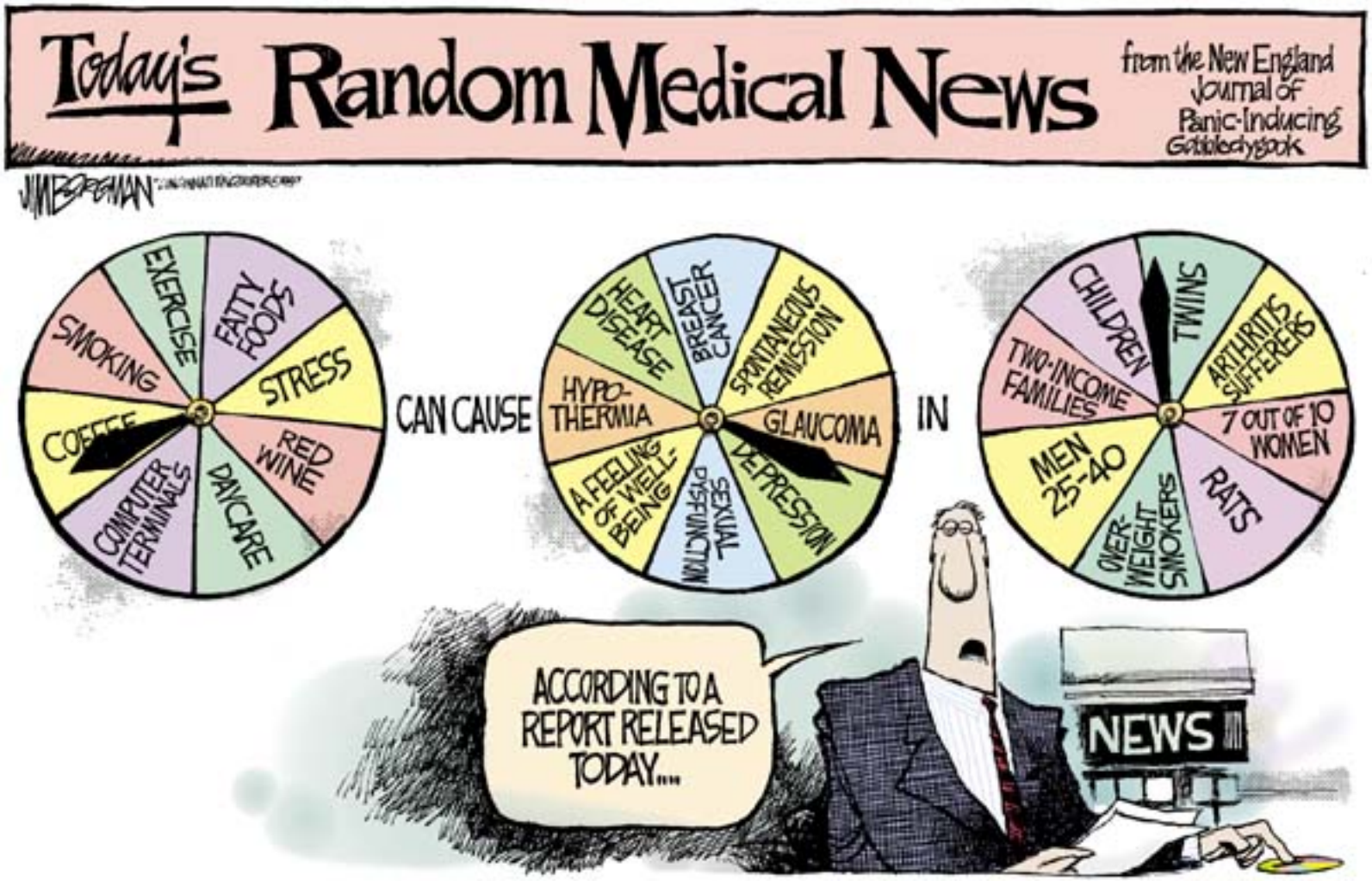
Causal Inference

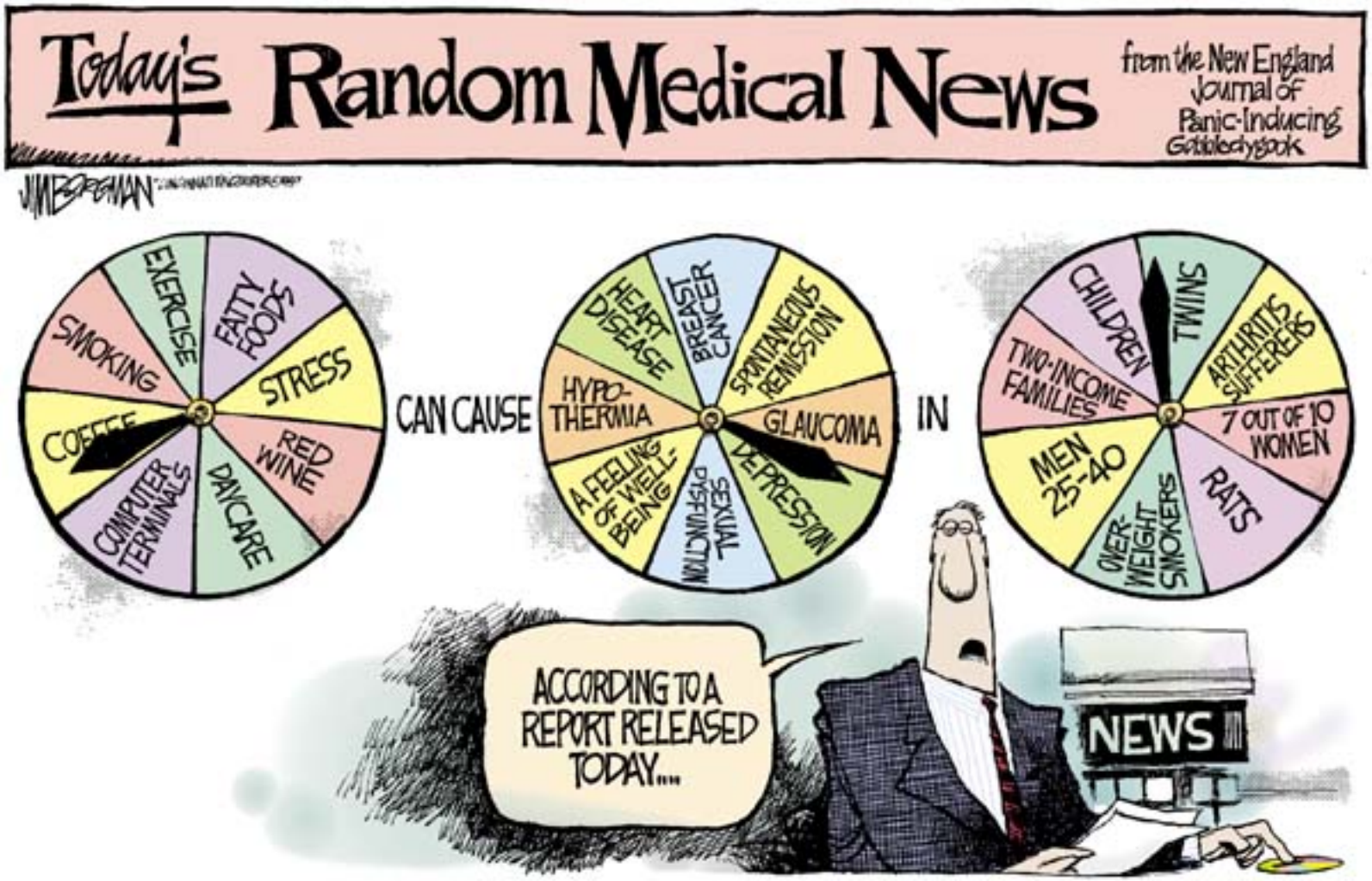
Justin Grimmer

Associate Professor
Department of Political Science
University of Chicago

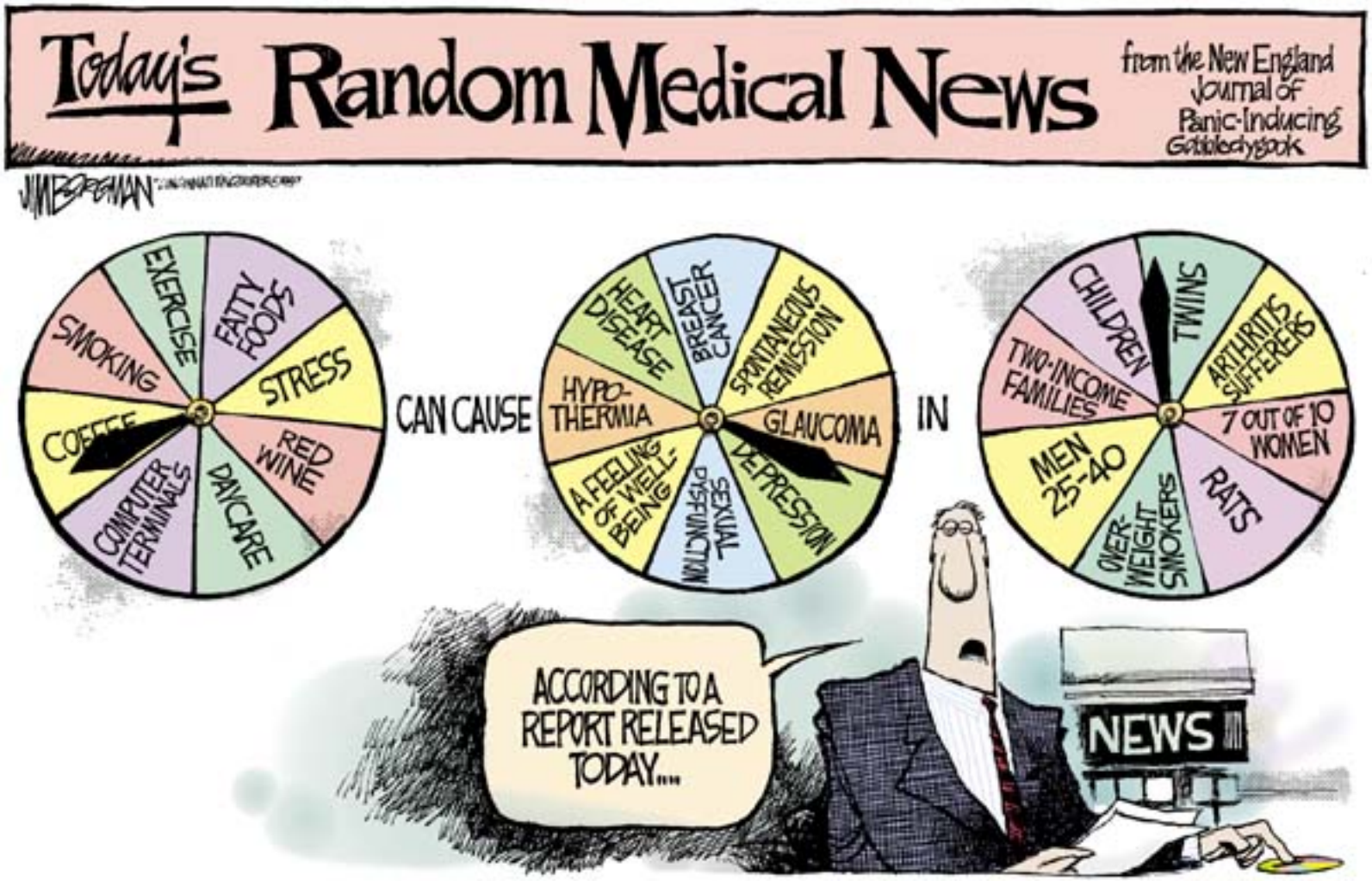
April 16th, 2018



Lancet 2001: negative correlation between coronary heart disease mortality and level of vitamin C in bloodstream (controlling for age, gender, blood pressure, diabetes, and smoking)



Lancet 2002: no effect of vitamin C on mortality in controlled placebo trial (controlling for nothing)



Lancet 2003: comparing among individuals with the same age, gender, blood pressure, diabetes, and smoking, those with higher vitamin C levels have lower levels of obesity, lower levels of alcohol consumption, are less likely to grow up in working class, etc.

Observational Studies

- Randomization forms gold standard for causal inference, because it balances observed and unobserved confounders
- Cannot always randomize so we do observational studies, where we adjust for the observed covariates and hope that unobservables are balanced
- Better than hoping: design observational study to approximate an experiment
  - "The planner of an observational study should always ask himself: How would the study be conducted if it were possible to do it by controlled experimentation" (Cochran 1965)

The Good, the Bad, and the Ugly

Treatments, Covariates, Outcomes

- Randomized Experiment: Well-defined treatment, clear distinction between covariates and outcomes, control of assignment mechanism
- Better Observational Study: Well-defined treatment, clear distinction between covariates and outcomes, precise knowledge of assignment mechanism
  - Can convincingly answer the following question: Why do two units who are identical on measured covariates receive different treatments?
- Poorer Observational Study: Hard to say when treatment began or what the treatment really is. Distinction between covariates and outcomes is blurred, so problems that arise in experiments seem to be avoided but are in fact just ignored. No precise knowledge of assignment mechanism.

The Good, the Bad, and the Ugly

How were treatments assigned?

- Randomized Experiment: Random assignment
- Better Observational Study: Assignment is not random, but circumstances for the study were chosen so that treatment seems haphazard, or at least not obviously related to potential outcomes (sometimes we refer to these as natural or quasi-experiments)
- Poorer Observational Study: No attention given to assignment process, units self-select into treatment based on potential outcomes

The Good, the Bad, and the Ugly

What is the problem with purely cross-sectional data?

- Difficult to know what is pre or post treatment.
- Many important confounders will be affected by the treatment and including these "bad controls" induces post-treatment bias.
- But if you do not condition on the confounders that are post-treatment, then often only left with a limited set of covariates such as socio-demographics.

The Good, the Bad, and the Ugly

Were treated and controls comparable?

- Randomized Experiment: Balance table for observables.
- Better Observational Study: Balance table for observables. Ideally sensitivity analysis for unobservables.
- Poorer Observational Study: No direct assessment of comparability is presented.

The Good, the Bad, and the Ugly

Eliminating plausible alternatives to treatment effects?

- Randomized Experiment: List plausible alternatives and experimental design includes features that shed light on these alternatives (e.g. placebos). Report on potential attrition and non-compliance.
- Better Observational Study: List plausible alternatives and study design includes features that shed light on these alternatives (e.g. multiple control groups, longitudinal covariate data, etc.). Requires more work than in experiment since there are usually many more alternatives.
- Poorer Observational Study: Alternatives are mentioned in discussion section of the paper.

Good Observational Studies

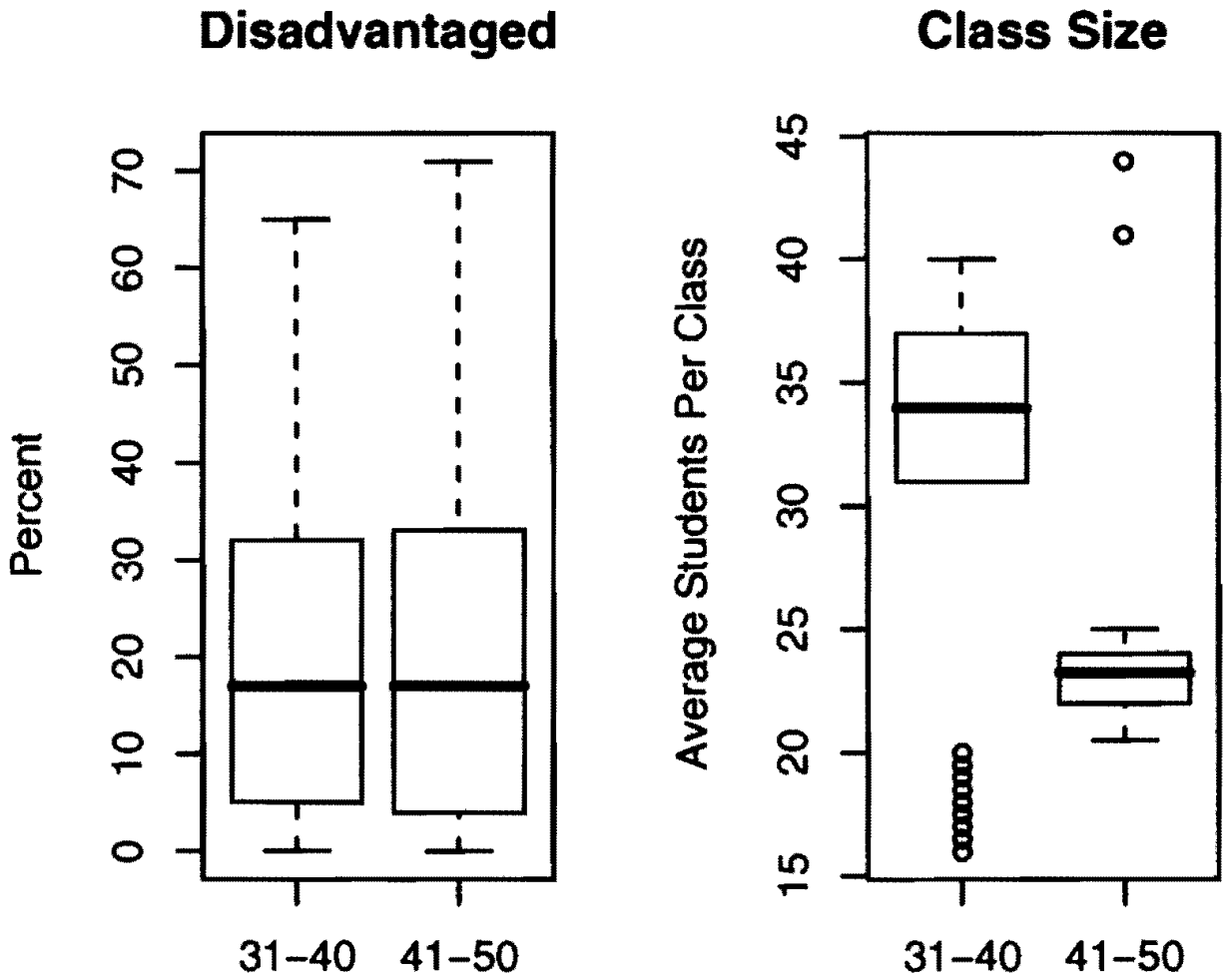
Design features we can use to handle unobservables:

- Design comparisons so that unobservables are likely to be balanced (e.g. sub-samples, groups where treatment assignment was accidental, etc.)
- Unobservables may differ, but comparisons that are unaffected by differences in time-invariant unobservables
- Instrumental variables, if applied correctly
- Multiple control groups that are known to differ on unobservables
- Sensitivity analysis and bounds

Good Observational Studies

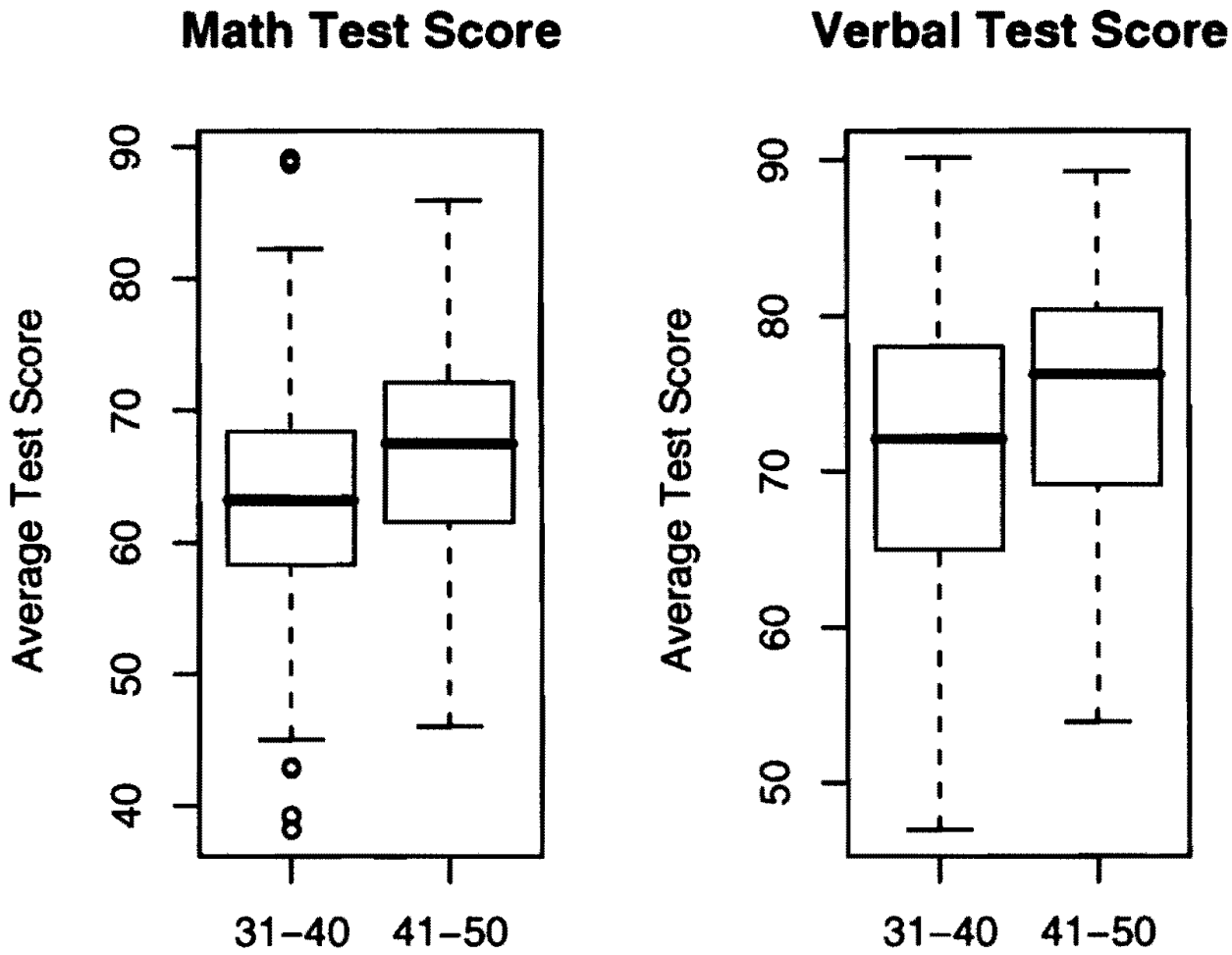
1. An observational study should be structured to resemble an experiment.
2. Adjustment for observed covariates should be simple, transparent, and convincing.
3. The most plausible alternatives to the treatment effect should be anticipated and addressed.
4. The analysis should address possible biases from unmeasured covariates.

Class Size on Student Achievements



Angrist and Lavy (1999)

Class Size on Student Achievements



Angrist and Lavy (1999)

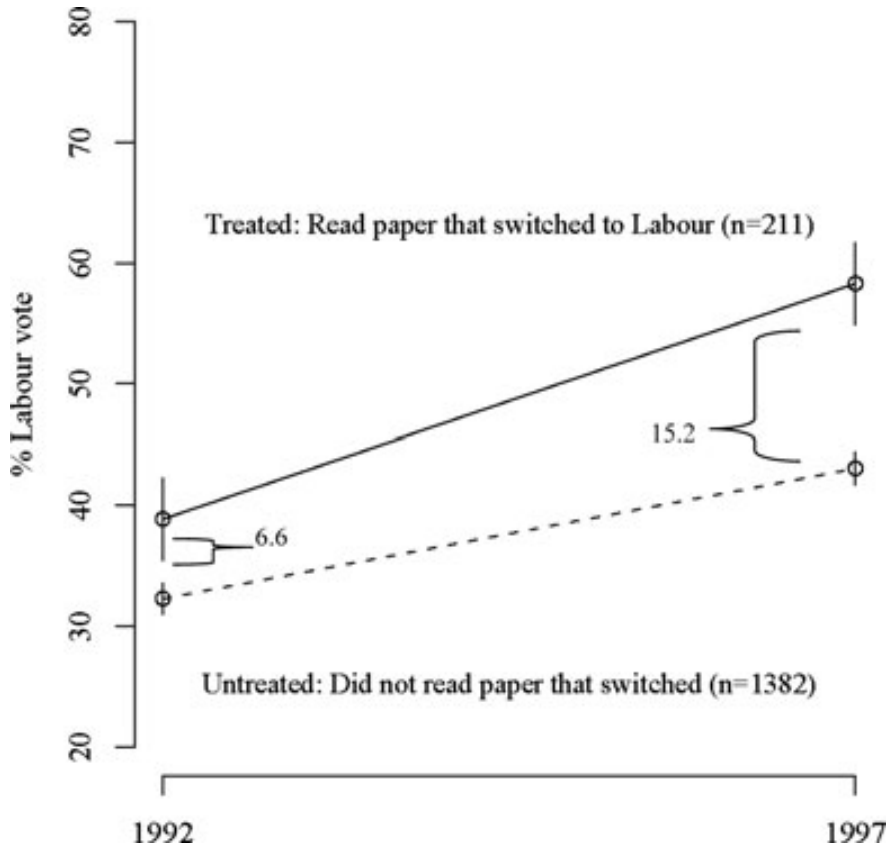
Seat Belts on Fatality Rates

Table 1.1 Crashes in FARS 1975–1983 in which the front seat had two occupants, a driver and a passenger, with one belted, the other unbelted, and one died and one survived.

	Driver Passenger	Not Belted Belted	Belted Not Belted
Driver Died	Passenger Survived	189	153
Driver Survived	Passenger Died	111	363

Evans (1986)

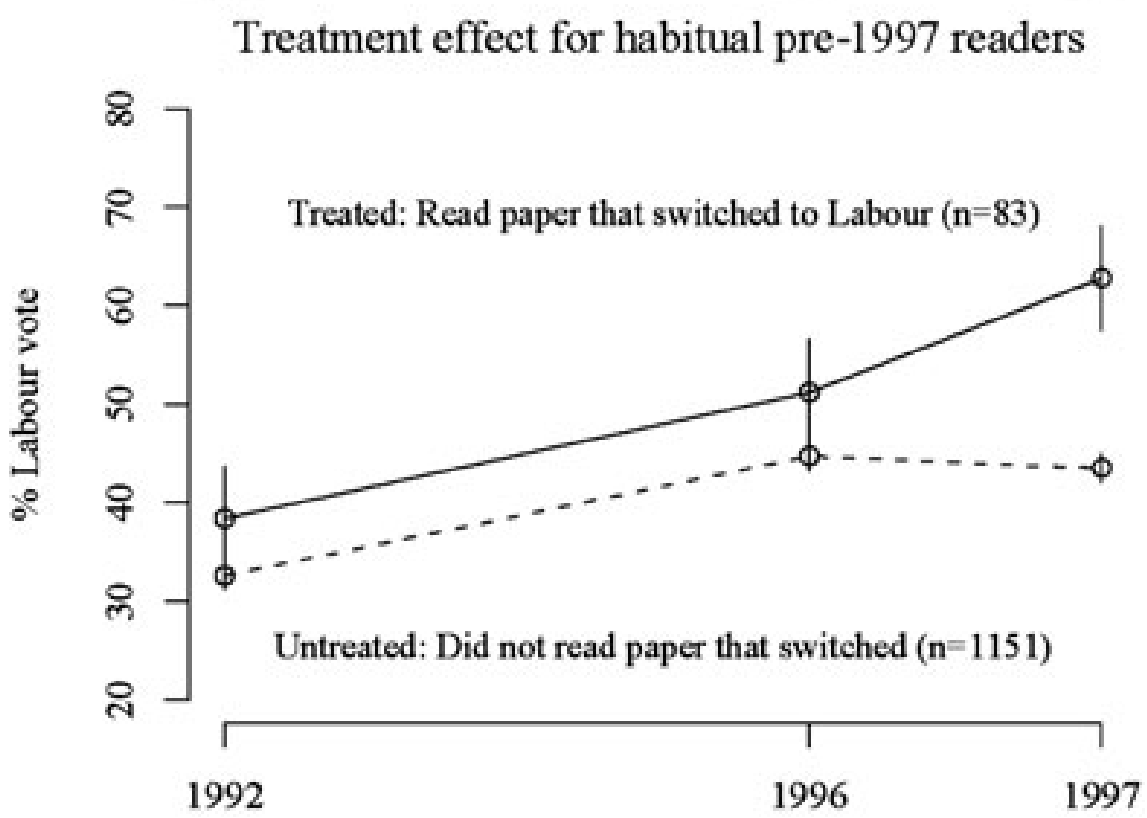
Persuasive Effect of Endorsement Changes on Labour Vote



This figure shows that reading a paper that switched to Labour is associated with an $(15.2 - 6.6 =)$ 8.6 percentage point shift to Labour between the 1992 and 1997 UK elections. Paper readership is measured in the 1996 wave, before the papers switched, or, if no 1996 interview was conducted, in an earlier wave. Confidence intervals show one standard error.

Ladd and Lenz (1999)

Persuasive Effect of Endorsement Changes on Labour Vote



Using the hypothetical vote choice question asked in the 1996 wave, this figure shows that the treatment effect only emerges after 1996. Habitual readers are those who read a paper that switched in every wave in which they were interviewed before the 1997 wave.

Ladd and Lenz (1999)

Adjustment for Observables in Observational Studies

- Subclassification
- Matching
- Propensity Score Methods
- Regression

Smoking and Mortality (Cochran (1968))

TABLE 1
DEATH RATES PER 1,000 PERSON-YEARS

Smoking group	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes	20.5	14.1	13.5
Cigars/pipes	35.5	20.7	17.4

Smoking and Mortality (Cochran (1968))

TABLE 2
MEAN AGES, YEARS

Smoking group	Canada	U.K.	U.S.
Non-smokers	54.9	49.1	57.0
Cigarettes	50.5	49.8	53.2
Cigars/pipes	65.9	55.7	59.7

Subclassification

To control for differences in age, we would like to compare different smoking-habit groups with the same age distribution

One possibility is to use subclassification:

- for each country, divide each group into different age subgroups
- calculate death rates within age subgroups
- average within age subgroup death rates using fixed weights (e.g. number of cigarette smokers)

Subclassification: Example

	Death Rates Pipe Smokers	# Pipe-Smokers	# Non-Smokers
Age 20 - 50	15	11	29
Age 50 - 70	35	13	9
Age + 70	50	16	2
Total		40	40

What is the average death rate for Pipe Smokers?

$15 \cdot (11/40) + 35 \cdot (13/40) + 50 \cdot (16/40) = 35.5$

Subclassification: Example

	Death Rates Pipe Smokers	# Pipe-Smokers	# Non-Smokers
Age 20 - 50	15	11	29
Age 50 - 70	35	13	9
Age + 70	50	16	2
Total		40	40

What is the average death rate for Pipe Smokers?

$15 \cdot (11/40) + 35 \cdot (13/40) + 50 \cdot (16/40) = 35.5$

Subclassification: Example

	Death Rates Pipe Smokers	# Pipe-Smokers	# Non-Smokers
Age 20 - 50	15	11	29
Age 50 - 70	35	13	9
Age + 70	50	16	2
Total		40	40

What is the average death rate for Pipe Smokers if they had same age distribution as Non-Smokers?

$15 \cdot (29/40) + 35 \cdot (9/40) + 50 \cdot (2/40) = 21.2$

Subclassification: Example

	Death Rates Pipe Smokers	# Pipe-Smokers	# Non-Smokers
Age 20 - 50	15	11	29
Age 50 - 70	35	13	9
Age + 70	50	16	2
Total		40	40

What is the average death rate for Pipe Smokers if they had same age distribution as Non-Smokers?

$15 \cdot (29/40) + 35 \cdot (9/40) + 50 \cdot (2/40) = 21.2$

Smoking and Mortality (Cochran (1968))

TABLE 3
ADJUSTED DEATH RATES USING 3 AGE GROUPS

Smoking group	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes	28.3	12.8	17.7
Cigars/pipes	21.2	12.0	14.2

Identification Under Selection on Observables

Identification Assumption

1. $(Y_1, Y_0) \perp\!\!\!\perp D|X$ *(selection on observables)*
2. $0 < \Pr(D = 1|X) < 1$ *with probability one (common support)*

Identfication Result

Given selection on observables we have

$$\begin{aligned} E[Y_1 - Y_0|X] &= E[Y_1 - Y_0|X, D = 1] \\ &= E[Y|X, D = 1] - E[Y|X, D = 0] \end{aligned}$$

Therefore, under the common support condition:

$$\begin{aligned} \tau_{ATE} &= E[Y_1 - Y_0] = \int E[Y_1 - Y_0|X]\, dP(X) \\ &= \int \big(E[Y|X, D = 1] - E[Y|X, D = 0]\big)\, dP(X) \end{aligned}$$

Identification Under Selection on Observables

Identification Assumption

1. $(Y_1, Y_0) \perp\!\!\!\perp D|X$ *(selection on observables)*
2. $0 < \Pr(D = 1|X) < 1$ *with probability one (common support)*

Identfication Result

Similarly,

$$\begin{aligned} \tau_{ATT} &= E[Y_1 - Y_0|D = 1] \\ &= \int \left(E[Y|X, D = 1] - E[Y|X, D = 0]\right) dP(X|D = 1) \end{aligned}$$

To identify τ_{ATT} the selection on observables and common support conditions can be relaxed to:

- $Y_0 \perp\!\!\!\perp D|X$ *(SOO for Controls)*
- $\Pr(D = 1|X) < 1$ *(Weak Overlap)*

Identification Under Selection on Observables

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	$E[Y_1 \mid X = 0, D = 1]$	$E[Y_0 \mid X = 0, D = 1]$	1	0
2			1	0
3	$E[Y_1 \mid X = 0, D = 0]$	$E[Y_0 \mid X = 0, D = 0]$	0	0
4			0	0
5	$E[Y_1 \mid X = 1, D = 1]$	$E[Y_0 \mid X = 1, D = 1]$	1	1
6			1	1
7	$E[Y_1 \mid X = 1, D = 0]$	$E[Y_0 \mid X = 1, D = 0]$	0	1
8			0	1

Identification Under Selection on Observables

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	$E[Y_1\|X=0,D=1]$	$E[Y_0\|X=0,D=1]=$	1	0
2		$E[Y_0\|X=0,D=0]$	1	0
3	$E[Y_1\|X=0,D=0]$	$E[Y_0\|X=0,D=0]$	0	0
4			0	0
5	$E[Y_1\|X=1,D=1]$	$E[Y_0\|X=1,D=1]=$	1	1
6		$E[Y_0\|X=1,D=0]$	1	1
7	$E[Y_1\|X=1,D=0]$	$E[Y_0\|X=1,D=0]$	0	1
8			0	1

$(Y_1, Y_0) \perp D|X$ implies that we conditioned on all confounders. The treatment is randomly assigned within each stratum of X:

$$E[Y_0|X=0,D=1] = E[Y_0|X=0,D=0] \text{ and}$$
$$E[Y_0|X=1,D=1] = E[Y_0|X=1,D=0]$$

Identification Under Selection on Observables

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	$E[Y_1\|X=0,D=1]$	$E[Y_0\|X=0,D=1]=$	1	0
2		$E[Y_0\|X=0,D=0]$	1	0
3	$E[Y_1\|X=0,D=0]=$	$E[Y_0\|X=0,D=0]$	0	0
4	$E[Y_1\|X=0,D=1]$		0	0
5	$E[Y_1\|X=1,D=1]$	$E[Y_0\|X=1,D=1]=$	1	1
6		$E[Y_0\|X=1,D=0]$	1	1
7	$E[Y_1\|X=1,D=0]=$	$E[Y_0\|X=1,D=0]$	0	1
8	$E[Y_1\|X=1,D=1]$		0	1

$(Y_1, Y_0) \perp D|X$ also implies

$$E[Y_1|X=0,D=1] = E[Y_1|X=0,D=0] \text{ and}$$
$$E[Y_1|X=1,D=1] = E[Y_1|X=1,D=0]$$

Subclassification Estimator

Identfication Result

$$\begin{aligned}\tau_{ATE} &= \int \Big(E[Y|X, D=1] - E[Y|X, D=0]\Big)\, dP(X) \\ \tau_{ATT} &= \int \Big(E[Y|X, D=1] - E[Y|X, D=0]\Big)\, dP(X|D=1)\end{aligned}$$

Assume X takes on K different cells $\{X^1, ..., X^k, ..., X^K\}$. Then the analogy principle suggests estimators:

$$\widehat{\tau}_{ATE} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k\right) \cdot \left(\frac{N^k}{N}\right); \; \widehat{\tau}_{ATT} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k\right) \cdot \left(\frac{N_1^k}{N_1}\right)$$

- N^k is # of obs. and N_1^k is # of treated obs. in cell k
- $\bar{Y}_1^k$ is mean outcome for the treated in cell k
- $\bar{Y}_0^k$ is mean outcome for the untreated in cell k

Subclassification Estimator

Identfication Result

$$\begin{aligned}\tau_{ATE} &= \int \Big(E[Y|X, D=1] - E[Y|X, D=0]\Big)\, dP(X) \\ \tau_{ATT} &= \int \Big(E[Y|X, D=1] - E[Y|X, D=0]\Big)\, dP(X|D=1)\end{aligned}$$

Assume X takes on K different cells $\{X^1, ..., X^k, ..., X^K\}$. Then the analogy principle suggests estimators:

$$\widehat{\tau}_{ATE} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k\right) \cdot \left(\frac{N^k}{N}\right); \; \widehat{\tau}_{ATT} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k\right) \cdot \left(\frac{N_1^k}{N_1}\right)$$

- N^k is # of obs. and N_1^k is # of treated obs. in cell k
- $\bar{Y}_1^k$ is mean outcome for the treated in cell k
- $\bar{Y}_0^k$ is mean outcome for the untreated in cell k

Subclassification by Age ($K = 2$)

X_k	Death Rate Smokers	Death Rate Non-Smokers	Diff.	# Smokers	# Obs.
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

What is $\widehat{\tau}_{ATE} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k \right) \cdot \left(\frac{N^k}{N} \right)$?

$\widehat{\tau}_{ATE} = 4 \cdot (10/20) + 6 \cdot (10/20) = 5$

Subclassification by Age ($K = 2$)

X_k	Death Rate Smokers	Death Rate Non-Smokers	Diff.	# Smokers	# Obs.
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

What is $\widehat{\tau}_{ATE} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k \right) \cdot \left(\frac{N^k}{N} \right)$?

$\widehat{\tau}_{ATE} = 4 \cdot (10/20) + 6 \cdot (10/20) = 5$

Subclassification by Age ($K = 2$)

X_k	Death Rate Smokers	Death Rate Non-Smokers	Diff.	# Smokers	# Obs.
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

What is $\widehat{\tau}_{ATT} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k \right) \cdot \left(\frac{N_1^k}{N_1} \right)$?

$\widehat{\tau}_{ATT} = 4 \cdot (3/10) + 6 \cdot (7/10) = 5.4$

Subclassification by Age ($K = 2$)

X_k	Death Rate Smokers	Death Rate Non-Smokers	Diff.	# Smokers	# Obs.
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

What is $\widehat{\tau}_{ATT} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k \right) \cdot \left(\frac{N_1^k}{N_1} \right)$?

$\widehat{\tau}_{ATT} = 4 \cdot (3/10) + 6 \cdot (7/10) = 5.4$

Subclassification by Age and Gender ($K = 4$)

X_k	Death Rate Smokers	Death Rate Non-Smokers	Diff.	# Smokers	# Obs.
Old, Male	28	22	4	3	7
Old, Female		24		0	3
Young, Male	21	16	5	3	4
Young, Female	23	17	6	4	6
Total				10	20

What is $\widehat{\tau}_{ATE} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k \right) \cdot \left(\frac{N^k}{N} \right)$?

Not identified!

Subclassification by Age and Gender ($K = 4$)

X_k	Death Rate Smokers	Death Rate Non-Smokers	Diff.	# Smokers	# Obs.
Old, Male	28	22	4	3	7
Old, Female		24		0	3
Young, Male	21	16	5	3	4
Young, Female	23	17	6	4	6
Total				10	20

What is $\widehat{\tau}_{ATE} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k \right) \cdot \left(\frac{N^k}{N} \right)$?

Not identified!

Subclassification by Age and Gender ($K = 4$)

X_k	Death Rate Smokers	Death Rate Non-Smokers	Diff.	# Smokers	# Obs.
Old, Male	28	22	4	3	7
Old, Female		24		0	3
Young, Male	21	16	5	3	4
Young, Female	23	17	6	4	6
Total				10	20

What is $\widehat{\tau}_{ATT} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k \right) \cdot \left(\frac{N_1^k}{N_1} \right)$?

$\widehat{\tau}_{ATT} = 4 \cdot (3/10) + 5 \cdot (3/10) + 6 \cdot (4/10) = 5.1$

Subclassification by Age and Gender ($K = 4$)

X_k	Death Rate Smokers	Death Rate Non-Smokers	Diff.	# Smokers	# Obs.
Old, Male	28	22	4	3	7
Old, Female		24		0	3
Young, Male	21	16	5	3	4
Young, Female	23	17	6	4	6
Total				10	20

What is $\widehat{\tau}_{ATT} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k \right) \cdot \left(\frac{N_1^k}{N_1} \right)$?

$\widehat{\tau}_{ATT} = 4 \cdot (3/10) + 5 \cdot (3/10) + 6 \cdot (4/10) = 5.1$

Matching is Not an Identification Strategy

Matching

When X is continuous we can estimate τ_{ATT} by "imputing" the missing potential outcome of each treated unit using the observed outcome from the "closest" control unit:

$$\widehat{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \left(Y_i - Y_{j(i)} \right)$$

where $Y_{j(i)}$ is the outcome of an untreated observation such that $X_{j(i)}$ is the closest value to X_i among the untreated observations.

We can also use the average for M closest matches:

$$\widehat{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \left\{ Y_i - \left(\frac{1}{M} \sum_{m=1}^{M} Y_{j_m(i)}, \right) \right\}$$

Works well when we can find good matches for each treated unit

Matching

When X is continuous we can estimate τ_{ATT} by "imputing" the missing potential outcome of each treated unit using the observed outcome from the "closest" control unit:

$$\widehat{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \left(Y_i - Y_{j(i)} \right)$$

where $Y_{j(i)}$ is the outcome of an untreated observation such that $X_{j(i)}$ is the closest value to X_i among the untreated observations.
We can also use the average for M closest matches:

$$\widehat{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \left\{ Y_i - \left(\frac{1}{M} \sum_{m=1}^{M} Y_{j_m(i)}, \right) \right\}$$

Works well when we can find good matches for each treated unit

Matching: Example with a Single X

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	6	?	1	3
2	1	?	1	1
3	0	?	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

What is $\widehat{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \left(Y_i - Y_{j(i)} \right)$?

Match and plugin in

Matching: Example with a Single X

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	6	?	1	3
2	1	?	1	1
3	0	?	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

What is $\widehat{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \left(Y_i - Y_{j(i)} \right)$?
Match and plugin in

Matching: Example with a Single X

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	6	9	1	3
2	1	0	1	1
3	0	9	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

What is $\widehat{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \left(Y_i - Y_{j(i)} \right)$?

$\widehat{\tau}_{ATT} = 1/3 \cdot (6 - 9) + 1/3 \cdot (1 - 0) + 1/3 \cdot (0 - 9) = -3.7$

Matching: Example with a Single X

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	6	9	1	3
2	1	0	1	1
3	0	9	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

What is $\widehat{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \left(Y_i - Y_{j(i)} \right)$?

$\widehat{\tau}_{ATT} = 1/3 \cdot (6 - 9) + 1/3 \cdot (1 - 0) + 1/3 \cdot (0 - 9) = -3.7$

Matching Distance Metric

"Closeness" is often defined by a distance metric. Let $X_i = (X_B, X_B, ..., X_{ik})'$ and $X_j = (X_{j1}, X_{j2}, ..., X_{jk})'$ be the covariate vectors for i and j.

A commonly used distance is the Mahalanobis distance:

$$MD(X_i, X_j) = \sqrt{(X_i - X_j)' \Sigma^{-1} (X_i - X_j)}$$

where Σ is the Variance-Covariance-Matrix so the distance metric is scale-invariant and takes into account the correlations. For an exact match $MD(X_i, X_j) = 0$.
Other distance metrics can be used, for example Stata's `nnmatch` by default uses the diagonal matrix of the inverse of the covariate variances (normalized Euclidean distance):

$$StataD(X_i, X_j) = \sqrt{(X_i - X_j)' diag(\Sigma_X^{-1}) (X_i - X_j)}$$

In `R`, Genetic matching uses (`GenMatch(Matching)`):

$$GeneticD(X_i, X_j) = \sqrt{(X_i - X_j)' (S^{-1/2})' \, W \, S^{-1/2} (X_i - X_j)}$$

where W is a $(k \times k)$ positive definite weight matrix with zeros in off-diagonals.

Mahalanobis Distance: Example

	X_1	X_2
Treated	0	0
Control A	2	2
Control B	1.8	0

$$X_T = \begin{pmatrix} 0 & 0 \end{pmatrix}' \quad X_A = \begin{pmatrix} 2 & 2 \end{pmatrix}' \quad X_B = \begin{pmatrix} 1.8 & 0 \end{pmatrix}'$$

$$\Sigma = \begin{pmatrix} & X_1 & X_2 \\ X_1 & 1 & .9 \\ X_2 & .9 & 1 \end{pmatrix}$$

Which control is closer?

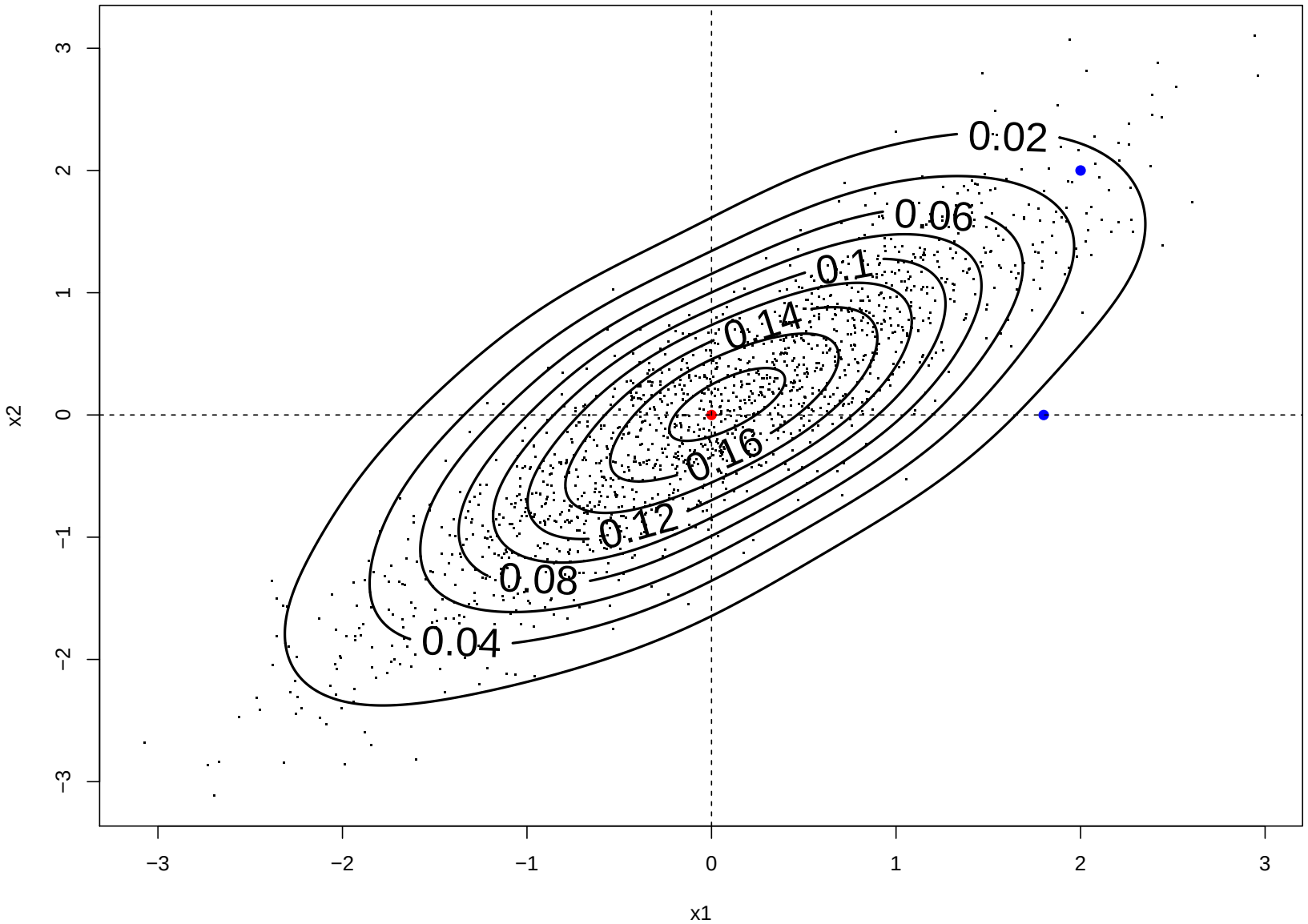
Mahalanobis Distance

$$X_T = \begin{pmatrix} 0 & 0 \end{pmatrix}' \quad X_A = \begin{pmatrix} 2 & 2 \end{pmatrix}' \quad X_B = \begin{pmatrix} 1.8 & 0 \end{pmatrix}' \quad \Sigma = \begin{pmatrix} & X_1 & X_2 \\ X_1 & 1 & .9 \\ X_2 & .9 & 1 \end{pmatrix}$$

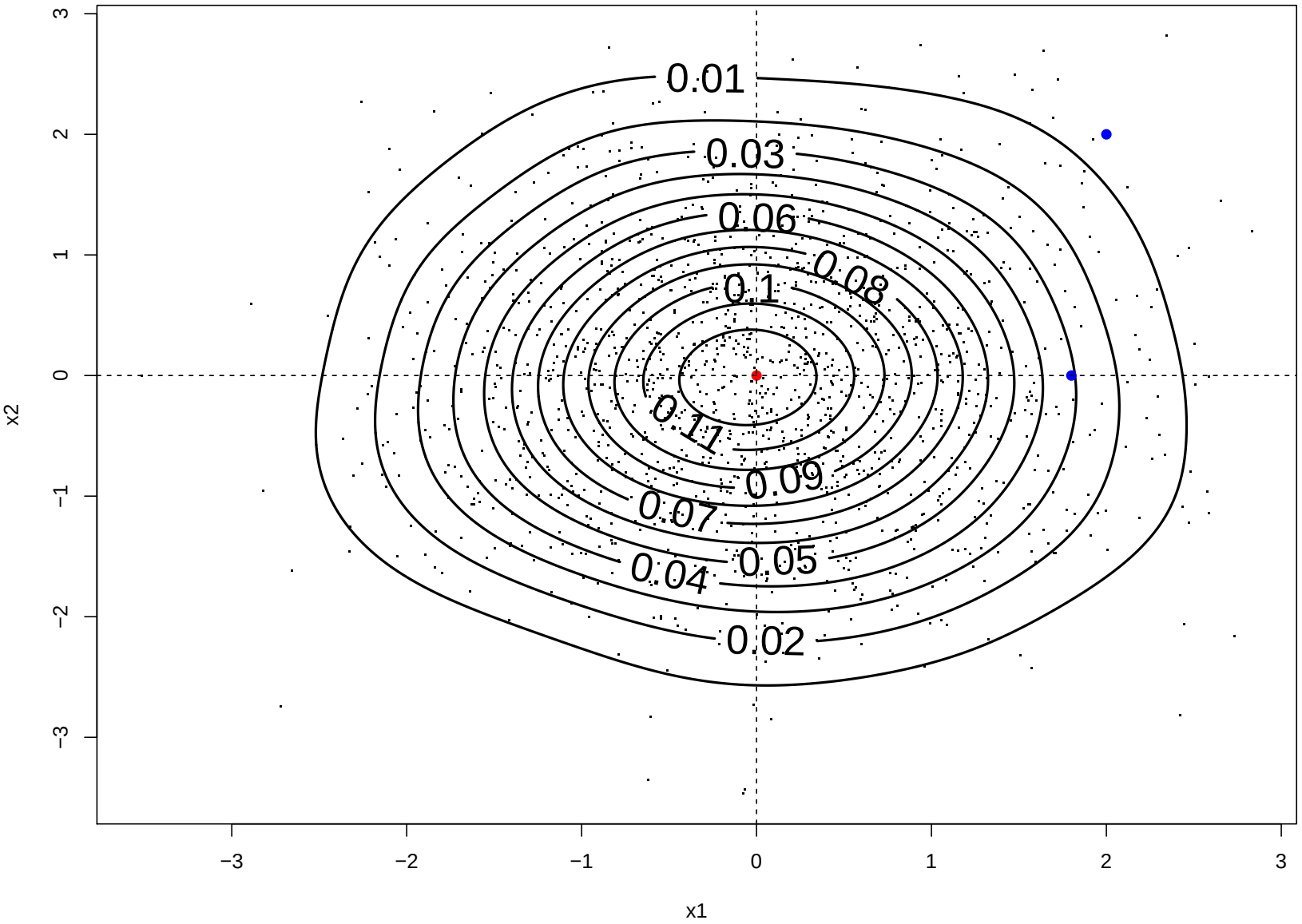
$$\begin{aligned} MD(X_A, X_T) &= \sqrt{(X_A - X_T)'\Sigma^{-1}(X_A - X_T)} \\ &= \sqrt{\left(\begin{pmatrix} 2 & 2 \end{pmatrix} - \begin{pmatrix} 0 & 0 \end{pmatrix}\right)' \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix}^{-1} \left(\begin{pmatrix} 2 & 2 \end{pmatrix} - \begin{pmatrix} 0 & 0 \end{pmatrix}\right)} \\ &= \sqrt{\begin{pmatrix} 2 & 2 \end{pmatrix}' \begin{pmatrix} 5.2 & -4.7 \\ -4.7 & 5.2 \end{pmatrix} \begin{pmatrix} 2 & 2 \end{pmatrix}} \\ &= 4.2 \\ MD(X_B, X_T) &= \sqrt{\begin{pmatrix} 1.8 & 0 \end{pmatrix}' \begin{pmatrix} 5.2 & -4.7 \\ -4.7 & 5.2 \end{pmatrix} \begin{pmatrix} 1.8 & 0 \end{pmatrix}} \\ &= 17 \end{aligned}$$

With $StataD(X_A, X_T) = \sqrt{(X_i - X_T)' diag(\Sigma_X^{-1})(X_i - X_T)}$ we find $StataD(X_A, X_T) = 84$ and $StataD(X_B, X_T) = 17$ since correlation is ignored.

Mahalanobis Distance



Normalized Euclidean Distance



Local Methods and the Curse of Dimensionality

Big Problem: in a mathematical space, the volume increases exponentially when adding extra dimensions.

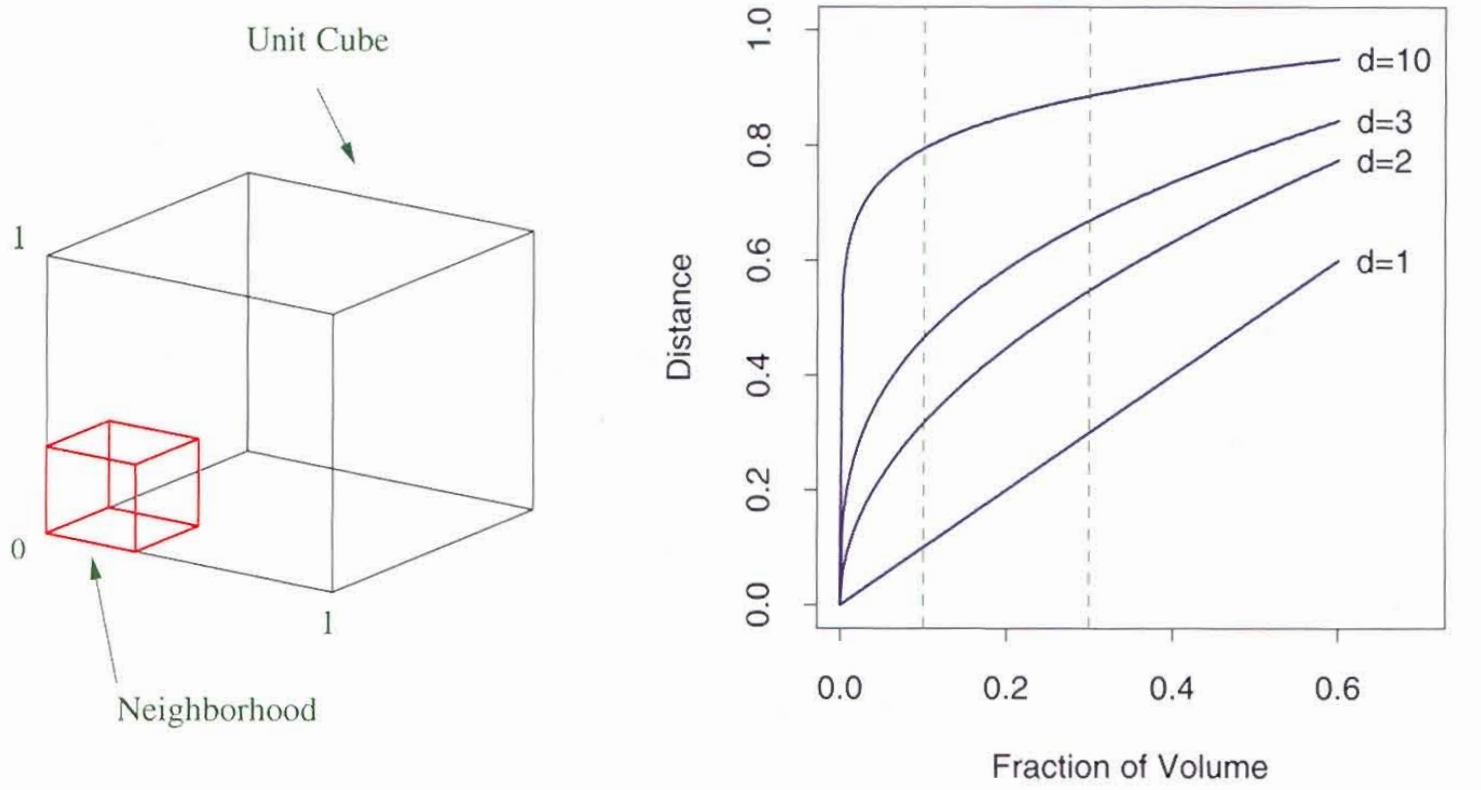


FIGURE 2.6. *The curse of dimensionality is well illustrated by a subcubical neighborhood for uniform data in a unit cube. The figure on the right shows the side-length of the subcube needed to capture a fraction r of the volume of the data, for different dimensions p. In ten dimensions we need to cover 80% of the range of each coordinate to capture 10% of the data.*

Local Methods and the Curse of Dimensionality

Big Problem: in a mathematical space, the volume increases exponentially when adding extra dimensions.

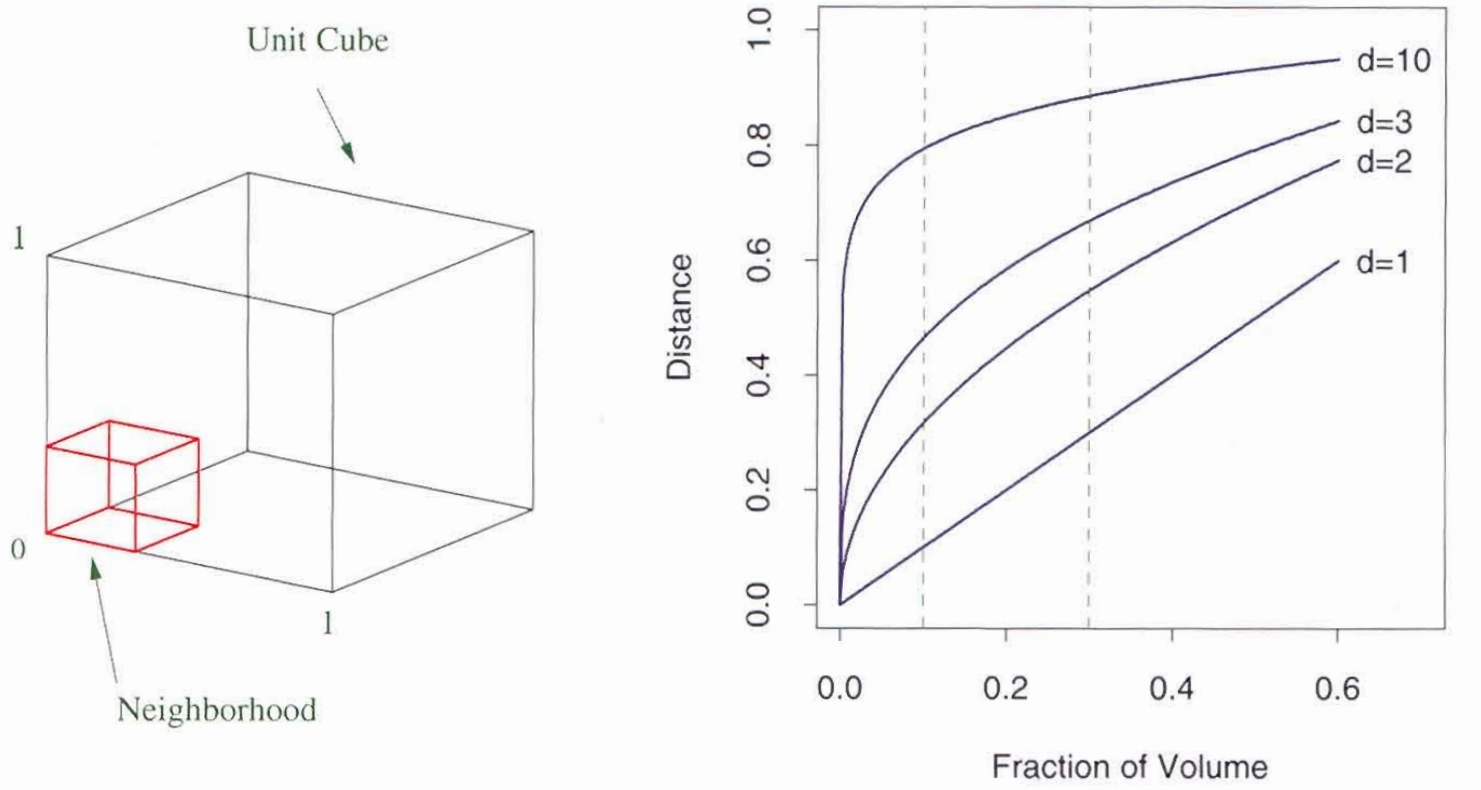
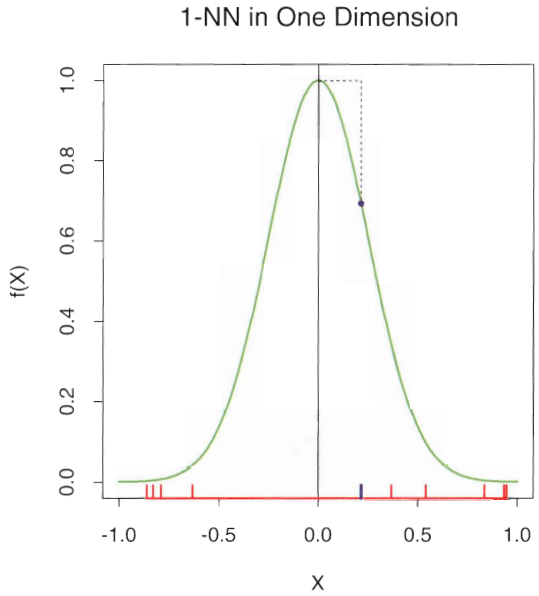
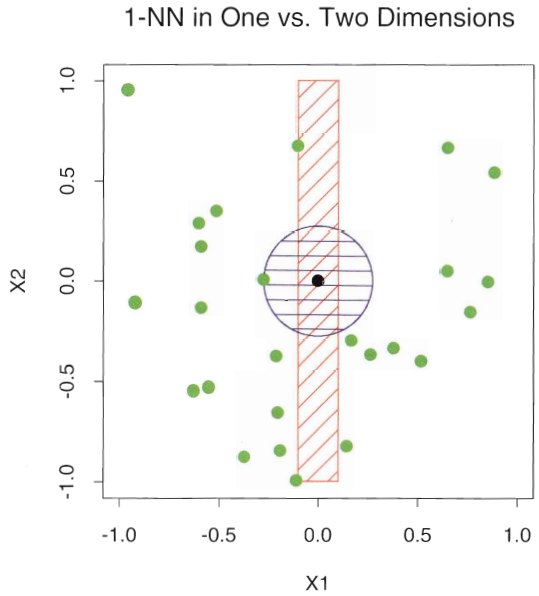
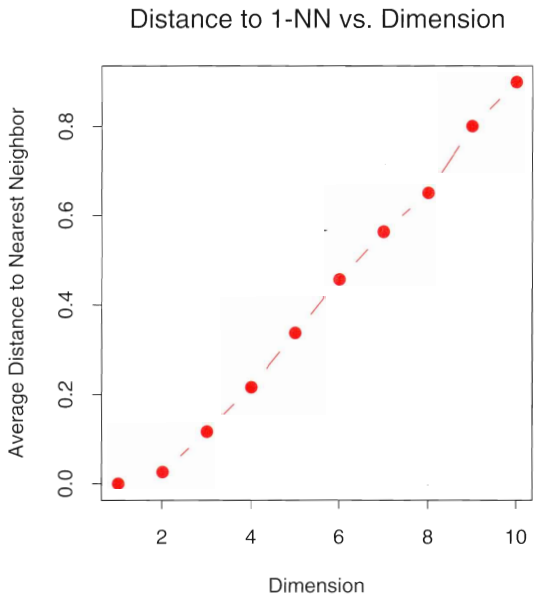


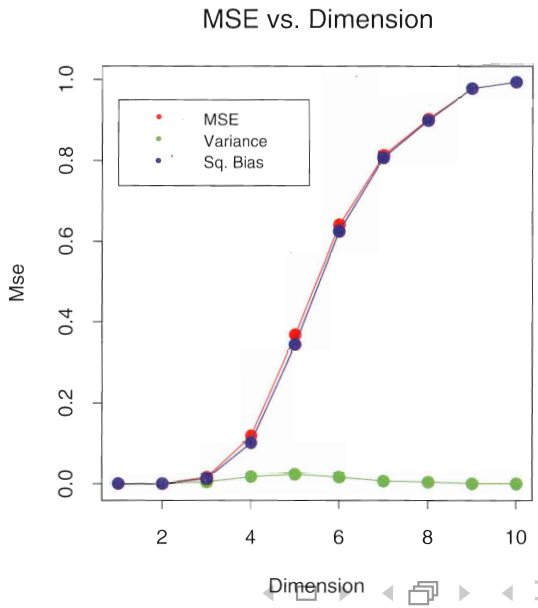
FIGURE 2.6. *The curse of dimensionality is well illustrated by a subcubical neighborhood for uniform data in a unit cube. The figure on the right shows the side-length of the subcube needed to capture a fraction r of the volume of the data, for different dimensions p. In ten dimensions we need to cover 80% of the range of each coordinate to capture 10% of the data.*

Local Methods and the Curse of Dimensionality









The E-Harmony Problem

Curse of dimensionality and on-line dating:

eHarmony matches you based on compatibility in the most important areas of life - like values, character, intellect, sense of humor, and 25 other dimensions.

The E-Harmony Problem

Curse of dimensionality and on-line dating:

eHarmony matches you based on compatibility in the most important areas of life - like values, character, intellect, sense of humor, and 25 other dimensions.

Suppose (for example) 29 dimensions are binary (0,1):

The E-Harmony Problem

Curse of dimensionality and on-line dating:

eHarmony matches you based on compatibility in the most important areas of life - like values, character, intellect, sense of humor, and 25 other dimensions.

Suppose (for example) 29 dimensions are binary (0,1):

Suppose dimensions are independent:

The E-Harmony Problem

Curse of dimensionality and on-line dating:

eHarmony matches you based on compatibility in the most important areas of life - like values, character, intellect, sense of humor, and 25 other dimensions.

Suppose (for example) 29 dimensions are binary (0,1):
Suppose dimensions are independent:
Pr(2 people agree) $= 0.5$

The E-Harmony Problem

Curse of dimensionality and on-line dating:

eHarmony matches you based on compatibility in the most important areas of life - like values, character, intellect, sense of humor, and 25 other dimensions.

Suppose (for example) 29 dimensions are binary (0,1):
Suppose dimensions are independent:
Pr(2 people agree) = 0.5

$$\Pr(\text{Exact}) = \Pr(\text{Agree})_1 \times \Pr(\text{Agree})_2 \times \cdots \times \Pr(\text{Agree})_{29}$$

The E-Harmony Problem

Curse of dimensionality and on-line dating:

eHarmony matches you based on compatibility in the most important areas of life - like values, character, intellect, sense of humor, and 25 other dimensions.

Suppose (for example) 29 dimensions are binary (0,1):
Suppose dimensions are independent:
Pr(2 people agree) = 0.5

$$\begin{aligned} \Pr(\text{Exact}) &= \Pr(\text{Agree})_1 \times \Pr(\text{Agree})_2 \times \cdots \times \Pr(\text{Agree})_{29} \\ &= 0.5 \times 0.5 \times \cdots \times 0.5 \end{aligned}$$

The E-Harmony Problem

Curse of dimensionality and on-line dating:

eHarmony matches you based on compatibility in the most important areas of life - like values, character, intellect, sense of humor, and 25 other dimensions.

Suppose (for example) 29 dimensions are binary (0,1):
Suppose dimensions are independent:
Pr(2 people agree) = 0.5

$$\begin{aligned} \Pr(\text{Exact}) &= \Pr(\text{Agree})_1 \times \Pr(\text{Agree})_2 \times \cdots \times \Pr(\text{Agree})_{29} \\ &= 0.5 \times 0.5 \times \cdots \times 0.5 \\ &= 0.5^{29} \end{aligned}$$

The E-Harmony Problem

Curse of dimensionality and on-line dating:

eHarmony matches you based on compatibility in the most important areas of life - like values, character, intellect, sense of humor, and 25 other dimensions.

Suppose (for example) 29 dimensions are binary (0,1):
Suppose dimensions are independent:
Pr(2 people agree) = 0.5

$$\begin{aligned}
\Pr(\text{Exact}) &= \Pr(\text{Agree})_1 \times \Pr(\text{Agree})_2 \times \cdots \times \Pr(\text{Agree})_{29} \\
&= 0.5 \times 0.5 \times \cdots \times 0.5 \\
&= 0.5^{29} \\
&\approx 1.8 \times 10^{-9}
\end{aligned}$$

The E-Harmony Problem

Curse of dimensionality and on-line dating:

eHarmony matches you based on compatibility in the most important areas of life - like values, character, intellect, sense of humor, and 25 other dimensions.

Suppose (for example) 29 dimensions are binary (0,1):
Suppose dimensions are independent:
Pr(2 people agree) = 0.5

$$\begin{aligned} \Pr(\text{Exact}) &= \Pr(\text{Agree})_1 \times \Pr(\text{Agree})_2 \times \cdots \times \Pr(\text{Agree})_{29} \\ &= 0.5 \times 0.5 \times \cdots \times 0.5 \\ &= 0.5^{29} \\ &\approx 1.8 \times 10^{-9} \end{aligned}$$

1 in 536,870,912 people

The E-Harmony Problem

Curse of dimensionality and on-line dating:

eHarmony matches you based on compatibility in the most important areas of life - like values, character, intellect, sense of humor, and 25 other dimensions.

Suppose (for example) 29 dimensions are binary (0,1):
Suppose dimensions are independent:
Pr(2 people agree) = 0.5

$$\begin{aligned} \Pr(\text{Exact}) &= \Pr(\text{Agree})_1 \times \Pr(\text{Agree})_2 \times \cdots \times \Pr(\text{Agree})_{29} \\ &= 0.5 \times 0.5 \times \cdots \times 0.5 \\ &= 0.5^{29} \\ &\approx 1.8 \times 10^{-9} \end{aligned}$$

1 in 536,870,912 people
Across many "variables" (events) agreement is harder

Matching with Bias Correction

Matching estimators may behave badly if X contains multiple continuous variables.
Need to adjust matching estimators in the following way:

$$\tilde{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} (Y_i - Y_{j(i)}) - (\widehat{\mu}_0(X_i) - \widehat{\mu}_0(X_{j(i)})),$$

where $\mu_0(x) = E[Y|X = x, D = 0]$ is the population regression function under the control condition and $\widehat{\mu}_0$ is an estimate of μ_0.
$X_i - X_{j(i)}$ is often referred to as the matching discrepancy.
These "bias-corrected" matching estimators behave well even if μ_0 is estimated using a simple linear regression (ie. $\mu_0(x) = \beta_0 + \beta_1 x$) (Abadie and Imbens, 2005)

Matching with Bias Correction

Each treated observation contributes

$$\mu_0(X_i) - \mu_0(X_{j(i)})$$

to the bias.
Bias-corrected matching:

$$\tilde{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \Big((Y_i - Y_{j(i)}) - (\widehat{\mu}_0(X_i) - \widehat{\mu}_0(X_{j(i)})) \Big)$$

The large sample distribution of this estimator (for the case of matching with replacement) is (basically) standard normal. μ_0 is usually estimated using a simple linear regression (ie. $\mu_0(x) = \beta_0 + \beta_1 x$).

In R: `Match(Y,Tr, X,BiasAdjust = TRUE)`

Bias Adjustment with Matched Data

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	6	?	1	3
2	1	?	1	1
3	0	?	1	10
4		0	0	2
5		9	0	3
6		1	0	8

What is $\tilde{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \Big((Y_i - Y_{j(i)}) - (\widehat{\mu}_0(X_i) - \widehat{\mu}_0(X_{j(i)})) \Big)$?

Bias Adjustment with Matched Data

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	6	9	1	3
2	1	0	1	1
3	0	1	1	10
4		0	0	2
5		9	0	3
6		1	0	8

What is $\tilde{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \Big((Y_i - Y_{j(i)}) - (\widehat{\mu}_0(X_i) - \widehat{\mu}_0(X_{j(i)})) \Big)$?

Estimate $\widehat{\mu}_0(x) = \beta_0 + \beta_1 x = 5 - .4x$. Now plug in:

$$\begin{aligned} \widehat{\tau}_{ATT} &= 1/3\{((6-9) - (\widehat{\mu}_0(3) - \widehat{\mu}_0(3))) \\ &+ ((1-0) - (\widehat{\mu}_0(1) - \widehat{\mu}_0(2))) \\ &+ ((0-1) - (\widehat{\mu}_0(10) - \widehat{\mu}_0(8)))\} \\ &= -0.86 \end{aligned}$$

Unadjusted: $1/3((6-9) + (1-0) + (0-1)) = -1$

Bias Adjustment with Matched Data

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	6	9	1	3
2	1	0	1	1
3	0	1	1	10
4		0	0	2
5		9	0	3
6		1	0	8

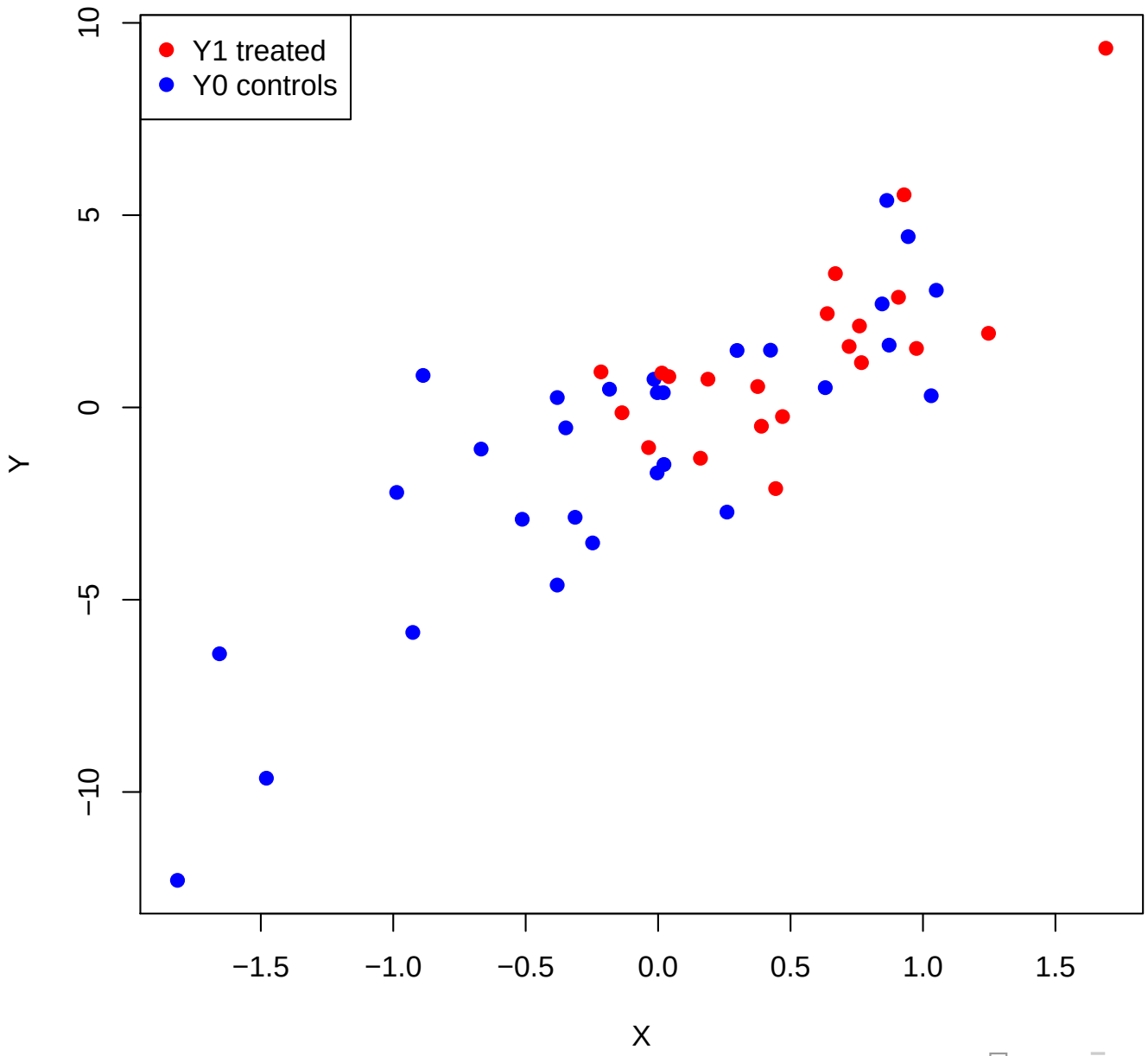
What is $\tilde{\tau}_{ATT} = \frac{1}{N_1} \sum_{D_i=1} \Big((Y_i - Y_{j(i)}) - (\widehat{\mu}_0(X_i) - \widehat{\mu}_0(X_{j(i)})) \Big)$?

Estimate $\widehat{\mu}_0(x) = \beta_0 + \beta_1 x = 5 - .4x$. Now plug in:

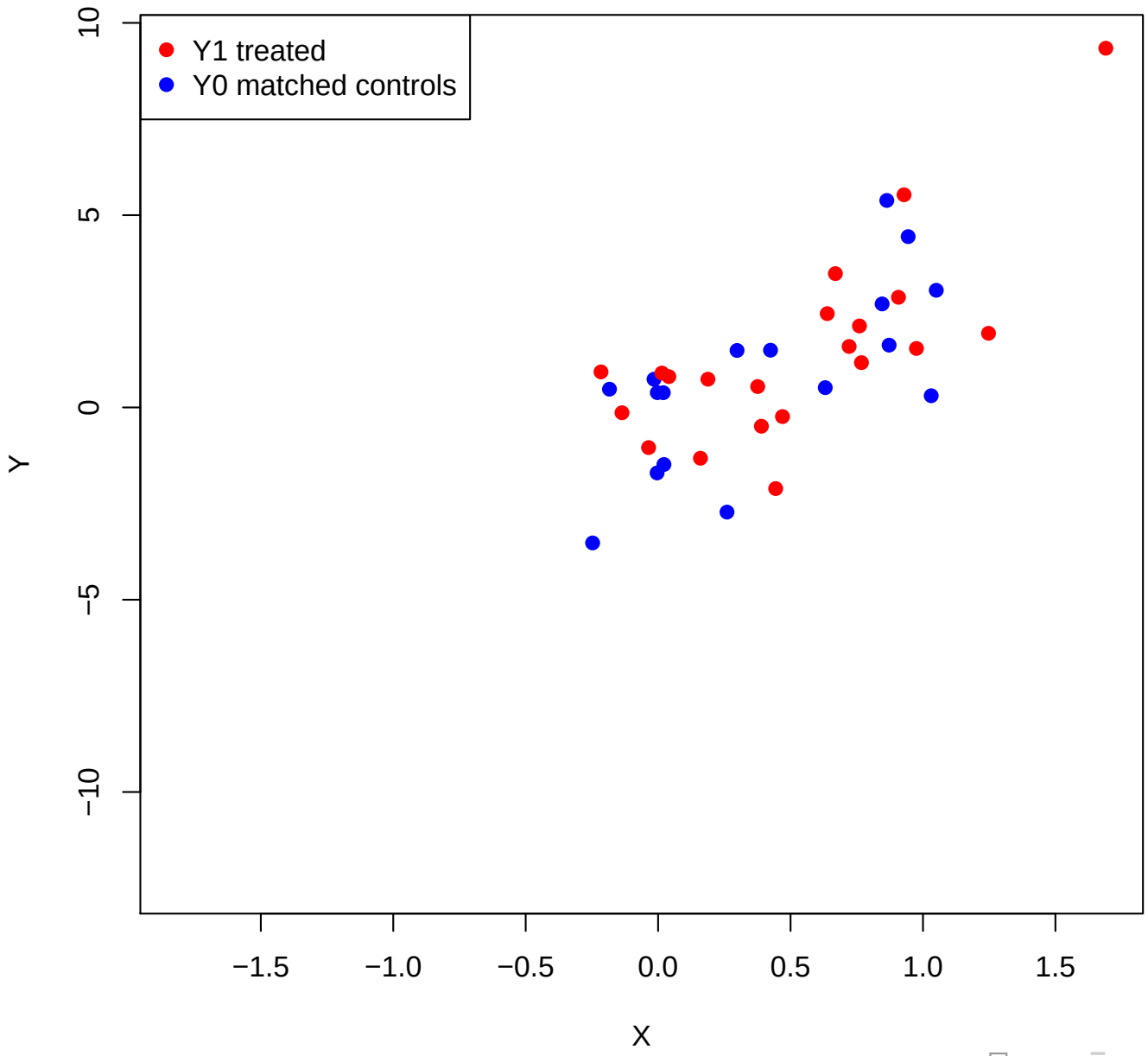
$$\begin{aligned} \widehat{\tau}_{ATT} &= 1/3\{((6-9) - (\widehat{\mu}_0(3) - \widehat{\mu}_0(3))) \\ &+ ((1-0) - (\widehat{\mu}_0(1) - \widehat{\mu}_0(2))) \\ &+ ((0-1) - (\widehat{\mu}_0(10) - \widehat{\mu}_0(8)))\} \\ &= -0.86 \end{aligned}$$

Unadjusted: $1/3((6-9) + (1-0) + (0-1)) = -1$

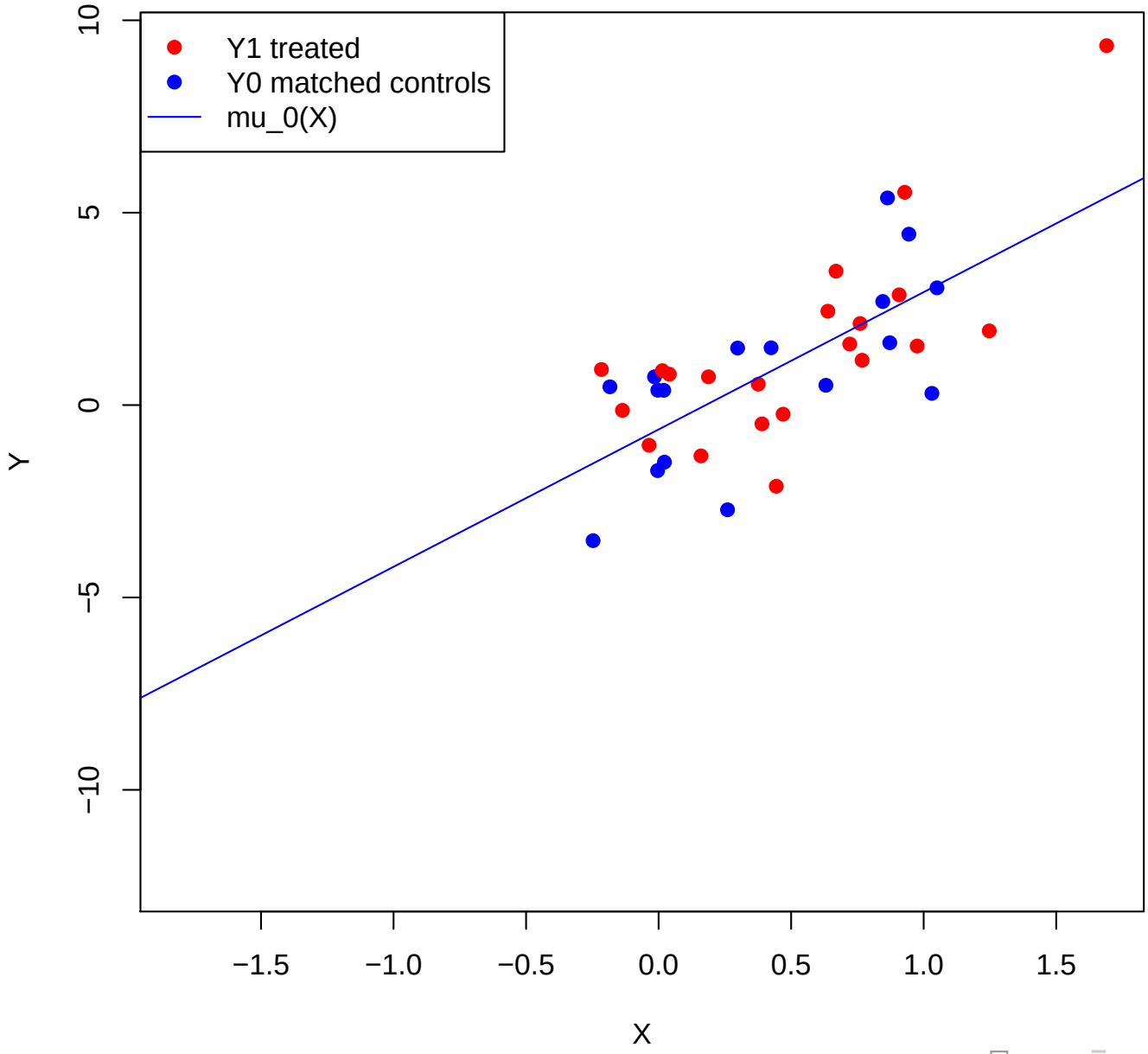
Before Matching



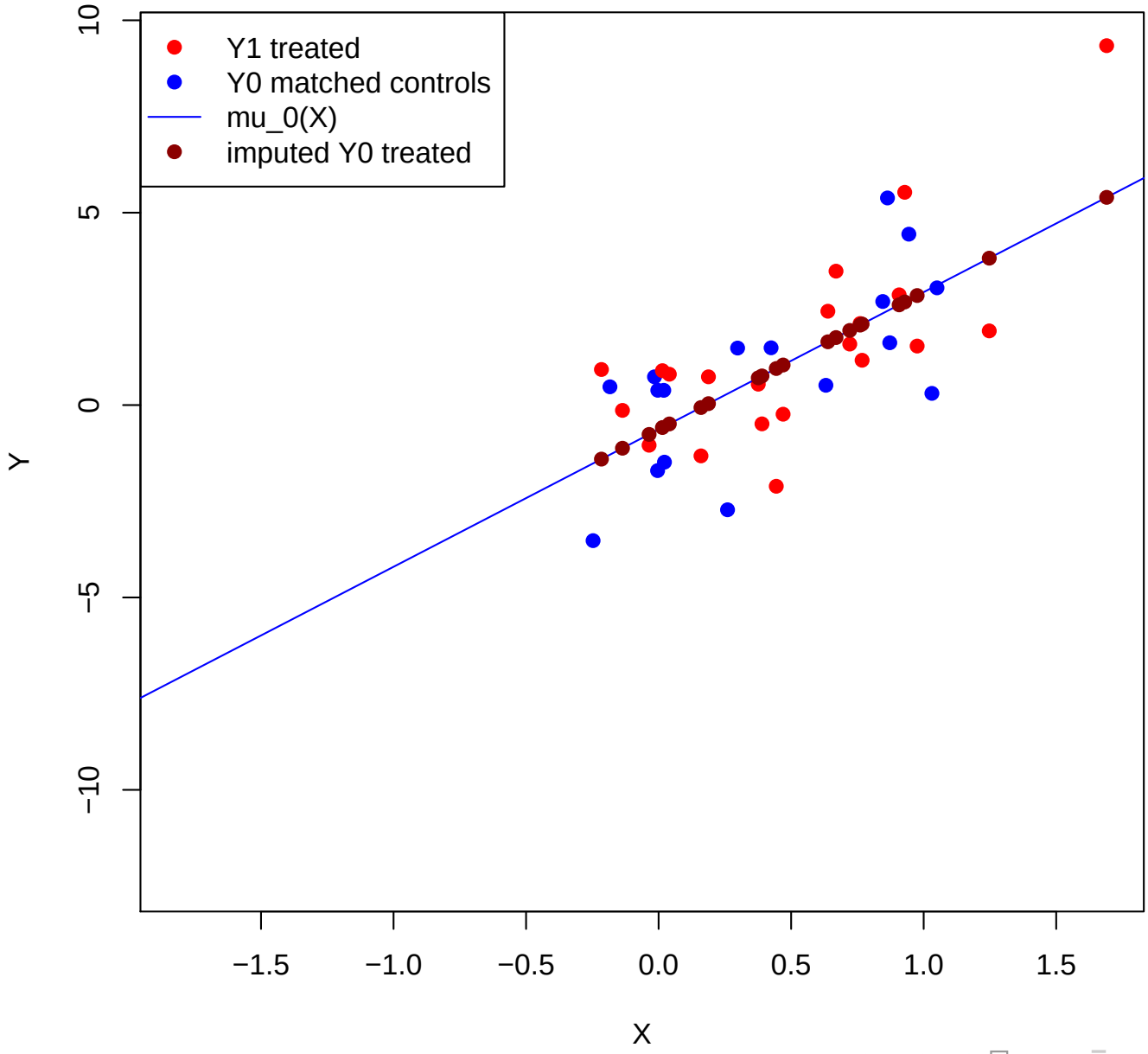
After Matching



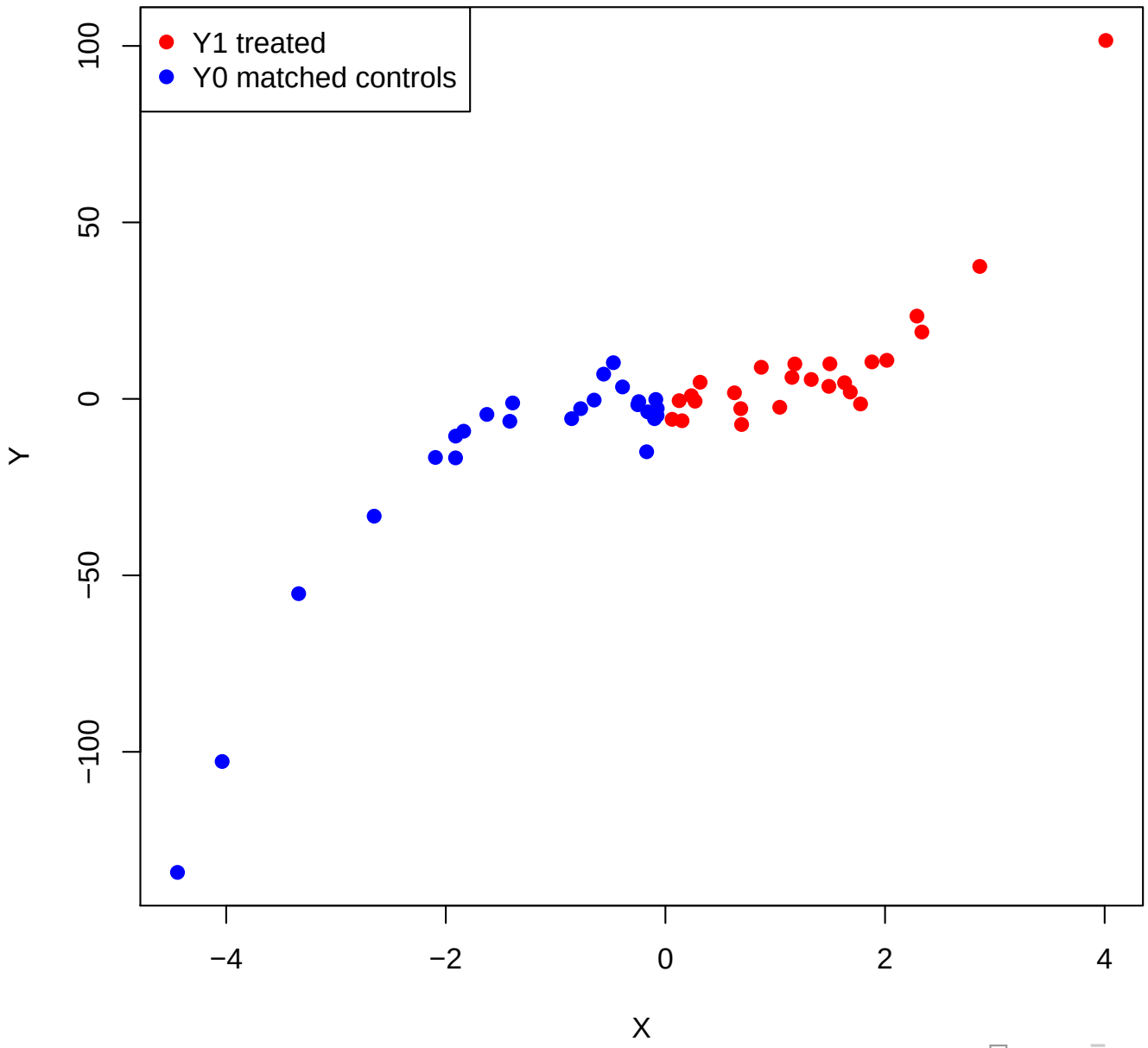
After Matching: Imputation Function



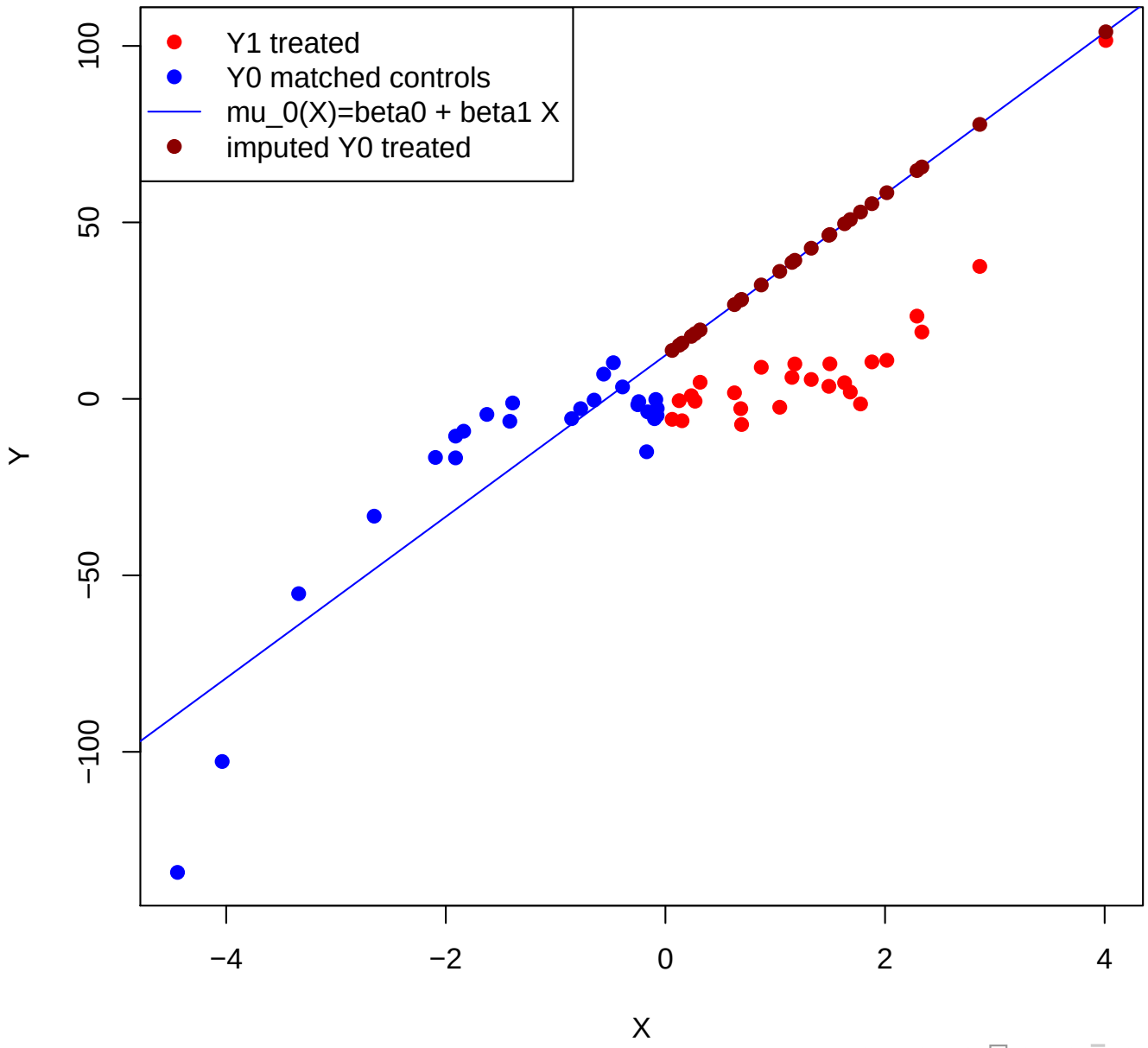
After Matching: Imputation of missing Y_0



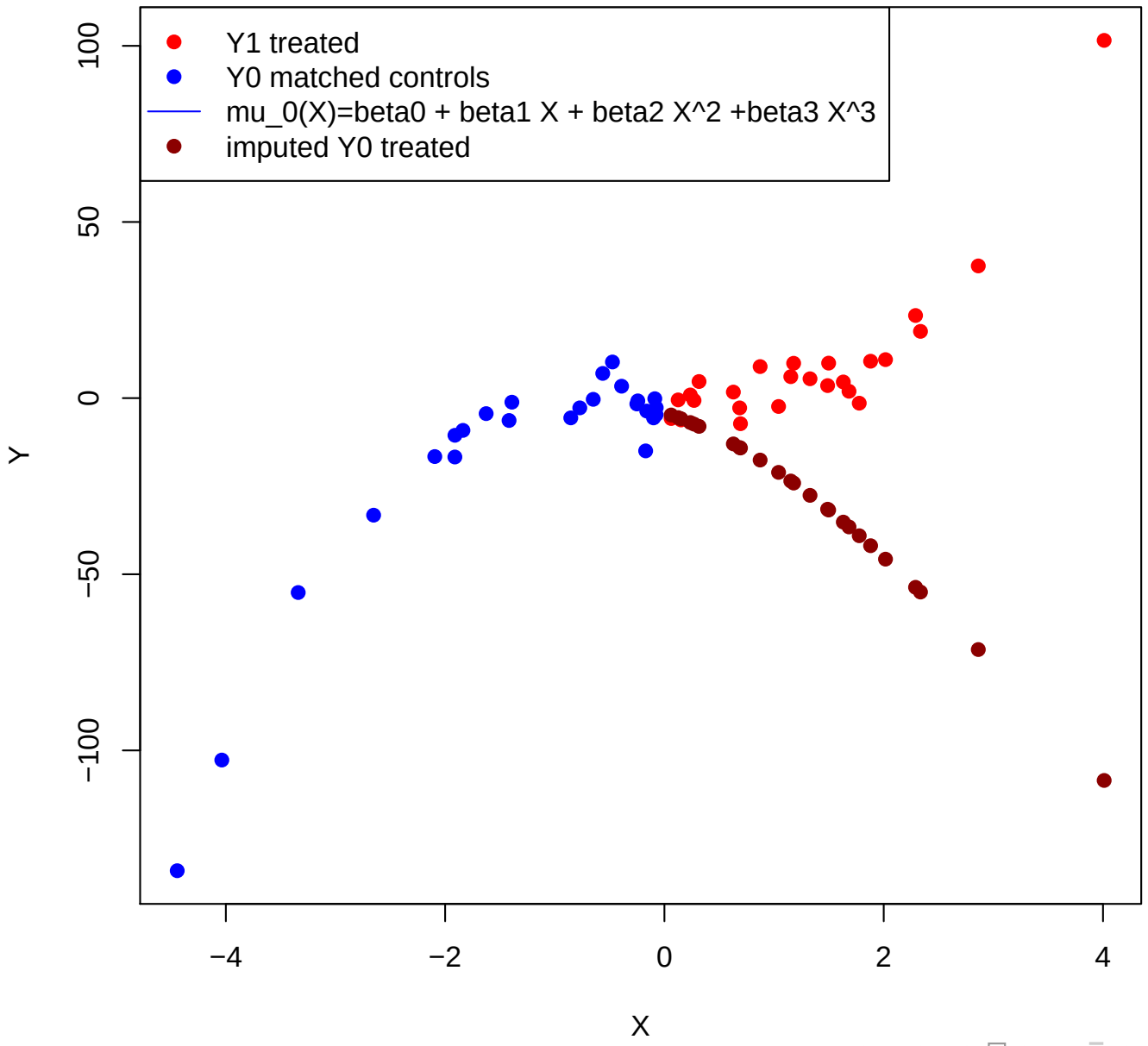
After Matching: No Overlap in Y_0



After Matching: Imputation of missing Y_0



After Matching: Imputation of missing Y_0



Choices when Matching

- With or Without Replacement?
- How many matches?
- Which Matching Algorithm?
  - Genetic Matching
  - Kernel Matching
  - Full Matching
  - Coarsened Exact Matching
  - Matching as Pre-processing
  - Propensity Score Matching
- Use whatever gives you the best balance! Checking balance is important to get a sense for how much extrapolation is needed
  - Should check balance on interactions and higher moments
- With insufficient overlap, all adjustment methods are problematic because we have to heavily rely on a model to impute missing potential outcomes.

Choices when Matching

- With or Without Replacement?
- How many matches?
- Which Matching Algorithm?
  - Genetic Matching
  - Kernel Matching
  - Full Matching
  - Coarsened Exact Matching
  - Matching as Pre-processing
  - Propensity Score Matching
- Use whatever gives you the best balance! Checking balance is important to get a sense for how much extrapolation is needed
  - Should check balance on interactions and higher moments
- With insufficient overlap, all adjustment methods are problematic because we have to heavily rely on a model to impute missing potential outcomes.

Choices when Matching

- With or Without Replacement?
- How many matches?
- Which Matching Algorithm?
  - Genetic Matching
  - Kernel Matching
  - Full Matching
  - Coarsened Exact Matching
  - Matching as Pre-processing
  - Propensity Score Matching
- Use whatever gives you the best balance! Checking balance is important to get a sense for how much extrapolation is needed
  - Should check balance on interactions and higher moments
- With insufficient overlap, all adjustment methods are problematic because we have to heavily rely on a model to impute missing potential outcomes.

Choices when Matching

- With or Without Replacement?
- How many matches?
- Which Matching Algorithm?
  - Genetic Matching
  - Kernel Matching
  - Full Matching
  - Coarsened Exact Matching
  - Matching as Pre-processing
  - Propensity Score Matching
- Use whatever gives you the best balance! Checking balance is important to get a sense for how much extrapolation is needed
  - Should check balance on interactions and higher moments
- With insufficient overlap, all adjustment methods are problematic because we have to heavily rely on a model to impute missing potential outcomes.

Balance Checks

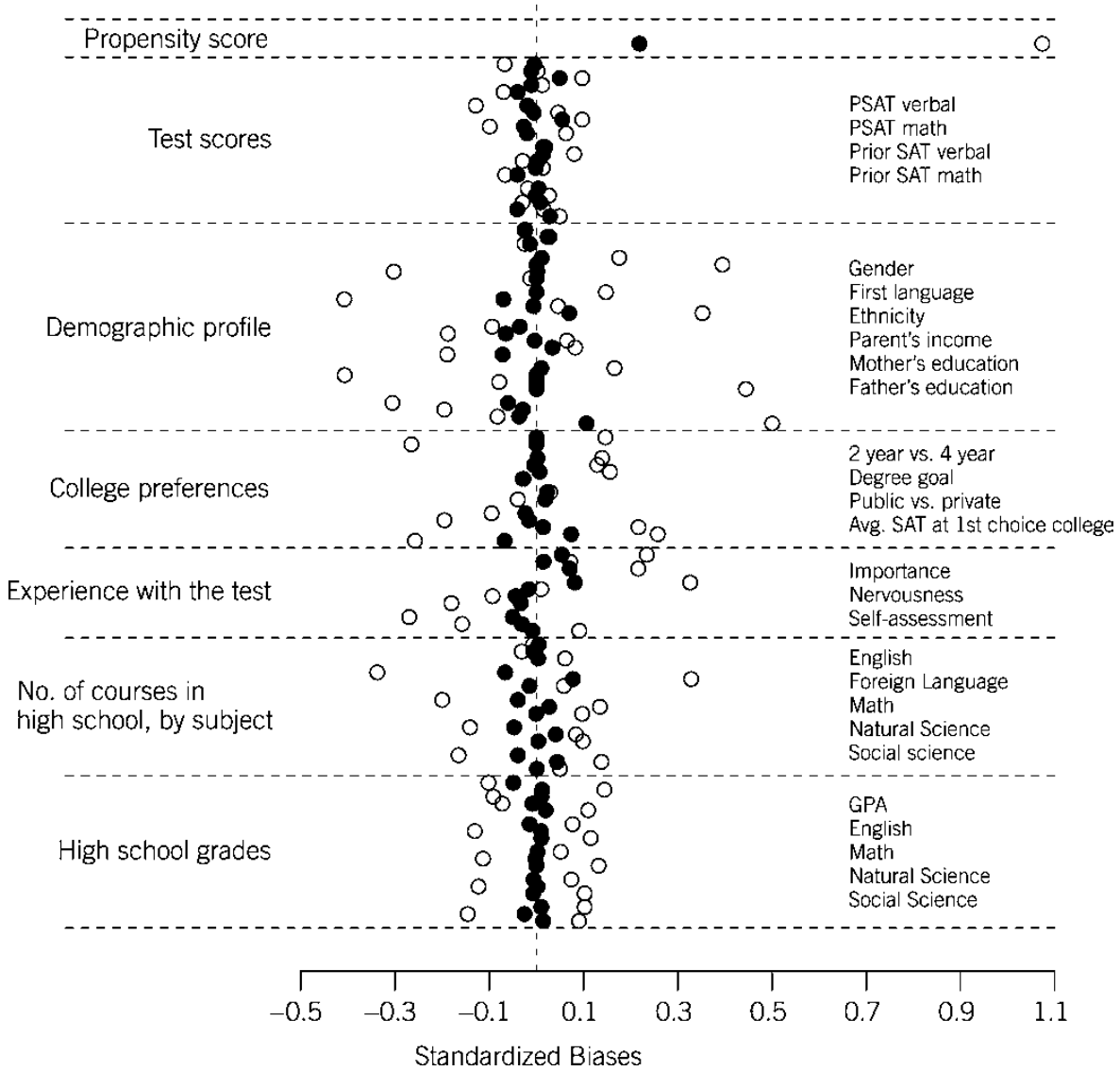


Figure 3. Standardized Biases Without Stratification or Matching, Open Circles, and Under the Optimal [.5, 2] Full Match, Shaded Circles.

Balance Checks

TABLE 2. Balance Summary Statistics and Tests: Russian and Chechen Sweeps

Pretreatment Covariates	Mean Treated	Mean Control	Mean Difference	Std. Bias	Rank Sum Test	K-S Test
Demographics						
Population	8.657	8.606	0.049	0.033	0.708	0.454
Tariqa	0.076	0.048	0.028	0.104	0.331	—
Poverty	1.917	1.931	−0.016	−0.024	0.792	1.000
Spatial						
Elevation	5.078	5.233	−0.155	−0.135	0.140	0.228
Isolation	1.007	1.070	−0.063	−0.096	0.343	0.851
Groznyy	0.131	0.138	−0.007	−0.018	0.864	—
War Dynamics						
TAC	0.241	0.282	−0.041	−0.095	0.424	—
Garrison	0.379	0.414	−0.035	−0.072	0.549	—
Rebel	0.510	0.441	0.070	0.139	0.240	—
Selection						
Presweep violence	3.083	3.117	−0.034	0.009	0.454	0.292
Large-scale theft	0.034	0.055	−0.021	−0.115	0.395	—
Killing	0.117	0.090	0.027	0.084	0.443	—
Violence Inflicted						
Total abuse	0.970	0.833	0.137	0.124	0.131	0.454
Prior sweeps	1.729	1.812	−0.090	−0.089	0.394	0.367
Other						
Month	7.428	6.986	0.442	0.130	0.260	0.292
Year	2004.159	2004.110	0.049	0.043	0.889	1.000

Note: 145 matched pairs. Matching with replacement.

Variance of the Matching Estimator

- The sampling variance of the matching estimator (conditional on covariates and treatment indicators) can be written as the following:

$$V(\hat{\tau}_{ATT}|\mathbf{X},\mathbf{D}) = \sum_{i=1}^{N} \lambda_i(\mathbf{X},\mathbf{W}) \cdot \sigma^2_{D_i}(X_i)$$

  where $\lambda_i(\mathbf{X},\mathbf{W})$ is just a set of weights which account for matching with replacement and multiple controls per treatment unit.
- σ_i^2 is the unit level variance. Abadie and Imbens suggest matching to estimate this quantity:
  - Let $v(i)$ be the closest unit to i with the same treatment indicator ($D_{v(i)} = D_i$). The sample variance of the outcome variable for these 2 units can be used to estimate $\sigma^2_{D_i}(X_i)$:

$$\hat{\sigma}^2_{D_i}(X_i) = (Y_i - Y_{v(i)})^2/2$$

  - Robust standard errors also available
- Do not use the bootstrap!

Variance of the Matching Estimator

- The sampling variance of the matching estimator (conditional on covariates and treatment indicators) can be written as the following:

$$V(\hat{\tau}_{ATT}|\mathbf{X},\mathbf{D}) = \sum_{i=1}^{N} \lambda_i(\mathbf{X},\mathbf{W}) \cdot \sigma^2_{D_i}(X_i)$$

  where $\lambda_i(\mathbf{X},\mathbf{W})$ is just a set of weights which account for matching with replacement and multiple controls per treatment unit.
- σ_i^2 is the unit level variance. Abadie and Imbens suggest matching to estimate this quantity:
  - Let $v(i)$ be the closest unit to i with the same treatment indicator ($D_{v(i)} = D_i$). The sample variance of the outcome variable for these 2 units can be used to estimate $\sigma^2_{D_i}(X_i)$:

$$\hat{\sigma}^2_{D_i}(X_i) = (Y_i - Y_{v(i)})^2/2$$

  - Robust standard errors also available
- Do not use the bootstrap!

Variance of the Matching Estimator

- The sampling variance of the matching estimator (conditional on covariates and treatment indicators) can be written as the following:

$$V(\hat{\tau}_{ATT}|\mathbf{X},\mathbf{D}) = \sum_{i=1}^{N} \lambda_i(\mathbf{X},\mathbf{W}) \cdot \sigma^2_{D_i}(X_i)$$

  where $\lambda_i(\mathbf{X},\mathbf{W})$ is just a set of weights which account for matching with replacement and multiple controls per treatment unit.
- σ_i^2 is the unit level variance. Abadie and Imbens suggest matching to estimate this quantity:
  - Let $v(i)$ be the closest unit to i with the same treatment indicator ($D_{v(i)} = D_i$). The sample variance of the outcome variable for these 2 units can be used to estimate $\sigma^2_{D_i}(X_i)$:

$$\hat{\sigma}^2_{D_i}(X_i) = (Y_i - Y_{v(i)})^2/2$$

  - Robust standard errors also available
- Do not use the bootstrap!

Useful Matching Functions

The workhorse model is the `Match()` function in the `Matching` package:

```
Match(Y = NULL, Tr, X, Z = X, V = rep(1, length(Y)),
    estimand = "ATT", M = 1, BiasAdjust = FALSE, exact = NULL,
    caliper = NULL,  replace = TRUE, ties = TRUE,
    CommonSupport = FALSE, Weight = 1, Weight.matrix = NULL,
    weights = NULL, Var.calc = 0,  sample = FALSE, restrict = NULL,
    match.out = NULL, distance.tolerance = 1e-05,
    tolerance = sqrt(.Machine$double.eps), version = "standard")
```

Default distance metric (`Weight=1`) is normalized Euclidean distance

- `MatchBalance(formu)` for balance checking
- `GenMatch()` for genetic matching

Identification with Propensity Scores

Definition

Propensity score is defined as the selection probability conditional on the confounding variables: $\pi(X) = \Pr(D = 1|X)$

Identification Assumption

1. $(Y_1, Y_0) \perp\!\!\!\perp D|X$ *(selection on observables)*
2. $0 < \Pr(D = 1|X) < 1$ *with probability one (common support)*

Identfication Result

Under selection on observables we have $(Y_1, Y_0) \perp\!\!\!\perp D|\pi(X)$*, ie. conditioning on the propensity score is enough to have independence between the treatment indicator and potential outcomes. Implies substantial dimension reduction.*

Identification with Propensity Scores

Proof.

Show that $\Pr(D = 1|Y_0, Y_1, \pi(X)) = \Pr(D = 1|\pi(X)) = \pi(X)$, implying independence of (Y_0, Y_1) and D conditional on $\pi(X)$.

$$\begin{aligned}
\Pr(D = 1|Y_1, Y_0, \pi(X)) &= E[D|Y_1, Y_0, \pi(X)] \\
&= E\left[E[D|Y_1, Y_0, X]|Y_1, Y_0, \pi(X)\right] \text{ (LIE)} \\
&= E\left[E[D|X]|Y_1, Y_0, \pi(X)\right] \text{ (SOO)} \\
&= E\left[\pi(X)|Y_1, Y_0, \pi(X)\right] \\
&= \pi(X)
\end{aligned}$$

Using a similar argument

$$\begin{aligned}
\Pr(D = 1|\pi(X)) &= E[D|\pi(X)] = E[E[D|X]|\pi(x)] \\
&= E[\pi(X)|\pi(X)] = \pi(X)
\end{aligned}$$

therefore $\Pr(D = 1|Y_1, Y_0, \pi(X)) = \Pr(D = 1|\pi(X))$ □

Identification with Propensity Scores

Proof.

Show that $\Pr(D = 1|Y_0, Y_1, \pi(X)) = \Pr(D = 1|\pi(X)) = \pi(X)$, implying independence of (Y_0, Y_1) and D conditional on $\pi(X)$.

$$\begin{aligned}
\Pr(D = 1|Y_1, Y_0, \pi(X)) &= E[D|Y_1, Y_0, \pi(X)] \\
&= E\left[E[D|Y_1, Y_0, X]|Y_1, Y_0, \pi(X)\right] \text{ (LIE)} \\
&= E\left[E[D|X]|Y_1, Y_0, \pi(X)\right] \text{ (SOO)} \\
&= E\left[\pi(X)|Y_1, Y_0, \pi(X)\right] \\
&= \pi(X)
\end{aligned}$$

Using a similar argument

$$\begin{aligned}
\Pr(D = 1|\pi(X)) &= E[D|\pi(X)] = E[E[D|X]|\pi(x)] \\
&= E[\pi(X)|\pi(X)] = \pi(X)
\end{aligned}$$

therefore $\Pr(D = 1|Y_1, Y_0, \pi(X)) = \Pr(D = 1|\pi(X))$ □

Identification with Propensity Scores

Proof.

Show that $\Pr(D = 1|Y_0, Y_1, \pi(X)) = \Pr(D = 1|\pi(X)) = \pi(X)$, implying independence of (Y_0, Y_1) and D conditional on $\pi(X)$.

$$\begin{aligned}
\Pr(D = 1|Y_1, Y_0, \pi(X)) &= E[D|Y_1, Y_0, \pi(X)] \\
&= E\left[E[D|Y_1, Y_0, X]|Y_1, Y_0, \pi(X)\right] \text{ (LIE)} \\
&= E\left[E[D|X]|Y_1, Y_0, \pi(X)\right] \text{ (SOO)} \\
&= E\left[\pi(X)|Y_1, Y_0, \pi(X)\right] \\
&= \pi(X)
\end{aligned}$$

Using a similar argument

$$\begin{aligned}
\Pr(D = 1|\pi(X)) &= E[D|\pi(X)] = E[E[D|X]|\pi(x)] \\
&= E[\pi(X)|\pi(X)] = \pi(X)
\end{aligned}$$

therefore $\Pr(D = 1|Y_1, Y_0, \pi(X)) = \Pr(D = 1|\pi(X))$ □

Identification with Propensity Scores

Proof.

Show that $\Pr(D = 1|Y_0, Y_1, \pi(X)) = \Pr(D = 1|\pi(X)) = \pi(X)$, implying independence of (Y_0, Y_1) and D conditional on $\pi(X)$.

$$\begin{aligned}
\Pr(D = 1|Y_1, Y_0, \pi(X)) &= E[D|Y_1, Y_0, \pi(X)] \\
&= E\left[E[D|Y_1, Y_0, X]|Y_1, Y_0, \pi(X)\right] \text{ (LIE)} \\
&= E\left[E[D|X]|Y_1, Y_0, \pi(X)\right] \text{ (SOO)} \\
&= E\left[\pi(X)|Y_1, Y_0, \pi(X)\right] \\
&= \pi(X)
\end{aligned}$$

Using a similar argument

$$\begin{aligned}
\Pr(D = 1|\pi(X)) &= E[D|\pi(X)] = E[E[D|X]|\pi(x)] \\
&= E[\pi(X)|\pi(X)] = \pi(X)
\end{aligned}$$

therefore $\Pr(D = 1|Y_1, Y_0, \pi(X)) = \Pr(D = 1|\pi(X))$ □

Identification with Propensity Scores

Proof.

Show that $\Pr(D = 1|Y_0, Y_1, \pi(X)) = \Pr(D = 1|\pi(X)) = \pi(X)$, implying independence of (Y_0, Y_1) and D conditional on $\pi(X)$.

$$\begin{aligned}
\Pr(D = 1|Y_1, Y_0, \pi(X)) &= E[D|Y_1, Y_0, \pi(X)] \\
&= E\left[E[D|Y_1, Y_0, X]|Y_1, Y_0, \pi(X)\right] \text{ (LIE)} \\
&= E\left[E[D|X]|Y_1, Y_0, \pi(X)\right] \text{ (SOO)} \\
&= E\left[\pi(X)|Y_1, Y_0, \pi(X)\right] \\
&= \pi(X)
\end{aligned}$$

Using a similar argument

$$\begin{aligned}
\Pr(D = 1|\pi(X)) &= E[D|\pi(X)] = E[E[D|X]|\pi(x)] \\
&= E[\pi(X)|\pi(X)] = \pi(X)
\end{aligned}$$

therefore $\Pr(D = 1|Y_1, Y_0, \pi(X)) = \Pr(D = 1|\pi(X))$ □

Identification with Propensity Scores

Proof.

Show that $\Pr(D = 1|Y_0, Y_1, \pi(X)) = \Pr(D = 1|\pi(X)) = \pi(X)$, implying independence of (Y_0, Y_1) and D conditional on $\pi(X)$.

$$\begin{aligned}
\Pr(D = 1|Y_1, Y_0, \pi(X)) &= E[D|Y_1, Y_0, \pi(X)] \\
&= E\left[E[D|Y_1, Y_0, X]|Y_1, Y_0, \pi(X)\right] \text{ (LIE)} \\
&= E\left[E[D|X]|Y_1, Y_0, \pi(X)\right] \text{ (SOO)} \\
&= E\left[\pi(X)|Y_1, Y_0, \pi(X)\right] \\
&= \pi(X)
\end{aligned}$$

Using a similar argument

$$\begin{aligned}
\Pr(D = 1|\pi(X)) &= E[D|\pi(X)] = E[E[D|X]|\pi(x)] \\
&= E[\pi(X)|\pi(X)] = \pi(X)
\end{aligned}$$

therefore $\Pr(D = 1|Y_1, Y_0, \pi(X)) = \Pr(D = 1|\pi(X))$ □

Matching on the Propensity Score

Corollary

If $(Y_1, Y_0) \perp\!\!\!\perp D|X$, *then*

$$E[Y|D=1, \pi(X)=\bar{\pi}] - E[Y|D=0, \pi(X)=\bar{\pi}] = E[Y_1 - Y_0|\pi(X)=\bar{\pi}]$$

Suggests a two step procedure to estimate causal effects under selection on observables:

1. Estimate the propensity score $\pi(X) = P(D=1|X)$ (e.g. using logit/probit regression, machine learning methods, etc)
2. Match or subclassify on propensity score.

Matching on the Propensity Score

Corollary

If $(Y_1, Y_0) \perp\!\!\!\perp D|X$, *then*

$$E[Y|D=1, \pi(X)=\bar{\pi}] - E[Y|D=0, \pi(X)=\bar{\pi}] = E[Y_1 - Y_0|\pi(X)=\bar{\pi}]$$

Suggests a two step procedure to estimate causal effects under selection on observables:

1. Estimate the propensity score $\pi(X) = P(D=1|X)$ (e.g. using logit/probit regression, machine learning methods, etc)
2. Match or subclassify on propensity score.

Estimating the Propensity Score

- Given selection on observables we have $(Y_1, Y_0) \perp\!\!\!\perp D | \pi(X)$ which implies the balancing property of the propensity score:

$$\Pr(X|D = 1, \pi(X)) = \Pr(X|D = 0, \pi(X))$$

- We can use this to check if our estimated propensity score actually produces balance: $P(X|D = 1, \widehat{\pi}(X)) = P(X|D = 0, \widehat{\pi}(X))$
- To properly model the assignment mechanism, we need to include important confounders correlated with treatment and outcome
- Need to find the correct functional form, miss-specified propensity scores can lead to bias. Any methods can be used (probit, logit, etc.)
- Estimate $\mapsto$ Check Balance $\mapsto$ Re-estimate $\mapsto$ Check Balance

Estimating the Propensity Score

- Given selection on observables we have $(Y_1, Y_0) \perp\!\!\!\perp D | \pi(X)$ which implies the balancing property of the propensity score:

$$\Pr(X|D = 1, \pi(X)) = \Pr(X|D = 0, \pi(X))$$

- We can use this to check if our estimated propensity score actually produces balance: $P(X|D = 1, \widehat{\pi}(X)) = P(X|D = 0, \widehat{\pi}(X))$
- To properly model the assignment mechanism, we need to include important confounders correlated with treatment and outcome
- Need to find the correct functional form, miss-specified propensity scores can lead to bias. Any methods can be used (probit, logit, etc.)
- Estimate $\mapsto$ Check Balance $\mapsto$ Re-estimate $\mapsto$ Check Balance

Estimating the Propensity Score

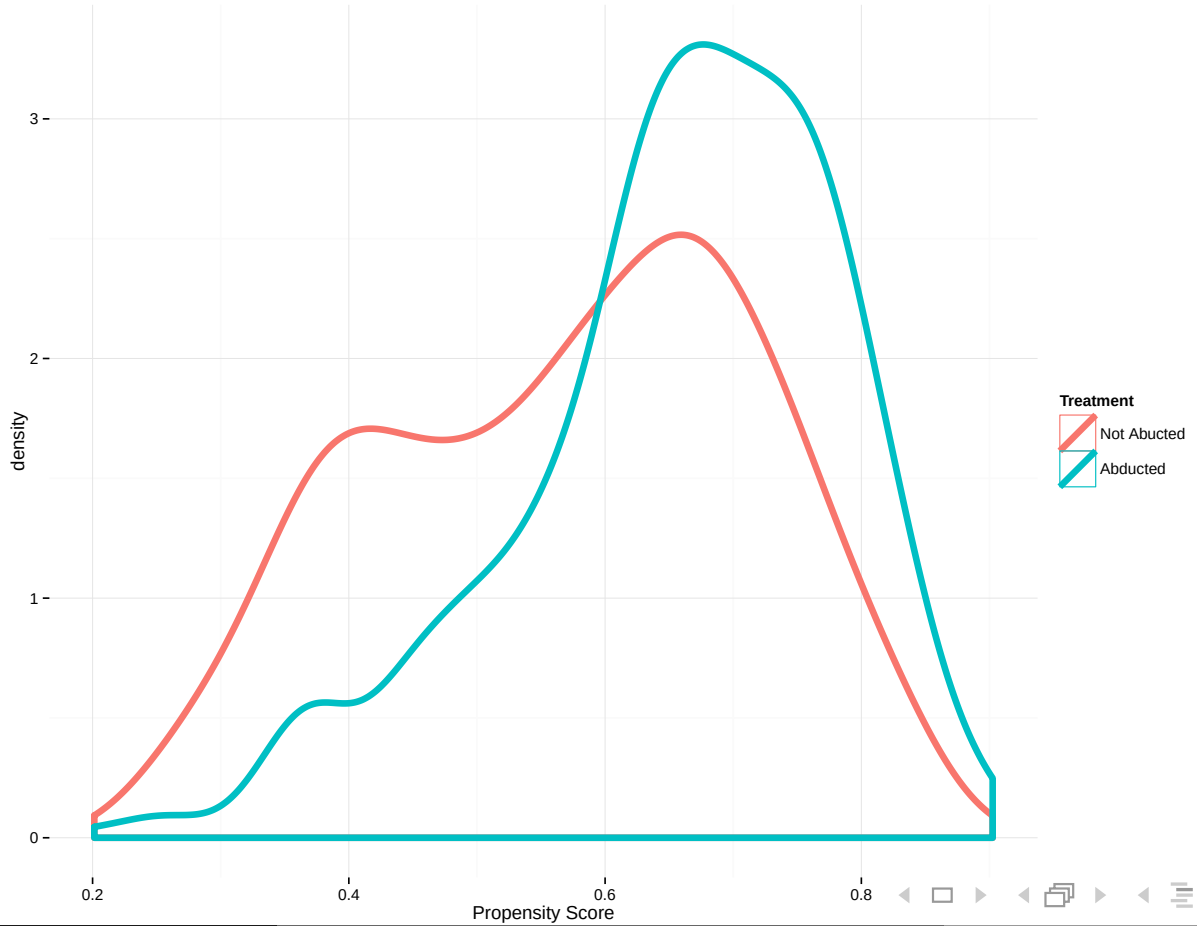
- Given selection on observables we have $(Y_1, Y_0) \perp\!\!\!\perp D | \pi(X)$ which implies the balancing property of the propensity score:

$$\Pr(X|D = 1, \pi(X)) = \Pr(X|D = 0, \pi(X))$$

- We can use this to check if our estimated propensity score actually produces balance: $P(X|D = 1, \widehat{\pi}(X)) = P(X|D = 0, \widehat{\pi}(X))$
- To properly model the assignment mechanism, we need to include important confounders correlated with treatment and outcome
- Need to find the correct functional form, miss-specified propensity scores can lead to bias. Any methods can be used (probit, logit, etc.)
- Estimate $\mapsto$ Check Balance $\mapsto$ Re-estimate $\mapsto$ Check Balance

Example: Blattman (2010)

```
pscore.fmla <- as.formula(paste("abd~",paste(names(covar),collapse="+")))
abd <- data$abd
pscore_model <- glm(pscore.fmla, data = data,
family = binomial(link = logit))
pscore <- predict(pscore_model, type = "response")
```



Weighting on the Propensity Score

Provided that the relevant moments exists, if $Y_1, Y_0 \perp\!\!\!\perp D|X$, then

$$\begin{aligned}
\tau_{ATE} &= E[Y_1 - Y_0] = E\left[Y \cdot \frac{D - \pi(X)}{\pi(X) \cdot (1 - \pi(X))}\right] \\
\tau_{ATT} &= E[Y_1 - Y_0|D = 1] = \frac{1}{P(D = 1)} \cdot E\left[Y \cdot \frac{D - \pi(X)}{1 - \pi(X)}\right]
\end{aligned}$$

Proof.

Consider

$$\begin{aligned}
& E\left[Y \frac{D - \pi(X)}{\pi(X)(1 - \pi(X))}\middle| X\right] \\
= \; & E\left[\frac{Y}{\pi(X)}\middle| X, D = 1\right]\pi(X) + E\left[\frac{-Y}{1 - \pi(X)}\middle| X, D = 0\right](1 - \pi(X)) \\
= \; & E[Y|X, D = 1] - E[Y|X, D = 0]
\end{aligned}$$

And the results of the proposition follow from integration over $\Pr(X)$ and $\Pr(X|D = 1)$. □

Weighting on the Propensity Score

$$\begin{aligned}
\tau_{ATE} &= E[Y_1 - Y_0] = E\left[Y \cdot \frac{D - \pi(X)}{\pi(X) \cdot (1 - \pi(X))}\right] \\
\tau_{ATT} &= E[Y_1 - Y_0 | D = 1] = \frac{1}{\Pr(D = 1)} \cdot E\left[Y \cdot \frac{D - \pi(X)}{1 - \pi(X)}\right]
\end{aligned}$$

How do we estimate this? Analogy principle suggest a two step procedure:

1. Estimate the propensity score ($\widehat{\pi}(X)$)
2. Use sample averages and the estimated propensity score to produce analog estimators of τ_{ATE} and τ_{ATT}:

$$\begin{aligned}
\widehat{\tau}_{ATE} &= \frac{1}{N} \sum_{i=1}^{N} Y_i \cdot \frac{D_i - \widehat{\pi}(X_i)}{\widehat{\pi}(X_i) \cdot (1 - \widehat{\pi}(X_i))}, \\
\widehat{\tau}_{ATT} &= \frac{1}{N_1} \sum_{i=1}^{N} Y_i \cdot \frac{D_i - \widehat{\pi}(X_i)}{1 - \widehat{\pi}(X_i)}
\end{aligned}$$

Weighting on the Propensity Score

$$\begin{aligned}
\tau_{ATE} &= E[Y_1 - Y_0] = E\left[Y \cdot \frac{D - \pi(X)}{\pi(X) \cdot (1 - \pi(X))}\right] \\
\tau_{ATT} &= E[Y_1 - Y_0 | D = 1] = \frac{1}{\Pr(D = 1)} \cdot E\left[Y \cdot \frac{D - \pi(X)}{1 - \pi(X)}\right]
\end{aligned}$$

How do we estimate this? Analogy principle suggest a two step procedure:

1. Estimate the propensity score ($\widehat{\pi}(X)$)
2. Use sample averages and the estimated propensity score to produce analog estimators of τ_{ATE} and τ_{ATT}:

$$\begin{aligned}
\widehat{\tau}_{ATE} &= \frac{1}{N} \sum_{i=1}^{N} Y_i \cdot \frac{D_i - \widehat{\pi}(X_i)}{\widehat{\pi}(X_i) \cdot (1 - \widehat{\pi}(X_i))}, \\
\widehat{\tau}_{ATT} &= \frac{1}{N_1} \sum_{i=1}^{N} Y_i \cdot \frac{D_i - \widehat{\pi}(X_i)}{1 - \widehat{\pi}(X_i)}
\end{aligned}$$

Identification under Selection on Observables: Regression

Consider the linear regression of $Y_i = \beta_0 + \tau D_i + X_i'\beta + \epsilon_i$.
Given selection on observables, there are mainly three identification scenarios:

1. Constant treatment effects and outcomes are linear in X
   - τ will provide unbiased and consistent estimates of ATE.
2. Constant treatment effects and unknown functional form
   - τ will provide well-defined linear approximation to the average causal response function $E[Y|D=1,X] - E[Y|D=0,X]$. Approximation may be very poor if $E[Y|D,X]$ is misspecified and then τ may be biased for the ATE.
3. Heterogeneous treatment effects (τ differs for different values of X)
   - If outcomes are linear in X, τ is unbiased and consistent estimator for conditional-variance-weighted average of the underlying causal effects. This average is often different from the ATE.

Identification under Selection on Observables: Regression

Consider the linear regression of $Y_i = \beta_0 + \tau D_i + X_i'\beta + \epsilon_i$.
Given selection on observables, there are mainly three identification scenarios:

1. Constant treatment effects and outcomes are linear in X
   - τ will provide unbiased and consistent estimates of ATE.

2. Constant treatment effects and unknown functional form
   - τ will provide well-defined linear approximation to the average causal response function $E[Y|D=1,X] - E[Y|D=0,X]$. Approximation may be very poor if $E[Y|D,X]$ is misspecified and then τ may be biased for the ATE.

3. Heterogeneous treatment effects (τ differs for different values of X)
   - If outcomes are linear in X, τ is unbiased and consistent estimator for conditional-variance-weighted average of the underlying causal effects. This average is often different from the ATE.

Identification under Selection on Observables: Regression

Consider the linear regression of $Y_i = \beta_0 + \tau D_i + X_i'\beta + \epsilon_i$.
Given selection on observables, there are mainly three identification scenarios:

1. Constant treatment effects and outcomes are linear in X
   - τ will provide unbiased and consistent estimates of ATE.

2. Constant treatment effects and unknown functional form
   - τ will provide well-defined linear approximation to the average causal response function $E[Y|D=1,X] - E[Y|D=0,X]$. Approximation may be very poor if $E[Y|D,X]$ is misspecified and then τ may be biased for the ATE.

3. Heterogeneous treatment effects (τ differs for different values of X)
   - If outcomes are linear in X, τ is unbiased and consistent estimator for conditional-variance-weighted average of the underlying causal effects. This average is often different from the ATE.

Identification under Selection on Observables: Regression

Identification Assumption

1. *Constant treatment effect:* $\tau = Y_{1i} - Y_{0i}$ *for all* i
2. *Control outcome is linear in* X*:* $Y_{0i} = \beta_0 + X_i'\beta + \epsilon_i$ *with* $\epsilon_i \perp\!\!\!\perp X_i$ *(no omitted variables and linearly separable confounding)*

Identfication Result

Then $\tau_{ATE} = E[Y_1 - Y_0]$ *is identified by a regression of the observed outcome on the covariates and the treatment indicator*

$Y_i = \beta_0 + \tau D_i + X_i'\beta + \epsilon_i$

Regression with Heterogeneous Effects

What is regression estimating when we allow for heterogeneity?

Suppose that we wanted to estimate τ_{OLS} using a fully saturated regression model:

$$Y_i = \sum_x B_{xi}\beta_x + \tau_{OLS} D_i + e_i$$

where B_{xi} is a dummy variable for unique combination of X_i.

Because this regression is fully saturated, it is linear in the covariates (i.e. linearity assumption holds by construction).

Regression with Heterogeneous Effects

What is regression estimating when we allow for heterogeneity? Suppose that we wanted to estimate τ_{OLS} using a fully saturated regression model:

$$Y_i = \sum_x B_{xi}\beta_x + \tau_{OLS} D_i + e_i$$

where B_{xi} is a dummy variable for unique combination of X_i. Because this regression is fully saturated, it is linear in the covariates (i.e. linearity assumption holds by construction).

Heterogenous Treatment Effects

With two X strata there are two stratum-specific average causal effects that are averaged to obtain the ATE or ATT.
Subclassification weights the stratum-effects by the marginal distribution of X, i.e. weights are proportional to the share of units in each stratum:

$$\begin{aligned}\tau_{ATE} &= (E[Y|D=1,X=1] - E[Y|D=0,X=1])\Pr(X=1) \\ &+ (E[Y|D=1,X=2] - E[Y|D=0,X=2])\Pr(X=2)\end{aligned}$$

Regression weights by the marginal distribution of X and the conditional variance of $V[D|X]$ in each stratum:

$$\begin{aligned}\tau_{OLS} &= (E[Y|D=1,X=1] - E[Y|D=0,X=1])\frac{V[D|X=1]\Pr(X=1)}{\sum_X Var[D|X=x]\Pr(X=x)} \\ &+ (E[Y|D=1,X=2] - E[Y|D=0,X=2])\frac{V[D|X=2]\Pr(X=2)}{\sum_X V[D|X=x]\Pr(X=x)}\end{aligned}$$

- So strata with a higher $V[D|X]$ receive higher weight. These are the strata with propensity scores close to .5
- Strata with propensity score close to 0 or 1 receive lower weight
- OLS is a minimum-variance estimator of τ_{OLS} so it downweights strata where the average causal effects are less precisely estimated.

Heterogenous Treatment Effects

With two X strata there are two stratum-specific average causal effects that are averaged to obtain the ATE or ATT.
Subclassification weights the stratum-effects by the marginal distribution of X, i.e. weights are proportional to the share of units in each stratum:

$$\begin{aligned}\tau_{ATE} &= (E[Y|D=1,X=1] - E[Y|D=0,X=1])\Pr(X=1) \\ &+ (E[Y|D=1,X=2] - E[Y|D=0,X=2])\Pr(X=2)\end{aligned}$$

Regression weights by the marginal distribution of X and the conditional variance of $V[D|X]$ in each stratum:

$$\begin{aligned}\tau_{OLS} &= (E[Y|D=1,X=1] - E[Y|D=0,X=1])\frac{V[D|X=1]\Pr(X=1)}{\sum_X Var[D|X=x]\Pr(X=x)} \\ &+ (E[Y|D=1,X=2] - E[Y|D=0,X=2])\frac{V[D|X=2]\Pr(X=2)}{\sum_X V[D|X=x]\Pr(X=x)}\end{aligned}$$

- Whenever both weighting components are misaligned (e.g. the PS is close to 0 or 1 for relatively large strata) then τ_{OLS} diverges from τ_{ATE} or τ_{ATT}.
- With constant effects we have $\tau_{OLS} = \tau_{ATE} = \tau_{ATT}$
- If linearity fails the intuition remains that linearity in X implies an implicit linearity in the underlying PS.

Conclusion: Regression

Is regression evil? ☹

- Its ease sometimes results in lack of thinking. So only a little. ☺
- For descriptive inference, very useful!
  - Good tool for characterizing the conditional expectation function (CEF)
  - But other less restrictive tools are also available for that task (machine learning)
- For causal analysis, always need to ask yourself if *linearly* separable confounding is plausible.
  - A regression is causal when the CEF it approximates is causal.
  - Still need to check common support!
  - Results will be highly sensitive if the treated and controls are far apart (e.g. standardized difference above .2)
- Think about what your **estimand** is: because of variance weighting, coefficient from your regression may not capture ATE if effects are heterogeneous

Conclusion: Regression

Is regression evil? ☹

- Its ease sometimes results in lack of thinking. So only a little. ☺
- For descriptive inference, very useful!
  - Good tool for characterizing the conditional expectation function (CEF)
  - But other less restrictive tools are also available for that task (machine learning)
- For causal analysis, always need to ask yourself if *linearly* separable confounding is plausible.
  - A regression is causal when the CEF it approximates is causal.
  - Still need to check common support!
  - Results will be highly sensitive if the treated and controls are far apart (e.g. standardized difference above .2)
- Think about what your **estimand** is: because of variance weighting, coefficient from your regression may not capture ATE if effects are heterogeneous

Conclusion: Regression

Is regression evil? ☹

- Its ease sometimes results in lack of thinking. So only a little. ☺
- For descriptive inference, very useful!
  - Good tool for characterizing the conditional expectation function (CEF)
  - But other less restrictive tools are also available for that task (machine learning)
- For causal analysis, always need to ask yourself if *linearly* separable confounding is plausible.
  - A regression is causal when the CEF it approximates is causal.
  - Still need to check common support!
  - Results will be highly sensitive if the treated and controls are far apart (e.g. standardized difference above .2)
- Think about what your **estimand** is: because of variance weighting, coefficient from your regression may not capture ATE if effects are heterogeneous

Dehejia and Wabha (1999) Results

Table 3. Estimated Training Effects for the NSW Male Participants Using Comparison Groups From PSID and CPS

	NSW earnings less comparison group earnings		*NSW treatment earnings less comparison group earnings, conditional on the estimated propensity score*					
			Quadratic in score[b]	*Stratifying on the score*			*Matching on the score*	
	(1) Unadjusted	*(2) Adjusted*[a]	*(3)*	*(4) Unadjusted*	*(5) Adjusted*	*(6) Observations*[c]	*(7) Unadjusted*	*(8) Adjusted*[d]
NSW	1,794	1,672						
	(633)	(638)						
PSID-1[e]	−15,205	731	294	1,608	1,494	1,255	1,691	1,473
	(1,154)	(886)	(1,389)	(1,571)	(1,581)		(2,209)	(809)
PSID-2[f]	−3,647	683	496	2,220	2,235	389	1,455	1,480
	(959)	(1,028)	(1,193)	(1,768)	(1,793)		(2,303)	(808)
PSID-3[f]	1,069	825	647	2,321	1,870	247	2,120	1,549
	(899)	(1,104)	(1,383)	(1,994)	(2,002)		(2,335)	(826)
CPS-1[g]	−8,498	972	1,117	1,713	1,774	4,117	1,582	1,616
	(712)	(550)	(747)	(1,115)	(1,152)		(1,069)	(751)
CPS-2[g]	−3,822	790	505	1,543	1,622	1,493	1,788	1,563
	(670)	(658)	(847)	(1,461)	(1,346)		(1,205)	(753)
CPS-3[g]	−635	1,326	556	1,252	2,219	514	587	662
	(657)	(798)	(951)	(1,617)	(2,082)		(1,496)	(776)

[a] Least squares regression: RE78 on a constant, a treatment indicator, age, age^2, education, no degree, black, Hispanic, RE74, RE75.

[b] Least squares regression of RE78 on a quadratic on the estimated propensity score and a treatment indicator, for observations used under stratification; see note (g).

[c] Number of observations refers to the actual number of comparison and treatment units used for (3)–(5); namely, all treatment units and those comparison units whose estimated propensity score is greater than the minimum, and less than the maximum, estimated propensity score for the treatment group.

[d] Weighted least squares: treatment observations weighted as 1, and control observations weighted by the number of times they are matched to a treatment observation [same covariates as (a)]. Propensity scores are estimated using the logistic model, with specifications as follows:

[e] PSID-1: Prob (T_i = 1) = F(age, age^2, education, education^2, married, no degree, black, Hispanic, RE74, RE75, RE74^2, RE75^2, u74*black).

[f] PSID-2 and PSID-3: Prob (T_i = 1) = F(age, age^2, education, education^2, no degree, married, black, Hispanic, RE74, RE74^2, RE75, RE75^2, u74, u75).

[g] CPS-1, CPS-2, and CPS-3: Prob (T_i = 1) = F(age, age^2, education, education^2, no degree, married, black, Hispanic, RE74, RE75, u74, u75, education*RE74, age^3).

Heckman, Ichimura, Smith, and Todd (1998)

- Study randomized evaluation of the Job Training Partnership Act (JTPA)
- 3 groups:
  - Experimental Treatment Group
  - Experimental Control Group
  - Eligible Non-Participants (ENPs): eligible for program but chose not to participate
- Heckman et al. studies under what conditions can covariate adjustment methods make the differences between ENPs and experimental controls diappear.

Heckman, Ichimura, Smith, and Todd (1998)

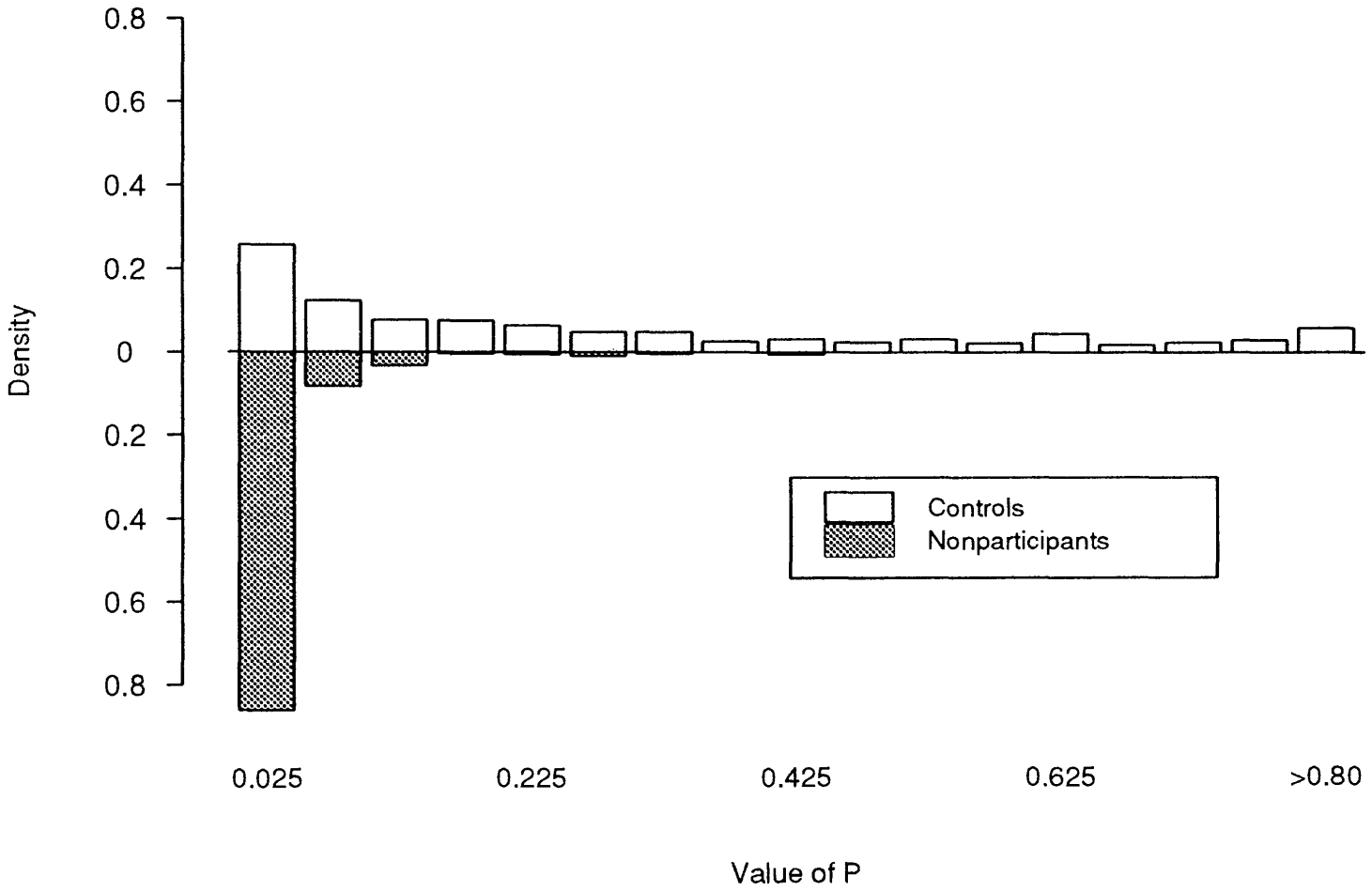


FIGURE 2.—Density of estimated probability of program participation for adult male controls and eligible nonparticipants.

Heckman, Ichimura, Smith, and Todd (1998)

Decomposes bias:

$$B = B_1 + B_2 + B_3$$

where

- B_1 is bias due to lack of common support
- B_2 is bias from the fact that distribution of outcome (Y_0) is different among ENPs and experimental controls
- B_3 is differences in outcomes that remain even after controlling for observed differences (hidden bias).

Heckman, Ichimura, Smith, and Todd (1998)

Decomposes bias:

$$B = B_1 + B_2 + B_3$$

where

- B_1 is bias due to lack of common support
- B_2 is bias from the fact that distribution of outcome (Y_0) is different among ENPs and experimental controls
- B_3 is differences in outcomes that remain even after controlling for observed differences (hidden bias).

Heckman, Ichimura, Smith, and Todd (1998)

Decomposes bias:

$$B = B_1 + B_2 + B_3$$

where

- B_1 is bias due to lack of common support
- B_2 is bias from the fact that distribution of outcome (Y_0) is different among ENPs and experimental controls
- B_3 is differences in outcomes that remain even after controlling for observed differences (hidden bias).

Characterizing Selection Bias

DECOMPOSITION OF MEAN SELECTION BIAS FOR THE BEST PREDICTOR MODEL FOR THE PROBABILITY OF PROGRAM PARTICIPATION[a]
Experimental Control and Elig. Nonparticipant (ENP) Samples, Adult Males,
508 Controls and 388 ENPs

Quarter	(1) Mean Difference[b] ($\hat{B}$)	(2) Nonoverlap Support[c] ($\hat{B}_1$)	(3) Density Weighting ($\hat{B}_2$)	(4) Selection Bias ($\hat{B}_3$)	(5) Average Bias ($\hat{\bar{B}}_{S_P}$)	(6) Experimental Treatment Impact	(7) Average Bias ($\hat{B}_{S_P}$) as of % of Treatment Impact[d]
Qtr1	−420	190[−45%]	−627[149%]	17[−4%]	29	5	566%
	(38)	(31)	(32)	(34)	(63)	(30)	
Qtr2	−352	209[−59%]	−581[165%]	19[−6%]	32	37	88%
	(47)	(41)	(45)	(35)	(65)	(33)	
Qtr3	−343	221[−65%]	−576[168%]	12[−3%]	20	57	35%
	(55)	(39)	(50)	(43)	(79)	(34)	
Qtr4	−294	234[−80%]	−568[194%]	41[−14%]	68	60	114%
	(57)	(40)	(46)	(42)	(79)	(34)	
Qtr5	−311	232[−75%]	−576[185%]	33[−10%]	54	44	121%
	(57)	(40)	(51)	(41)	(77)	(35)	
Qtr6	−334	223[−67%]	−573[172%]	16[−5%]	27	61	44%
	(63)	(45)	(51)	(44)	(81)	(34)	
Average of 1 to 6	−342	218[−64%]	−584[170%]	23[−7%]	38	44	87%
	(47)	(38)	(41)	(33)	(63)	(14)	

Characterizing Selection Bias

- Matching removes most of the bias, but because experimental estimates are small, point estimates would still be substantially biased. ☹
- Finds that the bias is roughly constant over time.
- Thus, difference-in-differences estimator (combined with matching) drives the bias to near 0. ☺
    - More about difference-in-differences estimator later in the quarter

Characterizing Selection Bias

- Matching removes most of the bias, but because experimental estimates are small, point estimates would still be substantially biased. ☹
- Finds that the bias is roughly constant over time.
- Thus, difference-in-differences estimator (combined with matching) drives the bias to near 0. ☺
    - More about difference-in-differences estimator later in the quarter

Characterizing Selection Bias

- Matching removes most of the bias, but because experimental estimates are small, point estimates would still be substantially biased. ☹
- Finds that the bias is roughly constant over time.
- Thus, difference-in-differences estimator (combined with matching) drives the bias to near 0. ☺
    - More about difference-in-differences estimator later in the quarter