Causal Inference
Introduction

José R. Zubizarreta
Harvard University

09/03/2023
CUSO Doctoral School in Statistics and Applied Probability
Saignelégier, Switzerland

Causal Inference Methods

José R. Zubizarreta
Harvard University

09/04/2023
CUSO Doctoral School in Statistics and Applied Probability
Saignelégier, Switzerland

Causal Inference
Designs

José R. Zubizarreta
Harvard University

09/04/2023
CUSO Doctoral School in Statistics and Applied Probability
Saignelégier, Switzerland

Outline

Outline

Learning what works

- Causal inference is about learning which treatments work
  - For whom?
  - When?
  - And why?

Learning what works

- Causal inference is about learning which treatments work
  - For whom?
  - When?
  - And why?
- Examples
  - Effectiveness of COVID-19 vaccines against hospitalizations?
  - Effect of raising interest rates on inflation?
  - Impact of a labor training program on earnings?

Learning what works

- Causal inference is about learning which treatments work
  - For whom?
  - When?
  - And why?
- Examples
  - Effectiveness of COVID-19 vaccines against hospitalizations?
  - Effect of raising interest rates on inflation?
  - Impact of a labor training program on earnings?
- A fertile (but busy!) research area
  - Theory
  - Methodology
  - Applications

Overview

- Causal inference
  - Which treatments work?
  - Form whom?
  - When?
  - And why?

Overview

- Causal inference
  - Which treatments work?
  - Form whom?
  - When?
  - And why?
- Interventions
  - Point exposures
  - Time-varying

Overview

- Causal inference
  - Which treatments work?
  - Form whom?
  - When?
  - And why?
- Interventions
  - Point exposures
  - Time-varying
- Designs
  - Experimental
  - Observational

Overview

- Causal inference
  - Which treatments work?
  - Form whom?
  - When?
  - And why?
- Interventions
  - Point exposures
  - Time-varying
- Designs
  - Experimental
  - Observational
- Strategies
  - Randomization
  - Observation, assumptions
    - E.g., instruments
- Core methods
  - Matching
  - Regression
  - Weighting

Overview

- Causal inference
  - Which treatments work?
  - Form whom?
  - When?
  - And why?
- Interventions
  - Point exposures
  - Time-varying
- Designs
  - Experimental
  - Observational
- Strategies
  - Randomization
  - Observation, assumptions
    - E.g., instruments
- Core methods
  - Matching
  - Regression
  - Weighting
- Sensitivity analyses
- Evidence integration
- Issues throughout
  - Missingness
  - Mismeasurement
  - Fairness...

Overview

- Causal inference
  - Which treatments work?
  - Form whom?
  - When?
  - And why?
- Interventions
  - Point exposures
  - Time-varying
- Designs
  - Experimental
  - Observational
- Strategies
  - Randomization
  - Observation, assumptions
    - E.g., instruments
- Core methods
  - Matching
  - Regression
  - Weighting
- Sensitivity analyses
- Evidence integration
- Issues throughout
  - Missingness
  - Mismeasurement
  - Fairness...
- Perspectives
  - Statistics, biostatistics
  - Economics, political science
  - Computer science

Overview

- Causal inference
  - Which treatments work?
  - Form whom?
  - When?
  - And why?
- Interventions
  - Point exposures
  - Time-varying
- Designs
  - Experimental
  - Observational
- Strategies
  - Randomization
  - Observation, assumptions
    - E.g., instruments
- Core methods
  - Matching
  - Regression
  - Weighting
- Sensitivity analyses
- Evidence integration
- Issues throughout
  - Missingness
  - Mismeasurement
  - Fairness...
- Perspectives
  - Statistics, biostatistics
  - Economics, political science
  - Computer science

Overview

- Causal inference
  - Which treatments work?
  - Form whom?
  - When?
  - And why?

Overview

- Causal inference
  - Which treatments work?
  - Form whom?
  - When?
  - And why?
- Interventions
  - Point exposures
  - Time-varying

Overview

- Causal inference
  - Which treatments work?
  - Form whom?
  - When?
  - And why?
- Interventions
  - Point exposures
  - Time-varying
- Designs
  - Experimental
  - Observational

Overview

- Causal inference
  - Which treatments work?
  - Form whom?
  - When?
  - And why?
- Interventions
  - Point exposures
  - Time-varying
- Designs
  - Experimental
  - Observational
- Strategies
  - Randomization
  - Observation, assumptions
    - E.g., instruments
- Core methods
  - Matching
  - Regression
  - Weighting

Overview

- Causal inference
  - Which treatments work?
  - Form whom?
  - When?
  - And why?
- Interventions
  - Point exposures
  - Time-varying
- Designs
  - Experimental
  - Observational
- Strategies
  - Randomization
  - Observation, assumptions
    - E.g., instruments
- Core methods
  - Matching
  - Regression
  - Weighting
- Sensitivity analyses
- Evidence integration
- Issues throughout
  - Missingness
  - Mismeasurement
  - Fairness...

Overview

- Causal inference
  - Which treatments work?
  - Form whom?
  - When?
  - And why?
- Interventions
  - Point exposures
  - Time-varying
- Designs
  - Experimental
  - Observational
- Strategies
  - Randomization
  - Observation, assumptions
    - E.g., instruments
- Core methods
  - Matching
  - Regression
  - Weighting
- Sensitivity analyses
- Evidence integration
- Issues throughout
  - Missingness
  - Mismeasurement
  - Fairness...
- Perspectives
  - Statistics, biostatistics
  - Economics, political science
  - Computer science

Overview

- Causal inference
  - Which treatments work?
  - Form whom?
  - When?
  - And why?
- Interventions
  - Point exposures
  - Time-varying
- Designs
  - Experimental
  - Observational
- Strategies
  - Randomization
  - Observation, assumptions
    - E.g., instruments
- Core methods
  - Matching
  - Regression
  - Weighting
- Sensitivity analyses
- Evidence integration
- Issues throughout
  - Missingness
  - Mismeasurement
  - Fairness...
- Perspectives
  - Statistics, biostatistics
  - Economics, political science
  - Computer science

Acknowledgments

- Work supported by the Alfred P. Sloan Foundation and PCORI

Outline

Outline

Causal evidence

- Causal inference requires careful thinking

Causal evidence

- Causal inference requires careful thinking
  - Adopt a multiplist strategy (Cook 1985) in which several complementary designs are used to rule out confounding effects
  - Integrate the evidence provided by different studies that vary in fundamental ways
- Design of experimental and observational studies

Causal evidence

- Causal inference requires careful thinking
  - Adopt a multiplist strategy (Cook 1985) in which several complementary designs are used to rule out confounding effects
  - Integrate the evidence provided by different studies that vary in fundamental ways
- Design of experimental and observational studies
  - Not unlike randomized experiments, observational studies can (and should) be carefully designed

Causal evidence

- Causal inference requires careful thinking
  - Adopt a multiplist strategy (Cook 1985) in which several complementary designs are used to rule out confounding effects
  - Integrate the evidence provided by different studies that vary in fundamental ways
- Design of experimental and observational studies
  - Not unlike randomized experiments, observational studies can (and should) be carefully designed
  - Consistent results from varied designs can gradually reduce, albeit not eliminate, uncertainty about unmeasured confounding

Big data and causal inference

- Algorithms and massive data sets have the potential to transform the way we do science and policy

Big data and causal inference

- Algorithms and massive data sets have the potential to transform the way we do science and policy
- The amount of available data is growing exponentially; algorithms and computations are increasingly more powerful

Big data and causal inference

- Algorithms and massive data sets have the potential to transform the way we do science and policy
- The amount of available data is growing exponentially; algorithms and computations are increasingly more powerful
- However, predictors are not founded on causal relationships, and the use of big data is not a substitute for thoughtful study design

Outline

Running example: impact of a labor training program

- Suppose that we want to estimate the effect of a labor training program on a given person i

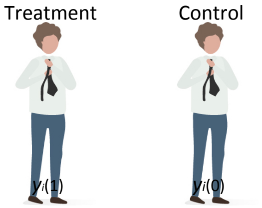
Designs by `rawpixel.com` / Freepik

Potential outcomes (Neyman 1923, Rubin 1974)

- Let $Y_i(1)$ be the *potential outcome* of person i under treatment and $Y_i(0)$ be his the potential outcome under control

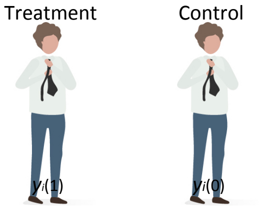
Potential outcomes (Neyman 1923, Rubin 1974)

- Let $Y_i(1)$ be the *potential outcome* of person i under treatment and $Y_i(0)$ be his the potential outcome under control
- In our example,
  - $Y_i(1) =$ income that the i-th person would have if he participates in the program
  - $Y_i(0) =$ income that the i-th person would have if he does not participate



Potential outcomes (Neyman 1923, Rubin 1974)

- Let $Y_i(1)$ be the *potential outcome* of person i under treatment and $Y_i(0)$ be his the potential outcome under control
- In our example,
  - $Y_i(1) =$ income that the i-th person would have if he participates in the program
  - $Y_i(0) =$ income that the i-th person would have if he does not participate

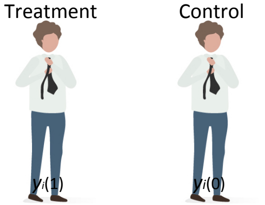


- $Y_i(1)$ is observed only for the people that receive the treatment; otherwise it is unobserved or counterfactual

Designs by rawpixel.com / Freepik

The fundamental problem of causal inference

- The causal effect of treatment compared to control for unit i is $Y_i(1) - Y_i(0)$
  (We can also write $Y_i(1)/Y_i(0)$.)
- These individual effects are not observable
- It is impossible to observe both the value of $Y_i(1)$ and $Y_i(0)$ for unit i



- This is what Holland (1986) called the “fundamental problem of causal inference”

Designs by rawpixel.com / Freepik

The science table

Unit	$Y(0)$	$Y(1)$	Z	Y	$Y(1) - Y(0)$
1	1568	2367	0	1568	799
2	2194	3295	0	2194	1201
3	2500	3503	0	2500	1003
4	5344	6343	0	5344	999
5	12780	12881	0	12780	101
6	1923	3024	1	3024	1101
7	5159	6182	1	6182	1023
8	1000	2000	1	2000	1000
9	2143	3197	1	3197	1044
10	15752	15953	1	15953	201

The science table

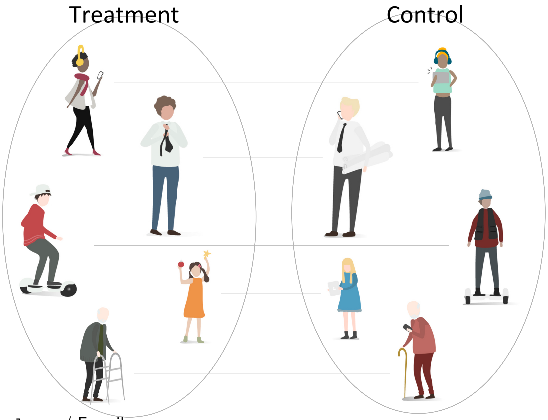
Unit	$Y(0)$	$Y(1)$	Z	Y	$Y(1) - Y(0)$
1	**1568**	**2367**	**0**	**1568**	**799**
2	2194	3295	0	2194	1201
3	2500	3503	0	2500	1003
4	5344	6343	0	5344	999
5	12780	12881	0	12780	101
6	1923	3024	1	3024	1101
7	5159	6182	1	6182	1023
8	1000	2000	1	2000	1000
9	2143	3197	1	3197	1044
10	15752	15953	1	15953	201

The statistician's table

Unit	$Y(0)$	$Y(1)$	Z	Y	$Y(1) - Y(0)$
1	**1568**	**?**	**0**	**1568**	**?**
2	2194	?	0	2194	?
3	2500	?	0	2500	?
4	5344	?	0	5344	?
5	12780	?	0	12780	?
6	?	3024	1	3024	?
7	?	6182	1	6182	?
8	?	2000	1	2000	?
9	?	3197	1	3197	?
10	?	15953	1	15953	?

The statistical solution

- The statistical solution to the fundamental problem of causal inference is to randomly assign the treatment and estimate an average treatment effect



Designs by rawpixel.com / Freepik

Notation and estimand

- Define the treatment assignment indicator

$$Z_i := \begin{cases} 1 & \text{if person } i \text{ is assigned to the program} \\ 0 & \text{if person } i \text{ is not assigned} \end{cases}$$

Notation and estimand

- Define the treatment assignment indicator

$$Z_i := \begin{cases} 1 & \text{if person } i \text{ is assigned to the program} \\ 0 & \text{if person } i \text{ is not assigned} \end{cases}$$

- Write the observed outcome as

$$Y_i := Z_i Y_i(1) + (1 - Z_i) Y_i(0)$$

Notation and estimand

- Define the treatment assignment indicator

$$Z_i := \begin{cases} 1 & \text{if person } i \text{ is assigned to the program} \\ 0 & \text{if person } i \text{ is not assigned} \end{cases}$$

- Write the observed outcome as

$$Y_i := Z_i Y_i(1) + (1 - Z_i) Y_i(0)$$

- Let $\boldsymbol{X}_i$ be a vector of observed covariates and U_i an unobserved covariate

Notation and estimand

- Define the treatment assignment indicator

$$Z_i := \begin{cases} 1 & \text{if person } i \text{ is assigned to the program} \\ 0 & \text{if person } i \text{ is not assigned} \end{cases}$$

- Write the observed outcome as

$$Y_i := Z_i Y_i(1) + (1 - Z_i) Y_i(0)$$

- Let $\boldsymbol{X}_i$ be a vector of observed covariates and U_i an unobserved covariate
- We wish to estimate the sample average treatment effect on the treated

SUTVA

- This notation implies the Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1980)

SUTVA

- This notation implies the Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1980)
- This assumption has two components

SUTVA

- This notation implies the Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1980)
- This assumption has two components
    - There is no interference between units

SUTVA

- This notation implies the Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1980)
- This assumption has two components
  - There is no interference between units
    - In our example, participation of one subject in the program does not affect the earnings of another subject

SUTVA

- This notation implies the Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1980)
- This assumption has two components
  - There is no interference between units
    - In our example, participation of one subject in the program does not affect the earnings of another subject
  - There are no versions of treatment beyond those encoded by Z

SUTVA

- This notation implies the Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1980)
- This assumption has two components
  - There is no interference between units
    - In our example, participation of one subject in the program does not affect the earnings of another subject
  - There are no versions of treatment beyond those encoded by Z
    - Subjects participate or not in the program and there are no “levels” of labor training

On the potential outcomes framework

- The potential outcomes model was developed by Neyman (1923) and Rubin (1974);

On the potential outcomes framework

- The potential outcomes model was developed by Neyman (1923) and Rubin (1974); for more background and history, see Holland (1986)

On the potential outcomes framework

- The potential outcomes model was developed by Neyman (1923) and Rubin (1974); for more background and history, see Holland (1986)
- See Robins (1986) and Hernán and Robins (2019) for an extension of the potential outcomes framework to longitudinal studies

On the potential outcomes framework

- The potential outcomes model was developed by Neyman (1923) and Rubin (1974); for more background and history, see Holland (1986)
- See Robins (1986) and Hernán and Robins (2019) for an extension of the potential outcomes framework to longitudinal studies
- For a discussion of the strengths of the potential outcomes framework see, e.g., Imbens and Wooldridge (2009):
  "The first advantage of the potential outcome framework is that it allows us to define causal effects before specifying the assignment mechanism, and without making functional form or distributional assumptions..."

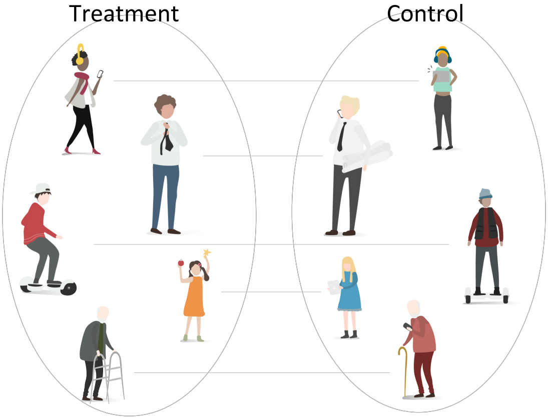
On the potential outcomes framework

- The potential outcomes model was developed by Neyman (1923) and Rubin (1974); for more background and history, see Holland (1986)
- See Robins (1986) and Hernán and Robins (2019) for an extension of the potential outcomes framework to longitudinal studies
- For a discussion of the strengths of the potential outcomes framework see, e.g., Imbens and Wooldridge (2009):
  "The first advantage of the potential outcome framework is that it allows us to define causal effects before specifying the assignment mechanism, and without making functional form or distributional assumptions..."

Outline

Randomized experiment

- In a randomized experiment, the treatment and control groups tend to be similar in terms of their observed and unobserved covariates



Designs by rawpixel.com / Freepik

"Table 1" for covariate balance

Table: Covariate balance after matching

Covariate	Mean treated	Mean controls	Standardized difference
Age	26.22	25.46	0.08
Education	10.23	10.23	0.00
Black	0.85	0.83	0.05
Hispanic	0.06	0.05	0.03
Married	0.20	0.20	0.00
No degree	0.70	0.70	0.00
Earnings 74	2040.29	2313.52	-0.05
Earnings 75	1484.92	1476.09	0.00

Randomization in experiments

- Randomized experiments are the "gold standard for causal inference," the most reliable approach for learning about the effects of treatments

Randomization in experiments

- Randomized experiments are the “gold standard for causal inference,” the most reliable approach for learning about the effects of treatments
  - Randomization yields an unbiased estimator — by design
    - Both observed and unobserved covariates are balanced in expectation across the treatment groups

Randomization in experiments

- Randomized experiments are the "gold standard for causal inference," the most reliable approach for learning about the effects of treatments
  - Randomization yields an unbiased estimator — by design
    - Both observed and unobserved covariates are balanced in expectation across the treatment groups
  - Randomization is the "reasoned basis for inference" (Fisher 1935)
    - The act of randomly assigning units to treatment physically induces a distribution that can be used for exact testing

The role of randomization for statistical control

- Randomization guarantees that $Y(1), Y(0) \perp\!\!\!\perp Z$

The role of randomization for statistical control

- Randomization guarantees that $Y(1), Y(0) \perp\!\!\!\perp Z$
- In words, a fair coin knows nothing about the individual units and is impartial in its treatment assignments

The role of randomization for statistical control

- Randomization guarantees that $Y(1), Y(0) \perp\!\!\!\perp Z$
- In words, a fair coin knows nothing about the individual units and is impartial in its treatment assignments
- Therefore

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

The role of randomization for statistical control

- Randomization guarantees that $Y(1), Y(0) \perp\!\!\!\perp Z$
- In words, a fair coin knows nothing about the individual units and is impartial in its treatment assignments
- Therefore

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

- Furthermore, $\mathbb{E}[Y(1) - Y(0) \mid Z = 1] = \mathbb{E}[Y(1) - Y(0)]$

The role of randomization for exact inference

- Randomization physically induces a distribution that can be used for exact testing, without distributional assumptions

The role of randomization for exact inference

- Randomization physically induces a distribution that can be used for exact testing, without distributional assumptions
- Consider the sharp null hypothesis: $H_0 : Y_i(1) = Y_i(0)$ for all i

The role of randomization for exact inference

- Randomization physically induces a distribution that can be used for exact testing, without distributional assumptions
- Consider the sharp null hypothesis: $H_0 : Y_i(1) = Y_i(0)$ for all i
- General testing procedure (Fisher 1935, Rosenbaum 2002)

The role of randomization for exact inference

- Randomization physically induces a distribution that can be used for exact testing, without distributional assumptions
- Consider the sharp null hypothesis: $H_0 : Y_i(1) = Y_i(0)$ for all i
- General testing procedure (Fisher 1935, Rosenbaum 2002)
    1. Assume H_0 holds, so Y_i is fixed

The role of randomization for exact inference

- Randomization physically induces a distribution that can be used for exact testing, without distributional assumptions
- Consider the sharp null hypothesis: $H_0 : Y_i(1) = Y_i(0)$ for all i
- General testing procedure (Fisher 1935, Rosenbaum 2002)
  1. Assume H_0 holds, so Y_i is fixed
  2. Consider the set of possible treatment assignments Ω from where $\boldsymbol{Z}$ was selected

The role of randomization for exact inference

- Randomization physically induces a distribution that can be used for exact testing, without distributional assumptions
- Consider the sharp null hypothesis: $H_0 : Y_i(1) = Y_i(0)$ for all i
- General testing procedure (Fisher 1935, Rosenbaum 2002)
  1. Assume H_0 holds, so Y_i is fixed
  2. Consider the set of possible treatment assignments Ω from where $\boldsymbol{Z}$ was selected
  3. Calculate the observed value, say T, of the test statistic $t(\boldsymbol{Z}, \boldsymbol{y})$

The role of randomization for exact inference

- Randomization physically induces a distribution that can be used for exact testing, without distributional assumptions
- Consider the sharp null hypothesis: $H_0 : Y_i(1) = Y_i(0)$ for all i
- General testing procedure (Fisher 1935, Rosenbaum 2002)
  1. Assume H_0 holds, so Y_i is fixed
  2. Consider the set of possible treatment assignments Ω from where $\boldsymbol{Z}$ was selected
  3. Calculate the observed value, say T, of the test statistic $t(\boldsymbol{Z}, \boldsymbol{y})$
  4. Compute the probability of a value of $t(\boldsymbol{Z}, \boldsymbol{y})$ at least as large as the one observed under H_0

$$\Pr_{H_0,\mathcal{D}}(t(\boldsymbol{Z}, \boldsymbol{y}) \geq T) = \sum_{\boldsymbol{z} \in \Omega} \mathbb{1}_{\{t(\boldsymbol{z}, \boldsymbol{y}) \geq T\}} \Pr(\boldsymbol{Z} = \boldsymbol{z})$$

The role of randomization for exact inference

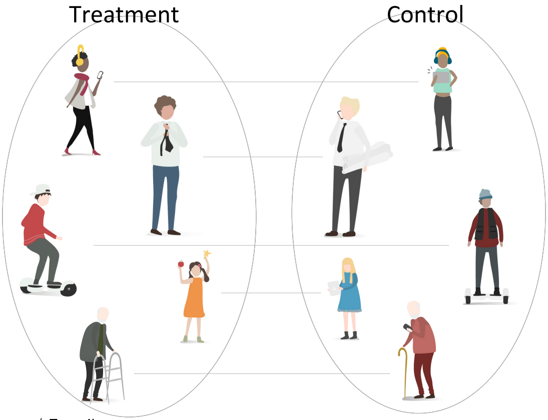
- Randomization physically induces a distribution that can be used for exact testing, without distributional assumptions
- Consider the sharp null hypothesis: $H_0 : Y_i(1) = Y_i(0)$ for all i
- General testing procedure (Fisher 1935, Rosenbaum 2002)
  1. Assume H_0 holds, so Y_i is fixed
  2. Consider the set of possible treatment assignments Ω from where $\boldsymbol{Z}$ was selected
  3. Calculate the observed value, say T, of the test statistic $t(\boldsymbol{Z}, \boldsymbol{y})$
  4. Compute the probability of a value of $t(\boldsymbol{Z}, \boldsymbol{y})$ at least as large as the one observed under H_0

$$\Pr_{H_0,\mathcal{D}}(t(\boldsymbol{Z}, \boldsymbol{y}) \geq T) = \sum_{\boldsymbol{z} \in \Omega} \mathbb{1}_{\{t(\boldsymbol{z},\boldsymbol{y}) \geq T\}} \Pr(\boldsymbol{Z} = \boldsymbol{z})$$

- The following example is by Rosenbaum (2010)

Matched pairs design

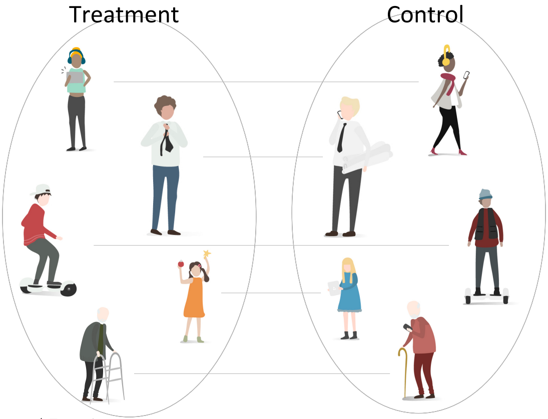
- We consider a matched pairs design $\mathcal{D}$ such that units are randomized to treatment within matched pairs
- This is one possible treatment assignment in $\boldsymbol{z} \in \Omega$



Designs by rawpixel.com / Freepik

Matched pairs design

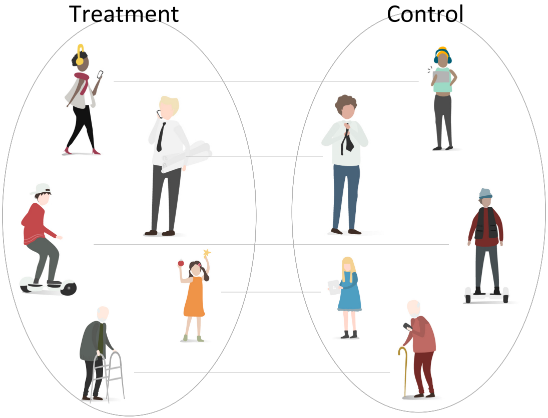
- We consider a matched pairs design $\mathcal{D}$ such that units are randomized to treatment within matched pairs
- This is another possible treatment assignment



Designs by rawpixel.com / Freepik

Matched pairs design

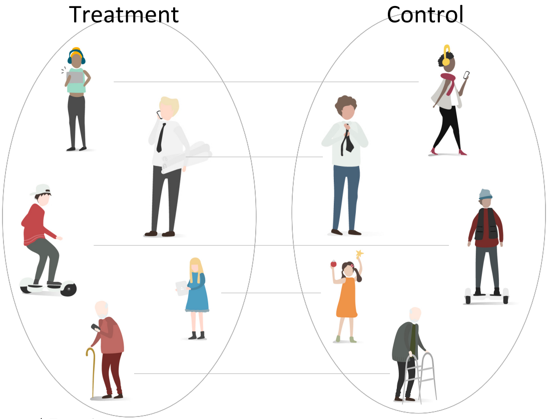
- We consider a matched pairs design $\mathcal{D}$ such that units are randomized to treatment within matched pairs
- And another one...



Designs by rawpixel.com / Freepik

Matched pairs design

- We consider a matched pairs design $\mathcal{D}$ such that units are randomized to treatment within matched pairs
- With 5 pairs, there are $2^5 = 32$ possible assignments



Designs by rawpixel.com / Freepik

1. Assume H_0 holds

Pair	Z	X_{age}	X_{black}	X_{hisp}	Y_{earn}	ΔY
1	1 0	17 18	1 1	0 0	3024 1568	1456
2	1 0	25 25	1 1	0 0	6128 2194	3988
3	1 0	25 25	1 1	0 0	0 45	-45
4	1 0	28 22	1 1	0 0	3197 5344	-2147
5	1 0	33 28	1 1	0 0	15953 12780	3173

1. Assume H_0 holds

Pair	Z	X_{age}	X_{black}	X_{hisp}	Y_{earn}	ΔY
1	**1**	**17**	**1**	**0**	**3024**	**1456**
	0	**18**	**1**	**0**	**1568**	
2	1	25	1	0	6128	3988
	0	25	1	0	2194	
3	1	25	1	0	0	-45
	0	25	1	0	45	
4	1	28	1	0	3197	-2147
	0	22	1	0	5344	
5	1	33	1	0	15953	3173
	0	28	1	0	12780	

2. Consider the set of possible treatment assignments Ω

Assignment	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5
1	1	1	1	1	1
2	1	1	1	1	0
3	1	1	1	0	1
4	1	1	1	0	0
5	1	1	0	1	1
6	1	1	0	1	0
7	1	1	0	0	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
31	0	0	0	0	1
32	0	0	0	0	0

2. Consider the set of possible treatment assignments Ω

Assignment	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5
1	**1**	**1**	**1**	**1**	**1**
2	1	1	1	1	0
3	1	1	1	0	1
4	1	1	1	0	0
5	1	1	0	1	1
6	1	1	0	1	0
7	1	1	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮
31	0	0	0	0	1
32	0	0	0	0	0

2. Consider the set of possible treatment assignments Ω

Assignment	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5
1	1	1	1	1	1
2	**1**	**1**	**1**	**1**	**0**
3	1	1	1	0	1
4	1	1	1	0	0
5	1	1	0	1	1
6	1	1	0	1	0
7	1	1	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮
31	0	0	0	0	1
32	0	0	0	0	0

2. Consider the set of possible treatment assignments Ω

Assignment	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5
1	1	1	1	1	1
2	1	1	1	1	0
3	**1**	**1**	**1**	**0**	**1**
4	1	1	1	0	0
5	1	1	0	1	1
6	1	1	0	1	0
7	1	1	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮
31	0	0	0	0	1
32	0	0	0	0	0

2. Consider the set of possible treatment assignments Ω

Assignment	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5
1	1	1	1	1	1
2	1	1	1	1	0
3	1	1	1	0	1
4	1	1	1	0	0
5	1	1	0	1	1
6	1	1	0	1	0
7	1	1	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮
31	0	0	0	0	1
32	**0**	**0**	**0**	**0**	**0**

3. Calculate the observed value of the test statistic

Assignment	Pair difference in outcomes					Mean diff.
	ΔY_1	ΔY_2	ΔY_3	ΔY_4	ΔY_5	$\overline{\Delta Y}$
1	1456	3988	-45	-2147	3173	1285
2	1456	3988	-45	-2147	-3173	15.8
3	1456	3988	-45	2147	3173	2143.8
4	1456	3988	-45	2147	-3173	874.6
5	1456	3988	45	-2147	3173	1303
6	1456	3988	45	-2147	-3173	33.8
7	1456	3988	45	2147	3173	2161.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
31	-1456	-3988	45	2147	3173	-15.8
32	-1456	-3988	45	2147	-3173	-1285

3. Calculate the observed value of the test statistic

Assignment	Pair difference in outcomes					Mean diff.
	ΔY_1	ΔY_2	ΔY_3	ΔY_4	ΔY_5	$\overline{\Delta Y}$
1	**1456**	**3988**	**-45**	**-2147**	**3173**	**1285**
2	1456	3988	-45	-2147	-3173	15.8
3	1456	3988	-45	2147	3173	2143.8
4	1456	3988	-45	2147	-3173	874.6
5	1456	3988	45	-2147	3173	1303
6	1456	3988	45	-2147	-3173	33.8
7	1456	3988	45	2147	3173	2161.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
31	-1456	-3988	45	2147	3173	-15.8
32	-1456	-3988	45	2147	-3173	-1285

3. Calculate the observed value of the test statistic

Assignment	Pair difference in outcomes					Mean diff.
	ΔY_1	ΔY_2	ΔY_3	ΔY_4	ΔY_5	$\overline{\Delta Y}$
1	1456	3988	-45	-2147	3173	1285
2	**1456**	**3988**	**-45**	**-2147**	**-3173**	**15.8**
3	1456	3988	-45	2147	3173	2143.8
4	1456	3988	-45	2147	-3173	874.6
5	1456	3988	45	-2147	3173	1303
6	1456	3988	45	-2147	-3173	33.8
7	1456	3988	45	2147	3173	2161.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
31	-1456	-3988	45	2147	3173	-15.8
32	-1456	-3988	45	2147	-3173	-1285

3. Calculate the observed value of the test statistic

Assignment	Pair difference in outcomes					Mean diff.
	ΔY_1	ΔY_2	ΔY_3	ΔY_4	ΔY_5	$\overline{\Delta Y}$
1	1456	3988	-45	-2147	3173	1285
2	1456	3988	-45	-2147	-3173	15.8
3	**1456**	**3988**	**-45**	**2147**	**3173**	**2143.8**
4	1456	3988	-45	2147	-3173	874.6
5	1456	3988	45	-2147	3173	1303
6	1456	3988	45	-2147	-3173	33.8
7	1456	3988	45	2147	3173	2161.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
31	-1456	-3988	45	2147	3173	-15.8
32	-1456	-3988	45	2147	-3173	-1285

3. Calculate the observed value of the test statistic

Assignment	Pair difference in outcomes					Mean diff.
	ΔY_1	ΔY_2	ΔY_3	ΔY_4	ΔY_5	$\overline{\Delta Y}$
1	1456	3988	-45	-2147	3173	1285
2	1456	3988	-45	-2147	-3173	15.8
3	1456	3988	-45	2147	3173	2143.8
4	1456	3988	-45	2147	-3173	874.6
5	1456	3988	45	-2147	3173	1303
6	1456	3988	45	-2147	-3173	33.8
7	1456	3988	45	2147	3173	2161.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
31	-1456	-3988	45	2147	3173	-15.8
32	**-1456**	**-3988**	**45**	**2147**	**-3173**	**-1285**

4. Compute $\Pr_{H_0,\mathcal{D}}(t(\boldsymbol{Z},\boldsymbol{y}) \geq T)$

y	$\Pr_{H_0,\mathcal{D}}(\overline{\Delta Y} = y)$	$\Pr_{H_0,\mathcal{D}}(\overline{\Delta Y} \geq y)$
2161.8	0.03125	0.03125
2143.8	0.03125	0.0625
1579.4	0.03125	0.09375
1561.4	0.03125	0.125
1303.0	0.03125	0.15625
1285.0	0.03125	0.1875
892.6	0.03125	0.21875
⋮	⋮	⋮
-2143.8	0.03125	0.96875
-2161.8	0.03125	1

4. Compute $\Pr_{H_0,\mathcal{D}}(t(\mathbf{Z},\mathbf{y}) \geq T)$

y	$\Pr_{H_0,\mathcal{D}}(\overline{\Delta Y} = y)$	$\Pr_{H_0,\mathcal{D}}(\overline{\Delta Y} \geq y)$
2161.8	0.03125	0.03125
2143.8	0.03125	0.0625
1579.4	0.03125	0.09375
1561.4	0.03125	0.125
1303.0	0.03125	0.15625
1285.0	**0.03125**	**0.1875**
892.6	0.03125	0.21875
⋮	⋮	⋮
-2143.8	0.03125	0.96875
-2161.8	0.03125	1

Randomization inference in experiments

- For valid inferences about the effects of a treatment in an experiment, it is sufficient to require random treatment assignment

Randomization inference in experiments

- For valid inferences about the effects of a treatment in an experiment, it is sufficient to require random treatment assignment
- Probability enters the experiment only through the random assignment — which is controlled by the experimenter

Outline

Randomized experiment

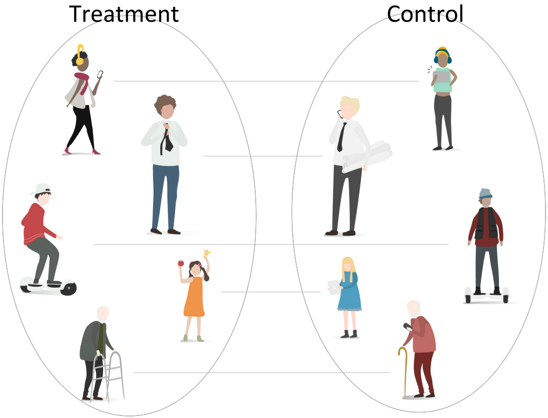
- A randomized experiment is the ideal way to investigate the causal effect of a treatment
  - The law of large numbers ensures that the treatment and control groups are comparable
  - Correct randomization inferences can be made due to random assignment

Randomized experiment

- A randomized experiment is the ideal way to investigate the causal effect of a treatment
  - The law of large numbers ensures that the treatment and control groups are comparable
  - Correct randomization inferences can be made due to random assignment
- However, a randomized experiment is not always possible...

Randomized experiment

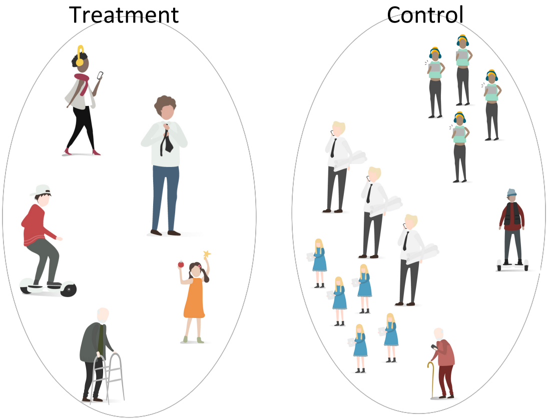
- In a randomized experiment, the treatment and control groups tend to be similar in terms of their observed and unobserved covariates



Designs by rawpixel.com / Freepik

Observational study

- In an observational study, by contrast, assignment is not at random and groups tend to differ systematically in their covariates



Designs by rawpixel.com / Freepik

Biases in observational studies

- An observational study is *biased* if the treated and control groups differ before treatment in ways that matter for the outcome

Biases in observational studies

- An observational study is *biased* if the treated and control groups differ before treatment in ways that matter for the outcome
- An overt bias is one that can be seen in the data at hand
  - For example, differences between the treated and control groups on age and education are overt biases

Biases in observational studies

- An observational study is *biased* if the treated and control groups differ before treatment in ways that matter for the outcome
- An overt bias is one that can be seen in the data at hand
  - For example, differences between the treated and control groups on age and education are overt biases
- A *hidden bias* is similar to an overt bias but cannot be seen because not all the relevant covariates were observed
  - If "ability" differs between groups even after we control for age, education and other covariates, then the study has a hidden bias

Two strategies

- Overt biases can be controlled by adjusting for, or balancing, observed covariates

Two strategies

- Overt biases can be controlled by adjusting for, or balancing, observed covariates
- Hidden biases require using other approaches, such as an instrumental variable or a discontinuity in treatment assignment

Cochran's basic advice

- Cochran (1965):
  - *"The planner of an observational study should always ask himself the question, 'How would the study be conducted if it were possible to do it by controlled experimentation?'."*

Outline

Outline

Under what conditions adjusting for overt biases works?

- Consider an outcome variable Y, treatment indicator Z, and observed covariates $\boldsymbol{X}$

Under what conditions adjusting for overt biases works?

- Consider an outcome variable Y, treatment indicator Z, and observed covariates $\boldsymbol{X}$
- Imagine $\boldsymbol{X}$ includes all the variables that are associated with both the treatment assignment and the potential outcomes

Under what conditions adjusting for overt biases works?

- Consider an outcome variable Y, treatment indicator Z, and observed covariates $\boldsymbol{X}$
- Imagine $\boldsymbol{X}$ includes all the variables that are associated with both the treatment assignment and the potential outcomes
- Then for individuals sharing a particular value of $\boldsymbol{X}$, treatment assignment would be essentially random

Under what conditions adjusting for overt biases works?

- Consider an outcome variable Y, treatment indicator Z, and observed covariates $\boldsymbol{X}$
- Imagine $\boldsymbol{X}$ includes all the variables that are associated with both the treatment assignment and the potential outcomes
- Then for individuals sharing a particular value of $\boldsymbol{X}$, treatment assignment would be essentially random
- Formally,

$$Y(0), Y(1) \perp\!\!\!\perp Z|\boldsymbol{X} \ \text{ for all } \ \boldsymbol{X}$$

Unconfoundedness

- The assumption that $Y(0)$, $Y(1)$ is conditionally independent of Z given $\boldsymbol{X}$ is called *ignorability*, *unconfoundedness*, *no unmeasured confounders*, *selection on observables*, or *exogeneity*
- We can also say that the potential outcomes are *missing at random (MAR)* given the observed covariates

Overlap

- Another assumption needed for adjusting for overt biases is the *overlap assumption*

$$0 < \Pr(Z = 1|\boldsymbol{X}) < 1 \ \text{ for all } \ \boldsymbol{X}$$

Identification

- We wish to estimate

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

Identification

- We wish to estimate

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

- Clearly,

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

Identification

- We wish to estimate

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

- Clearly,

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

- The first term can directly be estimated from the data

Identification

- We wish to estimate

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

- Clearly,

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

- The first term can directly be estimated from the data
- The second term can be identified under the above assumptions

$$\begin{aligned} \mathbb{E}[Y(0)|Z = 1] &= \mathbb{E}\left[\mathbb{E}[Y(0)|\boldsymbol{X}, Z = 1]|Z = 1\right] \\ &= \mathbb{E}\left[\mathbb{E}[Y(0)|\boldsymbol{X}, Z = 0]|Z = 1\right] \\ &= \mathbb{E}\left[\mathbb{E}[Y|\boldsymbol{X}, Z = 0]|Z = 1\right] \end{aligned}$$

Using regression

- Write

$$\mu_z(\boldsymbol{X}) := \mathbb{E}[Y(z)|\boldsymbol{X}]$$

Using regression

- Write
$$\mu_z(\boldsymbol{X}) := \mathbb{E}[Y(z)|\boldsymbol{X}]$$
- Under the above assumptions, $\mu_z(\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}, Z = z]$

Using regression

- Write

$$\mu_z(\boldsymbol{X}) := \mathbb{E}[Y(z)|\boldsymbol{X}]$$

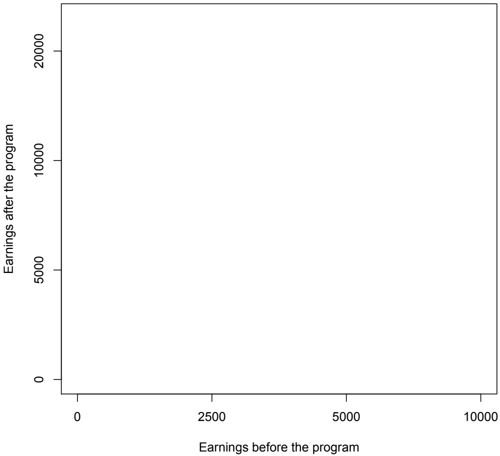
- Under the above assumptions, $\mu_z(\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}, Z = z]$
- So we can estimate the second term with

$$\bar{Y}^{imp}_{i:Z_i=1} := \frac{1}{n_1} \sum_{i:Z_i=1} \hat{\mu}_0(\boldsymbol{X})$$

for example using linear regression

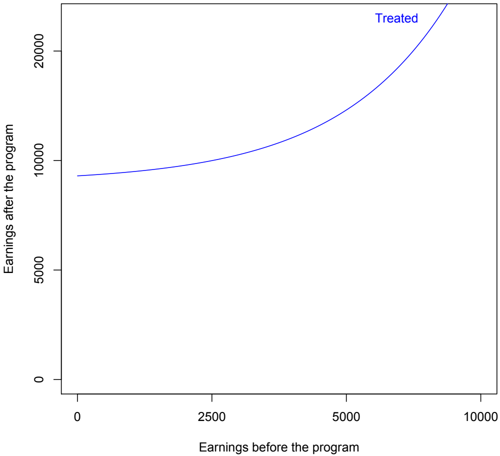
Extrapolation and fabrication

- See Hill (2011) for details



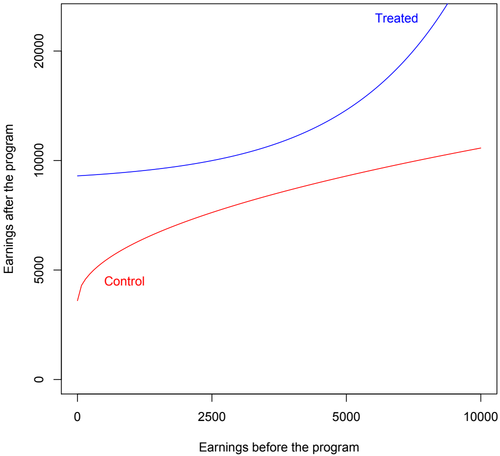
Extrapolation and fabrication

- See Hill (2011) for details



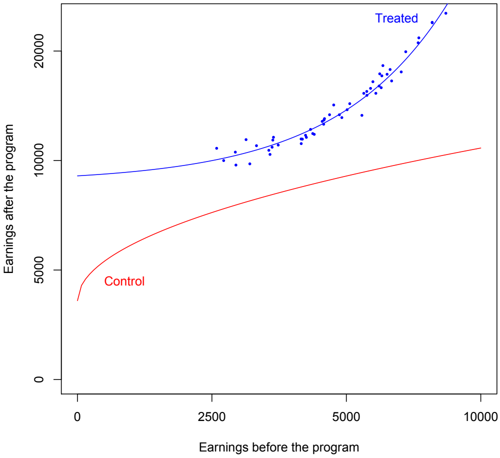
Extrapolation and fabrication

- See Hill (2011) for details



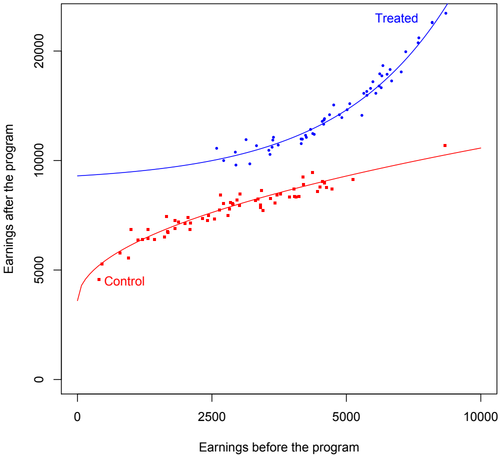
Extrapolation and fabrication

- See Hill (2011) for details



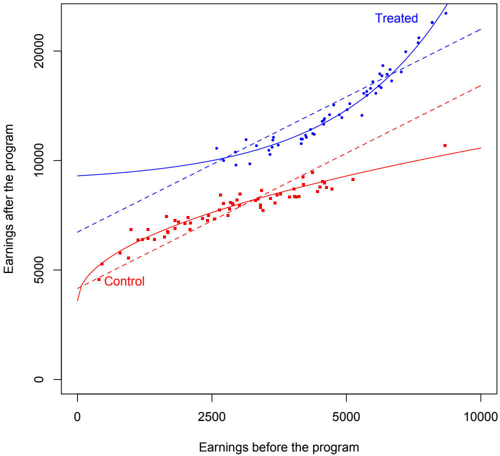
Extrapolation and fabrication

- See Hill (2011) for details



Extrapolation and fabrication

- See Hill (2011) for details



Concerns with regression approaches

- Potential lack of overlap in the covariate distributions between the treated and control groups

Concerns with regression approaches

- Potential lack of overlap in the covariate distributions between the treated and control groups
- In standard practice, covariate adjustments are done while looking at the results

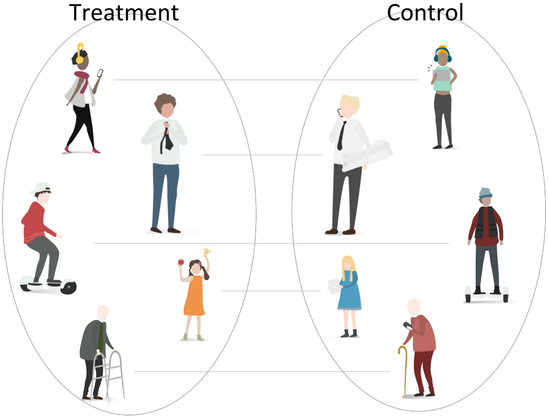
Concerns with regression approaches

- Potential lack of overlap in the covariate distributions between the treated and control groups
- In standard practice, covariate adjustments are done while looking at the results
- How does linear regression emulate key features of a randomized experiment?
  - E.g., covariate balance, study representativeness, self-weighted sampling, sample boundedness

Outline

Randomized experiment

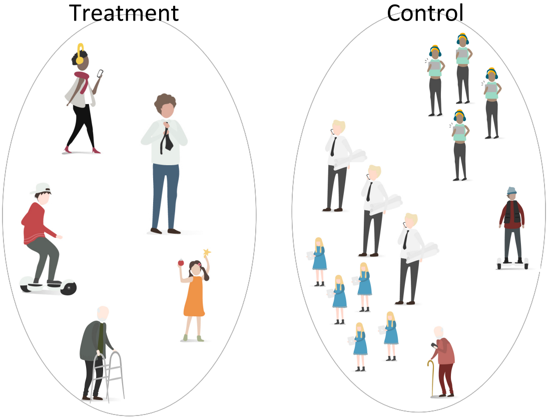
- In a randomized experiment, the treatment and control groups tend to be similar in terms of their observed and unobserved covariates



Designs by `rawpixel.com` / Freepik

Observational study

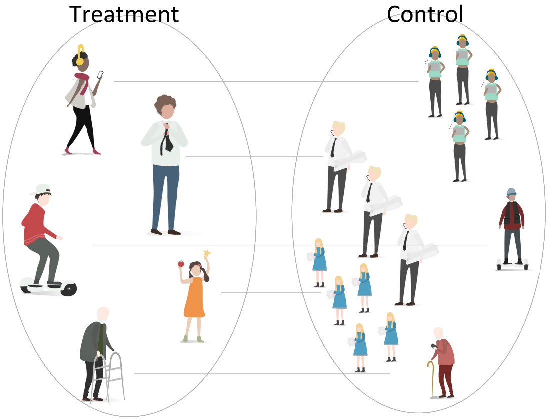
- In an observational study, by contrast, assignment is not at random and groups tend to differ systematically in their covariates



Designs by `rawpixel.com` / Freepik

Matching to approximate a randomized experiment

- With matching, we attempt to find the randomized experiment that is "hidden inside" the observational study



Designs by rawpixel.com / Freepik

"Table 1" for covariate balance

Table: Covariate balance after matching

Covariate	Mean treated	Mean controls	Standardized difference
Age	26.22	25.46	0.08
Education	10.23	10.23	0.00
Black	0.85	0.83	0.05
Hispanic	0.06	0.05	0.03
Married	0.20	0.20	0.00
No degree	0.70	0.70	0.00
Earnings 74	2040.29	2313.52	-0.05
Earnings 75	1484.92	1476.09	0.00

Why matching?

- Conceptual simplicity (comparing like with like while keeping the unit of analysis intact) (Rosenbaum and Silber 2001)

Why matching?

- Conceptual simplicity (comparing like with like while keeping the unit of analysis intact) (Rosenbaum and Silber 2001)
- Its adjustments are an interpolation instead of an extrapolation based on a parametric model (Imbens 2015)

Why matching?

- Conceptual simplicity (comparing like with like while keeping the unit of analysis intact) (Rosenbaum and Silber 2001)
- Its adjustments are an interpolation instead of an extrapolation based on a parametric model (Imbens 2015)
- It is conducted without using outcomes, thus preventing inappropriate manipulation of the data (Rubin 2008)

Why matching?

- Conceptual simplicity (comparing like with like while keeping the unit of analysis intact) (Rosenbaum and Silber 2001)
- Its adjustments are an interpolation instead of an extrapolation based on a parametric model (Imbens 2015)
- It is conducted without using outcomes, thus preventing inappropriate manipulation of the data (Rubin 2008)
- It can be used with varied strategies such as instrumental variables and discontinuity designs (Baiocchi et al. 2010, Keele et al. 2015)

Why matching?

- Conceptual simplicity (comparing like with like while keeping the unit of analysis intact) (Rosenbaum and Silber 2001)
- Its adjustments are an interpolation instead of an extrapolation based on a parametric model (Imbens 2015)
- It is conducted without using outcomes, thus preventing inappropriate manipulation of the data (Rubin 2008)
- It can be used with varied strategies such as instrumental variables and discontinuity designs (Baiocchi et al. 2010, Keele et al. 2015)
- Facilitates sensitivity analyses to biases due to unmeasured confounders (Rosenbaum 1987)

Some related methods

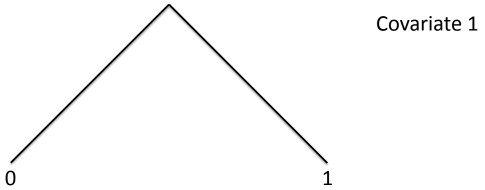
- Nearest neighbor matching (Rubin 1973; Abadie and Imbens 2006)
- Optimal matching (Rosenbaum 1989; Hansen 2004)
- Coarsened exact matching (Iacus et al. 2012)
- Genetic matching (Diamond and Sekhon 2013)
- Optimal matching with refined covariate balance (Pimentel et al. 2015)
- Covariate balancing propensity score (Imai and Ratkovic 2015)

Exact matching

- Ideally, we would match exactly for every covariate
- Impractical: 1 binary covariates renders 2 unit types

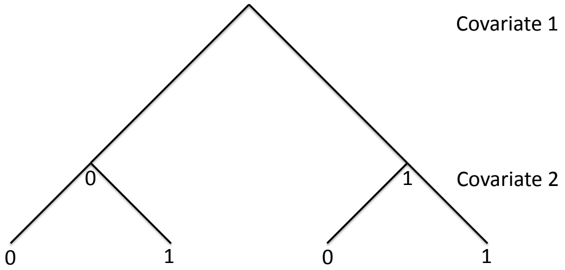
Exact matching

- Ideally, we would match exactly for every covariate
- Impractical: 1 binary covariates renders 2 unit types



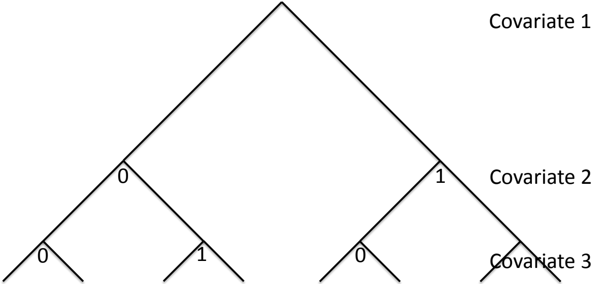
Exact matching

- Ideally, we would match exactly for every covariate
- Impractical: 2 binary covariates render 4 unit types



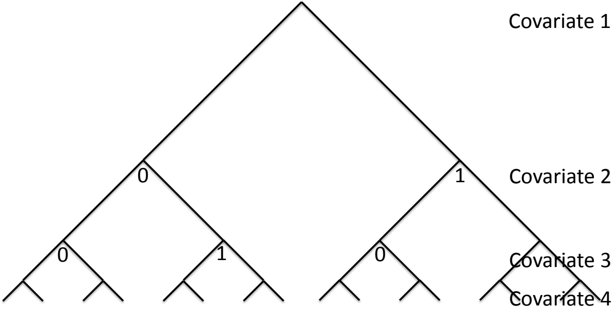
Exact matching

- Ideally, we would match exactly for every covariate
- Impractical: 3 binary covariates render 8 unit types



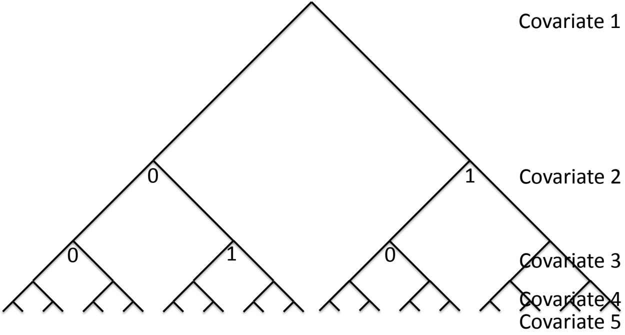
Exact matching

- Ideally, we would match exactly for every covariate
- Impractical: 4 binary covariates render 16 unit types



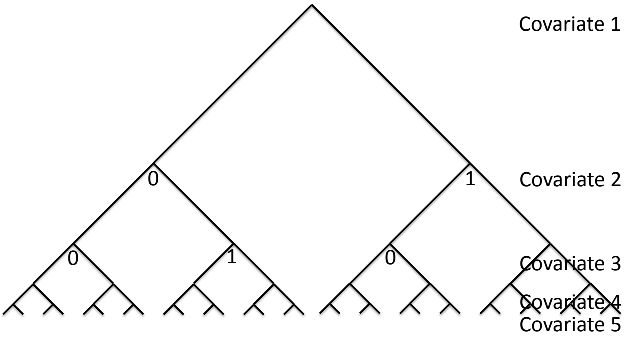
Exact matching

- Ideally, we would match exactly for every covariate
- Impractical: 5 binary covariates render 32 unit types



Exact matching

- Ideally, we would match exactly for every covariate
- Impractical: 20 binary covariates render over a million unit types



Balance on aggregate

- Randomization produces covariate balance, not perfect matches
- A fundamental tool for constructing matched sets is the propensity score, proposed by Rosenbaum and Rubin (1983)

The propensity score

- The propensity score is the conditional probability of treatment assignment given the observed covariates
  - $e(\mathbf{x}) = \Pr(Z = 1|\mathbf{x})$
- Informally, theorems 1 and 3 in Rosenbaum and Rubin (1983) state that
  - Matching on the propensity score tends to balance the P observed covariates used to estimate the score
  - For balancing the P covariates it suffices to balance the one-dimensional propensity score

Distances based on the propensity score

- Two common distances based on the propensity score are
  - $\delta_{t,c} = |e_t - e_c|$
  - $\delta_{t,c} = |\mathrm{logit}(e_t) - \mathrm{logit}(e_c)|$

Distances based on the propensity score

- Two common distances based on the propensity score are
  - $\delta_{t,c} = |e_t - e_c|$
  - $\delta_{t,c} = |\mathrm{logit}(e_t) - \mathrm{logit}(e_c)|$
- A more robust distance is the rank-based Mahalanobis distance with a caliper for penalty violations on the propensity score

$$\delta_{t,c} = \begin{cases} |\tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_c|' A \hat{\hat{\Sigma}}^{-1} A |\tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_c| & \text{if } |\mathrm{logit}(e_t) - \mathrm{logit}(e_c)| \leq c \\ \infty & \text{if } |\mathrm{logit}(e_t) - \mathrm{logit}(e_c)| > c \end{cases}$$

Nearest neighbor matching

- In its most basic form, this algorithm
  - First sorts the treated units in terms of the estimated propensity score (from highest to lowest, lowest to highest, or randomly)

Nearest neighbor matching

- In its most basic form, this algorithm
  - First sorts the treated units in terms of the estimated propensity score (from highest to lowest, lowest to highest, or randomly)
  - Then matches the first treated unit to the closest available control, making it no longer available for matching, and so on

Optimal matching

- Technically, nearest neighbor matching is a "greedy" algorithm in that it finds the best available control for each treated unit one at a time without considering global optimum

Optimal matching

- Technically, nearest neighbor matching is a “greedy” algorithm in that it finds the best available control for each treated unit one at a time without considering global optimum
- Instead, optimal matching finds the assignment of treated and control units that minimizes the global optimum, usually the total sum of covariate distances

Difficulties with propensity score matching...

- Theorems 1 and 3 above are stochastic properties; we don't have these guarantees for a given data set

Difficulties with propensity score matching...

- Theorems 1 and 3 above are stochastic properties; we don't have these guarantees for a given data set
- Finding an optimal match in terms of covariate distances does not imply that the matched groups will be balanced

Difficulties with propensity score matching...

- Theorems 1 and 3 above are stochastic properties; we don't have these guarantees for a given data set
- Finding an optimal match in terms of covariate distances does not imply that the matched groups will be balanced
- In practice this can involve a considerable amount of iteration or guesswork just to balance means

A general matching framework (Z. 2012; Z., et al. 2014; Kilcioglu and Z. 2016; Visconti and Z. 2018; Wang and Z. 2022)

- We solve

$$\min_{\boldsymbol{m}}\{\mathbb{D}(\boldsymbol{m}) - \lambda\mathbb{I}(\boldsymbol{m}) : \boldsymbol{m} \in \mathcal{M} \cap \mathcal{B} \cap \mathcal{R}\}$$

  where:
  - $\mathbb{D}(\boldsymbol{m})$ is the total sum of covariate distances between the matched groups
  - $\mathbb{I}(\boldsymbol{m})$ is the information content of the matched sample
  - λ is a scalar chosen by the investigator
  - $\mathcal{M}$, $\mathcal{B}$ and $\mathcal{R}$ are matching, balancing and representativeness constraints, respectively

Optimal matching (Rosenbaum 1989; Hansen 2004)

- We solve

$$\min_{\boldsymbol{m}}\{\mathbb{D}(\boldsymbol{m})-\lambda\mathbb{I}(\boldsymbol{m}) : \boldsymbol{m} \in \mathcal{M}\cap\mathcal{B}\cap\mathcal{R}\}$$

where:

  - $\mathbb{D}(\boldsymbol{m})$ is the total sum of covariate distances between the matched groups
  - $\mathbb{I}(\boldsymbol{m})$ is the information content of the matched sample (typically, the number of matched pairs)
  - λ is a scalar chosen by the investigator
  - $\mathcal{M}$, $\mathcal{B}$ and $\mathcal{R}$ are matching, balancing and representativeness constraints, respectively

Cardinality matching (Z. et al. 2014; Kilcioglu and Z. 2016; Visconti and Z. 2018; Niknam and Z. 2022)

- We solve

$$\min_{\boldsymbol{m}}\{\mathbb{D}(\boldsymbol{m}) - \lambda \mathbb{I}(\boldsymbol{m}) : \boldsymbol{m} \in \mathcal{M} \cap \mathcal{B} \cap \mathcal{R}\}$$

  where:
  - $\mathbb{D}(\boldsymbol{m})$ is the total sum of covariate distances between the matched groups
  - $\mathbb{I}(\boldsymbol{m})$ is the information content of the matched sample
  - λ is a scalar chosen by the investigator
  - $\mathcal{M}$, $\mathcal{B}$ and $\mathcal{R}$ are matching, balancing and representativeness constraints, respectively

Pseudo algorithm of standard matching methods

Algorithm 1 Matching with standard matching methods

Pseudo algorithm of standard matching methods

Algorithm 1 Matching with standard matching methods

0. Specify the covariate balance requirements (e.g., mean balance).

Pseudo algorithm of standard matching methods

Algorithm 1 Matching with standard matching methods

0. Specify the covariate balance requirements (e.g., mean balance).

Repeat:

Pseudo algorithm of standard matching methods

Algorithm 1 Matching with standard matching methods

0. Specify the covariate balance requirements (e.g., mean balance).

Repeat:

1. Estimate the propensity score or another summary of the covariates.

Pseudo algorithm of standard matching methods

Algorithm 1 Matching with standard matching methods

0. Specify the covariate balance requirements (e.g., mean balance).

Repeat:

1. Estimate the propensity score or another summary of the covariates.
2. Trim the extreme observations according to the summary measure.

Pseudo algorithm of standard matching methods

Algorithm 1 Matching with standard matching methods

0. Specify the covariate balance requirements (e.g., mean balance).

Repeat:

1. Estimate the propensity score or another summary of the covariates.
2. Trim the extreme observations according to the summary measure.
3. Match on the summary measure (e.g., using nearest neighbor matching).

Pseudo algorithm of standard matching methods

Algorithm 1 Matching with standard matching methods

0. Specify the covariate balance requirements (e.g., mean balance).

Repeat:

1. Estimate the propensity score or another summary of the covariates.
2. Trim the extreme observations according to the summary measure.
3. Match on the summary measure (e.g., using nearest neighbor matching).
4. Assess covariate balance.

Pseudo algorithm of standard matching methods

Algorithm 1 Matching with standard matching methods

0. Specify the covariate balance requirements (e.g., mean balance).

Repeat:

1. Estimate the propensity score or another summary of the covariates.
2. Trim the extreme observations according to the summary measure.
3. Match on the summary measure (e.g., using nearest neighbor matching).
4. Assess covariate balance.

Until:

Pseudo algorithm of standard matching methods

Algorithm 1 Matching with standard matching methods

0. Specify the covariate balance requirements (e.g., mean balance).

Repeat:

1. Estimate the propensity score or another summary of the covariates.
2. Trim the extreme observations according to the summary measure.
3. Match on the summary measure (e.g., using nearest neighbor matching).
4. Assess covariate balance.

Until:

The matched sample satisfies the covariate balance requirements.

Pseudo algorithm of cardinality matching

Algorithm 2 Matching with cardinality matching

Pseudo algorithm of cardinality matching

Algorithm 2 Matching with cardinality matching

0. Specify the covariate balance requirements (e.g., mean balance).

Pseudo algorithm of cardinality matching

Algorithm 2 Matching with cardinality matching

0. Specify the covariate balance requirements (e.g., mean balance).
1. Find the largest matched sample that satisfies the covariate balance requirements.

Pseudo algorithm of cardinality matching

Algorithm 2 Matching with cardinality matching

0. Specify the covariate balance requirements (e.g., mean balance).
1. Find the largest matched sample that satisfies the covariate balance requirements.
2. Rematch the balanced matched sample to minimize the covariate distances.

Remarks

- Cochran (1965) gives two basic pieces of advice for designing observational studies

Remarks

- Cochran (1965) gives two basic pieces of advice for designing observational studies
  - (i) "when selecting samples for study, make sure that they are large enough and have complete enough data to allow effects of practical importance to be estimated, and avoid treatment and control groups with large initial differences on confounding variables;"

Remarks

- Cochran (1965) gives two basic pieces of advice for designing observational studies
  - (i) "when selecting samples for study, make sure that they are large enough and have complete enough data to allow effects of practical importance to be estimated, and avoid treatment and control groups with large initial differences on confounding variables;"
  - (ii) "use both the statistician and the subject-matter expert in the planning stages."

Overview of forms of covariate balance

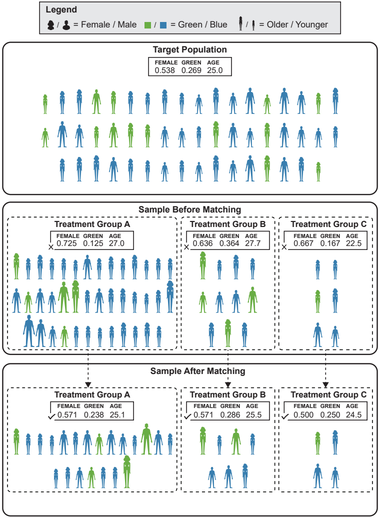
- From unit exact matching to aggregate mean balance
  - Exact matching
  - Near-exact (or almost exact) matching
  - Distributional balance
    - Joint distributions, marginal distributions
      The middle ground (say, all the two-way interactions of covariates)
  - Moments balance
    - Means, variances, skewnesses...

Using designmatch in R

```
>
>
>
>
>
>
>
> out = cardmatch(t_ind = t_ind, mom = mom,
solver)
   Building the matching problem...
   Gurobi optimizer is open...
   Finding the optimal matches...
   Optimal matches found
>
> # Number of matched groups
> length(out$t_id_1)
[1] 0
>
> # Assess mean balance
> tab = cbind(meantab(mom_covs, t_ind, t_id
4:6], round(mom_targets, 2))
> colnames(tab)[4] = "Target"
> tab
           Mean T  Mean C Std Dif  Target
age         26.22   25.46    0.08   25.82
education   10.23   10.23    0.00   10.35
black        0.85    0.83    0.05    0.84
hispanic     0.06    0.05    0.03    0.06
married      0.20    0.20    0.00    0.19
nodegree     0.70    0.70    0.00    0.71
re74      2040.29 2313.52   -0.05 2095.57
re75      1484.92 1476.09    0.00 1532.06
>
```

```
# Treatment indicator
t_ind = treatment

# Moment balance: constrain differences in means to be at most .1 st
mom_covs = cbind(age, education, black, hispanic, married, nodegree,
mom_tols = round(absstddif(mom_covs, t_ind, .05), 2)
mom_targets = apply(mom_covs[t_ind==1, ], 2, mean)
mom = list(covs = mom_covs, tols = mom_tols, targets = mom_targets)

# Solver options
t_max = 60*5
solver = "gurobi"
approximate = 0
solver = list(name = solver, t_max = t_max, approximate = approximat

# Match
out = cardmatch(t_ind = t_ind, mom = mom, solver = solver)

# Indices of the treated units and matched controls
t_id_1 = out$t_id
c_id_1 = out$c_id

# Time
out$time/60

# Number of matched groups
length(out$t_id_1)

# Assess mean balance
tab = cbind(meantab(mom_covs, t_ind, t_id_1, c_id_1)[, 4:6], round(m
colnames(tab)[4] = "Target"
tab

#################################
# Step 2 of cardinality matching: use optimal matching (minimum dist
# the matched treated and control units in order to minimizes the to
# between matched pairs.  For this, use the function 'distmatch' whi
#################################

# New treatment indicator
t_ind_2 = t_ind[c(t_id_1, c_id_1)]
```

Profile matching for a target population [Cohn and Z., 2022; *Epidemiology*]

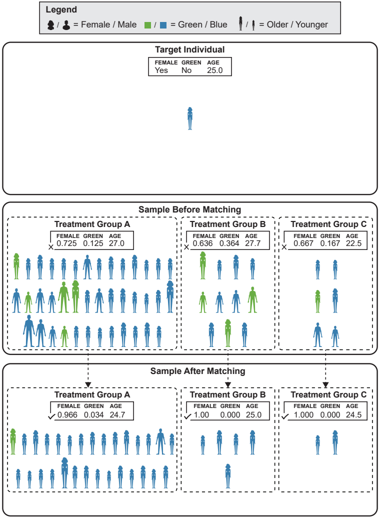


Illustrated by Xavier Alemañy

Profile matching with finite resolution [Cohn and Z., 2022; *Epidemiology*]

Illustrated by Xavier Alemañy

Profile matching for a target individual [Cohn and Z., 2022; *Epidemiology*]



Illustrated by Xavier Alemañy

Matching and weighting

- Matching can be viewed as a form of weighting that encodes an assignment between units
- Study design and interpretability in matching, versus statistical efficiency and computational tractability in weighting
- Matching approaches can be assisted with regression models in the spirit of doubly robust estimation (Robins et al. 1994; see also Rubin 1979, Abadie and Imbens 2011)

Balanced weighting

- Solve:

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & \sum_{i:Z_i=z} f(w_i) \\ \text{subject to} \quad & \left| \sum_{i:Z_i=z} w_i B_k(\boldsymbol{X}_i) - B_k^{target} \right| \leq \delta_k, \; k = 1, ..., K, \end{aligned}$$

where f is a convex function of the weights, $B_k(\boldsymbol{X}_i), k = 1, ..., K$, are regular functions of the covariates, and δ_k are scalars

Balanced weighting

- Solve:

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & \sum_{i:Z_i=z} f(w_i) \\ \text{subject to} \quad & \left| \sum_{i:Z_i=z} w_i B_k(\boldsymbol{X}_i) - B_k^{target} \right| \leq \delta_k, \; k = 1, ..., K, \end{aligned}$$

  where f is a convex function of the weights, $B_k(\boldsymbol{X}_i), k = 1, ..., K$, are regular functions of the covariates, and δ_k are scalars
- Stable balancing weights: $f(x) = (x - 1/n_z)^2$ and $\delta_k \in \mathbb{R}_0^+$; implemented in the `sbw` package in `R` (Z. 2015; Wang and Z. 2019)

Balanced weighting

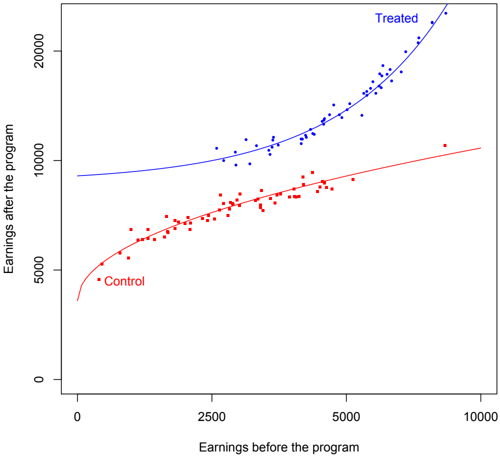
- Solve:

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & \sum_{i:Z_i=z} f(w_i) \\ \text{subject to} \quad & \left| \sum_{i:Z_i=z} w_i B_k(\boldsymbol{X}_i) - B_k^{target} \right| \leq \delta_k, \; k = 1, ..., K, \end{aligned}$$

  where f is a convex function of the weights, $B_k(\boldsymbol{X}_i), k = 1, ..., K$, are regular functions of the covariates, and δ_k are scalars
- Stable balancing weights: $f(x) = (x - 1/n_z)^2$ and $\delta_k \in \mathbb{R}_0^+$; implemented in the `sbw` package in R (Z. 2015; Wang and Z. 2019)
- See also Hainmueller (2012), Imai and Ratkovic (2014), Chan et al. (2016), Fan et al. (2016), Tan (2017), Zhao and Percival (2017), Athey et al. (2018), Hirshberg and Wager (2018), Zhao (2019)

Role of machine learning

- See Hill (2011) for details



Outline

Beyond association is not causation

- Can anything more be said about an observational study beyond *association is not causation*?

Sensitivity to hidden bias

- A sensitivity analysis is a statement about the *magnitude* of hidden bias that would need to be present to explain away a certain finding

Sensitivity to hidden bias

- A sensitivity analysis is a statement about the *magnitude* of hidden bias that would need to be present to explain away a certain finding
- Weak associations can be explained away by very small biases, but only a very large bias can explain a strong association in a large study

A model for sensitivity to hidden bias

- A study is free of hidden bias if the probability π_j that unit j gets the treatment is a function of the observed covariates describing the unit

A model for sensitivity to hidden bias

- A study is free of hidden bias if the probability π_j that unit j gets the treatment is a function of the observed covariates describing the unit
- There is *hidden bias* if two units with the same observed covariates $\boldsymbol{X}$ have different chances of assignment to treatment

A model for sensitivity to hidden bias

- A study is free of hidden bias if the probability π_j that unit j gets the treatment is a function of the observed covariates describing the unit
- There is *hidden bias* if two units with the same observed covariates $\boldsymbol{X}$ have different chances of assignment to treatment
- A sensitivity analysis asks: How would inferences about treatment effects be altered by hidden biases of various magnitudes?

A model for sensitivity to hidden bias

- A study is free of hidden bias if the probability π_j that unit j gets the treatment is a function of the observed covariates describing the unit
- There is *hidden bias* if two units with the same observed covariates $\boldsymbol{X}$ have different chances of assignment to treatment
- A sensitivity analysis asks: How would inferences about treatment effects be altered by hidden biases of various magnitudes?
- Suppose the π's differ at a given $\boldsymbol{X}$. How large would these differences have to be to alter the qualitative conclusions of a study?

Departure from from a study that is free of hidden bias

- Suppose we have units with the same $\boldsymbol{X}$ but possibly different π's, so $\boldsymbol{X}_i = \boldsymbol{X}_j$ but possibly $\pi_j \neq \pi_k$

Departure from from a study that is free of hidden bias

- Suppose we have units with the same $\boldsymbol{X}$ but possibly different π's, so $\boldsymbol{X}_i = \boldsymbol{X}_j$ but possibly $\pi_j \neq \pi_k$
- Then units j and k might be matched to form a matched pair to control overt bias due to $\boldsymbol{X}$

Departure from from a study that is free of hidden bias

- Suppose we have units with the same $\boldsymbol{X}$ but possibly different π's, so $\boldsymbol{X}_i = \boldsymbol{X}_j$ but possibly $\pi_j \neq \pi_k$
- Then units j and k might be matched to form a matched pair to control overt bias due to $\boldsymbol{X}$
- The odds that units j and k receive the treatment are $\frac{\pi_j}{1-\pi_j}$ and $\frac{\pi_k}{1-\pi_k}$ respectively and the odds ratio is the ratio of these odds

Departure from from a study that is free of hidden bias

- Suppose we have units with the same $\boldsymbol{X}$ but possibly different π's, so $\boldsymbol{X}_i = \boldsymbol{X}_j$ but possibly $\pi_j \neq \pi_k$
- Then units j and k might be matched to form a matched pair to control overt bias due to $\boldsymbol{X}$
- The odds that units j and k receive the treatment are $\frac{\pi_j}{1-\pi_j}$ and $\frac{\pi_k}{1-\pi_k}$ respectively and the odds ratio is the ratio of these odds
- Imagine that we knew that this odds ratio for units with the same $\boldsymbol{X}$ was at most some number $\Gamma \geq 1$

Departure from from a study that is free of hidden bias

- Suppose we have units with the same $\boldsymbol{X}$ but possibly different π's, so $\boldsymbol{X}_i = \boldsymbol{X}_j$ but possibly $\pi_j \neq \pi_k$
- Then units j and k might be matched to form a matched pair to control overt bias due to $\boldsymbol{X}$
- The odds that units j and k receive the treatment are $\frac{\pi_j}{1-\pi_j}$ and $\frac{\pi_k}{1-\pi_k}$ respectively and the odds ratio is the ratio of these odds
- Imagine that we knew that this odds ratio for units with the same $\boldsymbol{X}$ was at most some number $\Gamma \geq 1$

$$\frac{1}{\Gamma} \leq \frac{\pi_j(1-\pi_k)}{\pi_k(1-\pi_j)} \leq \Gamma \text{ for all } j, k \text{ with } \boldsymbol{X}_j = \boldsymbol{X}_k$$

Departure from from a study that is free of hidden bias

- Suppose we have units with the same $\boldsymbol{X}$ but possibly different π's, so $\boldsymbol{X}_i = \boldsymbol{X}_j$ but possibly $\pi_j \neq \pi_k$
- Then units j and k might be matched to form a matched pair to control overt bias due to $\boldsymbol{X}$
- The odds that units j and k receive the treatment are $\frac{\pi_j}{1-\pi_j}$ and $\frac{\pi_k}{1-\pi_k}$ respectively and the odds ratio is the ratio of these odds
- Imagine that we knew that this odds ratio for units with the same $\boldsymbol{X}$ was at most some number $\Gamma \geq 1$

$$\frac{1}{\Gamma} \leq \frac{\pi_j(1-\pi_k)}{\pi_k(1-\pi_j)} \leq \Gamma \text{ for all } j, k \text{ with } \boldsymbol{X}_j = \boldsymbol{X}_k$$

- If $\Gamma = 1$, then the study is free of hidden bias; for $\Gamma > 1$ there is hidden bias

Departure from from a study that is free of hidden bias

- Suppose we have units with the same $\boldsymbol{X}$ but possibly different π's, so $\boldsymbol{X}_i = \boldsymbol{X}_j$ but possibly $\pi_j \neq \pi_k$
- Then units j and k might be matched to form a matched pair to control overt bias due to $\boldsymbol{X}$
- The odds that units j and k receive the treatment are $\frac{\pi_j}{1-\pi_j}$ and $\frac{\pi_k}{1-\pi_k}$ respectively and the odds ratio is the ratio of these odds
- Imagine that we knew that this odds ratio for units with the same $\boldsymbol{X}$ was at most some number $\Gamma \geq 1$

$$\frac{1}{\Gamma} \leq \frac{\pi_j(1-\pi_k)}{\pi_k(1-\pi_j)} \leq \Gamma \text{ for all } j, k \text{ with } \boldsymbol{X}_j = \boldsymbol{X}_k$$

- If $\Gamma = 1$, then the study is free of hidden bias; for $\Gamma > 1$ there is hidden bias
- Γ is a measure of the degree of departure from a study that is free of hidden bias

Back to labor training example

- In our labor training example, we find the largest pair-matched sample that is balanced relative to the original treated group

Back to labor training example

- In our labor training example, we find the largest pair-matched sample that is balanced relative to the original treated group
- We obtain a point estimate equal to \$1886 with an associated *p*-value of 0.019

Back to labor training example

- In our labor training example, we find the largest pair-matched sample that is balanced relative to the original treated group
- We obtain a point estimate equal to \$1886 with an associated *p*-value of 0.019
- How large would need to be the influence of an unobserved covariate to explain away this significant effect estimate?

Back to labor training example

- In our labor training example, we find the largest pair-matched sample that is balanced relative to the original treated group
- We obtain a point estimate equal to \$1886 with an associated *p*-value of 0.019
- How large would need to be the influence of an unobserved covariate to explain away this significant effect estimate?

Γ	$p_{\min}$	$p_{\max}$
1	0.019	0.019
1.01	0.017	0.021
1.02	0.015	0.023
⋮	⋮	⋮
1.1	0.006	0.047
1.11	0.006	0.051

More on sensitivity analyses

- The first sensitivity analysis was proposed by Cornfield (1959)

More on sensitivity analyses

- The first sensitivity analysis was proposed by Cornfield (1959)
- Other approaches to sensitivity analyses include Rosenbaum and Rubin (1983), Lin et al. (1998), Robins et al. (2000), Gilbert et al. (2003), Imbens (2003), McCandless et al. (2007), VanderWeeele and Ding (2017)

More on sensitivity analyses

- The first sensitivity analysis was proposed by Cornfield (1959)
- Other approaches to sensitivity analyses include Rosenbaum and Rubin (1983), Lin et al. (1998), Robins et al. (2000), Gilbert et al. (2003), Imbens (2003), McCandless et al. (2007), VanderWeeele and Ding (2017)
- For extensions of the randomization-based inference approach to a weak null hypothesis see Fogarty (2019)

More on sensitivity analyses

- The first sensitivity analysis was proposed by Cornfield (1959)
- Other approaches to sensitivity analyses include Rosenbaum and Rubin (1983), Lin et al. (1998), Robins et al. (2000), Gilbert et al. (2003), Imbens (2003), McCandless et al. (2007), VanderWeeele and Ding (2017)
- For extensions of the randomization-based inference approach to a weak null hypothesis see Fogarty (2019)
- The sensitivity analyses methods discussed in this tutorial are implemented in the `R` packages `sensitivitymw` and `sensitivitymv` by Paul Rosenbaum

Outline

Two approaches that aim to control unobserved variation

- Under different assumptions, two devices to control for unobserved variation (balance unobserved covariates) are
  - Instruments
  - Discontinuities

Overview of instrumental variables

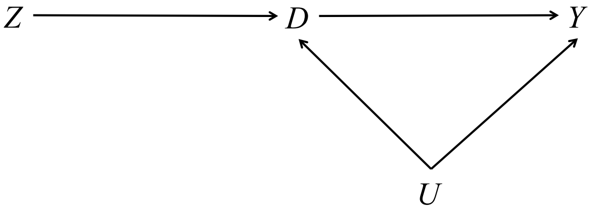
- What is an instrument?
  - A haphazard push to receive treatment which affects the outcome only through the treatment

Overview of instrumental variables

- What is an instrument?
  - A haphazard push to receive treatment which affects the outcome only through the treatment
- Main assumptions

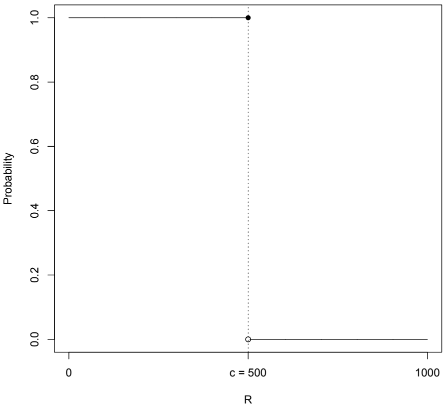
(R) The push is essentially random after adjusting for observed covariates

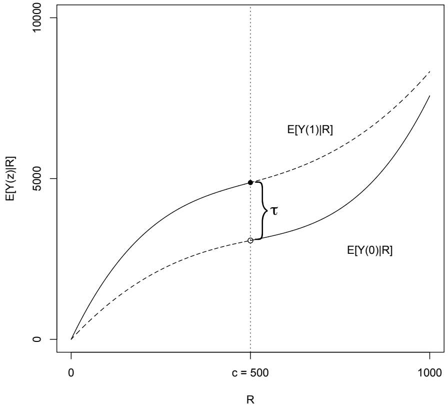
(E) The push affects the outcome only through the treatment (exclusion restriction)



Idea of a discontinuity design

- In our labor training example, imagine subjects with pre-treatment income below 500 are assigned to the program
- Pre-treatment income is the running variable R, 500 is the cutoff c





Outline

Some principles

- Focus on study design

Some principles

- Focus on study design
- Specify the target population
- State the key assumptions
- Avoid strong parametric specifications
- Characterize robustness of findings
- Be transparent

Matching for statistical control

Principle	Matching
Design-based	No outcomes
Target population	Covariate profile
Identification assumptions	Varied ones
Parametric specifications	Virtually none
Robustness	Easily assessed
Transparency	High emphasis

Causal Inference
Introduction

José R. Zubizarreta
Harvard University

09/03/2023
CUSO Doctoral School in Statistics and Applied Probability
Saignelégier, Switzerland