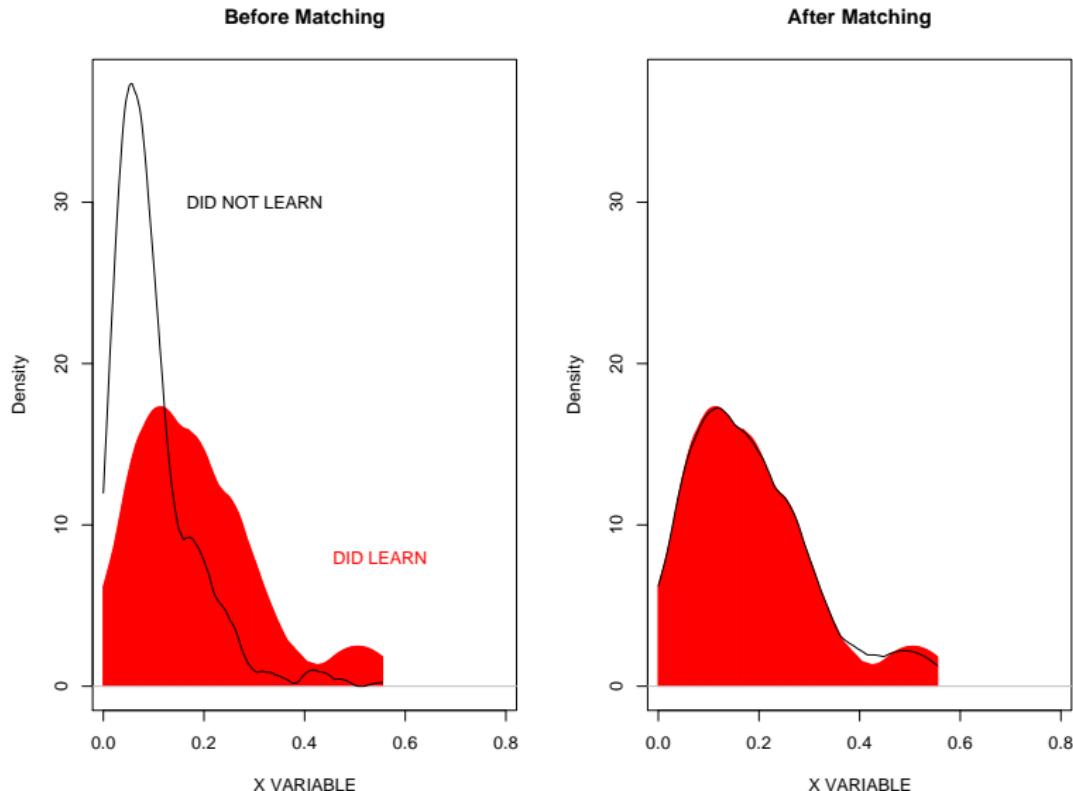
@

- ATT can then be estimated by

$$\bar{\tau}|(T = 1) = \mathbb{E}\{\mathbb{E}(Y|X, T = 1) - \mathbb{E}(Y|X, T = 0) | T = 1\}$$

where the outer expectation is taken over the distribution of $X|(T = 1)$

Observational Data: Matching



Estimands

- Sample Average Treatment Effect (SATE):

$$\tau^S = \frac{1}{N} \sum_{i=1}^N [Y_i(1) - Y_i(0)]$$

- Sample Average Treatment Effect on the Treated (SATT):

$$\tau_t^S = \frac{1}{N_t} \sum_{i: T_i=1} [Y_i(1) - Y_i(0)]$$

- Population Average Treatment Effect (PATE):

$$\tau^P = \mathbb{E} [Y(1) - Y(0)]$$

- Population Average Treatment Effect on the Treated (PATT):

$$\tau_t^P = \mathbb{E} [Y(1) - Y(0) | T = 1]$$

CATE

Conditional ATE (CATE): conditional on the sample distribution of sample covariates

- CATE:

$$\overline{\tau(X)} = \frac{1}{N} \sum_{i=1}^N \mathbb{E}[Y_i(1) - Y_i(0)|X_i]$$

- Condition on the treated (CATT):

$$\overline{\tau(X)_t} = \frac{1}{N_t} \sum_{i:T_i=1} \mathbb{E}[Y_i(1) - Y_i(0)|X_i]$$

Other Estimands

Conditional on whatever:

- ① Potential outcomes
- ② Potential response
- ③ Other latent attributes—e.g., probability of selection
- ④ (2) and (3) are probably related in practice

SATE, unbiased?

- Assume completely randomized experiment. Follow Neyman (1923/1990)
- $\hat{\tau} = \bar{Y}_t - \bar{Y}_c$
- The estimator $\hat{\tau}$ is unbiased for τ
- Expectation taken only over the randomization distribution
- Note: $\mathbb{E}[T_i | Y(0), Y(1)] = \frac{N_t}{N}$

SATE, unbiased?, proof

- Rerwrite $\hat{\tau}$:

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^N \left(\frac{T_i \cdot Y_i(1)}{N_t/N} - \frac{(1 - T_i) \cdot Y_i(0)}{N_c/N} \right)$$

- $\mathbb{E} [\hat{\tau} | Y(0), Y(1)] =$

$$\frac{1}{N} \sum_{i=1}^N \left(\frac{\mathbb{E}[T_i] \cdot Y_i(1)}{N_t/N} - \frac{\mathbb{E}[1 - T_i] \cdot Y_i(0)}{N_c/N} \right)$$

$$= \frac{1}{N} \sum_{i=1}^N (Y_i(1) - Y_i(0))$$

$$= \tau$$

Sample Variance

- $\mathbb{V}(\bar{Y}_t - \bar{Y}_c) = \frac{S_c^2}{N_c} + \frac{S_t^2}{N_t} - \frac{S_{tc}^2}{N}$
- S_c^2 and S_t^2 are the sample variances of $Y(0)$ and $Y(1)$:

$$S_c^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i(0) - \bar{Y}(0))^2$$

$$S_t^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i(1) - \bar{Y}(1))^2$$

$$\begin{aligned} S_{tc}^2 &= \frac{1}{N-1} \sum_{i=1}^N [Y_i(1) - Y_i(0) - (\bar{Y}(1) - \bar{Y}(0))]^2 \\ &= \frac{1}{N-1} \sum_{i=1}^N (Y_i(1) - Y_i(0) - \tau)^2 \end{aligned}$$

Alternative From

- $S_{tc}^2 = S_c^2 + S_t^2 - 2\rho_{tc} \cdot S_c \cdot S_t$

where

$$\rho_{tc} = \frac{1}{(N-1) \cdot S_c \cdot S_t} \sum_{i=1}^N (Y_i(1) - \bar{Y}(1)) \cdot (Y_i(0) - \bar{Y}(0))$$

- $\mathbb{V}(\bar{Y}_t - \bar{Y}_c) = \frac{N_t}{N \cdot N_c} \cdot S_c^2 + \frac{N_c}{N \cdot N_t} \cdot S_t^2 + \frac{2}{N} \rho_{tc} \cdot S_c \cdot S_t$

Sample Variance

- When is $\mathbb{V}(\bar{Y}_t - \bar{Y}_c)$ small [large]?
- An alternative: under the assumption of a constant treatment effect, $S_c^2 = S_t^2$, therefore:

$$\mathbb{V}_{\text{const}}(\bar{Y}_t - \bar{Y}_c) = \frac{s^2}{N},$$

where s_c^2 is the estimated version of S_c^2

- Most used is the conservative Neyman variance estimator:

$$\mathbb{V}_N(\bar{Y}_t - \bar{Y}_c) = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t}$$

Correlation version

- $\mathbb{V}_\rho = \frac{N_t}{N \cdot N_c} \cdot s_c^2 + \frac{N_c}{N \cdot N_t} \cdot s_t^2 + \frac{2}{N} \rho_{tc} \cdot s_c \cdot s_t$
- By Cauchy-Schwarz inequality, $|\rho_{tc}| \leq 1$, therefore

$$\begin{aligned}\mathbb{V}_{\rho=1} &= s_c^2 \cdot \frac{N_t}{N \cdot N_c} + s_t^2 \cdot \frac{N_c}{N \cdot N_t} + s_c \cdot s_t \times \frac{2}{N} \\ &= \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t} - \frac{(s_t - s_c)^2}{N}\end{aligned}$$

- If $s_c^2 \neq s_t^2$, then $\mathbb{V}_{\rho=1} < \mathbb{V}_N$

Improving the Bound

- In recent work, this bound been improved (the bound is sharp, but the estimator is still conservative): Aronow, Green, and D. K. Lee (2014)
- They use Hoeffding's inequality instead
- The new bound is sharp in the sense that it is the smallest interval containing all values of the variance that are compatible with the observable information
- For nice results linking Neyman with the common sandwich estimator, see: Samii and Aronow (2012)

Infinite Sample Case

- Let \mathbb{E} represent the expectation over the sample and the treatment assignment; \mathbb{E}_{SP} over the sampling; and \mathbb{E}_T over the treatment assignment
- $\tau^P = \mathbb{E}_{SP} [Y_i(1) - Y_i(0)]$
(note the i)

$$\begin{aligned}\mathbb{E}_{SP}[\tau^S] &= \mathbb{E}_{SP} [\bar{Y}(1) - \bar{Y}(0)] \\ &= \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{SP} [Y_i(1) - Y_i(0)] \\ &= \tau^P\end{aligned}$$

Infinite Sample Case

- Let $\sigma^2 = \mathbb{E}_{SP}[S^2]$. Note:

$$\sigma_{tc}^2 = \mathbb{V}_{SP}[Y_i(1) - Y_i(0)] = \mathbb{E}_{SP}[(Y_i(1) - Y_i(0) - \tau^P)^2]$$

- Definition of the variance of the unit-level treatment in the population implies:

$$\mathbb{V}_{SP}(\tau^S) = \mathbb{V}_{SP}[\bar{Y}(1) - \bar{Y}(0)] = \frac{\sigma_{tc}^2}{N} \quad (1)$$

- What is $\mathbb{V}(\hat{\tau})$, where $\hat{\tau} = \bar{Y}_t - \bar{Y}_c$?

Infinite Sample Case

$$\begin{aligned}\mathbb{V}(\hat{\tau}) &= \mathbb{E} \left[(\bar{Y}_t - \bar{Y}_c - \mathbb{E} [\bar{Y}_t - \bar{Y}_c])^2 \right] \\ &= \mathbb{E} \left[(\bar{Y}_t - \bar{Y}_c - \mathbb{E}_{SP} [\bar{Y}(1) - \bar{Y}(0)])^2 \right],\end{aligned}$$

With some algebra, and noting that

$$\mathbb{E}_T [\bar{Y}_t - \bar{Y}_c - (\bar{Y}(1) - \bar{Y}(0))] = 0$$

We obtain

$$\begin{aligned}\mathbb{V}(\hat{\tau}) &= \mathbb{E} \left[(\bar{Y}_t - \bar{Y}_c - \bar{Y}(1) - \bar{Y}(0))^2 \right] \\ &\quad + \mathbb{E}_{SP} \left[(\bar{Y}(1) - \bar{Y}(0) - \mathbb{E}_{SP} [Y(1) - Y(0)])^2 \right]\end{aligned}$$

Infinite Sample Case

$$\begin{aligned}\mathbb{V}(\hat{\tau}) &= \mathbb{E} \left[(\bar{Y}_t - \bar{Y}_c - \bar{Y}(1) - \bar{Y}(0))^2 \right] \\ &+ \mathbb{E}_{SP} \left[(\bar{Y}(1) - \bar{Y}(0) - \mathbb{E}_{SP}[Y(1) - Y(0)])^2 \right]\end{aligned}$$

Recall that

$$\mathbb{E}_T \left[(\bar{Y}_t - \bar{Y}_c - \bar{Y}(1) - \bar{Y}(0))^2 \right] = \frac{S_c^2}{N_c} + \frac{S_t^2}{N_t} - \frac{S_{tc}^2}{N}$$

And by simple results from sampling,

$$\mathbb{E}_{SP} \left[\frac{S_c^2}{N_c} + \frac{S_t^2}{N_t} - \frac{S_{tc}^2}{N} \right] = \frac{\sigma_c^2}{N_c} + \frac{\sigma_t^2}{N_t} - \frac{\sigma_{tc}^2}{N}$$

Infinite Sample Case

Also recall that

$$E_{SP} \left[(\bar{Y}(1) - \bar{Y}(0) - \mathbb{E}_{SP} [Y(1) - Y(0)])^2 \right] = \frac{\sigma_{tc}^2}{N}$$

Therefore,

$$\begin{aligned}\mathbb{V}(\hat{\tau}) &= \mathbb{E} \left[(\bar{Y}_t - \bar{Y}_c - \bar{Y}(1) - \bar{Y}(0))^2 \right] \\ &+ \mathbb{E}_{SP} \left[(\bar{Y}(1) - \bar{Y}(0) - \mathbb{E}_{SP} [Y(1) - Y(0)])^2 \right] \\ &= \frac{\sigma_c^2}{N_c} + \frac{\sigma_t^2}{N_t}\end{aligned}$$

Comments

- $E_{SP}[\hat{V}_N] = \mathbb{V}_{SP}$, under the SP model
- This does not hold for the other finite-sample variance estimators we have considered, although they are sharper for the finite-sample model
- \hat{V}_N is what is almost university used



← → C ⌂ <https://data.sfgov.org> ⌂ 7

San Francisco Data



App Showcase Help Developer About Analytics SFGov Sign Up Sign In



CITY & COUNTY OF SAN FRANCISCO

SFOpenBook



SFOpenBook: A clear look at San Francisco's fiscal and economic health

SFOpenBook provides easy access to a number of interactive tools, reports and other content to shed light on the City's economy, finances, and operational performance.



Case Data from San Francisco 311

Cases created since 7/1/2008 with location information. See what the top requests are in a neighborhood or district.



Restaurant Health Inspection Scores

History of Health Inspection Scores for SF Restaurants. Read more at <http://innovatesf.com/open-data-real-impact/>



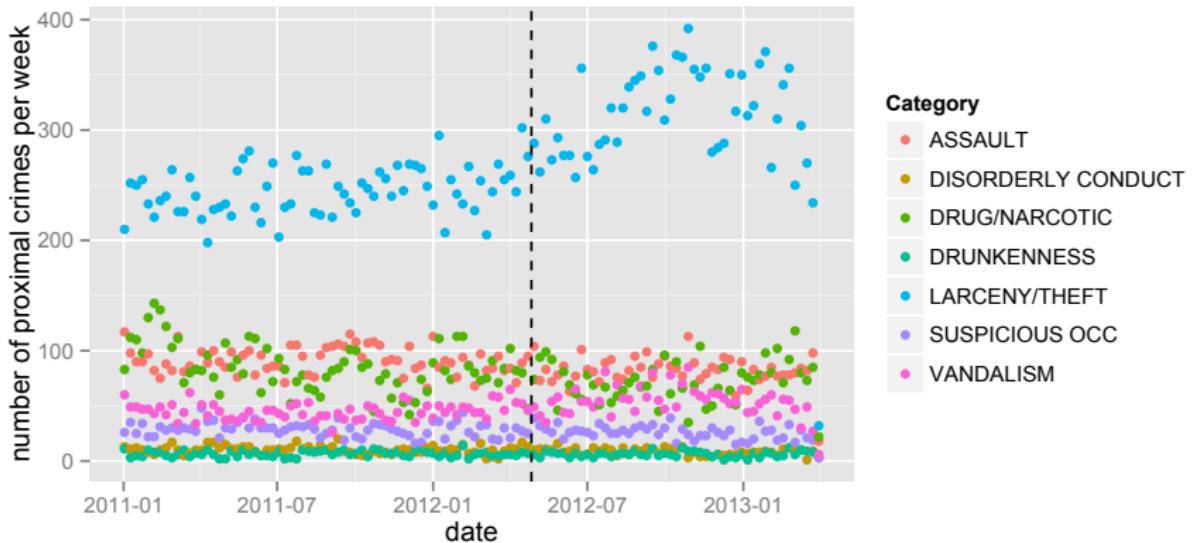
SFPD Incidents - Previous 3 Complete Months

Previous 3 completed months of incidents derived from SFPD Crime Incident Reporting system. To download complete calendar years: 2003-present:
<https://data.sfgov.org/Public->



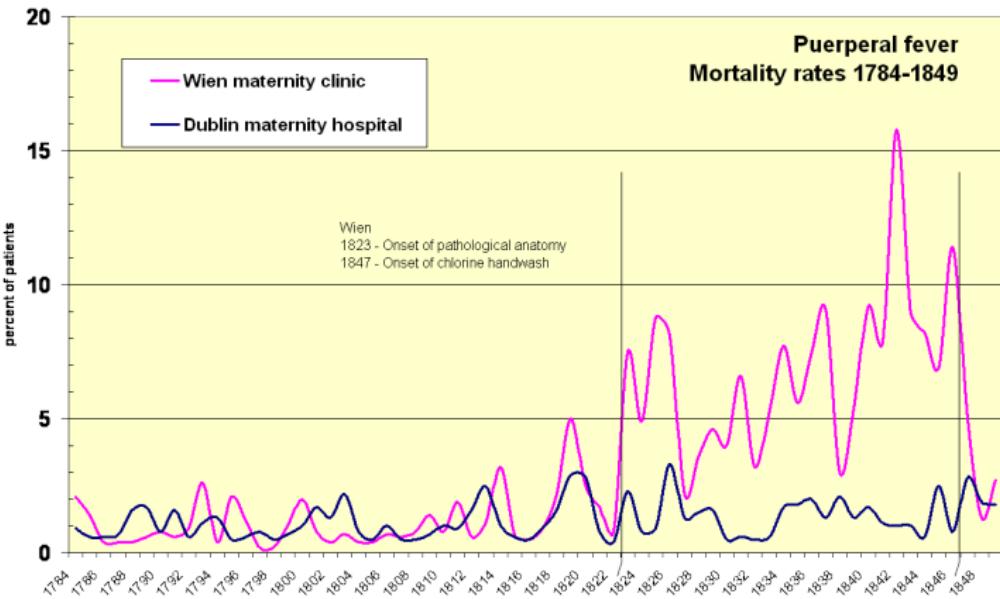
Businesses Registered In San Francisco - Active Locations

Use this link for the code classifications used in the data set:
<http://www.sftreasurer.org/Modules/ShwDocument.aspx?documentid=218>
Locations geocoded for ~75% of businesses. Does not include closed



A Comment on Design

- What would a hypothesis test look like here? What would it mean?
- What are the assumptions to make this a causal statement?
- Compare with Semmelweis and puerperal fever: see [Freedman, 2010, Chapter 20]
- Compare with the usual regression methods, such as this report: [[LINK](#)]
- Credit to: [Eytan Bakshy]



Kepler in a Stats Department

“I sometimes have a nightmare about Kepler. Suppose a few of us were transported back in time to the year 1600, and were invited by the Emperor Rudolph II to set up an Imperial Department of Statistics in the court at Prague. Despairing of those circular orbits, Kepler enrolls in our department. We teach him the general linear model, least squares, dummy variables, everything. He goes back to work, fits the best circular orbit for Mars by least squares, puts in a dummy variable for the exceptional observation - and publishes. And that’s the end, right there in Prague at the beginning of the 17th century.”

Freedman, D.A. (1985). Statistics and the scientific method. In W.M. Mason & S.E. Fienberg (Eds.), Cohort analysis in social research: Beyond the identification problem (pp. 343-366). New York: Springer-Verlag.

Making Inference with Assumptions

Fitting the Question

- The Fisher/Neyman approach required precise control over interventions
- Neyman relied on what because the dominant way to do inference
- Fisher, and before him Pierce, Charles Sanders Peirce, outlined the way that is probably closest to the data
- With the growth of computing, a renaissance in permutation inference

Lady Tasting Tea

- A canonical experiment which can be used to show the power of randomization inference.
- The example is a special case of trying to figure out if two proportions are different, which is a very common question.
- A lady (at Cambridge) claimed that by tasting a cup of tea made with milk she can determine whether the tea infusion or milk was added first (Fisher, 1935, pp. 11–25).
- The lady was B. Muriel Bristol-Roach, who was an alga biologist. [google scholar]

On Making Tea

- The precise way to make an optimal cup of tea has long been a contentious issue in China, India and Britain. e.g., George Orwell had 11 rules for a perfect cup of tea.
- On the 100th anniversary of Orwell's birth, the [Royal Society of Chemistry](#) decided to review Orwell's rules.

100 Years Later: On Making Tea

- The Society sternly challenged Orwell's advice that milk be poured in after the tea. They noted that adding milk into hot water **denatures the proteins**—i.e., the proteins begin to unfold and clump together.
- The Society recommended that it is better to have the chilled milk at the bottom of the cup, so it can cool the tea as it is slowly poured in (BBC, 2003).
- In India, the milk is heated together with the water and the tea is infused into the milk and water mix.
- In the west, this is generally known as *chai*—the generic word for tea in Hindi which comes from the Chinese character 茶 (pronounced *chá* in Mandarin) which is also the source word for tea in Punjabi, Russian, Swahili and many other languages.

Fisher's Experiment

The Lady:

- is given eight cups of tea. Four of the cups have milk put in first and four have tea put in first.
- is presented the cups in random order without replacement.
- is told of this design so she, even if she has no ability to discern the different types of cups, will select four cups to be milk first and four to be tea first.

The Fisher exact test follows from examining the mechanics of the randomization which is done in this experiment.

Fisher's Randomization Test I

There are 70 ways of choosing four cups out of eight: @

- If one chooses four cups in succession, one has 8, 7, 6, 5 cups to choose from. There are $\frac{8!}{4!} = 8 \times 7 \times 6 \times 5 = 1680$ ways of choosing four cups. @
- But this calculation has considered it a different selection if we choose cups in merely a different order, which obviously doesn't matter for the experiment. @
- Since four cups can be arranged in $4! = 4 \times 3 \times 2 \times 1 = 24$ ways, the number of possible choices is $\frac{1680}{24} = 70$.

Fisher's Randomization Test II

The Lady has:

- 1 way of making 4 correct and 0 incorrect choices @
- 16 ways of making 3 correct and 1 incorrect choices @
- 36 ways of making 2 correct and 2 incorrect choices @
- 16 ways of making 1 correct and 3 incorrect choices @
- 1 way of making 0 correct and 4 incorrect choices @
- These five different possible outcomes taken together add up to 70 cases out of 70.

Fisher's Randomization Test

Probability I

- The Lady has $p = \frac{1}{70}$ of correctly selecting the cups by chance. @
- The probability of her randomly selecting 3 correct and 1 incorrect cup is $\frac{16}{70}$ @
- But this proportion does not itself test the null hypothesis of no ability. @
- For that we want to know the chances of observing the outcome observed (3 correct and 1 wrong), plus the chances of observing better results (4 correct cups).

Fisher's Randomization Test

Probability II

- Otherwise, we could be rejecting the null hypothesis of no ability just because the outcome observed was itself rare although better results could have been frequently observed by chance. @
- In the case of the lady correctly identifying three cups correctly and one incorrectly, the p -value of the Fisher randomization tests is $\frac{16}{70} + \frac{1}{70} = \frac{17}{70} \cong 0.24$.

Fisher Exact Test

- This test is distribution (and model) free. @
- If the conditional randomization model is not correct (we will see an example in a bit), the level will still be still correct relative to the unconditional model.
- A test has level α if whenever the null is true, the chance of rejection is no greater than α . @
- But the size of the test is often less than its level if the randomization is not correct. The size of a test is the chance of rejection when the null is true.

Binomial Randomization

- This experimental design is not very sensitive. E.g., what if three cups are correctly identified? $p = \frac{17}{70} \cong 0.24$. @
- A binomial randomized experimental design would be more sensitive. @
- Follow binomial sampling: with the number of observations, $n = 8$, and the probability of having milk first be $p = 0.5$ for each cup. @
- The observations are independent.

Binomial Randomization II

- The chance of classifying correctly eight cups binomial randomized is $1 \text{ in } 2^8 = 256$, which is significantly lower than 1 in 70. @
- There are 8 ways in 256 ($p=0.032$) of incorrectly classifying only one cup. @
- Unlike in the canonical experiment, case of having both margins fixed, in the case of the binomial design, the experiment is sensitive enough to reject the null hypothesis if the lady makes a single mistake

Comments

- The extra sensitivity comes purely from the design, and not from more data. @
- The binomial design is more sensitive, but there may be some reasons to prefer the design with fixed margins. E.g., there is a chance with the binomial design that all cups would be treated alike.

Sharp Null

- Fisherian inference proceeds using a **sharp null**—i.e., all of the potential outcomes under the null are specified.
- Is the above true for the usual null hypothesis that $\bar{\tau} = 0$? Under the sharp null the equivalent is: $\tau_i = 0 \forall i$. @
- But note that we can pick any sharp null we wish—e.g., $\tau_i = -1 \forall i < 10$ and $\tau_i = 20 \forall i > 10$.

Sharp Null and Potential Outcomes

- The null hypothesis is: $Y_{i1} = Y_{i0}$
- The observed data is: $Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$
- Therefore, under the null:

$$Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0} = Y_{i0}$$

- But other models of the PO under the null are possible—e.g.,

$$Y_i = Y_{i0} + T_i \tau_0,$$

But this could be general: $Y = f(Y_{i0}, \tau_i)$, where τ is some known vector that may vary with i .

General Procedure

Test statistic: $t(T, r)$, a quantity computed from treatment assignment $T \in \Omega$ and the response r . For example, the mean. For a given $t(T, r)$, we compute a significance level:

- ① sharp null allows us to fix r , say at the observed value. @
- ② treatment assignment T follows a known randomization mechanism which we can simulate or exhaustively list. @
- ③ given (i), (ii), the observed value of the test statistic is known for all realizations of the random treatment assignment. @
- ④ we seek the probability of a value of the test statistic as large or larger than observed

Significance Level

Significance level is simply the sum of the randomization probabilities that lead to values of $t(T, r)$ greater than or equal to the observed value \tilde{T} . So,

$$P[t(T, r) \geq \tilde{T}] = \sum_{t \in \Omega} [t(t, r) \geq \tilde{T}] \times P(T = t),$$

where $P(T = t)$ is determined by the known randomization mechanism.

Further Reading

- See Rosenbaum (2002). *Observational Studies*, chp2 and
- Rosenbaum, P. R. (2002). “Covariance adjustment in randomized experiments and observational studies.” *Statistical Science* 17 286–327 (with discussion).
- Simple introduction: Michael D. Ernst (2004). “Permutation Methods: A Basis for Exact Inference.” *Statistical Science* 19:4 676–685.

SUTVA: Stable Unit Treatment Value Assumption

- We require that “the [potential outcome] observation on one unit should be unaffected by the particular assignment of treatments to the other units” (D. R. Cox, 1958, §2.4). This is called the **stable unit treatment value assumption** (SUTVA) (Rubin, 1978). @
- SUTVA implies that Y_{i1} and Y_{i0} (the potential outcomes for person i) in no way depend on the treatment status of any other person in the dataset. @
- **SUTVA is not just statistical independence between units!**

SUTVA: Stable Unit Treatment Value Assumption

No-interference implies:

$$Y_{it}^{L_i} = Y_{it}^{L_j} \quad \forall j \neq i$$

where L_j is the treatment assignment for unit i , and $t \in 0, 1$ denotes the potential outcomes under treatment and control.

SUTVA: Stable Unit Treatment Value Assumption

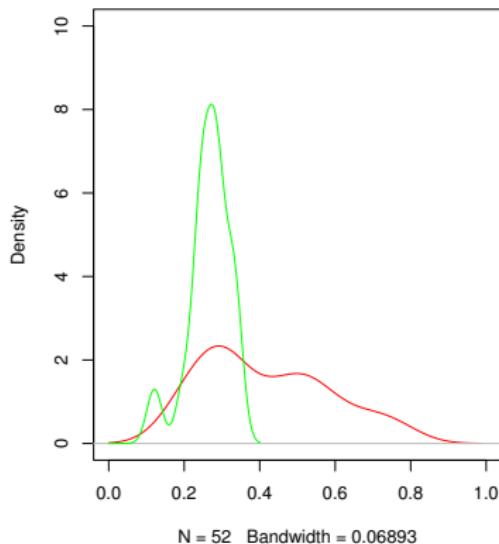
- Causal inference relies on a counterfactual of interest (Sekhon, 2004), and the one which is most obviously relevant for political information is “**how would Jane have voted if she were better informed?**”. @
- There are other theoretically interesting counterfactuals which, because of SUTVA, I do not know how to empirically answer such as “**who would have won the last election if everyone were well informed?**”.

Another Example: Florida 2004 Voting Technology

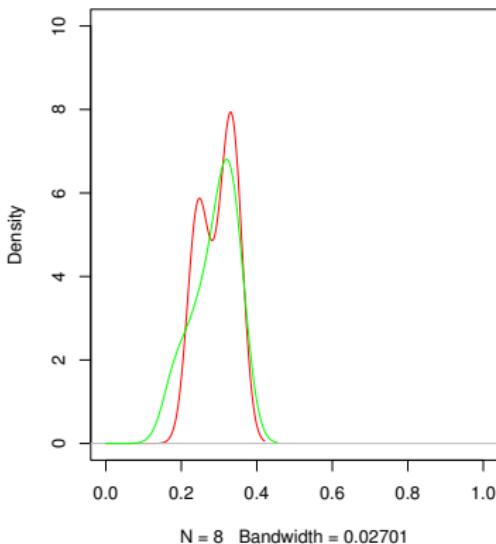
- In the aftermath of the 2004 Presidential election, many researchers argued that the **optical** voting machines that are used in a majority of Florida counties caused John Kerry to receive fewer votes than “Direct Recording Electronic” (**DRE**) voting machines (Hout et al.). @
- Hout et al. used a regression model to arrive at this conclusion. @
- Problem: **The distributions of counties were are profoundly imbalanced**

Dem Registration Pre/Post Matching

`density(x = dtam$reg04p.dem[Tr], from = 0)`



`density(x = dtam$reg04p.dem[io], from = 0)`



Densities of Democratic Registration proportions by county.
Green lines are DREs and red lines optical counties.

Average Treatment Effect for the Treated

Estimate 0.00540

SE 0.0211

p-value 0.798

balance on all other covariates, example:

	optical	dre	pvalue (t-test)
pre : Dem Reg	.551	.361	.000
post: Dem Reg	.369	.365	.556

See [http://sekhon.berkeley.edu/papers/
SekhonOpticalMatch.pdf](http://sekhon.berkeley.edu/papers/SekhonOpticalMatch.pdf) for details

Matching

- The nonparametric way to condition on X is to exactly match on the covariates. @
- This approach fails in finite samples if the dimensionality of X is large. @
- If X consists of more than one continuous variable, exact matching is inefficient: matching estimators with a fixed number of matches do not reach the semi-parametric efficiency bound for average treatment effects (Abadie and G. Imbens, 2006). @
- An alternative way to condition on X is to match on the probability of being assigned to treatment—i.e., the propensity score (P. R. Rosenbaum and Rubin, 1983). The propensity score is just one dimensional.

Propensity Score (pscore)

- More formally the propensity score is:

$$Pr(T = 1|X) = \mathbb{E}(T|X)$$

- It is a one-dimensional balancing score
- It helps to reduce the difficulty of matching
- If one balances the propensity score, one balances on the confounders X
- But if the pscore is not known, it must be estimated.
- How do we know if we estimate the correct propensity score?
- It is a tautology—but we can observe some implications

Neyman/Fisher versus OLS

- Compare the classical OLS assumptions with those from a canonical experiment described by Fisher (1935): “The Lady Tasting Tea.”
- Compare the classical OLS assumptions with those from the Neyman model

Classical OLS Assumptions

- A1. $Y_t = \sum_k X_{kt}\beta_k + \epsilon_t$, $t = 1, 2, 3, \dots, n$ $k = 1, \dots, K$, where t indexes the observations and k the variables.
- A2. All of the X variables are nonstochastic.
- A3. There is no deterministic linear relationship between any of the X variables. More precisely, the $k \times k$ matrix $\sum X_t X'_t$ is non-singular for every $n > k$.
- A4. $\mathbb{E}[\epsilon_t] = 0$ for every t , $t = 1, 2, 3, \dots, n$. Since every X is assumed to be nonstochastic (A2), (A4) implies that $\mathbb{E}[X_t \epsilon_t] = 0$. (A4) always holds if there is an intercept and if (A1)–(A3) hold.
- A5. The variance of the random error, ϵ is equal to a constant, σ^2 , for all values of every X (i.e., $\text{var}[\epsilon_t] = \sigma^2$), and ϵ is normally distributed. This assumption implies that the errors are independent and identically distributed.

Back to Basics

All of the assumptions which are referred to are listed on the assumptions slide.

The **existence** of the least squares estimator is guaranteed by A1-A3. These assumptions guarantee that $\hat{\beta}$ exists and that it is unique.

The **unbiasedness** of OLS is guaranteed by assumptions 1-4.

So, $\mathbb{E}(\hat{\beta}) = \beta$.

For **hypothesis testing**, we need all of the assumptions: A1-A5.

These allow us to assume $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$.

For **efficiency** we require assumptions A1-A5.

Correct Specification Assumption

The correct specification assumption implies that $\mathbb{E}(\epsilon_t|X_t) = 0$.

Why?

Because we are modeling the conditional mean.

$$Y_t = \mathbb{E}(Y_t|X_t) + \epsilon_t$$

Then

$$\epsilon_t = Y_t - \mathbb{E}(Y_t|X_t)$$

and

$$\begin{aligned}\mathbb{E}(\epsilon_t|X_t) &= E[Y_t - \mathbb{E}(Y_t|X_t)|X_t] \\ &= \mathbb{E}(Y_t|X_t) - E[\mathbb{E}(Y_t|X_t)|X_t] \\ &= \mathbb{E}(Y_t|X_t) - \mathbb{E}(Y_t|X_t) \\ &= 0\end{aligned}$$

Remarks

- The regression function $\mathbb{E}(Y_t|X_t)$ is used to predict Y_t from knowledge of X_t .
- The term ϵ_t is called the “regression disturbance.” The fact $\mathbb{E}(\epsilon_t|X_t) = 0$ implies that ϵ_t contains no systematic information of X_t in predicted Y_t . In other words, all information of X_t that is useful to predict Y_t has been summarized by $\mathbb{E}(Y_t|X_t)$.
- The assumption that $\mathbb{E}(\epsilon|X) = 0$ is crucial. If $\mathbb{E}(\epsilon|X) \neq 0$, $\hat{\beta}$ is biased.
- Situations in which $\mathbb{E}(\epsilon|X) \neq 0$ can arise easily. For example, X_t may contain errors of measurement.

Recall Post-Treatment Bias

The lesson of the lagged differences example is to not ask too much from regression. See:

<http://sekhon.berkeley.edu/causalinf/R/difference1.R>.

@

Many lessons follow from this. One of the most important is the issue of post-treatment bias. Post-treatment variables are those which are a consequence of the treatment of interest. Hence, they will be correlated with the treatment unless their values are the same under treatment and control.

Some Intuition: Let's derive OLS with two covars

- See my Government 1000 lecture notes from Harvard or any intro text book for more details
- Let's restrict ourselves to two covariates to help with the intuition
- Our goal is to minimize $\sum_i^n (Y_i - \hat{Y}_i)^2$, where
$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$$
.
- We can do this by calculating the partial derivatives with respect to the three unknown parameters α , β_1 , β_2 , equating each to 0 and solving.
- To simplify: deviate the observed variables by their means. These mean deviated variables are denoted by y_i , x_{1i} and x_{2i} .

$$\text{ESS} = \sum_i^n (Y_i - \hat{Y}_i)^2 \quad (2)$$

Therefore,

$$\frac{\partial \text{ESS}}{\partial \beta_1} = \hat{\beta}_1 \sum_i^n x_{1i}^2 + \hat{\beta}_2 \sum_i^n x_{1i}x_{2i} - \sum_i^n x_{1i}y_i \quad (3)$$

$$\frac{\partial \text{ESS}}{\partial \beta_2} = \hat{\beta}_1 \sum_i^n x_{1i}x_{2i} + \hat{\beta}_2 \sum_i^n x_{2i}^2 - \sum_i^n x_{2i}y_i \quad (4)$$

These can be rewritten as:

$$\sum_i^n x_{1i}y_i = \hat{\beta}_1 \sum_i^n x_{1i}^2 + \hat{\beta}_2 \sum_i^n x_{1i}x_{2i} \quad (5)$$

$$\sum_i^n x_{2i}y_i = \hat{\beta}_1 \sum_i^n x_{1i}x_{2i} + \hat{\beta}_2 \sum_i^n x_{2i}^2 \quad (6)$$

Recall:

$$\sum_i^n x_{1i}y_i = \hat{\beta}_1 \sum_i^n x_{1i}^2 + \hat{\beta}_2 \sum_i^n x_{1i}x_{2i} \quad (7)$$

$$\sum_i^n x_{2i}y_i = \hat{\beta}_1 \sum_i^n x_{1i}x_{2i} + \hat{\beta}_2 \sum_i^n x_{2i}^2 \quad (8)$$

To solve, we multiply Equation 7 by $\sum_i^n x_{2i}^2$ and Equation 8 by $\sum_i^n x_{1i}x_{2i}$ and subtract the latter from the former.

Thus,

$$\hat{\beta}_1 = \frac{(\sum_i^n x_{1i}y_i)(\sum_i^n x_{x2i}^2) - (\sum_i^n x_{2i}y_i)(\sum_i^n x_{1i}x_{2i})}{(\sum_i^n x_{1i}^2)(\sum_i^n x_{2i}^2) - (\sum_i^n x_{1i}x_{2i})^2} \quad (9)$$

And

$$\hat{\beta}_2 = \frac{(\sum_i^n x_{2i}y_i)(\sum_i^n x_{x1i}^2) - (\sum_i^n x_{1i}y_i)(\sum_i^n x_{1i}x_{2i})}{(\sum_i^n x_{1i}^2)(\sum_i^n x_{2i}^2) - (\sum_i^n x_{1i}x_{2i})^2} \quad (10)$$

If we do the same for α we find that:

$$\hat{\alpha} = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 \quad (11)$$

The equations for the estimates of $\hat{\beta}_1$ and $\hat{\beta}_2$ can be rewritten as:

$$\hat{\beta}_1 = \frac{\text{cov}(X_{1i}, Y_i) \text{var}(X_{2i}) - \text{cov}(X_{2i}, Y_i) \text{cov}(X_{1i}, X_{2i})}{\text{var}(X_{1i}) \text{var}(X_{2i}) - [\text{cov}(X_{1i}, X_{2i})]^2}$$

And,

$$\hat{\beta}_2 = \frac{\text{cov}(X_{2i}, Y_i) \text{var}(X_{1i}) - \text{cov}(X_{1i}, Y_i) \text{cov}(X_{1i}, X_{2i})}{\text{var}(X_{1i}) \text{var}(X_{2i}) - [\text{cov}(X_{1i}, X_{2i})]^2}$$

To understand the post-treatment problem in the motivating example see:

<http://sekhon.berkeley.edu/causalinf/R/part1.R>

Post-Treatment Bias Revisited

- The discussion about post-treatment bias is related to why certain causal questions are difficult if not impossible to answer (Holland, 1986).
- For example, what is the causal effect of education on voting? How do we answer this using a cross-sectional survey? @
- A more charged examples: race and sex. What is post-treatment in gender discrimination cases?

Unbiasedness

Suppose (A1)-(A4) hold. Then $\mathbb{E}(\hat{\beta}) = \beta$ —i.e., $\hat{\beta}$ is an unbiased estimator of β . *Proof:*

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'Y \\ &= (\sum_{t=1}^n X_t X_t')^{-1} \sum_{t=1}^n X_t Y_t\end{aligned}$$

By (A1)

$$\begin{aligned}&= (\sum_{t=1}^n X_t X_t')^{-1} \sum_{t=1}^n X_t (X_t' \beta + \epsilon_t) \\ &= (\sum_{t=1}^n X_t X_t')^{-1} (\sum_{t=1}^n X_t X_t') \beta + (\sum_{t=1}^n X_t X_t')^{-1} \sum_{t=1}^n X_t \epsilon_t \\ &= \beta + (\sum_{t=1}^n X_t X_t')^{-1} \sum_{t=1}^n X_t \epsilon_t\end{aligned}$$

Unbiasedness

$$\hat{\beta} = \beta + \left(\sum_{t=1}^n X_t X_t' \right)^{-1} \sum_{t=1}^n X_t \epsilon_t$$

Given (A2) and (A4)

$$E \left[\left(\sum_{t=1}^n X_t X_t' \right)^{-1} \sum_{t=1}^n X_t \epsilon_t \right] = \left(\sum_{t=1}^n X_t X_t' \right)^{-1} \sum_{t=1}^n X_t E[\epsilon_t] = 0$$
$$\mathbb{E}(\hat{\beta}) = \beta$$

Unbiasedness may be considered a necessary condition for a good estimator, but is certainly isn't sufficient. Moreover, it is also doubtful whether it is even necessary. For example, in general, Maximum Likelihood Estimators are not unbiased—e.g., Logit, Probit, Tobit.

Variance-Covariance Matrix

Suppose Assumptions (A1)-(A5) hold. Then, the variance-covariance matrix of $\hat{\beta}$ is:

$$E \left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \right] = \sigma^2 (X'X)^{-1} \quad (12)$$

Proof: Variance-Covariance Matrix

$$\hat{\beta} = \beta + \left(\sum_{t=1}^n X_t X_t' \right)^{-1} \sum_{t=1}^n X_t \epsilon_t$$

$$\hat{\beta} - \beta = (X'X)^{-1} X' \epsilon$$

$$\begin{aligned} (\hat{\beta} - \beta)(\hat{\beta} - \beta)' &= (X'X)^{-1} X' \epsilon \left[(X'X)^{-1} X' \epsilon \right]' \\ &= (X'X)^{-1} X' \epsilon \left[\epsilon' X (X'X)^{-1} \right] \\ &= (X'X)^{-1} X' (\epsilon \epsilon') X (X'X)^{-1} \end{aligned}$$

$$\begin{aligned} \mathbb{E}(\hat{\beta} - \beta)(\hat{\beta} - \beta)' &= (X'X)^{-1} X' \mathbb{E}(\epsilon \epsilon') X (X'X)^{-1} \\ &= (X'X)^{-1} X' \sigma^2 I X (X'X)^{-1} \end{aligned}$$

$$\begin{aligned} \text{move } \sigma^2 \text{ and drop } I &= \sigma^2 (X'X)^{-1} X' X (X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1} \end{aligned}$$

The sigma-squared Matrix

Note that $\sigma^2 I$ is an $n \times n$ matrix which look like:

$$\begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & \sigma^2 \end{bmatrix}$$

Estimating Sigma-squared

Recall that:

Let $\hat{\epsilon}_i = Y_i - \hat{Y}_i$, then

$$\hat{\sigma}^2 = \frac{\sum \hat{\epsilon}_i^2}{n - k}, \quad (13)$$

where k is the number of parameters.

Heteroscedasticity I

Assumption **A5** on the assumptions slide makes the homogeneous or constant variance assumption. This assumption is clearly wrong in many cases, this is particularly true with cross national data.

We can either use an estimator which directly models the heteroscedasticity and weights each observations appropriately—e.g, Weighted Least Squares (WLS)—or we can use robust standard errors.

Heteroscedasticity II

The OLS estimator is unbiased (and consistent) when there is heteroscedasticity, but it is **not** efficient. We say that $\hat{\alpha}$ is an efficient unbiased estimator if for a given sample size the variance of $\hat{\alpha}$ is smaller than (or equal to) the variance of any other unbiased estimator.

Why is OLS still unbiased? Recall that when we proved that OLS is unbiased, the variance terms played no part in the proof—i.e., Assumption A5 played no part.

Univariate Matching

- Assume that we obtain conditional independence if we condition on one confounder, X_{iG} , where i indexes the observation and $G \in \{T, C\}$ denotes the treatment or control group. @
- Say we wish to compare two matched pairs to estimate ATT, $\tau_{i'} = Y_{i1} - Y_{i'0}$, where ' denotes the matched observation. @
- Matched pairs can be used to estimate τ because the treatment observation reveals Y_{i1} and the control observation Y_{i0} under conditional exchangeability. @
- Bias in the estimate of $\tau_{i'}$ is assumed to be some unknown increasing function of $|X_{iT} - X_{i'C}|$.

Univariate Matching

- We can then minimize our bias in $\hat{\tau}_{i'}$ and hence $\hat{\tau}$ by minimizing $|X_{iT} - X_{i'C}|$ for each matched pair.
- Can be minimized non-parametrically by “nearest neighbor matching” (NN). The case with replacement is simple. @
- With NN matching we just match the nearest $X_{i'C}$ to a given X_{iT} . @
- For ATT, we match every X_{iT} with the nearest $X_{i'C}$. @
- For ATC, we match every X_{iC} with the nearest $X_{i'T}$. @
- For ATE, we match every X_{iT} with the nearest $X_{i'C}$ AND match every X_{iC} with the nearest $X_{i'T}$. @
- Let's walk through some examples: nn1.R

Generalized Linear Models

Generalized linear models (GLMs) extend linear models to accommodate both non-normal distributions and transformations to linearity. GLMs allow unified treatment of statistical methodology for several important classes of models.

This class of models can be described by two components:
stochastic and systematic.

- Stochastic component:

- Normal distribution for LS: $\epsilon_i \sim N(0, \sigma^2)$
- For logistic regression:

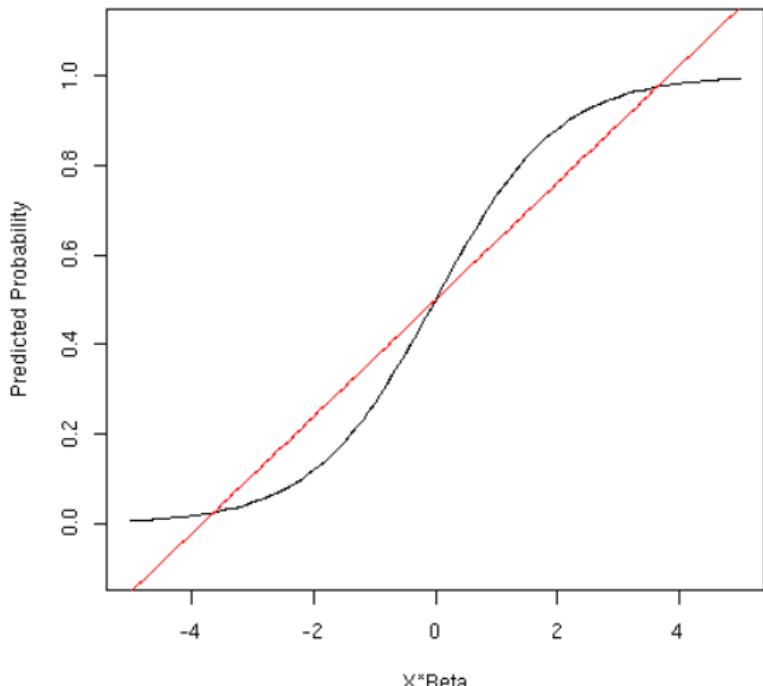
$$Y_{Bern}(y_i|\pi_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \quad \begin{cases} \pi_i & \text{for } y = 1 \\ 1 - \pi_i & \text{for } y = 0 \end{cases}$$

- Systematic Component:

- For LS: $E(Y_i|X_i) = X\beta$
- For logistic regression:

$$Pr(Y_i = 1|X_i) \equiv E(Y_i) \equiv \pi_i = \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)}$$

Logit Link vs. Identity Link



Logit Link vs. Identity Link

The black line is the logistic fitted values and the red line is from OLS (the identity link). For R code see, logit1.R

OLS goes unbounded, which we know cannot happen!

Defining Assumptions

A generalized linear model may be described by the following assumptions:

- There is a response Y observed independently at fixed values of stimulus variables X_1, \dots, X_k .
- The stimulus variables may only influence the distribution of Y through a single linear function called the *linear predictor* $\eta = \beta_1 X_1 + \dots + \beta_k X_k$.

Defining Assumptions

- The distribution of Y has density of the form

$$f(Y_i; \theta_i, \phi) = \exp[A_i\{Y_i\theta_i - \gamma(\theta_i)\}/\phi + \tau(Y_i, \phi/A_i)], \quad (14)$$

where ϕ is a *scale parameter* (possibly known). A_i is a *known* prior weight and parameter θ_i depends upon the linear predictor.

This is the exponential family of distributions. Most of the distributions you know are part of this family including the normal, binomial and Poisson.

Models of the Mean

The mean $\mu = E(Y)$ is a smooth invertible function of the linear predictor:

$$\mu = m(\eta), \tag{15}$$

$$\eta = m^{-1}(\mu) = I(\mu), \tag{16}$$

where the inverse function $I(\cdot)$ is called the *link function*

Models of the Mean: examples

For LS we use the identity link (also called the canonical link):
 $I(\mu) = \mu$. Thus,

$$X_i\beta = \eta_i = m(\mu_i) = \mu_i = E(Y_i)$$

For logistic regression we use the logit link:

$$I(\mu) = \log\left(\frac{\pi}{1-\pi}\right). \text{ Thus,}$$

$$X_i\beta = \eta_i = m(\mu_i) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} = \mu_i = E(Y_i) = Pr(Y_i = 1|X_i),$$

note that $\frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$ is the inverse logit link.

GLM: Gaussian

GLMs allow unified treatment of many models—e.g., normal:
The probability distribution function in the form one usually
sees the distribution is:

$$f(y; \mu) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(y - \mu)^2\right],$$

where μ is the parameter of interest and σ^2 is regarded, in this setting, as a nuisance parameter.

The following is the canonical form of the distribution
 $\theta = \mu$, $\gamma(\theta) = \theta^2/2$ and $\phi = \sigma^2$ so we can write:

$$\log f(y) = \frac{1}{\phi} \left\{ y\mu - \frac{1}{2}\mu^2 - \frac{1}{2}y^2 \right\} - \frac{1}{2} \log(2\pi\phi).$$

Notes on the Gaussian

- If ϕ , the variance or what is called the dispersion in the GLM framework, were known the distribution of y would be a one-parameter canonical exponential family.
- An unknown ϕ is handled as a nuisance parameter by moment methods. For example, note how $\hat{\sigma}^2$ is estimated in the least squares case.
- We can do this because θ and ϕ are orthogonal parameters. This means that we can estimate θ and then conditioning on this estimate, calculate ϕ . We don't have to jointly estimate both.

GLM: Poisson Example

For a Poisson distribution with mean μ we have

$$\log f(y) = y \log(\mu) - \mu - \log(y!),$$

where $\theta = \log(\mu)$, $\phi = 1$, and $\phi(\theta) = \mu = e^\theta$.

GLM: Binomial Example

For a binomial distribution with fixed number of trials a and parameter p we take the response to be $y = s/a$ where s is the number of “successes”. The density is

$$\log f(y) = a \left[y \log \frac{p}{1-p} + \log(1-p) \right] + \log \binom{a}{ay}$$

where we take $A_i = a_i$, $\phi = 1$, θ to be the logit transform of p and $\gamma(\theta) = -\log(1-p) = \log(1+e^\theta)$.

The functions supplied with **R** for handling generalized linear modeling distributions include `glm`, `gaussian`, `binomial`, `poisson`, `inverse.gaussian` and `Gamma`.

Logistic Regression

Let Y^* be a continuous unobserved variable such as the **propensity** to vote for a particular political party, say the Republican Party, or the **propensity** to be assigned to treatment.

Assume that Y_i^* has some distribution whose mean is μ_i . And that Y_i^* and Y_j^* are independent for all $i \neq j$, conditional on X.

Logistic Regression

We observe Y_i such that:

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* \geq \tau \text{ if } i \text{ is treated or votes Republican} \\ 0 & \text{if } Y_i^* < \tau \text{ if } i \text{ is NOT treated or does NOT vote Republican} \end{cases}$$

- Since Y^* is unobserved we usually define the threshold, τ , to be zero.
- Note that if Y_i^* is observed and that it is normally distributed conditional on X , we have LS regression. If only Y_i is observed and that Y^* is distributed standardized logistic (which is very similar to the normal distribution), we obtain the logistic regression model.
- Note that our guess of Y^* is $\hat{\mu}_i$ —or, equivalently, $\hat{\eta}_i$.

Logistic Regression Coefficients

- The estimated coefficients, $\hat{\beta}$, in a logistic regression can be interpreted in a manner very similar to that the coefficients in LS. But there are important differences which complicate the issue.
- We should interpret β as the regression coefficients of Y^* on X . So, $\hat{\beta}_1$ is what happens to Y_i^* when X_1 goes up by one unit when all of the other explanatory variables remain the same. **ICK! Causal**
- The interpretation is complicated by the fact that the link function is nonlinear. So the effect of a one unit change in X_1 on the predicted probability is different if the other explanatory variables are held at one given constant value versus another. Why is this?

Logistic Regression Assumptions and Properties

- The assumptions of logistic regression are very similar to those of least squares—see the assumptions slide. But there are some important difference. The most important is that in order to obtain consistency, logistic regression requires the homoscedasticity assumption while LS does not.
- We also need the no perfect separation assumption for the existence of the estimator.
- Another important difference is that logistic regression is **NOT** unbiased. But it is consistent.

Logistic Regression Loss Function and Estimation

The loss function for logistic regression is:

$$\sum_{i=1}^n -[Y_i \log(\pi_i) + (1 - Y_i) \log(1 - \pi_i)],$$

recall that $\pi_i = \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)}$.

The reasons for this loss function are made clear in a ML course. As will show this loss function is usually minimized—iterated weighted least squares.

See Venables and Ripley (2002) for more estimation details.

The Propensity Score

- The propensity score allows us to match on X without having to match all of the k -variables in X . We may simply match on the probability of being assigned to treatment—i.e., the propensity score. The propensity score is just one dimensional. @
- More formally the propensity score (also called the balancing score) is:

$$Pr(T_i = 1|X_i) = E(T_i|X_i)$$

The following theory is from Rosenbaum and Rubin (1983)

A1: Ignorability

Assumption 1: differences between groups are explained by observables (ignorability)

$$(Y_0, Y_1) \perp T | X \quad (A1)$$

Treatment assignment is “strictly ignorable” given X . This is equivalently to:

$$\Pr(T = 1 | Y_0, Y_1, X) = \Pr(T = 1 | X), \text{ or}$$

$$E(T | Y_0, Y_1, X) = E(T | X)$$

(A2) Overlap Assumption

Assumption 2: there exists common support for individuals in both groups (overlap)

$$0 < P(T = 1 | X = x) < 1 \quad \forall x \in X \quad (A2)$$

A1 and A2 together define strong ignorability (Rosenbaum and Rubin 1983).

@

A2 is also known as the Experimental Treatment Assignment (ETA) assumption

Observable Quantities

We observe the conditional distributions:

- $F(Y_1 | X, T = 1)$ and
- $F(Y_0 | X, T = 0)$

But not the joint distributions:

- $F(Y_0, Y_1 | X, T = 1)$
- $F(Y_0, Y_1 | X)$

Nor the impact distribution:

- $F(Y_1 - Y_0 | X, T = 1)$

Weaker Assumptions

For ATT, weaker assumptions can be made:

- A weaker conditional mean independence assumption on Y_0 is sufficient (Heckman, Ichimura, and Todd 1998).

$$E(Y_0 | X, T = 1) = E(Y_0 | X, T = 0) = E(Y_0 | X)$$

- A weaker support condition:

$$\Pr(T = 1 | X) < 1$$

This is sufficient because we only need to guarantee a control analogue for every treatment, and not the reverse

Pscore as a Balancing Score

Let $e(X)$ be the pscore: $\Pr(T = 1 | X)$.

$$F(X | e(X), T) = F(X | e(X)),$$
$$X \perp T | e(X)$$

That is, the distribution of covariates at different treatment/exposure levels is the same once one conditions on the propensity score

Balancing Score: intuition

We wish to show that if one matches on the true propensity score, asymptotically, the observed covariates, X , will be balanced between treatment and control groups. Assume that conditioning on the true propensity score does not balance the underlying X covariates. Then there exists x_1 and x_2 such that $e(X_1) = e(X_2)$, so that conditioning on $e(X)$, still results in imbalance of x_1 and x_2 . Imbalance by definition implies that $Pr(T_i = 1 | x_1) \neq Pr(T_i = 1 | x_2)$, but this contradicts $e(x_1) = e(x_2)$ since by the definition of the propensity score:

$$e(x_1) = Pr(T_i = 1 | x_1), \text{ and}$$

$$e(x_2) = Pr(T_i = 1 | x_2).$$

Therefore, the observed covariates X must be balanced after conditioning on the true propensity score. For details:

P. R. Rosenbaum and Rubin (1983), especially theorems 1–3.

Pscore as a Balancing Score

Proof:

$$\Pr(T = 1 \mid X, e(X)) = \Pr(T = 1 \mid X) = e(X)$$

$$\Pr(T = 1 \mid e(X)) = E\{E(T \mid X) \mid e(X)\} =$$

$$E\{e(X) \mid e(X)\} = e(X)$$

Balancing Score, finer

$b(X)$ is a balancing score, that is,

$$T \perp X \mid b(X),$$

if and only if $b(X)$ is finer than $e(X)$ —i.e., $e(X) = f\{b(X)\}$ for some function f . Rosenbaum and Rubin Theorem 2.

Suppose $b(X)$ is finer than $e(X)$. Then it is sufficient to show:

$$\Pr(T = 1 \mid b(X)) = e(X)$$

Note:

$$\Pr(T = 1 \mid b(X)) = E\{e(X) \mid b(X)\}$$

But since $b(X)$ is finer than $e(X)$:

$$E\{e(X) \mid b(X)\} = e(X)$$

Balancing Score, coarser

Now assume $b(X)$ is not finer than $e(X)$. So, there exists x_1 and x_2 s.t $e(x_1) \neq e(x_2)$ but $b(x_1) = b(x_2)$. But, by definition $\Pr(T = 1 | x_1) \neq \Pr(T = 1 | x_2)$, so that T and X are not conditionally independent given $b(x)$, and thus $b(X)$ is not a balancing score. Therefore, to be a balancing score $b(X)$ must be finer than to $e(X)$.

Pscore and Strong Ignorability

Theorem 3: If treatment assignment is strongly ignorable given X , it is strongly ignorable given any balancing score $b(x)$

- It is sufficient to show:

$$\Pr(T = 1 \mid Y_0, Y_1, b(x)) = \Pr(T = 1 \mid b(x))$$

- By Theorem 2, this is equiv to showing:

$$\Pr(T = 1 \mid Y_0, Y_1, b(x)) = e(x)$$

Pscore and Strong Ignorability II

Note:

$$\Pr(T = 1 \mid Y_0, Y_1, b(x)) = E\{\Pr(T = 1 \mid Y_0, Y_1, X) \mid Y_0, Y_1, b(X)\}$$

By assumption equals:

$$E\{\Pr(T = 1 \mid X) \mid Y_0, Y_1, b(x)\}$$

Which by definition equals:

$$E\{e(X) \mid Y_0, Y_1, b(X)\},$$

which since $b(X)$ is finer than $e(X)$, equals $e(X)$.

Pscore as Balancing Score III

We have shown (where, $Y_{0,1}$ is a compact way of writing Y_0, Y_1):

$$E(T \mid Y_{0,1}, \Pr(T = 1 \mid X)) =$$

$$E(E(T \mid Y_{0,1}, X) \mid Y_{0,1}, \Pr(T = 1 \mid X)),$$

So that

$$E(T \mid Y_{0,1}, X) = E(T \mid X) \implies$$

$$E(T \mid Y_{0,1}, \Pr(T = 1 \mid X)) = E(T \mid \Pr(T = 1 \mid X))$$

Correct Specification of the Propensity Score?

- It is unclear how one should correctly specify the propensity score.
- The good news is that the propensity score is serving a very clear purpose: matching on the propensity score should induce balance on the baseline covariates.
- This can be directly tested by a variety of tests such as difference of means. But difference of means are not by themselves sufficient.
- **THE BAD NEWS:** this only works for the observables.

Estimating the Propensity Score

- We may estimate the propensity score using logistic regression where the treatment indicator is the dependent variable. @
- We then match on either the predicted probabilities $\hat{\pi}_i$ or the linear predictor $\hat{\eta}_i$. It is in fact preferable to match on the linear predictor because the probabilities are compressed.
- For matching, we do not care about the parameters, $\hat{\beta}$. We only care about getting good estimates of $\hat{\pi}_i$.

Theory of Matching, revised

Both of the proceeding assumptions can be weakened if we are simply estimating ATT

Assumption 1: differences between groups are explained by observables

$$Y_0 \perp T | X \quad (A1)$$

Assumption 2: there exists common support for individuals in both groups

$$P(T = 1 | X = x) < 1 \quad \forall x \in X \quad (A2)$$

Obvious parallel changes for ATC.

Theory of Propensity Scores

PROBLEM: Simple matching approach fails in finite samples if the dimensionality of X is large.

An alternative way to condition on X is to match on the probability of being assigned to treatment—i.e., the propensity score. The propensity score is just one dimensional.

Defn: Propensity score is the probability that a person is treated/manipulated.

$$b(x) = P(T = 1|X = x)$$

Rosenbaum and Rubin (1983) show that we can reduce matching to this single dimension, by showing that the assumptions of matching (A1 and A2) lead to,

$$Y_0 \perp T | P(X)$$

Propensity Score Matching

Matching on $b(X)$ similar to matching on X :

Step 1: for each Y_{1i} find Y_{0j} which has similar (or proximate) $b(X)$.

One-one matching:

- Nearest-neighbor: choose closest control on $P(X_i)$
 - Caliper: same, but drop cases which are too far

Many-one: use weighted average of neighbors near $P(X_i)$

$$Y_{0i} = \sum_j W_{ij} Y_j$$

E.g., using a weighted method where

$$W \propto K \left(\frac{P_i - P_j}{h} \right)$$

Step 2: same, take average over matched values of Y_{1i} and Y_{0i}

Properties of Pscore Matching I

- The modeling portion of the estimator is limited to the model of $p(X_i)$. Estimation of this model requires **no** knowledge of the outcome. @
- Unlike in the regression case, there is a clear standard for choosing an optimal model; it is the model which balances the covariates, X . @
- The key assumption required is that no variable has been left unobserved which is **correlated** with **treatment assignment** *and* with the **outcome**. This is called the selection on observables assumption.

Properties of Pscore Matching II

- No functional form is implied for the relationship between treatment and outcome. No homogeneous causal effect assumption has been made. @
- Since we are interested in a lower dimensional representation of X_i , in particular $p(X_i)$, we do not need to estimate consistently any of the individual parameters in our propensity model—they don't even need to be identified.

Properties of Pscore Matching III

- Because we are taking a conditional expectation, we need to decide what to condition on. If we condition too little, we do not obtain strong ignorability. If we condition on variables which are not baseline variables, we also obtain biased estimates.

Practice of Matching (revisited)

Step 1: for each individual i in the treated group ($T=1$) find one or more individuals i' from the untreated group ($T=0$) who is equivalent (or proximate) in terms of X .

- (a) If more than one observation matches, take average
- (b) If no exact matches exist, either drop case (because of non-common support) or take closest / next best match(es)

Step 2:

$$\bar{\tau} = \frac{1}{N} \sum(Y_{1i}|X) - \frac{1}{N'} \sum(Y_{0i'}|X)$$

A simple example

i	T_i	X_i	Y_i	\mathcal{J}	$Y_i(0)$	$\hat{Y}_i(1)$	$K_M(i)$
1	0	2	7	{5}	7	8	3
2	0	4	8	{4, 6}	8	7.5	1
3	0	5	6	{4, 6}	6	7.5	0

i	T_i	X_i	Y_i	\mathcal{J}	$\hat{Y}_i(0)$	$Y_i(1)$	$K_M(i)$
4	1	3	9	{1, 2}	7.5	9	1
5	1	2	8	{1}	7	8	1
6	1	3	6	{1, 2}	7.5	6	1
7	1	1	5	{1}	7	5	0

where observed outcome is Y , covariate to match on is X , \mathcal{J} indexes

matches, $K_M(i)$ is number of times used as a match divided by $\#\mathcal{J}$;

$$\hat{Y}_i = 1/\#\mathcal{J}(i) \sum_{l \in \mathcal{J}(i)} Y_l.$$

The estimate of ATE is .1428. ATT is -0.25 . And ATC is $\frac{2}{3}$.

Alternative test statistics

Treatment may be multiplicative, rather than constant additive.
So the ratio of treated/non-treated should be constant: e.g.,

$$Y_1 = a Y_0$$

$$\log(Y_1) - \log(Y_0) = \log(Y_1/Y_0) = \log(a)$$

so,

$$\bar{\tau}_2 = \frac{1}{N} \sum \log(Y_{1,i}|X) - \frac{1}{N'} \sum \log(Y_{0,p}|X)$$

Robust way to discern differences (e.g., if you fear outliers distorting means) transform to ranks:

Let R_i be the rank of the i observation, pooling over treated and non-treated.

$$\bar{\tau}_3 = \frac{1}{N} \sum (R_{i|X}) - \frac{1}{N'} \sum (R_{p|X})$$

Comments

- If you think the variances are changed by treatment, but not the location, then test differences in variances.
- The point: your test statistic should be motivated by your theory of what is changing when manipulation / treatment occurs!
- One could in theory estimate any model one wishes on the matched data. Such as regression, logistic regression etc.

RDD Readings

- Lee, D., 2008. "Randomized experiments from non-random selection in U.S. house elections." *Journal of Econometrics* 142:2 675–697.
- Devin Caughey and Jasjeet S. Sekhon. "Elections and the Regression-Discontinuity Design: Lessons from Close U.S. House Races, 1942–2008." *Political Analysis*. 19 (4): 385–408. 2011.
- Thistlethwaite, D. and D. Campbell. 1960. "Regression-Discontinuity Analysis: An alternative to the ex post facto experiment." *Journal of Educational Psychology* 51 (6): 309–317.
- Gerber, Alan S. and Donald P. Green. "Field Experiments and Natural Experiments." *Oxford Handbook of Political Methodology*.
- Hahn J., P. Todd, and W.V. der Klaauw, 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica*.

Regression Discontinuity Design

Context: Individual receives treatment only iff observed covariate Z_i crosses known threshold z_0 .

Logic: compare individuals just above and below threshold

@

Key assumptions:

1. Probability of receiving treatment jumps discontinuously at z_0
2. Continuity of outcome at threshold

$E[Y_0|Z = z]$ and $E[Y_1|Z = z]$ are continuous in z at z_0

@3. Confounders vary smoothly with Z —i.e., there are no jumps at the treatment threshold z_0

Regression Discontinuity Design

If the potential outcomes are distributed smoothly at the cut-point, the RD design estimates the average causal effect of treatment at the cut-point, $Z_i = c$:

$$\tau_{RD} \equiv \mathbb{E}[Y_i(1) - Y_i(0)|Z_i = c] =$$

$$\lim_{Z_i \downarrow c} \mathbb{E}[Y_i(1)|Z_i = c] - \lim_{Z_i \uparrow c} \mathbb{E}[Y_i(0)|Z_i = c]$$

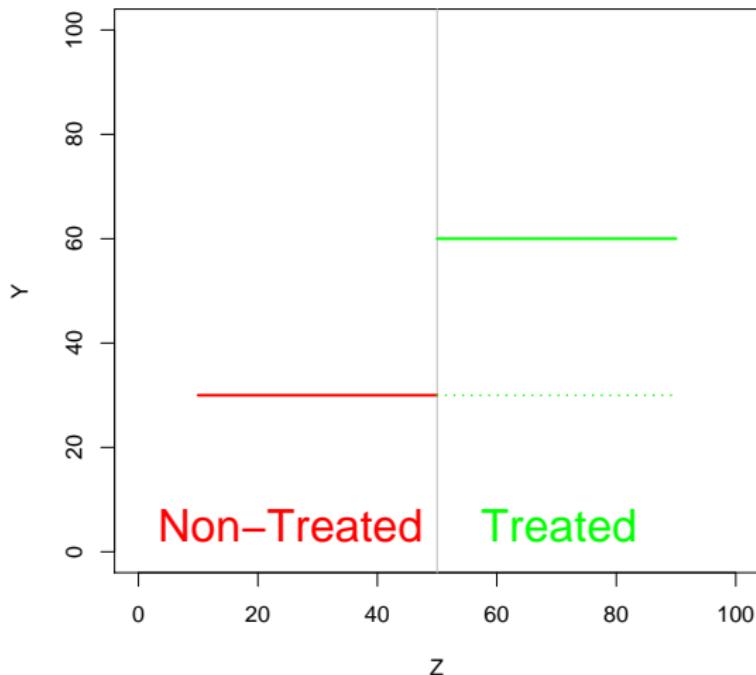
Two cases of RD

Constant treatment effect: Z_i does not independently impact outcome except through treatment

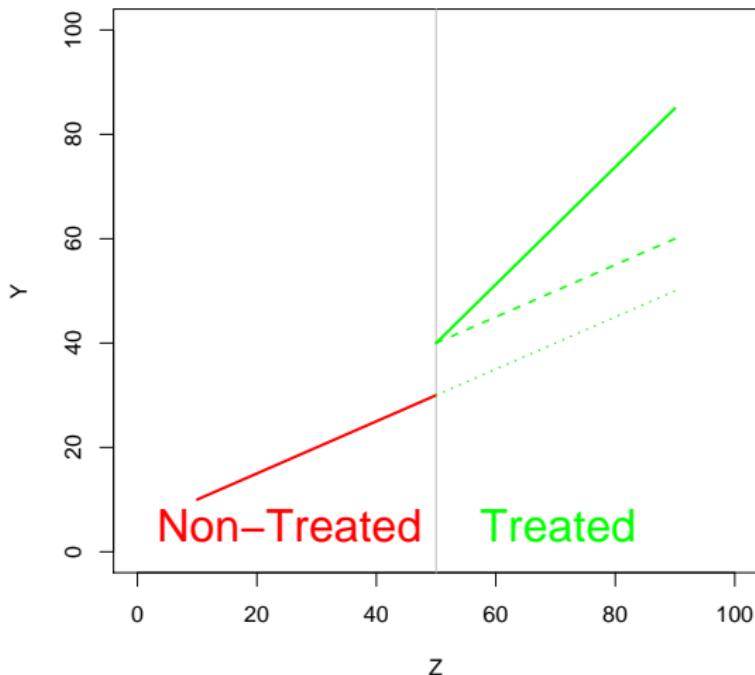
Variable treatment effect: Z_i also impacts outcome

Issue: We are estimating the Local Average Treatment Effect (LATE)

RDD with Constant Treatment Effects



RDD with Variable Treatment Effects



Example 1: Scholarship Winners

- From Kenya to the U.S. RDD has been used to estimate the effect of school scholarships.
- In these cases, z is the measure of academic performance which determines if one receives a scholarship—e.g., test scores, GPA.
- Of course, comparing everyone who wins a scholarship with everyone who loses is not a good way to proceed.
- Focus attention on people who just won and who just lost.
- This reduces omitted variable/selection bias, why?

RDD and Omitted Variable Bias I

- Let's assume that there is a homogeneous causal effect:

$$Y_i = \alpha + \beta T_i + \gamma X_i + \epsilon_i,$$

where T is the treatment indicator and X is a confounder.

@

- Then:

$$E(Y_i | T_i = 1) - E(Y_i | T_i = 0)$$

$$@ = [\alpha + \beta + \gamma E(X_i | T_i = 1) + E(\epsilon_i | T_i = 1)]$$

$$- [\alpha + 0 + \gamma E(X_i | T_i = 0) + E(\epsilon_i | T_i = 0)]$$

$$@ = \beta + \gamma [E(X_i | T_i = 1) - E(X_i | T_i = 0)]$$

True Effect + *Bias*

RDD and Omitted Variable Bias II

- Recall

$$\begin{aligned} & E(Y_i | T_i = 1) - E(Y_i | T_i = 0) \\ = & \beta + \gamma [E(X_i | T_i = 1) - E(X_i | T_i = 0)] \\ & \text{True Effect} + \text{Bias} \end{aligned}$$

- Since we assume that X varies smoothly with respect to Z (i.e., there are no jumps at the treatment threshold z_0): @

$$\begin{aligned} &= \beta + \gamma [(x^* + \delta) - x^*] \\ @ &= \beta + \gamma^* \delta \\ &\approx \beta \end{aligned}$$

Sharp vs. Fuzzy Design

- Sharp design: T_i is a deterministic function of Z_i
- Fuzzy design: T_i is a function of Z_i and an exogenous random variable:
- Fuzzy design requires observing Z_i separately from the random component
- We will cover this after instrumental variables (J. D. Angrist, G. W. Imbens, and Rubin, 1996)
- because intention-to-treat and IV for compliance correction can be used

Inc incumbency Advantage (IA)

- The **Personal Inc incumbency Advantage** is the candidate specific advantage that results from the benefits of office holding—e.g:
 - name recognition; incumbency cue
 - constituency service
 - public position taking
 - providing pork

@

- The **Incumbent Party Advantage**:
 - *ceteris paribus*, voters have a preference for remaining with the same party they had before
 - parties have organizational advantages in some districts

Problems with Estimating the IA

Empirical work

- agrees there is a significant incumbency advantage
- debates over source and magnitude

Theoretical work

- argues positive estimates of IA are spurious:
Cox and Katz 2002; Zaller 1998; Rivers 1988
- selection and survivor bias
- strategic candidate **entry** and **exit**

Estimates of Inc incumbency Advantage

Early empirical strategies:

- Regression model (Gelman-King)
~10%
- Sophomore-surge (Erikson; Levitt-Wolfram)
~7%

@

Design-based approaches

- Surprise loss of incumbents through death (Cox-Katz)
2%
- Redistricting and personal vote (Sekhon-Titiunik)
0%, Texas; 2%, California
- Regression-Discontinuity with close election (Lee)
10% increase in vote shares for incumbent party
40% increase in win probability for incumbent party

Regression-Discontinuity (RD) and Elections

Logic of applying Regression-Discontinuity to elections

- treatment: being elected (incumbent)
- assigned by getting 50%+1 of vote
- candidates near 50% randomly assigned to treatment @
- first applications: D. S. Lee, 2001, Pettersson-Lidbom, 2001

@

We show

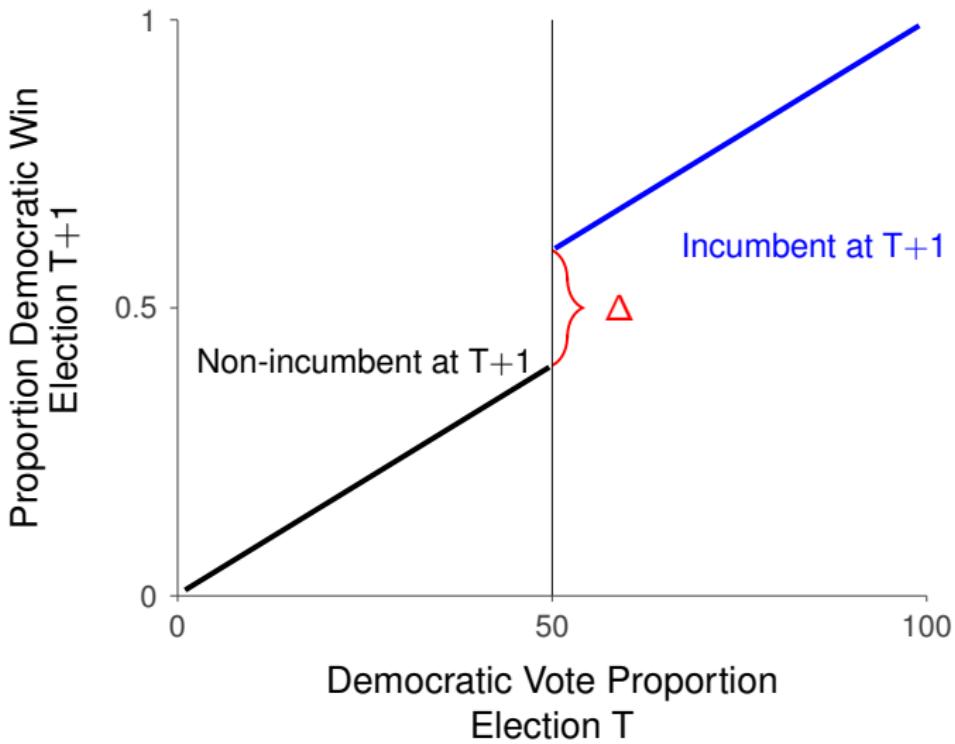
- RD doesn't work for U.S. House elections
- there are substantive implications

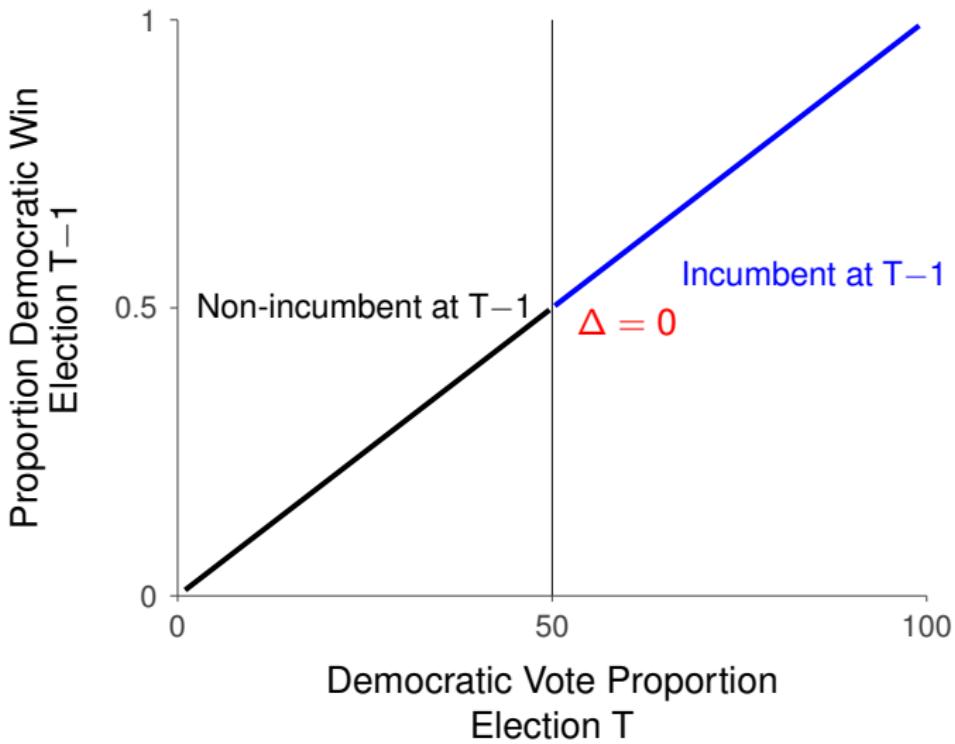
More Information and Cleaner Data

We created a new version of U.S. House elections dataset

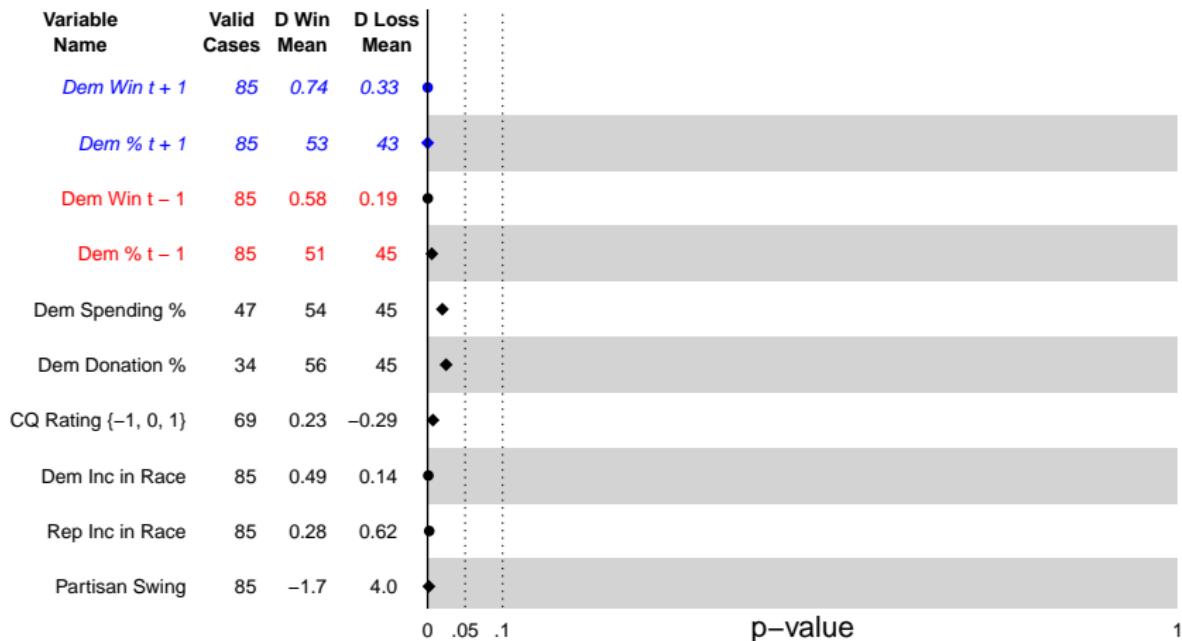
Key features:

- Corrected vote (and outcome) data, 1942–2008
validation of election results for every close race
- Incorporates 30 additional covariates
e.g., campaign expenditures, CQ scores
- Recounts, fraud, and other oddities:
random sample of 50% of close elections

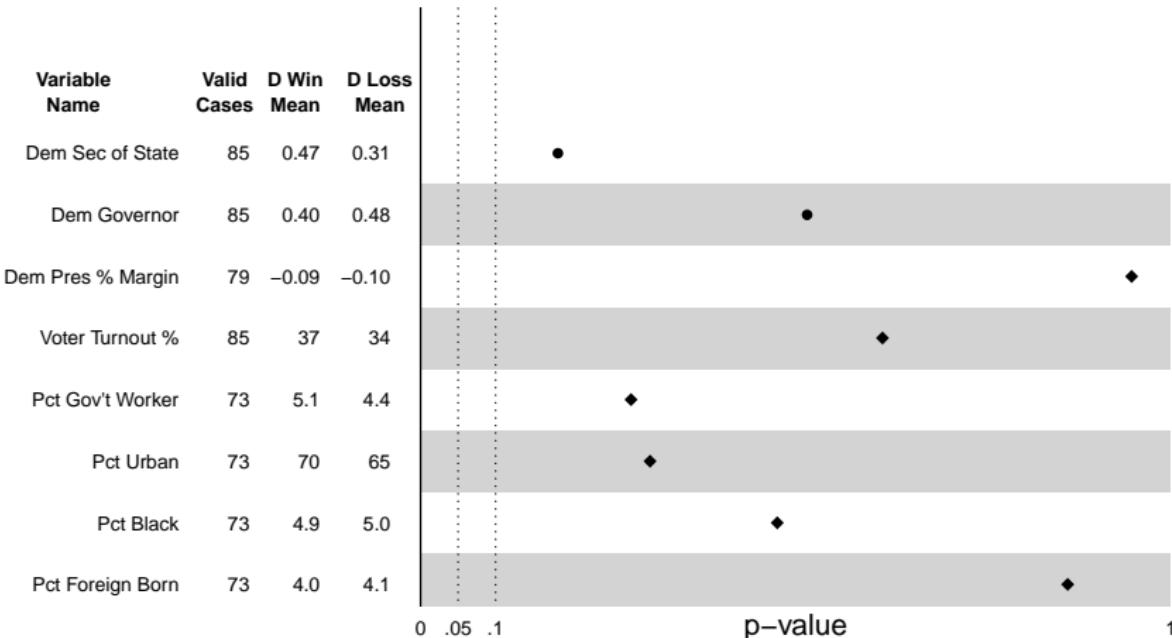




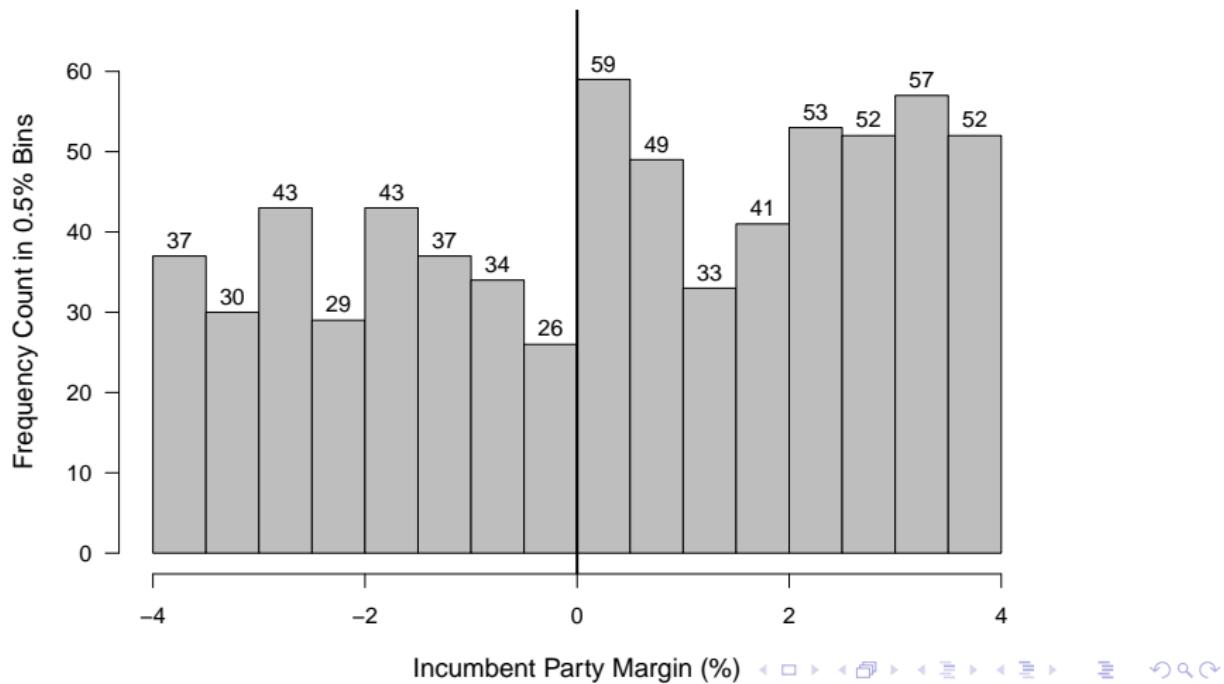
Covariate Balance



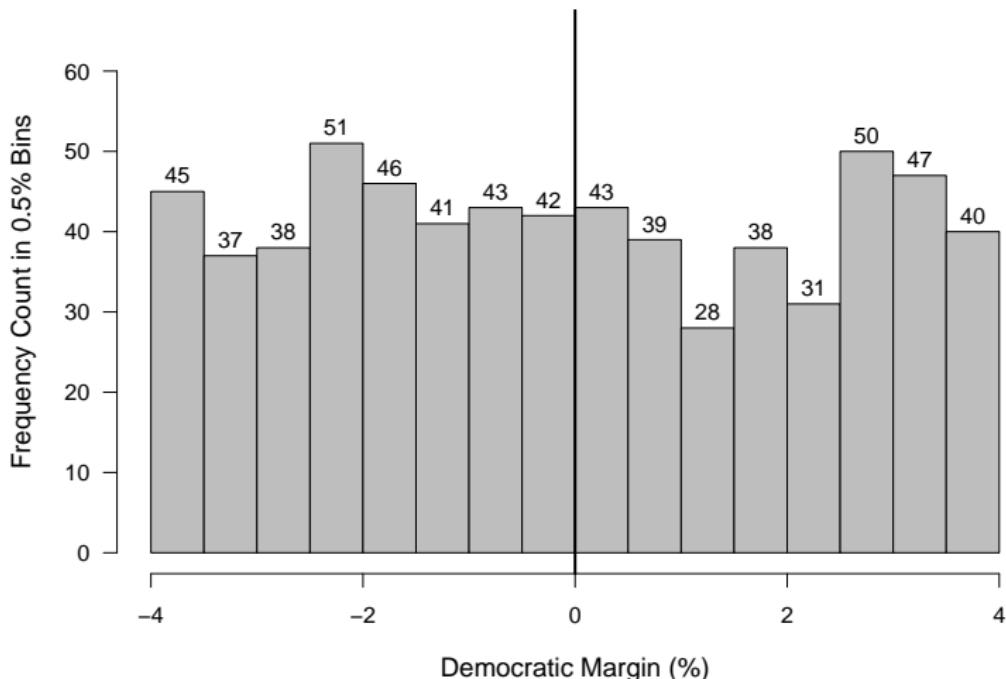
Covariate Balance



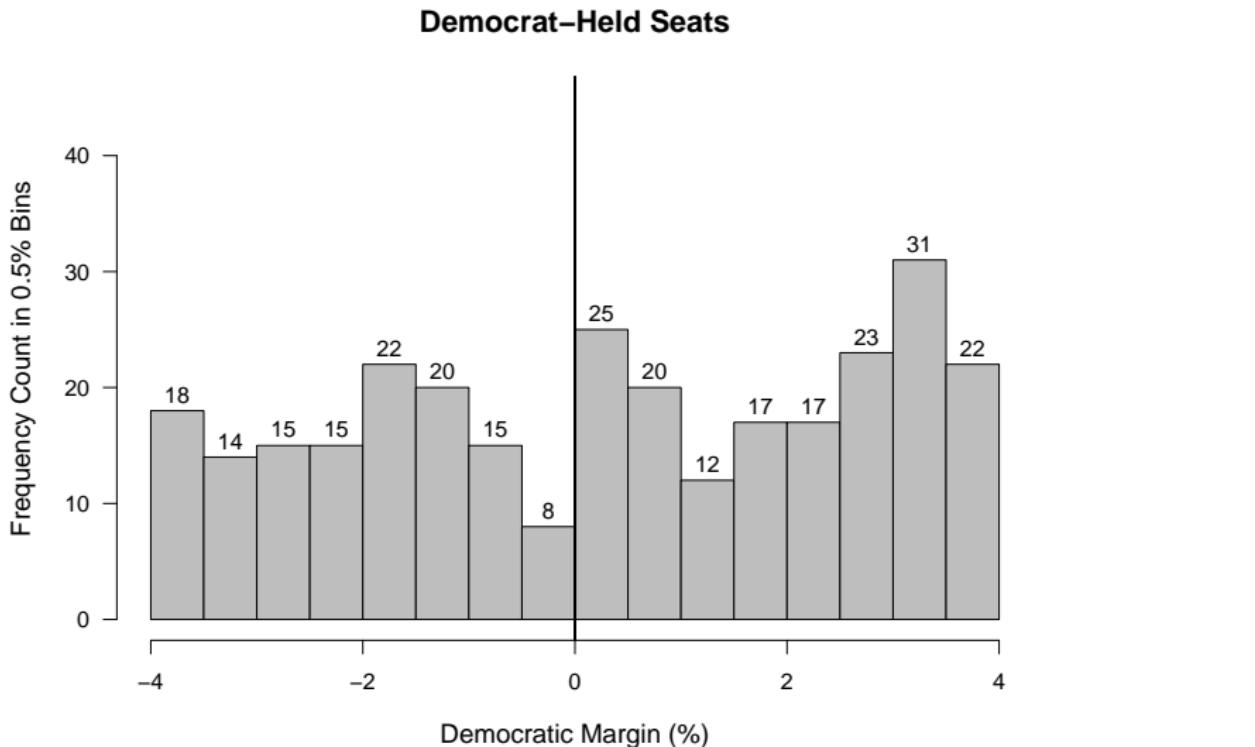
Incumbent Party's Margin in Close Elections



Democratic Margin in Close Elections

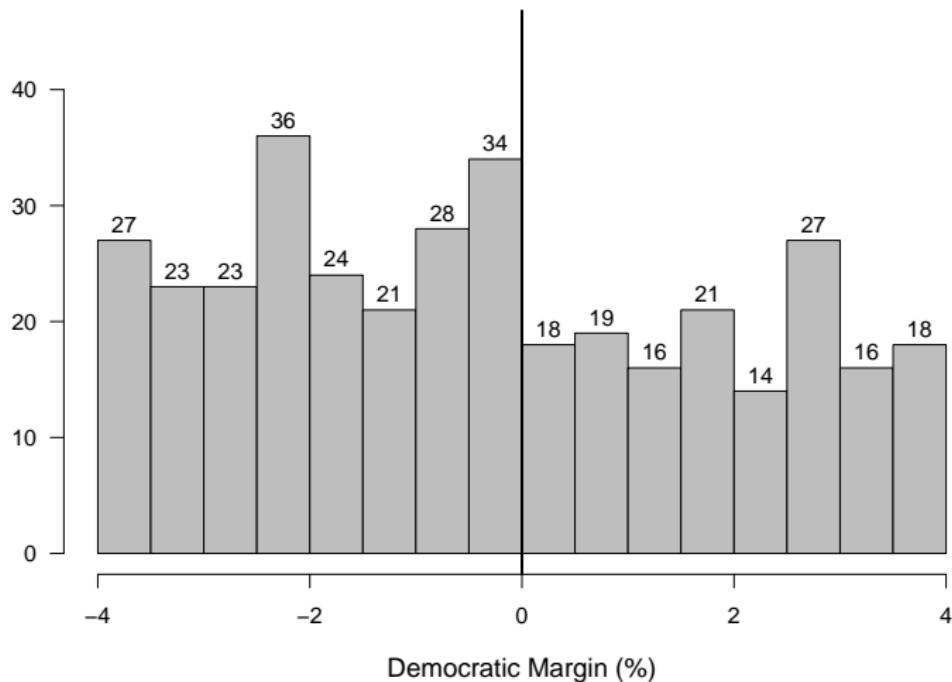


Broken Down by Incumbent Party



Broken Down by Incumbent Party

Republican–Held Seats



CQ Rating and Democratic Victory in Elections Decided by Less Than 0.5%

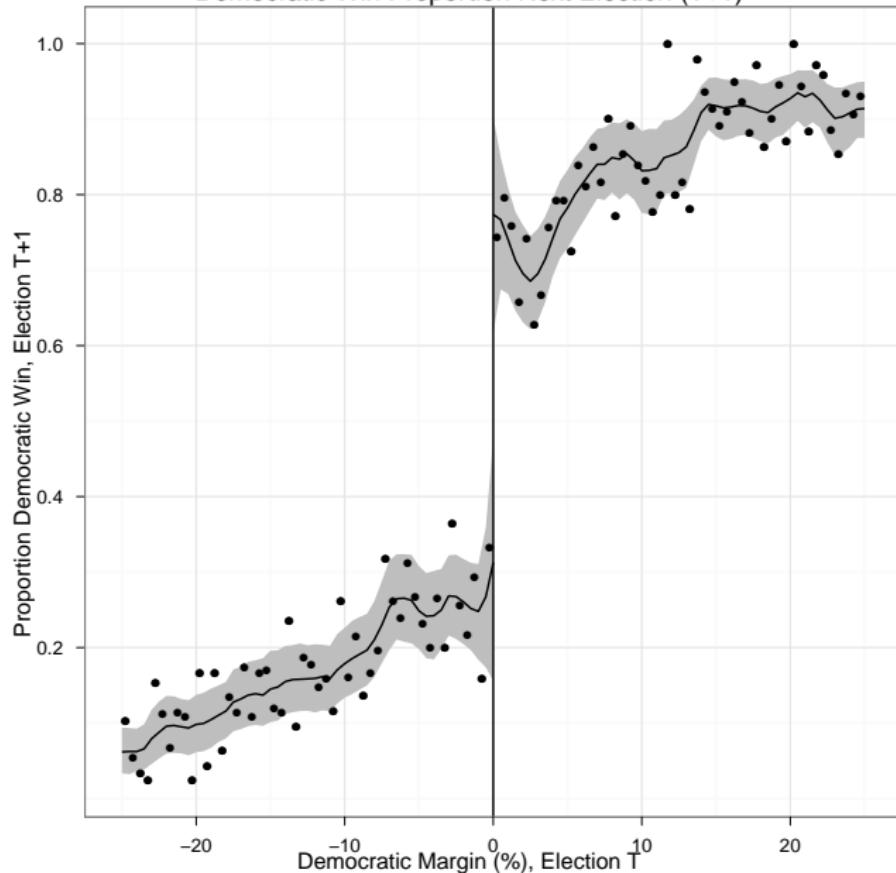
	Dem Win $t - 1$	Dem Win t
Rep Favored (23)	17%	30%
Tossup (25)	24%	52%
Dem Favored (21)	90%	71%

CQ Rating and Democratic Victory in Elections Decided by Less Than 0.5%

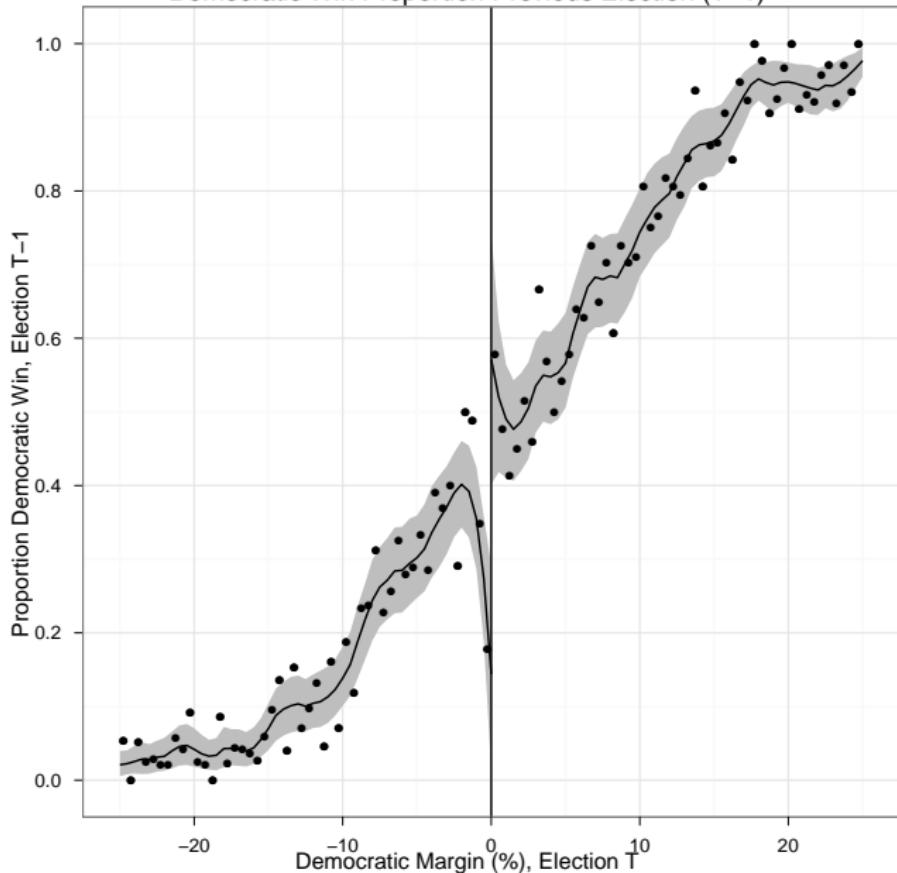
	Dem Win $t - 1$	Dem Win t
Rep Favored (23)	17%	30%
Tossup (25)	24%	52%
Dem Favored (21)	90%	71%

Even in tossup elections, incumbent party won about two-thirds of elections

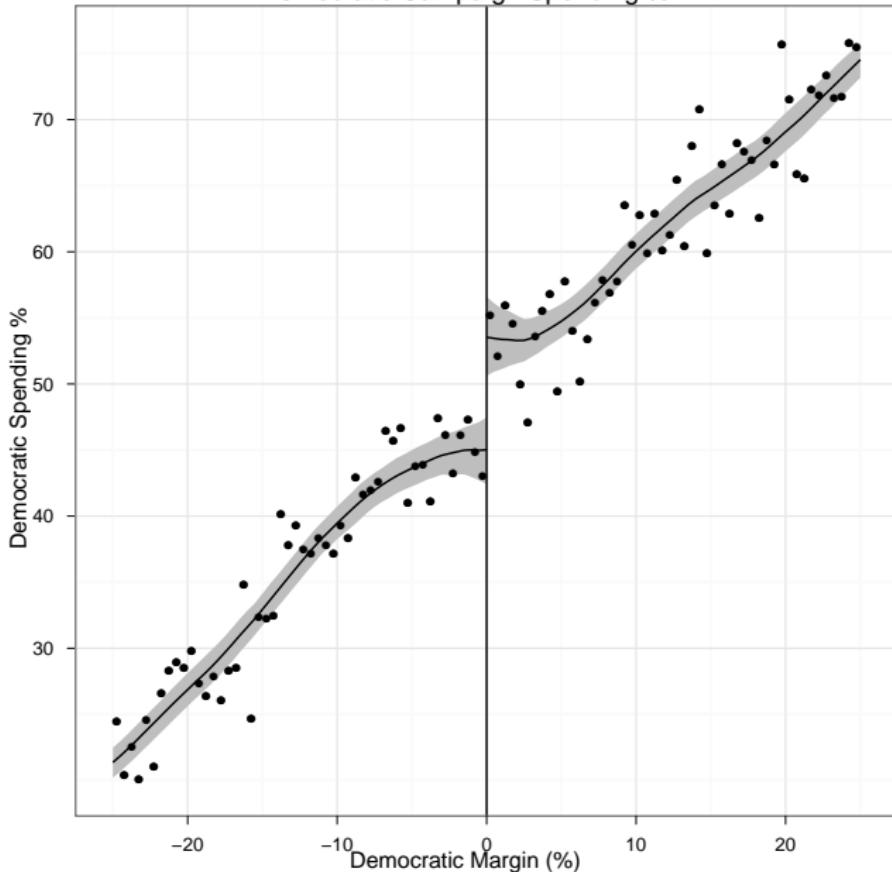
Democratic Win Proportion Next Election (T+1)



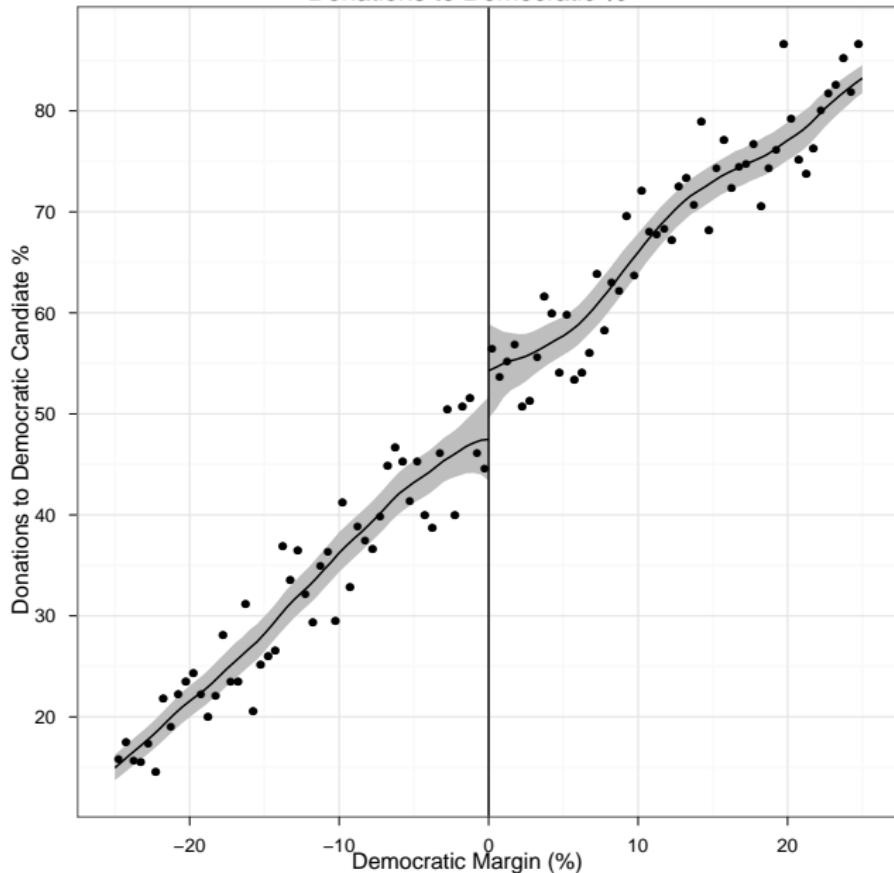
Democratic Win Proportion Previous Election (T-1)



Democratic Campaign Spending %



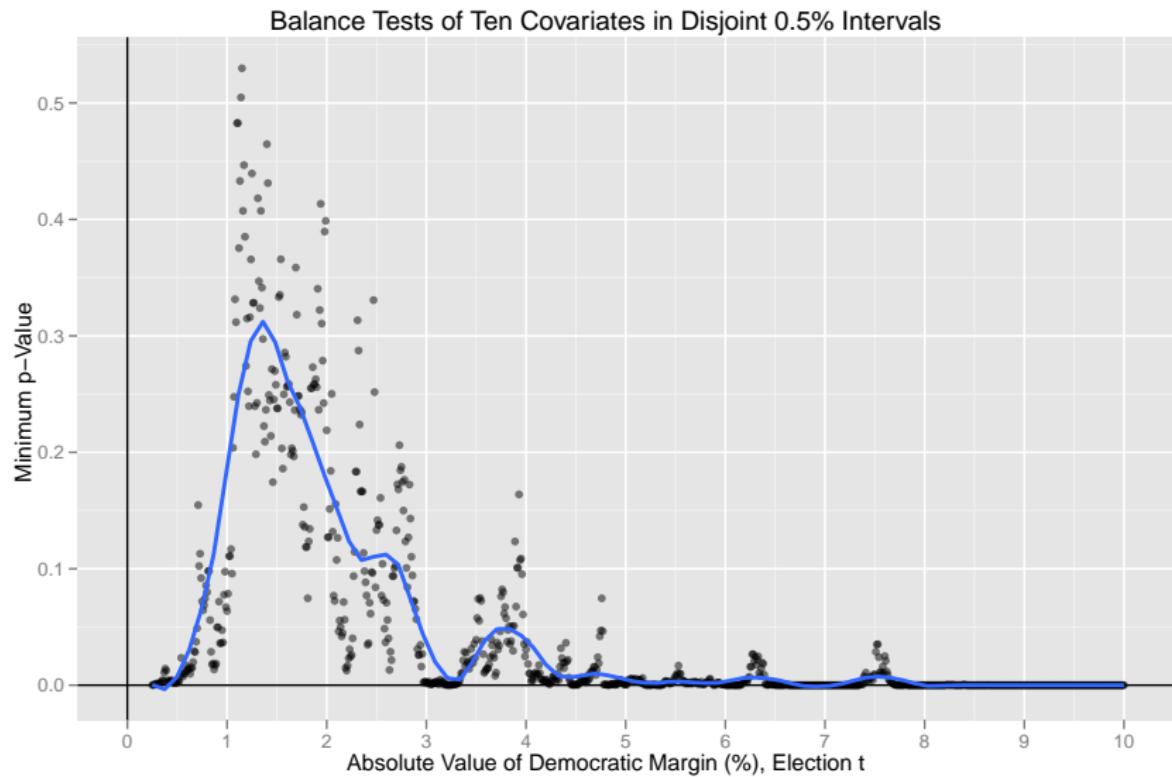
Donations to Democratic %



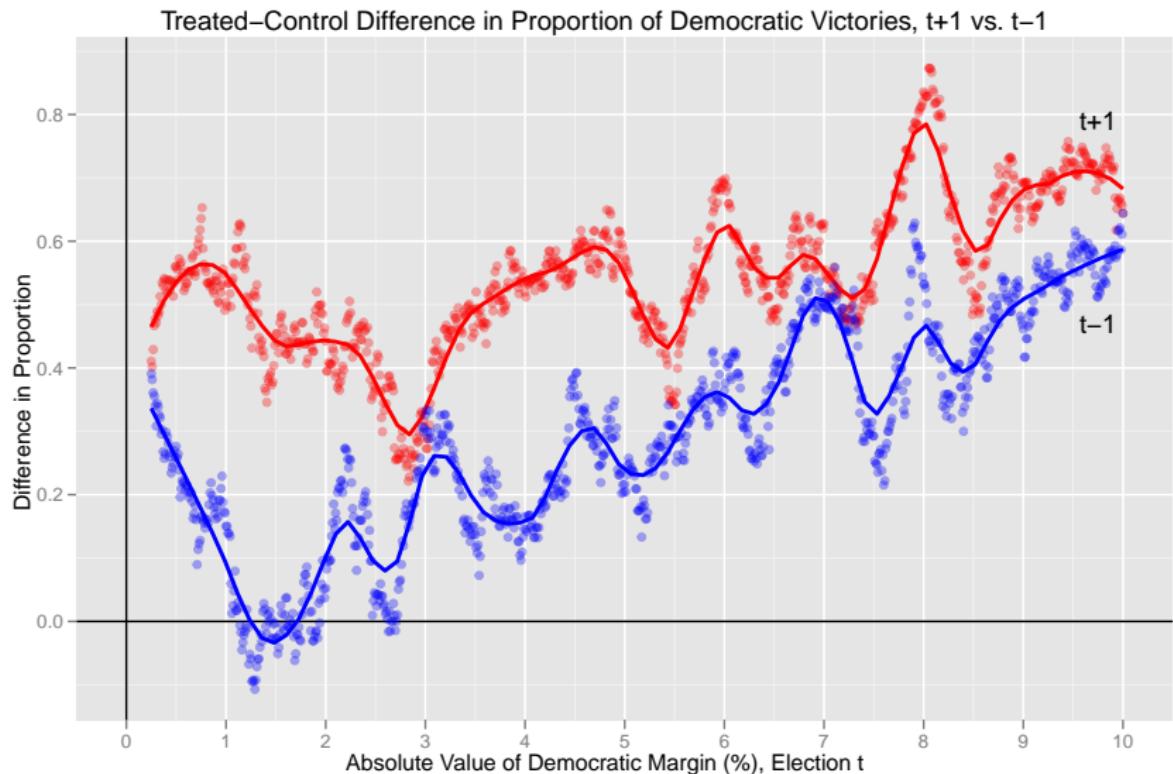
Divergence Figures

The following two figures plot the divergence of covariate distributions in the immediate neighborhood of the cut-point. Covariate differences are plotted against the mid-point of the disjoint interval tested (e.g., 1.5 for the interval $\{(-1.75, -1.25), (1.75, 1.25)\}$). Loess lines highlight the trends in the imbalance.

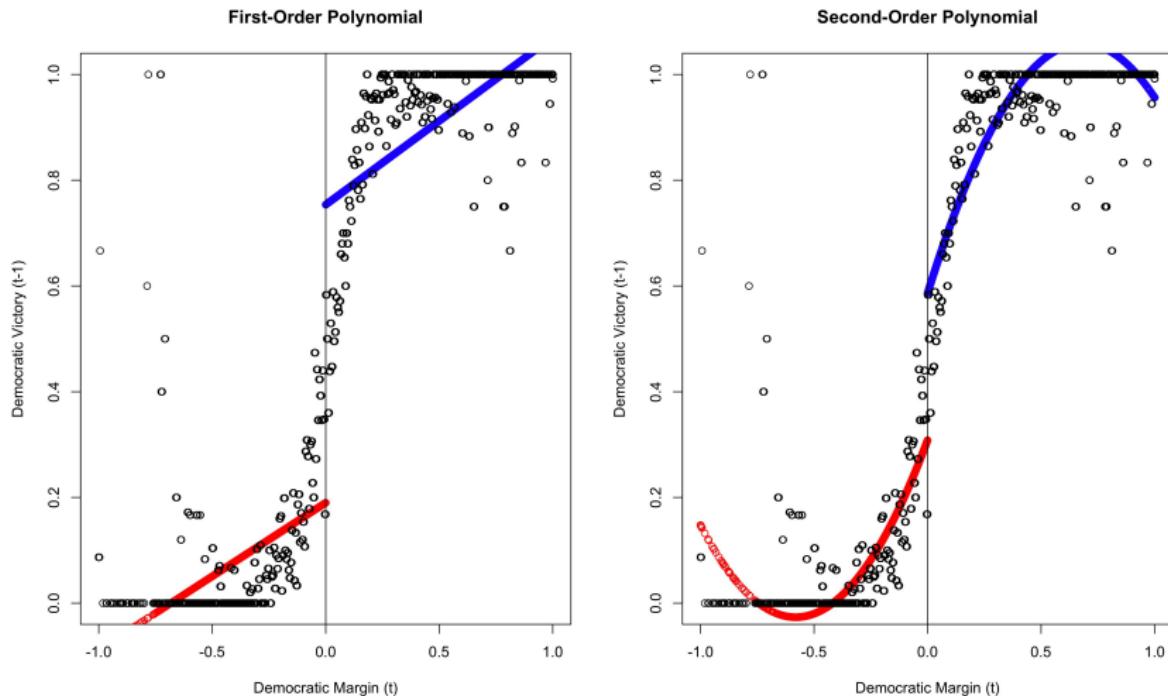
Divergence Figure 1



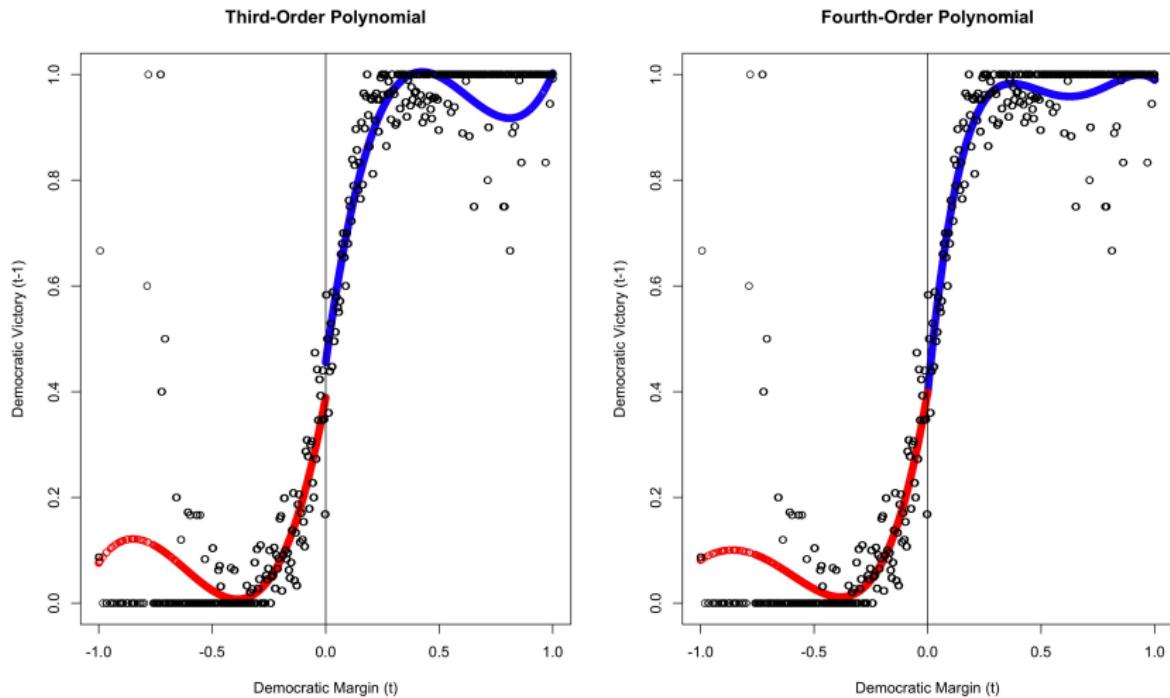
Divergence Figure 2



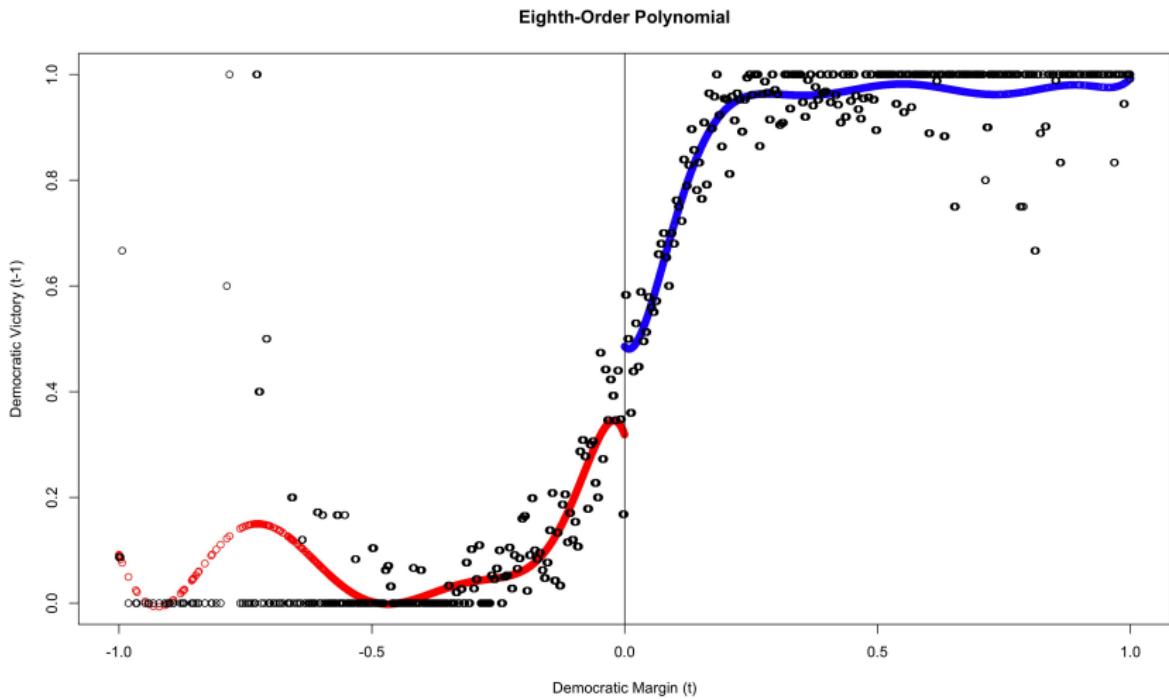
Lagged Democratic Win in RDD: 1st and 2nd-order polynomial



Lagged Democratic Win in RDD: 3rd and 4th-order polynomial



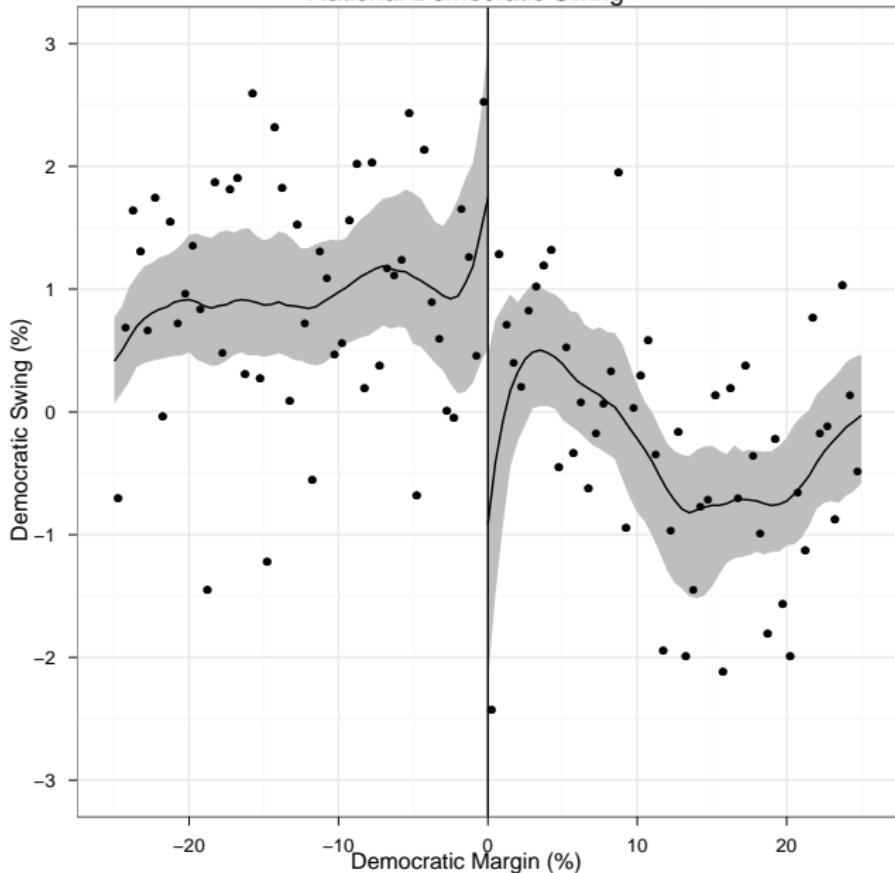
Lagged Democratic Win in RDD: 8th-order polynomial



National Partisan Swings

- Partisan swings are imbalanced:
 - 1958 (pro-Democratic tide): all 6 close elections occurred in Republican-held seats
 - 1994 (pro-Republican tide): all 5 close elections occurred in Democratic-held seats
 - Close elections do not generally occur in 50/50 districts

National Democratic Swing

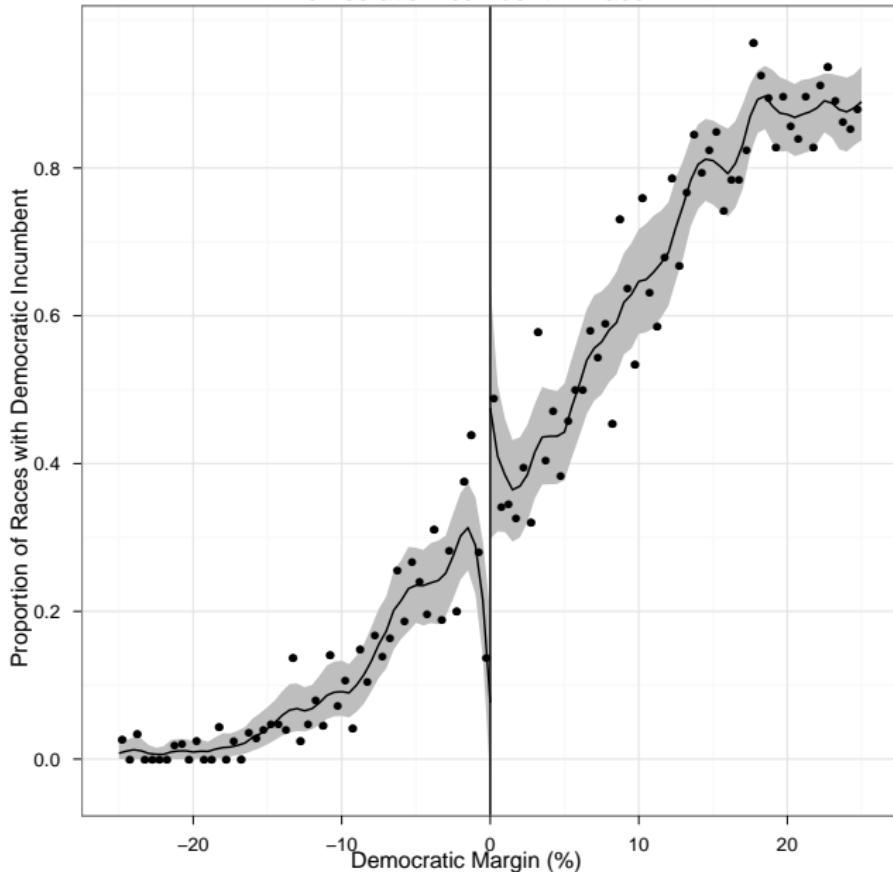


There is Strategic Exit

Strategic exit among incumbents

- 20% who barely win retire prior to next election
- but *all* candidates who barely won first election ran for reelection
@
- even outside of the RD window, previous election margin is predictive of retirements not due to running for higher office:
 - e.g., 1994: Equal percentage of Republicans and Democrats retired, but Republicans left to run for higher office (13 of 20) Democrats did not (7 of 27) @
 - 2006: 18 Republican open seats versus 9 Democratic open seats
 - 2010: In non-safe seats, 6 Republicans retired, but 15 Democrats retired

Democratic Incumbent in Race



Money and Incumbents in Close Elections

The imbalances in campaign resources occur because:

- When **incumbent candidates** run, they have more money: 62% of donations and 58% of expenditures

Money isn't predictive of winning, although they usually win

@

- In **open seats**, the incumbent party's candidates raise 50% of the money but spend 54%

Winners have more money:

winners raise 54% versus 41% ($p=0.01$)
spend 56% versus 49% ($p=0.02$)

Recounts?

New data on close races

- for random sample of half of all 130 elections in the 50% + / – 0.75 window
- detailed information from news coverage and administration records on each election

@

Recounts found in half of the elections (35 recounts)

- Imbalance on previous Democratic victory is greater in recounted elections: a treated-control difference of 0.56 versus 0.12
- In the 35 recounted elections, the first election recount winner was only changed **three** times @
- In each of the three cases, however, the incumbent representative was the ultimate winner

Benefits of RD

- Observable implications
- Still making weaker assumptions than model based estimators—e.g., Gelman-King

@

- Conjecture: RD will work better for elections with less professionalization and where cut point is less known
- Examples of smoothness with RD:
 - UK: Eggers and Hainmueller (2009)
 - Brazil: Hidalgo (2010)
 - California assembly elections
 - Canadian Parliamentary elections

Gelman–King Estimator

$$\mathbb{E}[V_{t+1}] = \beta_0 + \beta_1 P_t + \beta_2 (P_t \times R_{t+1}) + \beta_3 V_t,$$

where $V_t \in [0, 1]$ is the *Dem Share* in election t

$P_t \in \{-1, 1\}$ is the *Winning Party* in election t

and $R_{t+1} \in \{0, 1\}$ is a dummy variable indicating whether the *Incumbent Runs* in election $t + 1$

Modified Gelman–King Estimator

$$\begin{aligned}\mathbb{E}[V_{t+1}] = & \delta_0 + \delta_1 P_t + \delta_2 (P_t \times R_{t+1}) + \delta_3 M_t \\ & + \delta_4 M_t^2 + \delta_5 M_t^3 + \delta_6 M_t^4 \\ & + \delta_7 (M_t \times P_t) + \delta_8 (M_t^2 \times P_t) + \delta_9 (M_t^3 \times P_t) \\ & + \delta_{10} (M_t^4 \times P_t)\end{aligned}$$

Democratic Margin in election t , denoted M_t , is substituted in place of *Democratic Share* (V_t)

Lee's Model

Compare with Lee's specification, trivially modified so that $P_t \in \{-1, 1\}$ is substituted for *Democratic Victory* $\in \{0, 1\}$:

$$\begin{aligned}\mathbb{E}[V_{t+1}] = & \gamma_0 + \gamma_1 P_t + \gamma_2 M_t + \gamma_3 M_t^2 + \gamma_4 M_t^3 + \gamma_5 M_t^4 \\ & + \gamma_6 (M_t \times P_t) + \gamma_7 (M_t^2 \times P_t) + \gamma_8 (M_t^3 \times P_t) \\ & + \gamma_9 (M_t^4 \times P_t)\end{aligned}$$

Comments

- Models of campaign finance fail to accurately predict in the case they all agree on: e.g., Baron (1989), Synder (1989,1990), Erikson and Palfrey (2000)

@

- Why should elections be random?
e.g., close votes in legislatures

@

- Other designs may be better: redistricting (Sekhon and Titiunik 2012)

@

- Even when RD works, averaging over elections is delicate

Comments

- Similar results with open seats; but not much data
- Imbalance is so bad there is no support near the cutpoint
- Fuzzy RD will not work here: cannot observed random component separately from the endogenous one

Fraud?

- Only few cases of fraud described in secondary literature (e.g., Campbell 2005) also in RD window
- Given the imbalances of resources *a priori*, fraud isn't needed as an explanation

Other RD examples to read:

Make sure to read:

Eggers, Andy and Jens Hainmueller. 2009. "The Value of Political Power: Estimating Returns to Office in Post-War British Politics." *American Political Science Review* 104(4):513–533.

Mahalanobis Distance

- The most common method of multivariate matching is based on the Mahalanobis distance. The Mahalanobis distance measure between any two column vectors is defined as:

$$md(X_i, X_j) = \{(X_i - X_j)' S^{-1} (X_i - X_j)\}^{\frac{1}{2}}$$

where X_i and X_j are two different observations and S is the sample covariance matrix of X . @

- Mahalanobis distance is an appropriate distance measure if each covariate has an elliptic distribution whose shape is common between treatment and control groups (Mitchell and Krzanowski, 1985; Mitchell and Krzanowski, 1989).
- In finite samples, Mahalanobis distance will not be optimal.

Affine Invariance

- When can matching confounders make bias worse? e.g., what if the propensity score model is incorrect? @
- An affinely invariant matching method is a matching method which produces the same matches if the covariates X are affinely transformed. An affine transformation is any transformation that preserves collinearity (i.e., all points lying on a line initially still lie on a line after transformation) and ratios of distances (e.g., the midpoint of a line segment remains the midpoint after transformation).

Properties of Matching Algorithms

- All affinely invariant matching methods have the Equal Percent Bias Reduction (EPBR) property under some conditions. @
- If X are distributed with ellipsoidal distributions, then the EPBR property holds for affinely invariant matching methods (Rubin and Thomas, 1992). @
- There is an extension to a restricted class of mixtures (Rubin and Stuart, 2006): discriminant mixtures of proportional ellipsoidally symmetric distributions.

Equal Percent Bias Reduction (EPBR)

- Let Z be the expected value of X in the matched control group. Then we say that a matching procedure is EPBR if

$$E(X|T=1) - Z = \gamma \{E(X|T=1) - E(X|T=0)\}$$

for a scalar $0 \leq \gamma \leq 1$.

- We say that a matching method is EPBR for X because the percent reduction in the mean biases for each of the matching variables is the same. @
- In general, if a matching method is not EPBR, then the bias for some linear function of X is increased. @
- If the mapping between X and Y is not linear, EPBR is only of limited value.

When is EPBR a Good Property?

- In general, if a matching method is not EPBR, than the bias for a particular linear function of X is increased.
- We may not want EPBR if we have some specific knowledge that one covariate is more important than another. For example,

$$Y = \alpha T + X_1^4 + 2X_2,$$

where $X > 1$. In this case we should be generally more concerned with X_1 than X_2 .

- In finite samples, Mahalanobis distance and propensity score matching will not be optimal because X will not be ellipsoidally distributed in a finite sample even if that is its true distribution.

Can an Observational Study Recover the Experimental Benchmark?

- LaLonde (1986) examined a randomized job training experiment: National Supported Work Demonstration Program (NSW). @
- We know the experimental benchmark from the NSW. @
- LaLonde replaced the experimental controls with observational controls from the Current Population Survey @
- Could standard econometric techniques recover the experimental estimate? All failed (e.g., regression and IV).

The Controversy

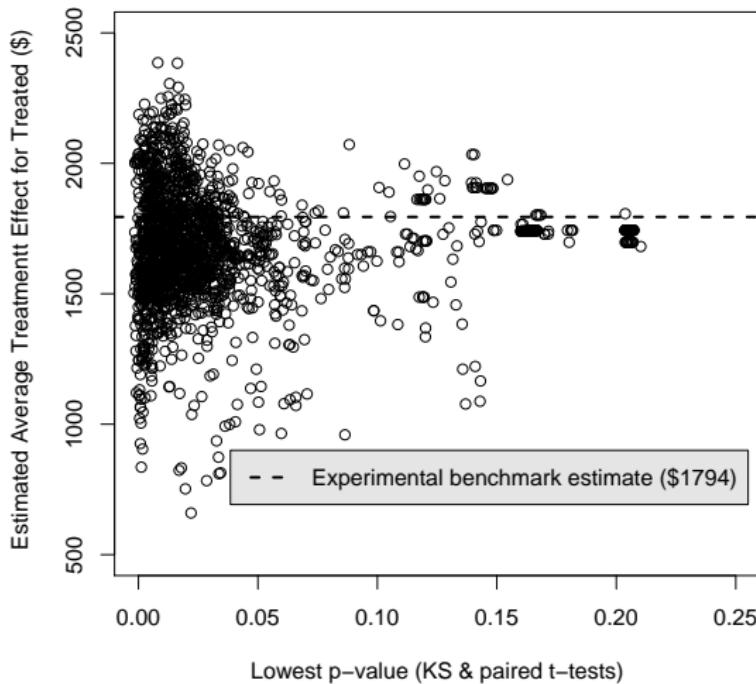
- Dehejia & Wahba used a subset for which 2-years pre-treatment earnings available, and claim propensity-score matching succeeds (1997; 1999). @
- Smith & Todd dispute Dehejia & Wahba's claims (2001). @
- The original question remains unresolved. Both sides have basically agreed to disagree: R. H. Dehejia and Wahba (2002), R. Dehejia (2005), J. Smith and P. Todd (2005a), and J. Smith and P. Todd (2005b).

Genetic Matching (GenMatch)

Genetic matching is a new general method for performing multivariate matching. GenMatch:

- maximizes the balance of observed potential confounders across matched treated and control units @
- uses an evolutionary search algorithm to determine the weight each covariate is given
- a genetic algorithm is used (Mebane and Sekhon, 2011)

Dehejia Wahba Sample



GenMatch

- Debate arises because existing matching methods fail to obtain reliable levels of balance in this dataset.
- GenMatch is able to reliably estimate the causal effects when other methods fail because it achieves substantially better balance.
- GenMatch in an hour produces vastly better balance than human researchers working away for ten years.

GenMatch

- No information about any outcome is used
- The method is nonparametric and does not depend on a propensity score (`pscore` can be included)
- Genetic matching can reduce the bias in X covariates in cases where conventional methods of matching **increase bias**. Who general is this?
- If selection on observables holds, but EPBR does not, has lower bias and MSE in the estimand in the usual simulations (will generally depend on loss function)

More General Method of Measuring Distance

- A more general way to measure distance is defined by:

$$d(X_i, X_j) = \left\{ (X_i - X_j)' (S^{-1/2})' W S^{-1/2} (X_i - X_j) \right\}^{1/2}$$

where W is a $k \times k$ positive definite weight matrix and $S^{1/2}$ is the Cholesky decomposition of S which is the variance-covariance matrix of X . @

- All elements of W are zero except down the main diagonal. The main diagonal consists of k parameters which must be chosen.
- This leaves the problem of choosing the free elements of W . For identification, there are only $k - 1$ free parameters.

Parameterization

- GenMatch uses the propensity score if it is known or if it can be estimated.
- The propensity score is estimated and its linear predictor, \hat{p} , is matched upon along with the covariates X once they have been adjusted so as to be uncorrelated with the linear predictor. @
- Combining is good because:
 - Propensity score matching is good at minimizing the discrepancy along the propensity score @
 - Mahalanobis distance is good at minimizing the distance between individual coordinates of X (orthogonal to the propensity score) (P. R. Rosenbaum and Rubin, 1985).

Optimization

- Many loss functions are possible. Such as:
 - minimize the largest discrepancy @
 - minimize the mean or median discrepancy @
 - minimize some other quantile @
 - restrict the above to only uniformly improving moves @
- Our recommended algorithm attempts to minimize the largest discrepancy at every step (minimizing the infinity norm). @
- For a given set of matches resulting from a given W , the loss is defined as the minimum p -value observed across a series of balance tests.

Measuring Balance

- There are many different ways of testing for balance and we cannot summarize the vast literature here. @
- The choice of balance will be domain specific. E.g., use randomization inference if you can as Bowers and Hansen (2006) do. @
- But the tests should be powerful and there should be many off them because different tests are sensitive to different departures from balance. @
- It is important the maximum discrepancy be small. p -values conventionally understood to signal balance (e.g., 0.10) are often too low to produce reliable estimates.

Balance Tests

- the p -values from these balance tests cannot be interpreted as true probabilities because of standard pre-test problems, but they remain useful measures of balance @
- By default, tests are conducted for all univariate baseline covariates, as well as their first-order interactions and quadratic terms. @
- The analyst may add tests of any function of X desired, including additional nonlinear functions and higher order interactions. @
- The tests conducted are t -tests for the difference of means and nonparametric bootstrap Kolmogorov-Smirnov distributional tests.

Genetic Optimization

- The optimization problem described above is difficult and irregular, and we utilize an evolutionary algorithm called GENOUD (Mebane and Sekhon, 2011) @
- Random search also works better than the usual matching methods, but is less efficient than GENOUD.

Monte Carlo Experiments

- Two Monte Carlos are presented.
- MC 1: the experimental conditions satisfy assumptions for EPBR
 - ① X covariates are distributed multivariate normal
 - ② propensity score is correctly specified
 - ③ linear mapping from X to Y
- MC 2: the assumptions required for EPBR are **not** satisfied:
 - ① X covariates are discrete and others are skewed and have point masses: they have the same distributions as the covariates of the LaLonde (1986) data.
 - ② propensity score is incorrectly specified
 - ③ mapping from X to Y is nonlinear

Experimental Condition 1: Multivariate Normal Distribution of Covariates

Estimator	Bias	RMSE	Bias Bias GM	MSE MSE GM
Raw	-604	.686	24.6	27.4
Mahalanobis (MH)	-8.63	.173	3.50	1.75
Pscore	-2.45	.210	.993	2.57
Pscore + MH	-5.96	.160	2.41	1.49
GenMatch	-2.47	.130		

Experimental Condition 2: Distribution of Lalonde Covariates

Estimator	Bias	RMSE	Bias Bias GM	MSE MSE GM
Raw	485	1611	19.0	18.2
Mahalanobis (MH)	-717	959	28.0	6.45
Pscore	512	1294	20.0	11.7
Pscore + MH	428	743	16.8	3.87
GenMatch	25.6	378		
@Random Search	112	493	4.38	1.30

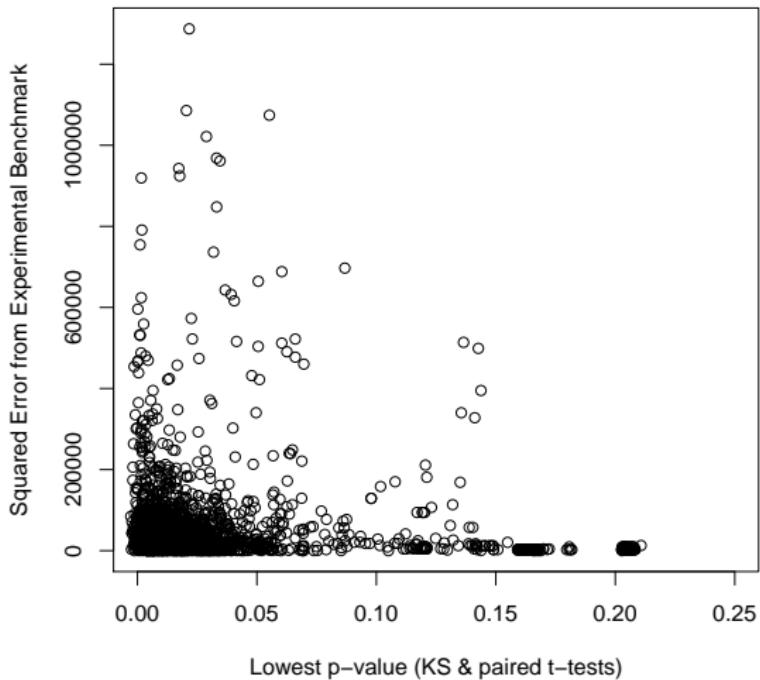
Summary of Monte Carlos

- Genetic matching reliably reduces both the bias and the MSE of the estimated causal effect even when conventional methods of matching **increase bias**.
- When the assumptions of the usual methods are satisfied, GenMatch still has lower MSE.
- **Caution 1:** selection on observables holds in both Monte Carlos
- **Caution 2:** loss function seemed to work. What about a different one?

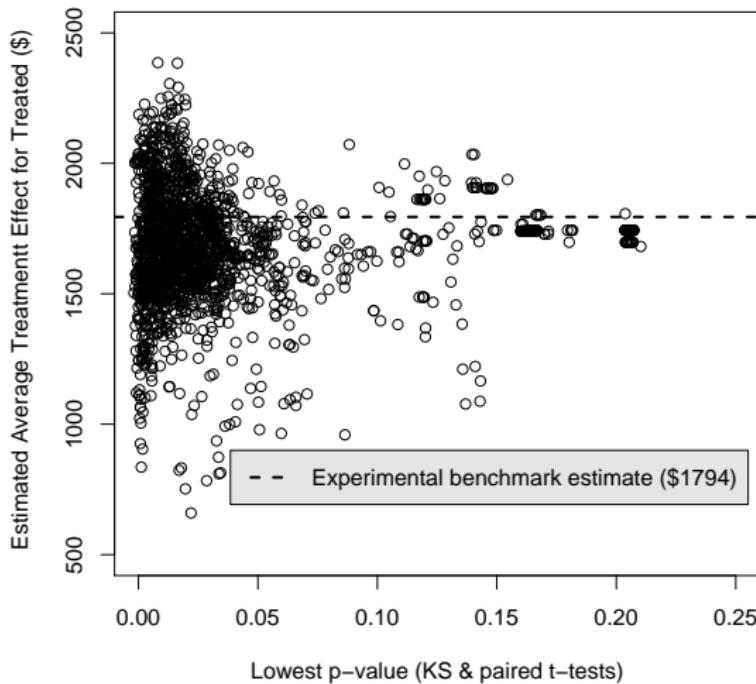
Summary of Monte Carlos

- Genetic matching reliably reduces both the bias and the MSE of the estimated causal effect even when conventional methods of matching **increase bias**.
- When the assumptions of the usual methods are satisfied, GenMatch still has lower MSE.
- **Caution 1:** selection on observables holds in both Monte Carlos
- **Caution 2:** loss function seemed to work. What about a different one?

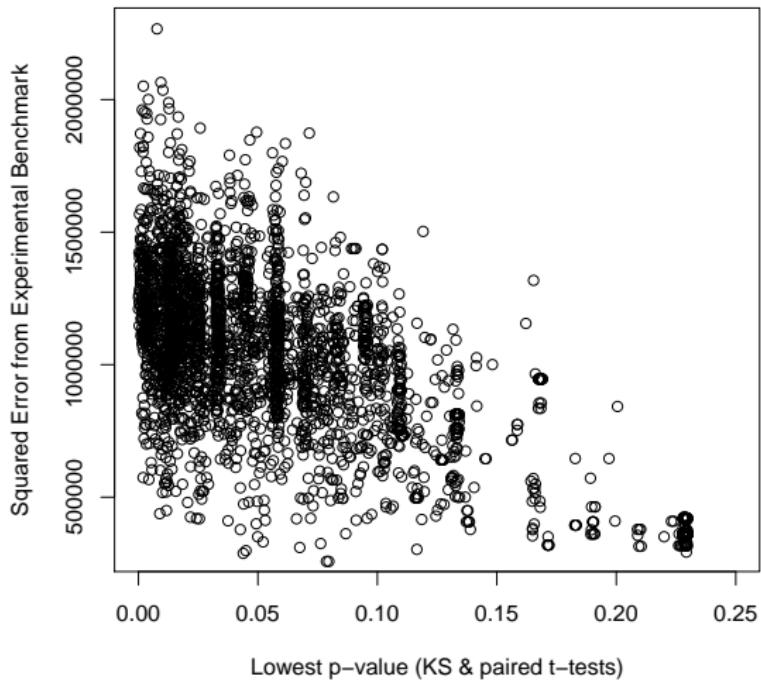
Dehejia Wahba Sample



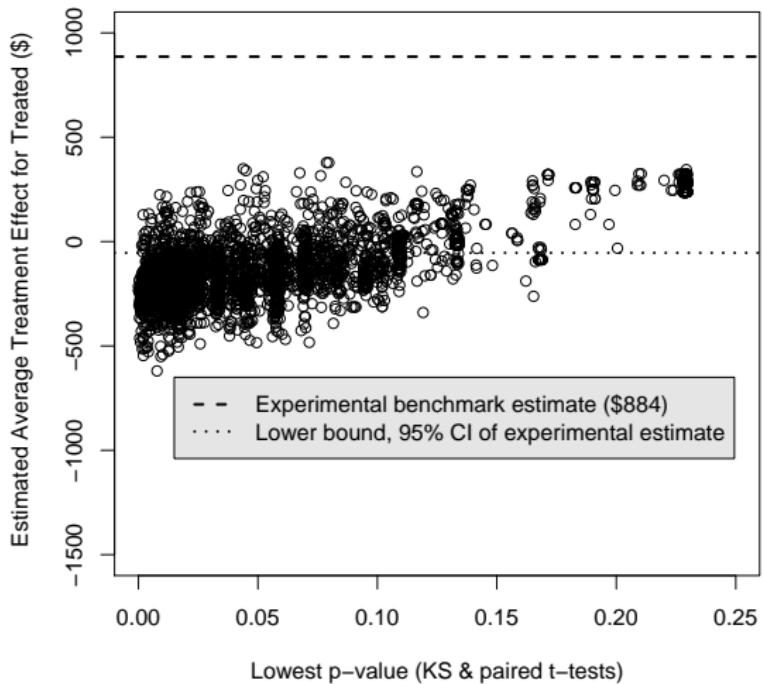
Dehejia Wahba Sample



Lalonde Sample



Lalonde Sample



Mahalanobis Distance

- Distributions of Xs are not ellipsoidal, so there's no reason to think MD will work @
 - Dehejia Wahba sample: balance is poor, but estimates are right on target!
 - LaLonde sample: balance exceeds conventional standards, but estimates are off the mark
- Lesson: Don't count on getting lucky. Instead, achieve best possible balance, and look at outcomes only once, after balance is attained.

GenMatch Summary

- GenMatch in an hour produces vastly better balance than human researchers working away for ten years.
- Machine learning can come to the rescue @
- But without the RCT, we would have a debate of how much to adjust
- Caution is warranted
- No amount of statistical modeling or algorithmic wizardry can resolve these questions

Opiates for the Matches

- Rubin: design trumps analysis
but design cannot be mass produced @
- Selection on observables assumed too readily:
 - no design
 - difficult questions about inference ignored
 - balance checks are rudimentary
 - no placebo tests @
- ties to Neyman-Rubin model weakening
- how well does it work in practice?
- statistical theory is of little guidance

Public Health and Medical Literature

- Use of matching has a longer history
- Many more experiments so a hope of calibration
- The interventions are actually done: another source of calibration @
- But resistance to comparing experiments with observational studies:
 - external validity
 - heterogeneous causal effect: casemix
 - private data (e.g., Hormone Replacement Therapy)

Non-Randomized Studies (NRS)

- Randomized Controlled Trials (RCTs) are used to evaluate clinical interventions.
- But RCTs are often unavailable and the only evidence is from NRS
- For cost effectiveness some prefer observational studies @
- Increasingly common to use propensity score matching, but:
 - selection on observables assumption may be false
 - parametric adjustment may be inadequate

Summary

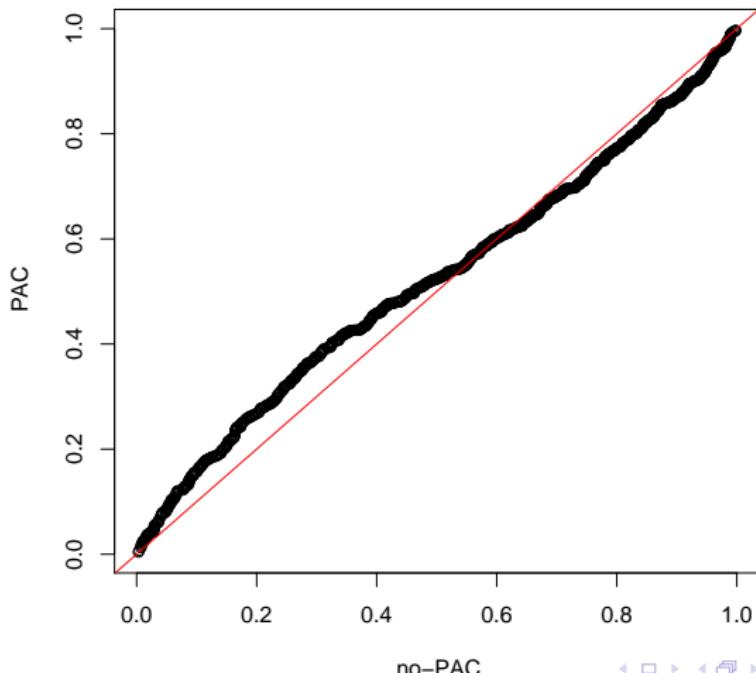
- Example: evaluation of Pulmonary Artery Catheterization (PAC) versus no-PAC. NRS found that PAC significantly increases mortality while the RCT found no difference in mortality. [Sekhon and Grieve 2008](#)
- [Genetic Matching](#) gives similar results to the RCT but using the NRS
- People failed to check balance after spending million of dollars collecting the data
- They didn't even plot the data

Pulmonary Artery Catheterization (PAC)

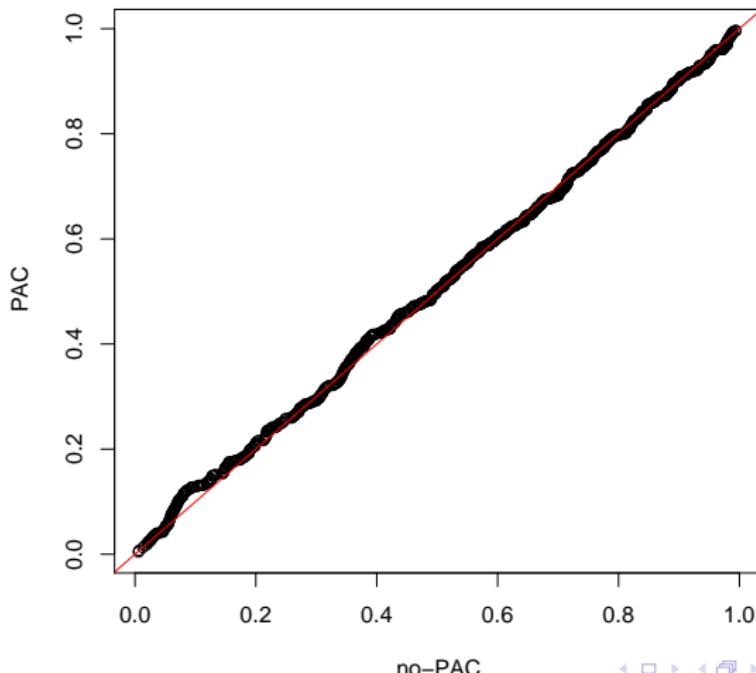
- PAC is an invasive cardiac monitoring device for critical ill patients (ICU)—e.g., myocardial infarction (ischaemic heart disease)
- Widely used for the past 30 years: spend \$2 billion in U.S. per year @
- RCT find no effect; **seven** NRS find that PAC increases mortality (e.g., Connors et al. JAMA 1996)
- RCT: causal effect really is zero
- Why do seven NRS disagree with RCTs?

- NRS all ICU admissions to 57 UK ICUs 2003-4
- 1052 cases with PAC; 32,499 controls (from a database of $N \approx 2M$) @
- Propensity score model: Age, reasons for admission, organ failure, probability of death, LOS, teaching hospital, size of unit @
- logistic regression, predict p (PAC) @
- match on propensity score @
- GenMatch: p score and the same underlying covariates multivariate matching algorithm, to max balance

Balance of baseline probability of Death, PSCORE



Balance of baseline probability of Death, GenMatch



Balance Measures

Variable	No PAC	PAC	T-test	KS P-value
IMProb				
Unmatched	30.8	56.2	0.00	0.00
Pscore	54.8		0.25	0.06
GenMatch	56.1		0.80	0.87
% admitted to hospital in London				
Unmatched	9.2	8.7	0.55	
Pscore	10.0		0.26	
GenMatch	8.7		1.00	

Balance Measures

Variable	No PAC	PAC	KS P-value
Age			
Unmatched	30.8	56.2	0.00
Pscore	54.8		0.06
GenMatch	56.1		0.87
% admitted for emergency surgery			
Unmatched	20.2	23.1	0.03
Pscore	21.3		0.32
GenMatch	23.7		0.66
% admitted to teaching hospital			
Unmatched	37.7	42.6	0.00
Pscore	44.5		0.38
GenMatch	42.6		1.00

Comparison of outcomes across methods PAC vs No PAC

Method	Odds Ratio	95% CI
RCT	1.13	(0.87 to 1.47)
GenMatch	1.10	(0.93 to 1.31)
@Pscore	1.22	(1.03 to 1.45)
Unmatched	3.51	(3.09 to 3.97)

Monte Carlo Setup

- Use data from PAC-Man RCT (Harvey et al. 2005)
 - Select treatment via a “true” propensity score
 - mapping from X to Y is nonlinear
 - Costs and QALYs generated assuming gamma and zero inflated Poisson distributions
 - Net benefits valued at £30,000 pounds per QALY
 - True INB=0
 - 1000 MCs
- For matching assume the propensity score is incorrect, but use the one used by Connors 1997.

Propensity Score Setup

Pscore based on previous study (Connors 1997):

$$\widehat{\text{pscore}} = \alpha + \alpha_1 \text{probmort} + \alpha_2 \text{misspmort} + \alpha_3 \text{emersurg} + \\ \alpha_4 \text{age} + \alpha_5 \text{elecsurg} + \alpha_6 \text{unit} + \alpha_7 \text{rate} + \alpha_8 \text{history} + \\ \alpha_9 \text{age} \times \text{probmort} + \alpha_{10} \text{age} \times \text{emersurg} + \\ \alpha_{11} \text{age} \times \text{elecsurg}$$

Using the same information what is relative performance of different matching methods? bias, RMSE compared to true INB

MC Results Incremental Net Benefit

Method	Bias	RMSE	$\frac{\text{Bias}}{\text{Bias GM}}$	$\frac{\text{RMSE}}{\text{RMSE GM}}$
GenMatch	-147	565		
Pscore	936	3039	6.4	5.4

True value: 0 pounds

Conclusion

- NICE (National Institute of Clinical Medicine) is coming to the U.S. in some form
- Cost effectiveness uses NRS more than FDA type regulators @
- The biases are profound, even when selection on observables holds
- Variance estimates profoundly underestimate the true uncertainty @
- Most extant objections are moral: I have a scientific concern
- What does this say about our field?

Conclusion

- Machine learning can help solve the easy problem of balance
- GenMatch achieves excellent covariate balance
- It gives similar results to the RCT
- But without the RCT, we would have a debate of how much to adjust
- Caution is warranted @
- Talking NICE into releasing observational data before parallel RCT results

The Problem

- When overcoming identification problems, we often turn to randomized controlled trials (RCTs)
- Often, though, RCTs are not a feasible, so we turn to observational studies, or non-random studies (NRSs)
- The problem, then, is how to combine this information in order to provide evidence for treatment effects in the full population of interest.
 - RCTs raise issues of Randomization Bias (Heckman and J. A. Smith, 1995): **poor external validity**
 - NRSs raise issues of Selection Bias, or non random assignment to treatment: **poor internal validity**

The Opportunity

- Explosion of data sources: administrative, electronic medical records (EMR), online behavior
- Population data is becoming more common and more precise
- How can it be used?
- Policy makers/firms:
“let’s just use the big data to make causal inference”
- Tension between identification vs. machine learning/prediction

The Problem

- Population effects cannot be estimated without assumptions:
 - Target population difficult to describe
 - Target population usually dynamic
 - General equilibrium effects as treatment is widely deployed (SUTVA violations)
- Our results can be used for the easier case of comparing two different experiments:
 - 1 Experiment 1: A vs. B and
 - 2 Experiment 2: A vs. C
 - 3 We wish to compare: B vs. C

Example

- A growing interest in the cost of health care
- Interest in system-wide issues—e.g., Dartmouth Atlas of Health Care
- Comparing the cost of UCLA versus Mayo Clinic is a causal question @
- Comparing the cost and health outcomes in the USA versus UK is also a causal question, but a less clear one
- We examine a simple case: conduct a cost effectiveness analysis (CEA) for one medical procedure

Estimands

- RCTs allow us to identify the Sample Average Treatment Effect (SATE), which is asymptotically equivalent to the Sample Average Treatment Effect on the Treated (SATT)
- We are often interested in the treatment effect for those who would receive treatment in practice, or the Population Average Treatment Effect on the Treated (PATT).

Our Method

We:

- develop a theoretical decomposition of the bias of going from SATT to PATT
- introduce a new method to combine RCTs and NRSs
 - Using Genetic Matching to maximize the internal validity
 - SATE → SATT
 - Using Maximum Entropy Weighting to maximize the external validity
 - SATT → PATT
- most importantly, provide placebo tests to validate the identifying assumptions.

Pulmonary Artery Catheterization (PAC)

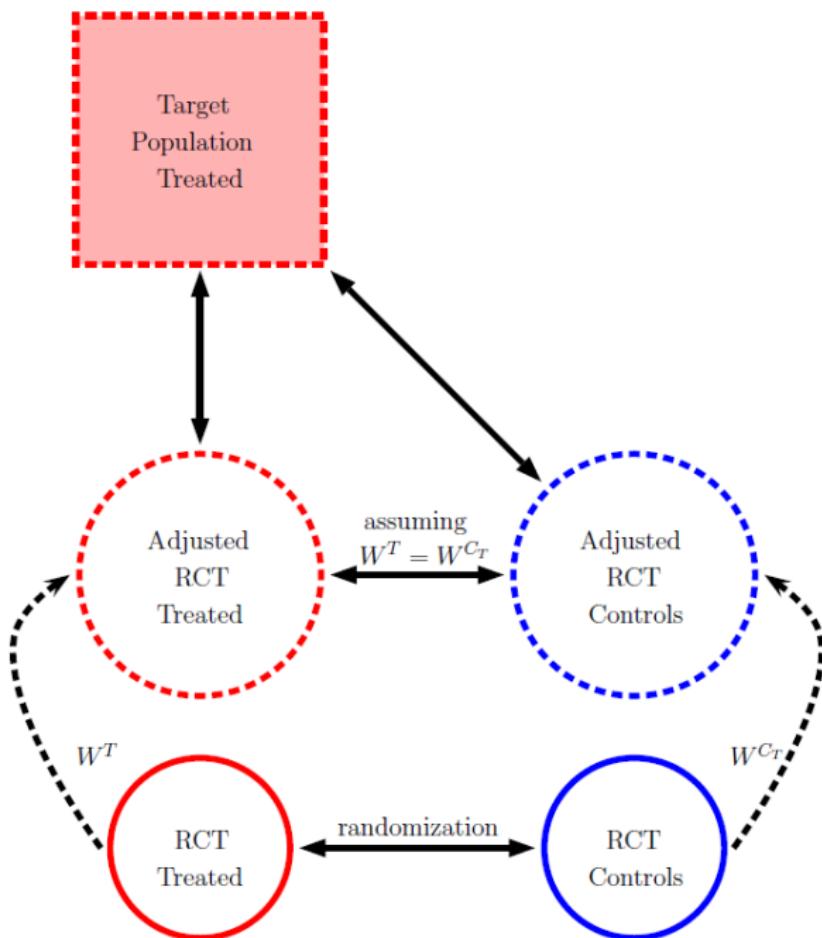
- PAC is an invasive cardiac monitoring device for critical ill patients (ICU)—e.g., myocardial infarction (ischaemic heart disease)
- Widely used for the past 30 years: spend \$2 billion in U.S. per year
- RCT find no effect; **seven** NRS find that PAC increases mortality (e.g., Connors et al. JAMA 1996)

Pulmonary Artery Catheterization

- RCT: a publicly funded, pragmatic experiment done in 65 UK ICUs in 2000-2004.
 - 1014 subjects, 506 who received PAC
 - No difference in hospital mortality ($p = 0.39$)
- NRS: all ICU admissions to 57 UK ICUs in 2003-2004
 - 1052 cases with PAC and 32,499 controls
 - One observational study was able to find no difference in hospital mortality ($p = 0.29$)
- However, the populations between the two studies differ, and we are interested in identifying PATT.

Schematic of SATE to PATT

In the next figure: double arrows indicate exchangeability of potential outcomes and dashed arrows indicate adjustment of the covariate distribution



We can think of the bias in our estimate of the PATT as a combination of bias due to a possible lack of internal validity of the estimate of SATT and bias due to poor external validity of the estimate from the RCT.

$$B = B_I + B_E$$

We can decompose of these biases in more detail as an extension of the Heckman, Ichimura, et al. (1998) decomposition of bias.

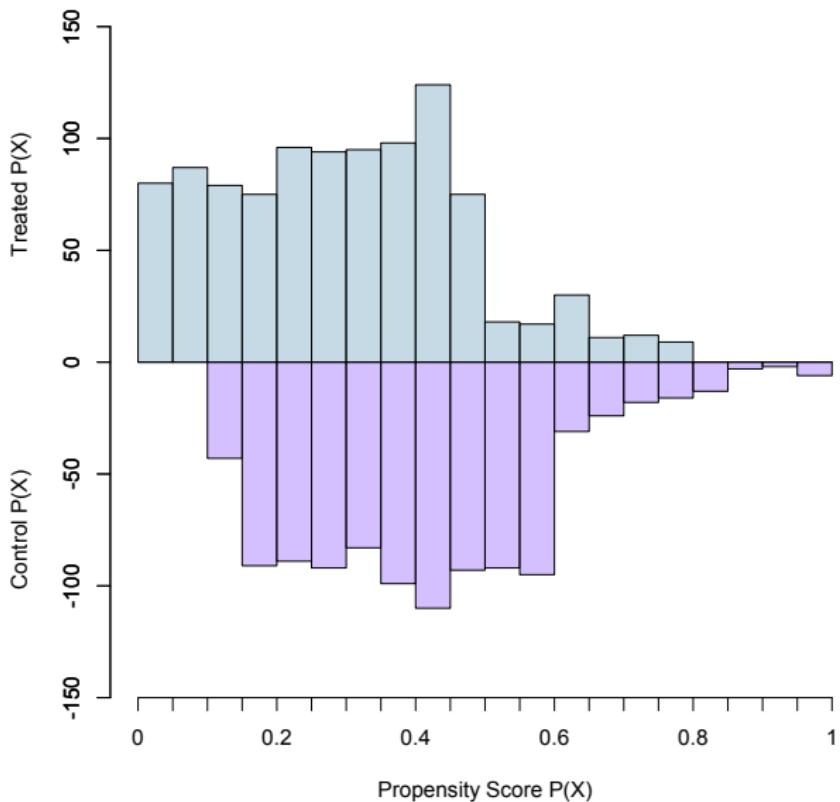
Some Definitions

- Let X denote a set of conditioning covariates in the sample population and W denote a set of conditioning covariates in the target population
- Let Y_1 and Y_0 denote the potential outcomes for a subject i
- Let $T \in (0, 1)$ be an indicator for whether or not subject i was in the treatment or control group
- Let $I \in (0, 1)$ be an indicator for whether or not subject i was in the sample population

$$\begin{aligned}
 B_I &= \mathbb{E}_{S_{I1} \setminus S_{I1}} \{ \mathbb{E}(Y_0 | X, T = 1, I = 1) \} \\
 &\quad - \mathbb{E}_{S_{I0} \setminus S_{I1}} \{ \mathbb{E}(Y_0 | X, T = 0, I = 1) \} \\
 &\quad + \mathbb{E}_{dF(T=1) - dF(T=0), S_{I1}} \{ \mathbb{E}(Y_0 | X, T = 0, I = 1) \} \\
 &\quad + \mathbb{E}_{dF(T=1), S_{I1}} \{ \mathbb{E}(Y_0 | X, T = 1, I = 1) - \mathbb{E}(Y_0 | X, T = 0, I = 1) \}
 \end{aligned}$$

The internal validity problem is due to bias in the RCT estimate of SATT.

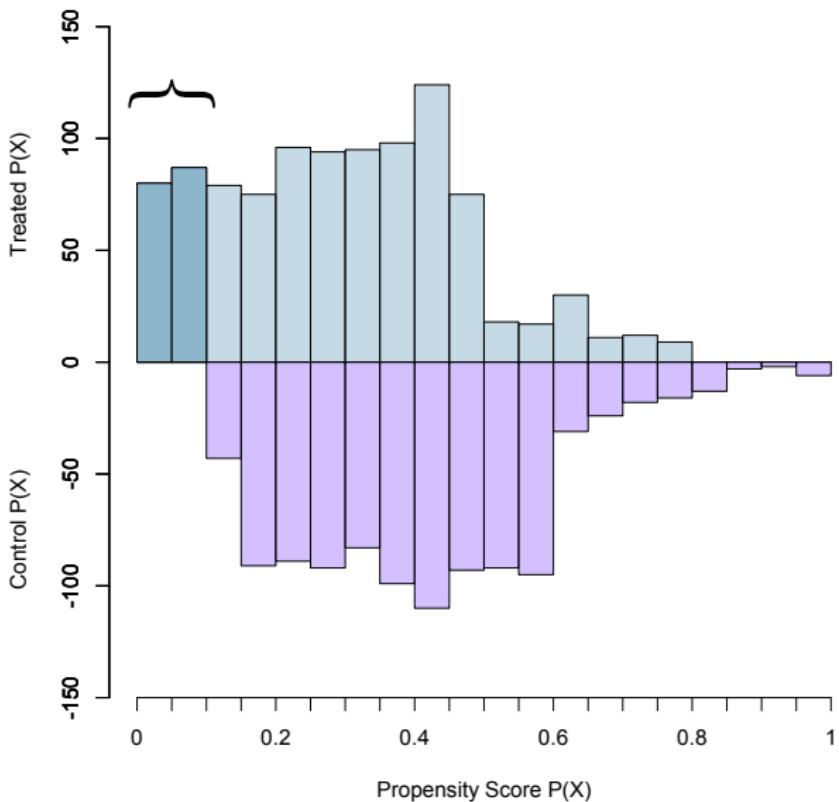
Example: Propensity Score Distributions



$$\begin{aligned}B_I &= \mathbb{E}_{S_{I1} \setminus S_{I0}} \{ \mathbb{E}(Y_0 | X, T = 1, I = 1) \} \\&\quad - \mathbb{E}_{S_{I0} \setminus S_{I1}} \{ \mathbb{E}(Y_0 | X, T = 0, I = 1) \} \\&\quad + \mathbb{E}_{dF(T=1) - dF(T=0), S_{I1}} \{ \mathbb{E}(Y_0 | X, T = 0, I = 1) \} \\&\quad + \mathbb{E}_{dF(T=1), S_{I1}} \{ \mathbb{E}(Y_0 | X, T = 1, I = 1) - \mathbb{E}(Y_0 | X, T = 0, I = 1) \}\end{aligned}$$

This is bias due to treated units who are not in the overlap region of the sample treated and sample controls.

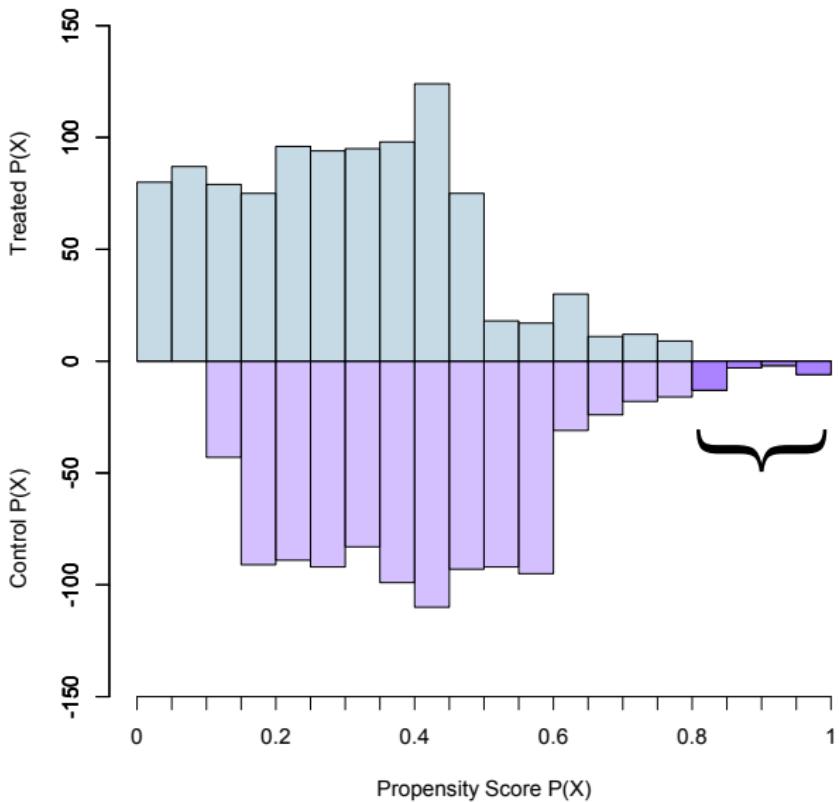
Example: Propensity Score Distributions



$$\begin{aligned}B_I &= \mathbb{E}_{S_{I1} \setminus S_{I0}} \{\mathbb{E}(Y_0|X, T=1, I=1)\} \\&\quad - \mathbb{E}_{S_{I0}} \{\mathbb{E}(Y_0|X, T=0, I=1)\} \\&\quad + \mathbb{E}_{dF(T=1) - dF(T=0), S_{I1}} \{\mathbb{E}(Y_0|X, T=0, I=1)\} \\&\quad + \mathbb{E}_{dF(T=1), S_{I1}} \{\mathbb{E}(Y_0|X, T=1, I=1) - \mathbb{E}(Y_0|X, T=0, I=1)\}\end{aligned}$$

This is bias due to controls units who are not in the overlap region of the sample treated and sample controls.

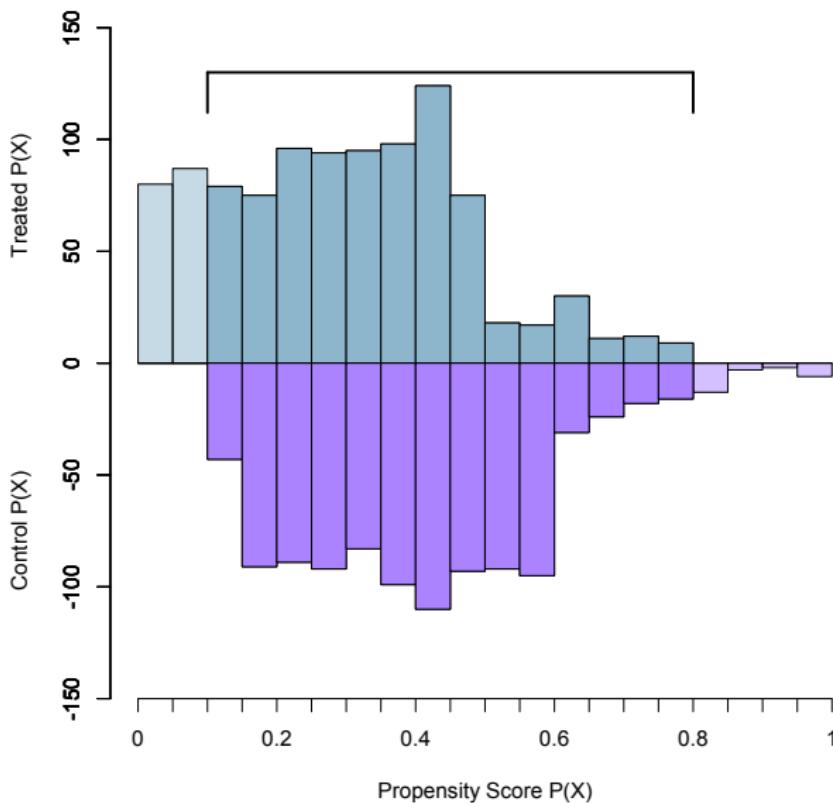
Example: Propensity Score Distributions



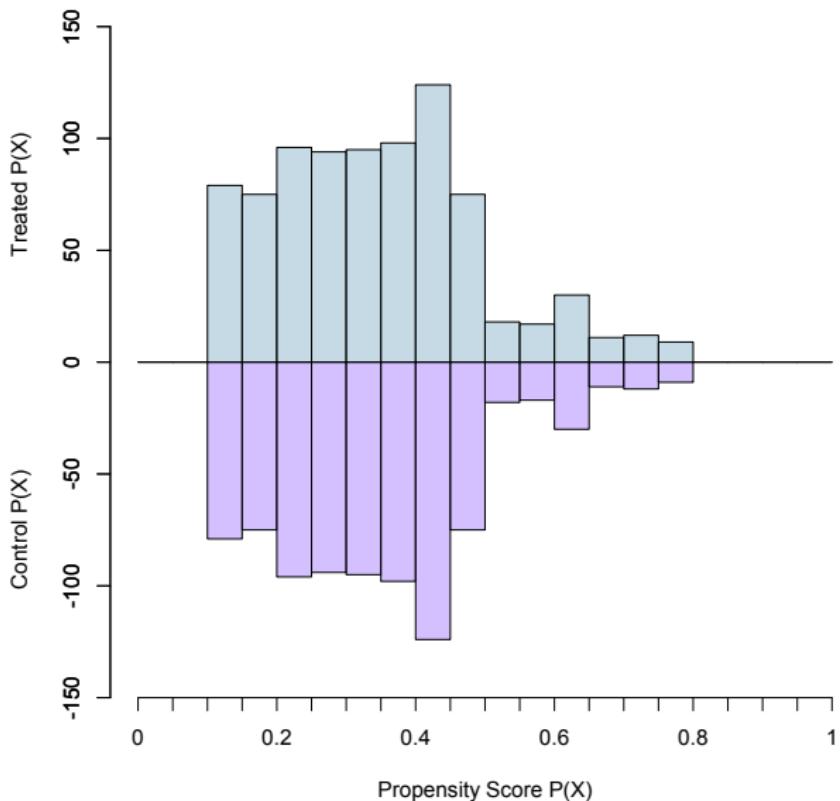
$$\begin{aligned}B_I &= \mathbb{E}_{S_{I1} \setminus S_{I1}} \{\mathbb{E}(Y_0|X, T=1, I=1)\} \\&\quad - \mathbb{E}_{S_{I0} \setminus S_{I1}} \{\mathbb{E}(Y_0|X, T=0, I=1)\} \\&\quad + \mathbb{E}_{dF(T=1) - dF(T=0), S_{I1}} \{\mathbb{E}(Y_0|X, T=0, I=1)\} \\&\quad + \mathbb{E}_{dF(T=1), S_{I1}} \{\mathbb{E}(Y_0|X, T=1, I=1) - \mathbb{E}(Y_0|X, T=0, I=1)\}\end{aligned}$$

This is bias due to imbalance of the sample treated and sample control units in the overlap region.

Example: Propensity Score Distributions



Example: Propensity Score Distributions



$$\begin{aligned}B_I &= \mathbb{E}_{S_{11} \setminus S_{I1}} \{\mathbb{E}(Y_0|X, T=1, I=1)\} \\&\quad - \mathbb{E}_{S_{10} \setminus S_{I1}} \{\mathbb{E}(Y_0|X, T=0, I=1)\} \\&\quad + \mathbb{E}_{dF(T=1) - dF(T=0), S_{I1}} \{\mathbb{E}(Y_0|X, T=0, I=1)\} \\&\quad + \mathbb{E}_{dF(T=1), S_{I1}} \{\mathbb{E}(Y_0|X, T=1, I=1) - \mathbb{E}(Y_0|X, T=0, I=1)\}\end{aligned}$$

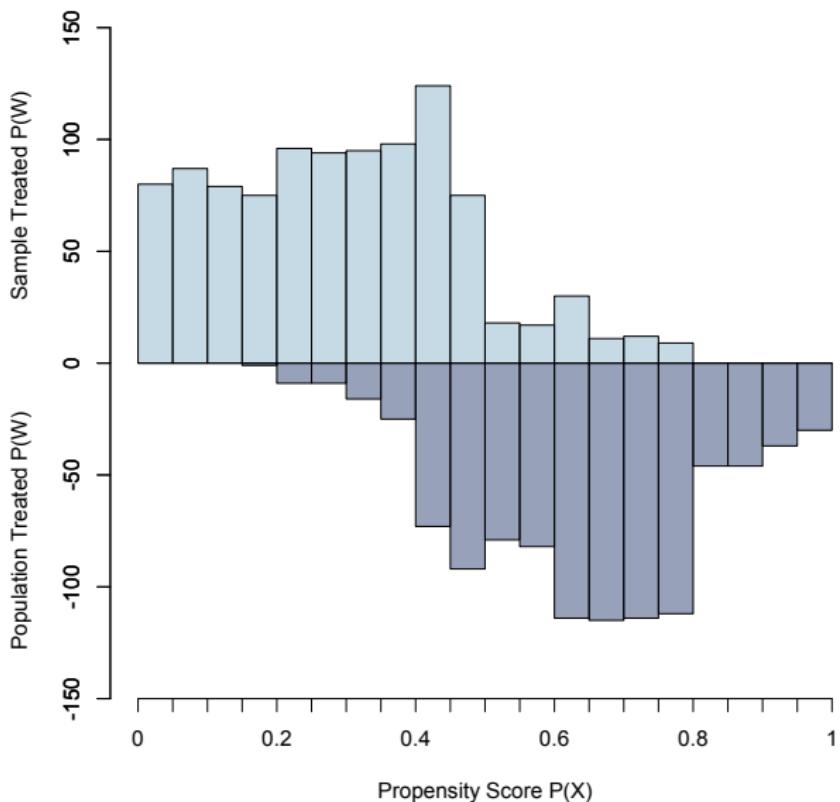
This is the usual definition of bias, or the selection issue. This should be minimized because we have an RCT, so there shouldn't be confounding of treatment.

$$\begin{aligned}
 B_E = & \mathbb{E}_{S_{11} \setminus S_{T1}} \{ \mathbb{E}(Y_i | W, T = 1, I = 1) \} \\
 & - \mathbb{E}_{S_{01} \setminus S_{T1}} \{ \mathbb{E}(Y_i | W, T = 1, I = 0) \} \\
 & + \mathbb{E}_{dF(I=1) - dF(I=0), S_{T1}} \{ \mathbb{E}(Y_i | W, T = 1, I = 0) \} \\
 & + \mathbb{E}_{dF(I=1), S_{T1}} \{ \mathbb{E}(Y_i | W, T = 1, I = 1) - \mathbb{E}(Y_i | W, T = 1, I = 0) \} \\
 & \text{for } i \in (0, 1)
 \end{aligned}$$

External validity problem:

- RCT is not a random sample of the population.
- Treatment can mean something different between the RCT and NRS.

Example: Propensity Score Distributions

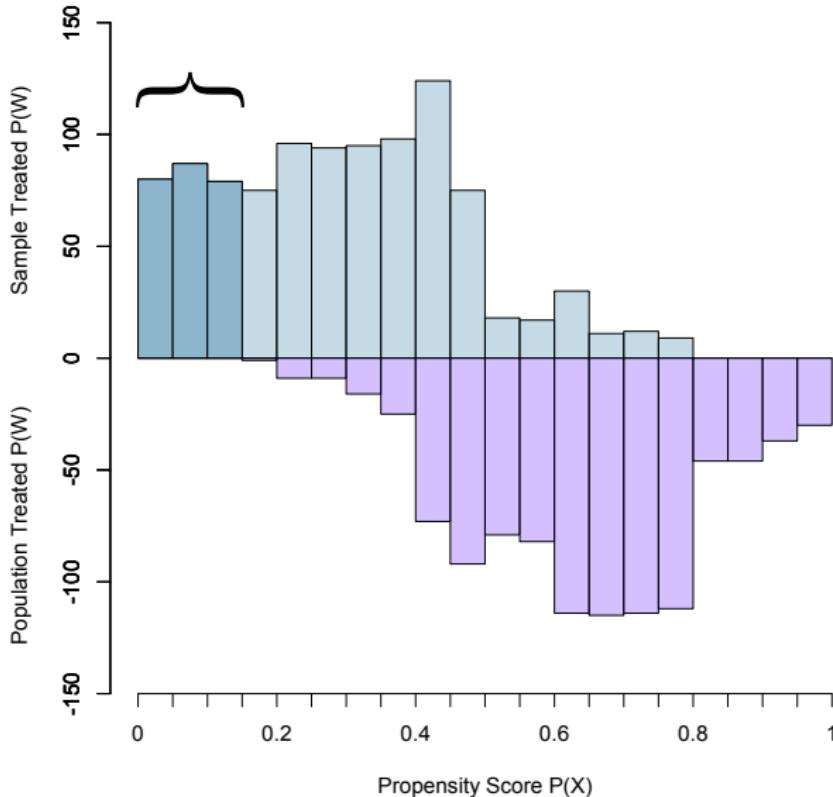


$$\begin{aligned}
 B_E = & \mathbb{E}_{S_{11} \setminus S_{T1}} \{ \mathbb{E}(Y_i | W, T=1, I=1) \} \\
 & - \mathbb{E}_{S_{01} \setminus S_{T1}} \{ \mathbb{E}(Y_i | W, T=1, I=0) \} \\
 & + \mathbb{E}_{dF(I=1) - dF(I=0), S_{T1}} \{ \mathbb{E}(Y_i | W, T=1, I=0) \} \\
 & + \mathbb{E}_{dF(I=1), S_{T1}} \{ \mathbb{E}(Y_i | W, T=1, I=1) - \mathbb{E}(Y_i | W, T=1, I=0) \}
 \end{aligned}$$

for $i \in (0, 1)$

This is bias due to sample treated units who are not in the overlap region of the sample treated and population treated.

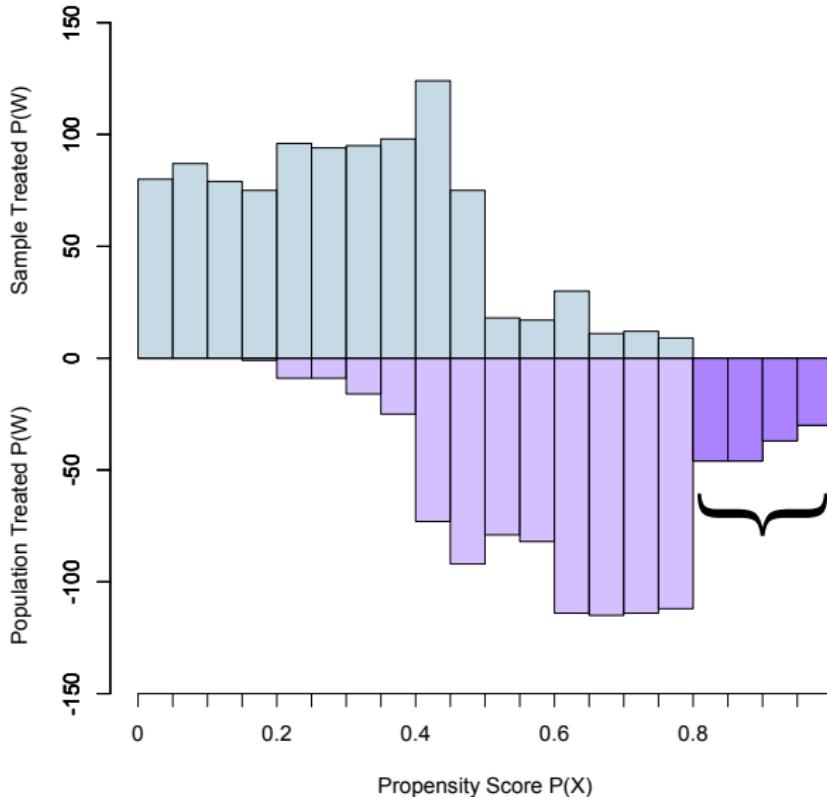
Example: Propensity Score Distributions



$$\begin{aligned}
 B_E = & \mathbb{E}_{S_{11} \setminus S_{T1}} \{ \mathbb{E}(Y_i | W, T=1, I=1) \} \\
 & - \mathbb{E}_{S_{01} \setminus S_{T1}} \{ \mathbb{E}(Y_i | W, T=1, I=0) \} \\
 & + \mathbb{E}_{dF(I=1) - dF(I=0), S_{T1}} \{ \mathbb{E}(Y_i | W, T=1, I=0) \} \\
 & + \mathbb{E}_{dF(I=1), S_{T1}} \{ \mathbb{E}(Y_i | W, T=1, I=1) - \mathbb{E}(Y_i | W, T=1, I=0) \} \\
 & \text{for } i \in (0, 1)
 \end{aligned}$$

This is bias due to population treated units who are not in the overlap region of the sample treated and population treated.

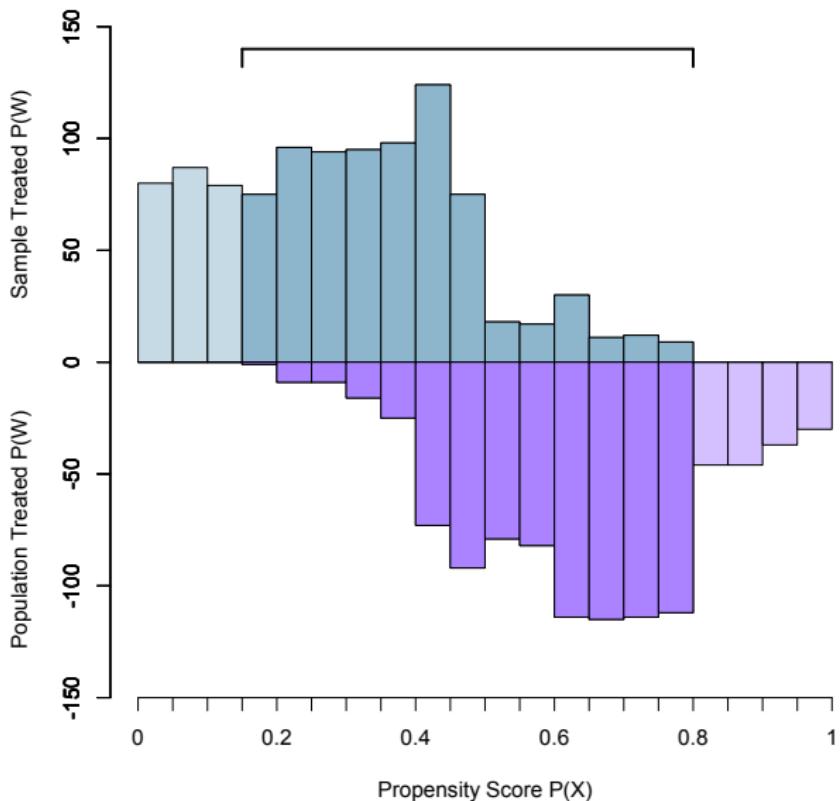
Example: Propensity Score Distributions



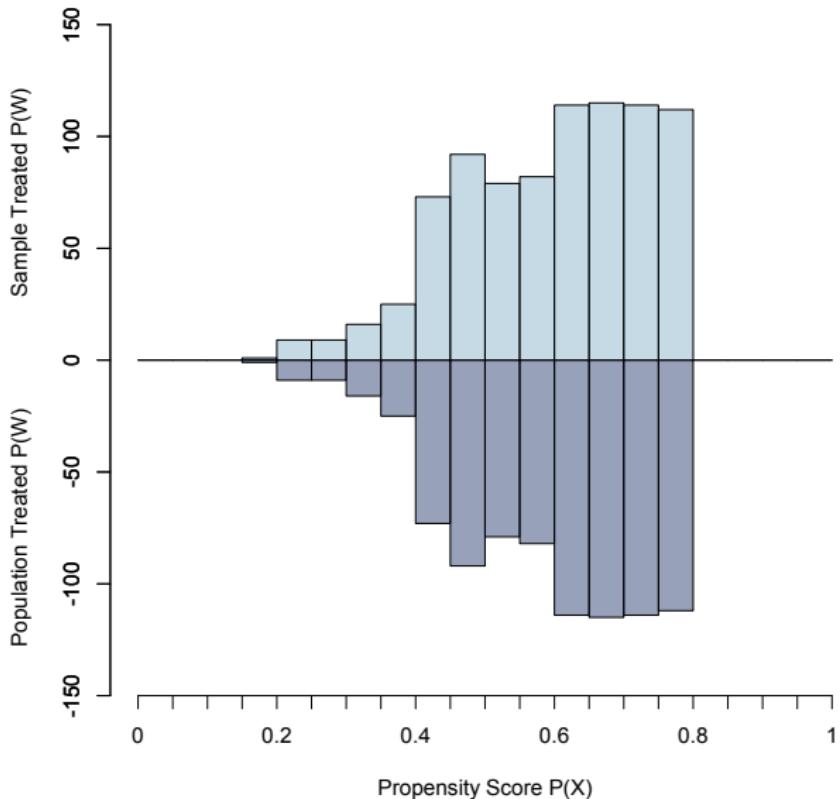
$$\begin{aligned}
 B_E = & \mathbb{E}_{S_{11} \setminus S_{T1}} \{ \mathbb{E}(Y_i | W, T=1, I=1) \} \\
 & - \mathbb{E}_{S_{01} \setminus S_{T1}} \{ \mathbb{E}(Y_i | W, T=1, I=0) \} \\
 & + \mathbb{E}_{dF(I=1) - dF(I=0), S_{T1}} \{ \mathbb{E}(Y_i | W, T=1, I=0) \} \\
 & + \mathbb{E}_{dF(I=1), S_{T1}} \{ \mathbb{E}(Y_i | W, T=1, I=1) - \mathbb{E}(Y_i | W, T=1, I=0) \} \\
 & \text{for } i \in (0, 1)
 \end{aligned}$$

This is bias due to imbalance of the sample treated and population treated units in the overlap region.

Example: Propensity Score Distributions



Example: Propensity Score Distributions



$$\begin{aligned}
 B_E = & \mathbb{E}_{S_{11} \setminus S_{T1}} \{ \mathbb{E}(Y_i | W, T=1, I=1) \} \\
 & - \mathbb{E}_{S_{01} \setminus S_{T1}} \{ \mathbb{E}(Y_i | W, T=1, I=0) \} \\
 & + \mathbb{E}_{dF(I=1) - dF(I=0), S_{T1}} \{ \mathbb{E}(Y_i | W, T=1, I=0) \} \\
 & + \mathbb{E}_{dF(I=1), S_{T1}} \{ \mathbb{E}(Y_i | W, T=1, I=1) - \mathbb{E}(Y_i | W, T=1, I=0) \} \\
 & \text{for } i \in (0, 1)
 \end{aligned}$$

This is the usual definition of bias, or the sample selection bias. There is nothing we can do to fix this, and the bias can be of opposite sign and any magnitude.

The Method

Let θ_s be a stratified treatment effect from the randomized trial, θ_{ws} be the treatment effect of reweighted strata, and θ be the true population treatment effect.

$$\theta_s \rightarrow \theta_{ws} \rightarrow \theta$$

- Estimate θ_s using Genetic Matching
- Reweight the strata using Maximum Entropy weighting to estimate θ_{ws}
- Run a placebo test to validate the identifying assumptions and provide evidence for how close θ_{ws} is to θ .

Some Definitions

- Let W denote a set of conditioning covariates, with the distribution of the population treated observation
- Let $Y_{s,t}$ denote the potential outcomes for a given subject in sample s and treatment t
- Let $T \in (0, 1)$ be an indicator for whether or not subject i was in the treatment ($T = 1$) or control ($T = 0$) group
- Let $S \in (0, 1)$ be an indicator for whether or not subject i was in the RCT ($S = 1$) or target population ($S = 0$)

Model

- Assume that both the RCT and the target population data are simple random samples from two different infinite populations
- Unlike the paper, we assume here that $W = W^t = W^c$
- The expectation $\mathbb{E}_{01}\{\cdot\}$ is a weighted mean of the W specific means, $\mathbb{E}(Y_{s1}|W, S = 1, T = 1)$, with weights according to the distribution of W in the treated target population, $Pr(W|S = 0, T = 1)$
- SATE is defined as:

$$\tau_{SATE} = \mathbb{E}(Y_{11} - Y_{10}|S = 1),$$

where the expectation is over the (random) units in $S = 1$ (the RCT)

Assumptions

A.1 Treatment is consistent across studies:

$$Y_{i01} = Y_{i11} \text{ and } Y_{i00} = Y_{i10}$$

A.2 Strong Ignorability of Sample Assignment for Treated:

$$(Y_{i01}, Y_{i11}) \perp\!\!\!\perp S_i | (W_i, T_i = 1) \quad 0 < Pr(S_i = 1 | W_i, T_i = 1) < 1$$

A.3 Strong Ignorability of Sample Assignment for Treated-Controls:

$$(Y_{i00}, Y_{i10}) \perp\!\!\!\perp S_i | (W_i, T_i = 1) \quad 0 < Pr(S_i = 1 | W_i, T_i = 1) < 1$$

A.4 SUTVA: no interference between units

Sufficient to identify:

$$\begin{aligned}\tau_{PATT} &= \mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i, S_i = 1, T_i = 1)\} \\ &\quad - \mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i, S_i = 1, T_i = 0)\}\end{aligned}$$

But the assumptions also imply:

$$\mathbb{E}(Y_i|S_i = 0, T_i = 1) - \mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i, S_i = 1, T_i = 1)\} = 0$$

Placebo Test

$$\mathbb{E}(Y_i|S_i = 0, T_i = 1) - \mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i, S_i = 1, T_i = 1)\} = 0$$

- The difference between the mean outcome of the NRS treated and mean outcome of the reweighted RCT treated should be zero
- If not 0, at least one assumptions has failed

@

- There is a similar placebo test for controls, however, it does not provide as much information
- Could fail due to lack of overlap, for example

Difference-in-Difference

An alternative design:

$$\begin{aligned}\tau_{PATT_{DID}} = & \mathbb{E}_{01}\{\mathbb{E}(Y|W, S=1, T=1) - \mathbb{E}(Y|W, S=1, T=0)\} \\ & - [\mathbb{E}_{01}\{\mathbb{E}(Y|W, S=1, T=1)\} - E(Y|S=0, T=1)]\end{aligned}$$

- The first difference is the adjusted experimental estimand and is intuitively a measure of the adjusted average effect
- The second difference is defined as the difference between the RCT treated and NRS treated
- Required that A3, A4, and part of A1 hold ($Y_{i00} = Y_{i10}$)

IV Alternative

- Our first method didn't used outcomes observed in NRS
- “Diff-in-Diff” estimator uses outcomes from NRS, and estimates PATT under relaxed sample ignorability assumption
- Alternative approach: use sample indicator S as IV under non-ignorable treatment assignment
- Need additional monotonicity assumption: T is increasing in S (conditional on W), with prob 1. Treatment incidence is (weakly) higher for everyone in the RCT sample.
Reasonable?
- Then, we can identify the following LATE:

$$\frac{\mathbb{E}[Y|S=1] - \mathbb{E}[Y|S=0]}{\mathbb{E}[T|S=1] - \mathbb{E}[T|S=0]}$$

which is consistent estimator for $\mathbb{E}[Y_1 - Y_0 | T_{s=1} > T_{s=0}]$

Statistical Methods

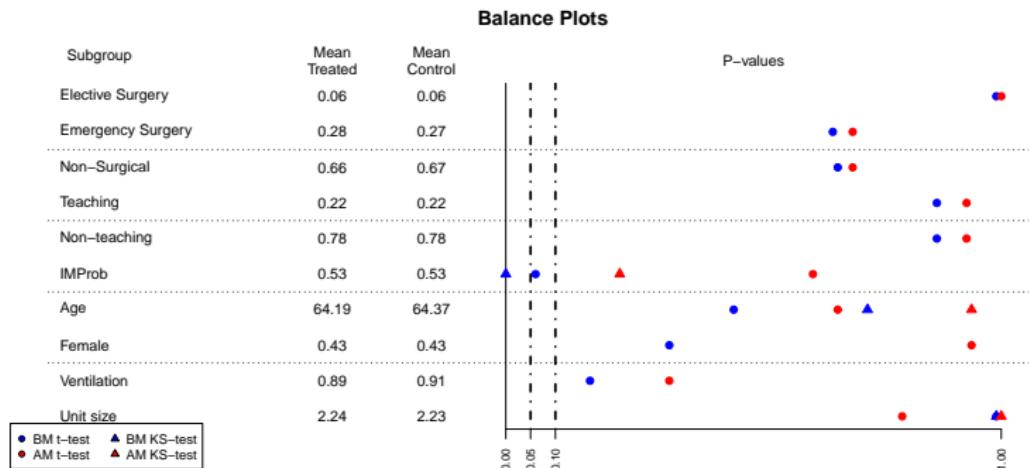
GenMatch:

- Matching method with automated balance optimization. @

Maximum Entropy weighting:

- Weighting method that assigns weights such that they simultaneously meet a set of consistency constraints while maximizing Shannon's measure of entropy.
- Consistency constraints are based on moments of the population based on the NRS.

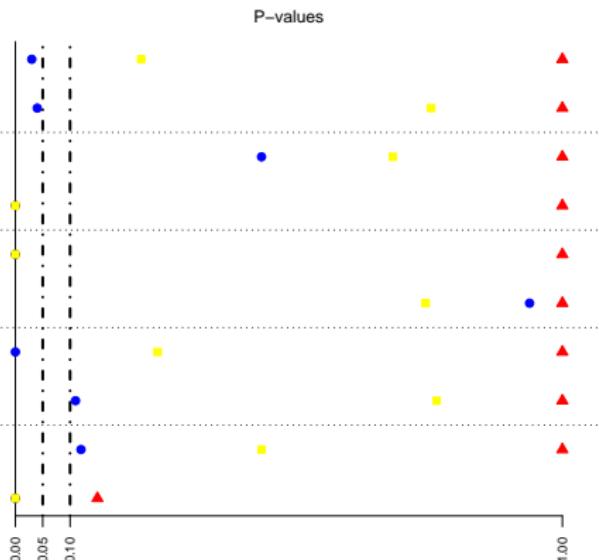
Covariate Balance in RCT



Balance Before and After Adjustment

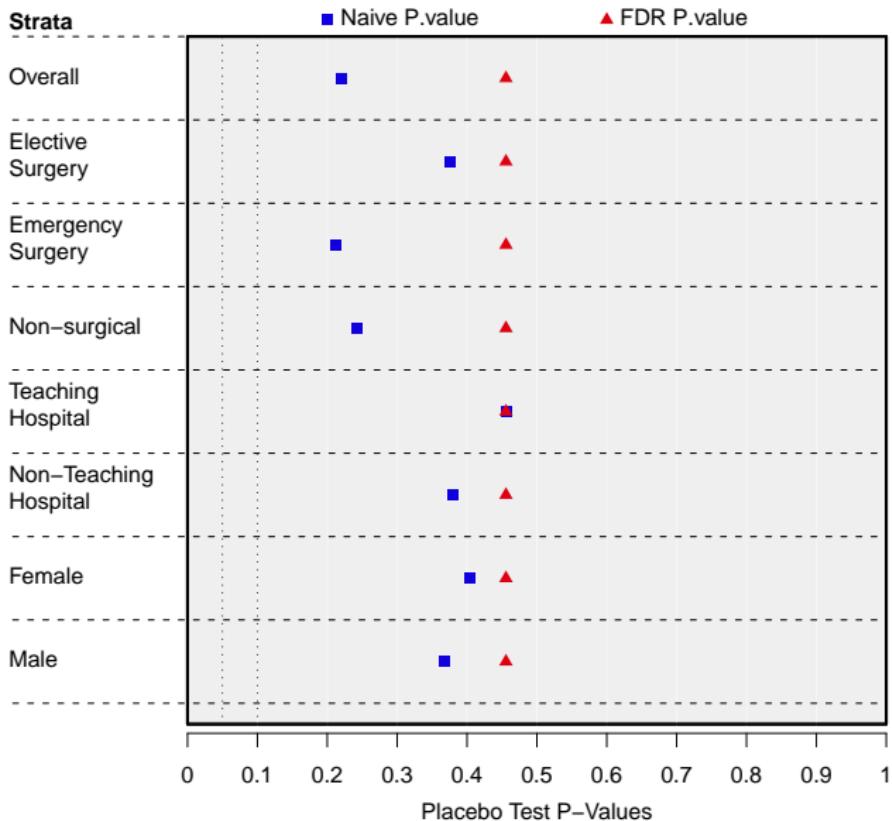
Subgroup	Mean RCT	Mean NRS	Mean RCT After Maxent Adjustment	Mean RCT After IPSW Adjustment
Elective Surgery	0.06	0.09	0.09	0.07
Emergency Surgery	0.28	0.23	0.23	0.24
Non-Surgical	0.66	0.68	0.68	0.69
Teaching	0.22	0.43	0.43	0.28
Non-teaching	0.78	0.57	0.57	0.72
IMProb	0.53	0.52	0.52	0.53
Age	64.19	61.86	61.86	63.49
Female	1.57	1.61	1.61	1.6
Ventilation	0.89	0.86	0.86	0.88
Unit size	2.24	2.66	2.45	2.28

● Before Adjustment ■ After IPSW Adjustment
▲ After Maxent Adjustment

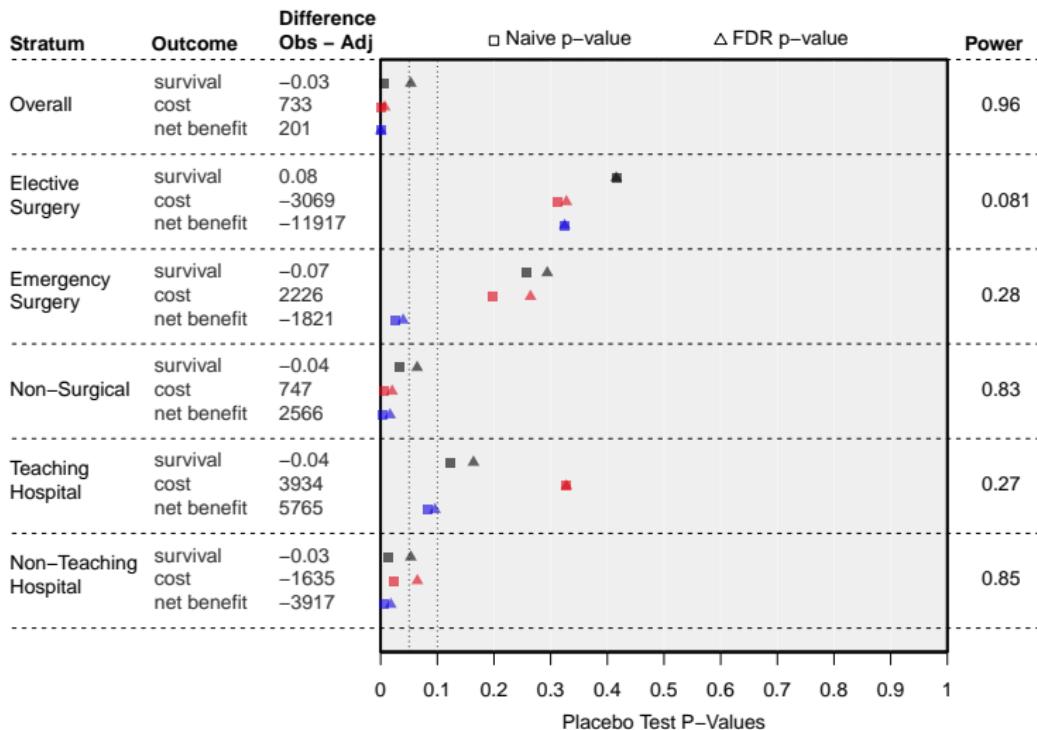


Placebo Tests

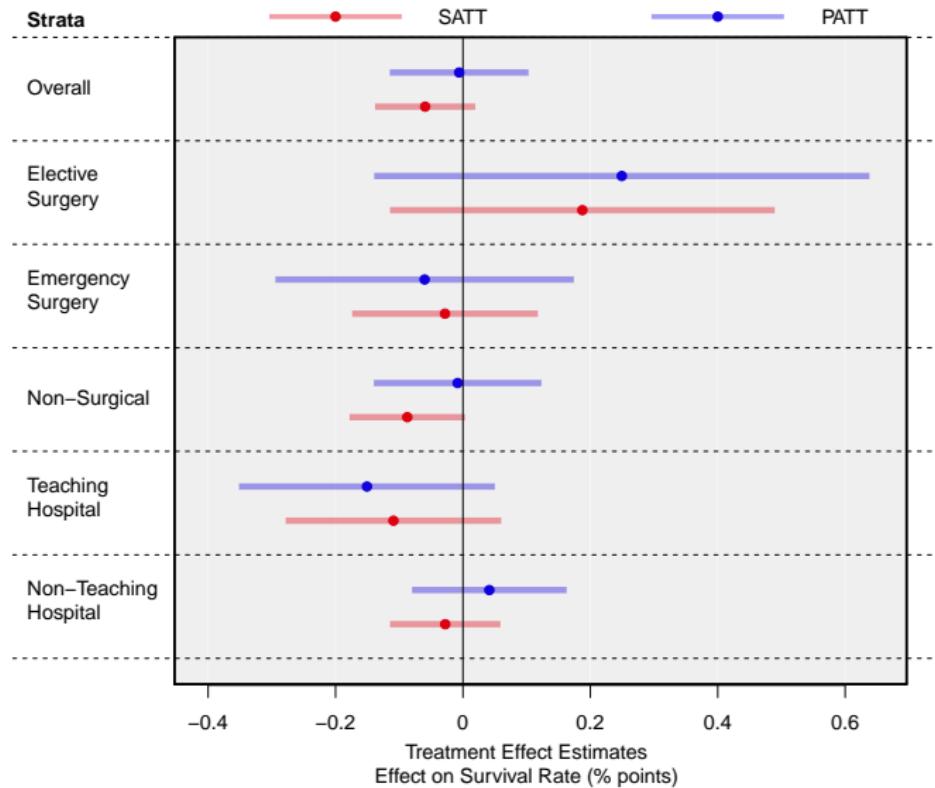
Mortality Placebos



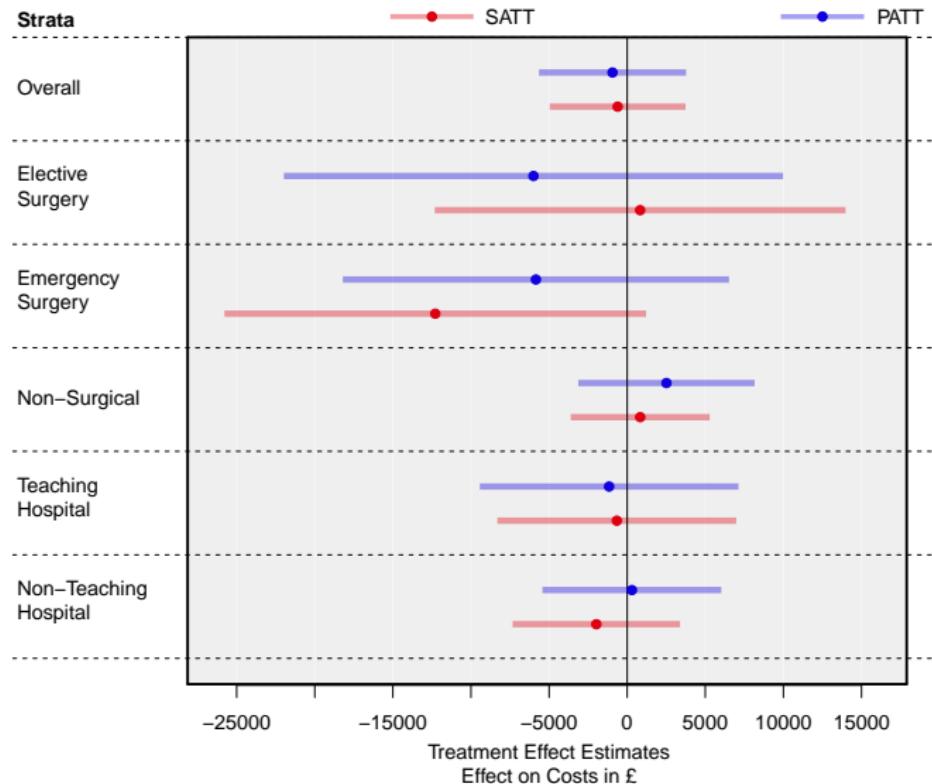
Placebo Tests



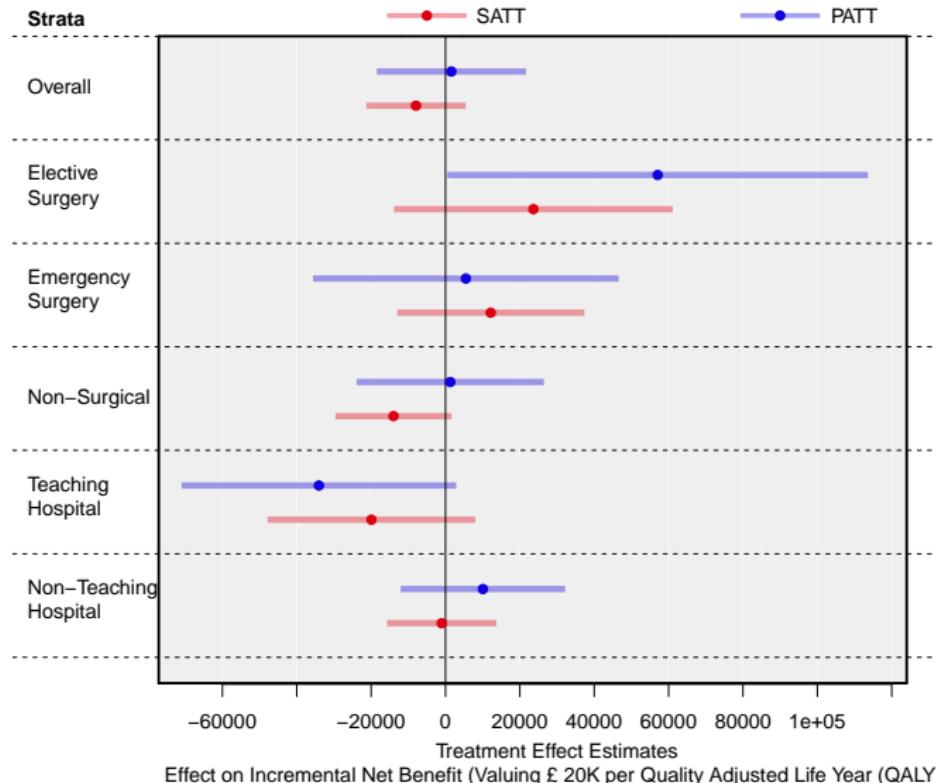
Population Treatment Effects on Hospital Survival Rates



Population Treatment Effects on Costs



Population Treatment Effects on Cost-Effectiveness



Conclusions and Implications

- We pass placebo tests for both costs and hospital mortality, thus validating our identifying assumptions for the PATT.
- This has implications for Cost-Effectiveness Analyses since CEAs are often based on observational studies

- See Hartman et al., forthcoming: [LINK]
- See the references in that paper on Maximum Entropy weighting methods. These are also used in Synthetic Matching—i.e., Abadie, Diamond, and Hainmueller, 2010; Hainmueller, 2012

Maximum Entropy

Jaynes defined the principle of maximum entropy as:

$$\max_{\mathbf{p}} S(\mathbf{p}) = - \sum_{i=1}^n p_i \ln p_i$$

$$s.t. \begin{cases} \sum_{i=1}^n p_i = 1 \\ \sum_{i=1}^n p_i g_r(x_i) = \sum_{i=1}^n p_i g_{ri} = a_r \quad r = 1, \dots, m \\ p_i \geq 0 \quad i = 1, 2, \dots, n \end{cases}$$

- Equation (1) is referred to as a natural constraint, stating that all probabilities must sum to unity.
- Equation (2), the m moment constraints, are referred to as the consistency constraints. Each a_r represents an r -th order moment, or characteristic moment, of the probability distribution.
- Equation (3) the final constraint ensures that all probabilities are non-negative. This is always met.

Balance Tests and the Propensity Score

Any given propensity score model may not be very good at balancing the X covariates. After matching on the pscore we need to test if the underlying X covariates have actually been balanced. You need to test more than just balance of the pscore.

@

The `MatchBalance()` function in `Matching` provides a number of tests, the two main ones being the t-test and the [Kolmogorov-Smirnov Test](#)—the KS test.

@

The `KS` test tries to determine if two distributions differ significantly. The KS-test has the advantage of making no assumption about the distribution of data—i.e., it is distribution free and non-parametric. However, because of this generality, other tests, such as the t-test, are more sensitive to certain differences—e.g., mean differences.

KS Test Details and Example

A great description of the standard KS test is offered at this webpage: <http://www.physics.csbsju.edu/stats/KS-test.html>.

//www.physics.csbsju.edu/stats/KS-test.html.

@But the standard KS test does not provide the correct p-value for noncontinuous variables. But the bootstrap KS does provide the correct p-values. And additional problem arises if we want to test if distribution from estimated propensity scores differ. We then need to do another bootstrap to take into account the distributions of the parameters in the propensity model.

Also see

<http://sekhon.berkeley.edu/causalinf/R/ks1.R>

The Bootstrap

Bootstrap methods are a good way of obtaining confidence intervals and other statistical estimates which require generally weaker assumptions than the usual analytical approaches. @
The bootstrap is also useful when we don't know of an analytical solution. We lack such solutions when:@

- ① we are using complicated statistics for which a sampling distribution is difficult to solve, such as two step estimators @
- ② we don't want to make population assumptions

Percentile Bootstrap Intervals

The basic setup is as follows:

- We have a sample of size n from a given from a probability distribution F

$$F \rightarrow (x_1, x_2, x_3, \dots, x_n) \odot$$

- The empirical distribution function \hat{F} is defined to be the discrete distribution that puts equal probability, $\frac{1}{n}$, on each value x_i . @
- A set X which is made up of various x_i has a probability assigned by \hat{F} :

$$\hat{\text{prob}}\{X\} = \#\{x_i \in X\}/n \odot$$

- We can make program via the plug-in principle. The plug-in estimate of a parameter $\theta = t(F)$ is $\hat{\theta} = t(\hat{F})$

Bootstrap Intervals Algorithm

- For a dataset (denoted “full sample”) iterate the following B times where B is a large number such as 1,000. @
 - 1 Generate a random sample of size n from x_1, \dots, x_n with replacement. This is called a bootstrap sample. @
 - 2 calculate the statistic of interest using this bootstrap sample, denoted $\hat{\theta}_b$. @
- The confidence interval for $\hat{\theta}$ (our full sample estimate) is obtained by taking the quantiles from the sampling distribution in our bootstrap samples: $\hat{\theta}_b$. @
- For example, for 95% CI, calculate the 2.5% and 97.5% quantiles of the distribution of $\hat{\theta}_b$.

Full Sample Estimate

- An estimate for the full sample quantity is provided by:

$$\hat{\theta}^* = \sum_{b=1}^B \hat{\theta}_b / B$$

This is an alternative to whatever we usually do in the full sample, whether it be the usual least squares estimates, the mean, median or whatever.

Bootstrapping Regression Coefficients

- For a dataset (denoted “full sample”) iterate the following B times. @
- Generate a random sample of size n from $\{Y, X\}$, with replacement. Denote this sample of data S^b . @
- Estimate your model using the dataset S^b . This results in a vector of bootstrap sample estimates of β denoted $\hat{\beta}_b$.

$$Y^b = \hat{\beta}^b X^b + \hat{\epsilon}^b @$$

- Create CIs and SEs in the usual bootstrap fashion based on the distribution $\hat{\beta}^b$.

Comments

So far we've talked about non-parametric bootstraps. And:

- We are treating the sample as a population and sampling from it. @
- does makes that assumption that the observations x_i are **independent**. It does not assume that they are homoscedastic. @
- can be applied to any function of the data which is smooth (this is a technical assumption) such as least squares regression. @
- the resulting estimates, point estimates, CIs and standard errors have $\frac{1}{\sqrt{n}}$ asymptotics just like the usual normal theory estimates. There are better bootstraps with, for example, $\frac{1}{n}$ asymptotics. Examples of such bootstraps are: calibrated bootstrap, percentile- t bootstrap and the bias corrected, BC_a , bootstrap.

Distributions of Difficult to Handle Quantities

The bootstrap provides an easy way to obtain the sampling distribution of quantities for which it is often not known or difficult to obtain sampling distributions. For example:

- CIs for the median or other quantiles @
- CIs and SEs for LMS coefficients @
- In a regression, is $\hat{\beta}_1 > \hat{\beta}_2$ @
- In a regression, is $\frac{\hat{\beta}_1}{\hat{\beta}_2} < \frac{\hat{\beta}_3}{\hat{\beta}_4}$ @
- Two stage regression estimation. Say a logit in the first stage and a linear model in the second. @
- Goodness-of-fit for two model which may not be nested.

Bootstrap Hypothesis Tests from CIs

This is the simplest way:

- Given test statistic $\hat{\theta}$, we are interested in obtaining the achieved significance level (ASL) of the test which is:

$$\text{ASL} = \text{Prob}_{H_0} \left\{ \hat{\theta}_b \geq \hat{\theta} \right\}$$

- Smaller the value of ASL, the stronger the evidence against H_0 .
- A confidence interval $(\hat{\theta}_{lo}, \hat{\theta}_{up})$ is the set of plausible values of θ having observed $\hat{\theta}$. Values not ruled out as being unlikely.

Difference of Means (case 1)

- For a dataset iterate the following B times. @
 - ① Generate a random sample of size n from x_1, \dots, x_n with replacement. Do the same for the m obs in z . @
 - ② calculate the test statistic of interest using this bootstrap sample, denoted $\hat{\theta}_b = \text{mean}(x_b) - \text{mean}(z_b)$. @
- The ASL under $H_0 = 0$ is

$$\hat{\text{prob}}(H_0) = \# \left\{ \hat{\theta}_b \geq 0 \right\} / B,$$

Which Side?

- Note that the previous ASL is based on the proportion of the empiric above 0. Or the portion which is most relevant if $\text{mean}(x) < \text{mean}(z)$ —i.e., $(\text{mean}(x) - \text{mean}(z)) < 0$
- The portion of the empiric below zero would be of interest in rejecting the null if $\text{mean}(x) > \text{mean}(z)$. If we decide which side to look at based on the full sample statistic, we need to adjust our ASL by dividing by the probability of choosing the side (generally .5). For details see the [bs1mc1.R], [bs1mc1b.R] files.

The forging algorithm is presented in the [bs1.R] file as case 1. The following cases are also presented in that R file. Please examine it closely.

The proceeding algorithm only works when the test statistic under the null is centered around zero. If not, we need to subtract the null from $\hat{\theta}_b$.

Difference of Means (case 2)

Here's another way to test the equivalent hypothesis.

- Bootstrap as in case 1, but now difference the full sample estimate which is $\hat{\theta}$ @
- The bootstrap ASL of interest if $mean(x) < mean(z)$,

$$\hat{prob}(H_0) = \# \left\{ \hat{\theta}_b - \hat{\theta} < \hat{\theta} \right\} / B,$$

else

$$\hat{prob}(H_0) = \# \left\{ \hat{\theta}_b - \hat{\theta} \geq \hat{\theta} \right\} / B,$$

Difference of Means (case 3)

The studentized version

- Bootstrap as in case 2, but let's studentize the quantities:

$$\hat{\theta} = \frac{\text{mean}(x) - \text{mean}(z)}{\sqrt{\text{var}(x) + \text{var}(z)}}$$

$\hat{\theta}_b$ is analogously defined

- The bootstrap ASL of interest if $\text{mean}(x) < \text{mean}(z)$

$$\hat{\text{prob}}(H_0) = \# \left\{ \hat{\theta}_b - \hat{\theta} < \hat{\theta} \right\} / B,$$

and for $\text{mean}(x) > \text{mean}(z)$

$$\hat{\text{prob}}(H_0) = \# \left\{ \hat{\theta}_b - \hat{\theta} \geq \hat{\theta} \right\} / B,$$

Difference of Means (case 4)

Here we can test if the two distributions F and G are equal. We will, however, use this just for the means though. The KS test is based on this bootstrap. See the code for [ks.boot]

- Draw B samples of size $n + m$ with replacement from $c(x, z)$. Call the first n observations x_b and the remaining m observations z_b .
- Construct $\hat{\theta}$ and calculate ASL as in case 2. [@](#)

This version is much like a [permutation](#) test except that we are (1) sampling with replacement and (2) we are learning about the distribution under the null by sampling instead of by construction.

Different Null Hypothesizes

- Code which adapts case four to use the KS test is provided at: [bs2.R]
- All four cases are based on the null hypothesis that the means in the two samples are equal. The null hypothesis need not be so restrictive for the first three cases.
- Why in the fourth case doesn't non-equality make sense? We could generalize (much like a sharp null).
- The following cases generalize the first three cases for the non-zero difference setting. For R code see: [bs1_null.R].

Difference of Means Redux (case 1*)

- Bootstrap as before. Note that the null hypothesis is H_0 @
- The bootstrap ASL of interest if
 $[mean(x) - mean(z)] < H_0$, is

$$\hat{\text{prob}}(H_0) = \# \left\{ \hat{\theta}_b \geq H_0 \right\} / B,$$

else

$$\hat{\text{prob}}(H_0) = \# \left\{ \hat{\theta}_b < H_0 \right\} / B.$$

Difference of Means Redux (case 2*)

- Bootstrap as before.
- The bootstrap ASL of interest if $[mean(x) - mean(z)] < H_0$, is

$$\hat{prob}(H_o) = \# \left\{ \hat{\theta}_b - \hat{\theta} < \hat{\theta} - H_0 \right\} / B,$$

else

$$\hat{prob}(H_o) = \# \left\{ \hat{\theta}_b - \hat{\theta} \geq \hat{\theta} - H_0 \right\} / B,$$

Difference of Means Redux (case 3*)

The studentized version

- Bootstrap as before and let's studentize the quantities:

$$\hat{\theta} = \frac{\text{mean}(x) - \text{mean}(z)}{\sqrt{\text{var}(x) + \text{var}(z)}}$$

$\hat{\theta}_b$ is analogously defined. And the studentized H_0 is:

$$H'_0 = \frac{\text{mean}(x) - \text{mean}(z) - H_0}{\sqrt{\text{var}(x) + \text{var}(z)}}$$

@

- The bootstrap ASL of interest if $[\text{mean}(x) - \text{mean}(z)] < H_0$, is

$$\hat{\text{prob}}(H_o) = \# \left\{ \hat{\theta}_b - \hat{\theta} < H'_0 \right\} / B,$$

and for $\text{mean}(x) > \text{mean}(z)$

$$\hat{\text{prob}}(H_o) = \# \left\{ \hat{\theta}_b - \hat{\theta} \geq H'_0 \right\} / B,$$

Parametric Bootstrap

The previous bootstraps assume that the observations are **independent**. But this may not be the case. Also we may want to assume that X is fixed. In these cases, one may estimate a parametric bootstrap if it can be assumed that the residuals of a parametric model are **independent** even if the original data is not. This often occurs with time-series models.

Parametric Bootstrap Algorithm

- Estimate the model in the full-sample:

$$Y = \hat{\beta}X + \hat{\epsilon}$$

- Iterate the following B times. @
- Generate a random sample of size n from $\hat{\epsilon}$, with replacement. Denote this $\hat{\epsilon}^b$. Create the bootstrap sample by adding the new $\hat{\epsilon}^b$ to the full sample X and $\hat{\beta}$ (there is NO estimation in this step, just addition):

$$Y^b = \hat{\beta}X + \hat{\epsilon}^b$$

- Estimate your model using the full sample X but the bootstrap sample Y^b . This results in a vector of bootstrap estimates of β estimates denoted β_b .

$$Y^b = \hat{\beta}^b X + \hat{\epsilon}^b$$

- Create CIs and SEs in the usual bootstrap fashion based on the distribution $\hat{\beta}^b$.

Two Nice Bootstrap Properties

Bootstrap CI estimates are:

- ① **Range Preserving.** The confidence interval cannot extend beyond the range which estimates of θ range in bootstrap samples. For example, if we are bootstrapping estimates of a proportion, our confidence interval cannot be less than 0 or greater than 1. The usual normal theory confidence intervals are not range persevering. @
- ② **Take into account more general uncertainty.** Many estimators, such as LMS, have the usual statistical uncertainty (due to sampling) but also have uncertainty because the algorithm used to obtain the estimates is stochastic or otherwise suboptimal. Bootstrap CIs will take this uncertainty into account. See the **R** help on the **lqs** function for more information about how the LMS estimates are obtained.

Bootstrap Readings

For bootstrap algorithms see Venables and Ripley (2002). Especially, p.133-138 (section 5.7) and p.163-165 (section 6.6). The **R** "boot" command is of particular interest. [@](#)

For a statistical discussion see Chapter 16 (p.493ff) of Fox (2002). [@](#)

For additional readings see:

- Efron, Bradley and Tibshirani, Robert J. 1994. *An Introduction to the Bootstrap*. Chapman & Hall. ISBN: 0412042312.
- Davison, A. C. and Hinkley, D. V. 1997. *Bootstrap Methods and their Applications*. Cambridge University Press.
- Go to jstor and search for Efron and/or Tibshirani.

The Search Problem

- Many search problems have an exponential asymptotic order: $O(c^N)$, $c > 1$ where N is the sample size.
 - Finding the optimal set of matches is such a problem.
@E.G., we want to estimate ATT and do matching without replacement:
 - with 10 treated and 20 control obs: 184,756 possible matches @
 - with 20 treated and 40 control obs: 13,784,652,8820 possible matches @
 - with 40 treated and 80 control obs: 1.075072e+23 @
 - in the LaLonde data with 185 treated and 260 control: 1.633347e+69 @
 - with 185 treated and 4000 control: computer infinity
- @Matching with replacement makes the search problem explode even more quickly.

The Solution

- It is impossible to search all possible matches in finite time.
- We need to make the problem tractable
- This is done by finding a solution which is “good enough”
- A variety of **global** search algorithms can do this including **simulated annealing** and **genetic optimization**
- Both are variants of random search

Difficult Optimization Problems

- Loss functions of statistical models are often irregular
- ML models are generally not globally concave
- As such, derivative methods fail to find the global maximum
- And Taylor series methods for SEs do not provide correct coverage

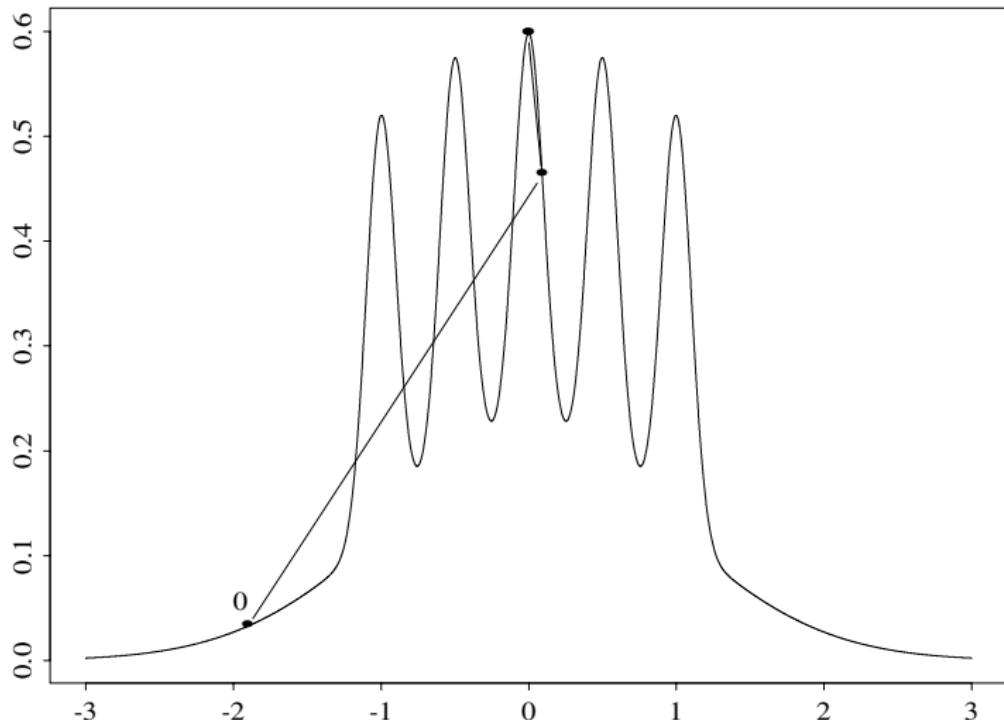
Normal Mixtures

Mixtures of Normal Densities provides a nice way to create difficult optimization surfaces.

The Claw:

$$\frac{1}{2}N(0, 1) + \sum_{m=0}^4 \frac{1}{10}N\left(\frac{m}{2} - 1, \left(\frac{1}{10}\right)\right)$$

The Claw



Normal Mixtures

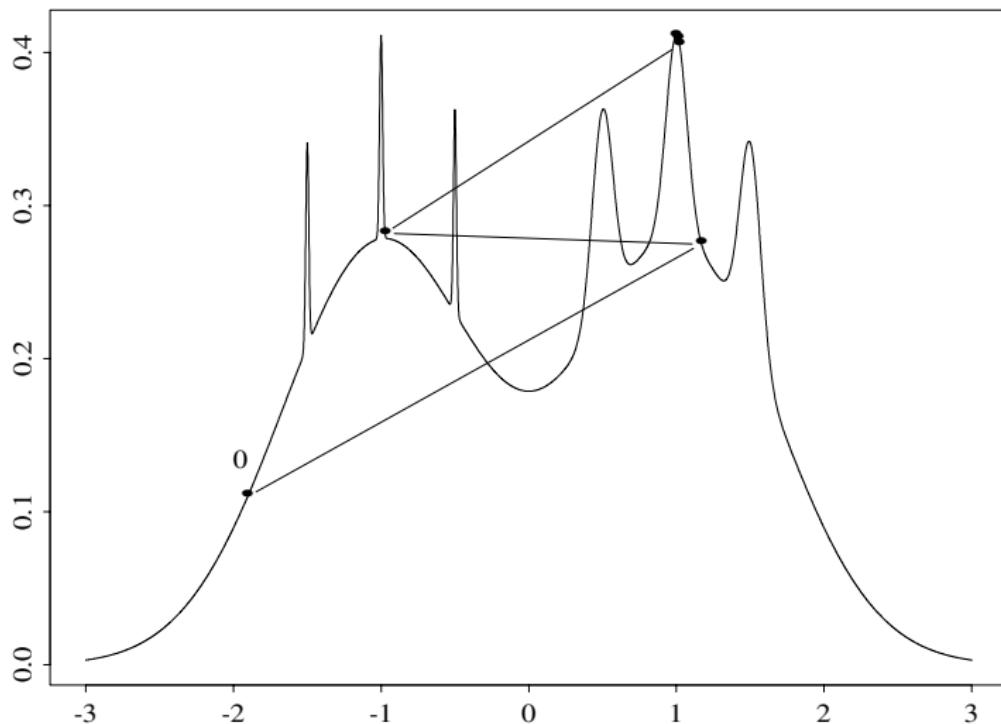
Asymmetric Double Claw:

$$\sum_{m=0}^1 \frac{46}{100} N(2m - 1, \frac{2}{3}) +$$

$$\sum_{m=1}^3 \frac{1}{300} N\left(\frac{-m}{2}, \frac{1}{100}\right) +$$

$$\sum_{m=1}^3 \frac{7}{300} N\left(\frac{m}{2}, \frac{7}{100}\right)$$

Asymmetric Double Claw



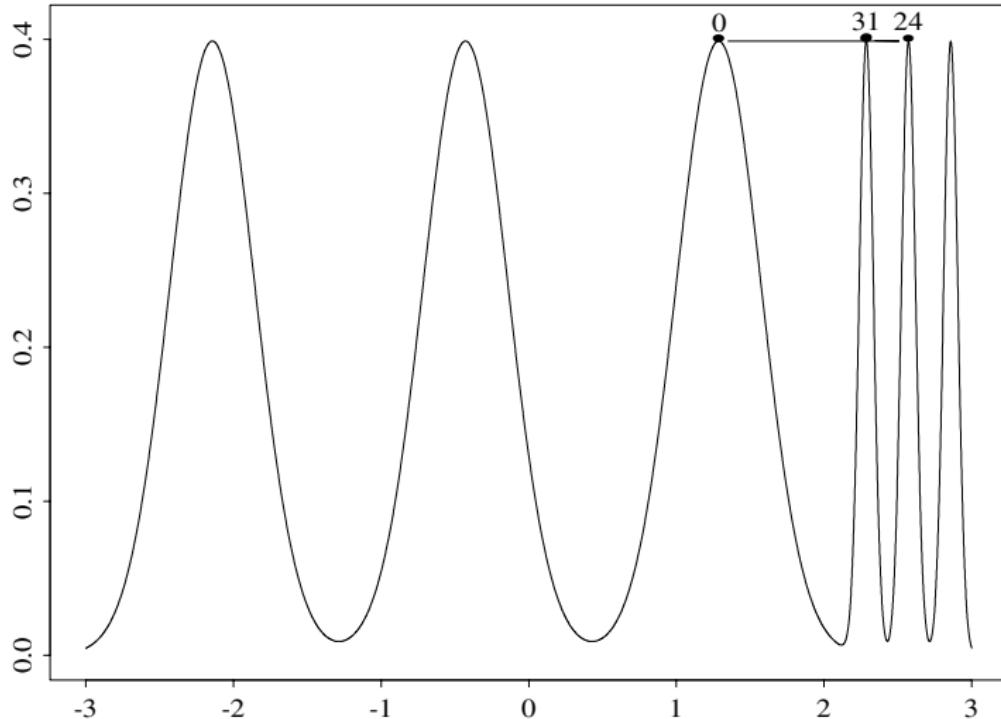
Normal Mixtures

Discrete Comb Density:

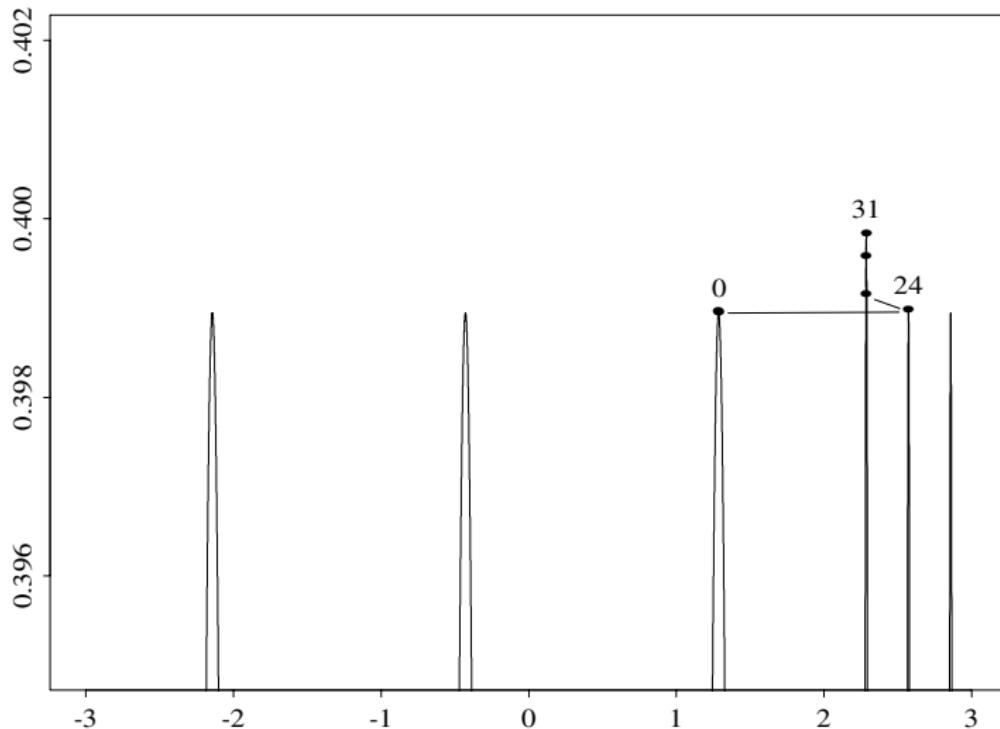
$$\sum_{m=0}^2 \frac{2}{7} N\left(\frac{(12m - 15)}{7}, \frac{2}{7}\right) +$$

$$\sum_{m=0}^2 \frac{1}{21} N\left(\frac{2m + 16}{7}, \frac{1}{21}\right)$$

Discrete Comb Density



Discrete Comb Density: Close Up



Evolutionary Search (EA)

- An EA uses a collection of heuristic rules to modify a population of trial solutions in such a way that each generation of trial values tends to be on average better than its predecessor. @
- The EA in GENOUD assumes a solution is a vector of real numbers, each number being a value for a scalar parameter of a function to be optimized.
- parameter==gene
- Each heuristic rule, or operator, acts on one or more trial solutions from the current population to produce one or more trial solutions to be included in the new population.

Description

- A GA uses a set of randomized genetic operators to evolve a finite population of finite code-strings over a series of generations (Holland 1975; Goldberg 1989; Grefenstette and Baker 1989)
- The operators used in GA implementations vary, but the basic set of operators can be defined as **reproduction**, **mutation** and **crossover** (Davis 1991; Filho, Treleaven and Alippi 1994).

General Description of Operators

- **reproduction**: a copy of an individual in the current population is reproduced into the new one. Analogous to asexual reproduction. @
- **mutation**: randomly alter parameter values (genes) so as to search the entire space @
- **crossover**: Create a new individual from one or more individuals from the current population. Splice some parameters from individual A with the parameter values from another individual. Analogous to sexual reproduction.
@
- All three can be combined to create mixed operators

The Actual Operators

Some notation:

$\mathbf{X} = [X_1, \dots, X_n]$ is the vector of *n parameters* X_i . \underline{x}_i is the lower bound and \bar{x}_i is the upper bound on values for X_i . x_i is the current value of X_i , and \mathbf{x} is the current value of \mathbf{X} .

$\mathbf{N} = \{1, \dots, n\}$. $p \sim U(0, 1)$ means that p is drawn from the uniform distribution on the $[0, 1]$ interval.

Operators 1-4

- 1 Cloning. Copy \mathbf{X}_t into the next generation, \mathbf{X}_{t+1} . @
- 2 Uniform Mutation. At random choose $i \in \mathbf{N}$. Select a value $\tilde{x}_i \sim U(\underline{x}_i, \bar{x}_i)$. Set $X_i = \tilde{x}_i$. @
- 3 Boundary Mutation. At random choose $i \in \mathbf{N}$. Set either $X_i = \underline{x}_i$ or $X_i = \bar{x}_i$, with probability 1/2 of using each value. @
- 4 Non-uniform Mutation. At random choose $i \in \mathbf{N}$. Compute $p = (1 - t/T)^B u$, where t is the current generation number, T is the maximum number of generations, $B > 0$ is a tuning parameter and $u \sim U(0, 1)$. Set either $X_i = (1 - p)x_i + p\underline{x}_i$ or $X_i = (1 - p)x_i + p\bar{x}_i$, with probability 1/2 of using each value.

Operators 5-7

- 5 Polytope Crossover. Using $m = \max(2, n)$ vectors \mathbf{x} from the current population and m random numbers $p_j \in (0, 1)$ such that $\sum_{j=1}^m p_j = 1$, set $\mathbf{X} = \sum_{j=1}^m p_j \mathbf{x}_j$. @
- 6 Simple Crossover. Choose a single integer i from \mathbf{N} . Using two parameter vectors, \mathbf{x} and \mathbf{y} , for the i set $X_i = px_i + (1 - p)y_i$ and $Y_i = py_i + (1 - p)x_i$, where $p \in (0, 1)$ is a fixed number. @
- 7 Whole Non-uniform Mutation. Do non-uniform mutation for all the elements of \mathbf{X} .

Operators 8

- 8 Heuristic Crossover. Choose $p \sim U(0, 1)$. Using two parameter vectors, \mathbf{x} and \mathbf{y} , compute $\mathbf{z} = p(\mathbf{x} - \mathbf{y}) + \mathbf{x}$. If \mathbf{z} satisfies all constraints, use it. Otherwise choose another p value and repeat. Set \mathbf{z} equal to the better of \mathbf{x} and \mathbf{y} if a satisfactory mixed \mathbf{z} is not found by a preset number of attempts. In this fashion produce two \mathbf{z} vectors.

Operators 9

- 9 Local-minimum Crossover. NOT USED IN GENMATCH.
- Choose $p \sim U(0, 1)$. Starting with \mathbf{x} , run BFGS optimization up to a preset number of iterations to produce \mathbf{x} . Compute $\mathbf{z} = p\mathbf{x} + (1 - p)\mathbf{x}$. If \mathbf{z} satisfies boundary constraints, use it. Otherwise shrink p by setting $p = p/2$ and recompute \mathbf{z} . If a satisfactory \mathbf{z} is not found by a preset number of attempts, return \mathbf{x} . This operator is extremely computationally intensive, use sparingly.

Markov Chains

GAs are Markov chains:

A Markov chain describes at successive times the states of a system. At these times the system may have changed from the state it was in the moment before to another or stayed in the same state. The changes of state are called transitions. The Markov property means that the conditional probability distribution of the state in the future, given the state of the process currently and in the past, depends only on its current state and not on its state in the past.

Markov Chains II

A Markov chain is a sequence of random variables X_1, X_2, X_3, \dots with the Markov property, that is, given the present state, the future state and past states are independent. Formally:

$$\Pr(X_{n+1} = x | X_n = x_n, \dots, X_1 = x_1, X_0 = x_0) = \\ \Pr(X_{n+1} = x | X_n = x_n)$$

The possible values of X_i form a countable set S which is the state space of the chain. Continuous time chains exist.

- For more on Markov chains cites in GENOUD article or
http:
`//random.mat.sbg.ac.at/~ste/diss/node6.html`
- A finite GA with random reproduction and mutation is an aperiodic and irreducible Markov chain.
- Such a chain converges at an exponential rate to a unique stationary distribution (Billingsley 1986, 128).
- the probability that each population occurs rapidly converges to a constant, positive value.
- Nix and Vose (1992; Vose 1993) show that in a GA where the probability that each code-string is selected to reproduce is proportional to its observed fitness, the stationary distribution strongly emphasizes populations that contain code-strings that have high fitness values.

Population Size

- asymptotically in the population size i.e., in the limit for a series of GAs with successively larger populations populations that have suboptimal average fitness have probabilities approaching zero in the stationary distribution, while the probability for the population that has optimal average fitness approaches one.
- The crucial practical implication from the theoretical results of Nix and Vose is that a GA's success as an optimizer depends on having a sufficiently large population.

Robustness of Randomization Inference

- Randomization inference is often more robust than alternatives
- An example of this is when method of moment estimators become close to unidentified in finite samples
- An illustration is provided by the Wald (1940) estimator, which is often used to estimate treatment effects with instruments

What is IV Used For?

- IV used for causal inference
- The most compelling common example: estimate the average treatment effect when there is one-way non-compliance in an experiment.
- Assumptions are weak in this case. In most other examples, the behavioral assumptions are strong.

One Way Non-Compliance

Only two types of units: compliers and never takers

We have five different parameters:

- (i) the proportion of compliers in the experimental population (α)
- (ii) the average response of compliers assigned to treatment (\bar{W})
- (iii) the average response of compliers assigned to control (\bar{C})
- (iv) the difference between \bar{W} and \bar{C} , which is the average effect of treatment on the compliers (\bar{R})
- (v) the average response of never-treat units assigned to control (\bar{Z})

Estimation I

- α can be estimated by calculating the proportion of compliers observed in the treatment group
- The average response of compliers to treatment, \bar{W} , is the average response of compliers in the treatment group
- \bar{Z} , the average response of never-treat units to control, is estimated by the average response among units in the treatment group who refused treatment

Estimation II

This leaves \bar{C} and \bar{R} .

For \bar{R} , the control group contains a mix of compliers and never-treat units. We know (in expectation) the *proportion* of each type there must be in control because we can estimate this proportion in the treated group.

α denotes the proportion of compliers in the experimental population, and assume that $\alpha > 0$. Under the model, the proportion of never-treat units must be $1 - \alpha$.

Denote the average observed responses in treatment and control by \bar{Y}^t , \bar{Y}^c , these are sample quantities which are directly observed.

Estimation III

$$E(\bar{Y}^c) = \alpha \bar{C} + (1 - \alpha) \bar{Z}.$$

Therefore,

$$\bar{C} = \frac{E(\bar{Y}^c) - (1 - \alpha) \bar{Z}}{\alpha}.$$

An obvious estimator for \bar{C} is

$$\hat{C} = \frac{\bar{Y}^c - (1 - \hat{\alpha}) \hat{Z}}{\hat{\alpha}}.$$

Then the only remaining quantity is \bar{R} , the average effect of treatment on the compliers—i.e., the Effect of Treatment on the Treated (ETT). This can be estimated by

$$\hat{W} - \hat{C} = \frac{\bar{Y}^t - \bar{Y}^c}{\hat{\alpha}}. \quad (17)$$

Instrumental Variables (IV)

- IV methods solve the problem of missing or unknown control **IF** the instruments are valid
- Simple example under the unnecessary assumption of constant effects. This assumption is only used to simply the presentation here:

$$\alpha = Y_{1i} - Y_{0i}$$

$$Y_{0i} = \beta + \epsilon_i,$$

where $\beta \equiv \mathbb{E}[Y_{0i}]$

- When we do permutation inference, we will assume under the null, unlike the Wald estimator, that the potential outcomes are fixed.

Towards the Wald Estimator I

- The potential outcomes model can now be written:

$$Y_i = \beta + \alpha T_i + \epsilon_i, \tag{18}$$

But T_i is likely correlated with ϵ_i .

- Suppose: a third variable Z_i which is correlated with T_i , but is unrelated to Y for any other reason—i.e., $Y_{0i} \perp\!\!\!\perp Z_i$ and $\mathbb{E}[\epsilon_i | Z_i] = 0$.
- Z_i is said to be an IV or “an instrument” for the causal effect of T on Y

Towards the Wald Estimator II

- Suppose that Z_i is dichotomous (0,1). Then

$$\alpha = \frac{\mathbb{E}[Y_i | Z=1] - \mathbb{E}[Y_i | Z=0]}{\mathbb{E}[T_i | Z=1] - \mathbb{E}[T_i | Z=0]} \quad (19)$$

- The sample analog of this equation is called the Wald estimator, since it first appear in Wald (1940) on errors-in-variables problems.
- More general versions for continuous, multi-valued, or multiple instruments.
- Problem: what if in finite samples the denominator in Eq 19 is close to zero?

Conditions for a Valid Instrument

- Instrument Relevance:

$$\text{cov}(Z, T) \neq 0$$

- Instrument Exogeneity:

$$\text{cov}(Z, \epsilon) = 0$$

- These conditions ensure that the part of X that is correlated with Z only contains exogenous variation
- Instrument relevance is testable
- Instrument exogeneity is **NOT** testable. It must be true by design

Weak Instruments

- A weak instrument is one where the denominator in Eq 19 is close to zero.
- This poses two distinct problems:
 - if the instrument is extremely weak, it may provide little or no useful information
 - commonly used statistical methods for IV do not accurately report this lack of information

Two Stage Least Squares Estimator (TSLS)

$$Y = \beta T + \epsilon \tag{20}$$

$$T = Z\gamma + v, \tag{21}$$

where Y , ϵ , v are $N \times 1$, and Z is $N \times K$, where K is the number of instruments

- Note we do not assume that $\mathbb{E}(T, \epsilon) = 0$, which is the central problem
- We assume instead $\mathbb{E}(v | Z) = 0$, $\mathbb{E}(v, \epsilon) = 0$, $\mathbb{E}(Z, \epsilon) = 0$, $\mathbb{E}(Z, T) \neq 0$

TSLS Estimator

$$\hat{\beta}_{iv} = (T' P_z T)^{-1} T' P_z Y, \quad (22)$$

where $P_z = Z(Z'Z)^{-1}Z'$, the projection matrix for Z . It can be shown that:

- $\text{plim } \hat{\beta}_{ols} = \beta + \frac{\sigma_{T,\epsilon}}{\sigma_T^2}$
- $\text{plim } \hat{\beta}_{iv} = \beta + \frac{\sigma_{T,\epsilon}}{\sigma_T^2}$

Quarter of Birth and Returns to Schooling

- J. Angrist and Krueger (1991) want to estimate the causal effect of education
- This is difficult, so they propose a way to estimate the causal effect of compulsory school attendance on earnings
- Every state has a minimum number δ of years of schooling that all students must have
- But, the laws are written in terms of the age at which a student can leave school
- Because birth dates vary, individual students are required to attend between δ and $\delta + 1$ years of schooling

Quarter of Birth

- The treatment T is years of education
- The instrument Z is quarter of birth (exact birth date to be precise)
- The outcome Y are earnings

The Data

- Census data is used
- 329,509 men born between 1930 and 1939
- Observe: years of schooling, birth date, earnings in 1980
- mean number of years of education is 12.75
- Example instrument: being born in the fourth quarter of the year, which is 1 for 24.5% of sample

How Much Information Is There?

- If the number of years of education is regressed on this quarter-of-birth indicator, the least squares regression coefficient is 0.092 with standard error 0.013
- if log-earnings are regressed on the quarter-of-birth indicator, the coefficient is 0.0068 with standard error 0.0027, so being born in the fourth quarter is associated with about $\frac{2}{3}\%$ higher earnings.
- Wald estimator: $\frac{0.0068}{0.092} \approx 0.074$

Estimates Using the Wald Estimator

Estimate	95% lower	95% upper
Simple Wald Estimator		
0.074	0.019	0.129
Multivariate TSLS Estimator		
0.074	0.058	0.090

Problem

- replace the actual quarter-of-birth variable by a randomly generated instrument that carries no information because it is unrelated to years of education.
- As first reported by Bound, Jaeger, and Baker (1995), the TSLS estimate incorrectly suggests that the data are informative, indeed, very informative when there are many instruments.
- The 95% confidence interval: **0.042, 0.078**
- But the true estimate is 0!

Method

- There are S strata with n_s units in stratum s and N subjects in total
- Y_{Csi} is the control ($T = 0$) potential outcome for unit i in strata s
- Y_{tsi} is the potential outcome for unit i in strata s if the unit received treatment $t \neq 0$
- Simplifying assumption (not necessary), effect is proportional:
$$Y_{tsi} - Y_{Csi} = \beta t.$$
- t is years of education beyond the minimum that are required by law

Instruments

- In stratum s there is a preset, sorted, fixed list of n_s instrument settings $h_{sj}, j = 1, \dots, n_s$, where $h_{sj} \leq h_{s,j+1} \forall s, j$
- $\mathbf{h} = (h_{11}, h_{12}, \dots, h_{1,n_1}, h_{2,1}, \dots, h_{S,n_s})^T$
- Instrument settings in \mathbf{h} are randomly permuted within strata
- Assignment of instrument settings, \mathbf{z} , is $\mathbf{z}=\mathbf{ph}$ where \mathbf{p} is a stratified permutation matrix—i.e., an $N \times N$ block diagonal matrix with S blocks, $\mathbf{p}_1, \dots, \mathbf{p}_S$.
- Block \mathbf{p}_s is an $n_s \times n_s$ permutation matrix—i.e., \mathbf{p}_s is a matrix of 0s and 1s s.t. each row and column sum to 1

Permutations

- Let Ω be the set of all stratified permutation matrices \mathbf{p} , so Ω is a set containing $|\Omega| = \prod_{s=1}^S n_s!$ matrices, where $|\Omega|$ denotes the number of elements of the set Ω
- Pick a random \mathbf{P} from Ω where $Pr(P = p) = \frac{1}{|\Omega|}$ for each $\mathbf{p} \in \Omega$
- Then $\mathbf{Z} = \mathbf{Ph}$ is a random permutation of \mathbf{h} within strata, so the i th unit in stratum s receives instrument setting Z_{si}

Outcomes

- For each \mathbf{z} there is a $t_{s\mathbf{z}}$ for each unit who then has an outcome $Y_{Csi} + \beta t_{s\mathbf{z}}$
- T_{si} is the dose, treatment value, for unit i in stratum s so $T_{si} = t_{s\mathbf{z}}$
- Let Y_{si} be the response for this unit, so $Y_{si} = Y_{Csi} + \beta T_{si}$
- Write $\mathbf{T} = (T_{11}, \dots, T_{S,n_s}^T)$ and
 $\mathbf{Y} = (Y_{11}, \dots, Y_{S,n_s}^T)$

Hypothesis Testing I

- We wish to test $H_0 : \beta = \beta_0$
- Let $\mathbf{q}(\cdot)$ be a method of scoring response such as their ranks within strata
- Let $\rho(\mathbf{Z})$ be some way of scoring the instrument settings such that $\rho(\mathbf{p}\mathbf{h}) = \mathbf{p}\rho(\mathbf{h})$ for each $\mathbf{p} \in \Omega$
- The test statistic is $U = \mathbf{q}(\mathbf{Y} - \beta_0(T))^T \rho(\mathbf{Z})$
- For appropriate scores, U can be Wilcoxon's stratified rank sum statistic, the Hodges-Lehmann aligned rank statistic, the stratified Spearman rank correlation, etc

Hypothesis Testing II

- If H_0 were true, $\mathbf{Y} - \beta_0 \mathbf{T} = \mathbf{Y}_C$ would be fixed, not varying with \mathbf{Z} : $\mathbf{q}(\mathbf{Y} - \beta_0 \mathbf{T}) = \mathbf{q}(\mathbf{Y}_C) = \mathbf{q}$ would also be fixed
- If the null is false, $\mathbf{Y} - \beta_0 \mathbf{T} = \mathbf{Y}_C + (\beta - \beta_0) \mathbf{T}$ will be related to the dose \mathbf{T} and related to \mathbf{Z}
- Our test amounts to looking for an absence of a relationship between $\mathbf{Y} - \beta_0 \mathbf{T}$ and \mathbf{Z} .

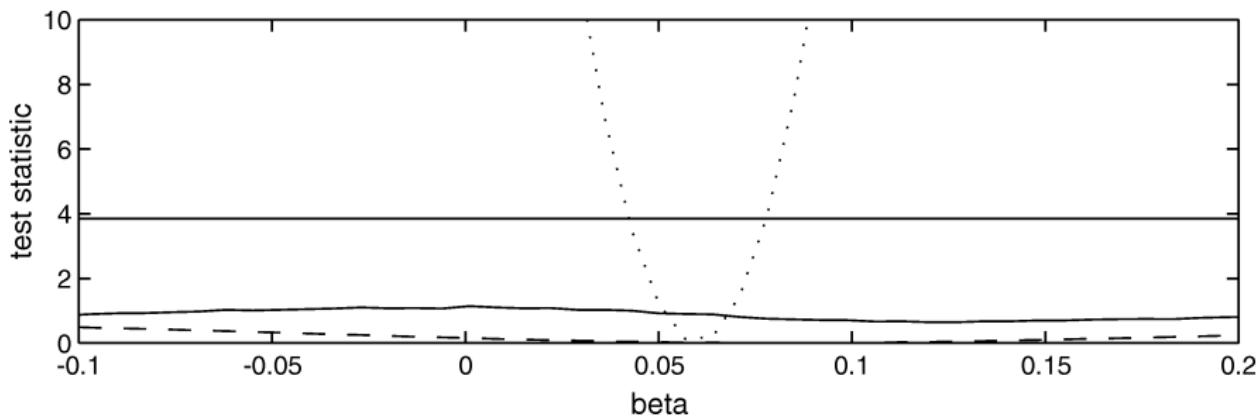
Exact Test

- An exact test computes $\mathbf{q}(\mathbf{Y} - \beta_0 \mathbf{T})$, which is the fixed value $\mathbf{q} = \mathbf{q}(\mathbf{Y}_C)$, in which case $U = \mathbf{q}^T \mathbf{P}\rho(\mathbf{H})$
- The chance that $U \leq u$ under H_0 is the proportion of $\mathbf{p} \in \Omega$ that $\mathbf{q}^T \mathbf{P}\rho(\mathbf{H}) \leq u$

Comparison of instrumental variable estimates with uninformative data

Procedure	95% lower	95% upper
TSLS	0.042	0.078
Permute ranks	-1	1
Permute log-earnings	-1	1

Results with Uninformative Instrument



Dotted line is TSLS; solid line is randomization test using ranks; and dashed line is randomization test using the full observed data

References

This treatment is based on:

- G. W. Imbens and P. Rosenbaum (2005): “Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education,” *Journal of the Royal Statistical Society, Series A*, vol 168(1), 109–126.
- J. Angrist and Krueger (1991): “Does compulsory school attendance affect earnings?” *QJE* 1991; 106: 979–1019.
- Bound, Jaeger, and Baker (1995): “Problems with Instrumental Variables Estimation when the Correlation Between the Instruments and the Endogenous Regressors is Weak,” *JASA* 90, June 1995, 443–450.

Background Reading

- Angrist, J.D. & Krueger, A.B. 2001. Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives*, 15(4), 69-85.

Example: Inc incumbency Advantage

- The purpose is to show that different identification strategies lead to different estimands.
- These differences are often subtle
- Some research designs are so flawed, that even random assignment would not identify the treatment effect of interest
- For details, see: Sekhon and Titiunik “Redistricting and the Personal Vote: When Natural Experiments are Neither Natural nor Experiments” <http://sekhon.berkeley.edu/papers/incumb.pdf>

Inc incumbency Advantage (IA)

- Extensive literature on the Inc incumbency Advantage (IA) in the U.S.
- One of the most studied topics in political science @
- Senior scholars e.g.:
Ansolabehere, Brady, Cain, Cox, Erikson, Ferejohn, Fiorina, Gelman, Jacobson, Nagler, Katz, Kernell, King, Levitt, Mayhew, Niemi, Snyder, Stewart
- Junior/junior senior scholars e.g.:
Ashworth, Bueno de Mesquita, Carson, Desposato, Engstrom, Gordon, Huber, Hirano, Landa, Lee, Prior, Roberts

So much work

“I am convinced that one more article demonstrating that House incumbents tend to win reelection will induce a spontaneous primal scream among all congressional scholars across the nation.”

Charles O. Jones @ (1981)

@

A Cautionary Tale. This should be an easy problem: clear question, a lot of data, and what people thought was a clean identification strategy.

Inc incumbency Advantage (IA)

- The **Personal Inc incumbency Advantage** is the candidate specific advantage that results from the benefits of office holding—e.g:
 - name recognition
 - constituency service
 - public position taking
 - providing pork @
- Different from the **Incumbent Party Advantage**:
 - marginally, voters have a preference for remaining with the same party they had before (D. S. Lee, 2008b)
 - parties have organizational advantages in some districts

Problems with Estimating the IA

- Empirical work agrees on there being an advantage, but much disagreement about the sources of the IA
- Theoretical work argues that selection issues are the cause of the positive estimates of IA
(e.g., G. W. Cox and Katz, 2002; J. Zaller, 1998):

Main concerns: strategic candidate entry and exit, and survival bias

Exploiting Redistricting

- Ansolabehere, Snyder, and Stewart (2000) use the variation brought about by decennial redistricting plans to identify the causal effect of the personal IA
- After redistricting, most incumbents face districts that contain a combination of old and new voters @
- They compare an incumbent's vote share in the new part of the district with her vote share in the old part of the district. @
- They hope to avoid some of the selection issues because the electoral environment is constant @
- They analyze U.S. House elections at the county level from 1872 to 1990
- Others have also used this design and found similar results Desposato and Petrocik (2003); Carson, Engstrom and Roberts (2007).

Using Redistricting to Estimate Incumbency

- We propose a new method for estimating incumbency advantage (IA)
- Previous redistricting designs result in biased estimates of incumbency @
- The wrong potential outcomes were used
- Leads to bias even if voters were redistricted randomly! @
- In practice, a selection on observables assumption must be made
- The selection on observables assumption in prior work is theoretically implausible and fails a placebo test @
- Use redistricting to answer other questions as well—e.g., how voters respond to candidates' race and ethnicity

Where we are Going

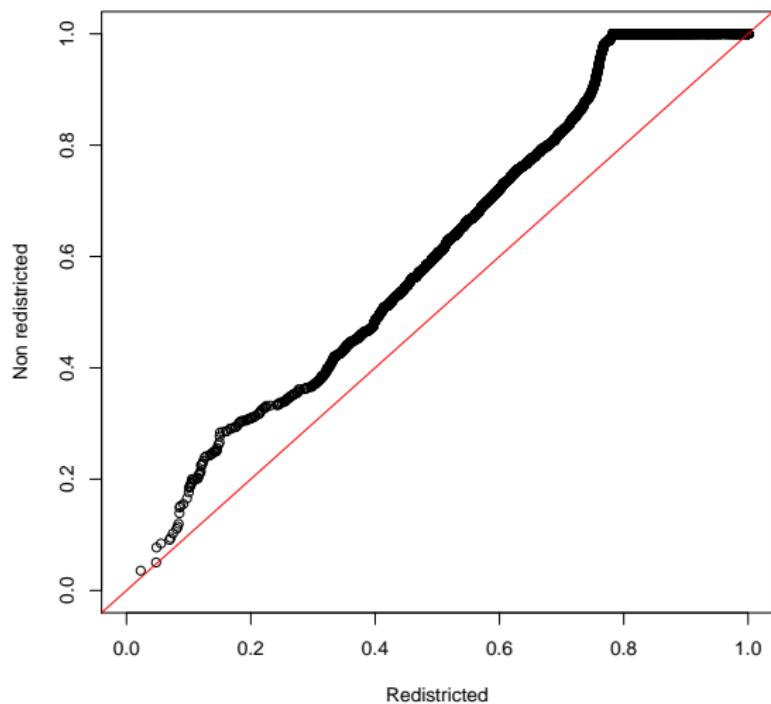
One interpretation of our results:

- No (TX) or little (CA) personal incumbency advantage in U.S. House elections
- Significant and large incumbent party effects @

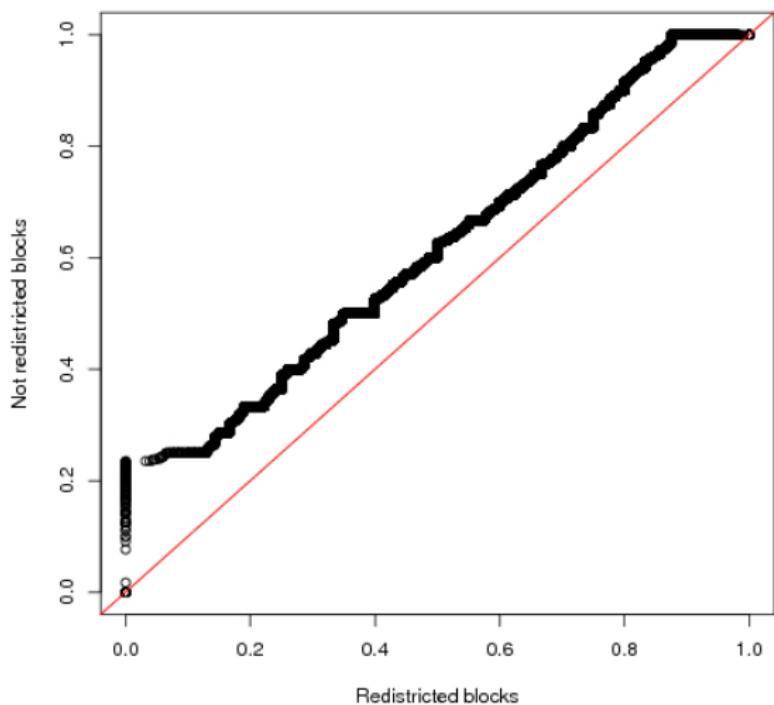
Results are more consistent with:

- Voters learn the type of new incumbent very quickly @
- Unless there is a mismatch between the partisanship of voters and their incumbent. Then, voters take more time to learn the type of their incumbent

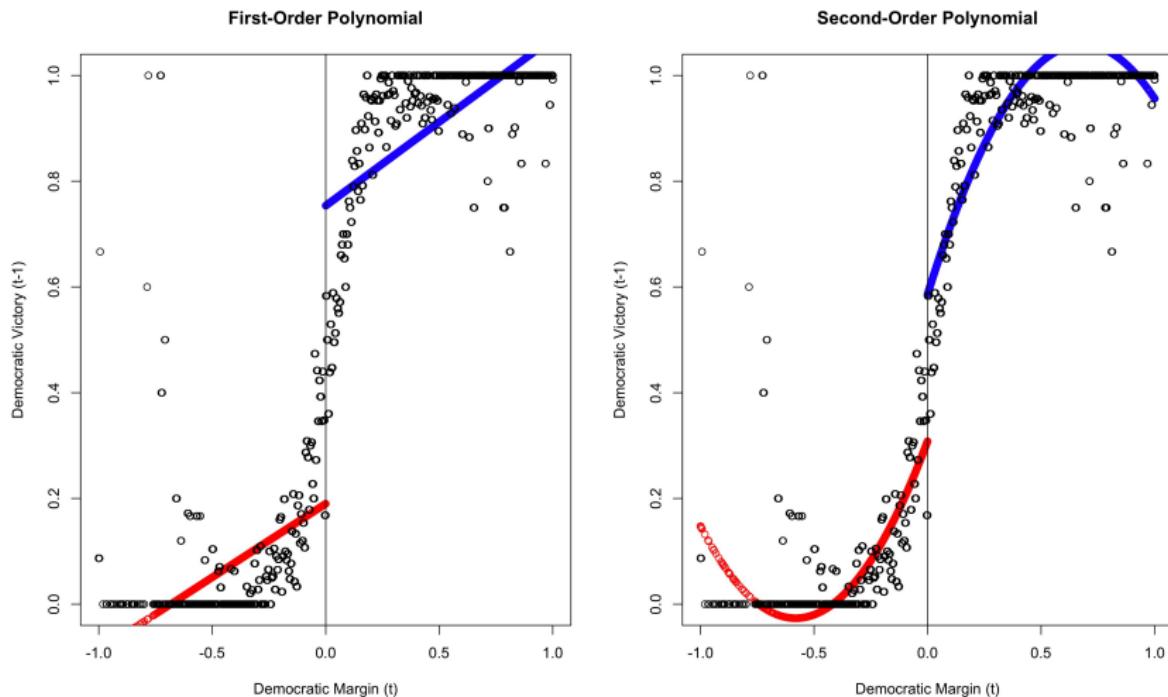
Baseline Vote Share for Incumbent House Member, Texas



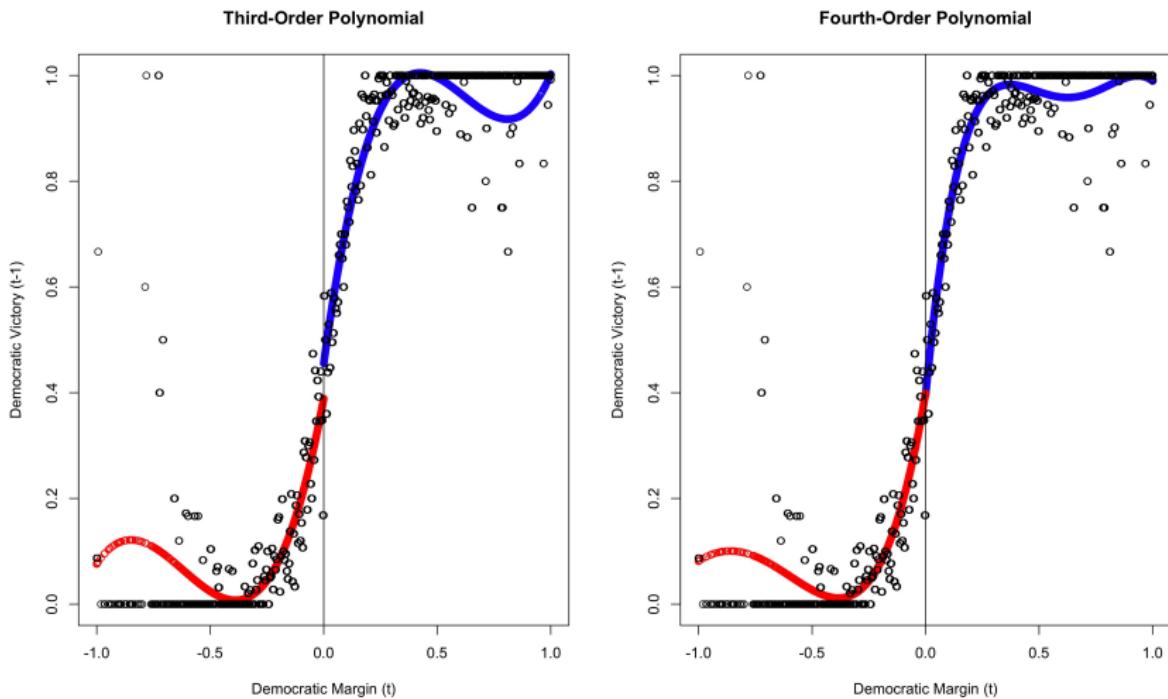
Baseline Vote Share for Incumbent House Member, California



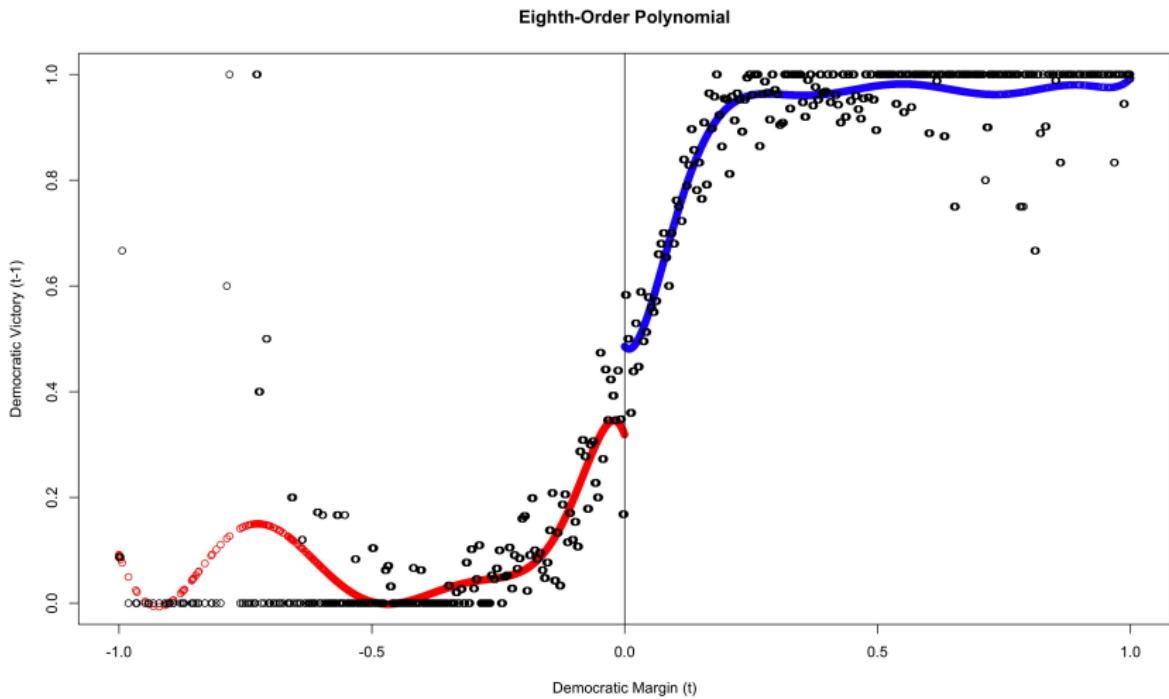
Lagged Democratic Win in RDD: 1st and 2nd-order polynomial



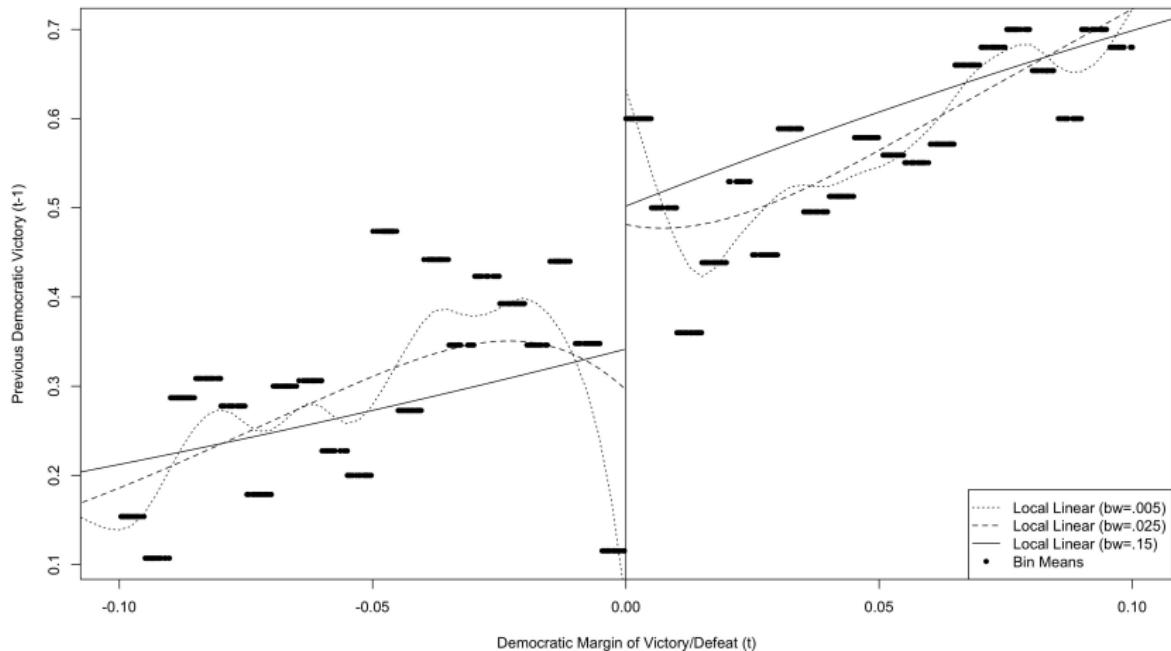
Lagged Democratic Win in RDD: 3rd and 4th-order polynomial



Lagged Democratic Win in RDD: 8th-order polynomial



Lagged Democratic Win in RDD: local regression



The Basic Setup

- Redistricting induces variation in two dimensions:
 - **time**: voters vote both before and after redistricting
 - **cross-sectional**: some voters are moved to a different district while others stay in the district

@

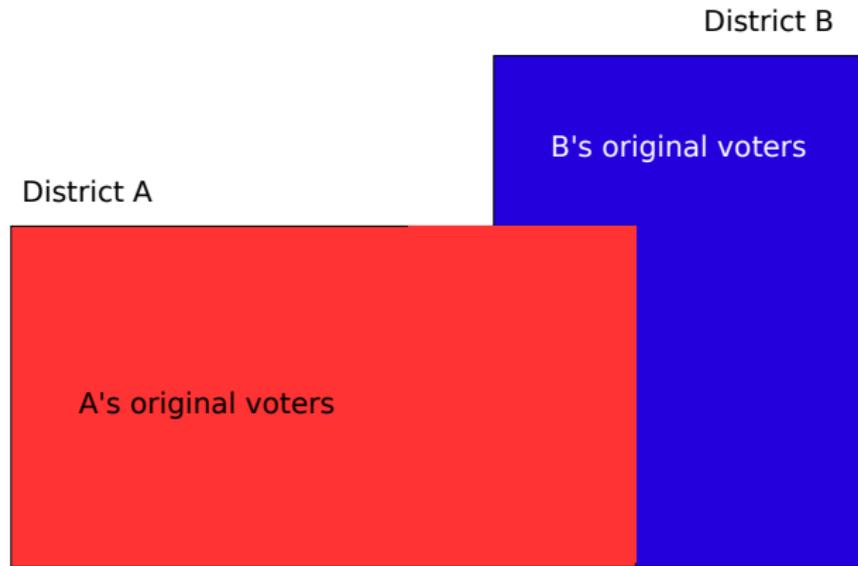
- Can estimate IA by comparing the behavior of voters who are:
 - moved to a new district (**new voters**)
 - to the behavior of voters whose district remains unchanged (**old voters**).

New Voters vs. Old Voters

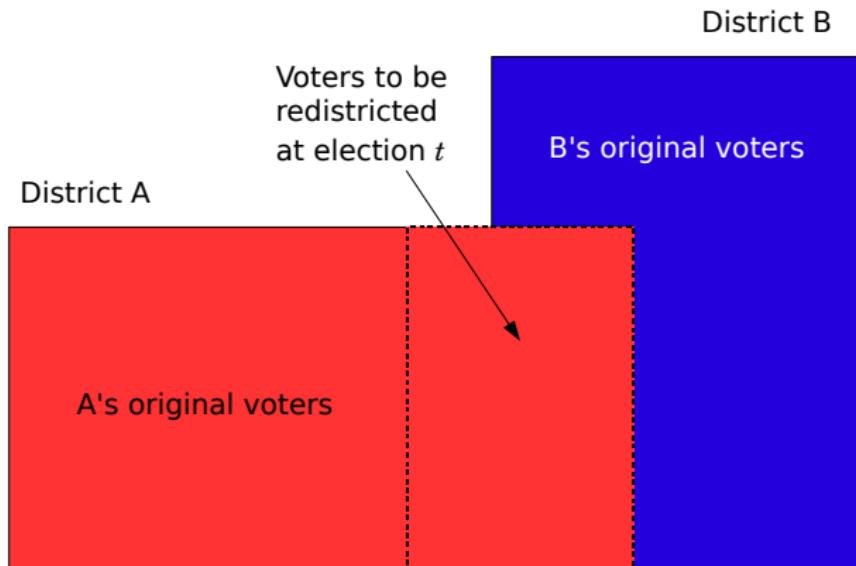
Although this comparison seems intuitive, there are important complications:

- New voters are naturally defined as the voters whose district changes between one election and another @
- BUT there is an ambiguity in the way in which old voters are defined
- Old voters could be either:
 - the electorate of the district to which new voters are moved (**new neighbors**), or
 - the electorate of the district to which new voters belonged before redistricting occurred (**old neighbors**)

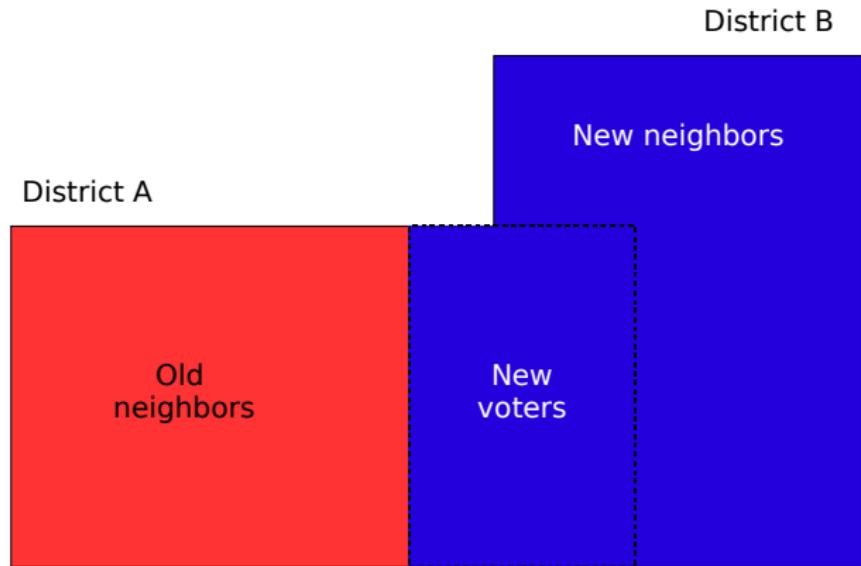
Before one-time redistricting



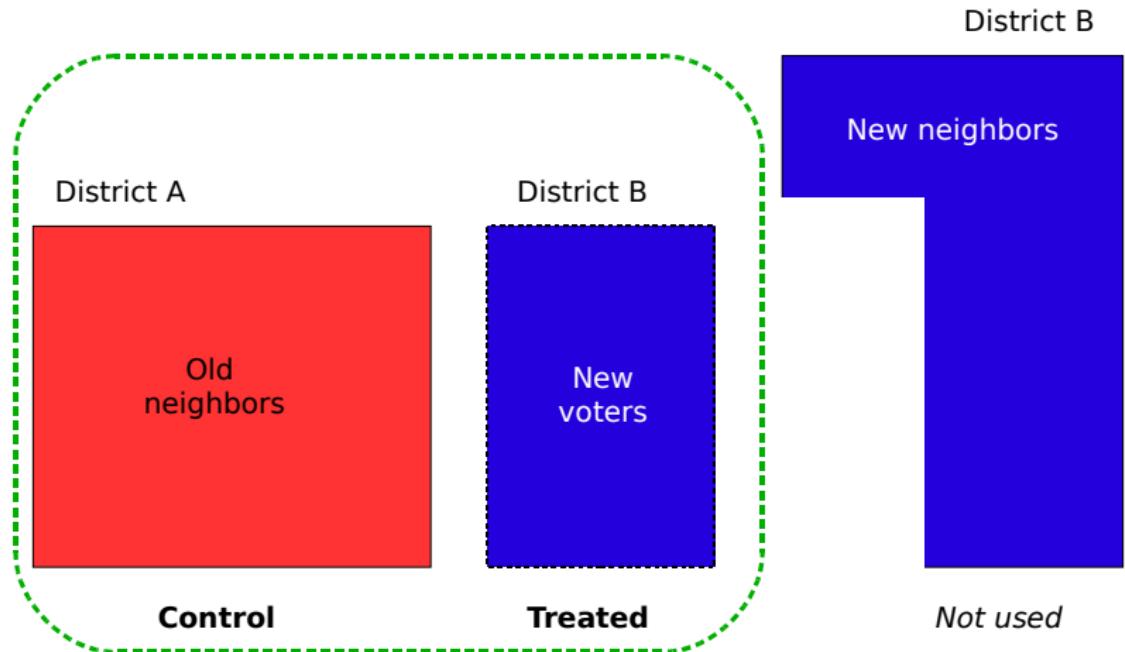
One-time redistricting



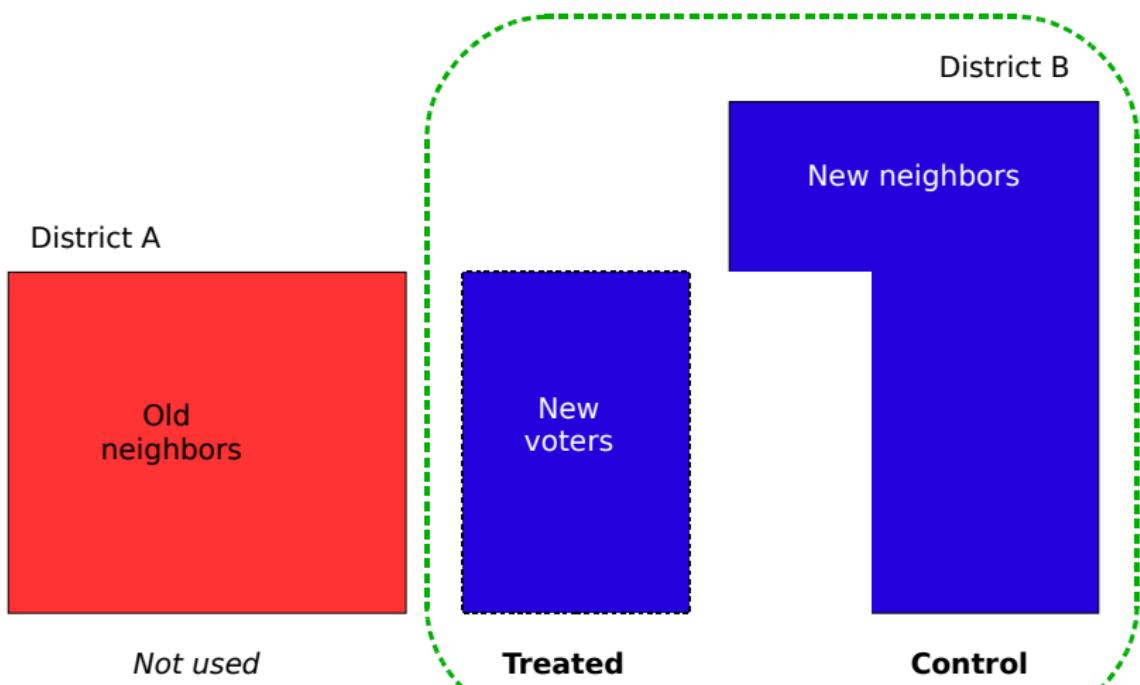
After one-time redistricting



First identification strategy: old-neighbors design



Second identification strategy: new neighbors design

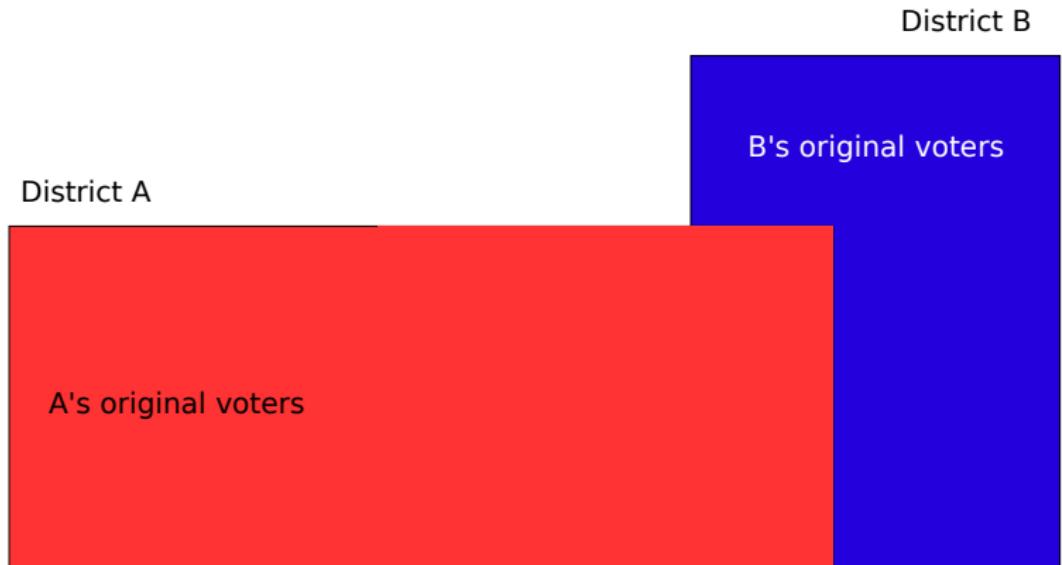


Two Obvious Designs

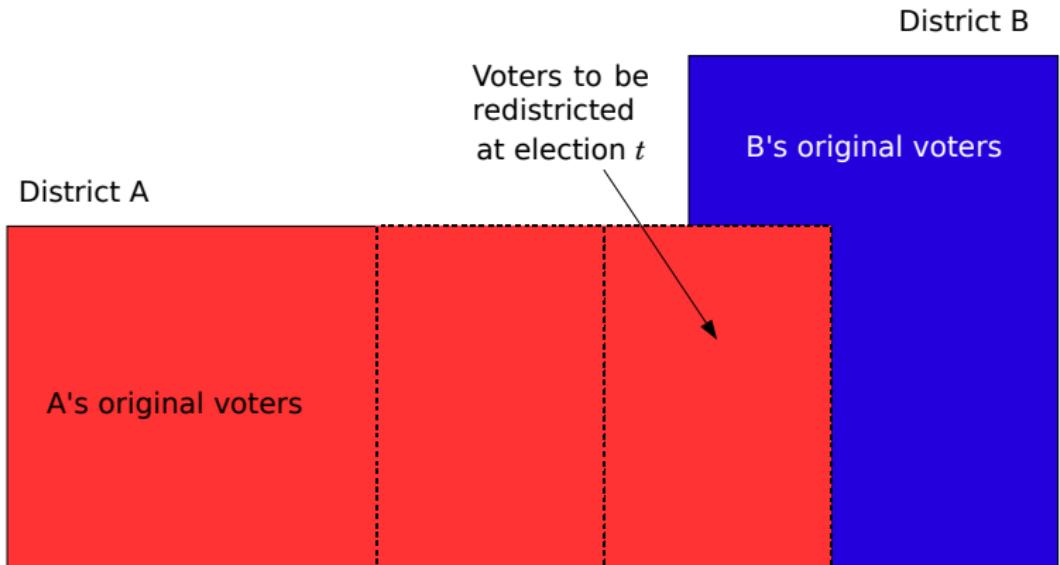
There are then two obvious designs under the randomization where everyone is in a certain district at $t - 1$ and some precincts are randomly moved to another district at t . @

- Compare new voters with old neighbors
- Compare new voters with new neighbors—this is the Anscombe, Snyder, and Stewart (2000) design @
- Randomization ensures exchangeability for the first design, but not the second
- This occurs because the history of new voters with new neighbors is not balanced by randomization

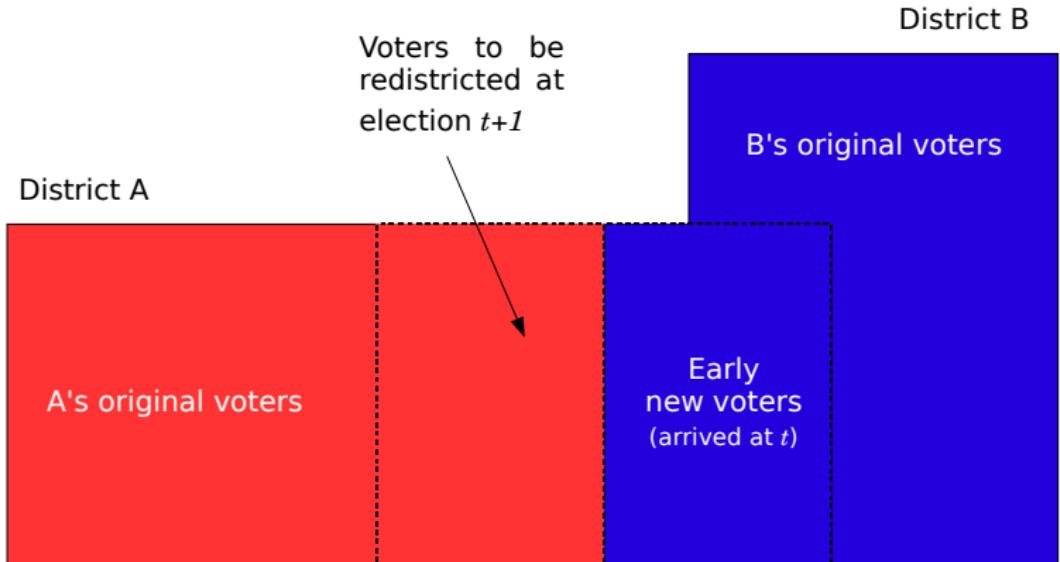
Before two-time redistricting (election $t-1$)



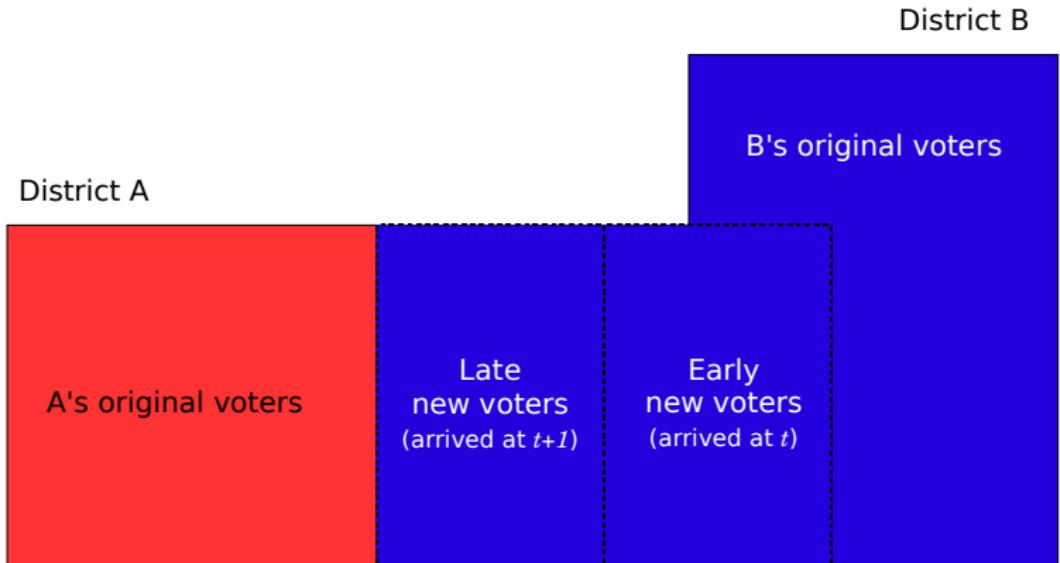
Before two-time redistricting (election $t-1$)



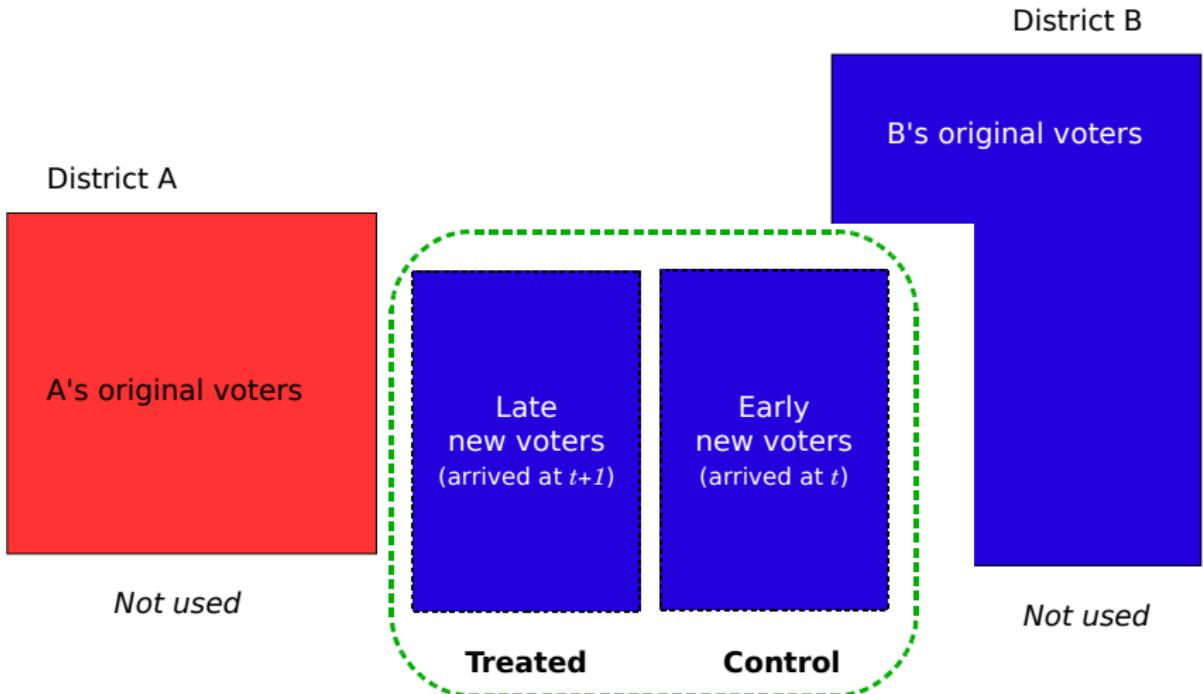
Two-time redistricting (election t)



Two-time redistricting (election $t+1$)



After two-time redistricting (election $t+1$)



More Formally

- Let T_i be equal to 1 if precinct i is moved from one district to another before election t and equal to 0 if it is not moved @
- Let D_i be equal to 1 if precinct i has new voters in its district at t and equal to 0 otherwise @
- Let $Y_{00}(i, t)$ be precinct i 's outcome $T_i = 0$ and $D_i = 0$ it is not moved and does not have new neighbors @
- Let $Y_{01}(i, t)$ be precinct i 's outcome if $T_i = 0$ and $D_i = 1$ the precinct is not moved and has new neighbors @
- Let $Y_{11}(i, t)$ be precinct i 's outcome if $T_i = 1$ and $D_i = 1$ the precinct is moved and has new neighbors

Fundamental Problem of Causal Inference

For each precinct, we observe only one of its three potential outcomes:

$$\begin{aligned} Y(i, t) = & Y_{00}(i, t) \cdot (1 - T_i) \cdot (1 - D_i) + \\ & Y_{01}(i, t) \cdot (1 - T_i) \cdot D_i + \\ & Y_{11}(i, t) \cdot T_i \cdot D_i \end{aligned}$$

@

We can estimate two different ATT's:

$$ATT_0 \equiv E[Y_{11}(i, t) - Y_{00}(i, t) | T_i = 1, D_i = 1] @$$

$$ATT_1 \equiv E[Y_{11}(i, t) - Y_{01}(i, t) | T_i = 1, D_i = 1]$$

Identification of ATT_0

$$ATT_0 \equiv E[Y_{11}(i, t) - Y_{00}(i, t) | T_i = 1, D_i = 1] @$$

Is identified if:

$$E[Y_{00}(i, t) | T_i = 1, D_i = 1] = E[Y_{00}(i, t) | T_i = 0, D_i = 0]$$

ATT_0 requires that voters who stay in A and voters who are moved from A to B would have the same average outcomes if they hadn't been moved.

Identification of ATT_1

$$ATT_1 \equiv E[Y_{11}(i, t) - Y_{01}(i, t) \mid T_i = 1, D_i = 1]$$

@

Is identified if:

$$E[Y_{01}(i, t) \mid T_i = 1, D_i = 1] = E[Y_{01}(i, t) \mid T_i = 0, D_i = 1]$$

ATT_1 requires that voters who are originally in B and voters who are moved from A to B would have the same average outcomes if A 's voters would not have been moved even though they would be in different districts.

Randomization does not imply that B 's old voters are a valid counterfactual for B 's new voters

Identification of ATT_1

Required for identification:

$$E[Y_{01}(i, t) | T_i = 1, D_i = 1] = E[Y_{01}(i, t) | T_i = 0, D_i = 1]$$

Since this is not guaranteed to hold under randomization, one must make the following assumption (even with randomization):

$$E[Y_1(i, t) | T_i = 1, D_i = 1, X] = E[Y_1(i, t) | T_i = 0, D_i = 1, X]$$

where X is a vector of observable characteristics which correct for imbalance on district specific history.

The Best Design: Multiple Redistrictings

Let $W_{i,t+1} = 1$ if precinct i is moved from district A to district B at election $t + 1$.

@

Let $W_{i,t+1} = 0$ if precinct i is moved from A to B at election t .

@

Let $Y_0(i, t + 1)$ denote the outcome of i at election $t + 1$ if $W_{i,t+1} = 0$

@

Let $Y_1(i, t + 1)$ denote the outcome of i at election $t + 1$ if $W_{i,t+1} = 1$

@

The parameter of interest ATT_B is

$$ATT_B \equiv E[Y_1(i, t + 1) - Y_0(i, t + 1) | W_{i,t+1} = 1]$$

The parameter of interest ATT_B is

$$ATT_B \equiv E[Y_1(i, t+1) - Y_0(i, t+1) | W_{i,t+1} = 1]$$

which is identified by

$$E[Y_0(i, t+1) | W_{i,t+1} = 1] = E[Y_0(i, t+1) | W_{i,t+1} = 0]$$

@ By randomization we have

$$E[Y_0(i, t-1) | W_{i,t+1} = 1] = E[Y_0(i, t-1) | W_{i,t+1} = 0]$$

@

Randomization along with the following stability assumption provides identification:

$$\begin{aligned} E[Y_0(i, t+1) - Y_0(i, t-1) | W_{i,t+1} = 1] &= \\ E[Y_0(i, t+1) - Y_0(i, t-1) | W_{i,t+1} = 0] \end{aligned}$$

Selection on Observables

- Redistricting was not actually randomized
- We, like previous work, make a selection on observables assumption:

$$E [Y_0(i, t-1) | W_{i,t+1} = 1, X] = E [Y_0(i, t-1) | W_{i,t+1} = 0, X]$$

@

- We conduct a placebo test which is enabled by our two-redistricting research design @
- We use the multiple redistrictings which were done in Texas between 2002 and 2006. We have merged VTD-level election, registration and census data with candidate data such as quality measures and Nominate.

Placebo Test: Key Covariates

@

- Treated: VTDs to be redistricted in 2004 @
- Control: VTDs to remain in same district in 2004 @
- Placebo test:
 - Match in 2000 Congressional district @
 - Match in 2000 VTD-level covariates: presidential vote, senate vote, house vote, governor vote, turnout, registration, etc. @
 - In 2002 there should be no significant difference between our treated and control groups in outcomes @
- Our set of covariates pass the test
- “Presidential vote” fails the test

Estimation

- Genetic Matching (GenMatch) is used to achieve covariate balance @
- Hodges-Lehmann interval estimation, bivariate overdispersed GLM estimation and Abadie-Imbens CIs are calculated post-matching. Substantive inferences remain unchanged.
- Hodges-Lehmann intervals are presented in the following.

Balance Tests, TX

Variable	Before Matching			After Matching		
	mean diff	D-stat	KS-pval	mean diff	D-stat	KS-pval
Dem Pres. vote '00	.0447	.100	0.00	.00459	.0337	0.953
Dem House vote '00	.159	.305	0.00	.00693	.0344	0.678
Dem House vote '98	.127	.340	0.00	.00585	.0368	0.996
Dem Senate vote '00	.0426	.120	0.00	.00576	.0317	0.846
Dem Governor vote '98	.0305	.0974	0.00	.00510	.0241	0.942
Dem Att. Gen. vote '98	.0353	.141	0.00	.00683	.0358	0.868
Dem Compt. vote '98	.0304	.208	0.00	.00499	.0373	0.994
Voter turnout '00	.0331	.102	0.00	.00607	.0327	0.943
Voter turnout '98	.028	.199	0.00	.0111	.0378	0.235
Registration '00	.0308	.157	0.00	.00736	.0608	0.601

The mean difference are the simple differences between treatment and control, the D-statistic is the largest difference in the empirical QQ-plot on the scale of the variable, and the KS-pvalue is from the bootstrapped Kolmogorov-Smirnov test.

Placebo Tests with All Key Covariates, TX

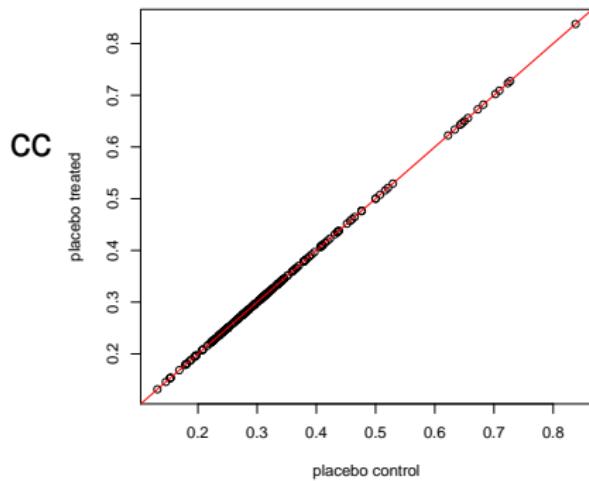
	Estimate	95% CI	p.value
Incumbent vote '02	0.00245	-0.00488 0.00954	0.513
@Turnout '02	0.00334	-0.00443 0.0112	0.412

Placebo Test TX: Presidential vote only

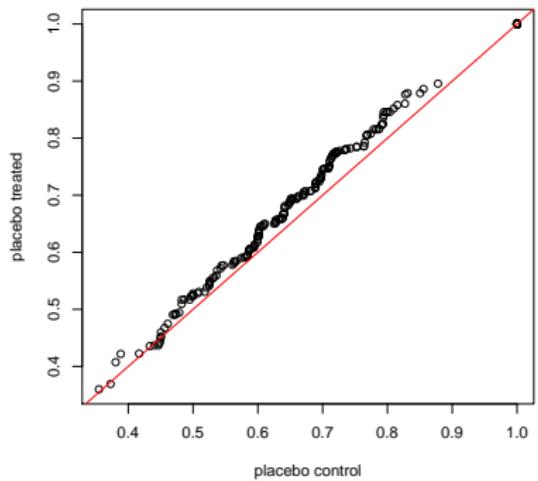
What happens if placebo test is done matching on presidential vote only? @

- Achieve excellent balance on 2000 presidential vote @
- But there is still a significant effect in 2002, before redistricting occurs @
- Past presidential vote is not sufficient to satisfy this placebo test.
- The significant effect estimated in the placebo test includes the actual estimate of Ansolabehere, Snyder, and Stewart (2000) in its confidence interval.

QQ plot:
2000 Presidential Vote



QQ Plot:
2002 Incumbent Vote



Placebo Test TX: Presidential vote only

	Estimate	95% CI	p.value
Incumbent vote '02	0.0285	0.0160 0.0413	0.00
@Turnout '02	0.00246	-0.0180 0.0218	0.831

Incumbency Advantage TX: Same Party

	Estimate	95% CI	p.value
Two-Time Redistricting Design			
Incumbent vote '04	0.00637	-0.00428	0.0177
Incumbent vote '06	0.00843	-0.00938	0.0258
One-Time Redistricting Design			
Incumbent vote '04	0.00214	-0.00807	0.0124
Incumbent vote '06	0.00472	-0.00539	0.0149

Incumbency Advantage TX: Different Party

	Estimate	95% CI	p.value
Incumbent vote '04	0.119	0.0595	0.191
Incumbent vote '06	0.0389	0.00973	0.0692

Incumbency Advantage CA

	Estimate	95% CI	p.value
Same-Party, One-Time Redistricting Design			
Incumb vote '02	0.0219	0.0171	0.0266
Incumb vote '04	0.0240	0.0195	0.0284
Incumb vote '06	-0.0072	-0.0129	-0.0015
Different-Party, One-Time Redistricting Design			
Incumb vote '02	0.1025	0.0925	0.1122
Incumb vote '04	0.1020	0.0932	0.1109
Incumb vote '06	0.0307	0.0186	0.0428

Summary

- Our results are consistent with theoretical arguments that existing positive estimates are plagued by selection problems (Ashworth and Bueno de Mesquita, 2007; G. W. Cox and Katz, 2002; J. Zaller, 1998). @
- Voters learn the type of new incumbent very quickly
- Find a significant incumbent party effect. Consistent with the results of Lee (2008). @
- Finding a positive incumbency effect is much less common outside of the U.S. Some work even documents a negative incumbency effect (Linden, 2004; Uppal, 2005).

A Cautionary Tale

- This should be an easy problem: clear question, a lot of data, and what people thought was a clean identification strategy
- “**Natural experiments**” require careful theoretical and statistical work to make valid inferences @
- Real experiments require that researchers a priori design the study so that randomization will ensure the identification of the causal effect @
- Different ideal experiments imply different identifying assumptions and hence different experimental designs

Hodges-Lehmann estimator

- Assume additive treatment effect @
- Subtract the hypothesized treatment effect from the treated responses, and test hypothesis using these adjusted responses @
- Hodges-Lehmann (HL) estimate is the value that when subtracted from the responses yields adjusted responses that are free of treatment effect, in the sense that the (Wilcoxon Signed Rank) test statistic equals its expectation in the absence of treatment effect @
- Wilcoxon Signed Rank test statistic: form matched treated-control pairs, rank absolute value of treated-control differences within matched pairs, sum ranks for pairs in which treated response is higher than control response

Effect of Incumbent Ethnicity on Turnout: White to Hispanic Incumbent

	Estimate	95% CI	p.value
Incumbent vote '04	-0.00674	-0.0165 0.00551	0.233
Incumbent vote '06	0.0167	-0.0075 0.0391	0.177
Turnout '04	-0.0406	-0.0699 -0.00696	0.0187
Turnout '06	-0.0139	-0.0413 0.0151	0.340

Estimating the Effect of Incumbent Ethnicity on Registration: White to Hispanic Incumbent

	Estimate	95% CI	p.value	
Hispanic Reg '04	-0.0197	-0.0338	-0.00692	0.00335
Hispanic Reg '06	-0.0279	-0.0429	-0.0138	0.000143
Non-Hispanic Reg '06	-0.00233	-0.0393	0.0332	0.912
Non-Hispanic Reg '04	-0.00449	-0.0381	0.0256	0.784

Effect of Incumbent Race on Turnout: White to White Incumbent

	Estimate	95% CI	p.value	
Incumbent vote '04	0.00555	−0.00503	0.0166	0.268
Incumbent vote '06	0.012	−0.000877	0.0267	0.071

Inverse Probability Weighting

- Inverse probability weighting: Treated observations are weighted by $\frac{1}{P(T_i = 1 | X_i)}$ and control observations by $\frac{1}{1 - P(T_i = 1 | X_i)}$
- This is a property of nature, not simply our creation.
- Example: imagine propensity score is 0.2, for every treated subject there are 9 control subjects.
- What would happen if everyone received treatment? What happens to the other four control subjects once they are assigned treatment?

Three Parts, PRE-Intervention

The PRE-INTERVENTION distribution can be decomposed into 3 conditional probabilities:

$$P(Y = 1, T = 1, X = x) = P(Y = 1 | T = 1, X = x) \times P(T = 1 | X = x) \times P(X = x) @$$

$P(Y = 1 | T = 1, X = x)$ describes how the outcome depends on treatment and X .

@

$P(T = 1 | X = x)$ describes how subjects choose treatment prior to intervention

@

$P(X = x)$ describes the prior distribution of the covariates

Three Parts, POST-Intervention

The POST-Intervention distribution is denoted by $P'::$

$$P'(Y = 1, T = 1, X = x) = P'(Y = 1 \mid T = 1, X = x) \times \\ P'(T = 1 \mid X = x) \times P'(X = x) @$$

$P'(X = x) = P(X = x)$ because X is pretreatment

@ $P'(Y = 1 \mid T = 1, X = x) = P(Y = 1 \mid T = 1, X = x)$
because of the selection on observables assumption.

@

$P'(T = 1 \mid X = x) \neq P(T = 1 \mid X = x)$. Why?

Treatment distribution

After the intervention, $P'(T = 1 | X = X)$ is either 0 or 1.

@Therefore, the post-intervention distribution is:

$$P'(Y = 1, T = 1, X = x) = \quad (23)$$

$$P(Y = 1 | T = 1, X = x) \times P(X = x) \odot$$

Equivalently:

$$P'(Y = 1, T = 1, X = x) = \frac{P(Y = 1, T = 1, X = x)}{P(T = 1 | X = x)} \quad (24)$$

Eq. 23 leads to regression. Eq. 24 leads to IPW. Since (23) and (24) are asymptotically equivalent, they must be identified by the same assumptions.

IPW Readings

- Freedman, D.A. and Berk, R.A. (2008). Weighting Regressions by Propensity Scores. *Evaluation Review* 32,4 392-409.
- Kang, J.D.Y. and Schafer, J.L. (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science* 22,4 523-539.
- Robins, J., Sued, M., Lei-Gomez, Q. and Rotnitzky (2007). Comment: Performance of Double-Robust Estimators When “Inverse Probability” Weights are Highly Variable. *Statistical Science* 22,4 544-559.

Collaborative Targeted Maximum Likelihood

- Double robust estimator use knowledge of Y , but they show poor properties when the pscore is close to 0 or 1.
- Even if the pscore model is correct, one is often better off not adjusted by the pscore and just modelling Y —e.g., [Freedman and Berk 2008 @](#)
- Even true for bounded doubly robust estimators
- Influence curve may be bounded, but MSE is still increased by using the pscore [@](#)
- C-TMLE helps ([van der Laan and Gruber 2009](#)); extends targeted ML
- One can extend C-TMLE to trim data as matching does—[adjusted collaborative targeted maximum likelihood](#)
- Incorporate various ways of trimming including that of [Crump, Hotz, Imbens, and Mitnik \(2006\)](#)

The Problem

- $O = (X, T, Y)$
X are confounders
T is the binary treatment indicator (or censoring model)
Y is the outcome @
- Estimand: $\psi = E_X [E[Y | T = 1, X] - E[Y | T = 0, X]]$
- Two nuisance parameters:
 $Q(T, X) \equiv E[Y | T, X]$
 $g(T, X) \equiv P(T | X)$ @
- Inverse probability weighting:

$$\psi^{IPTW} = \frac{1}{n} \sum_{i=1}^n [I(T_i = 1) - I(T_i = 0)] \frac{Y_i}{g_n(T_i, X_i)}$$

Alternatives

- Double Robust-IPTW

$$\begin{aligned}\psi^{DR} = & \frac{1}{n} \sum_{i=1}^n \frac{[I(T_i = 1) - I(T_i = 0)]}{g_n(T_i | X_i)} (Y_i - Q(T, X)) \\ & + \frac{1}{n} \sum_{i=1}^n Q_n(1, X_i) - Q_n(0, X_i)\end{aligned}$$

@

- Collaborative Targeted Maximum Likelihood

$$\psi^{C-TMLE} = \frac{1}{n} \sum_{i=1}^n Q_n^*(1, X_i) - Q_n^*(0, X_i)$$

where Q_n^1 adjusts a preliminary non-parametric estimate of Q_n^0 by

$$\epsilon_n^k = \arg \max P_n Q_g^{k-1}(P_n)(\epsilon)$$

All estimation via machine learning—e.g., a super learner that is based on a convex combination of non-parametric estimators with \sqrt{n} oracle properties.

The Collaborative Part

Q_n^1 adjustes a preliminary non-parametric estimate of Q_n^0 by

$$\epsilon_n^k = \arg \max P_n Q_g^{k-1}(P_n)(\epsilon)$$

Define our update based on $g()$ as:

$$h_1 = \left(\frac{I[T=1]}{g_1(1|X)} - \frac{I[T=0]}{g_1(0|X)} \right)$$

ϵ_1 is fitted by regression Y on h with offset Q_n^0 , with the influence curve

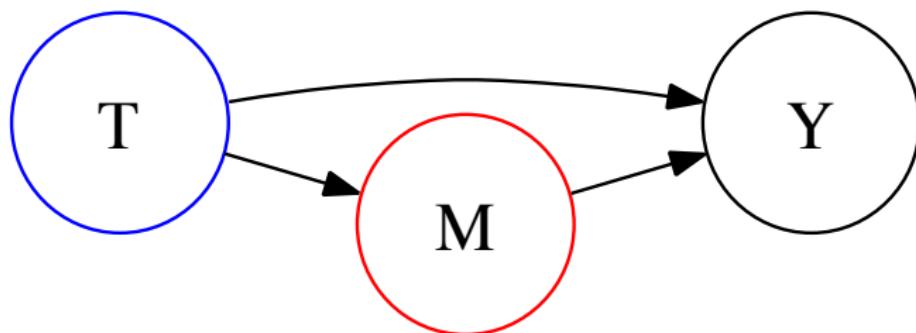
$$IC(P_0) = D^*(Q, g_0, \psi_0) + IC_{g_0}$$

The Adjusted Part

- If the assignment part ($g()$) near the bounds, trim as in [Crump, Hotz, Imbens, and Mitnik \(2006\)](#).
- maximize the efficiency of the estimand as determined by the influence—e.g., return the lowest MSE estimand closest to ATE. @
- estimate all parts by super learner: combine non-parametric estimators via maximum likelihood cross validation @
- downside: understandable balance checks are not possible since $g()$ now only corrects for the residuals of $Q()$

Causal Mediation

Examines the roles of intermediate variables that lie in the causal paths between treatment and outcome variables:



T is a treatment, M is a mediator, Y is an outcome

Mediation Analysis is Popular

- Researchers want to know how and why treatment has an effect not simply that it does
- Very popular—e.g., Baron and Kenny (1986), cited 19,988 times. Linear version goes back to Haavelmo (1943) and Simon (1943)
- New assumptions have been proposed that provide nonparametric identification—e.g., Robins and Greenland (1992), Pearl (2001), Robins (2003), Petersen, Sinisi and van der Laan (2006), Hafeman and VanderWeele (2010), Imai et al. (2010).

Mediation Analysis is Problematic

- These assumptions are usually implausible
- I provide analytical results which demonstrate that these assumptions are fragile: when the assumptions are slightly weakened, identification is lost
- Troubling because researchers claim it is essential to estimate causal mediation effects to test substantive theories

Example: Religious Practice in India

- Joint work with Pradeep Chhibber @
- Do Muslims political leaders have greater trust in a political context than other religious leaders?
- Can leaders use religious symbols to politically mobilize supporters?
- We are especially interested in differences between Muslims and Hindus
- Religious practice differs sharply between Muslims and Hindus, and between Sunnis and Shiites
- Conjecture: Islamic religious leaders, particularly Sunni, have greater political role than Hindu religious leaders

Why People Care?

Current Question:

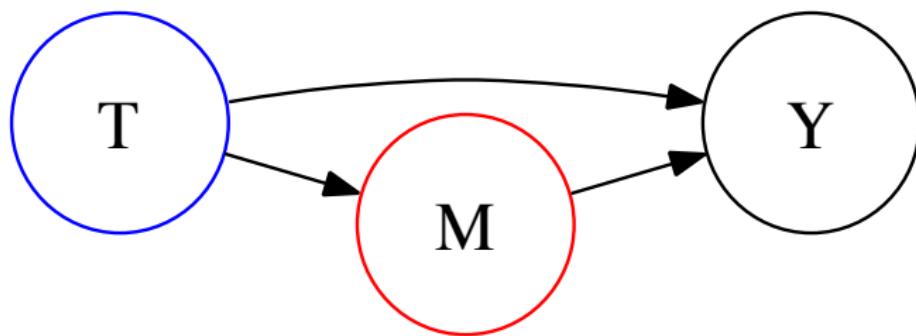
- Concern about the role of Imams in radicalization
- This role may vary across different Islamic groups and contexts

@

Historical Question:

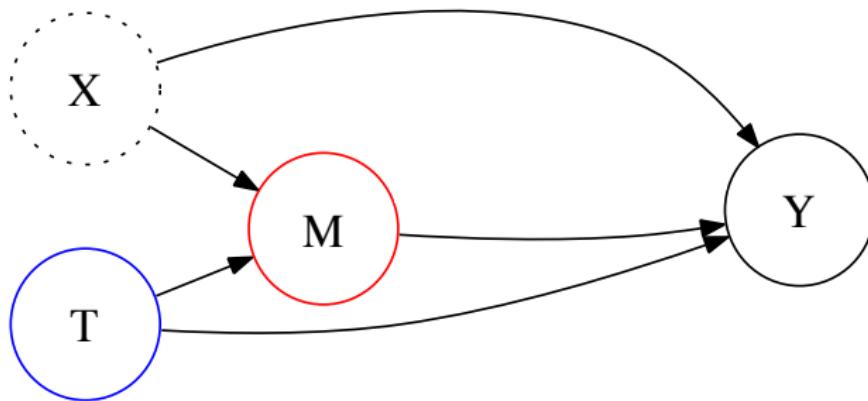
- Long running debate on the political role of Imams. Islamic organization was never centralized—no Pope.
- But Christians and Muslim both regularly launched cross-continental crusades in the last millennium.

Simple Graph



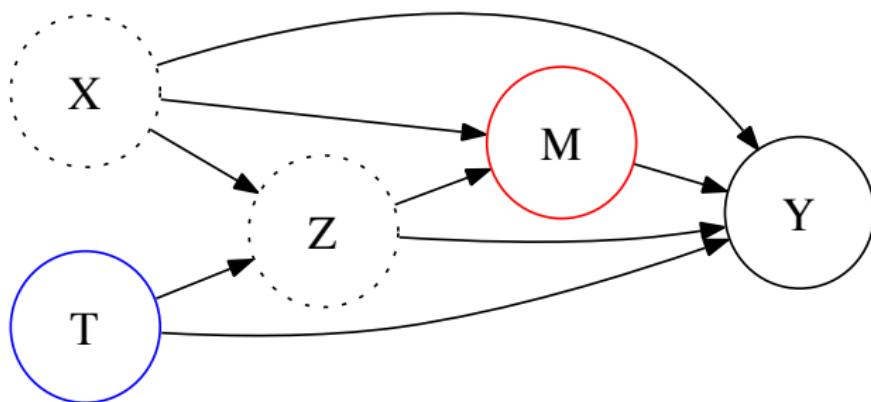
T is a treatment, M is a mediator, Y is an outcome

Graph with Pre-Treatment Variables



X are possibly unobserved pre-treatment variables

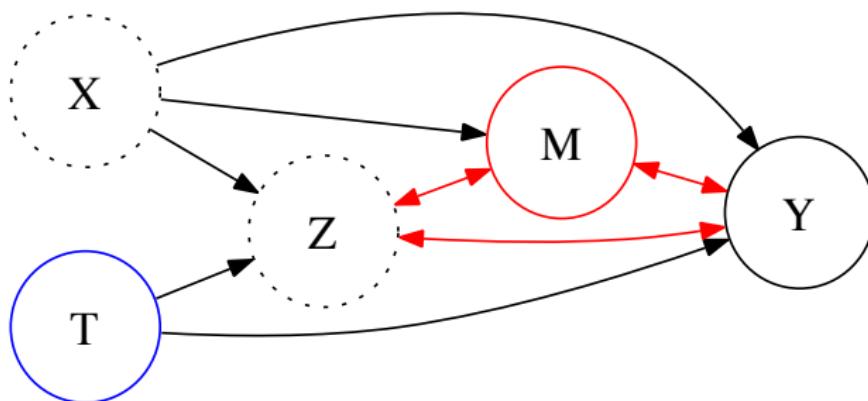
Graph with Other Moderators



X are possibly unobserved pre-treatment variables

Z are possibly unobserved post-treatment variables

We Still Have Simplifications



X are possibly unobserved pre-treatment variables

Z are possibly unobserved post-treatment variables

Some Notation

- Binary treatment: $T_i \in \{0, 1\}$
- Mediator: $M_i \in \mathcal{M}$
- Outcome: $Y_i \in \mathcal{Y}$
- Pre-treatment covariates: $X_i \in \mathcal{X}$
- Post-treatment covariates: $Z_i \in \mathcal{Z}$
- Potential outcomes (standard case): Y_{i1}, Y_{i0} ,
$$Y_i = Y_{i1}T + Y_{i0}(1 - T)$$
- Potential mediators: $M_i(t)$ where $M_i = M_i(T_i)$
- Potential outcomes: $Y_i(t, m)$ where $Y_i = Y(T_i, M_i(T_i))$

Defining Causal Mediation Effects

- Total Causal Effect:

$$\tau_i \equiv Y_i(1, M_i(1)) - Y_i(0, M_i(0))$$

@

- Natural Indirect Effect:

$$\delta_i(t) \equiv Y_i(t, M_i(1)) - Y_i(t, M_i(0))$$

Causal effect of the change in M_i on Y_i that *would be* induced by treatment while holding treatment constant

@

- Controlled Effect of the Mediator:

$$\text{controlled} \equiv Y_i(t, m) - Y_i(t, m')$$

$$m \neq m'$$

Defining Direct Effects

- Direct effects:

$$\zeta_i(t) \equiv Y_i(1, M_i(t)) - Y_i(0, M_i(t))$$

- Causal effect of T_i on Y_i holding mediator constant at its potential value that would be induced when $T_i = t$: @the natural direct effect (Pearl 2001) or the pure/total direct effect (Robins and Greenland 1992). @
- Total effect equals mediation (indirect) effect + direct effect:

$$\tau_i = \delta_i(t) + \zeta_i(1 - t) = \frac{1}{2} \{ \delta_i(0) + \delta_i(1) + \zeta_i(0) + \zeta_i(1) \}$$

Identification of Average Causal Mediation Effect

- Average Causal Mediation Effect (ACME):

$$\bar{\delta}(t) \equiv \mathbb{E}(\delta_i(t)) = \mathbb{E}\{Y_i(t, M_i(1)) - Y_i(t, M_i(0))\}$$

- Issue: we can never observe $Y_i(t, M_i(1-t))$ @

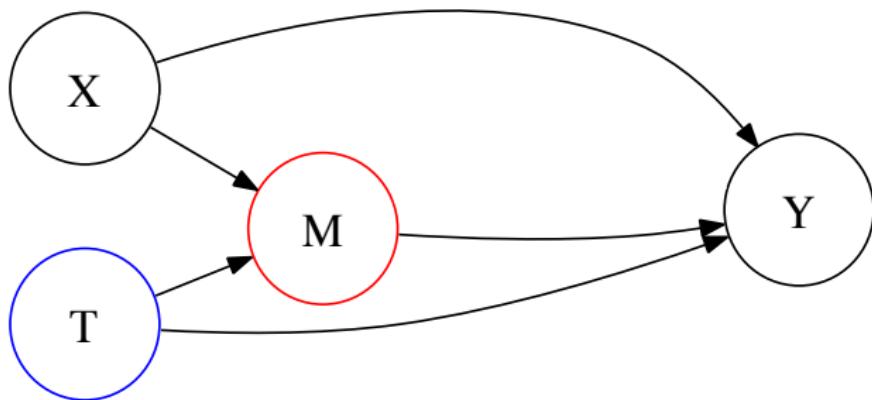
- One assumption, Sequential Ignorability (e.g., Imai et al. 2010):

$$\{Y_i(t', m), M_i(t)\} \perp\!\!\!\perp T_i \mid X_i = x \quad (25)$$

$$Y_i(t', m) \perp\!\!\!\perp M_i(t) \mid T_i = t, X_i = x, \quad (26)$$

for $t, t' = 0, 1; 0 < \Pr(T_i = t \mid X_i = x)$ and
 $0 < p(M_i = m \mid T_i = t, X_i = x), \forall x \in X$ and $m \in M$

Sequential Ignorability Graph



Pearl's (2001) Version of Sequential Ignorability

In order to identify $\bar{\delta}(t^*)$:

$$Y_i(t, m) \perp\!\!\!\perp M_i(t^*) \mid X_i = x, \quad (27)$$

② $p(Y_i(t, m) \mid X_i = x)$ and (28)

$p(M_i(t^*) \mid X_i = x)$ are identifiable,

for all $t = 0, 1$.

Other Post-Treatment Variables

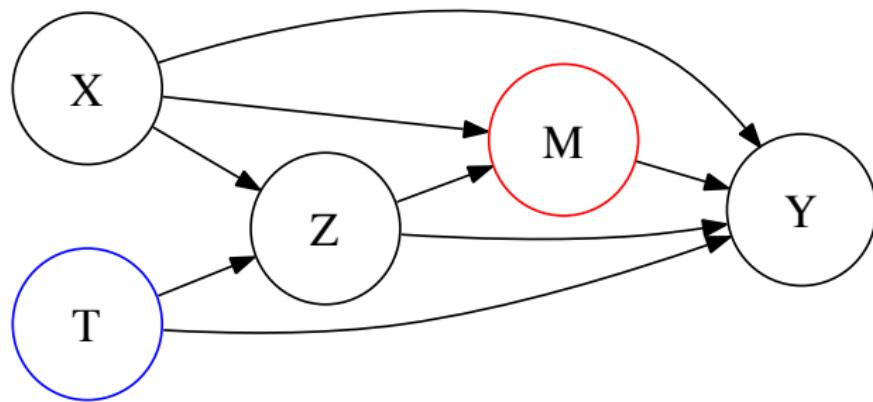
$$\begin{aligned} \{Y_i(t', m), M_i(t)\} &\perp\!\!\!\perp T_i \mid X_i = x \\ Y_i(t, m) &\perp\!\!\!\perp M_i(t) \mid T_i = t, Z_i = z, X_i = x \end{aligned} \tag{29}$$

Equation 29 is not sufficient to identify ACME without further assumptions. Robins' (2003) "no-interaction" assumption:

$$Y_i(1, m) - Y_i(0, m) = B_i,$$

where B_i is a random variable independent of m

Post-Treatment DAG



Nonparametric Bounds and Sensitivity Tests

- Imai et al. (2010) derive sharp nonparametric bounds for ACME when their particular assumption does not hold
- The nonparametric bounds are uninformative when the assumptions are weakened
- I derive **nonparametric** sensitivity tests to allow for partial weakening, and the bounds for ACME for the Robins and Greenland (1992) assumption.

Nonparametric Bounds

- Assume randomization of T :
 $\{Y_i(t', m), M_i(t)\} \perp\!\!\!\perp T_i$
- For Sequential Ignorability do not assume (26):

$$Y_i(t', m) \perp\!\!\!\perp M_i(t) \mid T_i = t, X_i = x$$

- For Robins (2003), no longer assume:

$$Y_i(1, m) - Y_i(0, m) = B_i,$$

but still assume:

$$Y_i(t, m) \perp\!\!\!\perp M_i(t) \mid T_i = t, Z_i = z, X_i = x$$

Nonparametric Bounds

In the case of binary M, Y :

$$\max \left\{ \begin{array}{l} -P_{001} - P_{011} \\ -P_{000} - P_{001} - P_{100} \\ -P_{011} - P_{010} - P_{110} \end{array} \right\} \leq \bar{\delta}(1) \leq \min \left\{ \begin{array}{l} P_{101} + P_{111} \\ P_{000} + P_{100} + P_{101} \\ P_{010} + P_{110} + P_{111} \end{array} \right\}$$

where $P_{ymt} \equiv \Pr(Y_i = y, M_i = m \mid T_i = t), \forall y, m, t \in \{0, 1\}$

the bounds will always include zero

Sensitivity Tests

- The identification assumptions ensure that within each treatment group the mediator is assigned independent of potential outcomes:

$$\frac{\Pr(Y_i(t, m) \mid M_i = 1, T_i = t')}{\Pr(Y_i(t, m) \mid M_i = 0, T_i = t')} = 1$$

$\forall t, m$

- Sensitivity test: calculate sharp bounds for a fix deviation from this ratio

A Series of Experiments

- Measure reported trust in political leaders, randomize religious versus secular leaders
- Prompt people to vote, randomize religious versus secular prompt. Measure voter turnout and vote.
- Conduct experiments on different groups (Sunni/Shiite, Hindus, Sikhs) and at different locations (majority/minority)
@
- Natural experiment: Iranian Shiite Imams in Afghanistan
- More Experiments: Governments

Example Photos



Simple Trust Experiment

At a recent meeting celebrating India's democracy this leading politician (show photo) said:

"Politicians like me from different political parties try hard to represent the interests of the people who support us and vote for us."

Do you think he can be trusted?

Trust in Uttar Pradesh

	Muslim	Hindu
UP, Aligarh		
Estimate	15.1%	3.79%
p-value	0.000686	0.387
UP, Kanpur		
Estimate	12.6%	-1.20%
p-value	0.0118	0.791

Tamil Nadu

	Muslim	Hindu
Estimate	12.1%	-5.09%
p-value	0.02	0.21

Shiite Muslims

Uttar Pradesh, Lucknow

	Muslim	Hindu
Estimate	4.80%	4.72%
p-value	0.337	0.336

Get-Out-The-Vote (GOTV) Experiment

- 2009 general election for the 15th Lok Sabha (April/May)
- GOTV Experiment with three arms:
 - **control**: no contact prior to the election
 - **religious**: receive GOTV appeal with religious picture
 - **secular**: receive GOTV appeal with secular picture

Non-Religious Leader Appeal

This learned person has written a book. The book says that Politics affects whether the government looks after the interests of people like you and the interests of your community. He urged everyone to VOTE!

He wrote that if you as a citizen want to have your input in making politics and government work for your community, you need to VOTE in order to send a message to those in power. Your interests and the interests of your community will not be attended to if you do not VOTE.

GOTV Results: Hindu Vote Proportions

	Religious	Secular	Control
Bharatiya Janata Party	0.44	0.32	0.36
Janata Dal–S	0.22	0.42	0.27
Congress	0.28	0.22	0.29

GOTV Results: Muslim Vote Proportions

	Religious	Secular	Control
Bharatiya Janata Party	0.20	0.18	0.21
Janata Dal-S	0.18	0.03	0.26
Congress	0.59	0.70	0.44
Bahujan Samaj Party	0.00	0.06	0.08

Other Experiments: Triangulation

- Iranian Shiite Imams in Afghanistan, a natural experiment during Muharram
- Public Safety Canada: Doing things Berkeley won't

Conclusion

- Triangulation will never be logically sufficient
- Many moving parts: political trust, voting, the strategic relationship between political parties, majority/minority, Shiite vs. Sunni, levels of violence, media saturation

@

- The temptation of reductionism is widespread: logical positivism, a common language in which all scientific propositions can be expressed
- Causality and manipulation are separable concepts

Manipulation and Reductionism

- A good Mantra: “**No causation without manipulation**” @
- But this does not imply the Dogma of **reductionism** common in the literature @
- Classic Example:
 - **Incoherent:** “What is the causal effect of being white on income/grades/happiness?”
 - But this does not imply that race does not play a role in generating these outcomes @
- We test theories of how different religions are organized. Do Imams have a greater political role than other religious leaders? Are they more trusted?

What is Reductionism?

- **Extreme Reductionism:** every scientific sentence has a full translation in sense-datum
- **Moderate Reductionism:** each scientific sentence has its own separate empirical content @
- **Moderate Holism:** a scientific sentence need not imply empirical consequences by itself. A bigger cluster is usually needed.
- Classic Example: Copernican vs. Ptolemaic system.
@Better example: Āryabhaṭa (500CE)—diurnal motion of the earth, elliptic orbits calculated relative to the sun, correct explanation of lunar and solar eclipses, but still geocentric.

What's the Problem?

- Some want to weaken the Mantra of “**No causation without manipulation**” in order theorize about race, gender, religion
@
- Others justify estimating causal mediation effects because they are needed to test theories—e.g., natural direct effect, pure/total direct effect.
- Assumptions needed to estimate natural mediation effects are usually implausible.
- Good work on the many possible assumptions need to identify the mediation effect—e.g., Robins and Greenland (1992); Pearl (2001); Robins (2003); Petersen, Sinisi, van der Laan (2006); Imai, Keele, Tingley (2010), ...

Keep the Mantra; Drop the Dogma

- The central issue is key parts of a scientific theory may not touch data in the way reductionism suggests.
- The Mantra should be kept: we need to be clear about what we are estimating.
- But our theories can have quantities that are not manipulatable, even in theory. @
- We make progress the way science always has: finding cases where the theories make different predictions, falsification, triangulate, etc.

Philosophy of science is about as useful to scientists as
ornithology is to birds.

—Richard Feynman

Some Citations

- Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* 51 1173-1182.
- Imai, Kosuke, Luke Keele, and Teppei Yamamoto. (2010). "Identification, Inference, and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science*, Vol. 25, No. 1 (February), pp. 51-71.
- Pearl, J . (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (J. S. Breese and D. Koller, eds.) 411-420. Morgan Kaufman, San Francisco, CA.
- Pearl, J. (2010). An introduction to causal inference. *Int. J. Biostat.* 6 Article 7.
- Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In *Highly Structured Stochastic Systems* (P. J. Green, N. L. Hjort and S. Richardson, eds.) 70-81. Oxford Univ. Press, Oxford. MR2082403
- Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 3 143-155.

References I

-  Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010). “Synthetic control methods for comparative case studies: Estimating the effect of Californiaâs tobacco control program”. In: *Journal of the American Statistical Association* 105.490.
-  Abadie, Alberto and Guido Imbens (2006). “Large Sample Properties of Matching Estimators for Average Treatment Effects”. In: *Econometrica* 74.1, pp. 235–267.
-  Angrist, J and AB Krueger (1991). “Does compulsory school attendance affect earnings?” In: *Quarterly Journal of Economics* 106, pp. 979–1019.
-  Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin (1996). “Identification of Causal Effects Using Instrumental Variables”. In: *Journal of the American Statistical Association* 91.434, pp. 444–455.

References II

-  Ansolabehere, Stephen, James M. Snyder, and Charles Stewart (2000). "Old Voters, New Voters, and the Personal Vote: Using Redistricting to Measure the Incumbency Advantage". In: *American Journal of Political Science* 44.1, pp. 17–34.
-  Aronow, Peter M, Donald P Green, and Donald KK Lee (2014). "Sharp bounds on the variance in randomized experiments". In: *The Annals of Statistics* 42.3, pp. 850–871.
-  Ashworth, Scott and Ethan Bueno de Mesquita (2007). "Electoral Selection, Strategic Challenger Entry, and the Incumbency Advantage". [Working Paper](#).
-  BBC (2003). "How to make a perfect cuppa". June 25. URL: <http://news.bbc.co.uk/1/hi/uk/3016342.stm>.

References III

-  Berelson, Bernard R., Paul F. Lazarsfeld, and William N. McPhee (1954). *Voting: A Study of Opinion Formation in a Presidential Campaign.* Chicago: University of Chicago Press.
-  Blake, Tom, Chris Nosko, and Steven Tadelis (2014). *Consumer heterogeneity and paid search effectiveness: A large scale field experiment.* Tech. rep. National Bureau of Economic Research.
-  Bound, J., D. Jaeger, and R. Baker (1995). “Problems with Instrumental Variables Estimation when the Correlation Between the Instruments and the Endogenous Regressors is Weak”. In: *Journal of the American Statistical Association* 90, pp. 443–450.

References IV



Bowers, Jake and Ben Hansen (2006). "Attributing Effects to A Cluster Randomized Get-Out-The-Vote Campaign".
Technical Report #448, Statistics Department, University of Michigan.

[http://www-personal.umich.edu/~jwbowers/
PAPERS/bowershansen2006-10TechReport.pdf](http://www-personal.umich.edu/~jwbowers/PAPERS/bowershansen2006-10TechReport.pdf).



Campbell, Angus et al. (1960). *The American Voter*. New York:
John Wiley & Sons.



Cox, David R. (1958). *Planning of Experiments*. New York:
Wiley.



Cox, Gary W. and Jonathan N. Katz (2002). *Elbridge Gerry's Salamander: The Electoral Consequences of the Reapportionment Revolution*. New York: Cambridge University Press.

References V

-  Dehejia, Rajeev (2005). "Practical Propensity Score Matching: A Reply to Smith and Todd". In: *Journal of Econometrics* 125.1–2, pp. 355–364.
-  Dehejia, Rajeev H. and Sadek Wahba (2002). "Propensity Score Matching Methods for Nonexperimental Causal Studies". In: *Review of Economics and Statistics* 84.1, pp. 151–161.
-  Dehejia, Rajeev and Sadek Wahba (1997). "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs". Rejeev Dehejia, *Econometric Methods for Program Evaluation*. Ph.D. Dissertation, Harvard University, Chapter 1.
-  – (1999). "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs". In: *Journal of the American Statistical Association* 94.448, pp. 1053–1062.

References VI

-  Fisher, Ronald A. (1935). *Design of Experiments*. New York: Hafner.
-  Freedman, David A (2010). *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*. Cambridge University Press.
-  Hainmueller, Jens (2012). "Entropy Balancing: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies". In: *Political Analysis* 20.1, pp. 25–46.
-  Hartman, Erin et al. (forthcoming). "From SATE to PATT: Combining Experimental with Observational Studies to Estimate Population Treatment Effects". In: *Journal of the Royal Statistical Society, Series A*.
-  Heckman, James J., Hidehiko Ichimura, et al. (1998). "Characterizing Selection Bias Using Experimental Data". In: *Econometrica* 66.5, pp. 1017–1098.

References VII

-  Heckman, James J. and Jeffrey A. Smith (1995). "Assessing the Case for Social Experiments". In: *Journal of Economic Perspectives* 9.2, pp. 85–110.
-  Holland, Paul W. (1986). "Statistics and Causal Inference". In: *Journal of the American Statistical Association* 81.396, pp. 945–960.
-  Imbens, Guido W. and Paul Rosenbaum (2005). "Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education". In: *Journal of the Royal Statistical Society, Series A* 168, pp. 109–126.
-  LaLonde, Robert (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data". In: *American Economic Review* 76.4, pp. 604–620.

References VIII

-  Lee, David S. (2001). "The Electoral Advantage to Inc incumbency and Voters' Valuation of Politicians' Experience: A Regression Discontinuity Analysis of Elections to the US House". Working paper no. W8441, National Bureau of Economic Research, <http://emlab.berkeley.edu/users/cle/wp/wp31.pdf>.
-  Lee, David S. (2008a). "Randomized Experiments from Non-Random Selection in U.S. House Elections". In: *Journal of Econometrics* 142.2, pp. 675–697.
-  – (2008b). "Randomized Experiments from Non-random Selection in U.S. House Elections". In: *Journal of Econometrics* 142.2, pp. 675–697.
-  Linden, L. (2004). "Are incumbents really advantaged? The Preference for Non-Incumbents in Indian National Elections". Working Paper.

References IX

-  Lupia, Arthur (2004). "Questioning Our Competence: Tasks, Institutions, and the Limited Practical Relevance of Common Political Knowledge Measures". *Working Paper*.
-  McKelvey, Richard D. and Peter C. Ordeshook (1985a). "Elections with Limited Information: A Fulfilled Expectations Model Using Contemporaneous Poll and Endorsement Data as Information Sources". In: *Journal of Economic Theory* 36, pp. 55–85.
 - (1985b). "Sequential Elections with Limited Information". In: *American Journal of Political Science* 29.3, pp. 480–512.
 - (1986). "Information, Electoral Equilibria, and the Democratic Ideal". In: *Journal of Politics* 48.4, pp. 909–937.
-  Mebane, Walter R. Jr. and Jasjeet S. Sekhon (2011). "Genetic Optimization Using Derivatives: The rgenoud package for R". In: *Journal of Statistical Software* 42.11, pp. 1–26.

References X

-  Mill, John Stuart (1873). *Autobiography*. London: Longmans, Green and Co.
-  Mitchell, Ann F. S. and Wojtek J. Krzanowski (1985). "The Mahalanobis Distance and Elliptic Distributions". In: *Biometrika* 72.2, pp. 464–467.
-  – (1989). "Amendments and Corrections: The Mahalanobis Distance and Elliptic Distributions". In: *Biometrika* 76.2, p. 407.
-  Neyman, Jerzy (1923/1990). "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9". In: *Statistical Science* 5.4. Trans. Dorota M. Dabrowska and Terence P. Speed., pp. 465–472.
-  Pettersson-Lidbom, Per (2001). "Do Parties Matter for Fiscal Policy Choices? A Regression-Discontinuity Approach". Working paper, <http://courses.gov.harvard.edu/gov3009/fall01/Partyeffects.pdf>.

References XI

-  Przeworski, A. and H. Teune (1970). *The Logic of Comparative Social Inquiry*. New York: Wiley.
-  Rosenbaum, Paul R. and Donald B. Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects". In: *Biometrika* 70.1, pp. 41–55.
-  Rosenbaum, Paul R. and Donald B. Rubin (1985). "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score". In: *The American Statistician* 39.1, pp. 33–38.
-  Rubin, Donald B. (1978). "Bayesian Inference for Causal Effects: The Role of Randomization". In: *Annals of Statistics* 6.1, pp. 34–58.
-  Rubin, Donald B. and Elizabeth A. Stuart (2006). "Affinely Invariant Matching Methods with Discriminant Mixtures of Proportional Ellipsoidally Symmetric Distributions". In: *Annals of Statistics* 34.4, pp. 1814–1826.

References XII

-  Rubin, Donald B. and Neal Thomas (1992). "Affinely Invariant Matching Methods with Ellipsoidal Distributions". In: *Annals of Statistics* 20.2, pp. 1079–1093.
-  Samii, Cyrus and Peter M Aronow (2012). "On equivalencies between design-based and regression-based variance estimators for randomized experiments". In: *Statistics & Probability Letters* 82.2, pp. 365–370.
-  Sekhon, Jasjeet S. (2004). "Quality Meets Quantity: Case Studies, Conditional Probability and Counterfactuals". In: *Perspectives on Politics* 2.2, pp. 281–293.
-  Smith, Jeffrey A. and Petra E. Todd (2001). "Reconciling Conflicting Evidence on the Performance of Propensity Score Matching Methods". In: *AEA Papers and Proceedings* 91.2, pp. 112–118.

References XIII

-  Smith, Jeffrey and Petra Todd (2005a). "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" In: *Journal of Econometrics* 125.1–2, pp. 305–353.
-  Smith, Jeffrey and Petra Todd (2005b). "Rejoinder". In: *Journal of Econometrics* 125.1–2, pp. 365–375.
-  Sniderman, Paul M. (1993). "The New Look in Public Opinion Research". In: *Political Science: The State of the Discipline II*. Ed. by Ada W. Finifter. Washington, DC: American Political Science Association.
-  Uppal, Yogesh (2005). "The (Dis)advantaged Incumbents: Estimating Inc incumbency Effects in Indian State Legislatures". In: *Center for the Study of Democracy. Symposium: Democracy and Its Development*. Paper G05-06.

References XIV

-  Zaller, John (1998). "Politicians as Prize Fighters: Electoral Selection and Incumbency Advantage". In: *Party Politics and Politicians*. Ed. by John Geer. Baltimore: Johns Hopkins University Press.
-  Zaller, John R (1992). *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.